CLASSE DI SCIENZE
Corso di Perfezionamento in Fisica

Tesi di Perfezionamento

Deep Learning techniques for the observation of the Higgs boson decay to bottom quarks with the CMS experiment

Candidato:
Leonardo Giannini

Relatori:
Prof. Luigi Rolandi
Prof. Andrea Rizzi

Anno Accademico 2018/2019

Introduction

The Higgs boson was discovered at the CERN LHC by both the ATLAS and CMS collaborations in 2012 with a mass near 125 GeV. The characterization of the newly discovered particle has been one of the principal goals of the LHC experiments since. The main result reported here marks an important step in the effort of characterizing the Higgs boson: this thesis describes the first observation of the $H\to b\bar{b}$ decay with CMS data.

The measurement of the $H\to b\bar{b}$ decay mode probes directly the Yukawa coupling of the Higgs boson to down-type quarks. Consequently, it is a fundamental test of the mechanism that generates the masses of the fermions, and of the consistency of the Higgs properties with the Standard Model hypothesis.
The $H\to b\bar{b}$ decay observation comes after the bosonic decay modes of the Higgs, $\gamma\gamma$, ZZ, and WW, and the fermionic decay into $\tau\tau$ were all firmly established. The $t\bar{t}H$ production mode has also been observed, thus probing directly the coupling to up-type quark. The $H\to b\bar{b}$ observation is therefore closing a chapter in the Higgs Physics at the LHC: with all the most accessible production and decay modes now observed, the focus is shifting to rare decay modes, precision measurements and differential cross-section measurements.

The analysis presented uses 2017 data collected by the CMS experiment at $\sqrt{s} = 13$ TeV, which corresponds to an integrated luminosity of 41.3 fb^{-1}. The vector boson associated production mode (VH) with 0, 1 or 2 charged leptons in the final state is targeted, as it's the most sensitive for the $H\to b\bar{b}$ decay. The Higgs boson signal is extracted via a likelihood fit and an excess of 3.3 standard deviations over the background-only hypothesis is measured (with 3.1 standard deviations expected), corresponding to a signal strength $\mu = 1.08 \pm 0.34$. The analysis is combined with previous results for $VH(b\bar{b})$, reaching 4.8 (4.9 expected) standard deviations for the $VH(b\bar{b})$ process, with a signal strength of 1.01 ± 0.22. In combination with results targeting different production modes, namely the VBF $H(b\bar{b})$ analysis using Run 1 data, the $t\bar{t}H(b\bar{b})$ with 2016 data and the inclusive search for $H(b\bar{b})$ in the boosted regime, a significance of 5.6 (5.5 expected) standard deviations is reached, corresponding to a signal strength of 1.04 ± 0.20.

The heavy usage of Deep Learning techniques that I largely developed in my Ph.D. work was a crucial element for the observation of the $H\to b\bar{b}$ decay. Four different deep neural networks have been used: for tagging b jets, which are the reconstructed objects originating from the $H\to b\bar{b}$ decay; for the calibration of their energy and momentum; for background classification in control regions; and for discriminating the signal from the backgrounds. Machine Learning already played an important role in the previous searches for $VH(b\bar{b})$, but with 2017 data Deep Learning was introduced and very quickly became fundamental.
Deep Learning techniques are now becoming more and more important at the LHC, not only at the analysis level, but also because they are starting to be an integral part of the reconstruction algorithms in CMS. Hence, an important part of the thesis is dedicated to Deep Learning techniques, and their application to the b jets is shown as a use case.

The thesis is structured as follows: the Standard Model framework, with a focus on the Higgs mechanism, is described in Chapter 1. Also, a summary of the most important results achieved at the LHC on the properties of the Higgs boson is given. Chapter 2 is dedicated to the experimental apparatus: after a description of the LHC machine, the most important features of the CMS detector are presented. The reconstruction of physics objects is performed in multiple steps: lower level objects' reconstruction is included in Chapter 2. Higher level objects which are then used in the analysis are described in Chapter 3.

Chapter 4 introduces Deep Learning concepts whose application is present both in Chapter 5 and Chapter 6.

Chapter 5 is dedicated exclusively to Deep Learning applications: the object under study are the b jets, which are produced at the LHC both in the H→$b\bar{b}$ decay and in a number of background processes. Two tasks are important in analyses with b jets in the final state: the correct reconstruction of the jet momentum and the ability to separate b jets from jets originating from gluons, light quarks and charm quarks. We usually talk about "b jet energy regression" for the calibration of the jet transverse momentum and "b tagging" for the discrimination of b jets from the other hadronic jets.
The b jet energy regression used in the VH($b\bar{b}$) analysis is described in detail. Subsequently, a Deep Learning based b tagging algorithm, which, unlike most of the tagging algorithms uses only reconstructed tracks but no reconstructed secondary vertices, is presented. The algorithm, called "DeepVertex", exploits the ability of Deep Neural Networks to learn from raw data, aiming to infer the secondary vertex properties from tracks and clusters of tracks in the hidden layers of the network.
My work focused on the development of the regression Deep Neural Network in parallel with the ETH group searching for di-Higgs production, then on the validation with data of the trained model for the VH($b\bar{b}$) analysis and potentially for the CMS collaboration. I also carried out the development and optimization of DeepVertex in simulation, which has now reached results useful for the entire CMS collaboration and is ready for deployment in data.

Chapter 6 covers the VH($b\bar{b}$) analysis with 2017 data and the combination with previous analyses. My first contribution to the analysis was the aforementioned b jet regression and its validation. Subsequently, I worked on the optimization and the inference of the Deep Neural Networks used in the multivariate analysis together with other members of the analysis team. The analysis relies heavily on Deep Learning, both for signal discrimination and to isolate background sources, thus improving the background modeling.

The outlook for H→$b\bar{b}$ and conclusions are in Chapter 7. In this last chapter a preview of the search for the Higgs boson decay into muons using the full Run 2 data collected by CMS is also presented. The H $\rightarrow \mu\mu$ decay is important as it's the most viable channel to probe the decay to the second generation of fermions at the LHC. I was involved in the search for H$\rightarrow \mu\mu$ in the VBF production channel, and in particular in the optimization of DNN discriminators, thanks to my previous experience. The analysis uses a similar strategy as VH($b\bar{b}$). Deep Learning techniques similar to the ones applied in VH($b\bar{b}$) turned out to be the best solution to maximize the sensitivity.

Contents

Chapter 1

The Standard Model and the Higgs Boson

The Standard Model of particle physics is the theoretical model that describes all known fundamental particles and the interactions among them. Among those, the Higgs boson was discovered last, in 2012, with a mass near 125 GeV, by the ATLAS and CMS collaborations. After the discovery of the Higgs boson, the LHC experiments are focusing both on the search for new physics and the measurement of the Higgs boson properties. The observation of the $H \to b\bar{b}$ decay, which is the focus of this thesis, constitutes an important step in the Higgs boson characterization effort, as it was the only way to probe directly the coupling of the Higgs to down-type quarks.
A brief overview of the Standard Model and of the state-of-the-art research on the Higgs boson properties are given in this chapter, to motivate the Higgs characterization effort and the analysis that led to the $H \to b\bar{b}$ observation. The characterization of the Higgs boson will continue for the entire lifetime of the LHC and at future colliders.

1.1 The Standard Model

The Standard Model (SM) is a renormalizable Quantum Field Theory which describes the interactions among fundamental matter components.
Three of the four known fundamental interactions are part of the SM: the electromagnetic interaction, the weak interaction and the strong interaction. The gravitational interaction is not included, but it is much weaker and is not expected to contribute to the physical processes currently investigated in high energy physics.
SM matter is constituted by a few fundamental particles spin-half particles, the fundamental fermions, described by spin-half fields. Spin-1 particles, or vector bosons, described by vector fields, are the mediators of the interactions. Finally, the Higgs boson, with spin 0, is described by a scalar field.

1.1.1 The fundamental particles

The particle content of the SM is represented in figure 1.1. The fundamental fermions can be divided into two groups: the quarks, which interact via the strong, electromagnetic and weak interactions, and the leptons, which don't interact via the strong force.

Ordinary matter is composed of one lepton with a negative unitary charge, the electron (e) and two types of quarks, the up-quark (u) with charge +2/3 and the down-quark (d) with charge -1/3, which form protons and neutrons. The electron, the neutral electron neutrino (ν_e), which is produced in nuclear β decays, together with the u and d quarks, are known as the first generation of fermions. For each of the first generation particles, two copies that

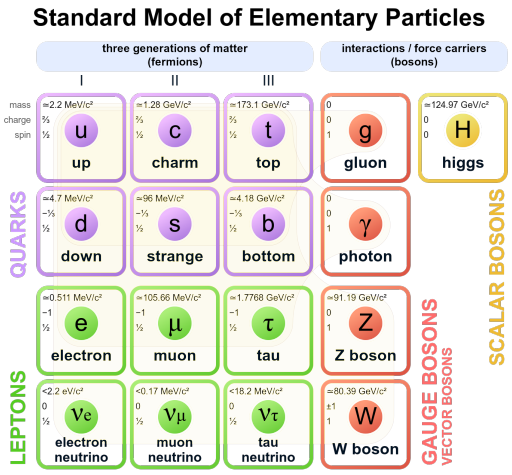


FIGURE 1.1: A pictorial representation of the SM particles [1].

have the same interactions, but larger masses [1], can be found in nature. These additional particles are known as the second and third generations. Antiparticles have exactly the same mass, but opposite quantum numbers. They were first predicted, then observed for each fermion.

The mediators of the three interactions, also known as gauge bosons, are the photon (γ), which is the mediator of the electromagnetic interaction; the gluons (g) for strong interaction, which, like the photon, are massless; and the three massive gauge boson which carry the weak interaction. These are the charged W^+ and W^-, both with a mass of ~ 80 GeV and opposite electric charges, and the neutral Z^0 boson, with a mass of ~ 91 GeV.

The Higgs boson is nowadays known to have a mass of ~ 125 GeV, it is electrically neutral and it can interact with all the fermions and gauge bosons.

1.1.2 Interactions and Gauge group

The strong, weak and electromagnetic interactions are introduced into the Quantum Field Theory framework via a local gauge invariance requirement. In Gauge theories, such as the SM, the Lagrangian that describes the field dynamics is required to have an internal symmetry under a Lie group, and the symmetry is local, i.e. space-time dependent. For each generator of the Lie group, a corresponding field called the gauge field must be introduced in the theory. The gauge symmetry also fixes the interactions of the matter fields, which are arranged into representations of the gauge group and assigned quantum numbers that fix their interaction properties.

In the SM, the description of the three fundamental interactions [2, 3, 4] is based on the local gauge symmetry group

$$U(1)_Y \otimes SU(2)_L \otimes SU(3)_c$$

$SU(3)_c$ is the color group, which describes the strong interaction or Quantum Chromodynamics (QCD). The quarks are a triplet under $SU(3)_c$, therefore the quarks exist in three

[1] For the neutrinos, the flavor eigenstates are not coincident with the mass ones. Experimental results require at least two neutrinos to have non-zero mass and a mass hierarchy is predicted. However, individual neutrino masses are yet to be measured and by convention neutrinos are described by flavor.

copies, with different color charges. The group has eight generators, which correspond to the eight gluon fields. However, free quarks cannot be observed in nature. Only colorless bound states, in the form of mesons and baryons can be observed. The leptons have no color charge, and they are singlets under $SU(3)_c$.

The $SU(2)_L$ together with $U(1)_Y$ describes the electromagnetic and weak interactions. $SU(2)_L$ has three generators, so three gauge fields are introduced into the Lagrangian. Spin-half fields behave differently under $SU(2)_L$ depending on their transformation properties under the Lorentz group: left-handed chirality fermions are all doublets under $SU(2)_L$, while right-handed chirality fermions are singlets. Specifically, the quark sector is described by two singlets (u_R, d_R) and one doublet ($q_L = \binom{u_L}{d_L}$) under $SU(2)_L$, while the lepton sector is described by one singlet (e_R) and one doublet ($\ell_L = \binom{e_L}{\nu_L}$). As already mentioned, each has three copies, one per generation. The right-handed neutrinos (ν_R) are not included in the SM description.

The $U(1)_Y$ hypercharge symmetry corresponds to a local phase invariance: the group has just one generator and the hypercharge quantum number defines the interaction properties for each field.

The particles' representations and their quantum numbers assignments are summarized in table 1.1.

	q_L	u_R	d_R	ℓ_L	e_R
$U(1)_Y$	1/6	2/3	-1/3	-1/2	-1
$SU(2)_L$	doublet	singlet	singlet	doublet	singlet
$SU(3)_c$	triplet	triplet	triplet	singlet	singlet

Table 1.1: Quantum numbers and representations associated to each generation of SM fermions

Once the symmetry and the quantum numbers of each matter field are chosen, the interactions are determined: the Lagrangian that describes the particle dynamics contains terms quadratic in the fields or their derivatives. The matter fields transform linearly under the gauge group, and in order to preserve the gauge local gauge invariance, all the field space-time derivatives are promoted to covariant derivatives: $\partial_\mu \rightarrow D_\mu$. Given the SM gauge structure, the covariant derivative will be:

$$D_\mu = \partial_\mu + ig' Y B_\mu(x) + igW^a_\mu(x)T^a + ig_S G^a_\mu(x)t^a$$

Where g', g and g_S are respectively the coupling of the hypercharge field B_μ, of the electroweak interactions, mediated by the three W^a_μ fields, and of the strong interaction, mediated by the eight G^a_μ fields. The vector boson fields, just like the fermion and scalar fields, are all functions of the space-time (the 4-vector x), as explicitly written in the above formula. The Y, T^a and t^a are the generators of the symmetry group in the representation required by the field.
The covariant derivative, when substituted in quadratic terms, produces the interaction vertices between fermions and vector bosons. This description of the interaction vertices, also called minimal coupling substitution, is predictive when used in perturbation theory.

Gauge invariant kinetic terms for the gauge bosons are also part of the Lagrangian, in the form:

$$\mathcal{L} = -\frac{1}{4} G^a_{\mu\nu} G^{a\mu\nu} - \frac{1}{4} W^a_{\mu\nu} W^{a\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu}$$

where the field strength tensors $G^a_{\mu\nu}$, $W^a_{\mu\nu}$ and $B^{\mu\nu}$ appear. The field strength tensor is defined e.g. for the weak gauge fields as:

$$W^a_{\mu\nu} = \partial_\mu W^a_\nu + \partial_\nu W^a_\mu - igf^{abc} W^b_\mu W^b_\nu$$

The third term appears only if the group corresponding to the gauge field is non-abelian; g is the coupling that is found also in the covariant derivative. In the SM the $SU(2)_L$ and $SU(3)_c$ are non-abelian, while the hypercharge group is abelian. The third term generates both cubic and quartic interaction vertices between gauge fields, for the strong and weak interaction gauge bosons.

The gauge bosons have no mass terms, as those would not be invariant under the gauge symmetry. A possible solution to give mass to the gauge bosons is the Higgs mechanism. This mechanism also clarifies the interplay of the $SU(2)_L$ and $U(1)_Y$ gauge bosons in the weak and electromagnetic interactions that we observe.

1.1.3 The Higgs mechanism

Mass terms for the gauge bosons can be introduced by a spontaneous symmetry breaking, while keeping the theory renormalizable. In the simplest case, spontaneous symmetry breaking requires the presence of a scalar field with a positive vacuum expectation value. In the SM case, the spontaneous symmetry breaking is realized via a complex scalar field which spontaneously breaks the $U(1)_Y \otimes SU(2)_L$ gauge symmetry [5, 6, 7, 8, 9, 10]. The $SU(3)_c$ symmetry remains unbroken.

A complex scalar field

$$\Phi(x) = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}$$

which behaves as doublet under $SU(2)_L$ and has charge +1/2 under $U(1)_Y$ is introduced.

The additional Lagrangian terms due to the scalar field will be :

$$\mathcal{L} = (D_\mu \Phi)^\dagger (D^\mu \Phi) + V(\Phi)$$

With a potential term like $V(\Phi) = V(\Phi^\dagger\Phi) = -\mu^2\Phi^\dagger\Phi + \lambda(\Phi^\dagger\Phi)^2$ the minimum of the potential, also called vacuum expectation value, will be $v = \left(\frac{\mu^2}{2\lambda}\right)^{1/2}$ and the Φ field, with appropriate gauge fixing, can be rewritten as

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + \mathrm{H}(x) \end{pmatrix}$$

A positive vacuum expectation value automatically fixes the masses of three gauge bosons: the Lagrangian terms containing the covariant derivative $(D_\mu\Phi)^\dagger(D^\mu\Phi)$, after making explicit the three generators of $SU(2)_L$ and the Y operator, can be rewritten as:

$$\frac{1}{2}\partial_\mu \mathrm{H}(x)\partial^\mu \mathrm{H}(x) + \frac{1}{8}(gW^3_\mu - g'B_\mu)^2 \cdot (v + \mathrm{H}(x))^2$$
$$+ \frac{1}{8}g^2(W^1_\mu - iW^2_\mu)(W^{1\mu} + iW^{2\mu}) \cdot (v + \mathrm{H}(x))^2$$

The first term is the kinetic term for the physical Higgs boson field H(x). The second term contains a linear combination of the of hypercharge gauge field B_μ and of the third gauge

field W^3_μ. The combination can be rewritten as the Z_μ with mass $M_Z = \frac{v^2(g^2+g'^2)}{4}$, describing the Z^0. The third term contains the weak gauge fields W^1_μ and W^2_μ, which are combined to give the W^+ and W^- fields. A mass $M_W = \frac{g^2v^2}{4}$ is therefore predicted for the charged W. The second and third term contain also the cubic and quartic interaction vertices of the Higgs with vector bosons. The couplings are proportional to the square of the vector boson masses.

The symmetry breaking pattern is $U(1)_Y \otimes SU(2)_L \rightarrow U(1)_{em}$. Three of the four gauge bosons acquire a positive mass, while the photon remains massless, as the $U(1)_{em}$ symmetry remains unbroken. As already seen, the Z_μ is a linear combination of the B_μ field and of the W^3_μ field. The orthogonal combination is the massless photon field A_μ, which is the mediator of the electromagnetic interaction. The photon doesn't have an interaction vertex with the Higgs boson, but the $H\rightarrow \gamma\gamma$ decay can happen via loops. The matter-photon coupling can be derived from the covariant derivative: the couplings are proportional to a common electric charge and the quantum numbers are given by the operator $Q = Y + T^3$, i.e. the sum hypercharge operator and the third generator of $SU(2)_L$s.

The physical Higgs field, H(x), describes a scalar particle with mass $m_H = \sqrt{2\lambda}\cdot v$, which can be derived from the potential term $V(\Phi)$. The Higgs mass is a free parameter of the model as it depends on the λ introduced by the Higgs potential, the vacuum expectation value is instead known to be $v = 246$ GeV because from the Fermi constant G_F, which depends in turn on M_W and g. The value of v can be extracted e.g. from the muon lifetime measurement, as the Fermi constant is $G_F = 1/\sqrt{2}v^2$.

Moreover, the presence of the Φ field allows mass terms proportional to the vacuum expectation value for the fermions, while preserving the local gauge invariance [11, 12]. Mass terms for the fermions are also not allowed in the Lagrangian, as they would take the form $m(\bar{\psi}_R\psi_L + \bar{\psi}_L\psi_R)$, which is not invariant under $SU(2)_L$. Gauge invariant mass terms are introduced via the interaction of the fermions with the Higgs field.

The mass terms for the leptons will look like:

$$\sim y_e\bar{e}_R\Phi^\dagger\ell_L + y_e\bar{\ell}_L\Phi e_R = y_e\frac{v+\mathrm{H(x)}}{\sqrt{2}}(\bar{e}_Re_L + \bar{e}_Le_R)$$

generating mass terms $y_e v\,/\,\sqrt{2}$, and analogous interaction terms with H(x) with coupling proportional to the lepton mass.
The Higgs boson gives mass both to the down type and up type quarks with Yukawa interactions. The mass terms for the down type quarks are totally analogous to the lepton masses, while in the mass terms for the up type quarks the conjugated field $\widetilde{\Phi} = \frac{1}{\sqrt{2}}\binom{v+H(x)}{0}$ is used. These terms take the form:

$$\sim y_u\bar{u}_R\widetilde{\Phi}^\dagger q_L + y_u\bar{q}_L\widetilde{\Phi}u_R = y_u\frac{v+\mathrm{H(x)}}{\sqrt{2}}(\bar{u}_Ru_L + \bar{u}_Lu_R)$$

As a result, the Higgs boson interacts with the entire fermion sector trough Yukawa-like vertices, with a coupling proportional to the fermion mass. It should be noted that the Yukawa couplings are actually matrices, but in the case of leptons the fields can be rearranged in such a way that weak interaction eigenstates are coincident with the mass eigenstates. In the case of the quarks, this is not possible as the up and down Yukawa matrices should be simultaneously diagonalized. As the quarks generations are defined as the mass eigenstates,

a unitary matrix describing the mixing of the mass eigenstates in the weak interactions mediated by W bosons becomes part of the interaction Lagrangian. This is the so-called Cabibbo-Kobayashi-Maskawa (CKM) matrix.

1.2 The Higgs Boson at the LHC

A Higgs-like particle, which is consistent with the SM Higgs according to the current experimental data, was discovered in 2012 at the LHC [13, 14, 15]. The mass of the new particle, approximately 125 GeV, sets the boundaries of today well-know Higgs phenomenology, as described in the next paragraph.

1.2.1 Higgs phenomenology: production and decay modes

In the SM, the phenomenology of the Higgs boson decays depends crucially on its mass, which defines the branching fractions. The production modes instead have a milder Higgs mass dependence.

In proton-proton collisions at the center-of-mass energies currently reached by the LHC (up to 13 TeV) four main production mechanisms are expected. The gluon-gluon fusion production mode has the largest cross section, followed by vector boson fusion, associated WH and ZH production, and production in association with a $t\bar{t}$ or $b\bar{b}$ pair [16]. The leading Feynman diagrams are reported in figures 1.2, 1.3. The total production cross section in proton-proton collisions depends on the center-of-mass energy. As the cross section is dominated by the gluon-gluon fusion, this production mechanism drives the increase in the cross section as a function of the center-of-mass energy ($\sqrt{s}$).

The gluon-gluon fusion (ggF) is the dominant production mode with a cross section of approximately ∼85% of the total. The leading diagram involves a quark loop: the main contribution to the SM amplitude arises from the top quark loop, though the amplitude is potentially sensitive to the presence of new massive particles with non zero color charge.

The vector boson fusion (VBF) has a cross section of about a tenth of the gluon-gluon fusion one. The leading diagrams involve a qq scattering in the t or in the u channel, with a vector boson exchange and the emission of a Higgs boson. Since the momentum exchange is typically lower than the center-of-mass energy of the two quarks, the channel is characterized by two separated high-rapidity quarks in the final state, detectable as high rapidity jets. Their presence can therefore serve as a signature of the VBF production channel. Additionally, as VBF is a pure electroweak process, low hadronic activity is expected in the rapidity gap between the two jets, where the Higgs decay products are typically found.

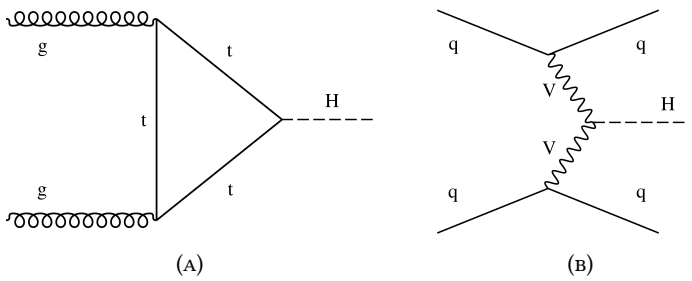


FIGURE 1.2: Leading Feynman diagrams for Higgs boson production via ggF (A) and VBF (B).

The Higgs-Strahlung (VH) has a slightly smaller cross section compared to VBF, but the presence of a vector boson in the final state helps to separate Higgs events from the background when the Higgs decays to two quarks. The presence of final state charged leptons or neutrinos is exploited in $H \to b\bar{b}$ searches.
The main contributions to the VH cross section come from quark initiated processes ($qq \to VH$, fig 1.3 A). A minor contribution to the ZH production comes from gluon initiated processes ($gg \to ZH$, fig. 1.3 B, C) whose contribution to the total ZH cross section is around 15%, but they can help to increase the sensitivity to high-p_T Higgs bosons.

Finally, the $t\bar{t}H$ associated production allows a direct measurement of Higgs coupling to the top quark. Its contribution to the total cross section is of about 1%. The final states are characterized by a relatively higher jet multiplicity due to the decay of the top quarks.

Other worth-mentioning processes are the $b\bar{b}H$ associated production, which is not target of direct searches and the single-top associated production, which is predicted to have a very low cross section. Both contributions are however taken into account with their expected cross-sections when measuring global Higgs properties.

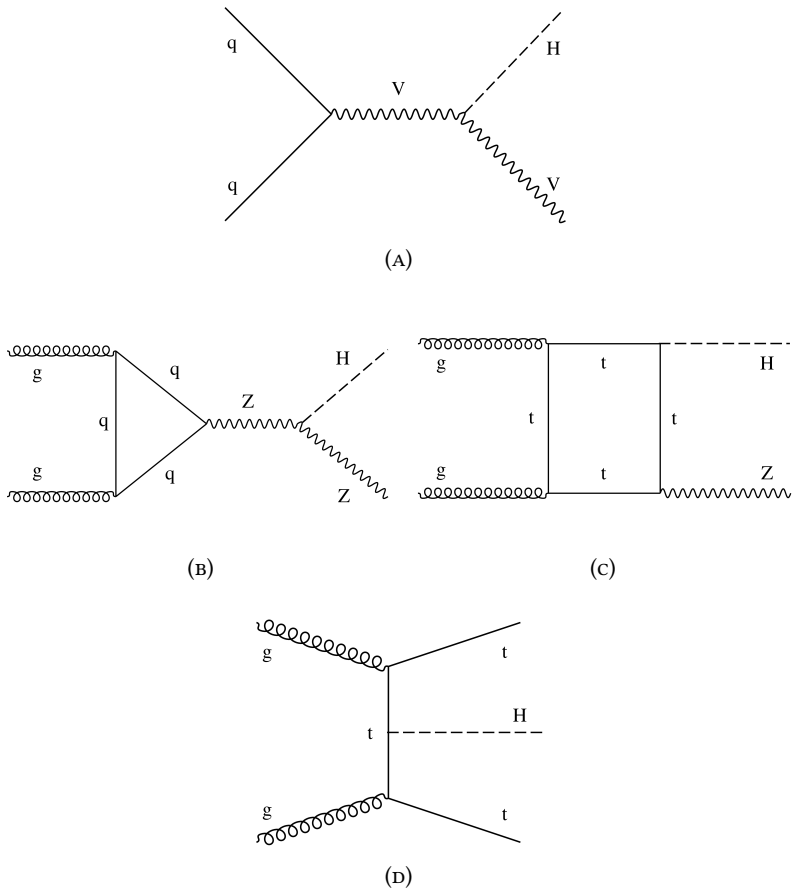
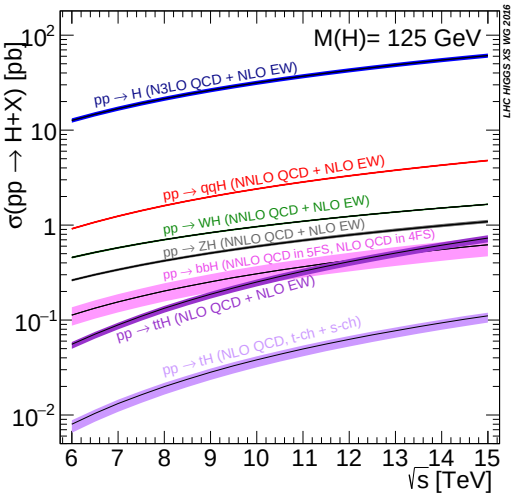


FIGURE 1.3: Leading diagrams for the VH production channel (A). Gluon initiated processes (B,C) are important when looking for a high-p_T Higgs decaying to hadrons; Example leading diagram for associated production with top quark pairs (D).

The most accurate predictions for the production cross sections in each mode are reported in figure 1.4. The accuracy level in perturbation theory both for corrections due to the

strong interactions, computed in perturbative QCD, and electroweak corrections is also reported in figure 1.4: the theoretical uncertainties affect mostly the ggF and $t\bar{t}H$ modes, due to the large QCD corrections. The central values recommended for the analysis at 13 TeV are listed in the adjacent table.



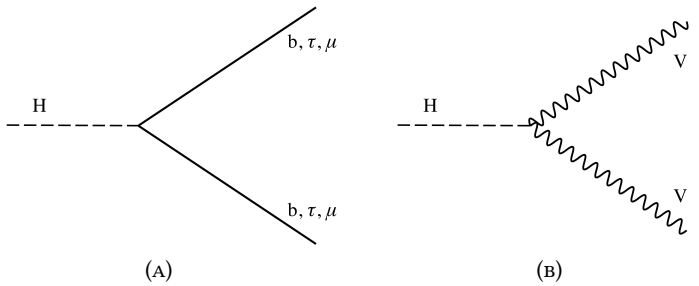
(A)

Production Mode	Cross Section [pb]
ggH	$48.58\ ^{+2.93}_{-3.77}$
VBF	3.78 ± 0.08
VH	2.26
W^+H	0.84 ± 0.02
W^-H	0.533 ± 0.011
ZH	$0.884\ ^{+0.036}_{-0.031}$
$t\bar{t}H$	$0.51\ ^{+0.03}_{-0.05}$

(B)

FIGURE 1.4: Production cross sections for $\sqrt{s}$ ranging from 7 to 14 TeV, for m_H = 125 GeV (A): the predicted central values for a Higgs boson of mass m_H = 125 GeV by production mode at $\sqrt{s}$ = 13 TeV are in (B). [17]

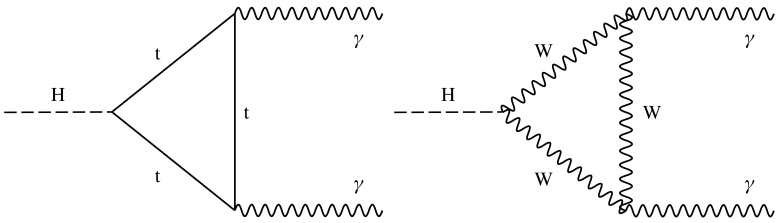
The Standard Model predicts the Higgs boson decay amplitude and its branching ratio in each final state. For a Higgs boson with a mass of approximately 125 GeV the total decay amplitude is expected to be of a ∼4 MeV. The Higgs boson decays into pairs of fermions through Yukawa-like interactions, with a relative branching ratio proportional to the fermion mass m_f at leading order, and into massive gauge boson pairs (figure 1.5) with couplings proportional to the square of the boson mass. Gluon-gluon and $\gamma\gamma$ final state are also possible via fermionic loops, or W loops in the $\gamma\gamma$ case (figure 1.6).



(A) (B)

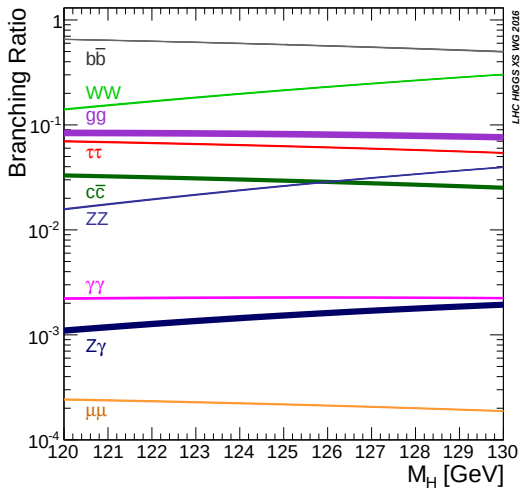
FIGURE 1.5: Higgs decay vertices at leading order.

The predicted branching ratios for a Higgs boson in the mass range 120-130 GeV are shown in figure 1.7 (A). The values for a Higgs mass of 125 GeV are listed in the adjacent table 1.7 (B).



(A) (B)

FIGURE 1.6: Higgs decays to photon pairs via loop diagrams.



Decay Channel	Branching Ratio [%]
$H\to b\bar{b}$	$58.24\ ^{+0.72}_{-0.74}$
$H\to \tau\tau$	6.27 ± 0.10
$H\to c\bar{c}$	$2.89\ ^{+0.16}_{-0.06}$
$H\to \mu\mu$	$(2.18 \pm 0.04)\cdot 10^{-2}$
$H\to WW$	$21.37\ ^{+0.03}_{-0.05}$
$H\to gg$	8.19 ± 0.42
$H\to ZZ$	2.62 ± 0.04
$H\to \gamma\gamma$	0.227 ± 0.005
$H\to Z\gamma$	0.153 ± 0.009

(A) (B)

FIGURE 1.7: Decay Branching Fractions for the Higgs boson mass in range 120–130 GeV (A). The SM predicted branching ratios for a Higgs boson of mass 125 GeV (B). [17]

1.2.2 Experimental tests

This thesis reports the observation of the H→ $b\bar{b}$ decay using CMS data [18]. This result, together with the ATLAS independent observation of the same decay, came in 2018, 6 years after the Higgs discovery in 2012. With this observation, a big chapter in the measurements of the Higgs boson properties, which started with the discovery and continued in Run 1 and throughout the LHC Run 2, has been closed. All the decay modes sought since the beginning of the LHC are now firmly established.

The LHC Run 1 was highlighted by the discovery of the Higgs via the H→ 4ℓ and H→ $\gamma\gamma$ decay channels and the detection of its mass, which was the missing parameter in the theory, near 125 GeV.

The decay to vector bosons were subsequently observed with a significance of at least 5 standard deviations (σ), both in H →ZZ→ 4ℓ and H→WW→ $2\ell 2\nu$ channels by the ATLAS and CMS collaborations [19, 20, 21, 22]. The H→ $\gamma\gamma$ channel was also observed with similar precision in Run 1 [23, 24].

The spin and parity of the Higgs boson were also tested exploiting the H→ $\gamma\gamma$, H→ZZ→ 4ℓ and H→WW→ $2\ell 2\nu$ channels [25, 26, 27]. The observations disfavor the spin-2 hypothesis and, assuming that the Higgs boson has spin zero, are consistent with the pure scalar hypothesis, $J^P = 0^+$, as predicted by the SM, while disfavoring the pseudoscalar hypothesis. The spin 1 hypothesis is excluded by the decay into photon pairs (Landau-Yang selection rules).

After the LHC Run 1, the ATLAS and CMS collaborations published two combination papers: the first with a combined mass measurement and the second with a combined measurement of the couplings. Overall, a good consistency between the data and the SM predictions was observed, just like in the input measurements performed by the collaborations in exclusive decay and production modes.

Given the larger significances of the observations of the decays to vector bosons, the consistency of the experimental data with the spontaneous symmetry breaking in the electroweak sector was one of the clear answers given by the LHC just with Run 1 data. On the other hand the consistency of Yukawa couplings was yet to be probed with similar accuracy. None of the fermionic decays had reached the 5σ threshold in Run 1, tough the data was consistent with the SM expectation in all channels.

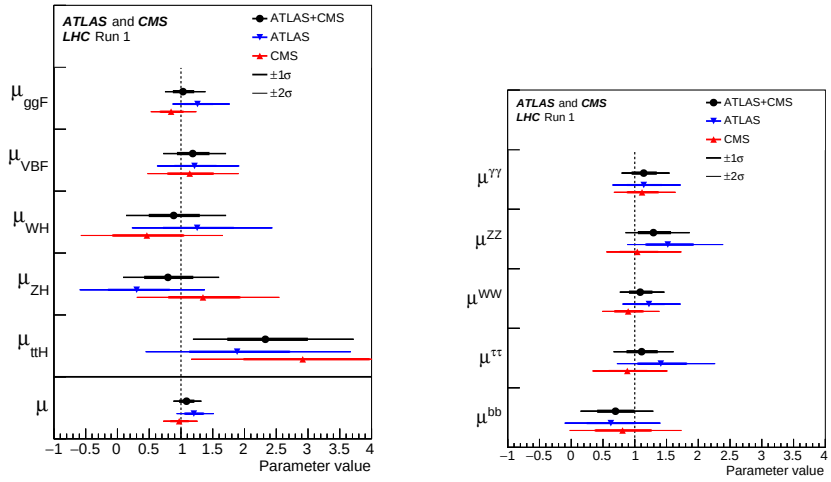
Run 2 data were fundamental to complete the picture, giving access to the main fermionic couplings.

Run 1 legacy papers: LHC combined results

The LHC Run 1 Higgs results are summarized in two combination papers using data from both ATLAS and CMS experiments [28, 29].

A combined mass measurement was performed exploiting the full Run 1 luminosity, which is of about 5 fb^{-1} at $\sqrt{s}$ = 7 TeV and of about 20 fb^{-1} at $\sqrt{s}$ = 8 TeV for both the ATLAS and CMS. Mass measurements are performed through the H→ $\gamma\gamma$ and H→ZZ→ 4ℓ decay channels, thanks to the good reconstructed mass resolution in the final states. The resulting combined mass is

$$m_H = 125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{syst.}) \text{ GeV}$$



(A) (B)

FIGURE 1.8: Best-fit results for the production signal strengths combining ATLAS and CMS measurements (A). Results for the branching ratio signal (B). The error bars indicate the 1σ and 2σ intervals.

The couplings to SM fermions and bosons were subsequently extracted assuming a Higgs mass of 125.09 GeV [29]. Individual analyses addressing specific decay modes and published separately by the two experiments were used as input for the combination.

Several parametrizations were used for the combined fits. The agreement between the SM prediction and the relative measurement can e.g. be described by the signal strength modifier μ. For each production mode i and decay channel f the production and decay signal strengths are defined as:

$$\mu_i = \frac{\sigma_i}{\sigma_{i,SM}} \qquad \text{and} \qquad \mu^f = \frac{BR^f}{BR^f_{SM}}$$

LHC analyses don't allow the disentanglement of production and decay modes, so the input of each analysis will be a signal strength modifier $\mu_i^f = \mu_i \cdot \mu^f$, which takes into account both contributions.

The global signal strength measurement, performed assuming the same μ_i and μ^f for each process, gives as a result a best-fit value of

$$\mu = 1.09^{+0.11}_{-0.10} = 1.09^{+0.7}_{-0.7}(\text{stat.})\,^{+0.4}_{-0.4}(\text{exp.})\,^{+0.7}_{-0.7}(\text{theory})$$

Analogous measurements are performed treating independently each production signal strength, assuming SM branching ratios, and each branching ratio signal strength, assuming SM production cross sections. The fit results are reported in figure 1.8.

An alternative parametrization uses coupling modifiers for each SM interaction vertex. The loop amplitudes can be assigned loop-specific coupling modifiers.
Alternatively fits with only the SM vertices modifiers can be performed. In this case the SM assumptions about the relative contribution of each coupling to the amplitudes are necessary. The coupling to each SM particle individually is tested assuming the relative consistency of the loop amplitudes ($\gamma\gamma$ final state, ggF production mode) with the SM.

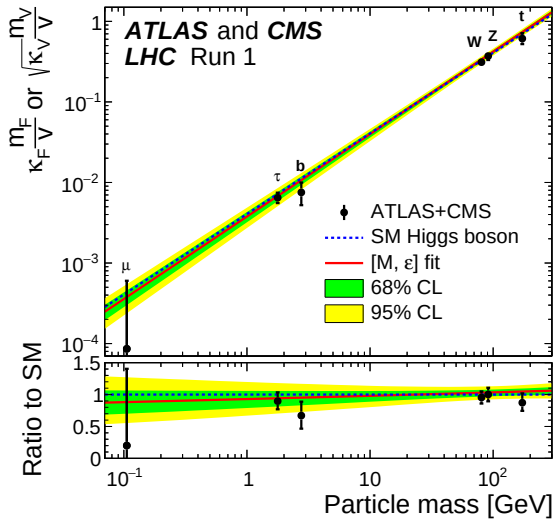


FIGURE 1.9: Best fit result for the coupling modifiers measurement for each SM particle; the dashed line indicates the expected value as function of the SM particle mass.

In the latter parametrization coupling modifiers k_i are such that for each vertex we have couplings:

$$y_i = \kappa_i \cdot y_{i,\mathrm{SM}}\,,$$

and combinations of the squared modifiers appear in the amplitude. With such a parametrization a fit that assesses the consistency of the data with the SM vertices as a function of the vacuum expectation value v is performed. The result is reported in figure 1.9, showing good consistency with the expectations for both the couplings to fermions, and to vector bosons with relatively better precision.

Combined fits with fewer assumptions are also performed: a fit treating all the signal strengths μ_i^f independently (23 parameters) and analogous fits to signal strengths ratio to the strength of the reference process ggF $H \to ZZ$. The results show overall a good consistency with the SM.

LHC Run 2: The quest for the couplings to fermions

The LHC Run 2 started in 2015 with a few proton-proton collisions data. Subsequently, unprecedentedly large amounts of data were collected: $\sim$36 fb^{-1} per experiment in 2016, $\sim$41 fb^{-1} in 2017, and more recently $\sim$60 fb^{-1} in 2018.

The larger production cross sections and the larger amount of integrated luminosity allowed also the observation of some of the decay and production modes, which were missing after Run 1. The Higgs $\to \tau\tau$ decay observation was reached with 2016 data by both CMS and ATLAS [30, 31] collaborations. Actually, the combination of the results from both experiments already yielded a 5.5σ significance with Run 1 data [29], but Run 2 data allowed the independent observation by each experiment. The $t\bar{t}H$ production mode was also observed both by CMS, with only 2016 data, and ATLAS soon afterwards [32, 33].

The observation of the $H\to \tau\tau$ decay and of the $t\bar{t}H$ production mode probed respectively the Yukawa coupling to charged leptons and to up-type quarks. Assuming the same mass

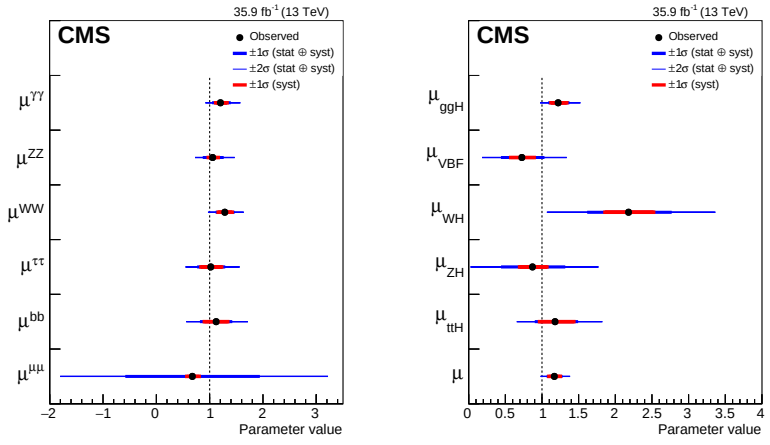
generation mechanism for the three generations of fermions, the only missing piece was the Higgs Yukawa coupling to down-type quarks, which could be measured via the H→ $b\bar{b}$ decay only.

The evidence for H→ $b\bar{b}$ was reached using the 2016 data [34, 35], but the channel was not observed before the analysis presented in this thesis. The observation of the H→ $b\bar{b}$ was achieved by both ATLAS and CMS in 2018 with 2017 data [36, 18]. The details of the CMS analysis and the techniques used are detailed in the chapters 5,6.
The most sensitive production mode for H→ $b\bar{b}$ is the VH, which exploits the leptonic decays of the vector boson and allowed the attainment of the observation.

Other important measurements performed with Run 2 data are in agreement with previous results. The mass was measured again by each experiment: ATLAS measured m_H = 124.97± 0.19(stat)±0.13(syst) GeV [37]. CMS used the ZZ→ 4ℓ decay mode: a mass of 125.26±0.20(stat)±0.08(syst) GeV, with reduced systematic uncertainty, was measured. [38].

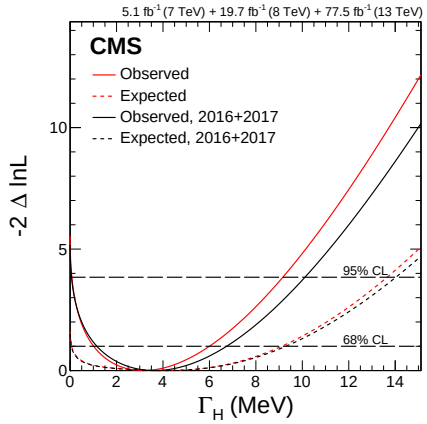
Analogously to the Higgs combination in [29], a combination of the coupling measurement was performed by each experiment using the data collected at 13 TeV [39]. The Combination performed by CMS uses 35.9 fb^{-1} collected in 2016, while a combination performed using ATLAS data uses different dataset depending on the specific channel. The results are compatible with Run 1 and with a SM Higgs boson in both cases, improving on the Run 1 combination. The measured signal strengths in the hypothesis of SM production modes and SM branching fractions are shown in figures 1.10 (A), (B).

The total width of the Higgs is also a challenging measurement at the LHC. The measurement can be performed comparing the on-shell and off-shell Higgs Boson production rates in the ZZ decay mode. The Run 2 CMS results [40] improve on the upper limit set in Run 1, and a lower limit is also set for the first time. The observed width is found to be in interval [0.08, 9.16] MeV ([0.0, 13.7] MeV expected) at 95% confidence level, as shown in figure 1.11.



(A) (B)

FIGURE 1.10: Summary plot showing the signal measured signal strength per decay channel at $\sqrt{s}$ = 13 TeV (A) and the measured signal strength per production mode at $\sqrt{s}$ = 13 TeV (b).



(A)

FIGURE 1.11: Likelihood scans Γ_H. The left plot (A) presents results both combining Run 1 and Run2 data (in red) and using Run 2 data only (black).

Beyond Run 2: Higgs precision measurements

After Run 2, with all the main decay modes and production channels firmly established, the physics of the Higgs boson at the LHC is entering a time of precision measurements and search for rare decays.

Precision measurements are motivated by the need to go beyond the SM. The last piece missing from the SM was the Higgs, which now appears to be consistent with the theoretical hypothesis. However, we already know that the SM is not able to describe all the experimental data. We have the experimental evidence of dark matter for a long time, but its presence is yet to be explained at the fundamental level. Neutrino masses, which are necessary given the experimental evidence of the neutrino oscillations are not explained with the Higgs mechanism. The mass of the Higgs boson itself, now that it has been found, opens the question of why it is close to the electroweak scale (the so-called "hierarchy problem"). Deviations from the Standard Model predictions could appear in precision measurements and hint at physics beyond the Standard Model.

An outlook for the most important Higgs properties measurements for the LHC Run 3 and beyond is given in appendix E.

Chapter 2

The CMS experiment at the LHC

The analysis described in this thesis uses data samples collected by the CMS detector during the LHC Run 2. This chapter is aimed at describing the main features of the LHC machine and of the CMS experiment with its subdetectors. The reconstruction of stable particles coming out of the collisions is also covered. Physics objects used at the analysis level are instead described in chapter 3.

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a 27 km long circular hadron accelerator and collider. Mainly protons, but also heavy ions are accelerated and collided. It was installed in the existing underground tunnel previously used for the Large Electron Positron collider (LEP), located at the border between Switzerland and France, at the CERN laboratories [41].

The LHC has a design center-of-mass energy ($\sqrt{s}$) of 14 TeV and instantaneous luminosity of 10^{34} cm^{-2} s^{-1}, representing a big step both in the energy and luminosity frontier compared to previous colliders.

The target luminosity of 10^{34} cm^{-2} s^{-1}, which requires high beam intensities, lead to the design of a particle-particle collider. For this reason, two separate rings with opposite magnetic optics host two counter-rotating particle beams. The beams are guided around the accelerator ring by a strong magnetic field (by design B_{max} = 8.33 T) maintained by 1232 superconducting dipole magnets. The maximum energy of the beams is limited by the maximum B field of the dipole magnets. In addition, a total of 392 quadrupole magnets are used to focus the beam.

The design of the magnets had to comply with the pre-existing LEP tunnel. The tunnel has eight straight sections and eight arcs, which would ideally be longer in a hadron collider to maximize the center-of-mass energy. Furthermore, an important feature inherited from the LEP tunnel design is the diameter of the tunnel arcs (3.7 m). As a consequence of the limited space, twin bore superconducting dipole magnets, also known as "two-in-one", were adopted. A cross section of the main LHC bending dipoles is shown in figure 2.1. The two beams are separated by 19.4 cm.

The arcs are instrumented with the main superconducting bending magnets, while the straight section host collision points with detectors and/or utilities: four collision points, with two aiming at the maximum luminosity; two beam injectors and two beam dump facilities; radiofrequency cavities and collimation systems.

Before injection into the LHC, the protons are accelerated in various steps that gradually increase their energy, as shown in figure 2.2. The chain starts with a linear accelerator

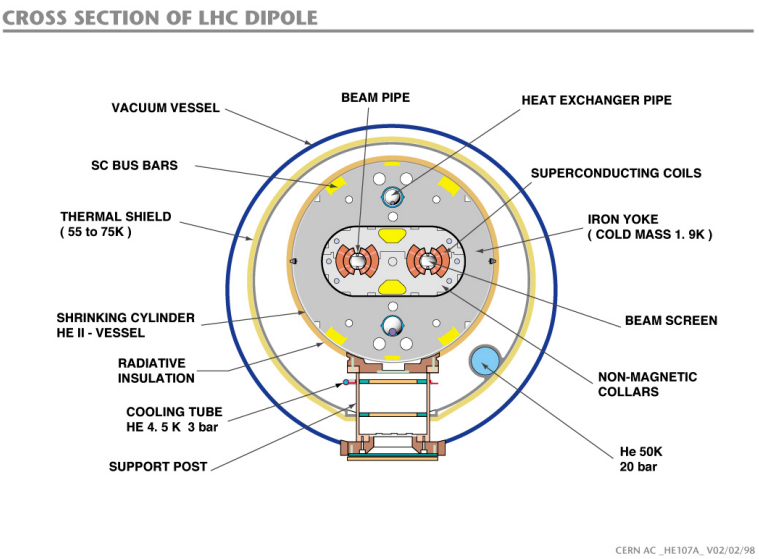


FIGURE 2.1: Cross section of a dipole magnet of the LHC. [42]

followed by three circular accelerators of increasing size. The protons are finally injected into the main ring with an energy of 450 GeV. Acceleration is achieved as the beam repeatedly crosses radiofrequency cavities. The magnetic field that guides the beams grows synchronously with the energy of the protons. Once the maximum field is reached, the beams are brought into collision at four interaction points.

The protons are accelerated in bunches made out of $\sim 10^{11}$ protons each. The bunches are spaced 25 ns in time in the nominal design, so collisions happen every 25 ns when the beams are stable. The bunches are $\sim$ 30 cm long, or 1 ns in time, and are squeezed in the transverse plane at the interaction points to a size of order ~ 10 μm.

The beams orbit around the LHC for about 12 hours in stable conditions. The beam intensity is gradually lost, primarily due to collisions. After the beams are partially depleted, they are dumped and the LHC is refilled. The duration of the LHC fill is chosen to maximize the luminosity integrated over time.

Four different experiments with different characteristics and purposes are located at the four interaction points, allowing a full exploitation of the physics potential of the LHC machine. ATLAS (A Toroidal LHC ApparatuS) and CMS (Compact Muon Solenoid) are general-purpose detectors designed to investigate a wide range of physics topics. Their focus includes the Higgs boson and the exploration of the energy frontier in a quest for new physics at the TeV scale. These are the two high luminosity experiments that receive the maximum luminosity delivered by the LHC. ALICE (A Large Ion Collider Experiment) is a heavy-ion experiment, designed to study the physics of the strong interaction at extremely high energy densities, in a phase of matter called quark-gluon plasma. The Large Hadron Collider beauty (LHCb) experiment specializes in the precise measurements of CP-violating observables in order to search for indirect evidence of new physics.

Other experiments designed to fully exploit the LHC collision are: TOTEM, now part of the CMS collaboration, whose detectors are placed symmetrically at $\sim$ 200 m on both sides of the CMS collision point along the beamline; LHCf made of two detectors similarly positioned 140 meters away from the ATLAS collision point; MoEDAL with detectors near

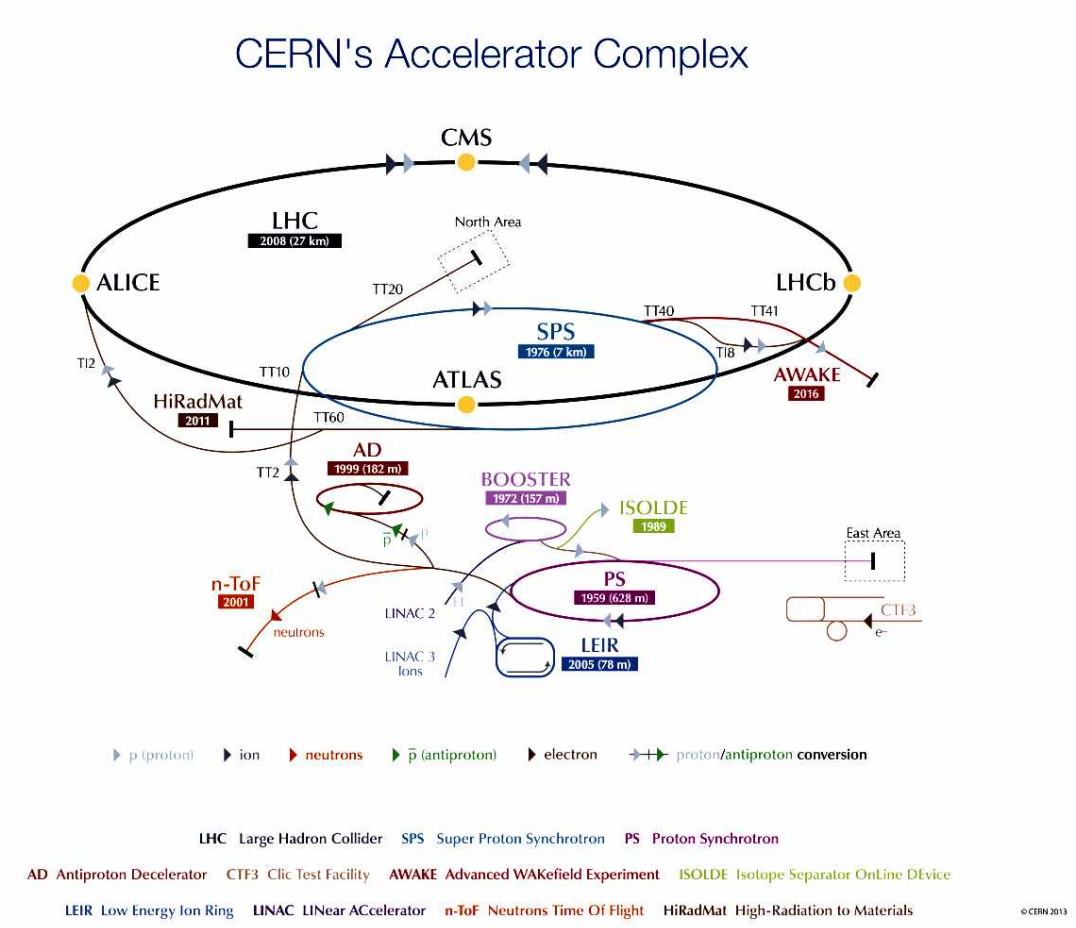


FIGURE 2.2: The CERN accelerator complex. [43]

LHCb to search magnetic monopoles.

Operational history

The LHC began its planned research program in the spring of 2010 with a center-of-mass energy of 7 TeV. By the end of 2011, the CMS experiment had collected a total integrated luminosity of 5.6 fb^{-1} with a record peak luminosity of 4.0×10^{33} cm^{-2} s^{-1}. In 2012, the center-of-mass energy was increased to 8 TeV and higher instantaneous luminosities were achieved. The total integrated luminosity collected by CMS during this year amounted to 22 fb^{-1} with a record peak luminosity of 7.7×10^{33} cm^{-2} s^{-1}.

In both years the LHC was operated with a bunch spacing of 50 ns corresponding to a collision frequency of 20 MHz. At the beginning of 2013, the LHC was shut down in order to prepare the collider to run at higher energy and luminosity. The accelerator was reactivated in early 2015, operating at a center-of-mass energy of 13 TeV.

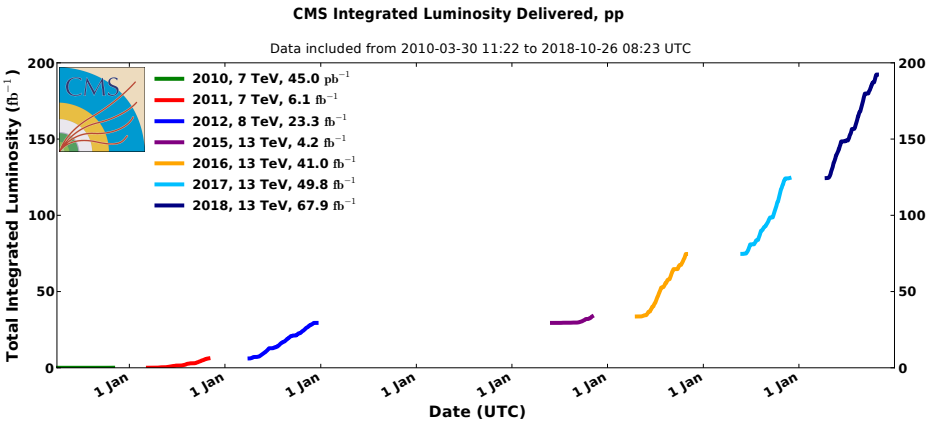
In 2015 the LHC reached a luminosity of 5×10^{33} $cm^{-2}s^{-1}$ and an integrated luminosity of about 4 fb^{-1}, with a bunch spacing of 25 ns, corresponding to a 40 MHz collision frequency.

During the years 2016-2018 the majority of the Run 2 data was delivered and collected, all with 40 MHz collision frequency. The LHC was successfully operated in proton-proton

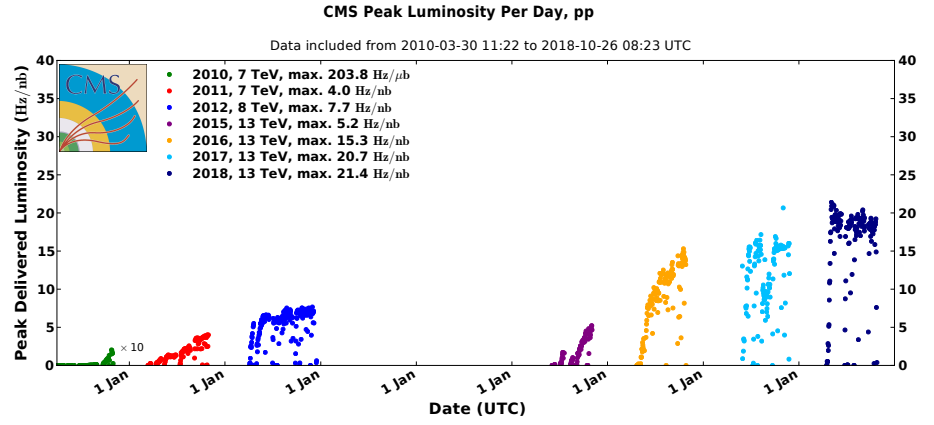
mode approximately from April to November each year, and increasingly higher instantaneous luminosities were achieved. The record luminosity was 1.5×10^{34} cm^{-2} s^{-1} in 2016, $\sim 2.1 \times 10^{34}$ cm^{-2} s^{-1} in 2017 and 2018, as measured by CMS. The design luminosity of 10^{34} cm^{-2} s^{-1} was achieved and exceeded since 2016. The integrated luminosities measured by CMS were 41 fb^{-1} in 2016, 49 fb^{-1} in 2017 and 68 fb^{-1} in 2018. Figure 2.3 shows the instantaneous (A) and integrated (B) luminosities delivered by LHC and measured by CMS since the LHC startup.

Both ATLAS and CMS were able to successfully collect and reconstruct the vast majority of the delivered luminosity, so each experiment has now $\sim$140 fb^{-1} of data that are "good for physics". The data successfully collected by both ATLAS and CMS are now almost one order of magnitude larger than the Run 1 data.

2018 marked the end of the LHC Run 2. A new shutdown phase is currently ongoing. The collider will again deliver new data from 2021 (Run 3), aiming at 300 fb^{-1} in the following three years [1].



(A)



(B)

FIGURE 2.3: Integrated (A) and instantaneous (B) luminosity recorded by the CMS experiment per year [44]. The instantaneous luminosity is quoted in Hz/nb equivalent i.e. 10^{33} cm^{-2} s^{-1} .

The high instantaneous luminosities are achieved by squeezing the proton bunches at the interaction point as much as possible in the transverse plane. The narrower the beam and

[1] Following the COVID-19, the plan for Run 3 is now to start in 2022 and collect 200 fb^{-1}.

the more protons in it, the higher the instantaneous luminosity. When the bunch cross one another multiple proton-proton collisions take place. The multiple collisions are usually referred to as "pileup" collisions. The distributions of the number of reconstructed interactions, or pileup profiles, are shown per year in figure 2.4. The higher luminosities are an advantage for physics, thanks to the higher rates of rare processes that are interesting. However the unavoidably larger pileup is an obstacle for the data taking and reconstruction.

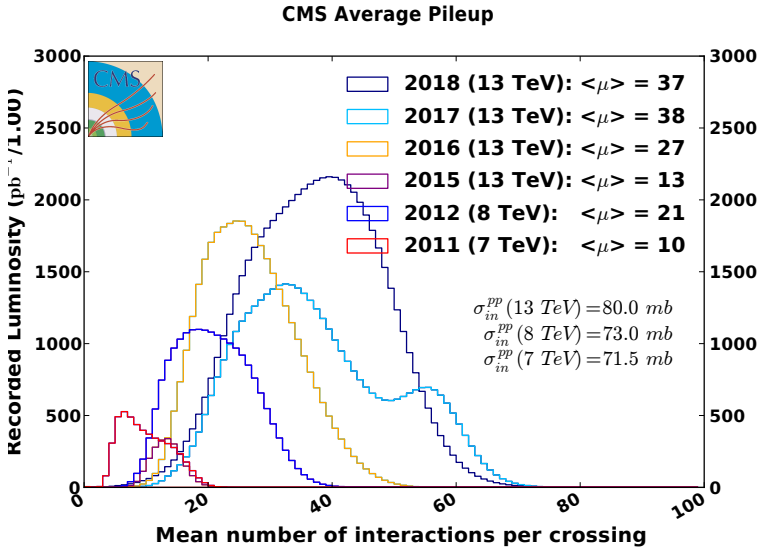


FIGURE 2.4: Distribution of the number of reconstructed pileup vertices measured by CMS in all the data taking years [44].

Reference Frame

The coordinate system used by the experiments at the LHC has its origin fixed at the nominal collision point. The x axis points towards the center of the LHC ring, the y axis points upwards and the z axis points along the counter-clockwise beam direction. The azimuthal angle ϕ is measured from the positive x direction in the xy plane and the polar angle θ is measured from the positive z direction. The coordinate r usually indicates the distance from the beam line ($r = \sqrt{x^2 + y^2}$)

In a typical collision, the center-of-mass of the interaction process is boosted along the z axis with respect to the laboratory frame. The kinematics of the collision products are therefore conveniently described by the coordinates (p_T , y, ϕ, m). Here, ϕ indicates the azimuthal angle, m the invariant particle mass, p_T the transverse momentum given by $p_T = p \sin\theta = \sqrt{p_x^2 + p_y^2}$, and y the rapidity defined as

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

The transverse momentum, the azimuthal angle and the mass are invariant under boosts along the z direction, while the rapidity is simply additive. The difference in rapidity between two particles is therefore invariant under boosts along the z direction.
The rapidity can be approximated for ultra-relativistic particles by the pseudo-rapidity

$$\eta = \frac{1}{2} \ln\left(\frac{|p| + p_z}{|p| - p_z}\right) = -\ln\left(\tan\frac{\theta}{2}\right) ,$$

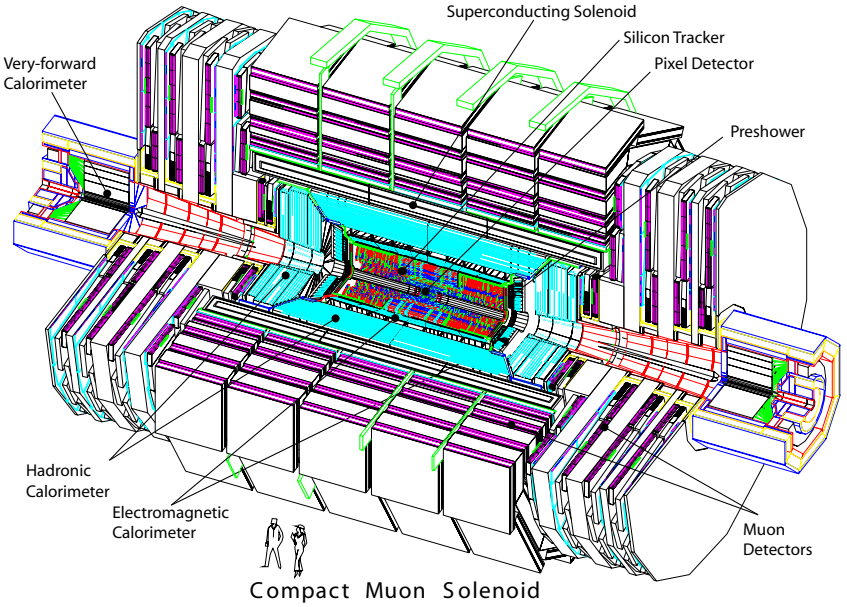


FIGURE 2.5: A three dimensional view of the CMS detector.

which is computed using just the polar angle θ.

2.2 The CMS Experiment

In the following sections a brief summary of the main features of the CMS detector is given; a detailed description can be found in the following references [45, 46].

The main feature of the detector is the central superconducting solenoid which provides a uniform magnetic field of 3.8 T along the z axis. The solenoid is 13 m long with a 6 m diameter. Several subdetectors are installed both inside and outside it. The solenoid contains from inside out, as shown in picture 2.5, the silicon tracker, the electromagnetic and the hadron calorimeter. Outside the magnet coil, the iron return yoke of the magnet hosts the muon spectrometer. The structure of the subdetectors consists of two regions: the barrel ($|\eta| \lesssim 1.2$), made of cylinder-shaped subdetectors positioned at increasing radii and the endcaps ($|\eta| \gtrsim 1.2$) where disk-shaped subdetectors are placed along the z axis, to ensure hermeticity. Forward sampling calorimeters extend the pseudorapidity coverage to high values ($|\eta| \sim 5$). Overall the CMS detector is 21.6 m long, has a diameter of 14.6 m, and weighs 12500 Tons in total.

Generally speaking, each subdetector is designed to perform a specific task and to identify and reconstruct a specific type of stable particle coming out of the collision. Muons are identified by the muon system, the energy of electrons and photons is measured by the electromagnetic calorimeter. Quarks and gluons undergo the "hadronization" process: they can be detected as a set of stable or almost stable hadrons, photons from π^0 decays, and other particles, which form a jet. Neutral hadrons in the jets can only be detected by the hadron calorimeter. All the charged particles, either isolated or in jets, are detected by the tracker with excellent spatial resolution. The magnet provides a large bending power, that

is fundamental to the reconstruction of the charged particles momenta. Neutrinos are detected as missing transverse energy.

Good performances at high particle flux and radiation-hardness are mandatory for the detectors and electronics. The very high collision rate and pileup are a challenge also for the data acquisition system, so a multi-level trigger system is necessary.

Several upgrades were performed to ensure optimal data taking through the end of Run 2. The three-layer pixel detector was replaced with a four-layer high-data-rate design; the first level of the trigger system (Level-1 Trigger) and the hadron calorimeter photo-detectors and electronics were also upgraded.
Additional upgrades are currently being applied in preparation for Run 3 and major changes are planned for the Phase 2 upgrade [47], after 2023, when the higher instantaneous luminosities and pileup rates will require the substitution of many subdetectors.

2.2.1 The Tracker

The tracker [48, 49] constitutes the inner part of CMS and is designed to provide a precise and efficient measurement of the charged particle tracks and of the primary and secondary interaction vertices. It is immersed in an almost homogeneous magnetic field of 3.8 T provided by the CMS solenoid.

The tracker has to be light, both in the active and dead material, in order not to alter the trajectories of charged particles and to provide the best possible resolution. It also has to be fast enough to take data every 25 ns (40 MHz). High granularity is necessary due to the high particle multiplicity. However, a fast and granular detector needs adequate numbers of readout channels and an efficient cooling, which result in dead material.

Different technologies are used to satisfy these requirements as best as possible: a silicon pixel detector is installed in the inner region, closest to the interaction point, while silicon microstrip detectors are used in the outer region. The total length of the tracker is 5.8 m and its diameter 2.5 m, and the angular coverage reaches up to $|\eta| = 2.5$. The layout of the original CMS tracker is shown in figure 2.5.

The pixel detector was changed with respect to the one in figure 2.6. The current pixel tracker was installed during the 2016 Technical stop [50]. It has four layers and three endcap disks, ensuring better redundancy and no performance loss at higher instantaneous luminosity with respect to the old one, which had three barrel layers and two endcap disks. The current pixel detector is planned to be used through the LHC Run 3 together with the original outer tracker.

The four barrel concentric cylindrical layers (BPIX) have a length of about 55 cm and radii between 2.9 cm and 16 cm. Compared to the old CMS pixel barrel, there is one new layer at high radius. The radius of the innermost layer is reduced by $\simeq$1 cm while layers 2 and 3 are almost unchanged.
The three disks in the endcaps (FPIX) are located at each end of the central barrel detector, with a radial coverage ranging from 4.5 to 16.1 cm. The position of the first disk along the beam line is 29.1 cm from the interaction point while the second and third disks are located at 39.6 cm and 51.6 cm from the interaction point. Together with the four barrel pixel layers, they provide a four-hit coverage for all tracks over the pseudorapidity region $|\eta| < 2.5$, as

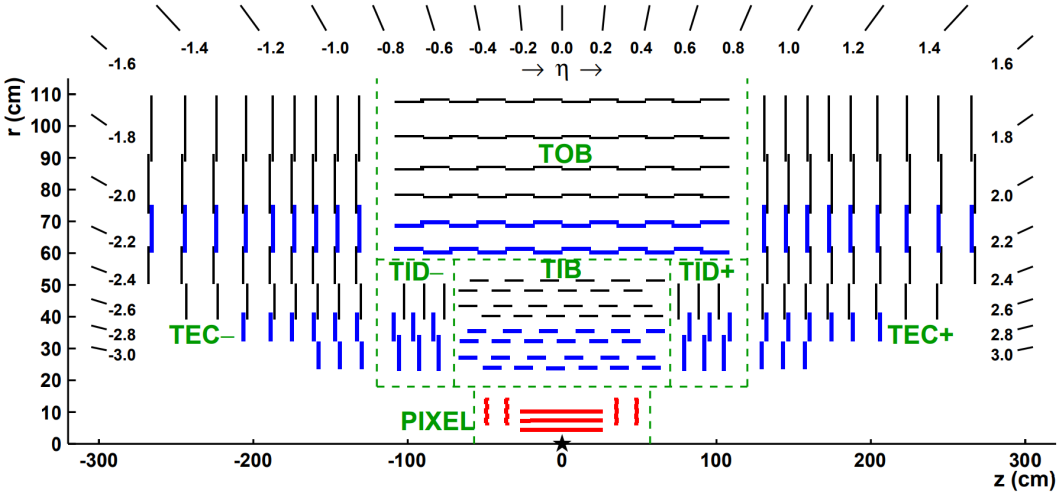
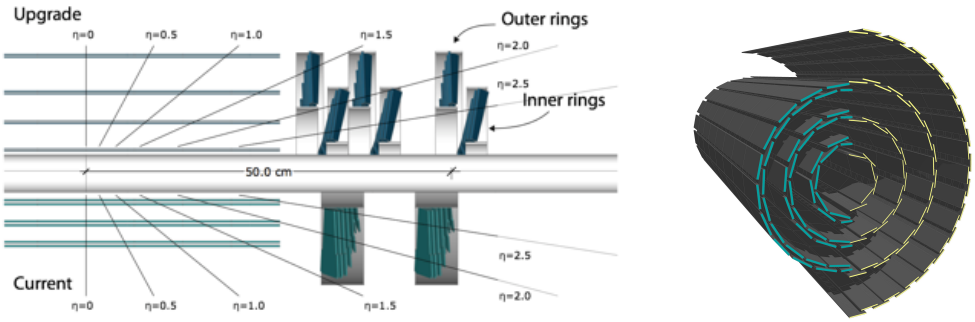


FIGURE 2.6: CMS original tracker layout.

shown in figure 2.7.

Each pixel has a surface of 100 × 150 μm^2 to obtain low cell occupancy (order 10 $^{-4}$ per pixel and collision) and a spatial resolution of about 10 μm in the transverse direction (ϕ for the barrel, with r given by the sensor position) and 15 μm in the longitudinal direction (z for the barrel).

The upgrade of the pixel detector featured also a reduction of the dead material in the tracker acceptance. The cooling tube diameter is significantly reduced, and the services of the BPIX are moved further out in the z direction, outside the active tracking volume.



(A) (B)

FIGURE 2.7: Comparison of the new pixel detector with the old one: $r \times z$ view (A) and comparison of the barrel cylinders (B).

The inner part of the silicon strip detector is made of 4 barrel layers (tracker inner barrel or TIB) positioned at radii ranging from 20 to 55 cm and 3 disks at each side (tracker inner disks or TIDs). The outer strip system is composed of 6 barrel layers positioned at radii up to 1.1 m (tracker outer barrel or TOB) and 9 disks for each endcap (tracker endcaps or TECs). All four regions (TIB, TID, TOB, TEC) are provided with both single-sided and double-sided microstrip modules. The double-sided modules are rotated with respect to the strip direction by a "stereo" angle of 100 μrad. They can therefore provide 3D measurements, though

with ∼10 times lower resolution along the strips.

The strips are oriented along the z axis in the barrel and along the r coordinate in the endcaps. The microstrip detector design spatial resolution is of about 20-50 μm in the transverse direction and about 200-500 μm along the strips for "stereo" hits.

2.2.2 The Calorimeters

The calorimeters are located outside the tracker and inside the magnetic solenoid. They are designed to measure the energy of electromagnetic and hadronic showers and, unlike the tracker, they are required to completely absorb the particles in the shower for optimal energy measurements. They are therefore required to be "heavy", which translates into a large number of radiation lengths X_0 for the ECAL and of interaction lengths λ_I for the HCAL. The full tracker material has a thickness of $\sim$ 1-2 X_0 and less than one λ_I for comparison.

The design of the magnet, whose radius is large enough to contain the tracker and both the ECAL and HCAL,s minimizes the amount of material encountered by particles before the calorimeters, thus improving the energy resolution.

ECAL

The electromagnetic calorimeter of CMS (ECAL) [51] is a homogeneous calorimeter made of 61200 $PbWO_4$ crystals mounted in the central barrel part, completed by 7324 crystals in each endcap. The ECAL barrel covers the central rapidity region ($|\eta| < 1.48$) and the two ECAL endcaps extend the coverage up to $|\eta| = 3$. A lead/silicon-strip preshower detector is also installed at pseudorapidities $1.6 < |\eta| < 2.6$. The crystals are all active material: they induce the shower and generate scintillation light to measure the shower energy. The scintillation light is detected by silicon avalanche photodiodes in the barrel region and vacuum phototriodes in the endcap region.

The use of the purpose-built high density crystals has allowed the design of a calorimeter which is compact, fast, has fine granularity, and is radiation resistant. The barrel crystals are 23 cm long, which corresponds to 26 X_0, while the endcap crystals are 22 cm long, for a total of (3+25) X_0 in the preshower+endcap. Electromagnetic showers of energies of 1 TeV are on average 98% contained both for electrons and photons. The length of the crystal corresponds also to 1 interaction length, therefore about one third of the charged hadrons start showering in the ECAL. The transverse dimensions of the crystals are equal to the Molière radius (2.2 cm) providing a very fine transverse granularity. The size is equivalent to 0.0174×0.0174 radians (1 degree) in the $\eta \times \phi$ plane for the barrel crystals. The endcap crystal transverse size is 3×3 cm^2.

The energy resolution of a calorimeter can be parametrized as:

$$\frac{\sigma_E}{E} = \frac{a}{E} \oplus \frac{b}{\sqrt{E}} \oplus c$$

where a is the noise term due the electronics and pileup, independently of the energy, b is the stochastic term which accounts mainly for the fluctuations in the photon conversions, and c is a constant term related to the energy scale calibration.

A typical measured $PbWO_4$ crystal energy resolution is of the order of

$$\frac{\sigma_E}{E} = \frac{0.12}{E/\text{GeV}} \oplus \frac{2.8\%}{\sqrt{E/\text{GeV}}} \oplus 0.3\%$$

The expected performances have been almost matched during the data taking.

HCAL

The CMS hadron calorimeter (HCAL) [52] is used to measure the energy of hadrons, and it is the only detector available to measure the energy of neutral hadrons. Its design ensures good hermeticity to allow the measurement of the missing transverse energy and angular coverage in the forward region for forward jets.
Four regions are instrumented with HCAL detectors: the barrel hadron calorimeter (HB) surrounds the electromagnetic calorimeter and covers the central pseudorapidity region up to $|\eta| = 1.3$; the two endcap hadron calorimeters (HE) cover up to $|\eta| = 3$. The Cherenkov calorimeter (HF) extends the coverage up to $|\eta| = 5$ in the forward region. An array of scintillators, the outer hadron calorimeter (HO), is located outside the magnet to catch the tails of the hadronic shower and avoid the misidentification of muons.

Contrary to ECAL, the HCAL is a sampling calorimeter: the energy is measured by scintillators alternated to brass plates used as absorbers in HB and HE. The presence of brass as an absorber guarantees the containment of the hadronic shower, with a thickness of at least six interaction lengths. Steel interlayered with a quartz-fiber Cherenkov calorimeter is used instead in the forward region. The HO detectors have no dedicated absorbers, but the shower is induced by the magnet and return yoke material.

The HCAL is coarser than the ECAL with modules of size 0.087×0.087 radians in the $\eta \times \phi$ plane for $|\eta| < 1.6$, corresponding to 5×5 ECAL crystals, and of size 0.17×0.17 for $|\eta| > 1.6$. The resolution is also worse: the combined ECAL+HCAL resolution measured in a pion test beam was $\sigma_E/E \simeq 110\%/\sqrt{E} \oplus 9\%$.

The HCAL photodetectors, readout system and electronics underwent an upgrade between Run 1 and Run 2 and through Run 2 [53]. The HF detector originally used photomultiplier tubes to collect the Cherenkov light, which are now replaced by multi-anode tubes to reduce the rate of anomalous signals; the HB and HE modules will use the recently installed silicon photomultipliers as photodetectors, allowing for an increase in the readout channels and better in-depth segmentation of the detector.

2.2.3 The Muon System

The muon system [54] is located outside the solenoid and covers the pseudorapidity range $|\eta| < 2.4$. Outside of the solenoid coil, the magnetic field flux is returned through a steel yoke. Three steel layers are present both in the barrel and in the endcaps, alternated with four layers of muon detectors, as shown in figure 2.8.
The muon system provides information to identify muons and to measure the momentum and charge of high-p_T muons. Additionally, two tasks are accomplished by the muon system thanks to its good time resolution: bunch crossing identification and muon-triggering.

Three different gaseous detector technologies are used: drift tube (DT) chambers and cathode strip chambers (CSC) detect muons in the regions $|\eta| < 1.2$ and $0.9 < |\eta| < 2.4$, respectively. They are supplemented by a system of resistive plate chambers (RPC) covering the

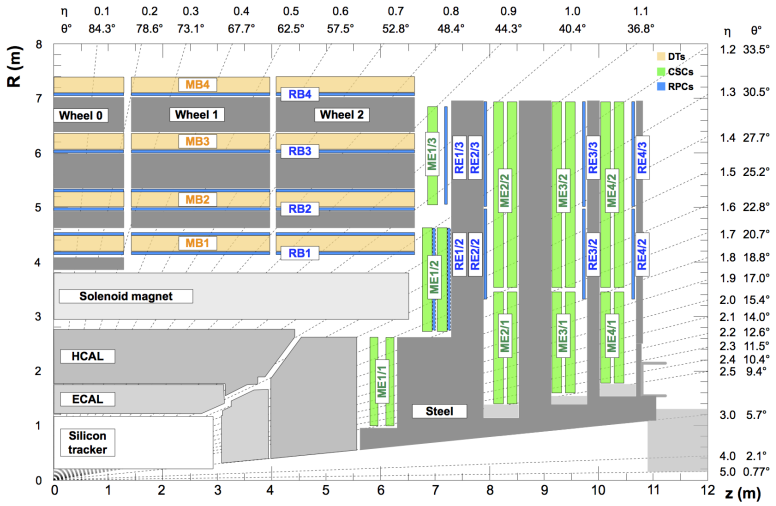


FIGURE 2.8: Longitudinal view of a quarter of the CMS muon system used in the CMS Run 2 (2015-2018). The various muon stations and the steel disks (dark grey areas) are shown. The 4 drift tube (DT, in light orange) stations are labeled MB ("muon barrel") and the cathode strip chambers (CSC, in green) are labeled ME ("muon endcap"). Resistive plate chambers (RPC, in blue) are in both the barrel and the endcaps of CMS (RB and RE stations).

range $|\eta| < 1.6$. Both the DTs and the CSCs are primarily tracking detectors, but have also good time resolution. The RPCs instead are mainly used for their good time resolution.

During the first run of LHC three layers of RPC detectors were installed in the endcap; the fourth layer was installed during the Long Shutdown 1 and was active in Run 2. The highest rapidity region ($1.6 < |\eta| < 2.4$) is currently being instrumented with two stations of Gas Electron Multiplier detectors (GEMs) [55], which can accomplish the same physics goals of the RPCs in a high radiation environment.

The DTs' layers or "stations" (at fixed r in the barrel), made out of 5 wheels, each divided into 12 sections in ϕ, consist of 8 layers of tubes measuring the position in the transverse plane and 4 layers in the longitudinal plane, except for the outermost station with only 4 layers of tubes in the transverse plane. The tubes have a section of 4.2×1.3 cm, with a conducting wire in the center. Each DT station provides a measurement of the muon position with a 100 μm resolution in $r \times \phi$ and 150 μm in the z direction. The DT drift cell was designed to provide a uniform electric field and constant drift velocity.

The CSCs' stations are located in the endcap at fixed z. CSCs are employed in the endcap regions because of the higher particle rates and the magnetic field properties. Each endcap has 4 stations of chambers perpendicular to the beam. A CSC consists of 6 layers, each measuring the muon position in 2 coordinates. Cathode strips are positioned radially and provide a precision measurement in the $r \times \phi$ plane. Anode wires provide a measurement in the radial direction. The CSC stations have a $r \times \phi$ resolution of about 75-150 μm and a z resolution of about 200 μm.

The resistive plate chambers (RPC) are located in both the barrel and endcap regions. The RPCs are double-gap chambers operated as gaseous detectors in avalanche mode. They are read out using strips oriented along z in the barrel and r in the endcaps. The spatial resolution is coarser, with a strip pitch of $\simeq 1$ cm and no segmentation in the orthogonal direction,

but the time resolution is very good (3 ns by design), thus ensuring a robust bunch crossing identification and efficient triggering.

The readout electronics of the muon chambers were also improved as part of the CMS trigger upgrade between Run 1 and Run 2.

2.2.4 The Trigger System

The 40 MHz rate of proton-proton collisions and the pileup make it impossible to process and store all the information provided by the detector. Most of the events are not interesting for physics analysis in any case, due the to fact that the total proton-proton cross section is more than 6 orders of magnitude larger compared to the cross section of interesting processes. The data needs to be reduced and selected trough a trigger system, whose crucial aspect is a fast and efficient real-time selection to record the useful events.
In CMS the data reduction happens in two steps: The Level-1 (or L1) Trigger [56] and the High Level Trigger (HLT) [57].

Level-1 Trigger

The Level-1 Trigger consists of programmable electronic devices, which process information coming from the calorimeters (both ECAL and HCAL) and from the muon system only. It reduces the event rate from an input of 40 MHz to an output of about 100 kHz, through a synchronous pipelined structure of processing elements. At every bunch crossing, each processing element sends its results to the next element and receives a new event to analyze. The detectors must be able to resolve in time two events and all the signals from a single event must be synchronized. During this process, the full detector data are stored in pipeline memories with limited size ($\sim$160 bunch crossings, or $\sim$4 μs latency).
An upgrade of the Level-1 trigger was necessary between Run 1 and Run 2 in order to maintain the same physics performances with higher collision rates and pileup. Because of the requirement of having the same sensitivity of Run 1 at the electroweak scale, the electronic devices for the calorimeter trigger, the muon trigger, and the global trigger were upgraded. The new design allows for more flexibility, and is more suitable to trigger on complex objects. FPGAs were introduced for this purpose.

High Level Trigger

The HLT further decreases the event rate from about 100 kHz to about 1 kHz for data storage. The HLT is implemented by a computer farm composed of more than 30000 CPUs running software modules similar to the ones used for the offline reconstruction. The full detector readout is available at the HLT, but in order to meet the timing requirements given by the input rate from L1, the events are reconstructed in multiple steps and rejected as soon as there is enough reconstructed information to make a decision.

Reconstructed physics objects, such as leptons, photons, jets, etc. are used in the HLT. A list of reconstruction algorithms and filters for one or more physics objects is called HLT Path. The path is characterized by the Level-1 "seed", i.e. the requirements passed by the event at Level-1, and by the HLT requirements.
An "HLT Menu" represents the set of trigger paths that, if enabled, contribute to a final OR of decisions that determines whether to reject or store an event. A single trigger path can require the presence of one or more physics objects of a particular type passing specific kinematic thresholds, and it can also mix physics objects.

The event rate of each trigger path should be maintained within the allowed limits given the expected instantaneous luminosity. Trigger paths with lower thresholds than those necessary to reduce the event rate can be kept in the HLT Menu with a "prescale" factor applied. They are useful to measure the efficiencies of higher threshold triggers.

2.2.5 Track Reconstruction

All the detector elements and the trigger have compelling time requirements, as they have to produce signals, transmit and process the data at rates dictated by the LHC bunch spacing. Once the data are stored, the offline reconstruction software reconstructs the data in steps and makes them available for analysis. The software is organized in modules operating on standardized data formats. The HLT is also running similar and in some cases the same modules. The next two sections are dedicated to a few fundamental parts of the offline reconstruction starting with the tracking algorithms.

The first step of the reconstruction process is referred to as local reconstruction both in tracking and for each subdetector. For the tracker, it consists of processing the detector readout to build hits. A tracker hit is the best estimate of the point where a charged particle has crossed the silicon layer, based on the released charge distribution Each trajectory is a sequence of hits in the tracker: tracking algorithms assign hits to tracks and aim at measuring with the best possible resolution the five parameters of the helix.

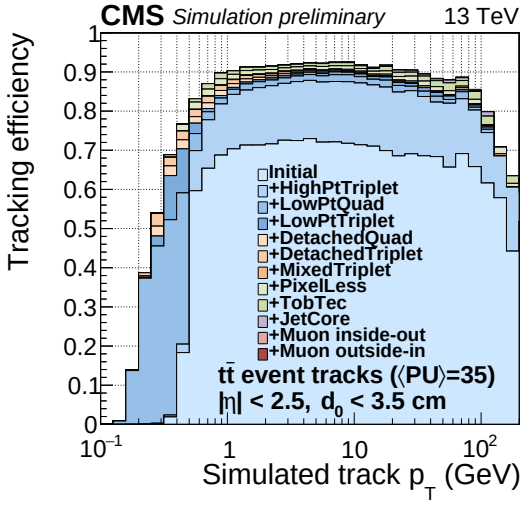
An iterative approach is employed for the CMS tracking: the track reconstruction is run several times, starting from easier tracks, i.e. non-displaced tracks and with relatively high p_T, and progressively moving to the more complex ones. After each iteration the tracks meeting quality criteria are kept and their hits are masked when running the following iterations. Each step allows an increase of the tracking efficiency, with low a low rate of fake tracks due to the quality requirements and the decreasing combinatorial complexity.

The contribution of each iteration to the final reconstructed tracks is shown in figure 2.9 as a function of p_T, with the various tracking iterations in different colors. Simulated events with the new pixel detector are used. The main tracking iterations are also listed in the adjacent table, 2.9 (B), with their seeding hits and target track. Specialized tracking iterations (jet core, muons) are included only in the plot, as final tracking steps. The plot shows that the tracking efficiency saturates at $\sim$ 90 % due to hadrons that undergo nuclear interactions in the tracker material, while it is known to be close to 100 % for muons. At low p_T the efficiency is lower due to the nuclear interactions. At high p_T, the efficiency decreases due to the fact that in the event topology chosen for the plot (top pair production) the high p_T tracks are found mainly in the core of high p_T jets. Often pixel hits are merged, thus making the seeding step of tracking inefficient.

Iterative Tracking

The algorithm which is run at every iteration is called the Combinatorial Track Finder (CTF), which is based on the combinatorial Kalman Filter [59, 60] technique. The CTF algorithm is composed of four steps, that are run at each iteration:

- **Seeding** - Seeding defines the initial parameters of the track and its uncertainties. The seeding needs at least three 3D hits (or two with an additional constraint on the origin of the track). Pixel hits are preferred for seeding because of the better resolution and the lower occupancy of the pixel detector compared to the strips. Pixel hits are also less affected by interaction in the tracker material. Strip "stereo" hits are also



(A)

Seeding	Target
4 pixel	prompt, high p_T (initial)
4 pixel	prompt, low p_T
3 pixel	prompt, high p_T
3 pixel	prompt, low p_T
4 pixel	displaced
3 pixel	displaced
3 pixel+strip	displaced
3 inner strip	displaced
3 outer strip	displaced

(B)

FIGURE 2.9: Iterative tracking efficiency as a function of the track p_T, at $\sqrt{s}$ = 13 TeV (A); Summary table of the seeding hits and of the target tracks for each of the main tracking iteration (B). [58]

used, but not in the initial tracking iterations.

The new pixel detector with four layers helped the seeding significantly: the best seeds have now 4 pixel hits. There are actually two kinds of seeds with 4 pixel hits as reported in table 2.9 (B): the first uses hits triplets with a compatible fourth hit, and the second propagates the triplet. In the subsequent iterations, a 3-hit seeding is run as a recovery for prompt tracks, then displaced tracks with several combinations of hits are considered.

- **Pattern recognition or Track finding** - The seed trajectories are propagated searching for compatible hits in the outer layers. The seeds are extrapolated up to the entire tracker and at each iteration the track parameters are updated. The material crossed by the track and the uncertainties of the hits are taken into account, and the hit reconstruction is also refined based on the compatible track. Ghost hits, i.e. layers without charge deposits, but whose the material is taken into account, are allowed to recover possible tracker inefficiencies.

  Once the track is completed another search is performed backwards starting from the outermost hit. This step is performed because the pattern recognition is more efficient than the seeding, as it handles correctly possible silicon modules overlap, and it can recover pixel hits which were not used at seeding time.

  In case multiple compatible hits are found when extrapolating the trajectory to a single layer, the algorithm will create one trajectory candidate for each hit and those are propagated independently. Eventually, only one track is retained based on the quality and the total number of hits.

- **Track fitting** - Once the tracks are built the trajectory is refitted using a Kalman filter inside-out, and a Kalman filter outside-in, with the latter step called also "smoothing

stage". The Kalman filter is first run starting from the innermost hit, with the trajectory state given by the inner hits. The errors are enlarged and all the hits inside out are added . For each valid hit, the estimated hit position uncertainty is also updated using the values of the track parameters. This first filter is followed by the smoothing stage: a second filter is initialized with the result of the first one, but with enlarged errors, and is run outside in. The helix parameters can be obtained from the weighted average of the parameters of these two filters at the surface associated with each hit. The parameters found at the innermost hit are used for extrapolations inwards, while the ones found at the outermost hit are used for extrapolation to calorimeters and the muon chambers.

- **Track selection** - Tracks are selected on the basis of the number of hits in the entire tracker,the number of 3D hits, the track normalized χ^2, and the distance from the primary vertex. The criteria are optimized as a function of p_T, η and the number of tracker layers with hits.

The efficiency of the tracking sequence is shown in figure 2.10 as a function of simulated η (A) and p_T (C), while the rate of fake tracks is shown in as a function of the reconstructed η (B) and p_T (D). The 2016 tracking performances are compared to the simulation of the detector with the new pixel detector installed (2017). Only "high purity" tracks are considered for these plots. The efficiency is 95% with the new pixel detector for central η, and an improved efficiency with respect to 2016 is visible in the range $1.5 < \eta < 2$. The efficiency is flat and close to 90% for the large region 1 GeV $< p_T <$ 100 GeV. The fake rate is also lower everywhere in the 2017 simulation. The tracking efficiencies and fake rate were also measured in data with similar results.

The reconstructed tracks are identified by five parameters: the resolution for two impact parameters, i.e. the distances of the point of closest approach to the origin, are plotted in figure 2.11. The resolution is of 100-200 μm in the transverse impact parameter, d_0 or d_{xy}, (A) and 100-500 μm in dz (B). The impact parameter resolution is largely improved with the new pixel tracker, assuming ideal detector operation, and it is crucial for b tagging (chapter 3) and the H$\rightarrow$b$\bar{\text{b}}$ analyses. The p_T resolution is 2-4% for p_T up to the TeV and doesn't depend as much from the new pixel detector.

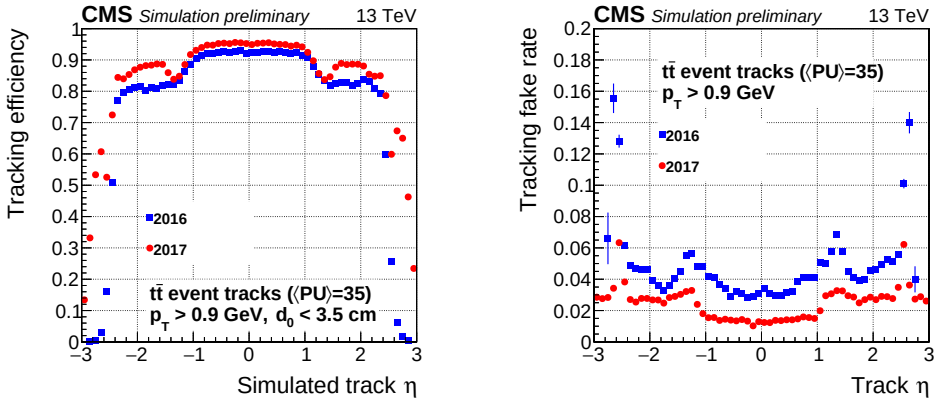
Primary vertex reconstruction

Reconstructed tracks are also used to measure the position of the primary vertices, which are the proton collision points. Tracks with good quality criteria and no significant displacement from the beam spot are clustered and fit to obtain the vertex position.

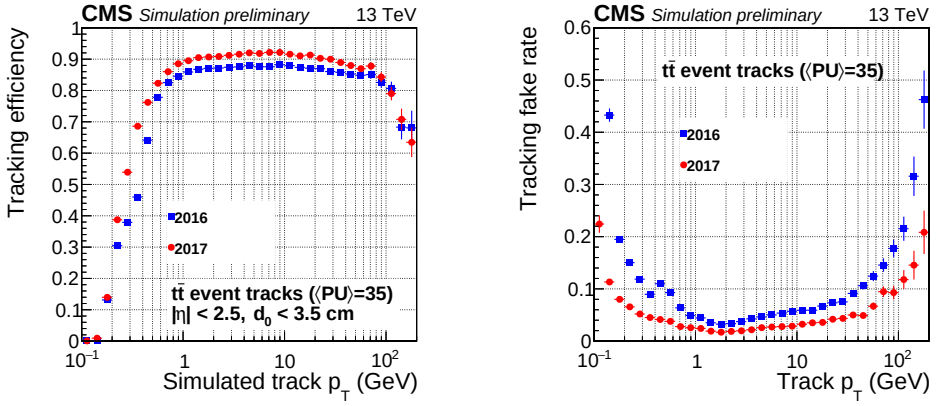
Tracks are clustered in z using the deterministic annealing algorithm [61]. Track clusters are then fitted using the Adaptive Vertex Fitter [62] algorithm. This algorithm is an iterative re-weighted Kalman filter that fits a candidate vertex starting from a collection of tracks. Tracks are re-weighted at each iteration so that the contribution of fake tracks gradually diminishes.

The primary vertex resolution depends on the number of associated tracks, as shown in figure 2.12 in $x \times y$ (A) and z (B). The resolution is about 10-50 μm.

There are multiple vertices in an event due to the pileup, and once they are reconstructed, one is chosen as the signal vertex and used for analysis. During the Run 1, the signal vertex was chosen as the primary vertex with the largest $\sum_{tracks} p_T^2$. This algorithm has been

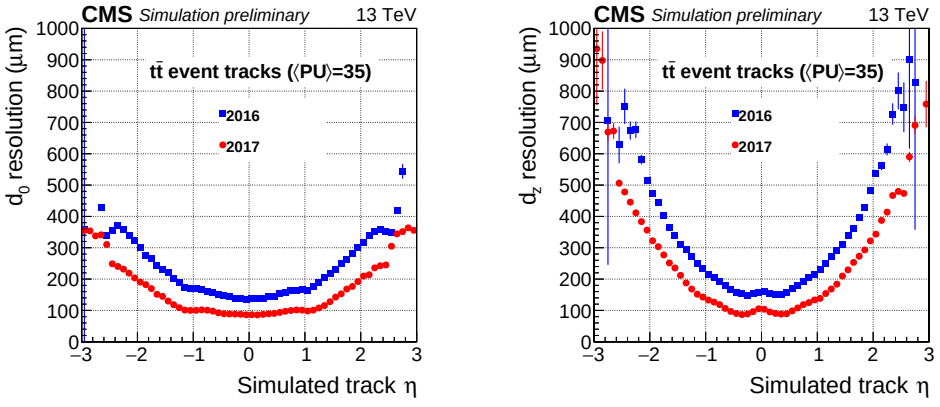


(A) (B)



(C) (D)

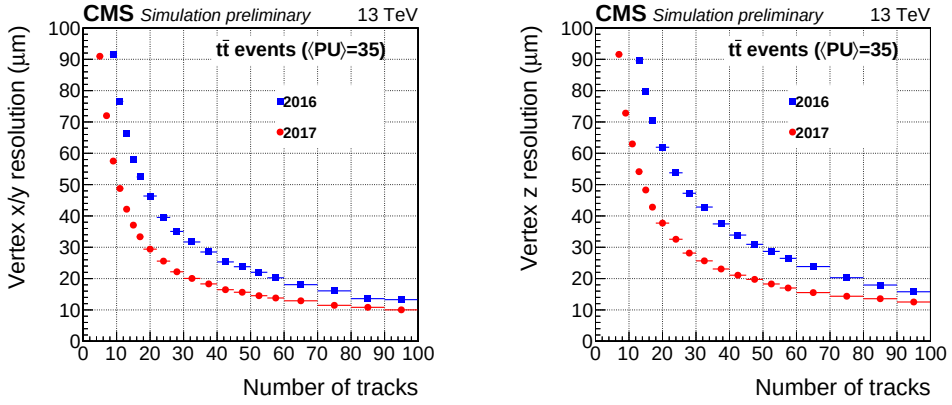
FIGURE 2.10: Track reconstruction efficiency as a function of simulated track η (A), p_{T} (C) for 2016 and 2017 detectors. The 2017 detector shows better performance than 2016 over the entire p_{T} and η spectrum. . The fake rate for both years as function of the reconstructed track η (B), p_{T} (D).



(A) (B)

FIGURE 2.11: Track d0 (transverse impact parameter) and dz (longitudinal impact parameter) resolution as a function of the simulated track η for 2016 and 2017 detectors The 2017 detector shows better performance than 2016 over all the η spectrum. .

improved during the Long Shutdown 1 in order to choose the vertex depending on the $\sum p_T^2$ of all the collision products, as reconstructed using the Particle Flow algorithm (2.2.6).



(A) (B)

FIGURE 2.12: Vertex transverse resolution as a function of the number of tracks used in the vertex fit. The 2017 detector shows better performance than 2016 detector .

2.2.6 Particle Flow reconstruction

The particle flow algorithm [63] is used by CMS to correlate all single detector measurements, to identify the final state particles and measure their properties. The basic elements given as input to the particle flow algorithm are tracks, including the ones reconstructed in the inner tracker and the muon tracks, and calorimeter clusters. The basic particle flow elements are first connected by a geometrical link to form a block (PF block). A link distance is also defined at this stage to quantify the quality of the link. Only the nearest neighbors are considered for link building, in order to reduce the computing time.

All the elements in a PF block are then combined to identify and measure the particle candidates (PF candidates). The PF candidates are in turn used to build higher level objects that are used at the analysis level.

Several types of links between PF elements are possible.

A track can be linked to a calorimeter cluster if its extrapolated position is within the cluster area, given by the union of the areas of all the cluster cells in the $\eta - \phi$ plane HCAL and the barrel of ECAL, or in the $x \times y$ plane for the ECAL endcaps and the preshower. This area is expanded by up to one cell in each direction to take into account several sources of mismeasurement. The link distance is then defined as the distance between the track extrapolation and the calorimeter cluster center. In case several tracks are linked to one ECAL cluster, only the link with the smallest link distance is kept.

Electrons and photons are affected by bremsstrahlung and conversions in the tracker material ($\sim 1\ X_0$, dependent on η). The electron and photon reconstruction is carried out using both the ECAL and the tracker: roughly speaking the electron candidates are identified by a ECAL cluster and a track, while the photon candidates are identified by ECAL clusters with no linked tracks. The emission of bremsstrahlung photons is taken into account for the electron reconstruction both in tracking and in building the ECAL clusters. Bremsstrahlung photons are emitted in the direction tangent to the original trajectory, so they have the same

η, but different ϕ with respect to the electron. In the ECAL superclusters with an enlarged ϕ window are used. The tracking for the candidate electrons is also rerun using a different filter, called Gaussian summation filter (GSF) [64], to take into account the possibly large bremsstrahlung losses and kinks in the trajectories.
At the linking stage, in order to recover bremsstrahlung extra photons emitted by electrons, the tangents of the electron tracks (GSF tracks) are extrapolated to the ECAL. A cluster is linked to the track as a potential bremsstrahlung photon if the extrapolated tangent position is within the cluster and compatible in η.
Additionally, photons conversion to e^+e^- pairs are considered: if the candidate photon direction, obtained from the sum of the two track momenta, is compatible with a tangent to a GSF track, a link between the track and the two other tracks is created.

Links between ECAL and HCAL (or preshower) clusters are made when the cluster position in the more granular calorimeter (preshower or ECAL) is within the cluster envelope in the less granular calorimeter (ECAL or HCAL). The link distance is also defined as the distance between the two cluster positions, in the $\eta \times \phi$ plane for an HCAL-ECAL link, or in the $x \times y$ plane for an ECAL-preshower link. In case of multiple clusters linked to each other only the pair with minimum link distance is considered. ECAL clusters and superclusters are linked to each other if cells are shared.

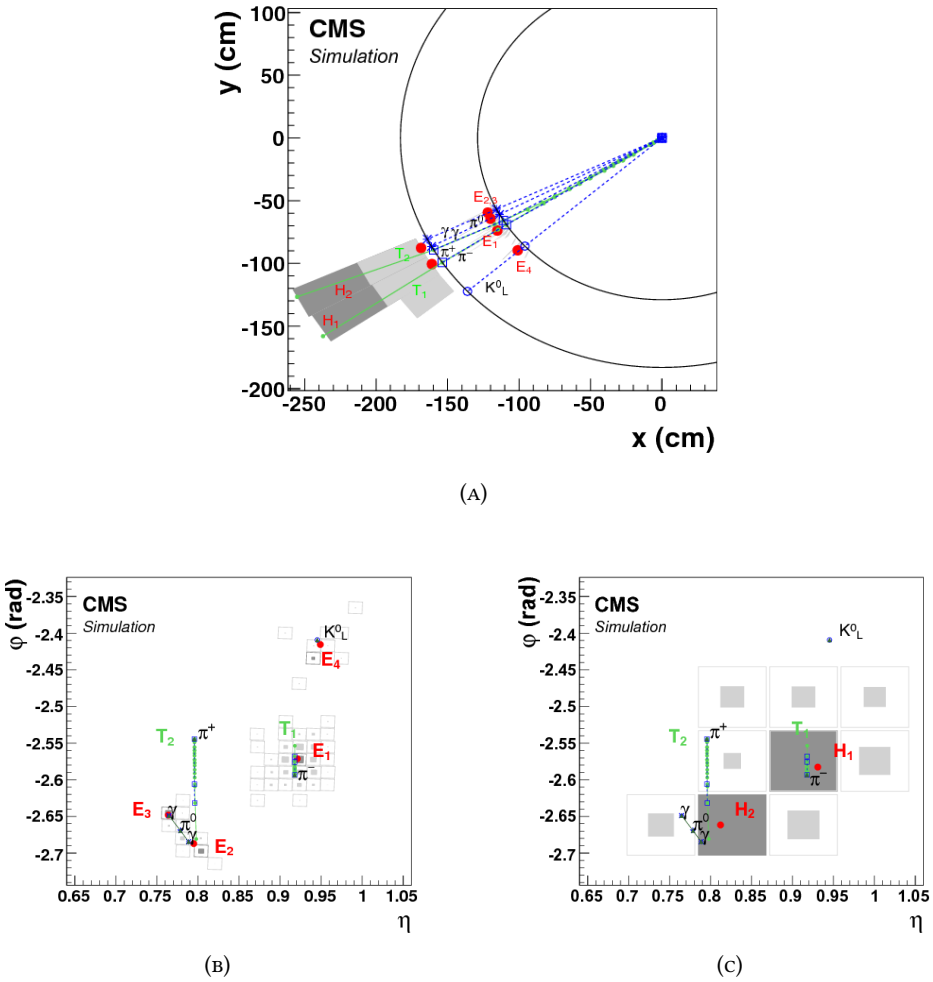
Nuclear interactions in the tracker material are also considered at the linking stage: tracks can be linked together in case they are coming from a common secondary vertex. Secondary vertices due to nuclear interactions are kept if they have at least three tracks and at most one incoming track, which points to the primary vertex.

Finally, links between tracks and muon tracks can be made: this is the first step in the muon reconstruction.

Figure 2.13 shows a simulated jet with only five particles for explanation purposes. The most common links can be found: the two charged pions produce tracks that are linked to calorimetric clusters. The π^- produces clusters both in the ECAL and in the HCAL. These are linked to each other and to the track. The π^+ doesn't produce an ECAL cluster. The HCAL cluster is displaced in ϕ due to the magnetic field with respect to the initial track direction: the magnetic field is necessary for momentum measurements in the tracker, but it can also help separate charged and neutral components in jets. The tracks are linked also the to other HCAL clusters, but with larger link distance, therefore the block includes 5 basic elements.The two photons from a π^0 form an ECAL cluster not linked to a track. The K^0 produces also an ECAL cluster with no links. These are two independent PF blocks.

Once all the links are available for the event the following steps are run, targeting increasingly complex reconstruction steps. The steps, in the same order they are executed, can be summarized as:

- μ **reconstruction** - Muon candidates are identified and reconstructed, and the corresponding basic elements are removed from each PF block. Different quality criteria are applied for isolated muons and muons in jets, in order to take care of charged hadron misidentification.
- **electron & isolated** γ **reconstruction** - Electrons are identified and reconstructed, with the aim of recovering the energy of all the bremsstrahlung photons. The energy of the electrons is measured with a combination of the ECAL energy and the GSF track, while the direction is taken as the GSF track direction. Isolated photons,



(A)

(B) (C)

FIGURE 2.13: Display of five particles in a jet. The transverse plane view (A) shows the tracks in green; the circumferences indicate the ECAL and HCAL surfaces. The ECAL (B) and HCAL (C) clusters in the $\eta \times \phi$ plane are also shown.

converted or unconverted, are identified. The corresponding tracks and ECAL or preshower clusters are removed from the PF element collections. Tracks identified as coming from photon conversions are also removed, even if not associated to a PF candidate.

- **Jet constituents reconstruction** - Charged hadrons, neutral hadrons, and photons in jets are targeted. Inside the tracker acceptance all the ECAL deposits not corresponding to tracks are identified as photons, while the HCAL clusters are identified as neutral hadrons. Precedence is given to photons which are more abundant in jets. The energies of photons and neutral hadrons are measured by ECAL or HCAL clusters inside the tracker acceptance. Outside of the tracker acceptance the photons are identified as ECAL clusters without a link to HCAL clusters, while ECAL+HCAL clusters are just identified as "hadronic showers". The energies are measured by HCAL+ECAL outside the tracker acceptance. Tracks linked to HCAL clusters are then used: at this stage if good compatibility between the sum of the tracks momenta and the cluster energy is found, no neutral particles are identified; in case there is an excess of HCAL energy, it can be attributed to neutrals. Very rarely, if the tracks have larger momentum sum, a search for muons with loose requirements is carried out, or tracks with

large momentum uncertainty are removed, as they could be mismeasured. The energies of charged hadrons are given by the weighted average of the track momenta and the calibrated cluster energy, where the tracker dominates.

- **Nuclear interactions cleaning** - Nuclear interaction links are used: an incoming track is identified, and it is used to refine the reconstruction.
- **Global post-processing** - The PF candidates are used to build the $p_{\mathrm{T}}^{\mathrm{miss}}$ vector, which is the opposite of the vectorial sum of all the candidate momenta. The post-processing targets events with artificially large $p_{\mathrm{T}}^{\mathrm{miss}}$ usually due to very high energy muons. Particles causing very high missing energy are inspected: in case of muon tracks compatible with cosmic rays, they are removed from the event. In case of poorly measured muons, i.e. if some of the elements are significantly different one from another, the element which gives the lower $p_{\mathrm{T}}^{\mathrm{miss}}$ is chosen. Particle misidentification is also considered: punch-trough charged hadrons can be rarely identified as muons and thus added to particle flow candidates twice, both as muons and neutral hadrons. Similarly genuine muons can fail the tight identification criteria and overlap with energetic neutral hadrons. The solution with lower $p_{\mathrm{T}}^{\mathrm{miss}}$ is kept in these cases.

The particle flow reconstruction is beneficial mostly to the jets and missing transverse energy measurements, whose reconstruction is described in chapter 3, and the identification and measurement of hadronic τ lepton decays, which are not covered in this thesis.

2.2.7 Simulation

The simulation is a crucial aspect of high energy physics experiments. In CMS it is used at analysis level and for testing the algorithms before deployment with data. The simulation targets relatively rare processes originating from a hard interaction between two proton components, which are signal or background for specific analyses. Single particles or hadronic jets can also be simulated for specific purposes.

Several theoretical, phenomenological and experimental inputs are necessary to build a simulation of the proton-proton collisions. Different techniques are used in particular to describe the QCD processes, whose phenomenology varies greatly at different energy scales (see e.g. [65]).

Proton-proton collisions are very complex and difficult to model accurately. Protons are composed of 3 quarks, called valence quarks, by virtual gluons and virtual quark anti-quark pairs coming from gluon splitting. All constituents of hadrons are generically called partons.

During high energy collisions, the protons behave as a collection of free partons and the hard scattering can be described at the level of parton interactions. The hadronic cross section σ_{pp} is calculated based on the QCD factorization theorem. The factorization theorem states that the hadronic cross-section σ_{pp} is a convolution of the partonic cross section $\hat{\sigma}_{i,j}$ with the parton distribution functions (PDFs) $f_i(x)$:

$$\sigma_{pp} = \int_{x_{min}}^{1} \sum_{i,j} f_i(x_1) f_j(x_2) \hat{\sigma}_{i,j}(x_1 p_1, x_2 p_2) dx_1 dx_2,$$

where function $f_i(x)$ is probability density that a parton of type i has a fraction x of the hadron energy.

Apart from the hard interaction, the other constituents of the proton can also interact. This usually results in a spray of softer particles, called underlying event (UE). Any high momentum particle involved in the collision will emit additional hard QCD radiation. Radiation from particles before the hard interaction is called initial-state-radiation (ISR), whereas radiation off particles produced in the collision is called final-state-radiation (FSR).
Quarks and gluons can emit additional radiation via the strong interaction. All the quarks and gluons go through the hadronization process, forming colorless hadrons. Finally, unstable particles are going to decay. A representation of all these elements is shown in figure 2.14.

The elements involved in the calculations of a process can be summarized as:

- The PDFs that are phenomenological functions computed using experimental information
- the hard scattering, computed perturbatively order by order
- the parton showering, used to simulate additional emissions in perturbative QCD
- the hadronization, describing the transition from colored particles to hadrons, treated using phenomenological models
- the decay of unstable particles, modeled based on experimental data.

The first two are usually included in Matrix Elements generators, while the last three are included in Parton Showering programs. Both use Monte Carlo techniques. The matching between these Matrix Elements generators and Parton Showering should be done in a way to avoid double counting of QCD radiation.

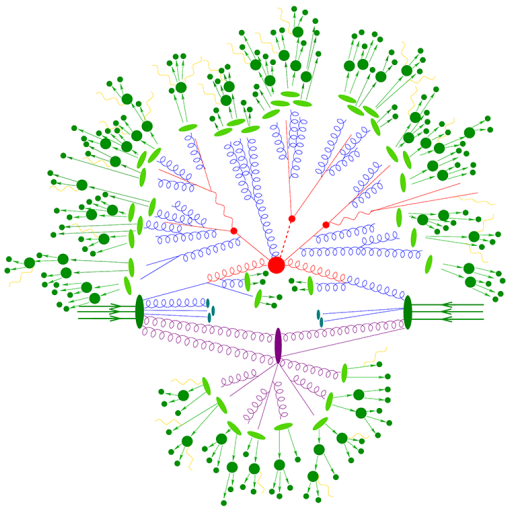


FIGURE 2.14: Representation of a proton-proton collision event. The red part includes the hard interaction and the decay of the products. Initial and final state radiation are in blue. A secondary interaction can take place, in purple, before the final-state partons hadronize. The hadronization is represented by the green blobs, and the hadron decay in dark green. Photon radiation is in yellow. [66]

Monte Carlo techniques are then used for the simulation of the interaction of the stable particles with the detector. The detector simulation is implemented using GEANT4 [67]. Pileup interactions are also added at this stage. A library of simulated hits of minimum bias events

is used to add a number of extra interactions onto the signal event according to a specified pileup scenario. Out-of-time pileup is modeled by modifying the timing of the detector hits when adding a minimum bias interaction.

Once the simulation of the detector is run, all the detector signals are converted to electronic signals in a format identical to the one used for data. From this point onwards the simulated events go through the same reconstruction steps as the collision data.

Chapter 3

Physics Objects reconstruction

Stable particles originating from proton-proton collisions are identified combining detector information with the particle flow algorithm. These stable particles are then used to build physics objects and high-level observables to be employed at the analysis level.
The present search focuses on a final state containing b jets coming from Higgs boson decays, leptons and missing energy from vector boson decays, which are used also at the trigger level. An overview of the high-level objects and of the performance of the algorithms used by CMS is given.

3.1 Isolated leptons

The leptonic decays of the W and the Z bosons are exploited in the VH($b\bar{b}$) analysis. In order to select leptons with high efficiency and purity, muons and electrons candidates obtained via the PF algorithm are employed. An important criterion used to select leptons originating from the W and Z decays is the isolation. The isolation distinguishes prompt leptons, such as the ones originating from W and Z boson decays, from leptons produced in hadron decays that usually are embedded in a jet. When using all the PF candidates the isolation is usually defined as:

$$Iso_{\mathrm{PF}} = \sum_{charged} p_{\mathrm{T}} + max(0, \sum_{neutral} p_{\mathrm{T}} + \sum_{\gamma} p_{\mathrm{T}} - p_{\mathrm{T}}^{\mathrm{PU}})$$

where the sum runs on the particle flow candidates contained in a cone of given $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ centered around the lepton direction, and $p_{\mathrm{T}}^{\mathrm{PU}}$ is a correction that takes pileup neutrals into account. Alternatively the charged particles only are used. The relative isolation $Iso_{\mathrm{PF}}/p_{\mathrm{T}}$ can also be used to select leptons at the analysis level.

3.1.1 Muons

The muon identification and reconstruction is the first step in the PF algorithm, as muons are identified almost unambiguously by the muon chambers. Particle flow muons are required to pass selection criteria, which depend on the isolation.

Muon tracks are reconstructed with an iterative approach, running a sequence of specific tracking algorithms. The inputs are tracker tracks and muon track segments; after each iteration the hits associated with the reconstructed tracks are removed.
The final muon collection is composed by:

- **Standalone muon** tracks, built using information from muon subdetectors only;
- **Tracker muon** tracks, built "inside-out" by matching tracks to segments in the muon system;

- **Global muon** tracks, built "outside-in" by matching standalone-muon tracks with tracker tracks: a combined fit is performed for optimal momentum resolution.

The muon transverse momentum is obtained from a refit of the tracker track and the matching hits in the muon system, which can constrain the very high momentum muons. For muons up to 200 GeV in p_T the inner track alone provides the best measurement. The momentum resolution for muons with transverse momenta up to approximately 100 GeV is 1% in the barrel and 3% in the endcap.
The resolution is measured with cosmic rays data, as the relative difference in q/p_T between the upper and the lower track segment, as shown in figure 3.1 for central barrel muons. Events with cosmic muons passing close to the interaction point with least one hit in the pixel detector are chosen, so both the upper and the lower leg look like collision muons. The quantity evaluated is the RMS of the relative q/p_T difference, which is computed as

$$\frac{1}{\sqrt{2}} \frac{(q/p_T)_{\text{upper}} - (q/p_T)_{\text{lower}}}{(q/p_T)_{\text{lower}}}.$$

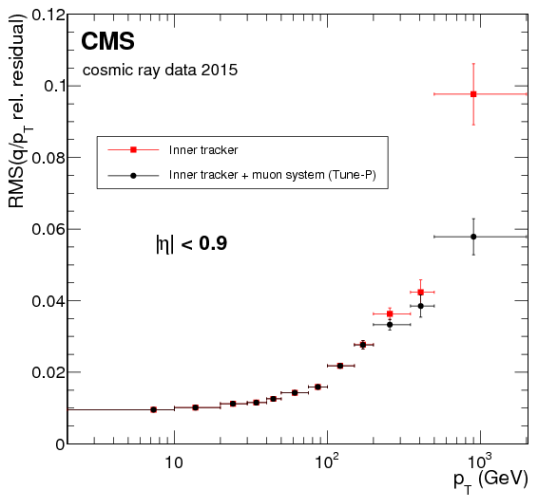


FIGURE 3.1: RMS of relative q/p_T difference as a function of p_T for cosmic rays recorded in 2015, using the inner tracker fit only (red) and including the muon system (black).

Several identification criteria were studied to increase the purity of the muon collection and reduce the rate of particles misidentified as muon, i.e. the fake rate. Three main identification types of muon identification are defined based on efficiency and fake rate. Loose, Medium and Tight muon identification correspond respectively to efficiencies of $\sim$ 99.7, 98.5 and 97 % in data.
"Loose muons" are selected by the PF algorithm and are also a tracker or a global muon.
"Medium muons" are Loose muons with a tracker track that uses hits from more than 80% of the inner tracker layers it traverses. Different selections based on the fit χ^2 and on the compatibility between the tracker track and the muon track or segment are applied for tracker and standalone muons.
"Tight muons" are loose global muons with a tracker track that uses hits from at least six layers of the inner tracker, including at least one pixel hit, and a segment matching in at least two of the muon stations. Other criteria applied are based on the global muon track fit χ^2 and on the compatibility with the primary vertex. These criteria suppress punch-through charged hadrons and muons produced in flight. The Tight muon identification is therefore specialized in prompt muons, while the Medium identification is used both for prompt and

muons in jets from heavy flavor decays.

The muon isolation is computed relative to muon p_T summing the energy coming from the particles in a cone of radius $\Delta R = 0.3$ or 0.4 around the muon. Tight and loose isolation working points are defined to achieve efficiencies of 95% and 98%, respectively.

The efficiency for muon identification and isolation is measured with the tag-and-probe method in $Z \to \mu\mu$ events, starting with tracker tracks as probes. The total muon efficiency ϵ_μ is measured in several steps and the contributions are factorized as $\epsilon_\mu = \epsilon_{trig} \times \epsilon_{track} \times \epsilon_{reco+ID} \times \epsilon_{iso}$, where the first term is taken into account if the muon is also used at the trigger level and the second is the tracking efficiency. The efficiencies measured for the identification (A,B) and isolation (C,D) for 2015 data are shown in figure 3.2. Scale factors, i.e. the data/MC ratio, for the muon efficiencies are also computed comparing data and simulation and applied to the analyses.

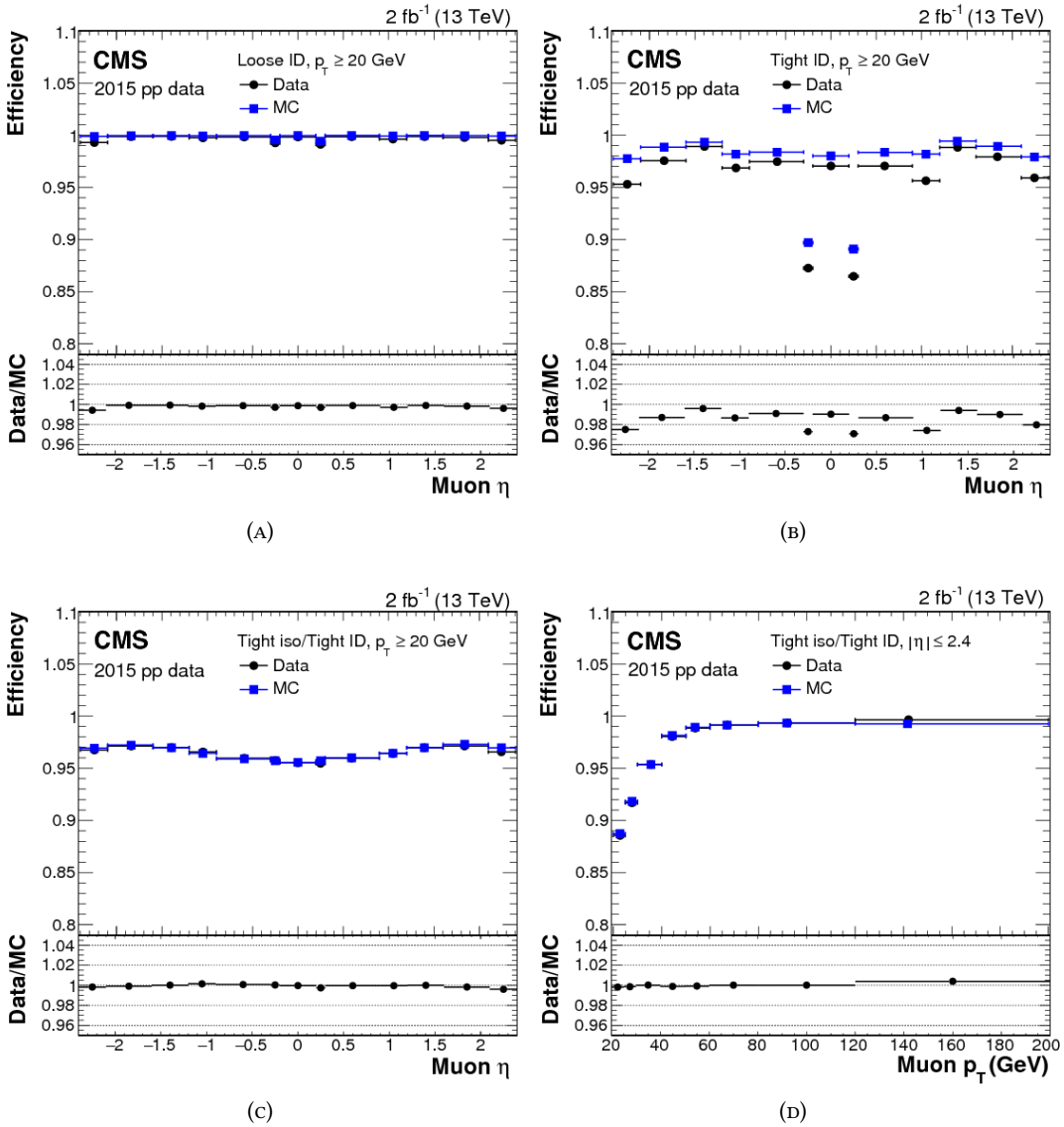


FIGURE 3.2: Tag-and-probe efficiency for the loose (A) and tight identification criteria (B), for the tight PF isolation working point on top of the tight identification versus η (C), p_T (D) for muons in the acceptance of the muon chambers..

3.1.2 Electrons

The electron reconstruction [68] is carried out as a part of the PF reconstruction, by matching tracker tracks with ECAL clusters, with the aim of recovering bremsstrahlung photons and possible photon conversions.

Unlike muons, whose charge is unambiguously measured, the electron charge measurement is affected by bremsstrahlung, in particular when the bremsstrahlung photons convert upstream in the detector. The methods used most commonly for the estimation of the charge are the curvature of the electron track, fitted with the GSF algorithm, the curvature of the Kalman Filter track matched to the electron and the different position of the ECAL cluster or supercluster with respect to the first hit of the GSF track.
The charge is chosen as the one given by at least two of these three estimates. The charge misidentification probability of this algorithm is predicted by simulation to be 1.5% for reconstructed electrons from Z boson decays without additional purity selections.

The electron momentum is measured using both the ECAL calibrated energy and the momentum of the GSF track, which are combined using a multivariate regression targeting the relative weight of the tracker and ECAL measurements. The resolution is optimal for low bremsstrahlung barrel electrons (2%) and ranges from 10 to 5% for showering endcap electrons of p_T up to 100 GeV, as shown in figure 3.3. The fraction of low bremsstrahlung and showering electrons depends on the η, due to the material crossed by the particles. In Z events about 60% of the electrons are showering and 25% are low bremsstrahlung, inclusively in η. The remaining fraction has very large bremsstrahlung or is badly reconstructed due to ECAL defects.

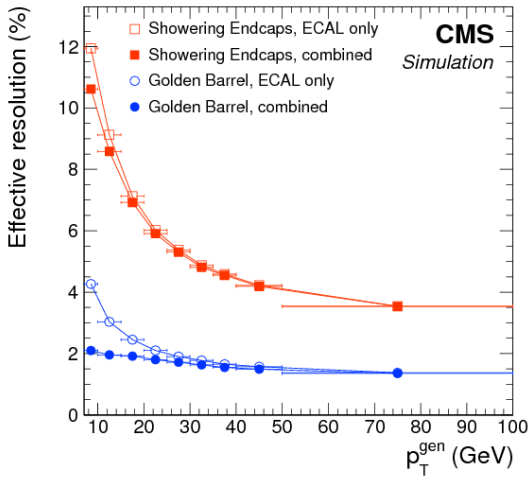
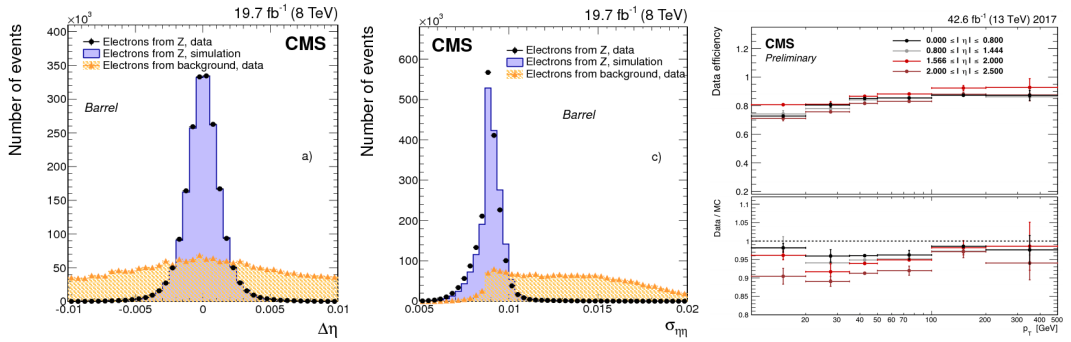


FIGURE 3.3: Resolution in electron momentum after combining the ECAL energy and track momentum, compared to the ECAL energy only, as a function of the generated electron p_T. "Golden" electrons in the barrel, with the best resolution, and showering electrons in the endcaps, with the worst resolution, are shown.

The identification of isolated electrons is aimed at separating them from photon conversions, or jets misidentified as electrons. Multivariate classifiers implemented via Boosted Decision Trees are mostly employed for this tasks. The variables used are observables that compare measurements obtained from the ECAL and the tracker, calorimetric only observables, and tracking only observables. Calorimetric observables make use of the transverse

shape of the ECAL deposits, the energy in HCAL and in the preshower, and help reject e.g. jets with large electro-magnetic components. Tracking observables improve the separation between electrons and charged hadrons, they are based on the GSF track and the matched Kalman Filter track. Example variables with good discriminating power are the lateral extension of the shower $\sigma_{\eta\eta} = \sqrt{\frac{\sum(\eta-\eta_i)^2 \cdot w_i}{\sum w_i}}$ of the ECAL superclusters, where the sum runs on the crystals surrounding the maximum energy one in the supercluster and the weight that depends on the crystal energy, and the $\Delta\eta$ between the position of the ECAL superclusters and the extrapolated track, as shown in figure 3.4 (A,B). Additionally, the isolation, which uses PF candidates for offline analysis and is computed within a cone of $\Delta R = 0.3$ or 0.4 around the electron direction, and the hit pattern and impact parameter, which discriminate against converted photons, are used for the electron selection.

The efficiencies for electron reconstruction, identification and isolation are also measured using the tag-and-probe method in $Z \rightarrow e^+e^-$ events. Comparing the results of tag-and-probe in data and simulation scale factors are obtained. The method requires one electron candidate, the "tag", to satisfy tight selection requirements. Different criteria are employed to define the tag electron, and it is found that the estimated efficiencies are almost insensitive to any specific definition of the tag. The probe is the second electron candidate. The efficiencies measured with 2017 data are shown in figure 3.4 (C) for a working point corresponding to 90% of signal efficiency including a selection on the isolation.



(A) (B) (C)

FIGURE 3.4: Distributions in the distance $\Delta\eta$ between the position of the ECAL superclusters and the extrapolated track (A), of $\sigma_{\eta\eta}$ (B) for barrel electrons at 8 TeV. Electron identification efficiency in data (C,top) and data to MC efficiency ratios (C,bottom) measured for the multivariate identification (Boosted Decision Trees) working point of 90% of signal efficiency including the cut on isolation. The error bars on the data/MC ratio represent the combined statistical and systematic errors [69].

3.2 Jets and Missing energy

As a result of QCD confinement particles carrying a color charge, such as quarks and gluons, cannot be observed free. Quarks and gluons produced in proton-proton collisions or in unstable particles decays are thus reconstructed as jets, i.e. a set of stable or almost stable hadrons and their decay products collimated in a narrow cone.
Jets originate not only from the hard scattering, but also from the underlying event and pileup collisions. These jets are present everywhere, even if the final state of the hard scattering has no quarks or gluons. Depending on the event topology and on the kinematics of these jets, they can be used in the analysis or tagged as pileup.
The missing energy is used to measure particles escaping the detector, such as neutrinos.

At the LHC the transverse component of the missing momentum, or missing transverse energy, is measured. The missing transverse energy resolution is dominated by the hadronic jets, whose resolution is worse than lepton and photon resolution.

3.2.1 Jets

In CMS jets are reconstructed from particle flow objects using the anti-k_T clustering algorithm [70], as implemented in the FASTJET package [71, 72]. The algorithm belongs to a class of sequential recombination algorithms, which includes also the k_T and Cambridge-Aachen algorithms [73, 74]. These three methods are prevalent in high energy physics nowadays and are adopted depending on the different analysis strategies. All the algorithms of this family are infrared and collinear safe: infrared safeness means that the results of the jet clustering are not altered if an arbitrary number of extra particles with momentum that tends to zero is included in the jet clustering; collinear safeness means that the clustering is not sensitive to a splitting of a particle into two collinear ones each taking a fraction of the momentum.

In these algorithms, for each pair of objects to be clustered (the PF objects for CMS, but also tracks or calorimeter clusters) the distance

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

is computed. ΔR_{ij} is defined as $\sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$ and ΔR is a fixed distance parameter which defines the cone typical amplitude. The pair that minimizes the distance d_{ij} is merged to form a new object. The distances are then computed with the new set of objects. At each step the pair that minimizes d_{ij} is merged, until a stopping condition is reached.

The sign of n characterizes each algorithm: in the k_T algorithm n is 1, in the Cambridge-Aachen it is 0, while anti-k_T algorithm it is -1. As a result, the anti-k_T algorithm clusters radiation around high p_T objects first, so that conical jets of radius equal to the distance parameter ΔR are typically produced, unless multiple hard objects are clustered.

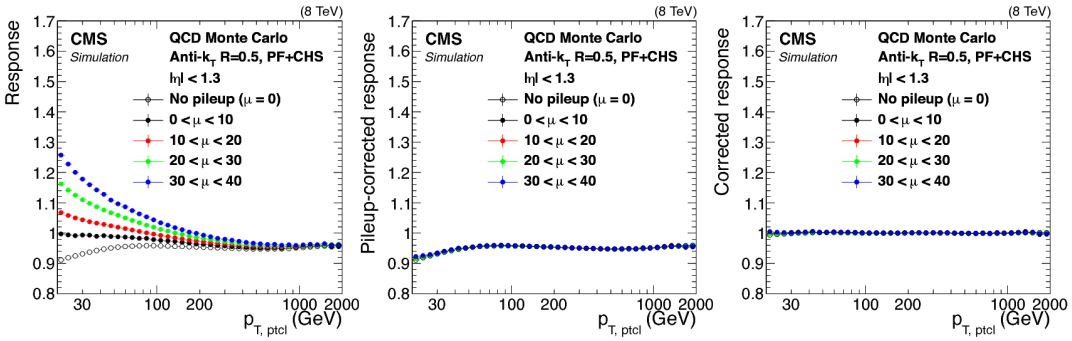
The standard clustering distance used at the CMS is $\Delta R = 0.4$ since Run 2, while it used to be 0.5 at 8 TeV. At the same time, larger jets, with a ΔR of 0.8, are used when looking for boosted heavy particles decaying into hadrons.

Jets built using the PF reconstruction have good angular resolution ($\delta\eta$, $\delta\phi$ in range 0.3-0.01 for p_T < 100 GeV, $\sim$ 0.01 for p_T > 100 GeV), while the energy needs to be corrected in multiple steps.
The corrections are called "jet energy corrections" (JEC), and are applied to the 4-momentum magnitude. They are usually derived using the jet transverse momentum, which is corrected as a function of the jet η and p_T.
CMS has adopted a factorized solution to the problem of jet energy corrections, where each level of correction takes care of a different effect. The approach is based on the final calibration at 8 TeV, documented in [75]. The jet calibration is repeated for each data taking year, in order to take into account the different conditions. For our analysis we use 2017 specific corrections.
The corrections can be briefly summarized as:

- **Pileup offset corrections**, determined in simulation. First, tracks coming from pileup vertices are removed (charged hadron subtraction, CHS). Then an offset correction is applied to account for residual contamination, determined from the per-event median energy density ρ computed with the k_T clustering algorithm. The corrections are parametrized as a function of η and p_T and the jet area. The pileup correction, jet area and ρ definitions are based on reference [76].
- **Simulated response corrections**, determined in simulation as a function of η and p_T by evaluating the response with respect to generator level jets.
- **Residual corrections**, determined in data and comparing data to simulation. The correction is derived in two main steps. The jet energy response is corrected as a function of η relatively to the better calibrated barrel region ($|\eta| < 1.3$), these being usually called "relative" jet corrections. Then the jet momentum is scaled in order to match a reference object within jet $\eta < 1.3$: these are usually called "absolute" corrections.

Residual data/MC corrections are measured in dedicated event topologies. Relative corrections are measured using dijet events, assuming p_T balance and absolute corrections are calibrated on data using the balancing of Z/γ+jet events. Analogous topologies are used to measure the difference in resolution between jets and simulation as a function of η and p_T and correct for it (JER correction). An analogous strategy is followed to apply at analysis level the corrections for the b jets momenta developed specifically for the VH($b\bar{b}$) analysis.

The impact of Jet Energy Corrections (JECs) on the jet response in simulation, without residual corrections, is shown in figure 3.5 for Run 1 at 8 TeV.



(A) (B) (C)

FIGURE 3.5: Ratio of measured jet p_T to particle-level jet p_T in QCD MC simulation at various stages of JEC: before any corrections (a), after pileup offset corrections (b), after all JECs (c). Here μ is the average number of pileup interactions per bunch crossing [75].

The typical jet energy resolution after applying the energy corrections is 15–20% at 30 GeV, about 10% at 100 GeV, and 5% at 1 TeV in the central rapidity region for jet clustered with $\Delta R = 0.4$.

Track Jets

Jets clustered using only tracks [77, 78] are also reconstructed. They are used to discriminate peculiar event topologies, usually when the hard scattering involving only the electroweak interactions from several background with larger additional hadronic activity.

Since the amount of additional radiation is expected to be soft, aiming to avoid the contributions from pileup interactions, only the charged tracks that clearly originate from the event primary vertex are used.
The tracks used for the soft activity jets are required to be high quality tracks and to have p_T > 300 MeV, have the minimum dz(PV) when associated with the signal primary vertex in the event and satisfy the condition dz(PV) < 2 mm.

After this track selection, a collection of "soft track-jets" is built clustering the above tracks with the anti-k_T clustering algorithm, with distance parameter $\Delta R = 0.4$. For analysis purposes, the variables most often used are the Soft-$H_T = \sum_{softjet} p_T$ and the number of soft jets above a certain p_T threshold.

3.2.2 Missing transverse energy

The missing transverse energy, or p_T^{miss} is defined as the negative vector p_T sum of all the PF candidates in the event. The p_T^{miss} therefore relies on the accurate measurement of all the PF objects: leptons, photons, jets, and unclustered energy, which is the contribution from the PF candidates not associated with any of the previous physics objects.
The reconstructed objects with worse resolution are the jets, so the jet energy corrections have to be propagated to the p_T^{miss} in the following way:

$$p_T^{miss} = p_T^{miss,\,raw} - \sum (p_{T,jet}^{corr} - p_{T,jet})$$

where $p_T^{miss,\,raw}$ is the uncorrected p_T^{miss}. The sum is over jets with p_T > 15 GeV.

To remove the overlap of jets with electrons and photons, jets with more than 90% of their energy associated with the ECAL are not included in the sum. In addition, if a muon reconstructed using the outer tracking system overlaps with a jet, its four momentum is subtracted from the four momentum of the jet, and the jet energy correction appropriate for the modified jet momentum is used in the p_T^{miss} computation.

To estimate the p_T^{miss} uncertainty, the uncertainties in the momenta of the all reconstructed objects are propagated to p_T^{miss} by varying the estimate of each PF candidate transverse momentum and recomputing p_T^{miss}. The JEC uncertainties are less than 3% for jets within the tracker acceptance and reach up to 12% for those outside. The jet energy resolution (JER) uncertainties typically in the range 5-20%. The uncertainty in the unclustered energy is evaluated based on the momentum resolution of each PF candidate, which depends on the type of the candidate. The largest contributions to the unclustered energy uncertainty are due to the PF neutral hadrons and PF candidates in the forward calorimeter (HF). As it depends on several objects, the total uncertainty on the p_T^{miss} measurement is expected to vary with the event topology.

The p_T^{miss} is sensitive also to the entire detector readout. For this reason dedicated filters are applied to remove anomalous high-p_T^{miss} events, which can arise because of a variety of reconstruction failures or detector noise and inefficiency. The filters include:

- **Calorimeter filters**, based on the isolation, electronic signal shape and timing in HCAL, or on noisy crystals and non-functioning electronics in the ECAL.
- **Beam halo filters**, which remove events with energy deposits along a line with constant ϕ, sometimes matching the muon CSC signals attributed to beam halo particles.

- **Reconstruction filters**, which veto events with poorly reconstructed high energy charged hadrons or muons, by looking at tracks with low quality and very high-p_T.

The effect of the filtering can be observed in figure 3.6, in the event topology with at least one jet and high p_T^{miss} (>250 GeV).

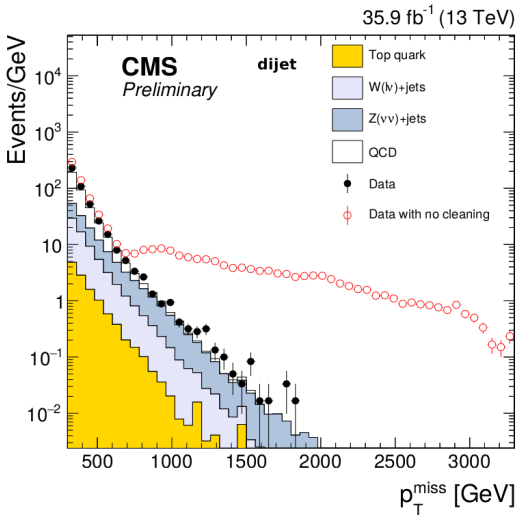


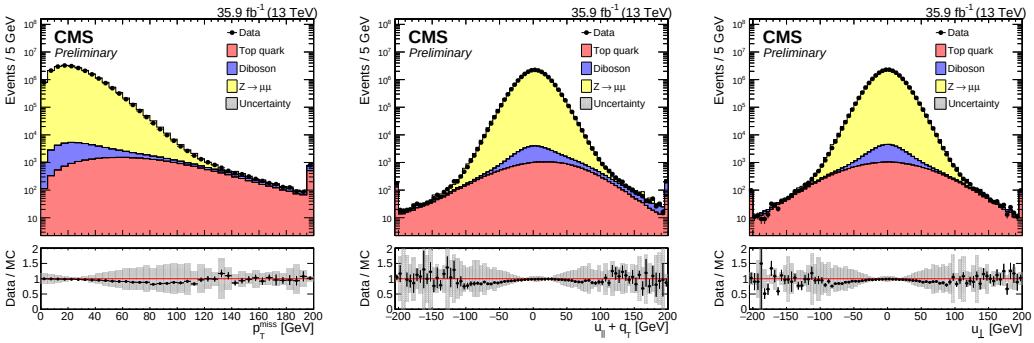
FIGURE 3.6: The p_T^{miss} without the event filtering algorithms applied, with the event filtering algorithms applied to data, and in simulation [79]..

The p_T^{miss} performance and resolution are usually measured in events with Z/γ + jets, where no genuine p_T^{miss} is expected. In such events the p_T^{miss} resolution is dominated by the hadronic activity, since the momentum resolution for leptons and photons is order 1%, compared to 5–20% for the jet momentum resolution. The p_T^{miss} for $Z \rightarrow \mu\mu$ + jets events is shown in figure 3.7 (A), the peak value is at about 30 GeV. In the $Z \rightarrow \mu\mu$ + jets topology one can also measure the performance in p_T^{miss} by comparing the momenta of the vector boson to that of the hadronic recoil system. The hadronic recoil system is defined as the vector p_T sum of all PF candidates except for the muons from the Z. The hadronic recoil vector is projected along the Z axis to evaluate the response, and the orthogonal component, as well as the difference between the recoil along the axis and the Z transverse momentum provide a similar estimate of the p_T^{miss} resolution, as shown in figures 3.7 (B,C). The intrinsic resolution, after removing the pileup contribution, is ~10 GeV for both the components of the recoil.

Other variables used at the analysis level include the p_T^{miss} computed using only the charged particles, called track MET or track p_T^{miss}, and the MHT defined as the negative vector sum of the p_T of the jets above a given threshold, which is useful in particular in the trigger.

Another important variable used at the analysis level is the p_T^{miss} significance. The significance is used to distinguish between events with genuine missing energy and those with spurious missing energy. The p_T^{miss} significance S is defined as a log-likelihood ratio with the two hypotheses given by the p_T^{miss} equal to the observed one and no genuine p_T^{miss}.In the reasonable Gaussian approximation S is a χ^2 distributed variable with two degrees of freedom, one for each p_T^{miss} component, and can be computed as

$$S = (\vec{p}_T^{miss})^T V^{-1} (\vec{p}_T^{miss}),$$



(A) (B) (C)

FIGURE 3.7: Distribution of p_T^{miss} (A) of the two orthogonal components of the hadronic recoil (B) and (C) in the Z→ $\mu\mu$ +jets topology. The points in the lower panel of each plot show the data to simulation ratio. The systematic uncertainties due to the jets and the unclustered energy are shown in the shaded band [79].

where V is the 2×2 covariance matrix of the total missing transverse energy obtained by propagating the uncertainties of all the hadronic objects in the event.

3.3 Identification of b jets

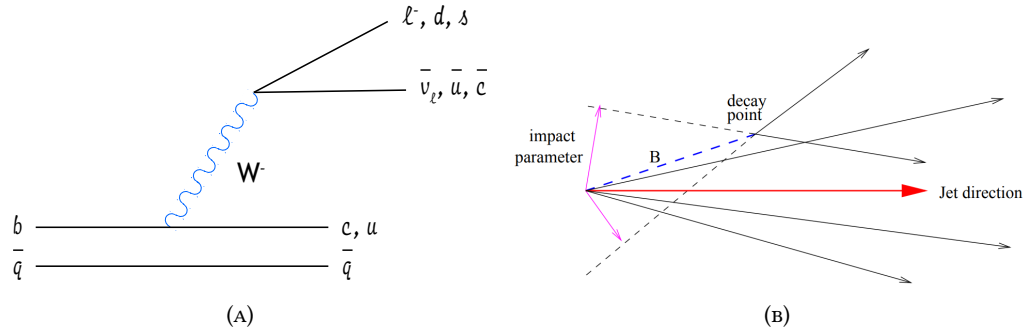
The most important tool used to tag signal jets in the VH($b\bar{b}$) analysis is the good capability of identifying b jets with the CMS detector. Jets are by far the most common object at hadron colliders. Jets from b quarks look overall very similar to the other jets and need no special treatment in their reconstruction. However, a B hadron is produced in the hadronization process of a b quark, and the reconstruction of its decay products inside the jet is the key to tag the b jets.

Jets containing B hadrons can be distinguished thanks to the B hadron long lifetime: $c\tau \simeq 500\ \mu m$. A B hadron with p_T = 50 GeV flies on average almost half a centimeter (L $\simeq \gamma c\tau$) after being produced.
The relatively long lifetime of B hadrons is due to the need for b quarks to decay weakly into lighter quarks (figure 3.8, (A)). The top quark final state would be favored, but it cannot be accessed kinematically due to the mass of ∼ 170 GeV of the top quark. The transition to lighter quarks, belonging to the second or the first family, comes with a sizeable suppression factor[1] and results in a longer lifetime. Often B hadrons decay into charmed D hadrons, which have in turn non negligible lifetimes ($c\tau \simeq 300\ \mu m$), so full decay chains can be found and in some cases reconstructed within b jets.

The B hadrons' long lifetime results in a sizeable impact parameter of the decay products with respect to the primary vertex, which are reconstructed, if charged, as tracks (figure 3.8, (B)). A secondary vertex or multiple secondary vertices can be reconstructed and their properties are highly discriminating variables.

Another property of the B hadron decay is the relatively high rate of lepton production from semileptonic decays (around 25%). These leptons can be identified thanks to their relatively

[1] Weak decays of the quarks can result in different family in the final state, because of the mixing of weak interaction eigenstates and mass eigenstates. The mixing components, i.e. the out of diagonal elements in the CKM matrix, are however small, i.e. < 10−2, hence the reduced decay rate.



(A) (B)

FIGURE 3.8: Representation of a b quark decay inside a B hadron (A). Representation of a B hadron decay inside a jet (B).

high p_T relative to the B flight direction, approximated by the b jet direction: leptons from B decays have order of GeV momenta relative to b jet direction, because of the B hadron mass ($\sim$ 5 GeV), while leptons in jets of other flavors tend to be closely aligned with the jet.

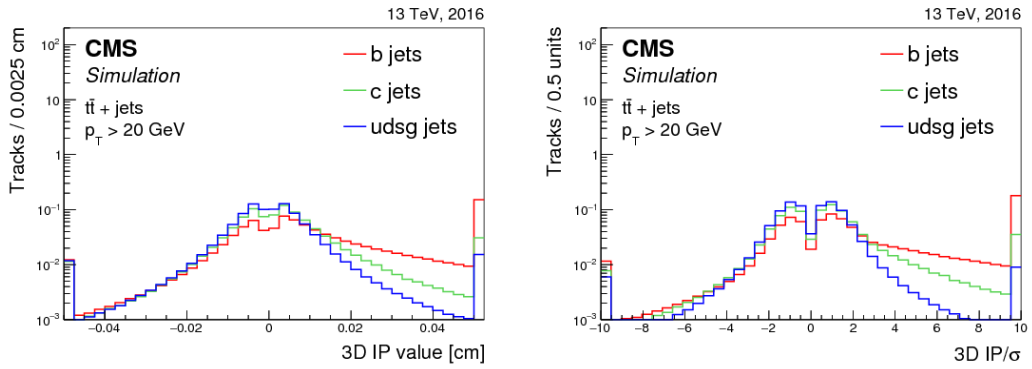
Jet b tagging observables

Jet b tagging relies primarily on the measurement of the impact parameters (IP) of the tracks associated with the jet with respect to the primary vertex. Track used for b tagging are selected by a dedicated algorithm requiring high quality tracks. Standard track requirements for b tagging in CMS include a p_T above 1 GeV, a minimum number of hits in the tracker and in the pixel detector, a loose compatibility with the primary vertex, a maximum distance from the jet axis and a maximum distance of the point of closest approach of the track to jet axis from the primary vertex. All the requirements help reject pileup and misreconstructed tracks.

The IP is independent, at first order, from the momentum of the B hadron because the angle between the track and jet directions is roughly proportional to $1/\gamma$, and the displacement of the secondary vertex is proportional to γ. The IP resolution for CMS tracks is $\sim$ 100 μm both in the transverse and longitudinal direction. The IP value used for tagging is both in three spatial dimensions (3D) or in the plane transverse to the beam line (2D). The signed IP is also used: the sign is positive (negative) if the scalar product of the jet direction with the impact parameter direction is positive (negative). The sign of the track impact parameter is expected to be positive if the track originates from the decay of a hadron outgoing from the primary vertex along the jet direction. The tracks from a B hadron decay are therefore expected to have positive impact parameters.
A commonly used variable to tag b jets is the significance of the track impact parameter (SIP):

$$\text{SIP} = \frac{\text{IP}}{\sigma_{\text{IP}}}$$

where σ_{IP} is the IP uncertainty. This observable takes care also of mismeasured tracks, which can have artificially large impact parameters. Figure 3.9 shows the signed 3D IP (A) and signed 3D SIP for tracks associated with jets of different flavors.
Other observables which are important for b tagging are related to secondary vertices (SV) associated with the jet. Two algorithms are used to reconstruct the secondary vertices in CMS. The first one uses the Adaptive Vertex Reconstruction (AVR) algorithm [80, 62] taking as input tracks clustered in the jet. The other algorithm, used more frequently since the LHC Run 2 is the Inclusive Vertex Finder (IVF) algorithm. In contrast with AVR, IVF uses as input all reconstructed tracks in the event with $p_T > 0.8$ GeV and a longitudinal IP <



(A) (B)

FIGURE 3.9: Distribution of the 3D impact parameter value (A) and significance (B) for tracks associated with jets of different flavors in $t\bar{t}$ events. Tracks are shown for jets with p_T > 20 GeV .

0.3 cm. This vertex finder was first used for low p_T b hadrons to resolve close-by decays [81].

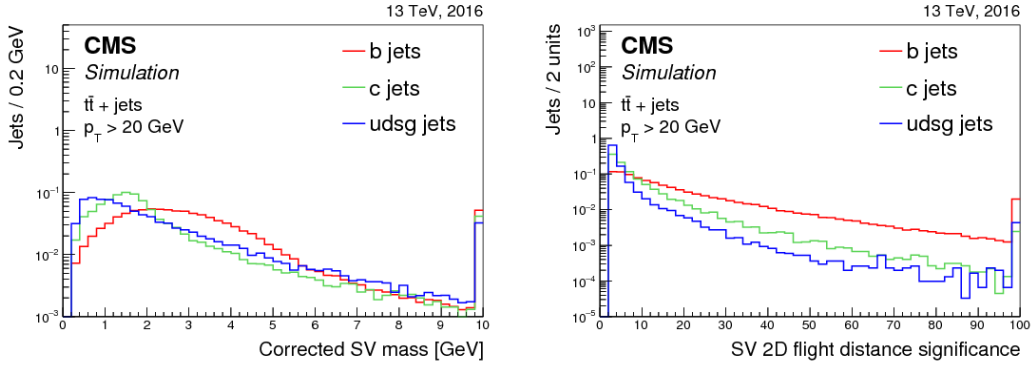
The IVF involves multiple steps which build and gradually clean the reconstructed vertices. The preliminary step of the IVF sequence is the seeding of the secondary vertices: tracks with a 3D impact parameter value of at least 50 μm and a 2D impact parameter significance of at least 1.2 are taken as seeds.

After a seed is found the algorithm includes four steps. Tracks are first clustered around seeding tracks: the compatibility between a seeding track and another track is evaluated using requirements on the distance at the point of closest approach (PCA) of the two tracks and the angle between them. A selection on the distance at PCA depending on the distance from the primary vertex is also applied. Subsequently, a fit to the track cluster is performed using the Adaptive Vertex Fitter and vertices close to the primary, as determined by their significance, are discarded. In case two vertices share a large fraction of the tracks, one can be kept and the other is discarded at this point. Next, tracks are reassigned based on the compatibility with the primary or the secondary vertex (track "arbitration"). Finally, the secondary vertex position is refitted after the track arbitration, if there are still two or more tracks associated with it. After refitting the secondary vertex positions, a second vertex cleaning is performed.

The secondary vertices are associated with the jets by requiring the angular distance between the jet axis and the secondary vertex flight direction to satisfy $\Delta R < 0.3$.
Among the SV related observables, the flight distance and direction, i.e. the vector between primary and secondary vertex, the SV mass and energy are included in b tagging algorithms. Figure 3.10 shows the mass and the 2D flight distance significance for secondary vertices associated with jets of different flavor.

CMS standard b tagging algorithms

Jet b tagging is one of the areas where Machine Learning is fundamental to have optimal performances. CMS standard algorithms, optimized with Machine Learning, rely both on secondary vertices and tracks. These algorithms were developed after and in parallel with simpler algorithms, based on a single observable or one type of observables, which are still useful to monitor the main observables. A description of those is first given.



(A) (B)

FIGURE 3.10: Distribution of the corrected secondary vertex mass (A) and of the secondary vertex 2D flight distance significance (B) for jets containing an IVF secondary vertex. The distributions are shown for jets of different flavors from $t\bar{t}$ events. Secondary vertices are shown if in jets with $p_T >$ 20 GeV .

Among the single-variable based algorithms the Track Counting (TC) algorithm sorts tracks in a jet by decreasing values of the IP significance. The first track has little discriminating power, however, the probability to have several tracks with high positive values is low for light-flavor and gluon jets. Therefore the two versions of the algorithm use the IP significance of the second and third ranked track as the discriminator value. The two versions are called Track Counting High Efficiency (TCHE) and Track Counting High Purity (TCHP) algorithm, respectively.

A natural extension of the TC algorithms is the combination of the IP information of several tracks associated with a jet in a cone of $\Delta R < 0.3$. Two discriminators are computed: the Jet Probability (JP) algorithm uses an estimate of the likelihood that all tracks associated with the jet come from the primary vertex while the Jet B Probability (JBP) algorithm gives more weight to the 4 tracks with the highest IP significance, with the number 4 chosen as it matches the average number of reconstructed charged particles from B hadron decays. The probability for a single track to originate from the primary vertex, P_i, is computed integrating resolution histograms derived in data and simulation using tracks with negative signed IP. The final probability is then computed as

$$P_{jet} = \Pi \cdot \sum_{i=0}^{n-1} \frac{(-\ln\Pi)^i}{i!} \qquad \text{where} \qquad \Pi = \prod_{i=1}^{n} \max(P_i, 0.005).$$

The probability is set to 0.5% for track probabilities below 0.5% in order to avoid a single track to drive P_{jet} close to zero.
Other single observable based discriminators use the secondary vertices. The Simple Secondary Vertex (SSV) algorithms use the significance of the SV flight distance as discriminating variable. If several vertices are present the one with the highest significance is used. As for the Track Counting algorithms, two SSV versions optimized for different purity exist: the High Efficiency version (SSVHE) uses vertices with at least two associated tracks, while for the High Purity version(SSVHP) at least three tracks are required.

This set of algorithms was developed before the beginning of the LHC data taking. Already during Run 1 multivariate discriminators with Machine Learning techniques were

employed: the CSV algorithm, which combines secondary vertices and tracks via a likelihood ratio was first developed. The performance of these algorithms is shown in figure 3.11 (A). A receiver operating characteristic curve, or ROC curve, comparing the efficiency of b jets and the mistag of light-flavor jets for the $t\bar{t}$ 7 TeV simulation is shown. The CSV algorithm outperforms all the algorithm, including JP.

Better performing versions of the CSV algorithms were developed afterward. The CSVv2 requires at least 2 tracks per jet compatible with the primary vertex. Additionally, any combination of two tracks compatible with the mass of the K_s^0 meson is rejected. The training of the algorithm is then performed in three independent vertex categories. The first vertex category contains jets with at least one associated reconstructed secondary vertex. The second, called "pseudo vertex", contains jets whose tracks with an IP significance larger than 2 can be combined in a pseudo vertex, allowing for the computation of a subset of SV observables. Otherwise, a "no vertex" category with track-based variables only is defined.

The variables used for the training include secondary vertex observables (2D flight distance significance, mass, number of tracks, energy and transverse momentum ratio with respect to the jet, etc.), variables relative to the track with the highest 2D SIP (η_{rel}, $p_{\text{T}}^{\text{rel}}$, decay length, etc.), the 3D SIP of the first four tracks, variables relative to the sum of the selected tracks, and the jet η, p_{T}. In Run 2 [82], the training was performed using a shallow neural network for each category, and separately for b jets versus light-flavor jets and b jets versus c jets. The outputs were then combined via likelihood ratios among the categories and a final re-weighting, with relative weights of 1:3 for the b versus c and b versus light-flavor jets, respectively.

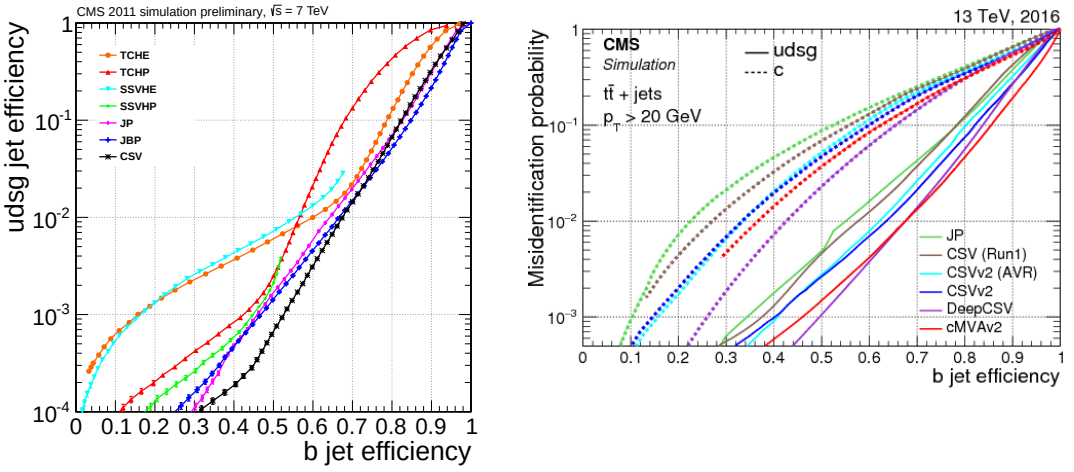
Finally, during Run 2 Deep Learning was introduced for b tagging. The DeepCSV algorithm was developed using a deep feed-forward neural network (see 4). The DeepCSV algorithm uses the same information as the CSVv2 one, but the training is performed using more events and a more flexible algorithm. This solves the entire b tagging problem in one step, i.e. a training including all categories and all jet flavors, and allows improved performances. The performance of b tagging algorithms used at CMS at 13 TeV are shown in figure 3.11 (B). The CMVA algorithm, which uses also leptonic decays information to improve on top of the CSVv2, but is outperformed by DeepCSV for high purity working points, is also shown.

Performance in data

The b tagging optimization relies on simulation for the optimization. After a discriminator is obtained the efficiency is measured in data for different working points for b, c and light quark jets. Scale factors are derived for the different jet flavors in bins of η and p_{T} by comparing data and simulation. Additionally, analyses that use the discriminator shape in multivariate analysis, like VH($b\bar{b}$), need a full reshaping of the b tagging discriminator in order to describe the data correctly.

In this case the scale factors are derived in bins of p_{T}, η only for light-flavor jets, and the discriminator value itself. The algorithm used to derive the scale factors is called "IterativeFit" [82]. The technique relies on two orthogonal regions which are very pure in jets of a particular flavor and where the scale factor for that flavor are measured.

The b jet scale factors are measured in a region enriched in dileptonic $t\bar{t}$, by requiring two same sign opposite charge leptons with invariant mass at least 10 GeV away from the Z boson mass, and exactly two jets. One jet is required to be b-tagged (tag jet), while the second



(A) (B)

FIGURE 3.11: Misidentification probability for c and light-flavor jets versus b jet identification efficiency for various b tagging algorithms applied to jets in $t\bar{t}$ events. [83, 82].

is used as a probe. The region is about 85-90% pure in $t\bar{t}$, depending also on the algorithm used to select the tag jet. The light-flavor scale factors jets are measured in a region enriched in Z+jets, and with exactly two jets, among which one is required to fail a b tagging criterion and the other is used as a probe. The region is 99.9% pure in Z + light-flavor jets.

The procedure is called "IterativeFit" because the distributions of the minority flavor jets (as expected from simulation) in the regions enriched in the opposite flavor are rescaled at each step with the scale factor obtained in the other region at the previous iteration and subtracted before computing the scale factor. Charm flavor scale factors are not considered in this procedure, but the c flavor contribution expected from simulation is subtracted from both regions.

In this method, the following systematic uncertainties are considered:

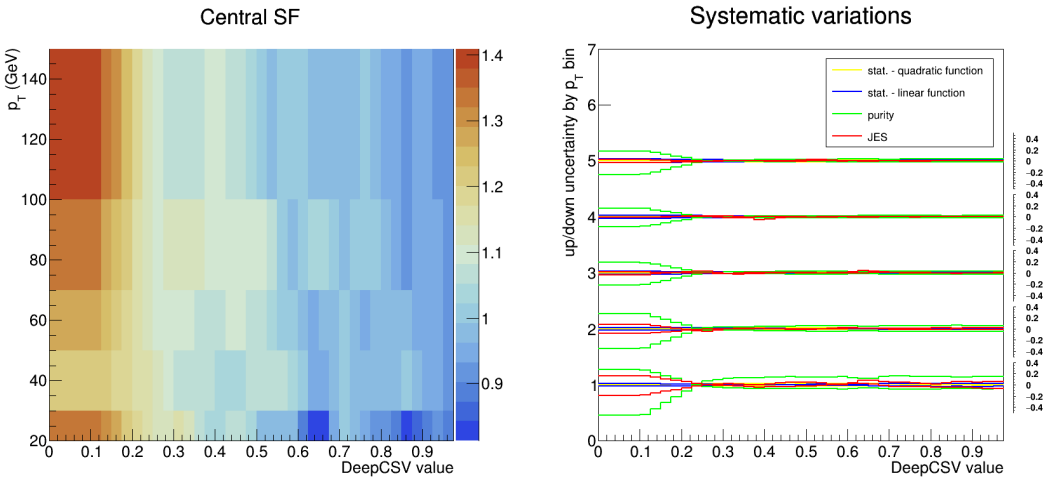
- Sample purity: the fraction of heavy-flavor jets in the sample is conservatively varied upwards and downwards by 20% when calculating the scale factor for light jets; the same variation is applied to non-b jets when measuring the b scale factor, which represents a realistic variation of the b jet purity.

- Jet energy scale: since the measurements are performed in bins of jet p_{T}, the fraction of jets in each bin may vary depending on the jet energy corrections. The scale factors are remeasured after varying the jet energies by 1 standard deviation about the nominal jet energy correction. The systematic effect due to this variation is less than 1%.

- Statistical uncertainty: the statistical uncertainties are treated using two functions (a quadratic and a linear one), representing two independent bias effects on the discriminator bin content. The scale factor value is varied in each discriminator bin according to $\sigma(x) \times f_i(x)$, where $\sigma(x)$ is the statistical uncertainty in the scale factor in a given bin and $f_i(x)$ is the linear or quadratic function value as a function of distance from the discriminator bin center. This allows us to obtain an envelope around the nominal scale factor.

For c jets the scale factor is set to unity. The scale factor uncertainty is obtained by doubling the aforementioned b jet scale factor relative uncertainties and adding them in quadrature to obtain a global relative uncertainty, which is in turn multiplied by a linear and a quadratic function, yielding two uncertainty sources.

The b jet scale factors obtained for 2017 data and used for the VH($b\bar{b}$) analysis are shown in figure 3.12 (A). The systematic uncertainties are shown in (B). The uncertainties are dominated by the jet flavor purity in the discriminator bins where the b flavor is disfavored ($\gtrsim$ 10%).



(A) (B)

FIGURE 3.12: Scale factor for b jets as a function of jet p_T and DeepCSV value for the scale factors used in the VH($b\bar{b}$) analysis (A); uncertainty on the b jet scale factors (B).

3.4 Pileup treatment

The presence of pileup interactions affects the reconstruction of jets in general, in particular the jet momentum resolution, and the use of vetoes on additional jet activity. It also affects the p_T^{miss} reconstruction, lepton isolation and b tagging. The effects on the jets and p_T^{miss} are addressed by the first step of the jet energy corrections, aimed specifically at pileup removal. Pileup is also taken into account when computing the isolation and for the b tagging track selections.

The problem of jets due to pileup collisions is also addressed: a multivariate technique to reject such pileup jets has been developed and applied to CHS jets. The identification of pileup jets is based on two characteristics of such jets. First, the majority of tracks associated with pileup jets do not come from the primary vertex, and secondly, pileup jets are clustered from particles originating from multiple collisions and tend to be broader than jets originating from one single quark or gluon. The multivariate discriminator therefore includes variables like the number of vertices in the event, the fraction of transverse momentum from particles coming from the primary, the energy spread and energy fractions in rings about the jet axis.

A loose working point is usually employed for pileup identification, corresponding to 99% efficiency for quark jets in the region at $|\eta| < 2.5$, 95% efficiency for quark jets at $|\eta| > 2.5$.

The purity is ∼95% inside the tracker acceptance, while in the region at $|\eta| > 2.5$ it drops to 30-40%, as shown in figure 3.13.

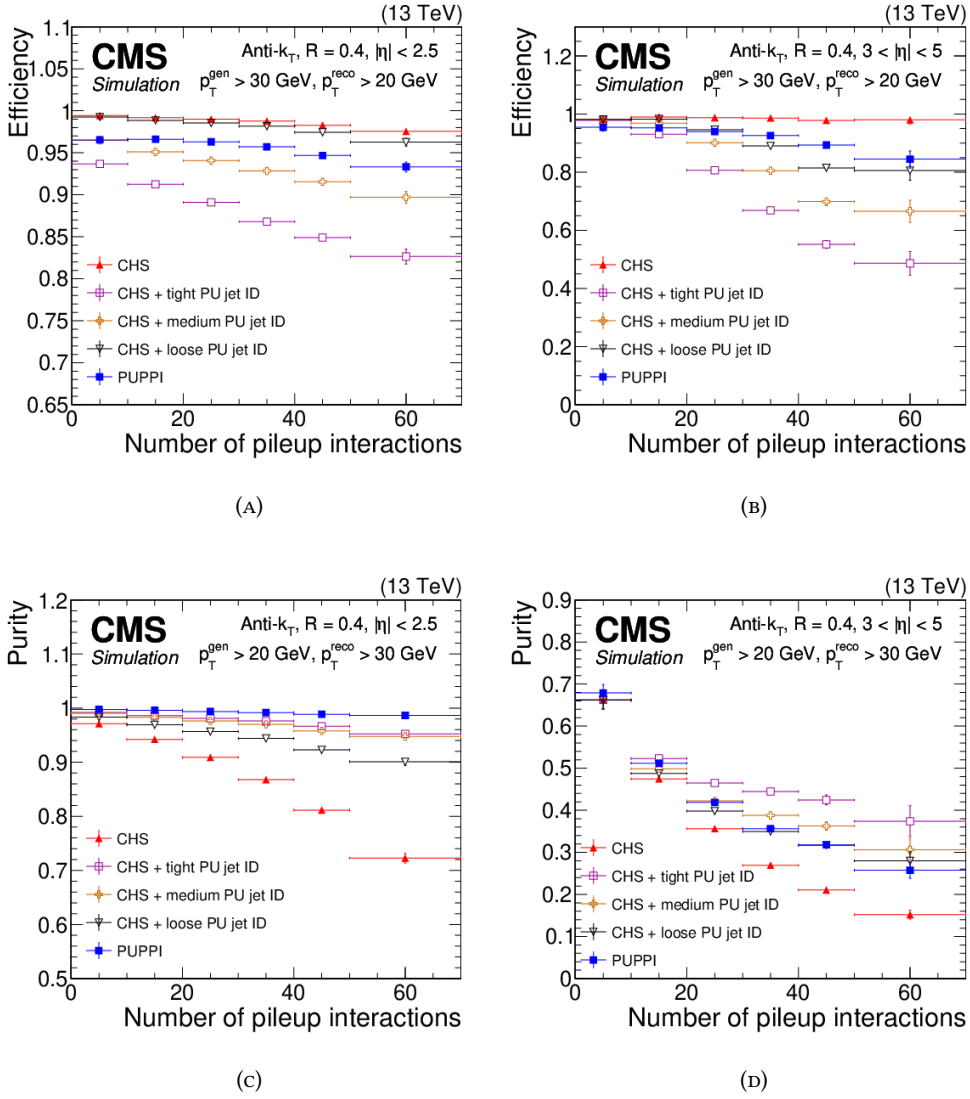


FIGURE 3.13: Leading vertex jet efficiency and purity in Z+jets simulation as a function of the number of pileup interactions. Plots are shown with AK4 jets having a p_T > 20 GeV and (left) $|\eta| < 2.5$ and (right) $|\eta| > 3$. ([84]) .

3.5 Trigger objects and PF objects

Objects used at the analysis level reconstructed with the best algorithms while for the ones used at the trigger level some parts of the reconstruction are removed or simplified in order to reduce the run time.

The L1 trigger uses the readout of the calorimeters and the muon system only. The muons are identified and their momentum measured with the muon chambers. The electrons and photons are reconstructed with ECAL clusters, with no distinction. The jets (and as a consequence the missing energy) are measured as "CaloJets", by clustering calorimeter energy readout. At the HLT the objects are reconstructed in steps of increasing complexity, so that

the most time consuming algorithms are run only on the events passing the previous requirements on simpler objects: the first step of the HLT may again be only calorimeter or muon-detector based. Tracking is greatly simplified in the HLT reconstruction: pixel only tracks, which are the tracking seeds, can be used. The primary vertex reconstruction can be based on the pixel tracks and not on the full tracks, and the knowledge of the beam spot is often necessary. In case more precise measurements are needed, pixels and strips are used, but with strategies aimed at reducing the run-time: regional tracking in detector areas of interest, reconstruction with a limited number of hits, and not all the iterations of tracking are usually run. HLT tracks are used also for the particle flow sequence which comes last in the trigger paths. The reconstructed PF objects are very similar to offline ones, so tighter cuts can be applied at this level.

Relevant for the VH($b\bar{b}$) analysis are the trigger paths using muons, electrons and p_T^{miss}. The muon and electron reconstruction at the HLT is very similar to the offline one. The efficiency is mainly limited by the L1 seeding and the isolation with respect to the offline. The missing energy reconstruction is based on the PF jets, but only the higher p_T jets are corrected to save computing time. The MHT is also used in the trigger as it's not modified by the lower p_T jets. In the offline analysis the efficiency is measured both in data and simulation as a function of the trigger requirements, allowing us to determine the threshold of full efficiency and to correct for discrepancies between data and simulation.

Chapter 4

Machine Learning and Deep Learning

This chapter introduces the Machine Learning and Deep Learning concepts applied in chapters 5 and 6. The work presented in this thesis was started when Deep Learning was relatively new in High Energy Physics, therefore some of the very first applications of Deep Learning in CMS publications are presented. Deep Learning is used for supervised learning tasks, training on simulated samples where the target is known. Deep learning is applied both to the global event topology and to jets, more specifically b-jets. Jets can be analyzed both by looking at their global properties and as a set of clustered particles. In the latter case, jets are complex objects, made up of simpler objects each with its own features, and Deep Learning techniques involving parameter sharing, used also for sequence processing, become very useful. The chapter structure and the content are partly inspired by reference [85], used as a guidance when approaching Deep Learning.

4.1 Introduction

Machine Learning (ML) is a field of computer science that uses statistical techniques to give computer systems the ability to learn from data. ML is suitable to solve problems that are difficult to describe with a set of rules. It can be considered instead as a form of applied statistics, which allows the approximation of complicated functions based on the input data distribution. A general and synthetic definition of Machine Learning is the one given by Mitchell in ref. [86]. Paraphrasing the definition: "a computer program is said to learn from experience with respect to some class of tasks and performance measure, if its performance at a task improves with experience".

The most common tasks for which ML was and is employed are e.g. classification and regression.

- In classification the computer algorithm has to separate data in different categories. To solve this task, the learning algorithm is set to produce e.g. a function $f : \mathbb{R}^n \rightarrow [0, 1]$ in the case of two categories.
- In regression tasks the computer program is asked to predict a numerical value given some input. To solve this task, the learning algorithm is asked to output a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

Other tasks, including unsupervised ones, such as clustering of unlabeled data, can be very different depending on the application domain and data types.

The performance metrics measure how well the algorithm performs on the task. The performance metrics must be distinguished from the loss or cost function. Both the performance

and the loss function measure the distance of the output value from the truth. However, the metric is chosen depending on the actual task, while the loss is the function of the parameters minimized at training time. In case the loss is minimized via gradient descent, its choice is based also on the possibility and the ease of calculating the derivatives.

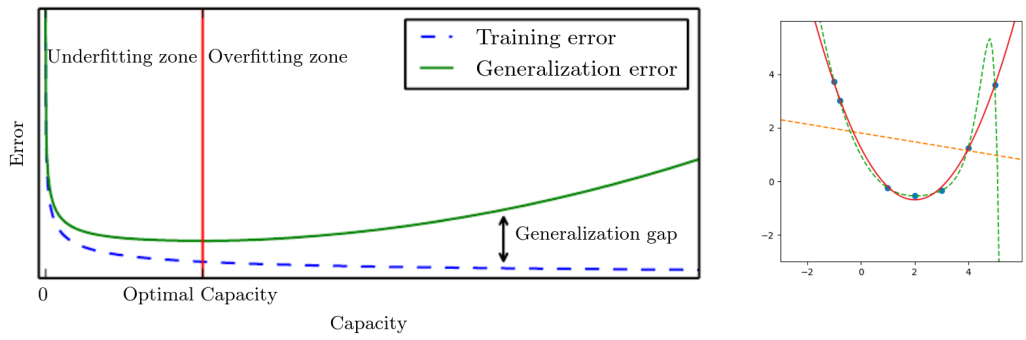
The experience is based on datasets similar to the ones we want to apply the algorithm to. The ML algorithm is usually trained for the task by "seeing" a dataset of examples, or data points. Sometimes multiple passes of the dataset are necessary for the optimization.

Generalization properties

Training a ML model reduces the cost or training error and improves the performance measure for the task: this also called optimization of the ML algorithm. One of the challenges for all ML algorithms is to achieve good performance on data not previously observed. The ability to perform well on new data points is called generalization.

In order to evaluate the generalization, a distinct dataset, called test set, is used. The test set data are assumed to be sampled from the same parent distribution of the training set and to be statistically independent. The difference between training error and test error is also called generalization gap. A successful ML algorithm needs to have a small generalization gap.

The minimization of the training error and the generalization gap corresponds to the two central challenges in Machine Learning: underfitting and overfitting (figure 4.1 (A)). Underfitting occurs when the model is not able to obtain a sufficiently low error value on the training set. Overfitting occurs when the gap between the training error and test error is too large.



(A) (B)

FIGURE 4.1: Typical error versus model capacity relationship. As the capacity increases a trained model can move to the underfitting regime to the overfitting regime, where the generalization gap increases (A). Image from [85]. Example of polynomial fit to data points generated with a parabolic model with low (yellow dashed), optimal (red) and too high (green dashed) capacity models (B).

Under-fitting and overfitting are connected to another important concept that describes ML algorithms, which is called the model capacity. A model's capacity can be thought of as its ability to fit different classes of functions. If a model capacity is too low, the algorithm will not be able to model the training data, so it will remain in the underfit regime. On the other hand, a high capacity model can easily overfit by "learn by heart" properties of the training

set that are not related to the test set data. This can be easily visualized when fitting data points with a polynomial (figure 4.1 (B)): the model with optimal capacity is able to capture the features of the data and generalize to new data, while polynomials with too high or too low degree either overfit or underfit the data.

The capacity of ML algorithms depends on a number of "hyper-parameters", e.g the degree for a polynomial, which are chosen before training and define the class of functions that can be learned. In order to choose the best performing hyper-parameters and capacity for a given task, an additional dataset, the so-called "validation" dataset, is employed. The validation dataset is equivalent to the test set, but it is used to choose the best model, in some cases also at training time.

Consequently, three independent training sets are usually necessary to optimize a ML learning algorithm and evaluate the final performance: the training set is used to minimize the loss function, the validation set serves as a guide during and after the training to choose the best model hyper-parameters, while the final performances are assessed on the test set.

Supervised and unsupervised learning

Machine Learning algorithms are usually classified as supervised or unsupervised learning algorithms. This distinction has practical consequences at training time, however, both kinds of algorithms are conceptually similar.

Supervised learning algorithms learn to produce an output that is known a priori, the target or label, based on an input distribution. Supervised learning algorithms are therefore built to estimate a probability $p(\mathrm{y} \mid \mathrm{x})$. The algorithm can often be seen as a function mapping the inputs into the outputs. The algorithm consists in optimizing the function based on a criterion and find the best parameters.
We can do this simply by using maximum likelihood estimation to find the best parameter vector $\vec{\theta}$ for a parametric family of distributions $p(\mathrm{y} \mid \mathrm{x}; \vec{\theta})$.

Unsupervised algorithms are those that experience only "features" but not a supervision signal, such as the target or label. The most common unsupervised algorithm include density estimation, i.e. learning $p(\mathrm{x})$ instead of $p(\mathrm{y} \mid \mathrm{x})$, learning to generate samples with a given distribution, data de-noising, clustering the data. Another important unsupervised learning task is to find the "best" representation of the data, which preserves all the information but makes the data simpler to use for another task.

ML in HEP at colliders

ML algorithms were used at colliders since the 1990s, at the LEP and the Tevatron experiments. The most common tasks are again classification and regression. Examples of classification were presented early on: b-tagging, pileup jet identification and removal, jet identification. A very important example is the calibration of the energy of b jets, which is described in the next chapter.

The input is usually a set of observables that describe (part of) a collision event. Historically a handful of "high level" variables, describing the collision event, were used. Different kinds of algorithms were adopted: initially Neural Networks (NNs) became popular. Subsequently, they were mostly replaced by Boosted Decision Trees (BDTs) for the LHC Run 1 analyses.

Neural Networks came back thanks to the success of Deep Learning (DL) in other fields and in several domains such as image classification or natural language processing. Deep Neural Networks (DNNs) can successfully replace BDTs in classification and regression tasks, but they also allow modeling a larger number of features, even raw features, and to train on a very large input dataset, if available.

With Deep Learning, a new trend of using "lower level" variables, e.g. single particle properties or even the detector signals, has started in HEP. New developments in this direction are the focus of chapter 5.

Other possible tasks that can be useful in HEP, which are nowadays more and more viable thanks to DL, tough not yet as common, are:

- Anomaly detection, which can be thought of as a way to look for unknown patterns and physics phenomena in the data, without a model assumption. Such algorithms would ideally allow searching for New Physics in a totally unbiased manner. Another possibility is to use such algorithms for data quality monitoring without human supervision.
- Denoising: In this type of task Machine Learning can be used to clean the data from noise information by inferring the distribution of both the interesting and the noisy component of the data.
- Data synthesis and sampling: this kind of task uses Machine Learning to generate new data samples. It could be useful in particular for a less computationally intense generation of Monte Carlo simulations.

Deep Learning offers therefore great tools, which the HEP community is starting to integrate among its practices. New developments are expected in the next few years.

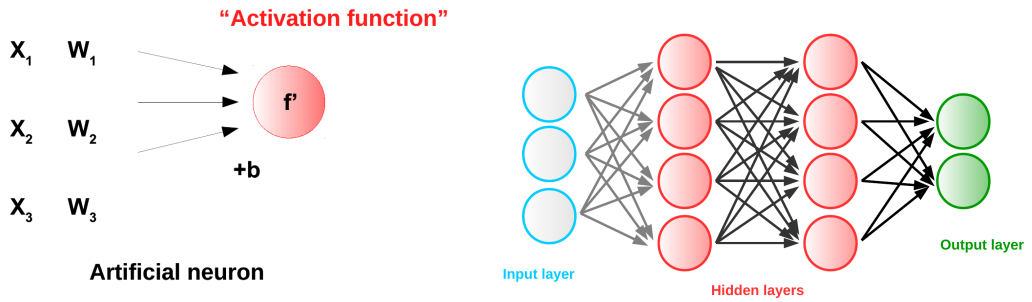
In the next sections the basic features of Neural Networks and Deep networks are described.

4.2 The feed-forward Neural Network

The feed-forward Neural Network, also called multilayer perceptron (MLP), is the basis of Deep Learning models. A feed-forward neural networks is a function $f : \mathbb{R}^n \to \mathbb{R}^m$, which maps the an input vector, typically with large dimensionality, into a simple output. E.g., in case of a classifier a large input vector is mapped into a category, yielding a real value between zero and one.

The feed-forward network function depends typically on a number of parameters $\vec{\theta}$ and the goal of training a neural network is to approximate as best as possible a desired function by tuning the parameters $\vec{\theta}$. The networks are built by applying a sequence of linear transformations and non-linear activation functions, hence they are indeed a very peculiar class of functions, however it can be demonstrated that given enough parameters they can approximate any given function [87].

The basic unit of feed-forward neural networks, which was initially inspired by biology, is the artificial neuron. As shown in figure 4.2 (A) an artificial neuron takes an input vector $\vec{x}$ and performs an affine transformation. An non-linear activation function is then applied. The mapping performed by the neuron can be written as:



(A) (B)

FIGURE 4.2: Artificial neuron schematic representation (A). Schematic representation of an artificial neural network with 2 hidden layers (B).

$$\vec{x} \to f'(\vec{w} \cdot \vec{x} + b)$$

The weight vector $\vec{w}$, and the bias b are the tunable parameters of the neuron. The activation f' is typically chosen in such a way that it has a threshold effect, mimicking the biological neuron (tanh was one of the popular activations early on in neural networks), but it is only required to be non-linear.

The feed-forward Neural Network is built by gathering multiple neurons in the same layer and stacking a number of layers, as shown in figure 4.2 (B). As shown by the colors in figure 4.2, the layers are often categorized as input, hidden and output layers. The input layers are the vector of features fed to the NN, the hidden layers are those whose ouput is fed to another layer, and the output layers are optimized to match a target value. One hidden layer is necessary to ensure the flexibility of the function, but typically deep neural networks have two or more hidden layers.

The parameters of the neural networks are tuned defining a cost function and minimizing the cost with respect to the parameters. The network is trained to match a desired output and the cost is defined as a distance of the estimated value from the target value.

An important aspect of the design of a DNN is the choice of the cost or loss function. As in other ML algorithms, the choice of the loss function is based on the principle of maximum likelihood. Minimizing the loss, which is the negative log-likelihood, is equivalent to maximizing the likelihood of the output distribution given the dataset. If based on the principle of maximum likelihood the choice of the loss function implies an assumption on the expected target data distribution. Under this hypothesis, the maximum likelihood criterion ensures the consistency of the result.

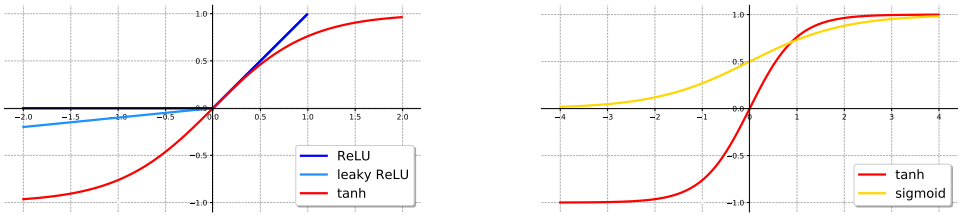
Different loss functions are employed depending on the problem of interest and correspondingly appropriate activations are used in the output layer.
In case e.g. of binary classification the binary cross-entropy, which corresponds to the maximum likelihood estimator of a Bernoulli distribution, is used together with the sigmoid activations. The sigmoid, defined as, $\frac{1}{1+e^{-x}}$, maps $\mathbb{R}$ into the interval $[0, 1]$.

Some common examples of output activation are also:

- The linear activation, for regression to a real valued number, using as loss function the mean squared error (MSE), i.e. the square of the mean difference from the target,

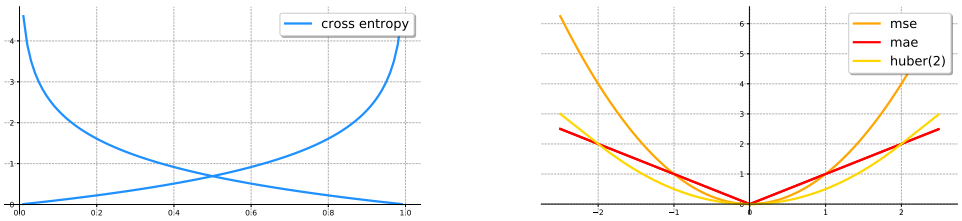
or the mean absolute error (MAE), i.e. the mean absolute value difference from the target. Both the loss functions are derived based on the maximum likelihood principle: the MSE loss provides an estimator of the mean under the gaussian hypothesis, while the MAE results in an estimator of the median.

- The softmax activation, defined for the i-th node as $\frac{e^{x_i}}{\sum_{\text{categories}} e^{-x_c}}$, together with the categorical cross-entropy. In this case the ML criterion is used under the assumption of a Multinoulli distribution.



(A) (B)

FIGURE 4.3: Examples of activation functions. The ReLU are LeakyReLU, typically used for hidden layers, are shown in (A), together with the tanh activation, which was formerly popular as activation for shallow networks. The sigmoid function, which is used as output layer activation for binary classifiers is shown in (B). The signoid is bound between 0 and 1. The sigmoid, compared to the tanh, which is bound between -1 and 1, is itself a rescaled version of the tanh, as $\sigma(x) = 1/2 + 1/2 \tanh(x/2)$..



(A) (B)

FIGURE 4.4: Some of the most common loss functions are shown. In (A) the binary cross-entropy, also called log-loss, is shown for both examples with label 0 and 1 as a function of the discriminator output, which is bound by the sigmoid activation between 0 and 1. In (B) the mean squared error ("mse"), which results in maximum likelihood estimator for the mean, the mean absolute error, which gives the maximum likelihood estimator for the median, and the "huber" loss, which is behaves like the mse in the central range and like the mae in the tails, thus becoming less sensitive to outliers..

The hidden layer activation function are by choice non-linear. Most neural networks nowadays use the Rectified linear unit activation (ReLU), which sets to zero the negative values and leaves unchanged the positive ones. Several functions similar to the ReLU, but avoid zero values, are also commonly used.

The neural networks are trained using gradient descent. The non linearity of the network causes the loss function to be non-convex. The training consist in moving to a very low value of the loss but there is no guarantee to reach the absolute minimum. Using negative log-likelihood is also helpful in the gradient descent, as the loss function doesn't usually have flat regions.

Universal approximation properties

One of the most important properties of feed-forward neural networks is their universal approximation capability. This properties holds for networks with just one or more hidden layers. Specifically, the universal approximation theorem [87, 88] states that a feed-forward network with a linear output layer and at least one hidden layer with any "squashing" activation function (such as the logistic sigmoid activation function) can approximate any Borel measurable function [1] from one finite-dimensional space to another with any desired non zero amount of error, provided that the network is given enough hidden units.

The original theorems were first stated in terms of units with activation functions that saturate, like the sigmoid or the tanh. Subsequently, universal approximation theorems have also been proved for a wider class of activation functions, which includes the ReLU units [90].

The obvious consequence of the universal approximation theorem is that a large enough, or capable enough, Neural Network will be able to learn any given function in realistic cases. However the training algorithm may not be able to learn due to underfitting or overfitting.

Model depth

The universal approximation theorems were demonstrated just for NNs with one hidden layer. However, these results don't consider the depth of the model or the efficiency in the optimization.

An important result regarding deep networks is presented in [91]: the authors show that functions representable with a deep neural networks (DNNs), with ReLU or similar activations, can separate a number of regions that is exponential in the depth of the network. In case of one hidden layer, an exponential number of hidden units is required. Choosing a model with more than one hidden layer can therefore be beneficial: with a deep model the number of nodes necessary to ensure optimal capacity is greatly reduced.

Another important feature of Deep Learning compared to simpler ML models is the empirically observed better generalization for a wide variety of tasks. Choosing a deep model implies that the function we want to learn is composition of several simpler functions. This can be interpreted as saying that we believe the learning problem consists of discovering a set of underlying factors of variation that can in turn be described in terms of other, simpler underlying factors of variation. Alternately, we can interpret the use of a deep architecture as expressing a belief that the function we want to learn is a computer program consisting of multiple steps, where each step makes use of the previous step output. The intermediate outputs are not always factors of variation but can be analogous to counters or pointers that the network uses to organize its internal processing.

Thanks to Deep Learning we can therefore talk about beating the "curse of dimensionality". Deep learning models, unlike linear models or boosted decision trees, can generalize to regions where there are few or no examples, provided that the function is a indeed a composition of simpler ones.

[1] The theorem is demonstrated for Borel measurable functions, defined e.g. in [89]. This set of function is very inclusive and covers all the practical applications.

Model choice

Another important result, which is valid for ML in general, and is useful to know when considering the NN architecture choice, is the so called "no free lunch" theorem [92]. The no free lunch theorem for Machine Learning states that, averaged overall possible data-generating distributions, every classification algorithm has the same error probability. This means that there is no ML algorithm that is better than another for all the tasks. However, in practice the algorithms that are known to perform well for a specific task can work similarly in tasks that are similar. The goal is therefore to find models and architectures that are suitable for the problems of interest. When we want to optimize a NN, one of the problems is finding the best architecture for the problem we want to solve, both from a capacity and efficiency point of view. A big part of Deep Learning research nowadays consists in experimenting with different kind of networks in order to find suitable architectures for given data formats and practical applications.

4.3 Training a Neural Network

The training procedure is based on the descent of the gradient of the loss function in the parameter ($\vec{\theta}$) space. The gradient is computed efficiently using the back-propagation algorithm described below, or its generalizations. Due to the large training sets necessary for good generalization, the gradient descent is performed using small chunks of data, called minibatches (or often just batches). This technique is called stochastic gradient descent (SGD).
Loss functions are usually non-convex for most neural networks. The training consists therefore in moving to lower values of the loss function in steps: iterative, gradient-based optimizers that drive the cost function to a very low value are used.
SGD applied to non-convex loss functions is not guaranteed to converge. The initialization of the parameters $\vec{\theta}$ is therefore important for this kind of algorithms: all weights should be initialized to small random values, several algorithms for the choice of the initialization were studied. The biases can be set to zero or to small positive values.

Back-propagation and stochastic gradient descent

The back-propagation algorithm [93], allows the information from the cost to then flow backward through the network in order to compute the gradient. Computing an analytical expression for the gradient is straightforward, but numerically evaluating such an expression can be computationally expensive. The back-propagation algorithm helps reduce the computation cost for computing the gradients.

The back-propagation is based on the chain rule of calculus: it computes the derivatives using the chain rule, with a specific order of operations that is highly efficient. The chain rule states that e.g.

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}.$$

This relationship holds also for function with vector inputs and we can transform a gradient via the Jacobian matrix.

Using the chain rule, one can write an analytical expression for the gradient of the loss with respect to any node in the networks. Many sub-expressions are repeated several times within the overall expression for the gradient. Any procedure that computes the gradient will need to choose whether to store these sub-expressions or to recompute them several

times. In many cases, computing the same sub-expression twice could be optimal to reduce memory consumption. The back-propagation instead, at least in its original design [93], reduces the number of common sub-expressions without regard to memory.
Other algorithms may be able to avoid some of the sub-expressions by performing simplifications on the computational graph, or may be able to limit the memory usage by recomputing rather than storing some sub-expressions.

The back-propagation can be applied to gradient descent or usually to the SGD. With SGD the gradient descent is performed in minibatches.SGD is necessary to overcome the computationally expensive gradient computation on large datasets. It assumes that the correct gradient can be estimated consistently using small sets of examples. Updating the gradient at each minibatch can also help escape local minima in case the function is not convex.

Back-propagation and SGD constitute the basic elements for optimizing a NN. An appropriate learning rate must be used. Additionally, algorithms that improve on the pure SGD and techniques that help reducing overfitting are commonly used nowadays in DL models.

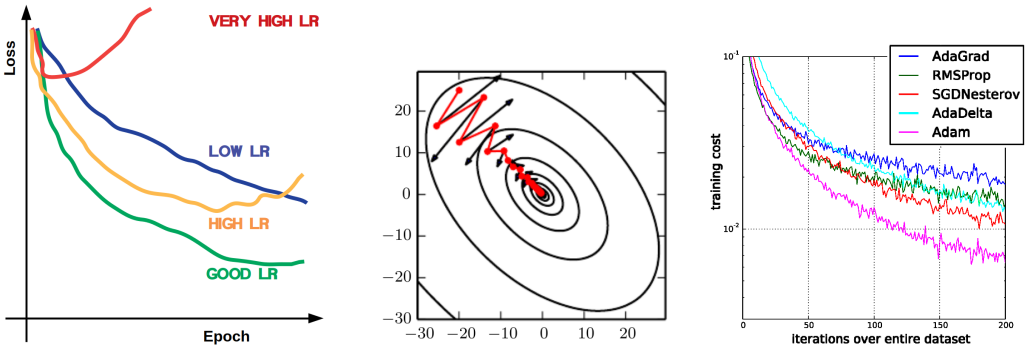
Optimization

The pure optimization of a model would be finding the global minimum of the cost function using gradient descent or other techniques. However, this is not possible for NNs and DNNs, because of the non-convexity of the loss functions. The goal of the model optimization is therefore to reach satisfactory performance according to the performance metrics chosen for the task. The loss function itself is a proxy of the performance metrics, which is minimized as the metrics improve. The training can be stopped when the performance metric stops improving.

The algorithms used to minimize the loss are based on stochastic gradient descent, where the gradient is computed via the back-propagation in minibatches and updated. With stochastic gradient descent the parameters are updated at each step by the gradient multiplied by the learning rate ϵ. The appropriate learning rate is an important parameter to ensure the convergence of the training. In case the learning rate is too large, a small enough value of the loss can be unreachable. On the other hand, a very small learning rate can slow down the convergence (see figure 4.5 (A)). In practice, the learning rate is monitored at training time and reduced based on the loss or other performance metrics to ensure better convergence.

Alternatively to the pure stochastic gradient descent, other methods like the ones employing momentum, are used to speed up and improve the convergence. If the momentum is used, the gradient is updated taking into account both the minibatch gradient with learning rate ϵ and the previous gradients. We define a velocity $\vec{v}$ that depends on the previous gradients: the velocity is updated as $\vec{v} \rightarrow \alpha\vec{v} - \epsilon\vec{\nabla}_\theta(\text{loss})$, where the parameter α can be tuned and is usually close to 0.9. Subsequently, the parameters are updated as $\vec{\theta} \rightarrow \vec{\theta} + \vec{v}$ at each step, instead of being updated based only on the last step, as $\vec{\theta} \rightarrow \vec{\theta} - \epsilon\vec{\nabla}_\theta(\text{loss})$. This behavior is exemplified in figure 4.5 (B).
This is especially useful for consistent but small gradients and noisy gradients.
Another momentum based algorithm is the one using the Nesterov momentum, where an update is performed based on the previous gradients, then the gradient is computed and the momentum is updated. Given the importance of the learning rate, several algorithms use adaptive learning rates. A different algorithm that is commonly used nowadays is Adam [94], "adaptive moments", which uses a rescaled version of momentum with additional bias

corrections.



(A) (B) (C)

FIGURE 4.5: Sketch of the typical behavior of a DNN loss function with several learning rates (A). Momentum in minimization (red) compared to the direction indicated by pure gradient descent (black) (B) [85]. Loss function behavior with several optimization algorithms (C) [94].

An important tool for the optimization is the so called batch normalization. Batch normalization [95] is one of the most recent innovations in optimizing deep neural networks. It's not an optimization algorithms that helps the descent of the gradient, but it is an adaptive re-parametrization, that was found in particular to help deep models.

Very deep models involve the composition of several functions, or layers. The gradient tells how to update each parameter, under the assumption that the other layers do not change. In practice, we update all the layers simultaneously. When we make the update, unexpected results can happen because many functions composed together are changed simultaneously, using updates that were computed under the assumption that the other functions remained constant.

Batch normalization takes a vector of activations $\vec{H}$ and normalizes each component H by updating it to (H-μ)/σ, where μ and σ are the mean and the standard deviation of the node output. The back-propagation runs trough both the computation of μ and σ and the normalization operation. Therefore, an update of the gradient cannot just move the output in one direction on average, i.e. to very large or small values, because that component would be removed by the standardization. Batch normalization regularizes the weights and also helps with the issues related to the first-order only approximation. At inference time, learned averages are used. Finally, two additional learnable parameters β and γ are used to renormalize the activations at training time.

Regularization

The optimization of a model is guided also via the so called "regularization" of the model. The regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error. Among the techniques that help regularization are weight norm penalties and constrained optimization. Weight norm penalties keep the weights small and help with gradient descent. Early stopping based on the performance on the validation dataset is also used to keep the model ability to generalize.

Other techniques include parameter sharing and ensemble methods. Parameter sharing is used under the assumption that some inputs have a symmetry property or should be treated

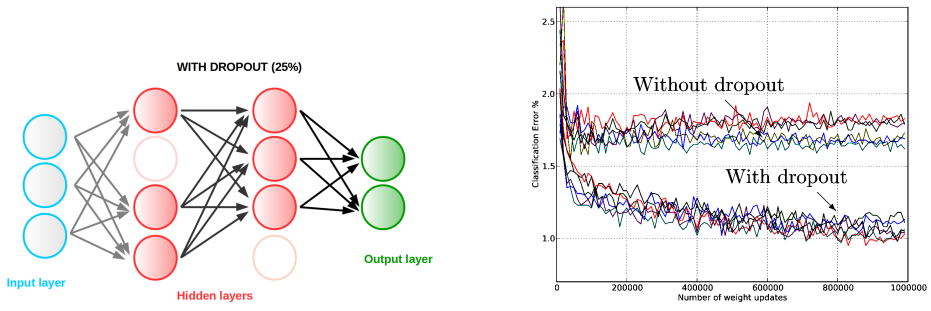
in the same way. One example is the usage of convolutional neural network or recurrent units, as outlined in section 4.4.

Ensemble methods include a typical DNN regularization method called "dropout". Dropout [96] provides a computationally inexpensive but powerful method of regularizing DNNs. In training with dropout, a fraction of the nodes of a layer are randomly zeroed at each iteration (figure 4.6). Dropout trains the ensemble consisting of all sub-networks that can be formed by removing non output units from an underlying base network.

With dropout one can train a very large ensemble of networks with little cost: each time we load a minibatch, we randomly sample a different binary mask to apply to all the input and hidden units in the network. The models share parameters, with each model inheriting a different subset of parameters from the parent neural network. The parameter sharing makes it possible to represent an exponential number of models with a low amount of memory. This would not be possible if all the models were to be initialized independently. The parameter sharing causes also most of the sub-networks to start at good settings of the parameters.

Additionally, dropout trains not just an ensemble of models, but an ensemble of models that share hidden units. This means each hidden unit must be able to perform well, regardless of which other hidden units are in the model. Hidden units must be prepared to be swapped and interchanged between models, thus providing a more solid regularization.

One big advantage of dropout is that it is very cheap computationally. The additional cost of dropout is only due to the binary masking, which is stored in memory and used for the gradient descent. Running inference in the trained model has the same cost per example as if dropout were not used, as it was demonstrated that one can run the inference on the final model with all the weights, but they must be multiplied by the dropout probability in the layer. Another advantage of dropout is that it works well with most DNNs trained with stochastic gradient descent.



(A) (B)

FIGURE 4.6: Example of a feed-forward DNN with dropout in the hidden layers (A). Example of test loss evolution with and without dropout, showing the benefits of the regularization (B) [96].

4.4 Deep Neural Network architectures

The recent success of Deep Learning in several applications came through the usage of suitable architectures, which were able to model different kinds of data. Using architectures suitable for a task allows the reduction of the number of parameters, thus improving the convergence of a network. This can be seen also as a form of model regularization.

An important factor in the choice of the architecture is the design of the connections between layers. In the feed-forward network layers described above, every input unit is connected to every output unit. Several deep models are able to use fewer connections and several parameters are shared between nodes, which helps also the convergence of the network. The strategy used to determine how to remove connections or share parameters depend on the specific data topology. For example, convolutional networks are used for data with a grid topology, as 2d images or fixed length time sequences. Another important example of parameter sharing are the recurrent networks, which are specialized in sequence processing, usually of variable length, with parameters sharing across the sequence elements.

4.4.1 Convolutional networks

The term convolutional network refers to feed-forward networks that use the convolution operation in their architecture. Convolutional networks [97] were historically very important in the development of Deep Learning, as thanks to convolutional networks unprecedented performances were reached in computer vision. The success of deep models in this task played a big role in the last wave of popularity of Deep Learning.

The most common type of convolution used in Deep Learning applications is the 2d discrete convolution applied to images. In the simplest case, e.g. a black and white picture, an image is a 2d matrix of pixels. Figure 4.7 (A) shows an example of the 2d convolution applied to a 2d image: a 2d filter (or kernel) moves along the x and y axes, the pixels of the image covered by the filter (orange) are multiplied each by the corresponding filter elements. The sum of the products for a given filter position becomes a pixel in the new filtered image (dark orange).

The operation can be written using 2-dimensional matrices for the image $I_{i,j}$ and the filter $F_{i,j}$. The result R of the filtering will be another 2-d matrix with elements:

$$R_{i,j} = \sum_{m,n} F_{m,n} \cdot I_{i+m,j+n}$$

This operation is actually called cross-correlation, and it is often implemented in Deep Learning libraries. The convolution operation, to be exact, is defined as

$$R_{i,j} = \sum_{m,n} F_{m,n} \cdot I_{i-m,j-n}$$

where the filters moves in the opposite direction. The filters are equivalent and flipped, but the two operations are completely equivalent from the point of view of the network optimization, as the elements of the filter matrix are parameters learned trough training.

The optimization of filters made out of a small number of common weights, instead of full layer to layer connections, allows a great simplification the optimization of networks that use large and sparse inputs such as images. Convolutional neural networks can also be

seen as feed-forward networks with parameters largely shared among different layers. The sharing is performed with translational invariance: this means that each element in a vector will be processed in the same way and in case of 2d matrices, like images made out of black and white pixels, each element will be propagated with the same weights.

Usually cascades of convolutional filters are applied in order to reduce the data dimensionality. The convolutional layers are followed by pooling layers, which reduce further the propagated information. The result of convolution and pooling optimized simultaneously is then fed to a feed-forward network, which receives a learned representation of the data optimized together with the feed-forward layers themselves.

In case of inputs of higher dimension, like Red-Blue-Green (RGB) images, processed as three layers of single color pixels, the 2d convolution is applied similarly, but the third dimension of the filter is fixed by the input size along the axis we don't convolve. In case of RGB images we have a tensor of dimension (M, N, 3), so the convolutional filter will have a dimension (m,n,3), with the dimension 3 fixed.

An easy analogy can be applied between convolutional neural networks and human vision: the cascading convolutional filters are often found to be learning several levels of detail of the image.

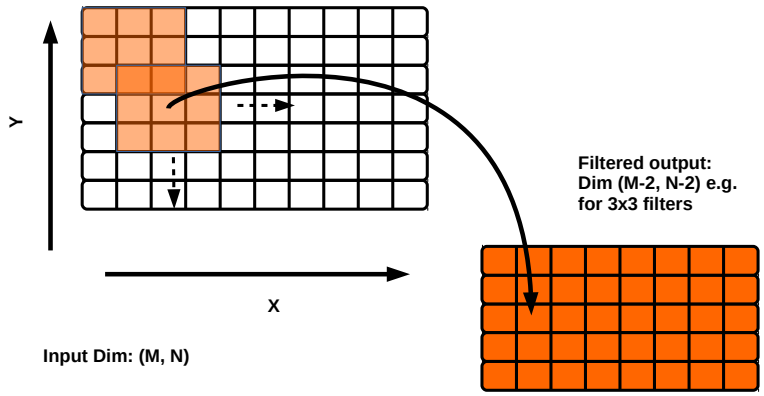
4.4.2 1x1 convolutional filters and weight sharing

Convolutional filters in one spatial dimension can also be used: in this case the filter runs over a vector of inputs and produces another vector. Convolutional filters are often used for processing natural language, where the sentences are sequences: one can think of a vector a representation of a sentence, and the filter output as another representation that can correlate close-by elements/words.

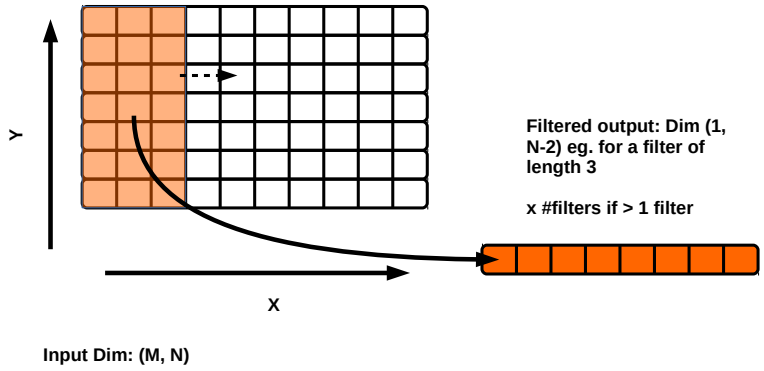
Convolution along one dimension can also be applied to 2d matrices, as in figure 4.7 (B). In this case the dimension of the filters along the convolution axis can be chosen, but the second dimension is fixed by the matrix dimension, analogously to the case of a simple vector, where the filter has dimension (n,1).
The convolution along one dimension allows the application of the same filters to a set of variables in a sequence and treat them all in the same way. The filter dimension allows the correlation of several elements that are neighbors in the sequence.

In case there is no reason to correlate single elements of the sequence, which can be the case in particle physics when e.g. using a set of particles in a jet, 1d filters with 1d convolution can be still be useful. A sketch of the result of applying such "1×1" filters is shown in figure 4.8. Each filter processes all the elements of the sequence in the same way, but keeps them independent, and each filter a new feature is built for each element of the sequence. A "1×1" filter is therefore equivalent to a dense layer shared across the sequence elements. Given a large number of filters we can change the representation by rebuilding new features - one per filter, defined in the same way for each object.

The application of such filters, "1×1", was first studied in [98] and it turns out to be a suitable tool for complex physics objects, such as the jets made out of particle flow candidates, when wants to use the PF candidates representation instead of the global one for Deep Learning applications.

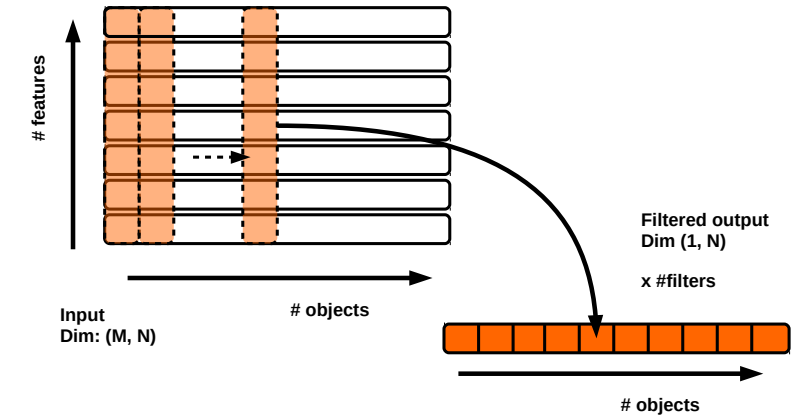


(A)

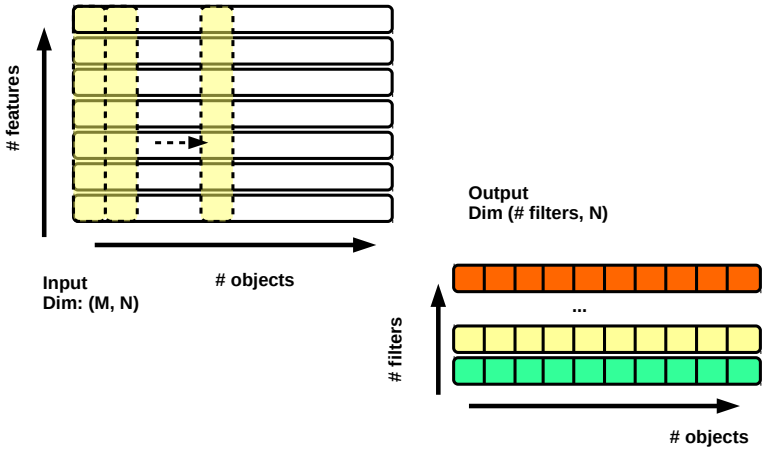


(B)

FIGURE 4.7: Representation of the application of convolutional filters on 2d-images. The 2d-convolution operation is shown in (A). The orange 3×3 filters runs on the images pixel in both directions and for each step, it produces a new pixel of the filtered image. The 1d-convolution applied to an image is shown in (B). The filter has dimension 3 for the convolution axis, while the second dimension is fixed by the image dimension. It is 1 for vectors. The result is a vector with one element for each convolution step. .



(A)



(B)

FIGURE 4.8: Representation of the application of 1x1 filters on 2d-inputs. One filter (A) produces a vector of outputs with an element for each step of the convolution. In case the convolution runs on elements of a sequence, the output has no correlation between neighbor elements but uses all of them in the same way. In case multiple filters are used each builds a new elements. As a result a new representation of the data, element by element in the sequence, is optimized by the filters..

"1 $\times$ 1" convolutional filters can also be seen as a way to optimize the representations of the features fed to the network, with each object being redefined in the same way. A downstream feed-forward network will be fed the new representation, which has the advantage of being optimized at training time, and by a ML algorithm, without human supervision.

4.4.3 Recurrent networks

Recurrent networks are a family of neural networks specialized in processing sequences. Most of them can handle sequences of variable length. The recurrent nodes are based on the principle of sharing parameters between different part of the network. Unlike convolution, where the parameters are usually shared between neighbors (e.g. a time sequence with 1d convolution), with recurrent networks parameters can be shared among members of a very long sequence. The sequence dimension is often refereed to as "time", even if the sequence is not in time, as recurrent networks are inspired mainly by temporal sequences.

Various types of recurrent networks can be used: networks with connections between hidden units, between the outputs of each hidden unit and the following one; multiple outputs can be produced, or just one for the last time step. For recurrent networks the underlying assumptions for sharing the parameters are two: all the sequence members are assumed to be treatable using the same weights and the behavior must be stationary, i.e. with no time dependence of the relationship between a member of the sequence and the previous one.

The computation of the gradient is similar to the one used for the usual gradient descent, however back-propagation has to take into account the internal loop. The gradient is computed with respect to the shared parameters as a function of all the inputs and the hidden layers outputs, depending on the network topology. This extension of the back-propagation, which makes recurrent networks more computationally intensive is often called back-propagation through time.

The LSTM Cell

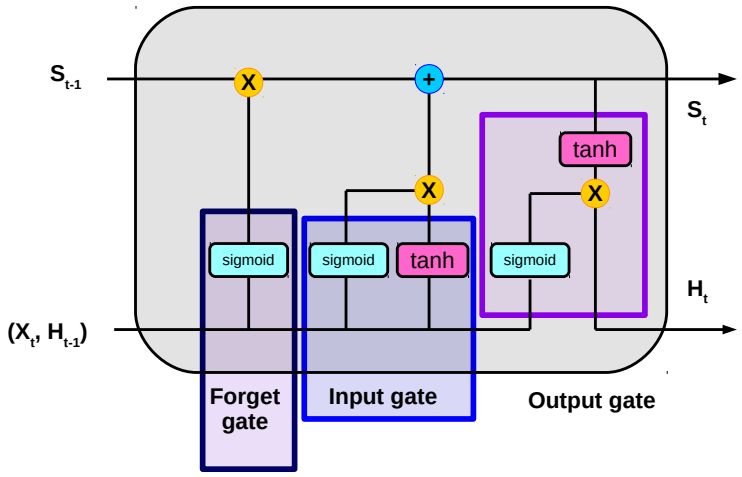
The most effective units in modeling both time sequences and other types of sequences as of today are the gated recurrent units. Among those the Long Short Term (LSTM) memory, which is used for b-tagging (chapter 5) is described here. The LSTM cell was introduced in [99].

Gated units in general introduce mechanism to accumulate or "remember" information and remove or "forget". The LSTM introduces a self-loop that avoids the problems related to the long term dependencies. The key element of the LSTM is the cell state, the horizontal line running through the top of the diagram, as shown in the sketch 4.9. It runs the entire chain with only some minor linear modifications. The cell state helps store information, as the values stored in it often just remain unchanged.

The first version included an input and output gate, while the current versions have one more gate, called "forget gate", which acts on the cell state, and modulates the update of the cell state based on the inputs.

The LSTM cell (figure 4.9), performs the following operations at each step of the sequence:

- **Forget gate:** the input and the cell outputs at the previous step are re-weighted and passed through a sigmoid.



(A)

FIGURE 4.9: Sketch of the LSTM cell. The S line represents the cell state, the input is labeled as X and the output is labeled as H. The index t denotes the step in the time sequence: the current step t, takes as input the current input vector X_t, and cell state from the previous time step S_{t-1}, and the outputs from the previous time step as a vector, H_{t-1}.

$$f_i^{(t)} = \sigma(b_i^f + \sum U_{ij}^f x_j^{(t)} + \sum W_{ij}^f h_j^{(t-1)})$$

This step of the LSTM is meant to decide what information we are going to keep or throw away from the previous step cell state. This decision is made by a sigmoid layer called the forget gate layer. It uses the vectors h_{t-1} and x_t as inputs, and outputs a number between 0 and 1 for each number in the cell state. A 1 would completely keep the previous cell state while a 0 would completely erase the previous cell state.

- **Input gate:** the gate similarly learns parameters that are used to re-weight the inputs and the output of the previous cell:

$$g_i^{(t)} = \sigma(b_i^i + \sum U_{ij}^i x_j^{(t)} + \sum W_{ij}^i h_j^{(t-1)})$$

In this step we decide what new information we are going to store in the cell state. A sigmoid layer called the input gate layer decides which values we update. Next, a tanh layer creates new candidate values, that could be added to the state. In the next step, we combine these two to create an update to the state.

- **Update of the cell state:**

the cell state is updated using the input gate results, the forget gate re-weighting and the external inputs passed to through the linear transformation and the tanh. The old cell state, $s^{(t-1)}$ is updated into the new cell state $s^{(t)}$.

The old state is multiplied by the modulator f_t, thus forgetting the things we decided to forget in the forget gate. Then we add $g_i^{(t)}$ times the tanh output. This is the new candidate value, scaled by how much we decided to update each state value.

$$s_i^{(t)} = f_i^{(t)} \cdot s_i^{(t-1)} + g_i^{(t)} \tanh(b_i^c + \sum U_{ij}^c x_j^{(t)} + \sum W_{ij}^c h_j^{(t-1)})$$

- **Output gate:** the inputs and previous outputs are passed through the output gate:

$$q_i^{(t)} = \sigma(b_i^o + \sum U_{ij}^o x_j^{(t)} + \sum W_{ij}^o h_j^{(t-1)})$$

- **Cell output computation:**

  Finally, we update the output. This output will be based on our cell state $s_i^{(t)}$, but use also the output gate $q_i^{(t)}$. The cell state is passed through a tanh (to push the values to be between -1 and 1) and multiplied by the output of the output gate. The output of the cell i is given by:

$$h_i^{(t)} = \tanh s_i^{(t)} \cdot q_i^{(t)}$$

Chapter 5

Deep Learning techniques applied to b jets

Jets originating form b quarks have peculiar characteristics that one can exploit to discriminate them from jets originating from light-flavor quarks and gluons, and to better reconstruct their momentum. Both tasks have been dealt with using ML and are now tackled with Deep Learning techniques. Two original Deep Learning applications, both involving b quark jets, are described in this chapter. The first application described is the momentum regression, the second one is a b tagging algorithm that aims at processing lower level data, and lets a DNN learn the secondary vertex information. A combination of this tagger, called "DeepVertex", with another state-of-the-art tagger, called "DeepJet", which aims at a single particle and secondary vertex description of the jet, is also presented. Both the regression and the DeepVertex tagger improve on the previously developed benchmark algorithms applied in physics analysis by the CMS collaboration. Furthermore, the combination of "DeepVertex" and "DeepJet" reaches unprecedented performance in simulation. The DNN regression was developed together with the ETH CMS group working on the search for Higgs pairs, but was deployed in data specifically for the VH($b\bar{b}$) analysis. The DeepJet algorithm was developed in parallel to DeepVertex by other groups, while DeepVertex and the combinations are presented for the first time in this thesis.

5.1 Properties and description of the b jets

The most important properties of b jets used for b tagging were highlighted already in chapter 3: b tagging makes use of the relatively long lifetime of the B hadrons in the jets, which produce significantly displaced tracks and possibly allows to reconstruct secondary vertices. Other useful properties of b jets used in b tagging, but even more useful in the momentum regression, are the harder fragmentation function compared to the other flavors, the mass of the B hadrons, larger compared to other hadrons present in the jets, and the significant rate of semileptonic decays.

Due to the harder fragmentation function, a higher fraction of the energy expected for the jet is absorbed by the B hadron, as shown in figure 5.1. As a result, a larger fraction of the jet momentum is carried by the tracks coming from the B hadron decay. For the regression the fact that some tracks are on average harder, thus better reconstructed, can help improve the momentum resolution of a b jet compared to a light-flavor jet.
The 5 GeV mass of the B hadron causes the tracks coming from the B hadron decay to have also larger transverse momenta relative to the jet axis. The high relative p_{T} information can be useful also in the regression, as in a few cases particles may leak the jet clustering cone. Finally, leptons in b jets, used also in b tagging thanks to their high relative p_{T} and displacement, are fundamental in the regression as they are accompanied by a neutrino, which is not reconstructed. The regression has the goal of recovering the neutrino, if necessary, and to improve the jet momentum resolution exploiting the other properties of the jet.

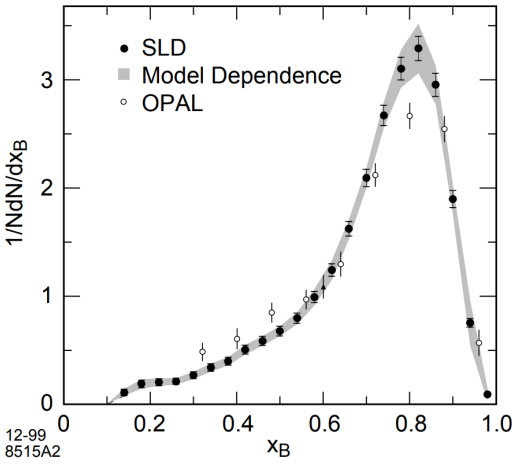


FIGURE 5.1: Fraction of B hadron energy with respect to b-quark energy [100].

All the above mentioned properties were exploited by CMS in Run 1 and at the beginning of Run 2 by building highly discriminating features, or "high level" variables, which aim to capture all the most useful properties of the jet for each task (see also section 3.3). This approach inevitably loses some of the properties that are present in a more complete and lower level jet description. Such a description can nowadays be exploited thanks to Deep Learning.

A more complete description of the jet is the one that uses single particles, taking advantage of the particle flow algorithm (see section 2.2.6). Single particles without selection or with a very loose selection can be sorted according to an importance order and used to build sequences. With this description both the b tagging properties and a global kinematic description of the jet, with its particles kinematics describing also the fragmentation function, are captured. This description is suitable both for b tagging and the b jet energy regression. All the algorithms presented in this chapter pursue a particle level description: the DNN based regression uses the PF candidates to build "particle level" jet images. On the other hand the b tagger presented, called "DeepVertex" uses tracks, i.e. charged PF candidates only, and no reconstructed secondary vertices. The algorithm is also compared to another state-of-the-art tagger, called "DeepJet", which uses single particle description of the jet, but uses also pre-reconstructed secondary vertices.
Given the performances achieved by all these algorithms, as of today, the description at the level of PF objects can be considered state-of-the-art, for b tagging and for the b jet regression. Even lower level representations of the data that retain the entire information, as "detector images", using the tracker, the calorimeter and muon detector hits, can be expected in the future.

5.2 DNN based b jet energy regression

The b jet energy regression is designed to provide the best possible estimate of the b jet momentum. It is fundamental in searches that use the invariant mass of two b jets, the most prominent one being the search for the $H \rightarrow b\bar{b}$ decay. A better resolution of the reconstructed invariant masses of the Higgs boson candidates allows the improvement of the signal-to-background discrimination and is used as input to multivariate techniques, as

shown in chapter 6.

The b jet energy regression was first used as a tool in the search for the H→ $b\bar{b}$ decay at the Tevatron [101], using neural networks with one hidden layer. BDT-based energy regressions were used prior to this result by the CMS Collaboration in searches for H→ $b\bar{b}$ in different production modes [102, 103, 35].

The regression presented here is implemented via a Deep Neural Network (DNN): the network is a feed-forward neural network with six hidden layers. The model employed has greater capacity compared to those used previously. The training uses a larger number of features: the particle flow candidates are not used directly, but jet composition and shape information are provided by building energy fractions by candidate type in rings of increasing radius around the jet axis.

A very large training dataset, made out of nearly 100 million Monte Carlo (MC) simulated jets, is employed for the training. The loss function, which combines a Huber [104] and two quantile [105] loss terms, allows the simultaneous training of point and dispersion estimators of the regression target. The method is validated on CMS data collected in 2017, and was successfully applied for the H→ $b\bar{b}$ observation.

5.2.1 Datasets

The DNN is trained on a simulated sample of $t\bar{t}$ events produced in pp collisions, generated at next-to-leading-order (NLO) accuracy in perturbative QCD with the POWHEG v2 program [106]. Simulated top quark pair production events are used as the top quark decays promptly into a b quark, which is revealed as a b jet, and a W boson. At $\sqrt{s}$ = 13 TeV, the $t\bar{t}$ production is a source of b jets that spans a large p_T spectrum and covers the full η acceptance of the tracking detector, where b jets properties can be measured.

The trained model is then tested on simulated events with b jets originating from several processes. The main test sample used was obtained from $t\bar{t}$ events not used at training time. The regression was then tested also on the Z($\ell\ell$)H($b\bar{b}$) production and the resonant HH production in the HH($b\bar{b}\gamma\gamma$) final state. Both are signal samples used in analysis with b jets in the final state. The ZH sample was generated with the MADGRAPH5_aMC@NLO generator [107] at NLO accuracy in perturbative QCD, while the di-Higgs was generated with MADGRAPH5_aMC@NLO at leading-order accuracy in perturbative QCD.

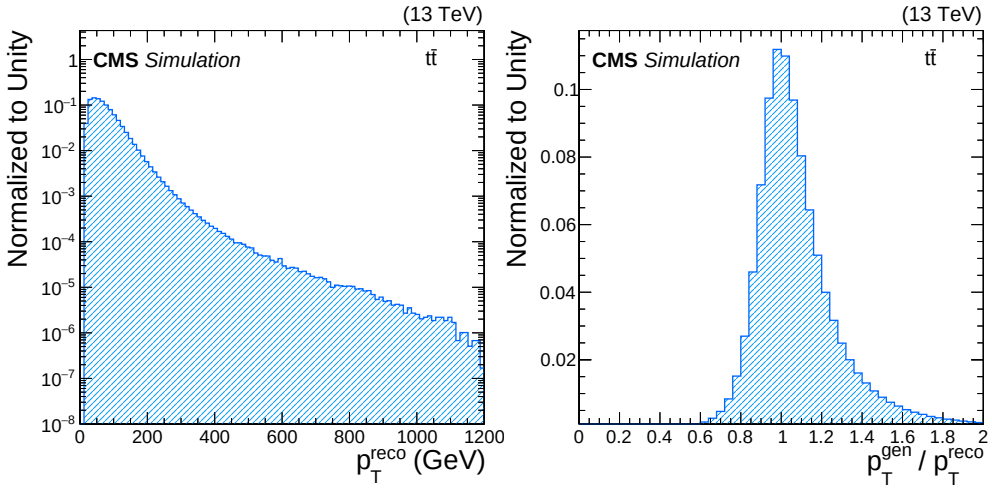
Finally, to validate the regression model on data, the DNN result for simulated jets was compared to the one obtained for jets recorded by the CMS detector with p_T balance. The events used for this validation exercise were recorded in 2017 and correspond to an integrated luminosity of 41 fb^{-1}. The simulated events come from a sample of Z bosons and up to two additional partons generated with MADGRAPH5_aMC@NLO at NLO accuracy in perturbative QCD.

For all simulated events, PYTHIA 8.2 [108] with the CP5 tune [109] is used for the parton showering and hadronization. The CMS detector response is simulated by the GEANT4 [67] package, and pileup interactions are added to the hard-scattering process according to the pileup distribution observed in data.

5.2.2 Inputs and targets

The target of the regression is the full b jet transverse momentum with neutrinos included: in order to train for that target we use "generator level jets", clustered from stable particles produced by the MC generator, which include the contribution from the neutrinos momenta. Generator level jets don't have the true b quark energy used in simulation, but they have much better resolutions compared to the "reconstructed jets", clustered from PF candidates. They are matched by cone to the reconstructed level jets.
The transverse momenta are called p_T^{gen} and p_T^{reco} in the following sections for the generator level and the reconstructed level jets, respectively.

In order to perform the regression training, the reconstructed b jets were matched to a generated b jet and were selected by applying a minimum threshold for transverse momentum ($p_T^{reco} > 15$ GeV and $p_T^{gen} > 15$ GeV) and by requiring the jet axis to be within the tracker's acceptance ($|\eta| < 2.5$). The transverse momentum p_T^{reco} is corrected with the baseline jet energy correction as described in section 3.2. Figure 5.2 (A) shows the distribution of transverse momentum, p_T^{reco}, for the selected b jets.



(A) (B)

FIGURE 5.2: (A) The p_T^{reco} distribution for reconstructed b jets in an MC $t\bar{t}$ sample. (B) Distribution of the regression target for the MC $t\bar{t}$ training sample.

The regression target used at training time, y, is defined as the ratio of the transverse momentum of the generator level jet, p_T^{gen}, to the one of the reconstructed jet, p_T^{gen}, corrected by the baseline jet energy corrections. Using this definition rather than p_T^{gen} directly has the effect of greatly reducing the variance of the target and producing a numerical value of order 1. When applied, the result becomes a correction, which is multiplied by p_T^{reco} to obtain the corrected p_T.

The distribution of the target for b jets from a MC simulated $t\bar{t}$ sample is shown in figure 5.2 (B). To improve the convergence of the training of the DNN, the target is further standardized by subtracting its median value and dividing it by its standard deviation. The reverse operation is performed when applying the training results.

The inputs are chosen with the primary goal of recovering the undetected energy fraction due to neutrinos, and allowing an improvement in the energy resolution. In order to preserve as much information as possible and perform a regression that uses all the particles in the jet, a particle level jet image was built, using energy fractions in rings of increasing radius about the jet axis. The energy fractions were added to an already established set of features, which was used already in the previous versions of the regression. The former set of inputs focused on leptons, secondary vertices and displaced tracks.

The list of inputs consists of the following features:

- jet kinematics: jet p_{T}, η, mass, and transverse mass m_{T}, defined as $m_{\mathrm{T}} = \sqrt{E^2 - p_Z^2}$;
- information about pileup interactions: the median energy density in the event, ρ, corresponding to the amount of transverse momentum per unit area that is added by overlapping collisions [76].
- information about semileptonic decays of B hadrons when an electron or muon candidate is clustered within a jet: the transverse component of lepton momentum perpendicular to the jet axis, the corresponding jet transverse momentum relative to the lepton candidate direction, the radial distance $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$ between the lepton candidate direction and the jet axis, and a categorical variable that encodes information about the lepton candidate's flavor;
- information about the secondary vertex, selected as the highest p_{T} displaced vertex linked to the jet: number of tracks associated to the vertex, transverse momentum and mass (computed assigning the pion mass to all reconstructed tracks forming the secondary vertex); the distance between the collision vertex and the secondary vertex computed in three dimensions with its associated uncertainty [49, 82];
- jet composition: largest p_{T} value of any charged hadron candidates, i.e. the leading track clustered in the jet, fractions of energy carried by jet constituents: electrons, photons, charged hadrons, neutral hadrons, and muons. These fractions are computed for the whole jet, and separately in five rings of ΔR around the jet axis (ΔR = 0 – 0.05, 0.05 – 0.1, 0.1 – 0.2, 0.2 – 0.3, 0.3 – 0.4);
- multiplicity of PF candidates clustered into the jet;
- information about jet energy sharing among the jet constituents computed as

$$\frac{\sqrt{\sum_i p_{\mathrm{T},i}^2}}{\sum_i p_{\mathrm{T},i}}$$

  where i runs over all jet constituents.

This results in a total of 43 input features. No additional preprocessing is performed, apart from the input normalization provided by batch normalization [110], which is used as the first layer of the DNN.

Unlike the $p_{\mathrm{T}}^{\mathrm{reco}}$ used in the target definition, all the input features are at the raw jet energy correction level (see section 3.2). This means that the jet total energy (raw jet energy) is the sum of all the energy fractions of the rings, when taking all the PF candidates together. The jet energy correction is only used in the definition of the target, while the jet energy resolution scale factor is not taken into account, but re-measured after comparing the results in data and simulation.

5.2.3 DNN loss function

The regression outputs are the estimated mean, and the 25 and 75% quantiles of the target distribution. The estimated mean is used as the correction to be applied to the reconstructed b jet energy, while half of the difference of the 75 and 25% quantiles is used as a per-jet estimator of the b jet energy resolution.

The Huber loss function is employed to learn the mean of the target distribution instead of the mean squared error because of its reduced sensitivity to the tails of the target distribution. It is defined as:

$$H_\delta(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| < \delta. \\ \delta|z| - \frac{1}{2}\delta^2, & \text{otherwise.} \end{cases}$$

where $z = y - \hat{y}$ represents the difference between the target and predicted values, and δ is set to 1 in our case.

To estimate the 25 and 75% quantiles of the target distribution, the quantile loss function is used:

$$\rho_\tau(z) = \begin{cases} \tau \cdot z, & \text{if } z > 0. \\ (\tau - 1) \cdot z, & \text{otherwise.} \end{cases}$$

where $\tau = 0.25$ (0.75) corresponds to the 25% (75%) quantile.

The full loss function can be therefore written as:

$$\text{loss}\,(\,\hat{y}(x),\ \hat{y}_{25\%}(x),\ \hat{y}_{75\%}(x)\,) \ = \\ = E_{(x,y)\sim p(x,y)}\,[\,H_1(y - \hat{y}(x)) \ + \ \rho_{0.25}(y - \hat{y}_{25\%}(x)) \ + \ \rho_{0.75}(y - \hat{y}_{75\%}(x))\,]$$

where $E_{(x,y)\sim p(x,y)}$ indicates the expectation value sampling (x, y) on the distribution $p(x, y)$, x indicates the set of input features, and $p(x, y)$ is the joint distribution of the input features and the target variables y in the training sample. The symbols $\hat{y}(x)$, $\hat{y}_{25\%}(x)$ and $\hat{y}_{75\%}(x)$ indicate the DNN outputs: $\hat{y}(x)$ is the estimator of the mean, $\hat{y}_{25\%}(x)$ and $\hat{y}_{75\%}(x)$ are the 25 and 75% quantile estimators, respectively.

The loss function minimized at training time is the sum of the three losses with each loss function having the same weight. The outputs are therefore correlated, and the simultaneous training for the targets, also called "multi-task" learning can help regularize the training. The results will be analyzed mainly for the estimator of the mean $\hat{y}(x)$, but it's worth keeping in mind that it is optimized together with the quantile estimators.

5.2.4 DNN architecture and hyperparameter optimization

The model used for this study is a feed-forward, fully connected DNN with 6 hidden layers, 43 input features and 3 outputs: the energy correction and the 25 and 75% quantiles. As mentioned above, a batch normalization layer is used to process the DNN input right before the first dense hidden layer.

Each hidden layer of the DNN is built from the following components:

- a dense layer, which outputs a linear combination of all outputs from the previous layer and adds a bias for each node in the layer;
- a batch normalization layer, which can rearrange the inputs to have zero-mean and unit-variance;

- a dropout unit: an operation that zeroes a fixed fraction of randomly chosen nodes, used as a regularization handle;
- an activation unit we chose the "Leaky" Rectified Linear Unit (LReLU) [111] with $\beta = 0.2$.

The slope $\beta = 0.2$ was chosen for the LReLU to allow for a nonvanishing gradient over the domain of the function [111]. The output layer has a linear activation function. It is important to note that usually the activation is part of the dense layer, but here we add it after the batch normalization layer, as suggested in the original implementation paper [110].

The DNN is implemented using the Keras package [112] with TensorFlow backend [113]. The back-propagation is uses stochastic gradient descent with the Adam optimizer [94].

The parameters and their values are: dropout rate = 0.1, learning rate = 0.001, and 6 hidden layers with [1024, 1024, 1024, 512, 256, 128] nodes. This architecture has therefore about 2.8 millions trainable parameters. The final configuration of the neural network is shown in figure 5.3. The final DNN configuration was obtained by the CMS ETH group working on the search for the SM di-Higgs production in the HH($b\bar{b}\gamma\gamma$) final state. To optimize the performance of the DNN, its hyperparameters were tuned using the cross-validation algorithm [114]. The mean validation loss was used as the figure of merit for the optimization over a five-fold splitting of the training sample. The hyper-parameters considered for the optimization are the depth of the network architecture, the dropout rate, and the gradient descent learning rate.

A check for possible bias due to the b jet p_T spectrum was also performed by the ETH group. The number of events in the jet p_T spectrum in the training sample spans six orders of magnitude, as shown in figure 5.2 (A). This means that, at training time, the DNN is fed many more jets with low values of p_T than with high values. A bias towards low-p_T jets can be expected when using such a sample. In order to check for a bias about 95% of the jets with p_T below 400 GeV were removed by extrapolating the shape of the high p_T region of the distribution down to low p_T. The DNN trained on this subsample of events showed no improvement for high p_T jets but did have up to 0.5% degradation of the inclusive relative jet energy resolution. The final training was therefore performed using the natural $t\bar{t}$ spectrum of b jets. The high statistics allow a satisfactory coverage the full phase space desired even at high p_T.

Several configurations of the DNN were tested before converging to the final architecture. In particular networks not using the energy fractions in rings around the jet axis, but processing all the particle flow candidates in the jet with LSTM nodes. The preliminary results were similar to the ones of the feed-forward network with the energy rings given as input, but the LSTM architecture was not pursued further due to the time constraints of the VH($b\bar{b}$) analysis.

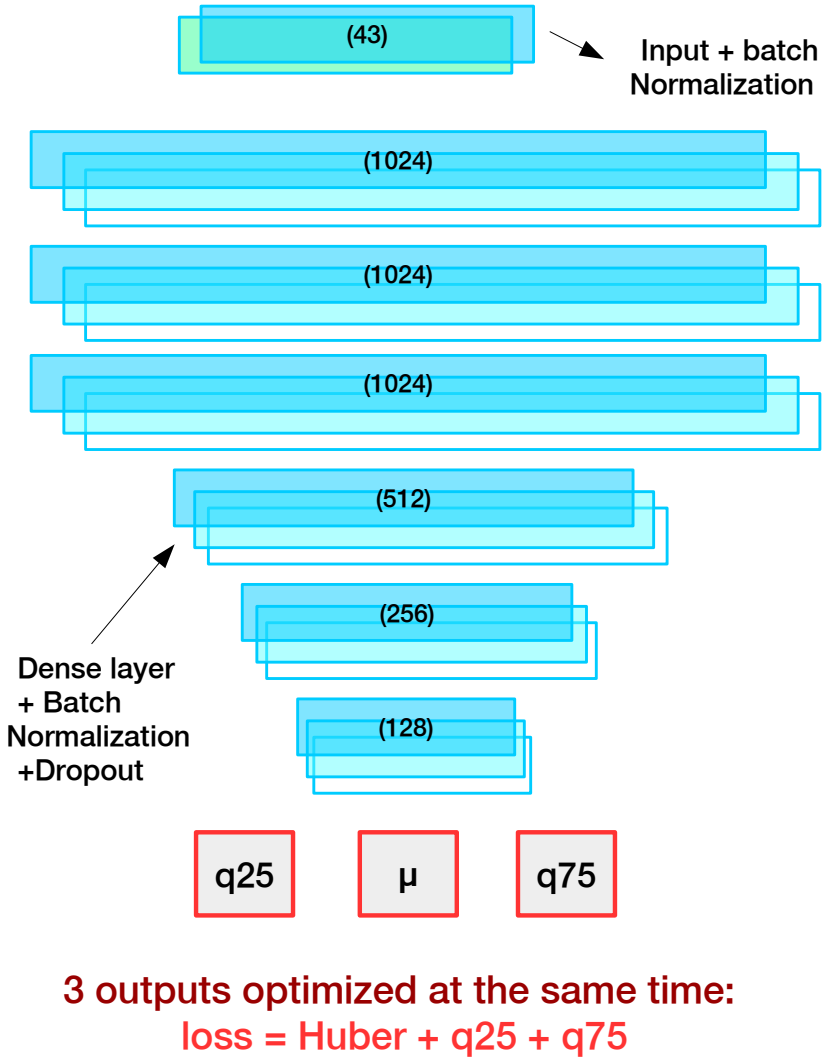


FIGURE 5.3: DNN architecture for training of the b jet energy regression.

5.2.5 Results in simulation

The performance of the b jet regression was evaluated by comparing the b jet resolution and scale (defined as the most probable values of the p_T^{gen}/p_T^{reco} distribution) before and after the energy correction on a test sample that is statistically independent from those used for training and validation. The performances were tested with $t\bar{t}$ simulated events independent of the training dataset. Compatible results were obtained with the other samples. The results are relative to the estimator of the mean $\hat{y}(x)$. Quantiles are used to quantify the performances, but they are not related to the quantile estimators obtained via the training, unless specified otherwise.

Figure 5.4 shows the 25, 40, 50, and 75% quantiles of the target distribution before and after applying the DNN b jet energy corrections, as a function of jet p_T, η, and ρ. The 40% quantile is added to the 25, 40 and 75% quantiles, as it has been found to be a good approximation of the most probable value of the target distribution. It can be seen that after DNN corrections, the distribution becomes narrower, and its median and 40% quantile exhibit a smaller dependence on jet p_T, η, and the median event energy density ρ.

The jet energy resolution, here denoted as s, is estimated as half the difference between the 75% (q_{75}) and 25% (q_{25}) quantiles of the target distribution. To quantify the resolution improvement, we compared the relative jet energy resolution, $\bar{s}$, defined as:

$$\bar{s} \equiv \frac{s}{q_{40}} = \frac{q_{75} - q_{25}}{2}\frac{1}{q_{40}}$$

where the resolution s is divided by q_{40}, the most probable value estimated as the 40% quantile of the target distribution. The relative improvement on $\bar{s}$ for b jets is on average 12%. Figure 5.5 shows the value of $\bar{s}$ obtained for b jets from the $t\bar{t}$ test sample as a function of the p_T^{gen} (A), η (B), and ρ (C). The lower panels in figure 5.5 show the relative improvements resulting from the DNN energy correction.

If we consider physics processes beside the $t\bar{t}$ production, the per-jet relative resolution improvement is consistent everywhere, with values of 12-18% for $p_T < 100$ GeV, falling to around 5-9% for $p_T > 200$ GeV.

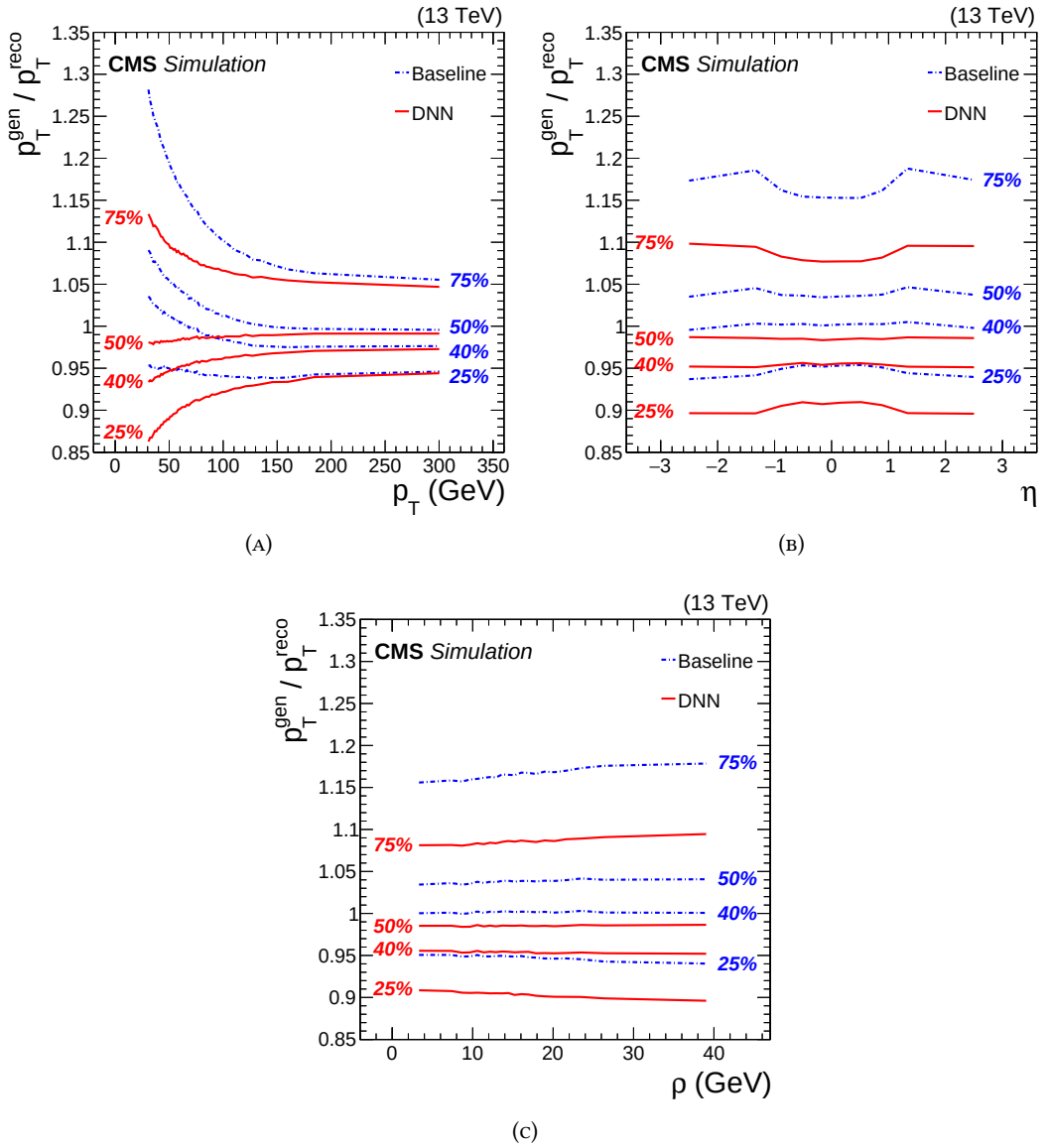
The resolution estimator was not used in the VH($b\bar{b}$) analysis presented in chapter 6, but it could be useful in future analyses. The resolution estimator is obtained using the the two quantile estimators $\hat{y}_{25\%}$ and $\hat{y}_{75\%}$ as:

$$\hat{s} \equiv \frac{1}{2}(\hat{y}_{75\%} - \hat{y}_{25\%}).$$

In order to check the consistency of the per-jet estimator, the correlation between the jet resolution s and the value of the per-jet resolution estimator, $\hat{s}$ was measured in bins of p_T. The estimator was found to be linearly correlated to the resolution of the jet, as expected. Deviations are compatible within 20% with the linear correlation. More information on the resolution estimator, which was not applied to data yet, and a summary of the regression training and results can be found in [115].

5.3 The b jet regression in data

After verifying the performance of the DNN based regression on a simulated test set, it is necessary to validate the algorithm with data. The validation of an algorithm in general



(A) (B)

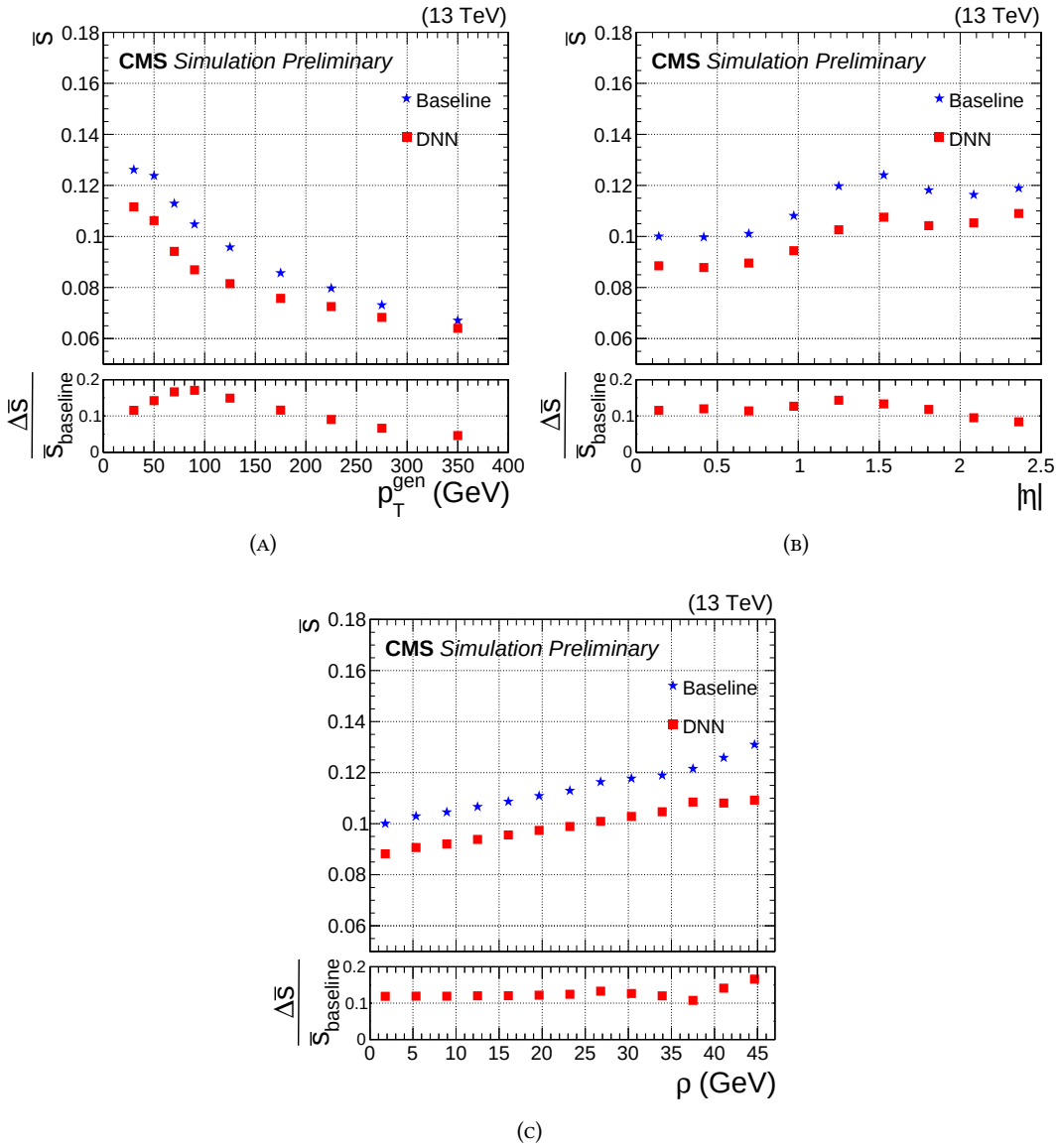
(C)

FIGURE 5.4: The 25, 40, 50, and 75% quantiles are shown for the b jet energy scale p_T^{gen} / p_T^{reco} distribution before (blue dashdot) and after (red solid) applying the regression correction as a function of jet p_T (A), η (B), and ρ (C).

consists in comparing data and simulation with the new technique applied and, if necessary, correct the simulation to better match the data. Discrepancies can arise from the mismodeling of the input features in simulation. The corrections, or scale factors, can be derived in categories or inclusively, depending on the analysis needs. In this case an inclusive scale factor is derived.

The DNN based regression can be treated as a flavor specific jet energy correction, similar to the ones described in section 3.2. In particular, the b jet energy regression covers the last step of the correction, as it is applied on top of the standard jet energy corrections.

As for the standard jet energy corrections, after applying a simulation based correction, in-situ measurements are performed to assess the need for residual corrections and a jet energy resolution scale factor. Given that the regression is a flavor specific correction, it is



(A) (B)

(C)

FIGURE 5.5: Relative jet energy resolution, $\bar{s}$, as a function of generator-level jet p_T^{gen} (left), η (center), and ρ (right) for b jets from ttbar MC events. The average pt of these b jets is 80 GeV. The blue stars and red squares represent $\bar{s}$ before and after the DNN correction, respectively. The relative difference $\Delta\bar{s}/\bar{s}_{baseline}$ between the $\bar{s}$ values before and after DNN corrections is shown in the lower panels.

necessary to validate it on a clean sample of b jets. A standard topology used for the in-situ measurement of the jet energy corrections, the "Z($\ell\ell$) + jets" final state, is used. In our case, jet b tagging is required in order to probe the jet energy regression with a relatively pure sample of b jets.
The validation procedure is described in the paragraph below.

5.3.1 Validation

The regression is validated with 2017 data selecting the Z($\ell\ell$)+b jet final state, assuming balance between the leading jet p_T, denoted as p_T^{j1}, and the Z boson p_T, denoted as $p_T^{\ell\ell}$. The Z($\ell\ell$) has much better resolution compared to the jets, therefore it can be used as a reference object to check the jet response.

The event selection follows the prescriptions for the "Z($\ell\ell$) + jets" final state used in other CMS jet energy corrections and resolution measurements [116]. A reconstructed Z candidate collinear with a jet is required. Additional hadronic activity, quantified by the ratio between the subleading jet and the Z boson $p_T^{j2}/p_T^{\ell\ell}$, here called α, is required to be suppressed. The events are enriched in b jets by requiring b tagging for the leading jet. The selections are reported in detail below.
The jet response is then evaluated as:

$$R_j = \frac{p_T^{j1}}{p_T^{\ell\ell}}.$$

In order to select "Z($\ell\ell$) + jets" events, trigger paths requiring two leptons are used. The trigger selections are the same of the 2 lepton channel of the VH($b\bar{b}$) analysis, as reported in chapter 6. The p_T thresholds for the two muons are 17 and 8 GeV respectively, and 23 and 12 GeV for the electrons. Loose isolation and identification criteria are also applied to the leptons at this level.
The offline selection of the leptons is also the same as for the 2-lepton selection of the VH($b\bar{b}$) analysis (see section 6.2.3). Both the electrons and the muons are selected using relatively loose isolation and identification criteria.

The jet selection is relatively loose: the leading must pass jet and pileup identification criteria, but no p_T threshold is required. The jet is required to be b-tagged, and a η selection is applied in order to improve the b jet purity. In summary, the event must have at least one jet with:

- loose PF Jet identification, loose pileup identification;
- Deep CSV Medium working point, corresponding to 70% b jet efficiency and 1 % light-flavor quark and gluon jet mistag;
- $|\eta| < 2.0$, chosen to have a robust b tagging.

A $\pm$ 20 GeV window about the Z mass is selected.
The leading jet and the candidate Z are required to be collinear and the hadronic activity, quantified as α to be suppressed. The selections applied are the following:

- $|\Delta\phi(\ell\ell, j1)| > 2.8$;
- $\alpha < 0.3$.

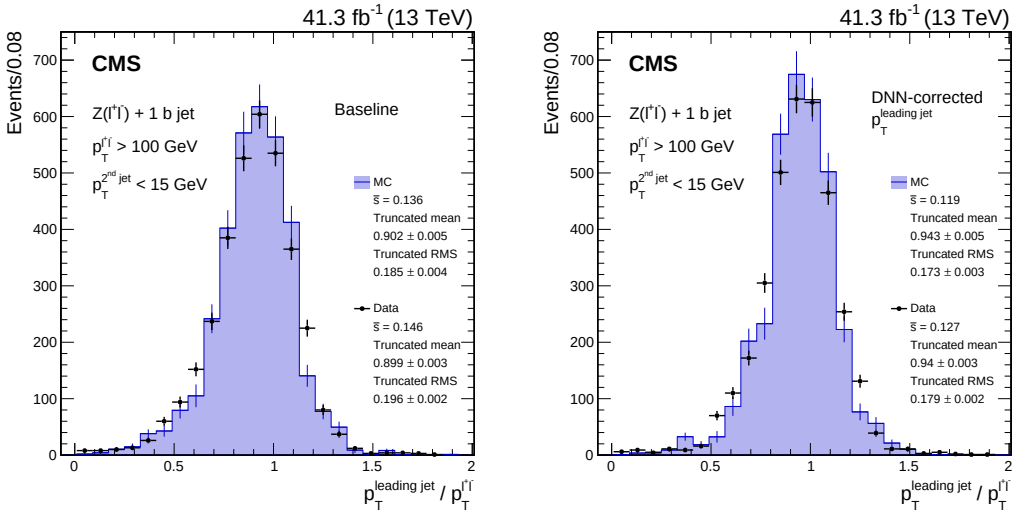
A p_T threshold is also required for the candidate Z boson. The $p_T^{\ell\ell}$ threshold usually adopted in jet energy corrections measurements is 30 GeV. Here the threshold is modified, due to the fact that jets are stored only if their p_T is above 15 GeV.
If the jet response parameters were measured as a function α with a 30 GeV threshold only, each α bin would have a different $p_T^{\ell\ell}$ spectrum. The jet response parameters would eventually be extracted as a function of $p_T^{\ell\ell}$, and consequently of the p_T of the leading jet itself. The $p_T^{\ell\ell}$ dependency is mitigated by applying a $p_T^{\ell\ell} > 100$ GeV cut, so that the $p_T^{\ell\ell}$ spectra and mean values are comparable in each α bin.

Approximate transverse momentum balance is expected between the Z boson and the leading jet after the selection. Given the imperfect balance, an extrapolation to the ideal case of 0 hadronic activity is usually performed for jet energy corrections measurements and

resolution measurements. The sample is divided in bins in α, and the R_j parameters are extracted in each bin and extrapolated to $\alpha = 0$, which corresponds to the perfect p_T balance hypothesis.

The bins in α used for the extrapolation here and in the following section are: $\alpha < 0.185$, $0.185 < \alpha < 0.245$ and $0.245 < \alpha < 0.3$. The bins include events where the p_T of the second jet, p_T^{j2}, is greater than 15 GeV. The subset of events with $p_T^{j2} < 15$ GeV is not included in the extrapolation (α is not computed and set to $\alpha = 0$), but the events have the best balance in our dataset and are used for qualitative comparisons.

We first check the response R_j in the "$\alpha = 0$" subset of events. In figure 5.6, we can observe a similar behavior in data and simulation before (A) and after (B) applying the regression. The mean of the distribution was measured to be compatible between data and MC, and it moves closer to one after the regression. The resolution, reported in the figure as $\bar{s}$ is worse in data before the regression, and improves similarly in data and simulation with regression. The fact that the resolution is worse in data both before and after the regression, points to the fact that, as for the standard jet energy correction, a resolution scale factor is necessary.



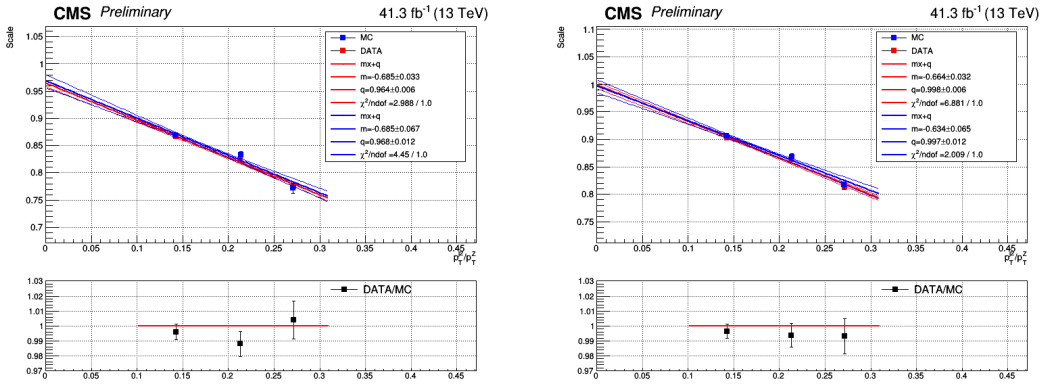
(A) (B)

FIGURE 5.6: Distribution of the ratio between the transverse momentum of the leading b tagged jet and that of the dilepton system from the decay of the Z boson. Distributions are shown before (A) and after (B) applying the b jet energy corrections. The $\bar{s}$ values of the core distributions are included in the figures. The black points and histogram show the distributions for data and simulated events, respectively.

For the extrapolation, the R_j parameter used is truncated mean of the response, computed integrating 98.5% of the distribution symmetrically, i.e. from the 0.75% quantile to the 99.25% quantile.

If we look just at the truncated mean of R_j, the effect of the DNN b-regression is found to be consistent in data and MC improving the p_T balance in all the α bins, as shown in figure 5.7), where the jet response is compared before (A) and after the DNN correction (B). The data points are in red, while the equivalent points for simulation is in blue. The lower panel shows the data/MC ratio in α bins. A simple extrapolation in α, as a function of the mean α for each α bin, can be performed to remove the extra-activity dependence, using a

linear function in this case. This extrapolation exercise shows that pre-regression the mean is extrapolated to ~ 0.95, while post-regression to ~ 1. This means that we have no bias on average and that the regression improves the jet response, as already seen in simulation. Moreover, the extrapolation of the truncated mean is consistent within the uncertainty in MC and data.



(A) (B)

FIGURE 5.7: Truncated mean in the α bins pre regression (A) and post-regression (B). The three points correspond to the three α bins: $\alpha < 0.185$, $0.185 < \alpha < 0.245$ and $0.245 < \alpha < 0.3$. The mean of the response depends linearly on the additional activity. The linear fit and the $\pm 1\sigma$ uncetainties are shown. The lower panel shows the data/MC ratio in α bins.

5.3.2 Resolution scale factor extraction

The measurement of the jet p_T resolution also involves an extrapolation to the ideal case of zero hadronic activity. To measure the jet p_T resolution, the observable $\sigma(p_T^{j1}/p_T^{\ell\ell})$ is expressed as:

$$\sigma_{\text{total}}(p_T^{j1}/p_T^{\ell\ell}) = \sigma_{\text{intrinsic}}(p_T^{j1}/p_T^{\text{gen,j1}}) \oplus \sigma_{\text{imbalance}}(p_T^{\text{gen,j1}}/p_T^{\ell\ell}).$$

The $\sigma_{\text{intrinsic}}(p_T^{j1}/p_T^{\text{gen,j1}})$ is the term we are interested in to measure the jet energy resolution. The term $\sigma_{\text{imbalance}}(p_T^{\text{gen,j1}}/p_T^{\ell\ell})$ can in turn be written as the sum of two independent contributions: $\sigma_{\text{ISR+FSR}} \oplus \sigma_{\text{PLI}}$. The first one depends on the event imbalance due to the extra jets in the event, both from ISR and FSR. The second one, usually called "particle level imbalance", is due to underlying event and to particles showering outside the jet cone.

The effect of extra jet activity is studied as a function of α. In the limit $\alpha \to 0$ the ISR and FSR components of $\sigma_{\text{imbalance}}(p_T^{\text{gen,j1}}/p_T^{\ell\ell})$ are expected to be 0, and the imbalance component is expected to converge to σ_{PLI}, which is approximately independent of α. The $\sigma_{\text{intrinsic}}(p_T^{j1}/p_T^{\text{gen,j1}})$ term is again in principle independent of α, except for spurious dependencies due to the phase space selection, which can be estimated in simulation from the $p_T^{j1}/p_T^{\text{gen,j1}}$ distributions.

Again, the resolution is evaluated as the truncated RMS of the response R_j, computed integrating 98.5% of the distribution symmetrically, i.e. from the 0.75% quantile to the 99.25% quantile.

The extrapolation is performed by fitting the resolutions as a function of the mean α for each α bin. The fit function used is:

$$f(\alpha) = c \cdot (1 + c_k \cdot \alpha) \oplus (m \cdot \alpha). \tag{5.1}$$

The first term is used for the $\sigma_{\text{intrinsic}} \oplus \sigma_{\text{PLI}}$ contribution, which is not strongly α-dependent. An α-dependent correction c_k is added to c, so that it becomes $c \cdot (1 + c_k \cdot \alpha)$. The second term $(m \cdot \alpha)$ is used for the $\sigma_{\text{ISR+FSR}}$ contribution to the resolution and depends linearly on α.

The parameters of the fitting function are not all free in the extrapolation fit: the parameter c_k is extracted from a fit to the MC intrinsic resolution performed similarly in α bins. A linear fit of the MC resolution is performed and c_k is fixed at m_0/q_0, where m_0 and q_0 are respectively the slope and the intercept of the linear model. The α dependency of $\sigma_{\text{intrinsic}} \oplus \sigma_{\text{PLI}}$, is therefore assumed to be proportional to the one of the MC intrinsic resolution. The parameters extracted from the extrapolation fit are therefore only c and m.
The jet energy resolution (JER) scale factor is finally measured by comparing the term c in MC and data. A scale factor $SF = c_{\text{data}}/c_{\text{MC}}$ should applied to b jets in simulation, enhancing the p_T difference between the jet and the generator level jet with the neutrino component included, which is the target of the regression.

The distributions of the b jets response R_j after the regression in the 3 α bins used in the extrapolation are shown in figure 5.8 for both data and MC. The MC intrinsic resolution distributions used to fix the c_k parameter of the model are in figure 5.9.

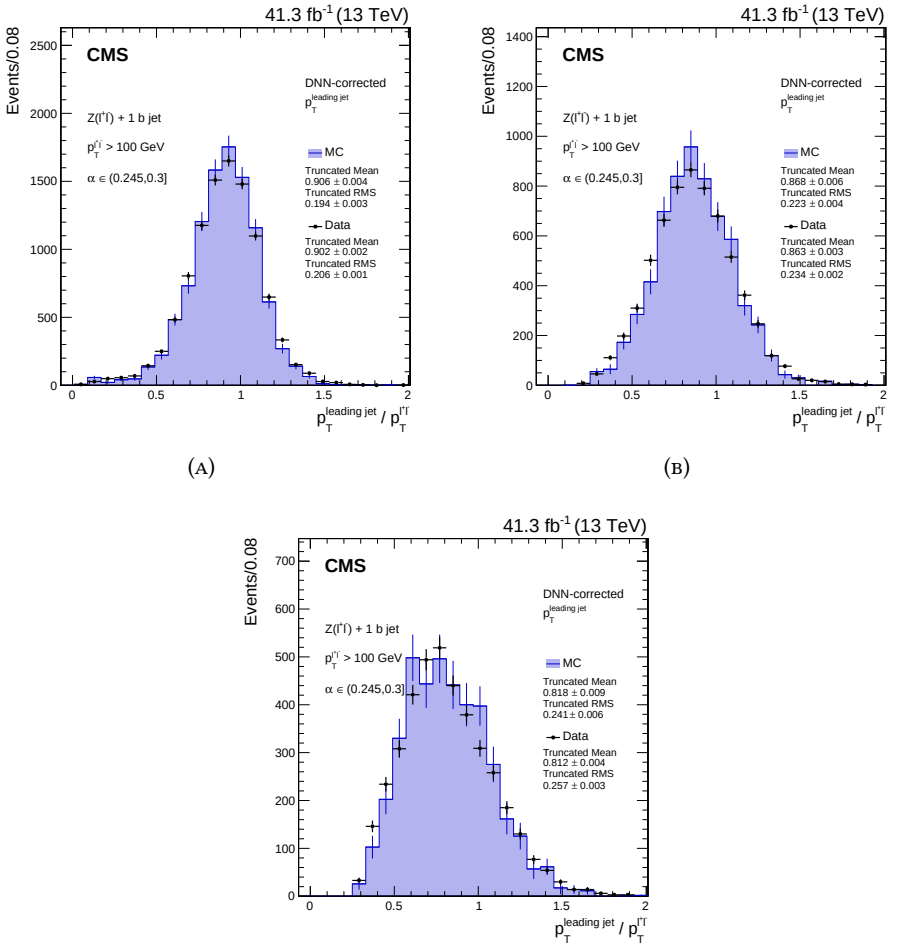


FIGURE 5.8: Distributions of the b jets response R_j after regression for data (black) and MC (blue) in the α bins $\alpha < 0.185$ (A), $0.185 < \alpha < 0.245$ (B) and $0.245 < \alpha < 0.30$ (C).

The extrapolation fit is shown in figure 5.10, where the MC intrinsic resolution points, which fix c_k are shown as green dots. Again the data points and are in red, while the equivalent points for simulation is in blue. The "$\alpha = 0$" bin points are also superimposed for comparison, and labeled as "0-bin": the value of *alpha* was roughly estimated by looking at the p_T^{j2} distribution.
The two contributions to the resolution extracted from the fit are shown as dashed green lines. The extrapolated intrinsic contribution is $c = 0.163 \pm 0.003$ in data and $c = 0.174 \pm 0.003$ in MC, pointing to a scale factor of $\simeq 1.07$. A conservative scale factor of 1.1 ± 0.1 is therefore used for the VH($b\bar{b}$) analysis. The 0.1 is meant to cover the both the statistical and the systematic uncertainty.

The measurement of the resolution scale factor is not refined further. The measurement can be considered satisfactory for the VH($b\bar{b}$) analysis, as the jets selected for this measurement are kinematically similar to the ones selected to build the Higgs boson candidate and there is not need to make more categories. Moreover, the number of events is just enough to perform the inclusive extrapolation. A conservative estimate of the uncertainty (the full scale factor) is meant to cover also the uncertainties due to the inclusive measurement.

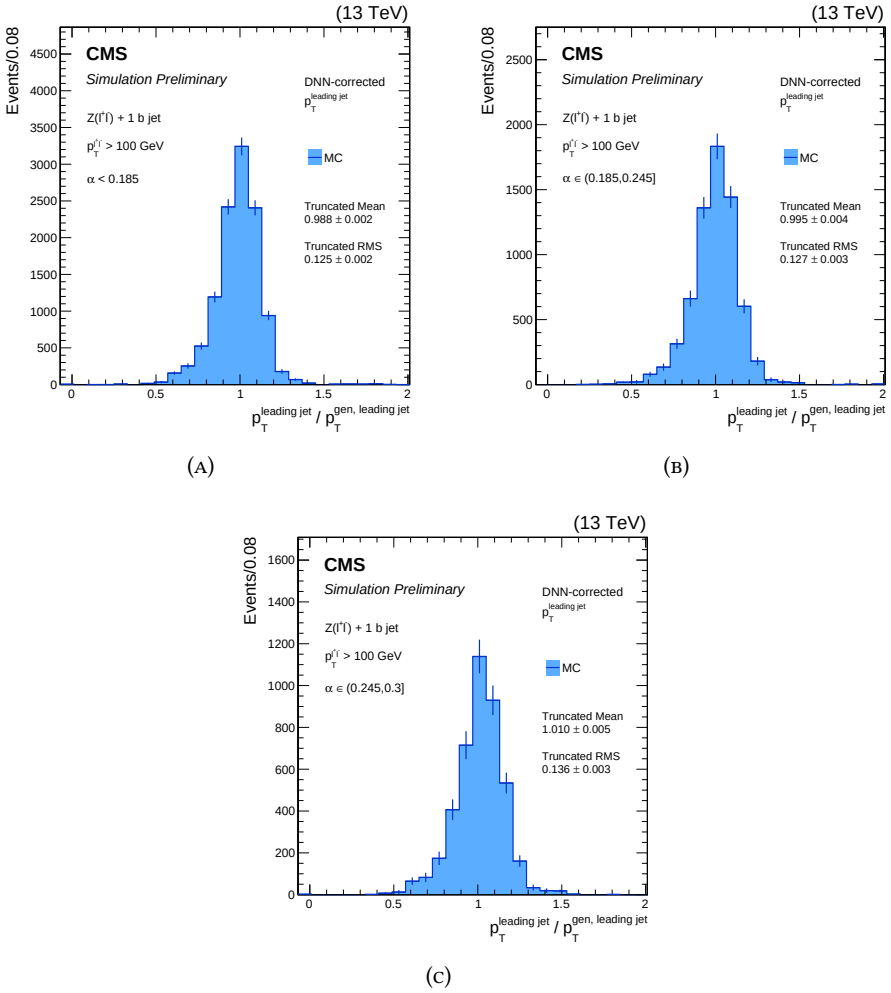


FIGURE 5.9: Distributions of the MC intrinsic resolution measured using the ratio $p_{\mathrm{T}}^{\mathrm{reco}}/p_{\mathrm{T}}^{\mathrm{gen}}$ in the α bins $\alpha < 0.185$ (A), $0.185 < \alpha < 0.245$ (B) and $0.245 < \alpha < 0.30$ (C).

A closure test is performed by applying the 1.1 scale factor only to the leading jet, and a good closure of the extrapolation within the statistical uncertainty is found (see figure 5.11). The second jet transverse momentum $p_{\mathrm{T}}^{\mathrm{j2}}$ can also be corrected with the resolution scale factor derived by the CMS collaboration for all flavor jets. This would also modify the content of the α bins. However usually the resolution scale factor is not applied to any jet for the extraction of the resolution scale factor itself, and we follow the same prescription.
An additional closure test with the standard JER resolution SF applied also to the extra jets in the event is performed. A better event closure is found in this case (see figure 5.12). For completeness, the pre-regression curve with no resolution scale factor is also reported in figure 5.13. The parameters estimated trough the extrapolations and the MC inputs m_0 and q_0 are reported for each case in table 5.1.

The distributions entering the first closure test are reported in figures 5.14 and 5.15. The distributions used to verify the event closure with smearing applied also to the extra jets are reported in figures 5.16 and 5.17.

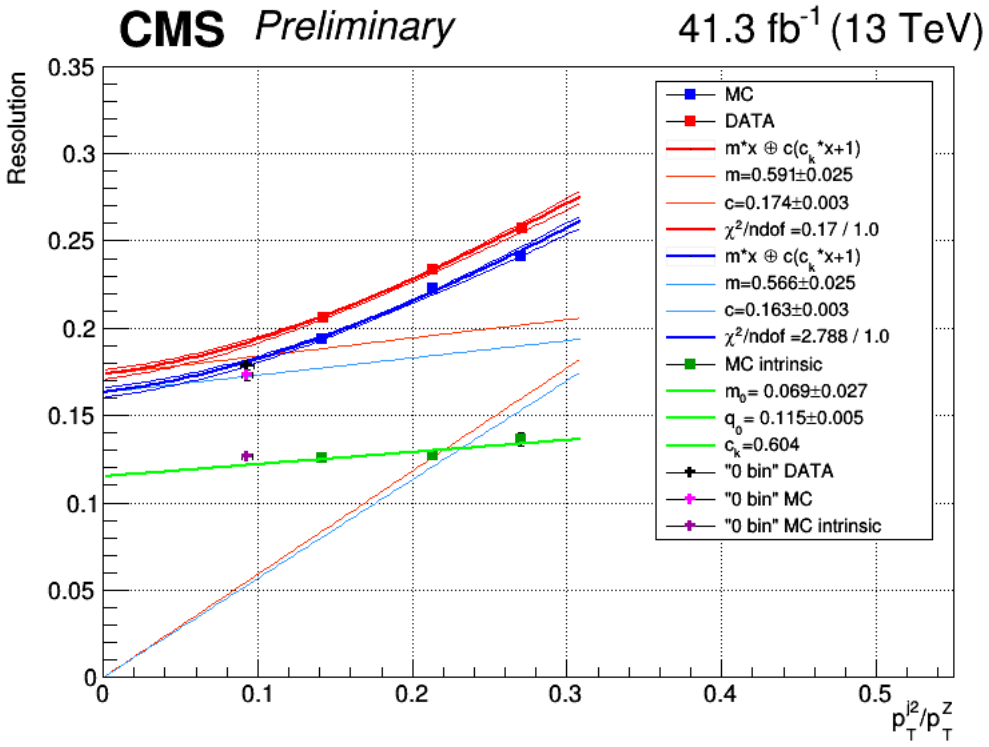


FIGURE 5.10: Extrapolation of the fitted resolutions with no resolution SF applied to the leading jet as a function of α. The data points are in red, and the MC points are in blue. The MC intrinsic resolutions which fix c_k, are shown as green dots.

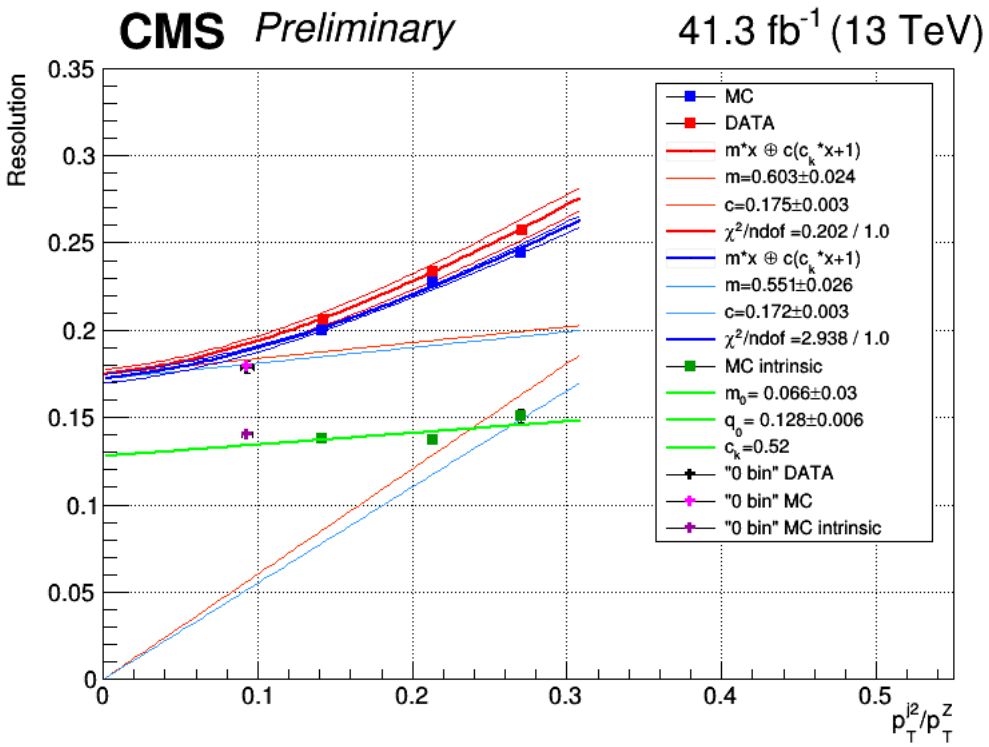


FIGURE 5.11: Extrapolation of the resolutions with 1.1 resolution SF applied to the leading jet as a function of α. A closure test with a 1.1 smearing factor applied only the leading jet is shown.

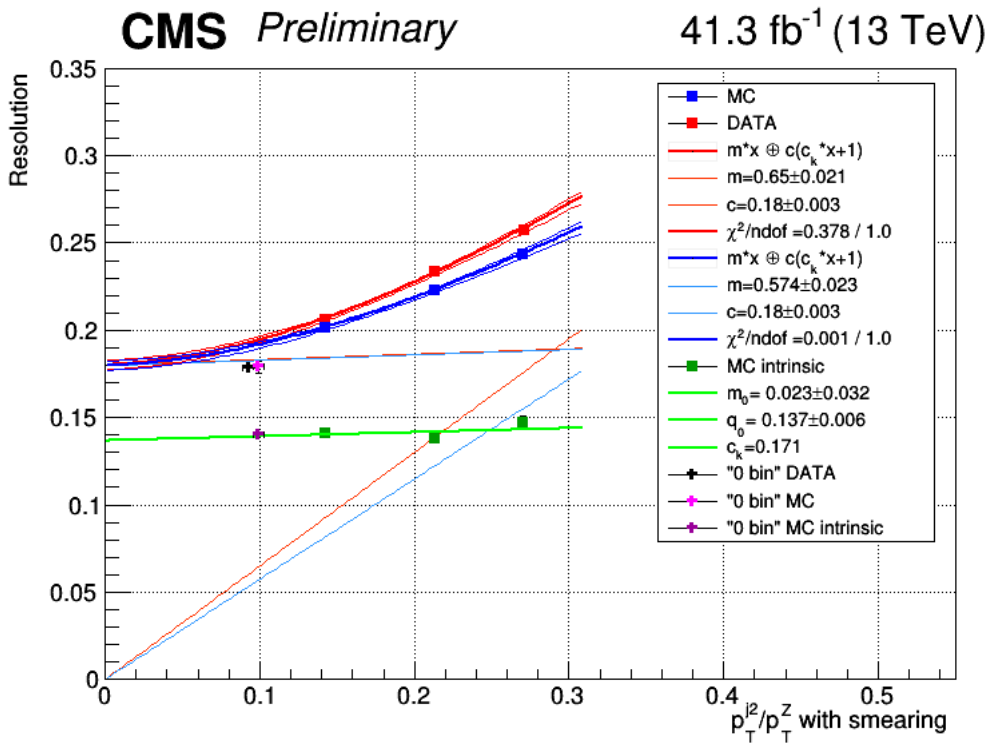


FIGURE 5.12: Extrapolation of the resolutions with 1.1 resolution SF applied to the leading jet as a function of α. A closure test with a 1.1 smearing factor applied to the leading jet and JER SF applied to the extra jets is shown.

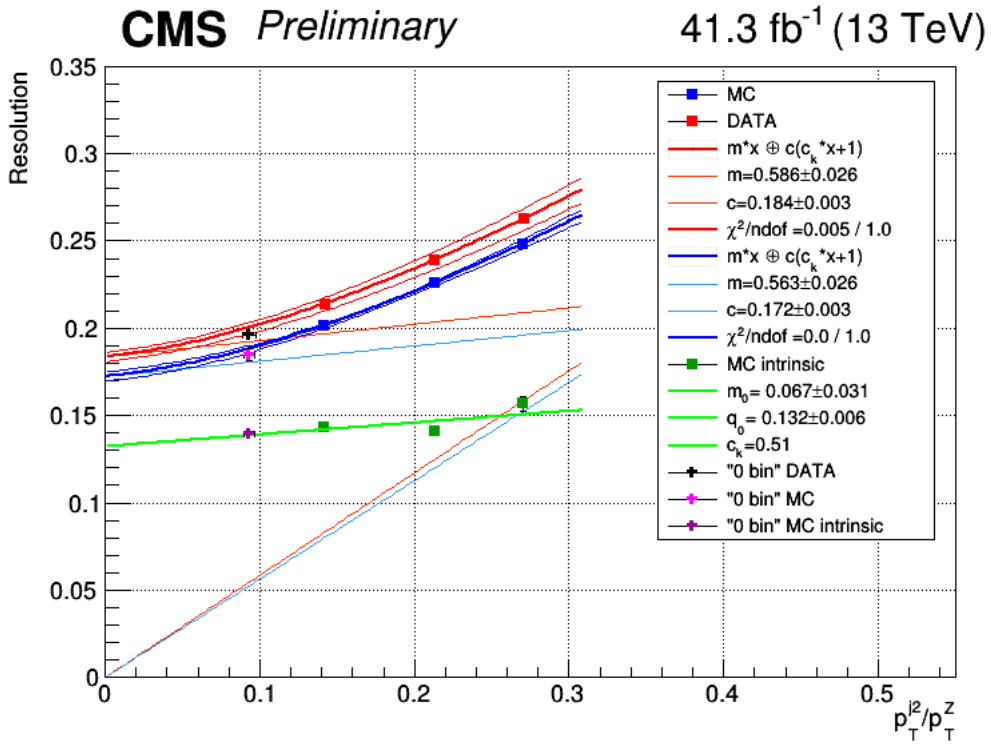


FIGURE 5.13: Extrapolation of the fitted resolutions with no resolution SF applied to the leading jet as a function of α and without applying the b jet energy regression.

TABLE 5.1: Extrapolation to 0 hadronic activity summary table. The parameters q_0, m_0 are used to determine the slope of the intrinsic resolution both in MC and data. The parameter q is fixed at 0.

	p_T^{j1} after regression, unsmeared, p_T^{j2} unsmeared	p_T^{j1} after regression, with 1.1 JER SF, p_T^{j2} unsmeared	p_T^{j1} after regression, with 1.1 JER SF, p_T^{j2} smeared	p_T^{j1} no regression, unsmeared, p_T^{j2} unsmeared
MC intrinsic				
m_0	0.069±0.027	0.066±0.030	0.030±0.032	0.067±0.031
q_0	0.115±0.005	0.128±0.006	0.137±0.006	0.132±0.006
MC reco				
m	0.566±0.025	0.551±0.026	0.574±0.023	0.563±0.026
c	0.163±0.003	0.172±0.003	0.180±0.003	0.172±0.003
Data				
m	0.591±0.025	0.603±0.024	0.650±0.021	0.586±0.026
c	0.174±0.003	0.175±0.003	0.180±0.003	0.184±0.003

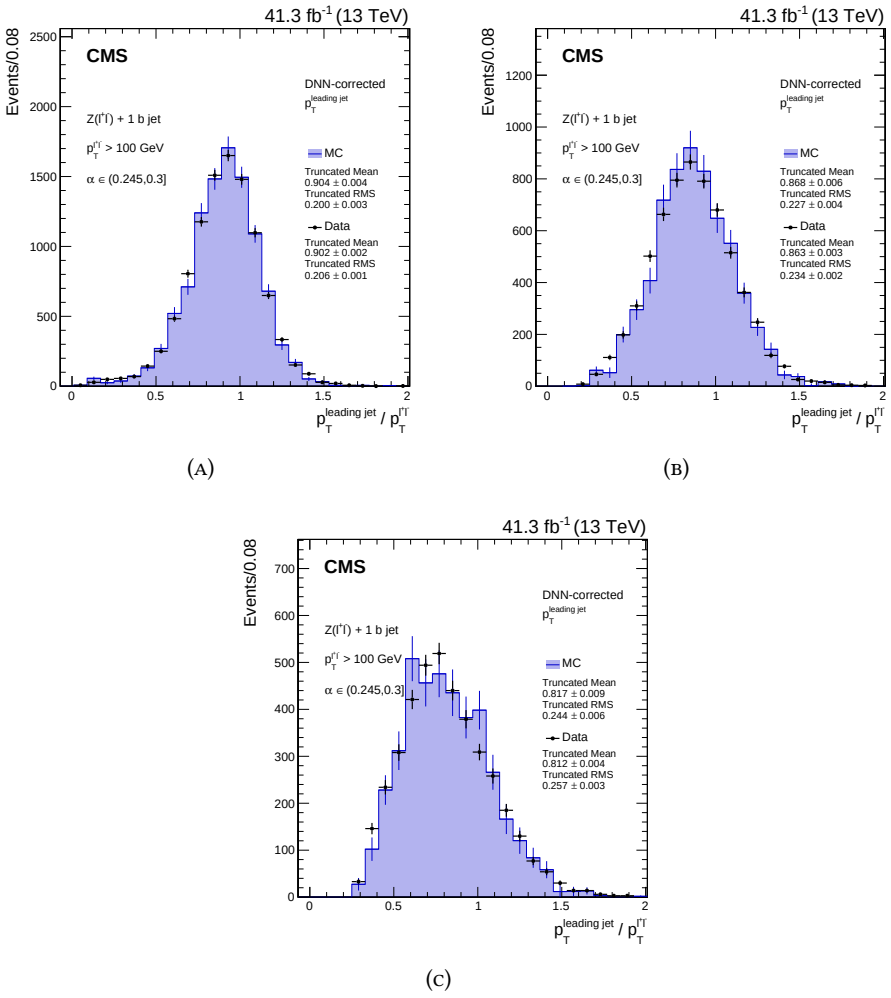


FIGURE 5.14: Distributions of the b jets response after regression with 1.1 resolution SF applied to the leading jet for data (black) and MC (blue) in the α bins $\alpha < 0.185$ (A), $0.185 < \alpha < 0.245$ (B) and $0.245 < \alpha < 0.30$ (C).

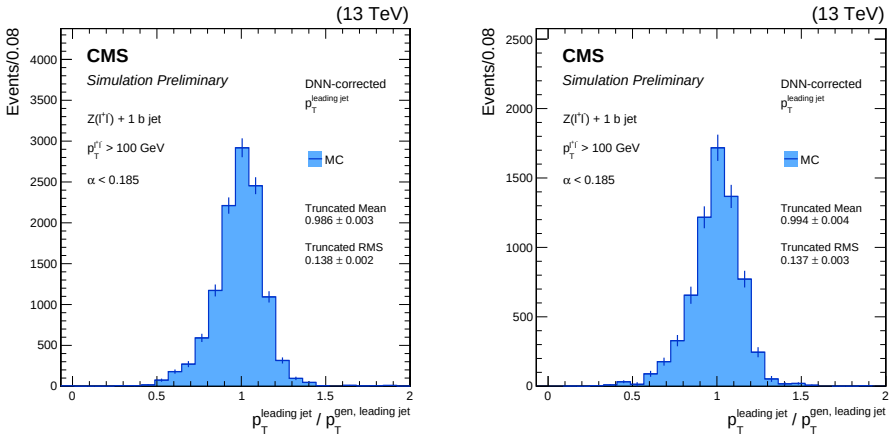


FIGURE 5.15: Distributions of the MC intrinsic resolution with 1.1 resolution SF applied to the leading jet measured using the ratio $p_{\mathrm{T}}^{\mathrm{reco}}/p_{\mathrm{T}}^{\mathrm{gen}}$ in the α bins $\alpha < 0.185$ (A), $0.185 < \alpha < 0.245$ (B) and $0.245 < \alpha < 0.30$ (C).

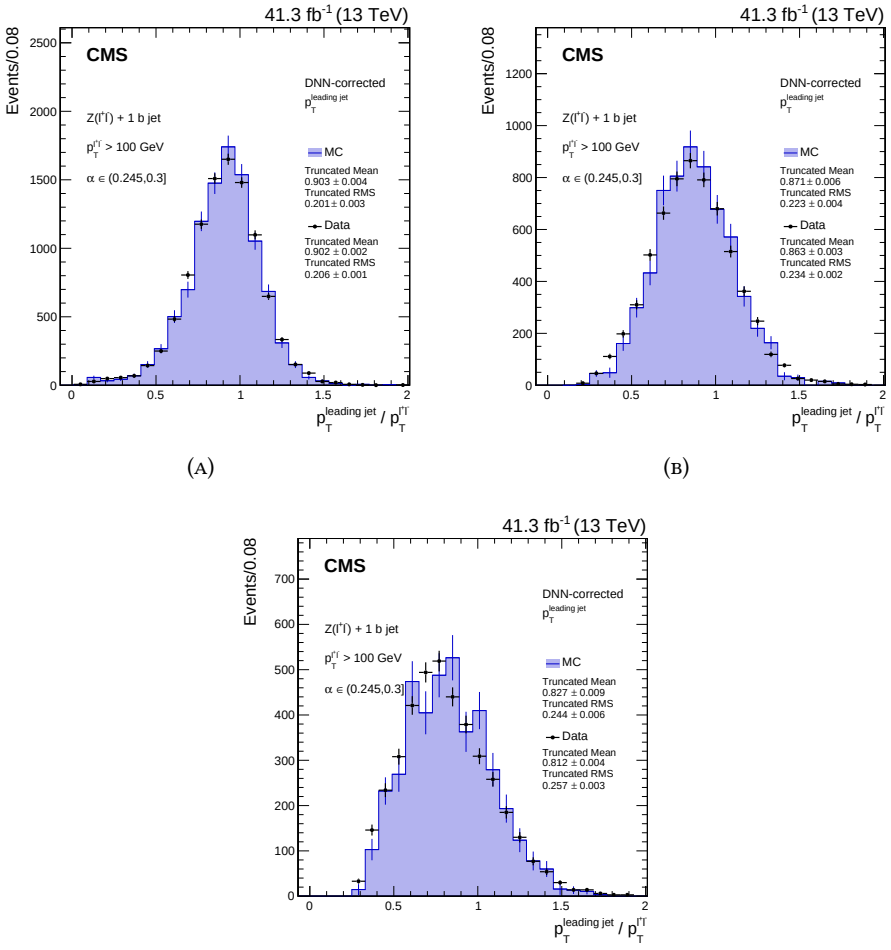


FIGURE 5.16: Distributions of the b jets response after regression with 1.1 resolution SF applied to the leading jet for data (black) and MC (blue) in the α bins $\alpha < 0.185$ (A), $0.185 < \alpha < 0.245$ (B) and $0.245 < \alpha < 0.30$ (C). The JER resolution SF is applied also to the extra jet used to define the α bins in this case.

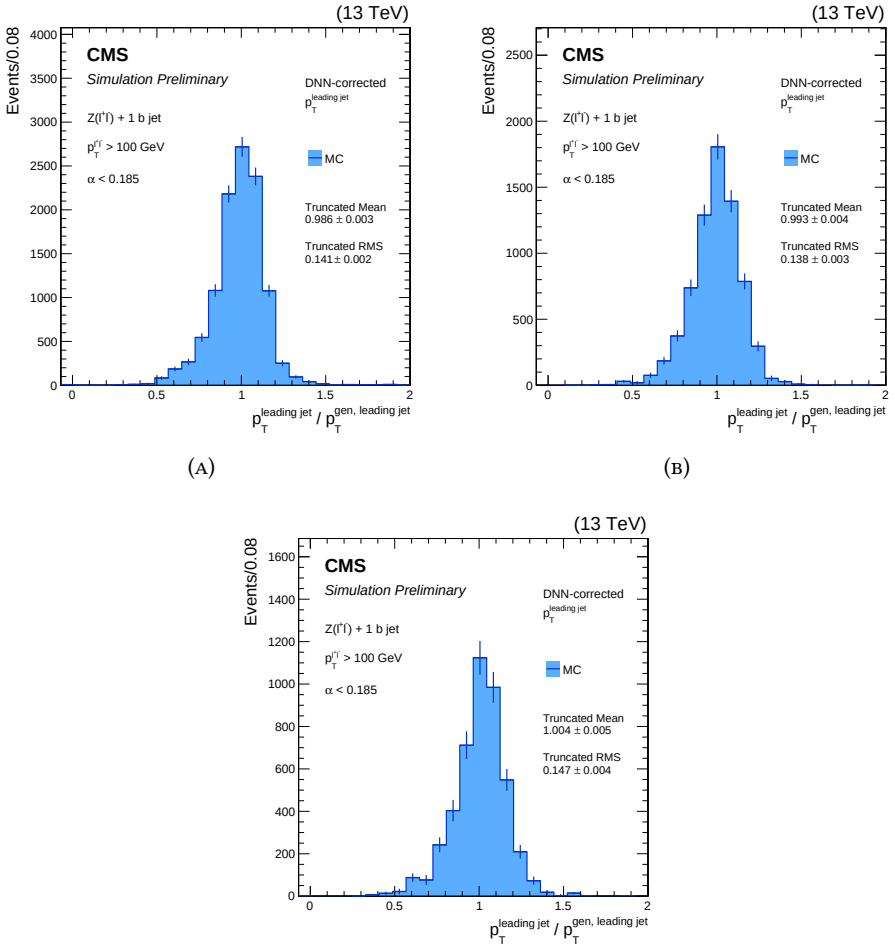


FIGURE 5.17: Distributions of the MC intrinsic resolution with 1.1 resolution SF applied to the leading jet measured using the ratio $p_{\mathrm{T}}^{\mathrm{reco}}/p_{\mathrm{T}}^{\mathrm{gen}}$ in the α bins $\alpha < 0.185$ (A), $0.185 < \alpha < 0.245$ (B) and $0.245 < \alpha < 0.30$ (C). The JER resolution SF is applied also to the extra jet used to define the α bins in this case.

5.4 Deep Vertexing

Heavy flavor jet identification has been already presented in chapter 3, with the standard tools developed and adopted by the CMS collaboration. This section presents a new b tagging algorithm based on Deep Learning, which aims to bring DL techniques at the vertex reconstruction level, and not to use the reconstructed secondary vertices built from tracks.

The tagger, using track-only information and outputting a discriminator value, is aimed to process the track information and infer the secondary vertices properties thanks to the capacity of a DNN. Since the secondary vertex properties are inferred in the hidden layers of a DNN, this algorithm is called "DeepVertex". Actually, no explicit secondary vertex reconstruction is performed, and the vertex properties cannot be retrieved, but in principle it is possible to have a DNN process lower level inputs (the tracks) and build itself the secondary vertex, hence the name DeepVertex. The fact that the secondary vertex is not explicitly reconstructed is not a problem if we want to approach just the jet flavor tagging task. At the same time, the fact that the secondary vertex is "learned" by the DNN during the optimization, can provide a more flexible definition of the secondary vertex and a representation more suitable for the DNN optimization.

The tagger was developed in parallel with the DeepJet tagger [117]. Both taggers aim for a particle level representation of the jet and let a DNN learn the discriminating features. The secondary vertex treatment is the most important difference between the two taggers. The DeepVertex tagger is based on tracks, coming from all the charged particles with no distinction among the different PF candidates. On the other hand the DeepJet tagger uses both charged and neutral PF candidates, with labels of the particle identification when they are important (i.e. electron and muon tracks). The DeepJet tagger uses also the reconstructed IVF secondary vertices, treated by the network as single particles.

In the following paragraphs both taggers are described: first DeepJet, briefly, then DeepVertex, in detail. Finally a combination of the two taggers is presented and discussed. The training of DeepJet was performed independently by people involved in the CMS b tagging working group, while the DeepVertex and combination were trained as a part of the work presented in this thesis.

5.4.1 Jet b tagging with DeepJet

The DeepJet algorithm, which exploits the capacity of deep models to process a large number of particles and features, is a recent innovation developed by the CMS Collaboration in jet b tagging. The representation of the jet used by DeepJet aims at capturing fully the description of the jet as produced by the PF algorithm. Up to 25 charged PF candidates and up to 16 neutral PF candidates per jet are used as input to the DNN. In addition the Inclusive Vertex Finder reconstruction of the secondary vertices is exploited: up to 4 vertices matched geometrically to the jet are fed to the neural network. Other variables used are the high level features that provide a global description of the jet suitable for b tagging, the so called tagging variables used since the development of the CSV algorithm, and the jet kinematics (p_T, η). The sequences of particles include respectively 16, 8 and 12 input features for the charged PF candidates, the neutral PF candidates and the secondary vertices. The global variables are 15 in total.

The network exploits the batch normalization as first layer, which is trained only for the first epoch, in order to rescale the input features on the fly. The sequences are processed

using 1×1 convolutional filters. For each collection of charged and neutral particles and vertices, separate 1×1 convolutional layers are trained: 4 hidden layers with 64, 32, 32 and 8 filters respectively for charge candidates and vertices and 3 hidden layers with 32,16 and 4 filters for neutral particles. The filters act on each particle or vertex individually. The compressed and transformed output is then separately fed into 3 LSTMs with 150, 50 and 50 output nodes respectively. The outputs from the LSTMs are merged with global jet properties which are first fed through one dense layer with 200 nodes before being passed to 7 subsequent hidden layers with 100 nodes each.

The target is the jet flavor with 6 flavor categories employed in the training. The categories are gluon, light-flavor quark, charm quark and bottom quark jets, with b quark jets further split into 3 categories: b with no leptons, b with leptonic decays and bb, which are then merged for evaluation purposes. The loss function is the categorical cross-entropy. The DNN is implemented using the Keras package with TensorFlow backend. The back-propagation uses stochastic gradient descent with the Adam optimizer.

A schematic representation of the DeepJet DNN is shown in figure 5.18, where the dimensionality of the inputs is also reported in the colored boxes. The total number of trainable parameters is about 265 000.

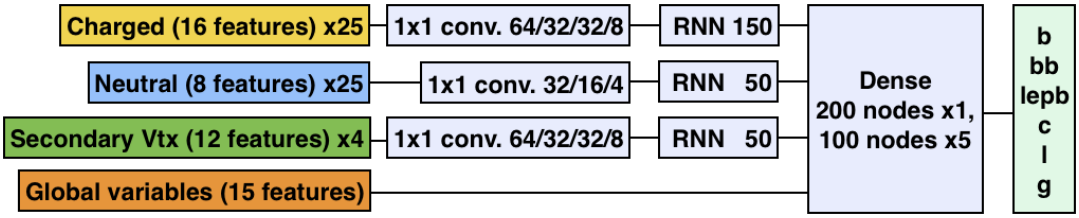


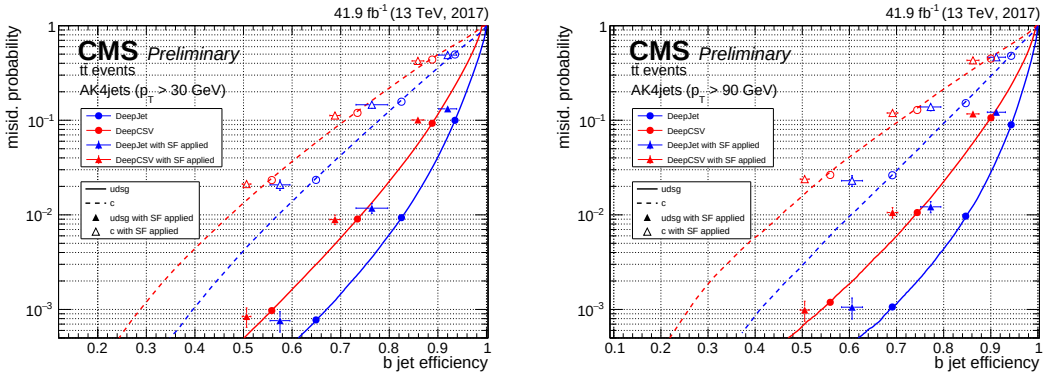
Figure 5.18: DNN architecture of the DeepJet network. The total number of trainable parameters is about 265K.

The DeepJet algorithm was trained on a sample of 100M jets. The jets come from $t\bar{t}$ production and QCD mutijet production simulated events. All flavors are included. The p_T, η and flavor distribution was built artificially drawing jets from different samples in order to avoid biased towards the event topology or the kinematics peculiar of a given flavor. More information on the training sample distribution can be found in the paragraph 5.4.4, as the sample was reproduced for the training of DeepVertex with the purpose of combining the two taggers.

The performance of DeepJet represents a step forward in the jet b tagging performances. DeepJet was found to improve on the standard algorithm both in $t\bar{t}$ and QCD mutijet simulation, against both charm quark, light-flavor quark and gluon jets. The performance of the b jet algorithm, described by the ROC curves both in $t\bar{t}$ simulation are shown in figure 5.19. The results were obtained using jets with $|\eta| < 2.5$, $p_T > 30$ GeV (A) and $p_T > 90$ GeV (B). The improvement of the blue ROC curves (DeepJet) with respect to the blue curves (DeepCSV) is sizeable everywhere. DeepCSV is actually also DNN based, but with higher level inputs are fed to a feed-forward DNN, while DeepJet uses a lower level description and Deep Learning techniques suitable to process the particles' sequences. Thanks to the more inclusive description of the jet the b tagging is also largely improved for high transverse momentum jets ($p_T > 100$ GeV).

The performance of DeepJet were also verified in data and similar scale factors to the ones of standard taggers were derived. As a result, DeepJet is currently the most performing b

tagging algorithm developed by the CMS collaboration. The performance of the b jet algorithm with scale factors for three standard working points are shown in figure 5.19. For the loose, medium and tight working points, the data-to-simulation scale factors have been applied and are represented by the triangles with error bars. Circular markers represent the performance of the respective working point in simulated samples.



(A) (B)

FIGURE 5.19: Performance of the CMS.DeepJet and DeepCSV algorithm. The ROC curves show the b tagging efficiency as a function of probability of misidentifying non-b jets as b jets. The results were obtained using jets with $|\eta| < 2.5$ and $p_T > 30$ (A) and 90 (B) GeV, from simulated top pair events. For the loose, medium and tight working points, the data-to-simulation scale factors have been applied and are represented by the triangles with error bars. Circular markers represent the performance of the respective working point in simulated data.

The development of DeepJet was fundamental also to check the impact of the dataset size on the training of such a deep model. It was found that the 100M jets were necessary to achieve the performances shown in 5.19. A training performed with 33M and 2M jets were found to have respectively 0.6 and 3% lower b tag efficiency at 10% mistag rate for light-flavor quark and gluon jets, and 2 and 8% worse b tag efficiency at 1% mistag rate for light-flavor quark and gluon jets, using a test sample of jets from $t\bar{t}$ simulated events with $|\eta| < 2.5$ and $p_T > 30$ GeV.

More information about the sample size dependence and the deployment of the DeepJet tagger can be found in reference [117]

5.4.2 Motivation for "Deep Vertexing"

A similar algorithm, but aiming for an even lower level description of the jet is DeepVertex, presented for the first time in this thesis. DeepVertex has some similarities to DeepJet. Both use sequences of objects: DeepVertex uses the tracks matched geometrically to the jet. The main difference between the two algorithms is in the handling of the secondary vertices. While DeepJet uses the IVF reconstructed vertices, in DeepVertex only tracks are fed to the network and the algorithm itself handles the raw information to tag the jets.

Several reasons lead us to bring Deep Learning to the secondary vertex reconstruction. Secondary vertices are one of the most discriminating features of b jets, but they are not all easily reconstructed. Secondary vertices with two or more well reconstructed tracks are the ideal case. In case we have decay chains even multiple vertices, one from the B hadron decay and a tertiary from the daughter D decay, can be reconstructed. In case some tracks

are misreconstructed or not reconstructed, it can be convenient to merge multiple vertices from decay chains, and to relax the criteria for matching tracks or merging candidate vertices. In any case, the vertex reconstruction is tuned to give the best efficiency, but some potential vertices are missed.

Deep Learning is an ideal tool for a more flexible secondary vertex definition, or to avoid a vertex definition at all. If the DNN has enough capability, multiple secondary vertex topologies can be used without need for tuning by hand the vertexing algorithm. Moreover, a DNN can exploit better tracks coming from a secondary vertex, but not used in a secondary vertex fit. Usually, only the impact parameters are used in this case, but we can benefit from correlating the impact parameters of multiple tracks and from looking at the position of the neighbor tracks.
When applied to the jet tagging a DNN can instead be optimized to "reconstruct" the secondary vertex inside the hidden layers of the network, starting from lower level features. In this case the secondary vertex explicit reconstruction is skipped, but it would not be necessary anyway, as the secondary vertex properties would be fed to a multivariate discriminator.

The latter approach, going end to end from raw features to jet tagging, can be seen a textbook application of Deep Learning: the network learns the ideal representation of the data in the upstream hidden layers - we know that the representation using secondary vertex features can be extremely useful - while the downstream portions of the network can use the learned representation to discriminate the jet flavor. This is in fact the approach used in "deep vertexing".
It can be compared e.g. to convolutional neural networks learning to classify images. The convolutional filters in the best performing networks usually put together and learn objects of increasing complexity as we go downstream in the network. The upstream layers "see" simple features, such as smaller colored pieces or edges, and the downstream layers combine them into larger pieces of an image that used to associate the images to a category.

5.4.3 Datasets

Several attempts were made to train the DeepVertex DNN. Initially, jets from $\mathrm{t\bar{t}}$ were used to study the network convergence and gauge the capability and performance in a sample rich of both b and light-flavor jets. The $\mathrm{t\bar{t}}$ events were generated at next-to-leading-order (NLO) accuracy in perturbative QCD with the POWHEG v2 program [106], analogously to the samples used for the b jet regression training.

Subsequently artificial samples were built drawing jet from both QCD multijet production samples and $\mathrm{t\bar{t}}$ samples with fixed flavor proportions and templated p_{T} and η distributions. This is done in order not to bias the training to jet kinematics, as in $\mathrm{t\bar{t}}$ samples b jets, coming directly from the top decays, have e.g. different p_{T} spectra compared to c and light-flavor jets. The QCD multijet process was simulated in several bins of transverse momentum of the leading jet using the generator available in PYTHIA 8.2. The processes used in the evaluation are again $\mathrm{t\bar{t}}$ production, using events independent from the training ones, and QCD multijet production, generated with the leading jet p_{T} in range [15-7000] GeV generated with the PYTHIA 8.2 generator.

For all simulated events, the standard PYTHIA 8.2 [108] with the CP5 tune [109] is employed for parton showering and hadronization. The detector is simulated by the GEANT4 [67]

package, and pileup interactions are added on top of the simulated primary vertex. No CMS data is used for this study, which is currently in a simulation-only development stage, but almost ready to be deployed for future usage in Physics analysis.

5.4.4 Training, validation and test samples

The training sample used to optimize the DeepVertex DNN was built to match the DeepJet training dataset. Jets were drawn from $t\bar{t}$ and QCD multijet simulated samples in order to fit a template distribution in p_T, η and jet flavor.

The flavor proportions are roughly 2:2:4:11 for b, c, light-flavor quark and gluon jets respectively, as shown in figure 5.20. Here and in the following plots the color convention with b jets in red, c jets in green and light-flavor quark and gluon jets in blue and light blue is adopted. b jets are not further categorized in the training of DeepVertex, and were treated inclusively also when building the DeepJet training sample.

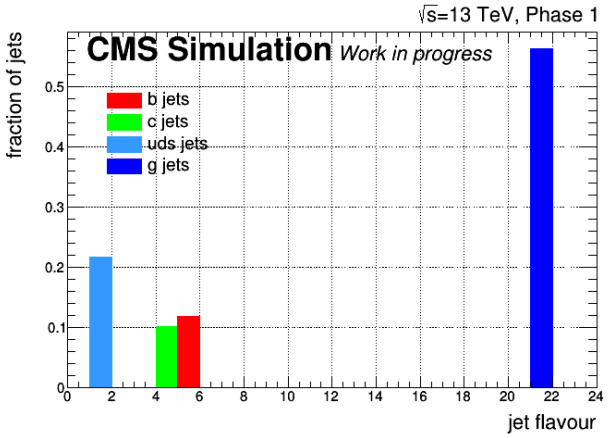
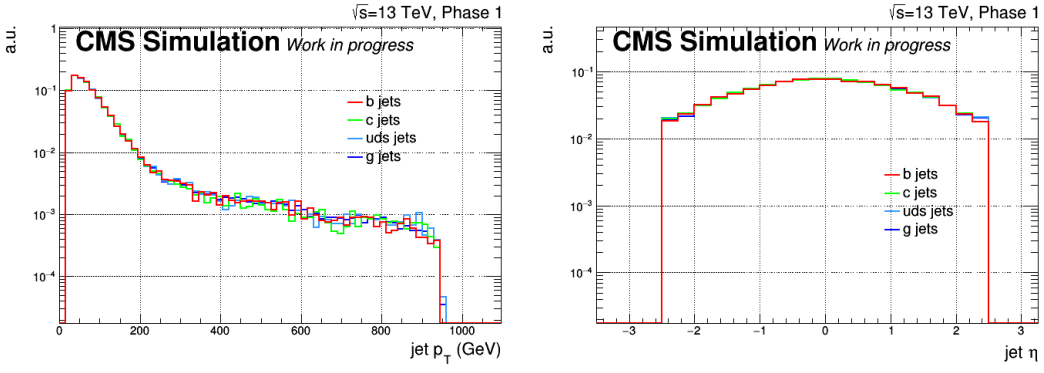


FIGURE 5.20: Flavor composition of the sample used for the training of DeepVertex. The four bins are filled with the fraction of jets per flavor. The bins are chosen according the convention: b quark jet - label "5", c quark jet - label "4", uds quark jet - label "1", gluon jet - label "21".

The kinematic distribution are equalized in jet p_T, η for all the flavor categories. The p_T is in range 20 - 1000 GeV, while $|\eta|$ is less than 2.5. The normalized distributions are shown in figure 5.21.



(A) (B)

FIGURE 5.21: Spectra of the jet used to build the sample by flavor.

The validation of the training uses two sets. A fraction of the training events, with the same kinematics and flavor distributions is used at training time to compute the loss and update the learning rate if needed. Another two validation sets, made of simulated jets from $t\bar{t}$ and QCD multijet production respectively, are used to select the best model based on the ROC curves. Finally, two test sets, made of simulated jets from $t\bar{t}$ and QCD multijet production, but statistically independent of the validation events, are used to quote the b tagging performances.

5.4.5 DeepVertex inputs

Our aim is to have a DNN solve the b tagging problem without being fed the secondary vertex information explicitly. However, the data must include the full information we could loose when reconstructing secondary vertices. We based our work on a track-only description of the jet.

Approaches using e.g. detector hits could work, but are yet to be studied. Tracks only, in the form of the track 5 parameters could be used, as they preserve the full information we want. Other possibilities include also using b tagging typical variables together with the track kinematics. This is the case in "DeepVertex": the impact parameters with respect to the primary vertex and other track features used in b tagging are used as input.
Regarding secondary vertices, it was chosen to include features that help the network infer the secondary vertex properties, instead of feeding only all the tracks independently. The secondary vertex reconstruction algorithm used by CMS (IVF, described in chapter 3), starts with the clustering of tracks about a track with significant displacement from the primary vertex. Such tracks are selected as they have higher probability to originate from a secondary and are called seeding tracks.

Clusters of tracks are built around the seeding tracks in order to keep compatible tracks, based on the track-to-track 3d distance at the point of closest approach (PCA). These clusters contain the secondary vertices. In the deep-vertexing algorithm we feed such clusters directly to a DNN, instead of using them for a fit. Both single tracks variables and relative track-to-track variables are used: single tracks variables include the kinematics and the impact parameter. Track-to-track variables include the distance at the point of closest approach, the angle between the tracks at PCA, etc.

The tracks and clusters of tracks are also chosen as inclusively as possible to let the DNN pick the important objects itself.

The full list of inputs can be summarized as follows.

- The jet 4-vector in the form (p_T, η, ϕ, m).
- The displaced tracks, used as seeding tracks, with their features. These tracks are selected using the criterion of 1 σ displacement from the primary vertex. The tracks are also required to have $p_T > 0.5$ GeV, and $dz < 0.5$ cm from the primary vertex and reduced $\chi^2 < 5$. The variables used for these tracks are:
  - The track 4-vector in the form (p_T, η, ϕ, m[1]).
  - The longitudinal and transverse impact parameter (dz, dxy).

[1] Particle flow information is used to determine the mass in case the particle is identified as a lepton.

- The impact parameter and significance in space (3D) and in the transverse plane (2D), both with and without sign. The sign is assigned to be positive if the impact parameter projection along the jet axis direction is in the positive or negative direction of the jet axis, with the primary vertex used as 0. The track probabilities used in standard b tagging (see section 3.3) for the 3D and 2D impact parameters are also used.
- Track quality information: the reduced χ^2 of the track fit, the number of pixel hits and the total number of tracker hits.
- Jet relative features of the track: the distance from the jet axis at point of closest approach between the track and the jet axis direction; the distance of this point form the primary vertex. Tracks whose minimum distance point from the PCA is more than 5 cm away from the jet axis are also removed from the seeding tracks collection.

The tracks in the sequence are sorted according to 3D signed impact parameter significance and up to 10 tracks are kept as seeding tracks in the jet. The sorting variable signed 3D SIP is shown in figure 5.22 for the ten tracks separately for the 4 jet flavor categories used at training time. The features per track are 21 in total.

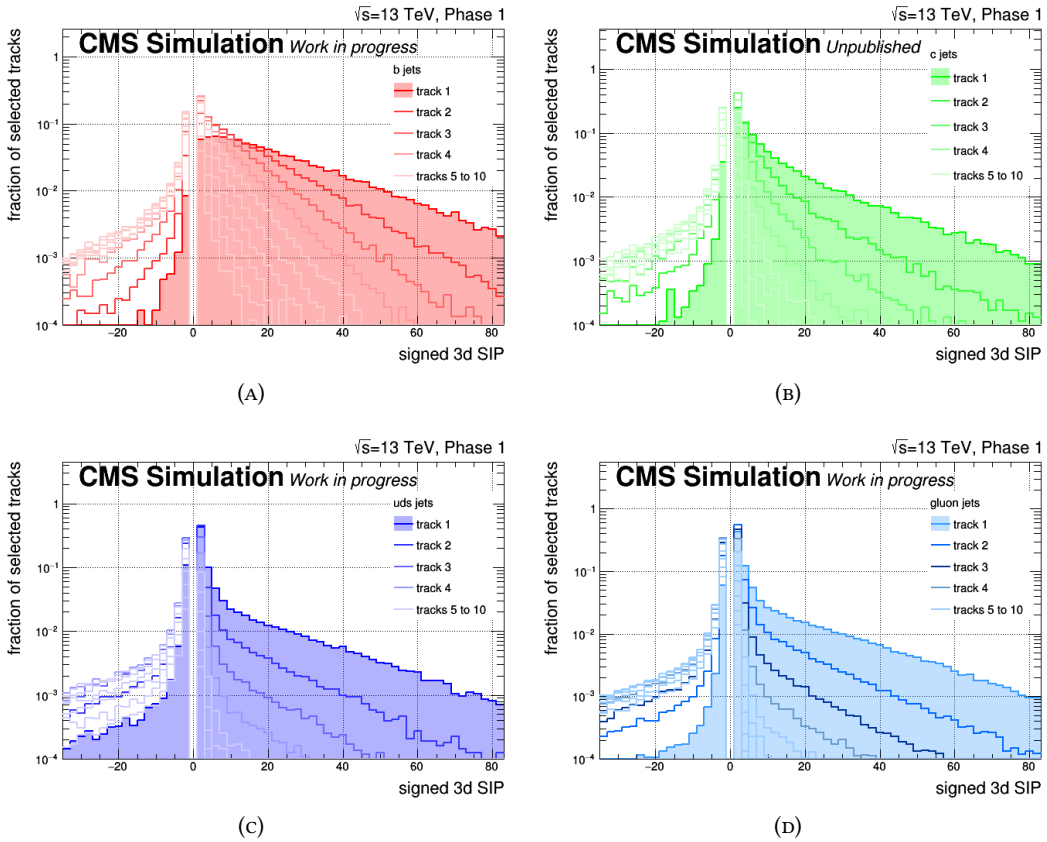


FIGURE 5.22: DNN input features - the jet 4 vector variables by jet flavor. All the distributions are normalized to unity.

- The tracks in the clusters with their features. For each displaced track a cluster of neighbors is built: the cluster is built using all the tracks with $p_T > 0.5$ GeV, and $dz < 0.5$ cm from the primary vertex, regardless of the jet angular distance and the

displacement. The tracks per cluster are sorted by distance at the point of closest approach from the seeding track and and up to 20 tracks are kept for the training. The PCA distance, used as sorting variable is shown in figure 5.23 for all the seeding tracks and all the flavors. We can notice the increasing peak and average PCA distance of the distribution from the 1st to the 20th track. Full repetition of tracks is allowed when building clusters: a seeding track can be in the cluster of another and tracks can be used in multiple clusters. The variables used for these tracks are:

- The track 4-vector in the form (p_T, η, ϕ, m[1]), the longitudinal and transverse impact parameter (dz, dxy).
- The impact parameter and significance in space (3D) and in the transverse plane (2D), all without sign.
- Track quality information: the reduced χ^2 of the track fit, the number of pixel hits and the total number of tracker hits.
- Jet relative features of the track: the distance from the jet axis at point of closest approach between the track and the jet axis direction; the distance of this point form the primary vertex.

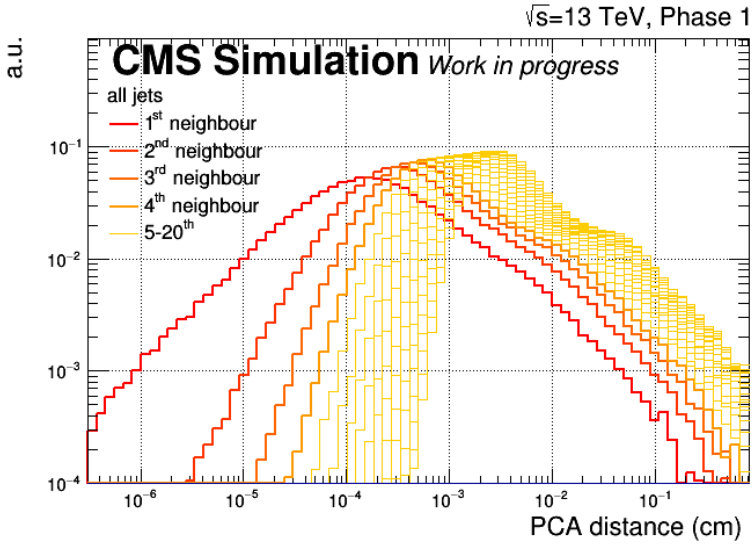


FIGURE 5.23: Shape of the input vectors used in the DeepVertex training.

For each track variable relative to the point of the closest approach to the seeding track,i.e. track-pairs variable, are used. These are:

- The PCA distance and its significance;
- the coordinates of the PCA, both on the neighbor and on the seeding track, in the form (x, y, z) and their uncertainties;
- the distance from the primary vertex of the two above points (in case of a secondary vertex it would be the decay point);
- jet relative variables: the distance if the PCA (the central one) from the jet axis, the $\Delta\eta$ and $\Delta\phi$ of the PCA direction from the jet axis direction, the scalar product between the jet direction and the direction given by the momentum sum of the two tracks.
- the scalar products between the track direction and the PCA on the track direction, both the seeding track and its neighbor, the scalar products between the

two tracks directions both in 3D and in the transverse plane, the scalar products between the two PCA directions both in 3D and in the transverse plane.

The distributions of all the inputs can be found in appendix A. The variables are separated by jet flavor and with the usual color convention. The features of all the seeding or neighbor tracks are merged into the same distribution.

5.4.6 Input data structure

Given the choice of the inputs the data format per jet is going to be made of three blocks. The first block contains the jet 4-vector, which can be useful mostly as context information. These are 4 variables.

The second block is made of up to 10 seeding tracks. In case less than 10 seeding tracks are selected in the jets the data is zero-padded to get to 10 input vector. For each track 21 variables are used: the shape of this matrix will be 10×21 per jet. The tracks are homogeneous objects: they can be sorted according to a criterion and treated as a sequence, or parameter sharing between tracks can be advantageous in the optimization.

The third block is made of clusters of tracks about the seeding tracks. For each seeding track we built clusters based on the 3d distance at PCA. These tracks are also homogeneous objects, which are treated as a sorted sequence. Tracks are sorted by distance from the seeding track and we take the first 20 tracks. For each track in the cluster we include variables relative to the track-seeding track pair, and variables relative to the track in the cluster itself. In total we have 36 variables. The shape of the tensor is therefore $10 \times 20 \times 36$. In this case we have a double folded sequence, so it can be again convenient to find techniques to share parameters.

The total number of inputs features for the network is $10 * 20 * 36 + 10 * 21 + 4 = 7414$ per jet. The vector and matrices fed to the DNN per jet can be visualized in figure 5.24. The lines of the matrix with dimensionality (21,10) contain homogeneous objects, the features of the tracks. The 10 matrices with dimensionality (36,20) are homogeneous with each other, and the lines of a matrix contain homogeneous objects, the neighbors of a seeding track. Given the large number of inputs per jet, the DNN implementation has to take into account the data structure and the homogeneity of the objects by using parameter sharing techniques.

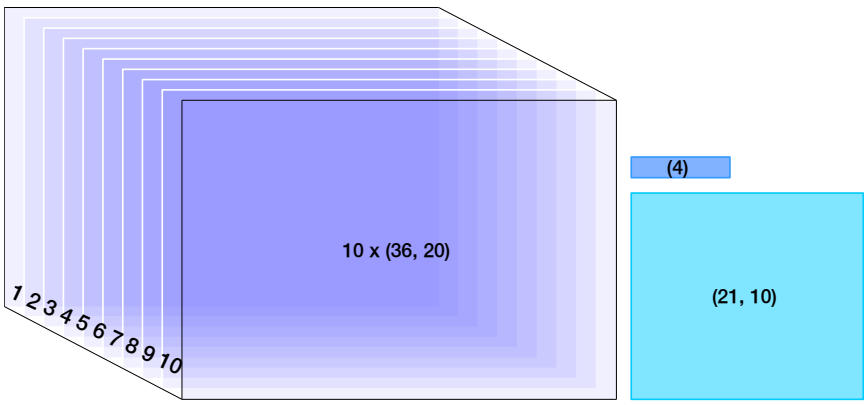


FIGURE 5.24: Shape of the input vectors used in the DeepVertex training.

5.4.7 DeepVertex implementation

Before being fed to the DNN some of the inputs are transformed. A logarithmic transformation was found to be useful for variables covering large ranges, but on average all close to 0, like the impact parameters. For the signed impact parameters the logarithm was applied to the absolute value and the sign was used to multiply again the result. The output distributions after these transformations and data standardization were applied, can be found in appendix A. The data standardization is applied inside the DNN itself as a custom layer, which applies the standardization, but has no trainable parameters.

The DeepVertex network architecture is chosen taking into account the characteristics of the inputs. The inputs contain a sequence of tracks and a sequence of sequences, the neighbors for each of the above tracks. The sequences are all sorted. Given the presence of sequences of objects, techniques for sharing parameters across objects are used. LSTM nodes can be used to process the sequence of neighbors for each track. The conv 1×1 filters can be used also where we have sequences, in order to re-optimize the input representation, both to expand and to reduce the features' dimensionality.

The network, with the shapes in input and output of each layer is shown in figure 5.25. The network has two separate branches: one for the seeding tracks and one for the clusters of neighbors. The 10 clusters matrices, with shape (36,20) are processed via a conv 1×1 layer with 64 filters. The layer is shared across the 10 clusters (magenta). The rearranged variables are passed through a shared LSTM. The LSTM is again the same for all the clusters (orange). Each cluster has 20 tracks and for each now 64 features. We use 64 LSTM nodes to collapse the track sequence dimensionality of the cluster. At this point we have 10 vectors of 64 features each. These are processed via conv 1×1 filters (red), just like the 10 seeding tracks features. A cascade of conv 1×1 with 64, 32, 32 and 8 filters is used on both sides of the DNN. The output is flattened, merged with the jet global features and fed into ad feed-forward network (grey).

The flavor categories are 4: b, c, light-flavor quark and gluon jet. The categorical cross-entropy loss function is used. The DNN is again implemented using the Keras package with TensorFlow backend. The back-propagation uses stochastic gradient descent with the Adam optimizer.

The total number of parameters of the DeepVertex network is 144 000. Most of them are actually in the dense part of the network, due to the large use of parameter sharing when processing the sequences attached to the jet.

The model is trained by using multiple CPUs for data loading and pre-processing and one GPU (initially NVIDIA Tesla K80, then NVIDIA Tesla T4) for the actual DNN training.

5.4.8 Hyperparameter optimization

The hyper-parameter optimization was not performed in a systematic way as for the regression. This was due mainly to practical considerations: the optimization of the DeepVertex model with 100M jets uses over 1 TB of data, and requires about 12 hours per epoch, due to the data loading time. The performances of the GPUs are not a limit in our case, as they are not even fully used for 100% of the time due to the data loading latency. A model needs about 30 epochs to be optimized, therefore when a single training is run a GPU is busy for two weeks.

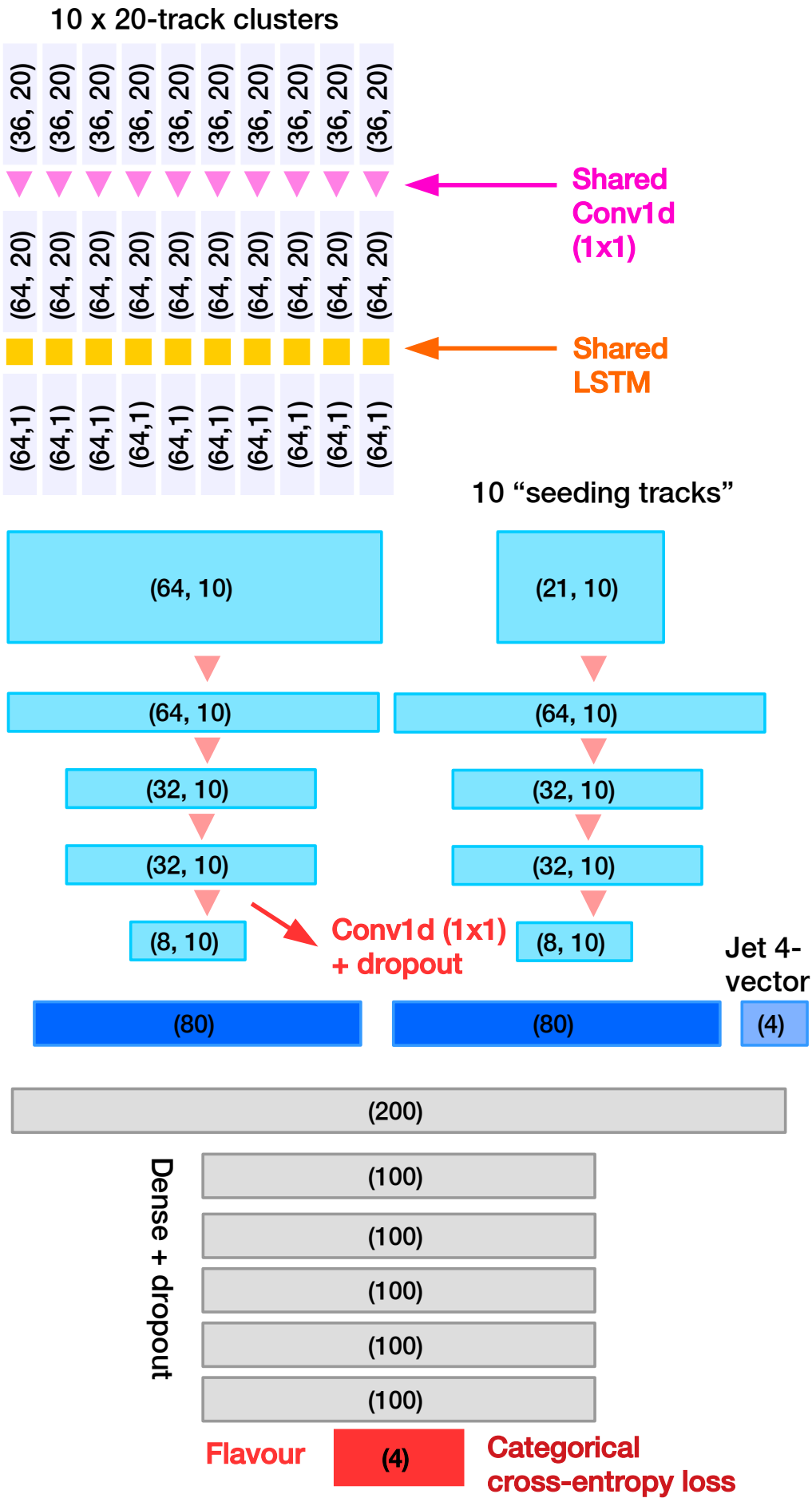


FIGURE 5.25: DNN architecture for training of DeepVertex.

Several experiments were performed, with datasets slightly different from the final one. In particular the model was trained with several loss functions aimed at optimizing the b tagging efficiency at high purity, varying the number of the hidden layers of the dense part of the DNN, making the DNN overall larger up to 250 000 free parameters. All the experiments lead to similar performances to the ones in the original model.

One of the modifications that was found to be most useful was the addition of the 1×1 convolutional filters before the LSTM nodes when processing the sequences of neighbors. The first model was updated to include these layers. Similar experiments were run also for the data pre-processing, leading to the already mentioned variables transformations and rescaling.

5.4.9 Results in simulation

The results of the DeepVertex training are presented as b tagging efficiency versus mistag rate ROC curve. The ROC curves are presented for b versus light-flavor quark and gluon jets and for b versus c jets. The ROC curves are shown inclusively for $\mathrm{t\bar{t}}$ and QCD multijet simulation for the data taking conditions of 2017 and 2018, after the new CMS pixel detector was installed (Run 2, Phase 1) in figure 5.26. The DeepVertex ROC curves are in green: the continuous line is the ROC curve showing the b versus light-flavor/gluon performances, while the dashed line shows the b versus c performances.

The DeepVertex tagger has b jet efficiencies at the standard working points comparable with the DeepJet ones in the 2017 $\mathrm{t\bar{t}}$ simulation: at 1% mistag (Medium WP) we have 82% efficiency, at 0.1% mistag (Tight WP) we have 65% efficiency. The results are similar also with the 2018 $\mathrm{t\bar{t}}$ simulation, shown in figure 5.26 (C). Comparing the DeepVertex ROC curve (green) with the DeepJet one (in blue both in 5.26 and 5.19) we can observe a consistent pattern: the DeepJet tagger has better b jet efficiency at loose and Medium working point, DeepJet and DeepVertex are even at the Tight WP, but for mistag rates < 0.1% the DeepVertex tagger is consistently better. Using such low mistag rates is not common in analysis nowadays, but could be useful in the future as more data will be collected.
The same ROC curves (b versus light/gluon) are shown in figure 5.26 (B,D) for the 2017 and 2018 QCD simulation. The performances are overall a bit worse for all the taggers, but we can observe that the DeepVertex tagger has better b jet efficiency for mistag rates $\lesssim$ 0.5%. Looking at the b versus charm jets ROC curves instead we can observe very good performances of the DeepVertex tagger, but the DeepJet tagger is better everywhere.

Similar ROC curve in jet p_{T} bins are reported in appendix B. The results are overall similar to the inclusive ones. Comparing the DeepVertex and DeepJet performance we can observe a pattern: DeepVertex is overall performing better at lower p_{T}, and the performances are worse for p_{T} > 70 GeV, while the DeepJet performances improve a bit at p_{T} close to 100 GeV and are much better at very high p_{T} (>200 GeV).

The DeepVertex ROC curve for the 2018 $\mathrm{t\bar{t}}$ simulation is also shown in comparison with the DeepCSV tagger in figure 5.27. Looking at the b versus light-flavor/gluon ROC, we can observe that the DeepVertex tagger has better performances than the previous generation of taggers.

5.4.10 Combination of DeepVertex and DeepJet

The DeepVertex and DeepJet model are both performing b taggers. The taggers are similar in terms of performance, but use different inputs. The ROC curves have also different trends. The DeepJet has higher efficiency at low and medium mistag working points (ranging from 10% to 1% mistag for light-flavor jets). DeepJet uses all the information available for b tagging, including the secondary vertices. Charged leptons are included in the charged particle flow candidates collection, but the IDs of the candidate are not explicitly fed to the network. Conversely, the DeepVertex tagger is better performing at tighter working points (with mistag rates < 1% for light jets). The DeepVertex tagger is track-only based. However, the fact that clusters of tracks and not reconstructed vertices are used can improve the efficiency in ambiguous and more difficult cases, i.e. at tighter working points. The performances of the two taggers are different also in p_T bins, with DeepJet being more performing at high p_T and DeepVertex being relatively better at low p_T.

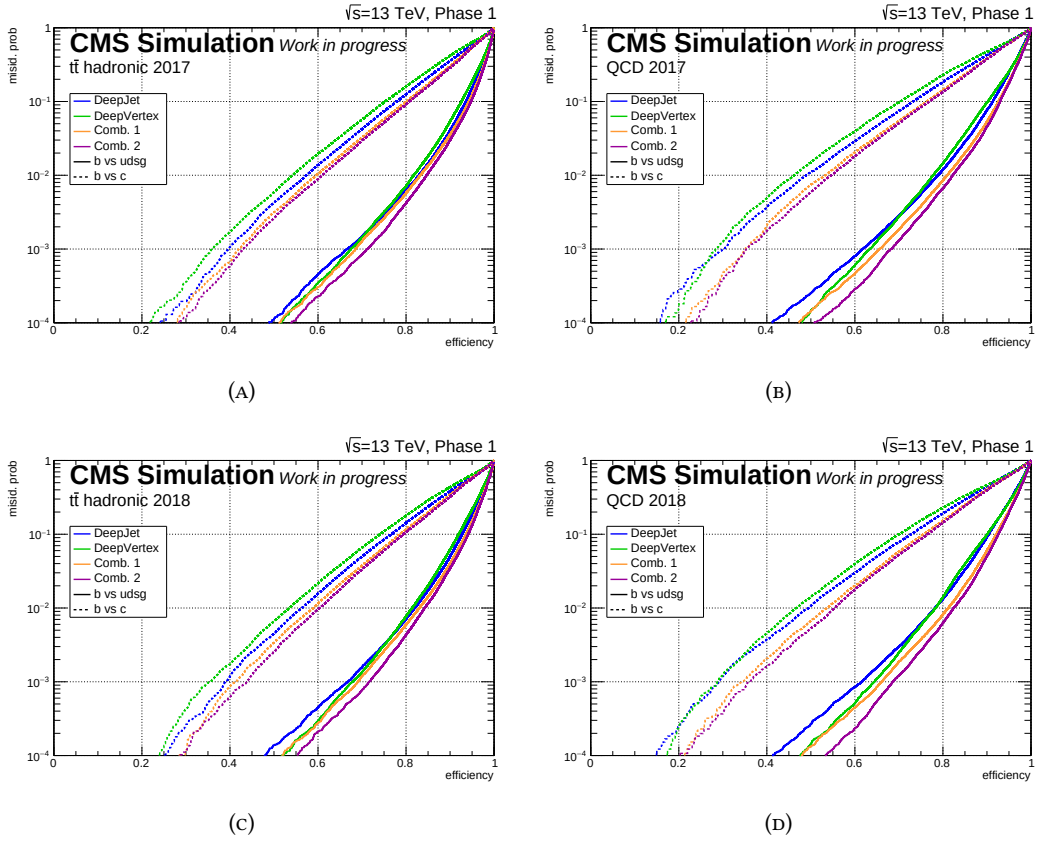
The different trends of the ROC curves and the differences between the two models, from the inputs to the network structure and the optimization, motivate a combination of the two taggers in order to further improve the b tagging performance.

The combination can be run at multiple levels. The simplest combination uses the categories outputs of the two taggers only as input. Alternatively the two architectures can be merged at the level of the outputs and some of the intermediate layers, initialized with the weights optimized in the single trained models, can be retrained in order to tune the model. Even all the weights of both networks can be unfrozen, but it is not guaranteed that training converges in that case. Finally the inputs of the two networks can be used all together to optimize a larger network. Also, in this case, it is not guaranteed that the model can be successfully optimized.

Two combinations are presented here: a high level one and a lower level one. The high level one combines the 6 categories outputs of DeepJet and the 4 categories outputs of DeepVertex. For the combination two hidden layers with 100 nodes each are added downstream. The lower level one combines the outputs of the last hidden layers before the dense part of each network. Regarding figures 5.25 and 5.18, the inputs of the grey block and the inputs of the "dense" block respectively are combined. The inputs of the last part of the network are therefore 265 from the DeepJet network and 164 from the DeepVertex network, 429 in total. Six hidden layers with [300, 200, 200, 100, 100, 100] nodes are added, and a dropout unit with dropout rate 0.1 is added after each layer. 4 categories with categorical cross-entropy loss are employed for the training of the combination, the Adam optimizer [94] is used again in both cases for the optimization. Everything is implemented in Keras with TensorFlow backend.

The results of the combinations are shown in figures 5.26 and in p_T in B: the purple lines show the performances of the lower level combination, while the orange line shows the high level one. We can see everywhere, in $t\bar{t}$ and QCD simulation, across the full jet p_T range, that the combination performs better than both taggers. The performances of the lower level combination are overall better in simulation everywhere, reaching e.g. in 2017 inclusive $t\bar{t}$ $\sim$ 85% efficiency at 1% mistag and $\sim$ 70% efficiency at 0.1% mistag rate for b versus light and gluon jets. The performances are better than all the other taggers also in b versus c efficiency versus mistag rate.

The ROC curves show that this combination is currently the best b tagger developed in CMS looking at the performances in simulation. Such a combination is therefore worth being validated in data for future analysis. Moreover the fact that the combination can improve the performance shows that we are still not using an optimal representation of the jet, which is worth working on in the future. Nevertheless, the usage of clusters of tracks instead of reconstructed vertices, being this the main difference between DeepVertex and DeepJet, can recover information and makes it worth using Deep Learning to perform the "reconstruction" of the secondary vertices from lower level objects, i.e. the tracks.



(A) (B)

(C) (D)

FIGURE 5.26: DNN results for both 2017 and 2018 simulated samples - Inclusive jet spectra for the samples $t\bar{t}$ hadronic and QCD are used.

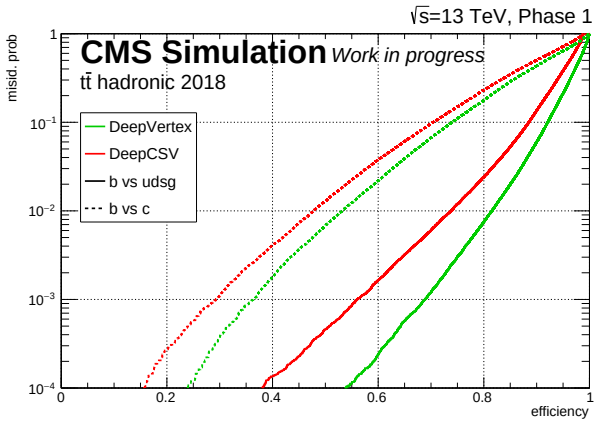


FIGURE 5.27: DeepVertex ROC curves compared with DeepCSV and with DeepJet in all hadronic $t\bar{t}$ simulation.

Chapter 6

Observation of the H $\to$ b$\bar{\text{b}}$ decay

This chapter focuses on the VH ($b\bar{b}$) analysis with 2017 data. The analysis was performed quickly after completing the analysis of the first bunch of the Run 2 data collected at 13 TeV in 2016, as the total integrated luminosity made it possible to finally reach the observation of the $H \to b\bar{b}$ decay. The strategy, the key improvements added for the 2017 data analysis, and the results are presented in this chapter. The combination with the other H ($b\bar{b}$) channels, which resulted in the observation, is also presented.

6.1 Introduction

The H$\to$ b$\bar{\text{b}}$ decay is the only hadronic decay mode of the Higgs boson we can currently probe with sensitivity at the level of the Standard Model (SM) expectation. The motivation to search for the H$\to$ b$\bar{\text{b}}$ decay is the fact that its cross section allows the direct measurement of the Yukawa coupling of the Higgs boson to a down-type quark, thus providing a test of the hypothesis that the Higgs field is the source of mass generation for fermions.

The b$\bar{\text{b}}$ final state has a few experimental advantages: in the SM a 125 GeV Higgs boson has the largest branching fraction ($\sim$ 58%) in the b$\bar{\text{b}}$ decay channel; additionally, b jets can be effectively tagged, thus removing a large amount of the backgrounds. However, tagging is not sufficient because the QCD multijet background is very large, even from b jet production only, and all the searches for H(b$\bar{\text{b}}$) need to target either associated production modes, or very specific kinematic regimes, or both, in order to be sensitive.
This is mandatory at the trigger level, as b-tagging techniques are used mostly off-line and saving events based on the presence of two b jets only is not feasible due to very high b jet cross section. Tight selections are necessary also offline to reject the backgrounds. Multivariate techniques are then fundamental to optimize the searches.

At the LHC the most sensitive production process in the search for the H$\to$ b$\bar{\text{b}}$ decay is when the Higgs boson is produced in association with a vector boson (VH). The vector boson decay into leptons is used both at the trigger level and in the offline event selection. The sensitivity of this channel is enhanced by requiring the vector boson have a large boost in the transverse plane: a $p_{\text{T}}^{\text{V}} \gtrsim 100$ GeV is typically required.

The presence of vector bosons in the final state highly suppresses the QCD multijet background, but not the W and Z production in association with jets, which is still very large, as shown in figure 6.1. Requiring a large boost of the vector boson allows a reduction of the vector boson + jets (V+jets) background and a better signal to background ratio, as shown in the bottom panel for the ZH process in figure 6.1. A large vector boson transverse momentum is also beneficial for the trigger and for the mass resolution: a high p_{T}^{V} translates into a large $p_{\text{T}}^{\text{miss}}$ in the Z($\nu\nu$)H(b$\bar{\text{b}}$) channel, and makes it accessible at the trigger level;

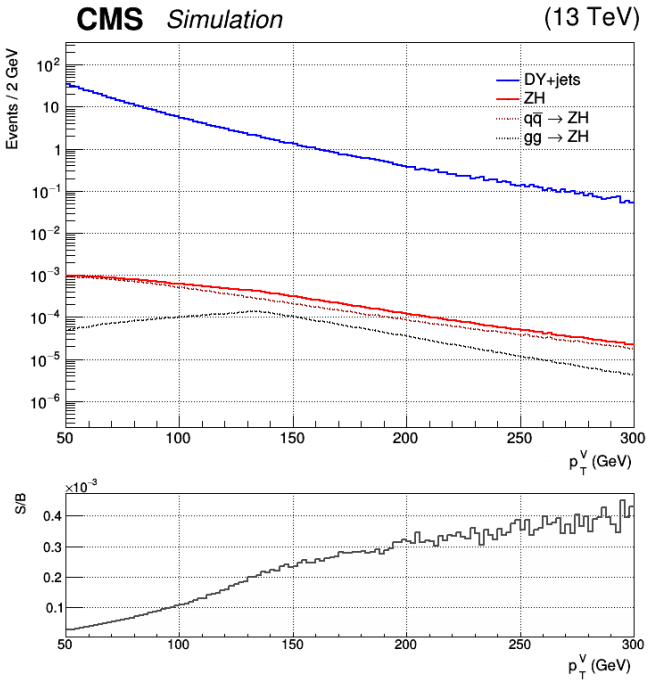


FIGURE 6.1: Top panel: differential cross section at $\sqrt{s}$ = 13 TeV for the Z($\ell\ell$)+jets SM production (blue) and the Z($\nu\nu$)H($b\bar{b}$) production (red) as a function of the generator Z p_T. The cross sections of the gg→ZH and the qq→ZH processes, which add up to the ZH total cross section (red), are shown by dotted lines. Bottom panel: S/B ratio, increasing at large p_T values.

on the other hand, the mass resolution of the reconstructed Higgs candidates is relatively better, as the jets from the Higgs boson have in turn larger transverse momenta and better p_T resolution.

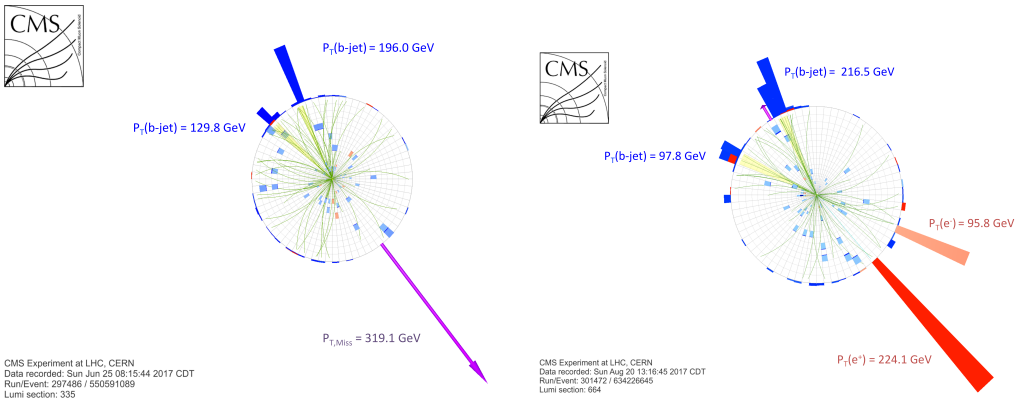
Another important production mode with good sensitivity to the H→ $b\bar{b}$ decay is the associated production with top quark pairs ($t\bar{t}H$), which was useful in combination with other final states to measure the coupling of the Higgs boson to the top quark. Searches in the VBF production mode have also been carried out. The sensitivity is similar to the one in the $t\bar{t}H$ production mode, but lower compared to the VH one. Other searches for H($b\bar{b}$) have been demonstrated to be possible. In particular, the search for inclusively produced H($b\bar{b}$) with large p_T [118] is worth mentioning. In this case the tiny phase space selected allows both to remove the backgrounds and to access the very high p_T tail of the Higgs production spectrum.

The next paragraphs focus on the VH($b\bar{b}$) analysis, which is then covered in detail for the 2017 data. The results obtained in the other channels are used for the combination (6.3.2) presented at the end of this chapter.

6.1.1 Signal characteristics

VH($b\bar{b}$) events have two b jets, with an invariant mass close to 125 GeV. The mass resolution is expected to be of roughly 20 GeV, corresponding to a relative resolution of $\sim$ 15 %.
The vector boson decay products and the dijet system are approximately back-to-back and balanced on the transverse plane, due to the large vector boson p_T requirement. Two isolated leptons (ℓ) of opposite charge and of the same flavor (e or μ) are expected for the

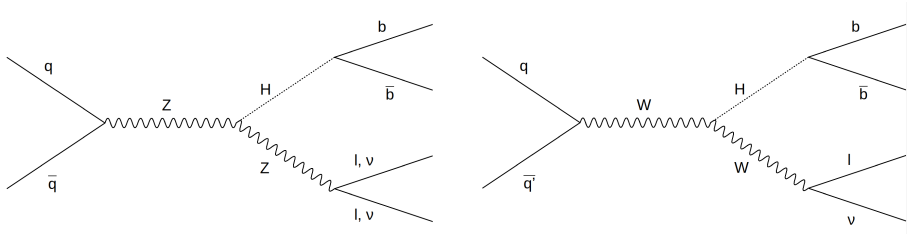
Z($\ell\ell$)H($b\bar{b}$) channel; large p_T^{miss} and no extra leptons are expected in the Z($\nu\nu$)H($b\bar{b}$) channel; one charged isolated lepton and a large p_T^{miss} are expected for W($\ell\nu$)H($b\bar{b}$). Signal like events are shown in figure 6.2, where leptons or p_T^{miss} recoil against the b jets, while two signal Feynman diagrams are shown in figure 6.3.



(A) (B)

FIGURE 6.2: Event candidate for the Z($\nu\nu$)H($b\bar{b}$) (A) and for the Z(ee)H($b\bar{b}$) production recorded by the CMS detector in 2017. The Higgs boson decays to two b quarks whose decays are characterized by jets in blue. The large missing transverse momentum due to neutrinos is in purple, the signals of electrons are in red.

Additional jets may arise from the initial state (ISR) or final state (FSR) radiation. FSR jets radiating from the b quarks are expected to be found close-by in angle with respect to the b jets, and should be taken into account in the Higgs mass reconstruction, while the ISR jets have some discriminating power for identifying signal events.
ZH events can be produced by qq scattering, as shown in fig 6.3, and also by gg scattering (see chapter 1). The latter process has a lower cross section (15% inclusively), but a harder p_T spectrum, making it important at high p_T^V, and on average more ISR jets.



(A) (B)

FIGURE 6.3: Example signal Feynman diagrams. Quark initiated processes only for ZH (A), and the WH process(B) are shown.

6.1.2 Backgrounds

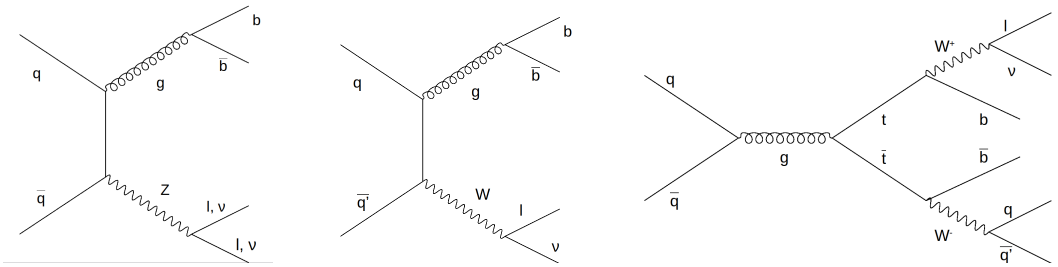
The signal cross section is about 1 pb, while most of the backgrounds cross sections are order of magnitudes larger, therefore a selection that greatly reduces them is necessary. The QCD multijet production, with the minimal requirement of H_T>200 GeV, has a cross section of order 10^6 pb. The V+jets production have cross sections of order 10^4 pb, the $t\bar{t}$ cross section is $\sim 10^3$ pb. Other processes like the single top and the diboson production have

cross sections in ranging form tenths to hundreds of pb.

One of the dominant background is the V+jets production process. In particular, the Z($\ell\ell$), W($\ell\nu$) or Z($\nu\nu$)+b jets have a signature that is identical to the signal one with the exception of the dijet invariant mass, and represent an irreducible background (see figure 6.4 A, B). Backgrounds due to Z and W+light-flavor jets can be reduced requiring jet b-tagging, but their contribution is sizeable their cross section, which is about 100 times larger compared to Z and W+b jets. Given the typical b-tagging efficiencies and fake rates, the V+light-flavor jets and V+b jets processes end up having similar yields after the event selection.

Another large background arises from $t\bar{t}$ production, in particular in final states with significant p_T^{miss}. $t\bar{t}$ events are characterized by two W bosons and two b jets in the final state. The dileptonic $t\bar{t}$ is a background for the Z($\ell\ell$)H($b\bar{b}$) search, but it can be reduced by applying requirements on the p_T^{miss} and on the dilepton invariant mass.
More importantly, the $t\bar{t}$ production is a background for the W($\ell\nu$)H($b\bar{b}$) and Z($\nu\nu$)H($b\bar{b}$). In particular, semileptonic $t\bar{t}$ events have the same signature as the signal ones except for the two extra jets (figure 6.4 C) and the dileptonic $t\bar{t}$ process contributes to the background in case a lepton is not identified.
On the other hand, the $t\bar{t}$ production looks similar to the Z($\nu\nu$)H($b\bar{b}$) signal one if at least one of the W bosons decays to leptons and the lepton from the W decay is outside the detector acceptance or is not reconstructed. A veto on the extra jets can help reduce the $t\bar{t}$ events in this case.

The production of a single top quark is also a significant background. A single top quark can be produced in association with a b quark, with a W boson or with a light quark. The final states come always with a W boson that can decay leptonically, a b jet, and other light or heavy-flavor jets.



(A) (B) (C)

FIGURE 6.4: Example leading order Feynman diagrams for some of the irreducible backgrounds. Quark initiated processes for Z/W +jets and $t\bar{t}$ are shown. Two b jets and leptons are present in all final states.

The signature of the diboson production (ZZ/WZ/WW) is very similar to the signal. Indeed, in case of a Z boson decaying to $b\bar{b}$, the events are analogous to the signal events, the only difference being that the dijet mass distribution peaks at m_Z = 91 GeV, instead of m_H = 125 GeV. Due to $\sim$ 20 GeV mass resolution the two peaks are partially overlapping. An optimal dijet mass resolution is therefore important to reduce this source of background. The VZ($b\bar{b}$) events can also be used as a "standard candle" to cross-check the VH analysis.

Finally, the QCD multijet production background does not have the same signature as the signal final states, but has a very large cross section. The presence of isolated leptons or

large p_T^{miss} and the requirement of large p_T^V are generally sufficient to reduce the large QCD multijet production enough to make it a negligible background.

6.1.3 Analysis strategy

The analysis can be summarized in two main steps. The first step is a tight event selection, while the second step consists of a multivariate analysis.

The event selection starts with the identification of leptons and vector bosons. Since there is a cross-contamination among the Z($\ell\ell$), the W($\ell\nu$) and the Z($\nu\nu$) + H signals, the three channels of the analysis, 0, 1 and 2-lepton, are labeled according to the number of selected isolated leptons.
The H $\rightarrow b\bar{b}$ decay is reconstructed from two jets: the most b-tagged jets are used. Further background rejection is achieved by exploiting signal properties as the resonant dijet mass, the back-to-back VH topology, and the reduced additional hadronic activity.

We can take as a reference for the signal and background contributions after the selection the event yields in table 6.5 (A), obtained using the 2017 simulation normalized to 41 fb^{-1}. The yields are listed by channel. The 2-lepton channel analysis is further split into two categories, low and high p_T^V, as the presence of a Z boson decaying into two leptons allows the removal of all the QCD multijet background, while in the other channel only the high p_T^V category is used.

The signal-to-background (S/B) ratios are close to 1% in the 0 and 2-lepton channels (high p_T^V), while they are ∼0.5% in the one lepton channel.
The backgrounds have multiple components everywhere: in the 0-lepton channel both the V+jets and the $t\bar{t}$ background are important. The Z+jets are dominant among the V+jets, but W+jets are present due not identified or out of the acceptance leptons. The $t\bar{t}$ and single top backgrounds are relatively prevalent in the 1-lepton channel compared to the W+jets. The Z+jets backgrounds are again due to Z($\ell\ell$) decays with leptons not identified. In the 2-lepton channel, the backgrounds are mainly due to Z+jets events.

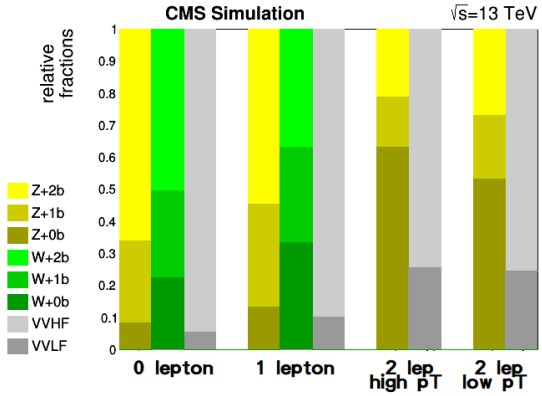
In each channel the V+jets process has been split into three components, depending on the number of true b jets in the simulated events: V+0b jets, V+1b jets, and V+2b jets. The b jets are counted as the number of simulated jets containing a B hadron and having $p_T > 30$ GeV and $|\eta| < 2.4$. The diboson (VV) background is also split into two categories, with the VV to Heavy flavor (VVHF) containing two b jets. The relative fractions of each component is shown in figure 6.5 (B). In the 2-lepton channel the b-tagging requirements in the event selection are looser, so the light-flavor component is relatively higher compared to the other channels.

It is also worth noticing that the event selection is tuned to maximize the S/B ratio, therefore a large fraction of the signal events is not used. After the event selection, if we take as a reference only the targeted decays of the vector bosons (Z($\nu\nu$), Z($\ell\ell$ and W($\ell\nu$))), the signal efficiencies are ∼3.9% (WH), ∼3.2% (qq→ZH), and ∼6.5% (gg→ZH) - 3.6% inclusively for the ZH process.

Following the event selection, the most discriminating variables in each channel are combined into a single discriminator to maximize the sensitivity to the signal. The multivariate analysis is necessary to achieve the best possible sensitivity, but it relies on a precise modeling of the backgrounds. The backgrounds are modeled using Monte Carlo simulated

Channel	0-lepton	1-lepton	2-lepton high p_T^V	2-lepton low p_T^V
Signal	93.4	144.7	54.1	103.7
Background	6041.4	26833.3	4643.5	38041.7
W+jets %	8.7	6.9	-	-
Z+jets %	38.4	1.3	91.5	77.9
$t\bar{t}$+single top %	48.6	90.8	6.7	21.1
diboson %	4.3	1	1.8	1
S/B %	**1.55**	**0.54**	**1.17**	**0.27**

(A)



(B)

FIGURE 6.5: Signal and background expected yields for the 2017 analysis by channel (A). The background components percentages are also listed. The relative fractions by flavor category for the V+jets and VV processes are shown (B). The background with top quarks is not split into flavor categories as the b jets selected are mostly true b jets. The full yields are reported in table 6.7.

samples. In order to verify the reliability of the background model in the signal region, high-purity control regions for the V+light-flavor jets and $t\bar{t}$ backgrounds are identified in data. Another region, enriched V+b jets production, orthogonal to the signal region in dijet mass, is used to constrain the V+1b jets, and V+2b jets backgrounds. Eventually, event yields or distributions of sensitive variables in the control regions are used in the final fit together with the discriminator in the signal region.

This two steps strategy was defined since the beginning of the LHC Run for the VH($b\bar{b}$) analysis with CMS data, and was employed with Run 1, 2016 and 2017 data.

6.1.4 Previous results: $H \to b\bar{b}$ decay evidence with CMS data

The search for VH($b\bar{b}$) started with 2011 data collected at 7 TeV, corresponding to 5.1 fb^{-1} [119]. The analysis was improved for the data taken in 2012, corresponding to 19.7 fb^{-1} [103]. The full Run 1 result showed an excess of events above the expected background, with a local significance of 2.1 standard deviations for a Higgs boson mass of 125 GeV, consistent with the expectation from a SM Higgs boson production. The corresponding signal strength relative to the SM predicted one was measured to be $\mu = 0.89 \pm 0.43$.

In 2016, CMS collected 35.9 fb^{-1} of pp data at 13 TeV. The analysis led to the evidence of the $H \to b\bar{b}$ decay with CMS data, with an excess of observed (expected) significance of 3.3σ (2.8σ). The H($b\bar{b}$) signal was extracted with a signal strength of $\mu = 1.2 \pm 0.4$ [35]. This was combined with the Run 1 analysis for an overall signal strength of $\mu = 1.06\ ^{+0.31}_{-0.29}$ with observed (expected) significance of 3.8σ (3.8σ).

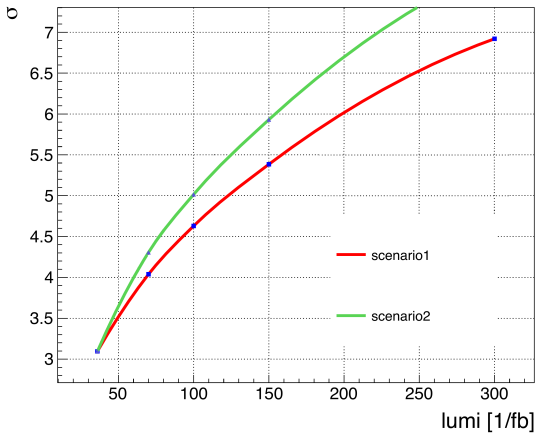
6.1.5 Projected sensitivity

After reaching the evidence of the Higgs boson decay to $b\bar{b}$, the target of the CMS collaboration was the observation of the decay, i.e. measuring an excess from the background-only hypothesis corresponding to 5 standard deviations (5σ). CMS collected ~41 fb^{-1} of data in 2017. An extrapolation of the expected sensitivity with all the available data was performed before analyzing the 2017 integrated luminosity.

The 2016 analysis was extrapolated to 70 fb^{-1}, which was roughly the available data after 2017, and to 100 fb^{-1}, with the full Run 2 in mind, and then combined with the Run 1 result. Two different scenarios were considered in this extrapolation exercise.

- Scenario 1: extrapolation with the 2016 analysis systematics unchanged, the statistical uncertainties scaled as $1/\sqrt{L}$.
- Scenario 2: the experimental uncertainties were scaled with the luminosity, while the theoretical, simulation correction and luminosity uncertainties were divided by 2.

The extrapolation of the 2016 analysis to higher luminosities is shown under the two scenarios in figure 6.6 (A), while the expected sensitivity in combination with the Run 1 results is shown in table 6.6 (B). The 5.0σ is reached with 70 fb^{-1} only under the scenario 2.



(A)

	70 fb^{-1} + Run1	100 fb^{-1} + Run1
scenario 1	4.7σ	5.2σ
scenario 2	5.0σ	5.6σ

(B)

FIGURE 6.6: Extrapolation of the 2016 VH($b\bar{b}$) analysis result to higher integrated luminosities under scenarios 1 and 2 (A); combination of the extrapolated results at 70 and 100 fb^{-1} with Run 1 VH ($b\bar{b}$) result.

This simple extrapolation exercise showed that simply replicating the 2016 analysis with the larger amount of data wouldn't have been sufficient to reach the $H \to b\bar{b}$ decay observation with the $\sim 41^{-1}$ of data available, but an improvement in sensitivity had to be sought during the analysis of 2017 data. An increase of about 10% in sensitivity would have been necessary in case the improvements were applied only to the 2017 integrated luminosity. In case they were extended to the 2016 data, an increase of the sensitivity of 5-7% would still have been necessary.

In the end, the CMS collaboration opted to publish the observation of the $H \to b\bar{b}$ decay after analyzing only the 2017 data and combining with the previously published results. However, the improvements developed for the 2017 data are now being ported to the full Run 2 luminosity, including the 2016 data, for a legacy Run 2 result.

6.2 The VH($b\bar{b}$) analysis including 2017 data

Several possibilities to improve the analysis were tested. The most important improvements were all based on Deep Learning techniques.

- Deep Learning was applied to b-tagging, as the DeepCSV tagger was employed to select the jets and build the Higgs boson candidate, while previously the CMVA tagger was used (see chapter 3).
- The b jet energy regression was upgraded to the DNN based regression described in chapter 5, while the previous analyses used a BDT based regression.
- Feed-forward DNNs were employed to build the multivariate discriminators per channel used to extract the signal significance.
- Deep Learning turned out to be useful also to improve the background modeling in the fit: a multi-category DNN was optimized to separate the different sources of background and the output of the discriminator helped to constrain individual background sources.

My contributions were important both for the regression, where I started testing deep learning tools and validated the final training, for the DNN training, and the application of the trained models to all the analysis channels.

Other improvements targeted the Higgs boson reconstruction together with the b jet energy regression. Final state radiation (FSR) jets partially overlapping with b jet used to build the Higgs candidates were merged for the invariant mass computation. The recovery of FSR was found to improve the relative dijet mass resolution, as the Higgs peak was shifted closer to 125 GeV. Finally, in the 2-lepton channel, where no genuine $p_{\mathrm{T}}^{\mathrm{miss}}$ was expected, a kinematic fit was used to constrain the dijet invariant mass and improve its resolution.

6.2.1 Datasets and simulated samples

The datasets employed, or "primary dataset" as defined in chapter 2, depend on the channel, as different trigger paths are used in each. Only good quality data collected during 2017 by CMS is used for the analysis. The data sample amounts to 41.3 fb^{-1}, corresponding to $\sim$ 83% of the total integrated luminosity delivered by the LHC and to $\sim$ 91% of the luminosity recorded by CMS. The integrated luminosities used for the analysis by LHC Run are listed in table 6.1.

Dataset	Integrated luminosity (fb^{-1})
Run2017B	~4.8
Run2017C	~9.6
Run2017D	~4.2
Run2017E	~9.3
Run2017F	~13.5
All 2017 data	41.3

TABLE 6.1: Integrated luminosities for each LHC Run used for the VH($b\bar{b}$) analysis. An uncertainty of 2.3% is assigned for the 2017 luminosity.

Signal and background processes are simulated with several Monte Carlo (MC) event generators, while the CMS detector response is simulated by the GEANT4 [67]. Simulated samples were centrally produced by the CMS collaboration and tuned to match the 2017 data taking conditions. For all samples, simulated additional proton-proton interactions (pileup) are added to the hard-scattering process with the multiplicity distribution matched to the 2017 data.

The quark induced ZH and WH signal processes are generated at NLO QCD accuracy using the POWHEG v2 [106] event generator extended with the MiNLO procedure [120, 121], while the gluon-induced ZH process is generated at LO accuracy with POWHEG v2. The Higgs boson mass is set to 125 GeV for all signal samples and the Higgs boson is forced to decay to $b\bar{b}$ pairs. The samples used are listed in detail in table 6.2.

Diboson background events are generated with MADGRAPH5_aMC@NLO generator [107] at NLO in perturbative QCD with the FxFx merging scheme [122] and up to two additional partons, and at LO with the PYTHIA 8.2 [108] generator, as backup option. NLO samples are used for the VZ($b\bar{b}$) cross-check analysis if available, while LO samples are used as background samples for the VH($b\bar{b}$) analysis and for the cross-check analysis in the Z($\nu\nu$)Z($b\bar{b}$) case.

The MADGRAPH5_aMC@NLO generator is used at LO accuracy with the MLM matching scheme [123] to generate V+jets events in inclusive and b quark enriched configurations. The V+heavy flavor component has a cross section order 100 times smaller than the inclusive V+jets production, therefore using a b-enriched configuration enhances the statistical power of the simulation in the most sensitive phase space of the analysis. The same generator and merging scheme is used for QCD multijet events. The $t\bar{t}$ and single top production processes in the tW and t channels are generated to NLO accuracy with POWHEG v2, while the s channel single top process is generated with MADGRAPH5_aMC@NLO.

The parton distribution functions used to produce all samples are the next-to-next-to-leading order (NNLO) NNPDF3.1 set [124]. For parton showering and hadronization, the matrix element generators are interfaced with PYTHIA 8.2 [108] with the CP5 tune [109].

The simulated background processes are listed in table 6.4. All the simulated background samples have equivalent luminosities of the order of 100 to 1000 fb^{-1}. For the Z+jets and W+jets processes, in particular, the samples are binned in H_T and "b-enriched" samples are used, in order to better cover the most sensitive analysis phase space, even if the low H_T bins of the simulation have equivalent luminosities close to the one of the data (41 fb^{-1}). On the other hand, the QCD-multijet simulation has an equivalent luminosity lower than the dataset one for H_T up to 2000 GeV and its equivalent luminosity ranges from 0.01 fb^{-1} ($200 < H_T < 300$ GeV), to $\sim$70 fb^{-1} for $H_T > 2000$ GeV. The QCD-multijet simulation is used just to make sure that the signal region is not affected by this type of background.

Monte Carlo reweighting

The production cross sections for the signal samples are rescaled as a function of the vector boson transverse momentum, p_T^V, to NLO electroweak accuracy, and to NNLO QCD accuracy inclusively and as described in reference [17]. The electroweak differential correction reduces the signal cross section by 10% at p_T^V near 100 GeV, and by $\sim$ 20% at p_T^V near 300 GeV

process	generator + PS	cross section (pb)	cross section × B.R. (pb)
W^+H, $W(\ell\nu)H(b\bar{b})$		0.840	0.053
W^-H, $W(\ell\nu)H(b\bar{b})$		0.533	0.034
$q\bar{q}ZH$, $Z(\ell\ell)H(b\bar{b})$	POWHEG	0.7612	0.049
$q\bar{q}ZH$, $Z(\nu\nu)H(b\bar{b})$	+	0.7612	0.089
$ggZH$, $Z(\ell\ell)H(b\bar{b})$	PYTHIA	0.1227	0.008
$ggZH$, $Z(\nu\nu)H(b\bar{b})$		0.1227	0.014

TABLE 6.2: Signal Monte Carlo samples by process, with m_H = 125 GeV. The generator and PS simulator are listed in the second column. The cross sections for the specific Higgs boson production mode are reported in the third column, while the cross sections times the branching ratios for both the Higgs and the vector boson decay are listed in the fourth column.

The production cross sections for the $t\bar{t}$ samples are rescaled to the NNLO prediction with the next-to-next-to-leading-log result obtained from TOP++ [125], while the V+jets samples are rescaled to the NNLO cross sections using FEWZ 3.1 [126].

In the V+jets samples, the p_T^V spectrum in data is observed to be softer than in simulation, as expected from higher order electroweak and QCD corrections to the production processes. Events in each channel are reweighted using a differential correction as a function of p_T^V, which reaches up to 10% for p_T^V near 400 GeV. After the above rescaling, an extra differential NLO/LO correction is applied as a function of the separation in η between the two jets from the candidate Higgs bosons, as in [35]. The NLO/LO ratio is calculated and applied separately for the $Z(\ell\ell)$+0b, $Z(\ell\ell)$+1b, $Z(\ell\ell)$+2b cases. The scale factor is ~ 1 for $\Delta\eta(jj) \lesssim 2$ and reaches up to ~ 1.5 for $\Delta\eta(jj)$ close to 4. The same scale factor is used for W and $Z(\nu\nu)$+jets samples.

The $t\bar{t}$ simulated samples require an extra correction on top of the above ones to account for an observed difference between data and simulation, consistent with [127]. The samples are reweighted as a function of the reconstructed p_T^V. A correction is derived from data in a dedicated control region in the 1-lepton channel, mainly populated by $t\bar{t}$ events, but also by W+jets. Linear reweighting functions for $t\bar{t}$, W+light-flavor jets, and the combination of W+bb and single top are extracted via a fit of the reconstructed p_T^V. The reweighting is applied to the $t\bar{t}$, W+jets and single top backgrounds in the 1-lepton channel, and to the $t\bar{t}$ simulation only in the 0 and 2-lepton channels. The linear reweighting is of order 10% at p_T^V = 200 GeV, and has a relative uncertainty of 10% (see table 6.3). The procedure was validated when performing the 2016 data analysis, and the correction derived as a function of the p_T^V was consistent with the one recommended centrally for the CMS data analysis to correct the top p_T spectrum.

Process	$t\bar{t}$	W+light-flavor	W+bb & single top
Fitted Slope (/GeV)	0.00061 ± 0.00008	0.00064 ± 0.00004	0.0016 ± 0.0001
Norm. preserving constant	1.103	1.115	1.337

TABLE 6.3: Linear corrections used for the 2017 analysis and normalization preserving constants. The normalization preserving constants maintain the process rate after the correction.

Process	generator + PS	generator level phase space selection (sub-samples)	cross section (pb)
WW	PYTHIA		115.3
WZ	PYTHIA		48.1
ZZ	PYTHIA		14.6
WW ($\ell\nu$qq)	MADGRAPH + PYTHIA		50.86
WZ ($\ell\nu$qq)	MADGRAPH + PYTHIA		10.88
WW ($\ell\ell$qq)	MADGRAPH + PYTHIA		3.69
DY+jets, Z($\ell\ell$)	MADGRAPH+ PYTHIA	$m_{\ell\ell}$ > 50 GeV	6571.9
		H_T>100 GeV, H_T bins, $m_{\ell\ell}$ > 50 GeV	270.0
		p_T^V>100 GeV, p_T^V bins, $m_{\ell\ell}$ > 50 GeV b enriched (hard scattering)	5.97
		p_T^V>100 GeV, p_T^V bins, $m_{\ell\ell}$ > 50 GeV b enriched (hadronization)	5.54
		H_T>100 GeV, H_T bins, $m_{\ell\ell}$ < 50 GeV	327.1
Z+jets, Z($\nu\nu$)	MADGRAPH+ PYTHIA	H_T>100 GeV, H_T bins	509.71
		p_T^V>100 GeV, p_T^V bins b enriched (hard scattering)	11.36
		p_T^V>100 GeV, p_T^V bins b enriched (hadronization)	10.10
W+jets, W($\ell\nu$)	MADGRAPH + PYTHIA		64057.4
W+jets, W($\ell\nu$)	MADGRAPH + PYTHIA	H_T>100 GeV, H_T bins	2274.46
		p_T^V>100 GeV, p_T^V bins b enriched (hard scattering)	15.34
		p_T^V>100 GeV, p_T^V bins b enriched (hadronization)	41.33
$t\bar{t}$, 2 lepton decays	POWHEG + PYTHIA		88.29
$t\bar{t}$, semileptonic decays	POWHEG + PYTHIA		365.34
$t\bar{t}$, hadronic decays	POWHEG + PYTHIA		377.96
single top, tW top	POWHEG + PYTHIA		35.85
single top, tW antitop	POWHEG + PYTHIA		35.85
single top, t-channel top	POWHEG + PYTHIA		136.02
single top, t-channel antitop	POWHEG + PYTHIA		80.95
single top, s-channel, leptonic decays	POWHEG + PYTHIA		3.354
QCD multijet	POWHEG + PYTHIA	H_T>200 GeV, H_T bins	1907121

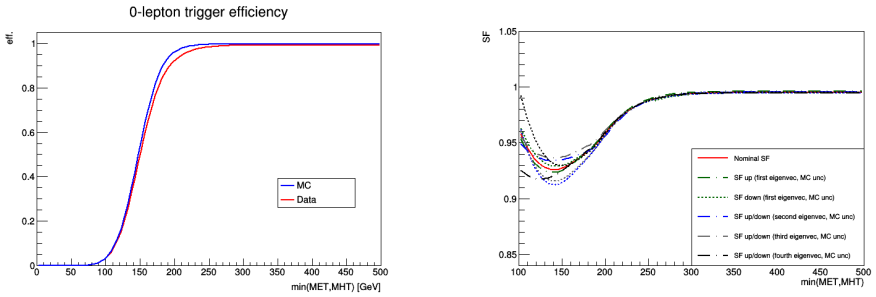
TABLE 6.4: Summary of the background Monte Carlo samples by process. The generator and PS simulator used are listed in the second column of the table. In case several samples are used, the phase space selections at the generator level are specified in the third column. The cross sections or the sum of the cross sections used in the analysis are reported in the fourth column. In case a process generated at LO is corrected to match the NLO cross section, the corrected cross section is reported. The sum of cross sections is reported in case sub-samples are generated several in bins of H_T or p_T^V.

6.2.2 Trigger strategy

The 1-lepton (e, μ) channels utilize single lepton triggers. The p_T thresholds are 27 GeV for the muon and 32 GeV for the electron trigger. The Z($\mu\mu$)H and Z(ee)H channels are instead based on dilepton triggers, with lower p_T thresholds compared to the single lepton triggers. The p_T threshold for the muons are respectively 17 and 8 GeV, with loose isolation requirements. For the electrons the p_T thresholds are 23 and 12 GeV, with loose identification and isolation required. The angular acceptances are $|\eta| < 2.5$, i.e. the full tracker coverage, for the electrons, and $|\eta| < 2.1$ for optimal muon triggering. Due to both the angular acceptances and the respective identification efficiencies, the 1-lepton triggers have signal efficiencies of approximately 95% for muons and 90% for electrons, while the dilepton trigger efficiency is 91% for muons and 96% for electrons in signal simulation.

The different trigger efficiencies between data and simulation are corrected for using scale factors. The scale factors are derived using the tag-and-probe method exploiting dilepton events from Z bosons decays, as described in chapter 3. The trigger efficiencies are measured after the application of the offline lepton identification and isolation selections.

In the 0-lepton channel the logical *OR* of two triggers is used. Both triggers require a p_T^{miss} and MHT both larger than 120 GeV, where p_T^{miss} is the opposite p_T vector sum of all the reconstructed particles in the event, while MHT is the opposite p_T vector sum of the jets with $p_T > 30$ GeV in this trigger path. The main trigger has no other requirements, while the second has the extra requirement of $H_T > 60$ GeV. The simulated trigger efficiency is corrected specifically for this analysis: the trigger efficiency is measured as a function of min(p_T^{miss}, MHT) both in data and simulation, selecting W($e\nu$)+jets events, and a correction is derived as the ratio of the two fitted functions for data and simulation. The events used are required to pass a single electron trigger, to have an isolated electron with $p_T > 37$ GeV and $\Delta\phi(e, p_T^{miss})$, two jets within the tracker acceptance, and to pass the p_T^{miss}-specific event filters.



(A) (B)

FIGURE 6.7: Trigger efficiency as function of min(p_T^{miss}, MHT) for the data and MC (A). The efficiency correction applied to simulation is shown as a function of the min(p_T^{miss}, MHT) (B). In addition to the nominal efficiency correction, the variation in the correction due to uncertainties in the function fitted to the efficiency are shown. The efficiency correction is close to 1, except in the trigger turn-on region, where the difference in efficiency between data and simulation is up to 8%.

Figure 6.7 (A) shows the trigger efficiency as a function of the offline min(p_T^{miss}, MHT) in data and simulation, obtained from a convolution of a crystal ball function and a step function fit. As shown by the data turn-on (red) the 0-lepton trigger efficiency reaches 100% above $p_T^{miss} \sim 220$ GeV. Figure 6.7 (B) shows the correction as a function of min(p_T^{miss}, MHT).

The relevant sources of uncertainty are determined by performing the eigenvector decompositions of the covariance matrices of the fitted functions, each having 5 parameters, in data and simulation. 5 eigenvectors are obtained and they are multiplied by the square root of the relative eigenvalue, then added to the fitted parameters' vector, yielding five sources of uncertainties per function. The most relevant ones are shown. The variations were then used to estimate a normalization uncertainty.

6.2.3 Event pre-selection and vector boson reconstruction

Events selected under the channel specific trigger paths are again selected offline based on the number of isolated leptons. Two opposite charge and same flavor leptons are required in the 2-lepton channel, while strictly one or zero leptons are required in the 1 and 0-lepton channels. The categories of the offline selection are mutually exclusive in order to prevent events selected by multiple trigger paths to be double counted.

Electrons in the 1 and 2-lepton channels are selected using a multivariate electron discriminator. The pseudorapidity range $1.44 < |\eta| < 1.57$ is vetoed. Two different working points based on the expected selection efficiency of either 90% (loose, WP90) or 80% (tight, WP80) are used. The loose WP90 working point is used in the event selection of the 2-lepton channel. The tighter WP80 working point is used in the 1-lepton channel. The p_T thresholds for the electrons are different in the two channels: in the 1-lepton channel the only lepton is required to have $p_T > 30$ GeV. For the 2-lepton channel the p_T thresholds are 20 GeV for both electrons.
Analogously, for muons loose identification criteria are used in the 2-lepton channel, and tight identification criteria are used in the 1-lepton channel. The muon p_T threshold in the 1-lepton channel is 25 GeV. For the 2-lepton channel 20 GeV p_T thresholds are required. All the muon candidates are also required to have $dxy < 0.05$, $dz < 0.2$ with respect to the primary vertex.

Isolation cones of radius 0.3 (0.4) in the (η,ϕ) plane around the electron (muon) momentum are used. Both muons and electrons in the 1-lepton channel are required to have a relative isolation smaller than 0.06. In the 2-lepton channel, the threshold is relaxed to 0.12 (0.15) for muons (electrons). Working points and isolation cuts for 2-lepton channels are generally looser because requiring two leptons eliminates almost all the QCD multijet background, whereas in the 1-lepton channels tighter cuts are necessary.

Candidate $Z(\nu\nu)$ decays are identified in the 0-lepton channel requiring $p_T^{\text{miss}} > 150$ GeV and no extra leptons.

The reduction factor of the preselection for the QCD multijet production is shown in figure 6.8. The colored bins show the efficiency after the trigger and the offline requirements. The QCD multijet background is reduced to only $\sim 10^3$ events in the 2-lepton channel, and it is easily removed when applying loose selections on the Z candidate p_T and mass. On the other hand, the 1 and 0-lepton channels need dedicated strategies.

In the 2-lepton channel, candidate Z($\ell\ell$) decays are reconstructed by combining the selected opposite charge electrons or muons, and by requiring $75 < M_{\ell\ell} < 105$ GeV. The dilepton candidate p_T, $p_T^{\ell\ell}$, is required to be larger than 50 GeV. The analysis is performed in two bins of $p_T^{\ell\ell}$. The "low p_T" category where $50 < p_T^{\ell\ell} < 150$ GeV and the "high p_T" category where

$p_{\mathrm{T}}^{\ell\ell} > 150$ GeV.

In the 1-lepton channel, candidate $W \to \ell\nu$ decays are identified primarily by the single isolated lepton. The transverse momentum $p_{\mathrm{T}}^{\mathrm{W}}$ and mass $\mathrm{M_T}$ of the W candidate are computed as:

$$p_{\mathrm{T}}^{\mathrm{W}} = \sqrt{(p_{\mathrm{T},x}^{\mathrm{miss}} + p_{\mathrm{T},x}^{\ell})^2 + (p_{\mathrm{T},y}^{\mathrm{miss}} + p_{\mathrm{T},y}^{\ell})^2} \qquad \text{and} \qquad \mathrm{M_T} = \sqrt{(p_{\mathrm{T}}^{\mathrm{miss}} + p_{\mathrm{T}}^{\ell})^2 - (p_{\mathrm{T}}^{\mathrm{W}})^2}$$

The analysis is performed in one category with $p_{\mathrm{T}}^{\mathrm{W}} > 150$ GeV.

In the 0-lepton channel, the transverse momentum of the Z candidate is defined as $p_{\mathrm{T}}^{\mathrm{Z}} = \min(p_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{MHT})$, as in the 0-lepton trigger path. The $Z(\nu\nu)$ analysis is performed in one $p_{\mathrm{T}}^{\mathrm{Z}}$ category with $p_{\mathrm{T}}^{\mathrm{Z}} > 170$ GeV.

The transverse momentum normalized distributions for the V+jets backgrounds and the signal after the preselection are shown in figure 6.9 for the 1-lepton (A) and 0-lepton (B) channel, after the event categorization. The $\mathrm{t\bar{t}}$ and the residual QCD multijet backgrounds are also shown. The distributions are shown with only the lepton preselection applied, thus showing the better S/B ratio at high $p_{\mathrm{T}}^{\mathrm{V}}$.
The $\mathrm{M_T}$ distributions for the 1-lepton channel are shown in figure 6.10 after the lepton preselection only (A) and with $p_{\mathrm{T}}^{\mathrm{W}} > 150$ GeV (B). For inclusive W+jets production, the distribution of $\mathrm{M_T}$ reflects the characteristic Jacobian peak and is effective at separating W events from the QCD multijet production at small values of the transverse mass. However, the discriminating power is not the same for the WH signal, due to the harder p_{T} spectrum. This is evident also for the W production itself when $p_{\mathrm{T}}^{\mathrm{W}} > 150$ GeV is required. Therefore, no selection is applied on $\mathrm{M_T}$, but the variable is used in the multivariate analysis.

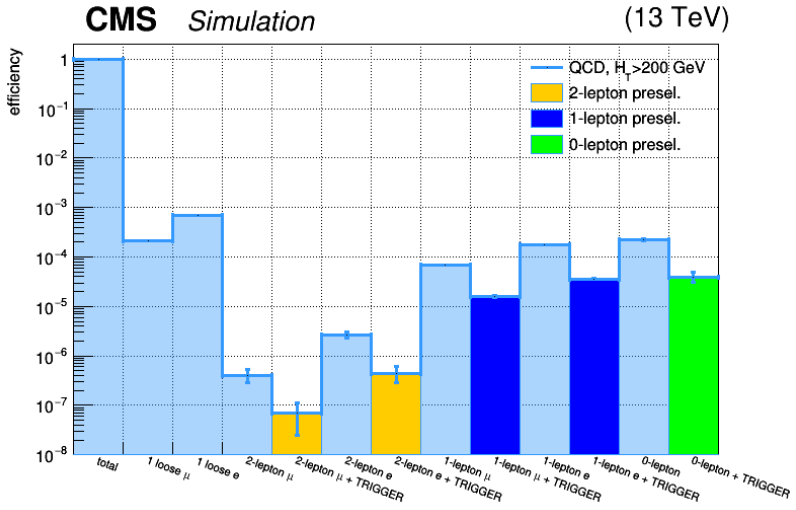
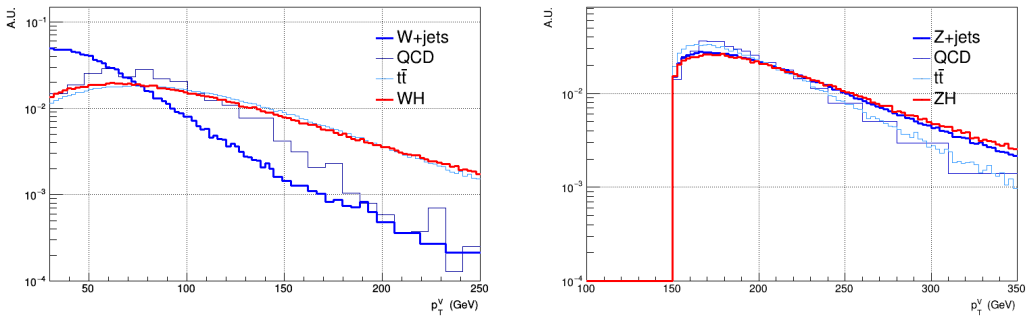
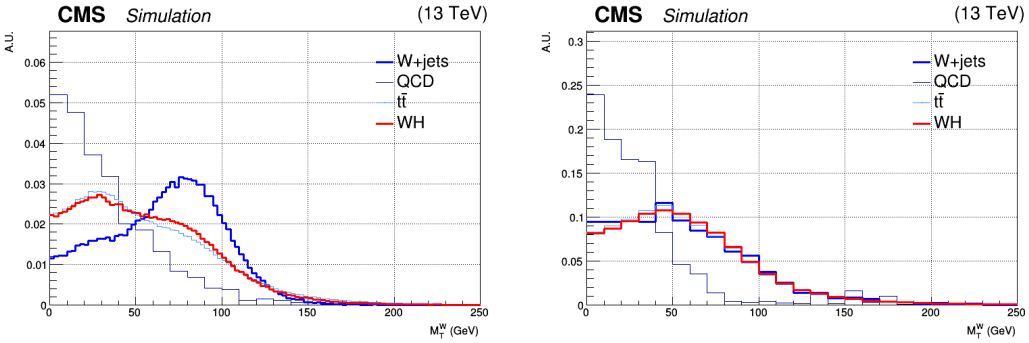


FIGURE 6.8: Background reduction of QCD multijet events based on lepton requirements only. The vector boson transverse momentum is not considered at this stage, except for the 0-lepton channel, where events with $p_{\mathrm{T}}^{\mathrm{miss}} > 150$ GeV only are preselected. The colored bins show the efficiency by channel and lepton flavor after the trigger and off-line lepton requirements.



(A) (B)

FIGURE 6.9: Vector boson transverse momentum distribution in signal , QCD, and V+jets in events in the 1-lepton channel (A) and in the 0-lepton channel (B). The $t\bar{t}$ background is included in the 1-lepton case. The distributions are normalized to unit area after the pre-selection.



(A) (B)

FIGURE 6.10: M_T distributions after the lepton selection for W+jets, signal and the $t\bar{t}$ background in the 1-lepton channel (A). The p_T^V > 150 GeV selection is applied in (B). The distributions are normalized to unit area after the selection.

6.2.4 Higgs candidate reconstruction

The jets in the event most likely originating from b quarks are used to identify the Higgs boson candidates.

Loose jet identification criteria are applied to reject misreconstructed jets resulting from detector noise, as well as jets primarily due to pileup. Jets that overlap geometrically (ΔR < 0.4) with preselected electrons or muons are discarded. Jets are required to be within the tracker acceptance ($|\eta|$ < 2.5) in order to perform b-tagging. In the 0-lepton channel, transverse momentum thresholds of 60 and 35 GeV are required for the leading and subleading jet respectively. In the 1-lepton channel a p_T threshold of 25 GeV is used for both jets, while a looser requirement, p_T > 20 GeV, is applied in the 2-lepton channel.

The jets used to reconstruct the Higgs boson candidate are selected as the most b-like jets in the event, using the DeepCSV discriminator. Additionally, the jets are required to meet at least the loose b-tagging working point criterion, corresponding to $\sim$ 10% efficiency for light quark and gluon jets and $\sim$ 90% efficiency for b jets. In the 0 and 1-lepton channel the jet with the highest b-tag score is also required to meet the tight b-tagging WP criterion, corresponding to $\sim$ 0.1% efficiency for light quark and gluon jets and $\sim$ 55% efficiency for b jets. The b jet regression described in chapter 5 is applied to b jets used to build the Higgs candidate in order to improve the mass resolution. The b jet energy regression is one of the Deep Learning improvements introduced for the 2017 dataset. Previously a BDT regression was employed. A comparison of the two techniques, using the dijet invariant mass distribution in Z($\ell\ell$)H(b$\bar{\text{b}}$) signal events, is shown in appendix C.

FSR recovery

In the hadronization process, the b quarks can emit radiation that is not clustered in the b jet and is detected as a final state radiation (FSR) jet. FSR jets are typically soft and collinear to the originating hadron, but not enough to be clustered in the same jet. They can be recovered effectively by looking at the jets close in angle to the b jets.

In order to recover the FSR emission and reconstruct the invariant mass of all the products of the Higgs boson decay, the 4-vectors of the Higgs candidates are corrected by adding the 4-vector of additional jets selected among those within ΔR < 0.8 of either Higgs candidate b jet and passing the p_T > 20 GeV and $|\eta|$ < 3.0 selection cuts. Recovering the FSR jets changes the selection acceptance and improves the relative mass resolution of signal events by $\sim$ 2% thanks to shift in the mass peak, without sculpting the background shape. The net effect of FSR recovery in Z($\ell\ell$)H(b$\bar{\text{b}}$) signal events is also shown in appendix C.

Kinematic fit in 2-lepton channel

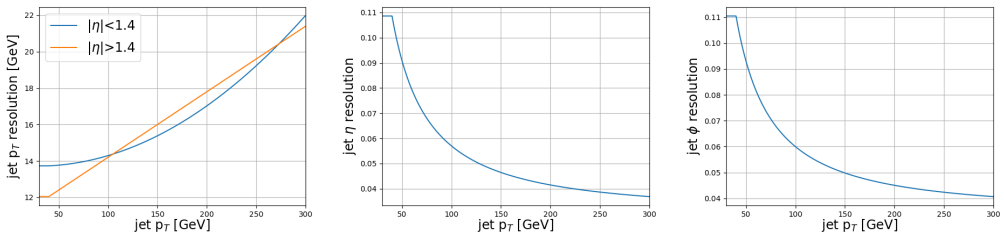
The resolution on the kinematic properties of final state objects can be improved also by applying kinematic constraints based on the target event topology and an event-by-event least square fitting technique, as shown in [128]. The kinematic constraints, which can be defined just for some of the particles or for the full event, are applied by means of Lagrange multipliers. The technique, usually called "kinematic fit", was tested successfully also in another VH(b$\bar{\text{b}}$) search in the 2-lepton channel [129].

The 2-lepton channel of the VH(b$\bar{\text{b}}$) analysis is well suited to the application of the kinematic fit. Two leptons and at least two jets are expected, but no genuine missing energy. The vector sum of the transverse momenta of all particles events should be zero. This kinematic

constraint can be used to improve the transverse momentum estimate of the jets, which have lower momentum resolutions compared to the leptons, and eventually of the dijet invariant mass. Additionally the fact that the leptons come from a Z boson decay can be exploited.

The final state objects considered for the kinematic fit and their properties given as input to the fit are listed below.

- Two Higgs candidate jets: b jets from the Higgs candidates are used, after the jet energy regression and FSR recovery are applied. The b jets resolutions in p_T, η, ϕ are based on standard recipes developed for the analysis in [130]. The resolutions used are shown in figure 6.11.
- Two lepton candidates: electrons and muons are used. The estimate of the per-lepton momentum uncertainty, which is used for the p_T variance, is the one measured for CMS standard objects. The angular resolutions, of order 10^{-4}, are considered negligible.
- The hadronic recoil vector: additional jets, which are not used to build the Higgs candidate, are summed and used. They are required to have $p_T > 20$ GeV, and to pass loose pileup rejection and jet identification criteria, and have $|\eta|$ up to 5. The covariance in p_x and p_y of the recoil vector is estimated using standard recipes. A value of 8 GeV is used as resolution in both directions for the χ^2 minimization.



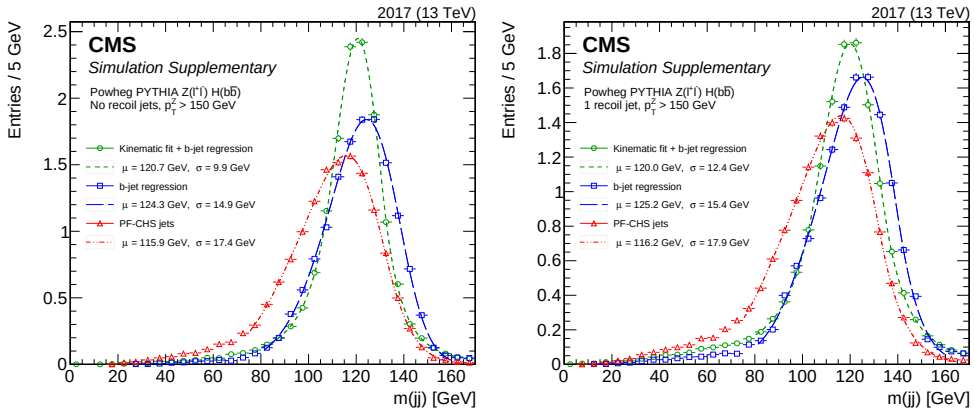
(A) (B) (C)

FIGURE 6.11: Jet transverse momentum and angular resolutions used in the fit. The p_T (A), η (B) and ϕ (C) resolution are shown as a function of the jet p_T.

The constraints applied are the nominal Z boson mass and the absence of genuine p_T^{miss}. The invariant mass of the 4-vector sum of the two lepton candidates is constrained to 91 GeV, by varying the momenta within their uncertainties. The transverse momentum of the dilepton+dijet+recoil system is then constrained to be zero, varying the jets 4-vector within their resolutions, which is equivalent to requiring no p_T^{miss}.

Dijet mass resolution

After all the event selection criteria are applied, the dijet mass (m(jj)) resolution is approximately 15%. After the regression, the FSR recovery and the kinematic fit, the m(jj) resolution is in the 8 - 13% range, depending on the channel, the p_T of the reconstructed Higgs boson and the number of extra jets. The dijet mass for the 2-lepton channel is shown in figure 6.12 in two categories, based on the presence of extra jets beside the two b jets used to build the Higgs candidate, to gauge the improvements due to the regression and the kinematic fit. The fitted resolution are reported in table 6.5



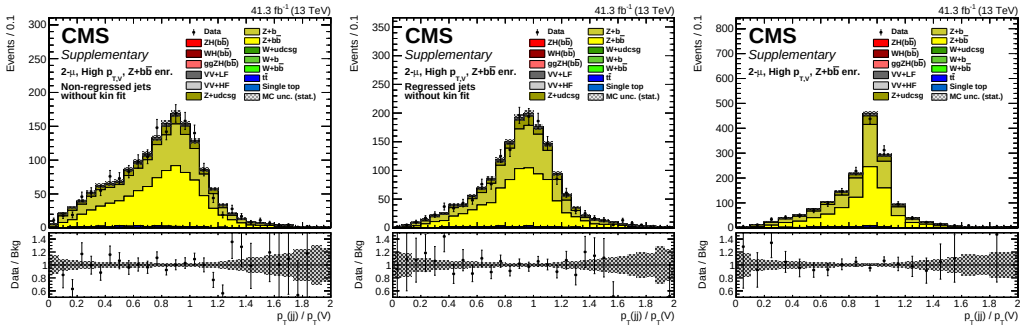
(A) (B)

FIGURE 6.12: Distributions of m(jj) for simulated signal samples in the 2-lepton channel for events with no exta jets (A) and with one recoiling jet (B). The distributions are shown before (red) and after (blue) the regression application, and after the kinematic fit procedure (green) is used after the regression. A Bukin function [131] fit is performed. The fitted mean and width are displayed in the legend [132].

p_T^V	N.ISR jets	σ_{baseline}	σ_{reg}	σ_{kinfit}
>150 GeV	0	17.4 GeV	14.9 GeV	9.9 GeV
>150 GeV	1	17.9 GeV	15.4 GeV	12.4 GeV
>150 GeV	>1	18.9 GeV	15.9 GeV	14.4 GeV

TABLE 6.5: Resolution on m(jj) in the 2-lepton channel high-p_T bin. The resolution is listed before the regression, after the regression and the FSR recovery and after the kinematic fit in 3 categories depending on the number of extra jets used in the kinematic fit.

Apart from the validation procedure used specifically for the b jet energy regression and described in chapter 5, all three improvements are validated in data by studying the p_T balance between the dijet and the dilepton system in samples of Z($\ell\ell$)+jets events containing at least one b-tagged jet, and by studying the top quark mass distribution in a high-purity sample of $t\bar{t}$ events [132]. The p_T balance distribution are shown in figure 6.13 for the Z($\ell\ell$)+jets events before the regression (A), after the regression (B) and after the kinematic fit (C).



(A) (B) (C)

FIGURE 6.13: Ratio of the dijet p_T to the p_T^V in the 2-lepton V+b jets control region [132]. The b jets in the center plot have been updated by the regression. The dijet p_T resolution is visibly improved from (A) to (B). In (C) the b jet energies are updated once again with a kinematic fit, which constrains the b jet energies using the full event, showing good data/MC agreement.

6.2.5 Signal Region selection

The signal regions are primarily defined by the dijet mass window: 60 < m(jj) < 160 GeVis required in the 0-lepton channel, while 90 < m(jj) < 150 GeV is required in the 1- and 2-lepton channels.

Other channel specific selection are applied channel by channel to reduce specific backgrounds.

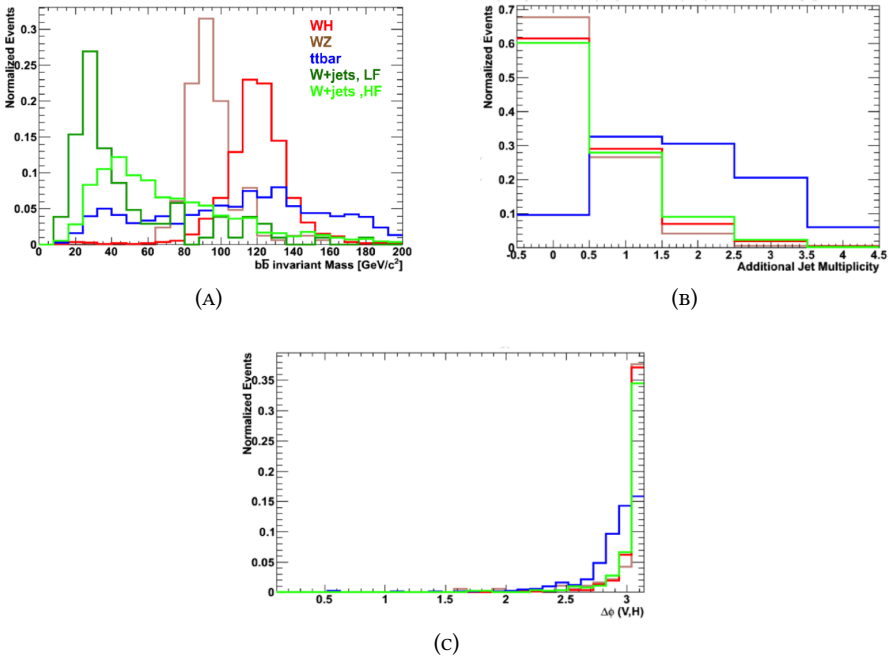
- Events with 2 or more additional jets are removed from the 1-lepton signal region to reduce the $t\bar{t}$ background; events with additional loosely identified leptons are removed from both the 1 and 0-lepton signal regions.
- In order to reject the QCD multijet background in the 0-lepton channel, an "anti-QCD" selection is applied. The "anti-QCD" selection aims to reject both neutrinos in jets and mismeasured jets, with the latter dominating at p_T^V > 170 GeV. $\Delta\phi(j, p_T^{miss})$ > 0.5 is required for all jets with p_T > 30 GeV.
- A selection on the $\Delta\phi$ between the vector boson and the candidate Higgs boson ($\Delta\phi$(V,jj)) is also used the for 0 and 1-lepton channel, to enforce the back-to-back topology requirement.
- Finally, the 0-lepton signal region is cleaned further by requiring the track-p_T^{miss}, computed with charged particles only, to be approximately aligned with the p_T^{miss}. Furthermore, in the 1-lepton signal region a maximum angle requirement between the lepton and the p_T^{miss} is imposed.

The selection criteria for the signal regions are summarized in table 6.6. The normalized distributions of some of the features are shown for signal and background in figure 6.14.

Variable / Channel	0-lepton	1-lepton	2-lepton
b-tag$_{j1}$, b-tag$_{j2}$	Tight+Loose	Tight+Loose	Loose+Loose
p_T^{j1}, p_T^{j2}	> (60,35)	> (25,25)	> (20,20)
p_T^{ℓ}	–	(> 25, > 30)	> 20
p_T(jj)	> 120	> 100	–
m(jj)	[60 – 160]	[90 – 150]	[90 – 150]
p_T^V	> 170	> 150	[50 – 150], > 150
$m_{\ell\ell}$	–	–	[75 – 105]
p_T^{miss}	> 170	–	–
N. extra jets	–	< 2	–
N. extra lepton	= 0	= 0	–
$\Delta\phi$(V,jj) (rad)	> 2.0	> 2.5	–
$\Delta\phi(p_T^{miss}$, track-$p_T^{miss})$ (rad)	< 0.5	–	–
$\Delta\phi(p_T^{miss}, \ell)$(rad)	–	< 2.0	–
anti-QCD	Yes	–	–

TABLE 6.6: Signal region selection criteria for each channel. The values listed for kinematic variables are in units of GeV, otherwise it is reported explicitly.

The total event yields of the simulated signal and background, by process, are reported in table 6.7, together with the total data yield. The numbers are all reported by channel, lepton flavor and p_T^V category. The discrepancy between the total MC yield and the data yield is covered by the uncertainties included in the fit. The QCD multijet contribution was determined to be negligible in simulation.



(A) (B)

(C)

FIGURE 6.14: Distributions of some of the features employed in the event selection. The 1-lepton channel m(jj), N. extra jets, $\Delta\phi$(V,jj) are shown. The selection criteria were optimized in previous iterations of the analysis.

Channel /Process	0-lepton	1-lepton (e)	1-lepton (μ)	2-lepton (ee) low p_T^V	2-lepton ($\mu\mu$) low p_T^V	2-lepton (ee) high p_T^V	2-lepton ($\mu\mu$) high p_T^V
WH	11.96	59.82	80.76	-	-	-	-
qqZH	61.53	1.52	2.61	14.63	20.67	32.99	50.23
ggZH	19.94	-	-	7.80	10.99	8.16	12.33
W+jets, 0b	117.47	254.9	362.75	-	-	-	-
W+jets, 1b	143.35	236.77	311.19	-	-	-	-
W+jets, 2b	265.07	288.45	391.97	-	-	-	-
Z+jets, 0b	193.75	18.09	30.18	1098.69	1589.47	6181.34	9608.18
Z+jets, 1b	595.88	50.36	65.77	283.12	384.80	2247.08	3587.16
Z+jets, 2b	1530.84	76.43	121.45	376.69	518.52	3077.49	4926.98
single top	310.15	1509.88	2004.31	6.11	12.44	112.73	161.18
$t\bar{t}$	2625.88	8835.16	12005.3	111.53	180.76	2965.82	4801.21
VV (Heavy flavor)	245.13	103.09	139.81	24.54	35.91	111.92	168.82
VV (Light flavor)	13.84	7.7	19.71	8.03	12.84	31.04	60.79
Signal	93.43	61.34	83.37	22.43	31.66	41.15	62.56
Background	6041.36	11380.84	15452.43	1908.72	2734.74	14727.41	23314.32
Total MC	6134.79	11442.18	15535.8	1931.15	2766.39	14768.56	23376.87
data	6892	11268	16054	1826	2695	13325	21737

TABLE 6.7: Signal region pre-selection event yields for each process by channel. The total data yield is also reported.

6.2.6 Multivariate analysis

The signal is extracted fitting a DNN trained to discriminate signal from background. Due to the presence of different background sources, all modeled using simulated samples, several control regions, each enriched in events from individual background processes are selected. As already pointed in the introduction a precise model of the backgrounds is needed as much as a robust discriminator in order to have the best possible sensitivity. A simultaneous binned-likelihood fit of the signal region, and of the control regions for all channels is used to extract the signal. The variables used in the fit are chosen depending on the region.

Three control regions are designed to be enhanced in $t\bar{t}$ events, V+light-flavor jets and V+b jets. They are labeled as $t\bar{t}$, V+LF and V+HF. All the control regions are mutually exclusive from each other and with respect to the signal regions.

The normalizations of the $t\bar{t}$, W+0/1/2 b jets and Z+0/1/2 b jets (see table 6.7) are treated as unbiased nuisance parameters in the signal extraction fit. The ratios of the fitted normalizations to those predicted by MC are usually referred to as Scale Factors (SF).
Different scale factors are used for the same processes in different channels, except for the W+jets ones in the 0-lepton channel, which are in common (i.e. correlated in the fits) with the 1-lepton channel, as they can model also potential residual differences in the physics object selection. However, in the 1 and 2-lepton channels the scale factors are correlated between muons and electrons, as the lepton efficiencies are taken into account as systematic uncertainties.
As some of the control regions are not pure in the processes we want to model, it can be necessary to further discriminate among processes in the control regions, in particular in the V+HF control regions, to better isolate V+b jets events.

The selections for each control region are listed by channel in the next paragraph. Data to Monte Carlo comparisons for a few selected variables in control regions are shown in figure 6.15.

0-lepton channel control regions

The $t\bar{t}$, Z+LF, Z+HF control regions selections for the 0-lepton channel are listed in table 6.8. All the regions are characterized by high $p_{\mathrm{T}}^{\mathrm{miss}}$. The selections can be compared to the signal region ones in table 6.6, as only the requirements that are different are listed here.

- The $t\bar{t}$ control region is defined by requiring at least two additional jets (besides the two b-tagged jets) with p_{T} >30 GeV, at least a medium working point (1% mistag - 70 % efficiency) b-tagged jet among the best b-tagged ones, and at least one isolated lepton.
- The Z+LF control region is defined inverting the b-tagging cut and removing the m(jj) window requirement. The remaining cuts are identical to the Z+HF control region.
- The Z+HF control region is the most similar to the signal region, only requiring an inverted m(jj) selection. The anti-QCD cut is used to increase the Z+b jets purity.

Variable / CR	$t\bar{t}$	Z+LF	Z+HF
m(jj) (GeV)	–	–	$\notin [60 - 160]$
b-tag$_{j1}$, b-tag$_{j2}$	Medium+Loose	<Medium+Loose	Tight+Loose
N. extra leptons	≥ 1	= 0	= 0
N. extra jets	≥ 2	≤ 1	≤ 1
$\Delta\phi(p_T^{miss}, \text{track-}p_T^{miss})$ (rad)	–	< 0.5	< 0.5
min $\Delta\phi$(j, p_T^{miss}) (rad)	$< \pi/2$	–	–
anti-QCD	–	Yes	Yes

TABLE 6.8: Definition of control regions for the 0-lepton channel. Only the selection that are different from the signal region ones are reported.

1-lepton channel control regions

The $t\bar{t}$, Z+LF, Z+HF control regions specific selections for the 1-lepton channel are listed in table 6.9. The criteria defining the control regions are the same for the 1-lepton (μ) and 1-lepton (e) selections.

- The $t\bar{t}$ control region is defined by requiring one tight b-tag and increasing the requirement on number of additional jets to 1 or more jets (besides the two b jets).
- The W+LF control region is defined by inverting the b-tagging requirement of the signal region, to enhance the light-flavor jets contribution. The m(jj) window requirement is also removed.
- The W+HF has the same b-tagging requirements as the $t\bar{t}$ control region and the signal region, and no additional jets. In addition, a dijet invariant mass window veto is applied to remove the overlap with the signal region.

Variable / CR	$t\bar{t}$	W+LF	W+HF
m(jj) (GeV)	< 250	< 250	< 250, > 90 $\notin [90 - 150]$
b-tag$_{max}$	>Tight	<Medium	>Tight
N. extra jets	> 1	–	= 0
N. extra leptons	= 0	= 0	= 0
p_T^{miss} significance	–	> 2.0	> 2.0
$\Delta\phi(p_T^{miss}, \ell)$(rad)	< 2	< 2	< 2

TABLE 6.9: Definition of control regions for the 1-lepton channel. Only the selection criteria that are different from the signal region ones are reported.

2-lepton channel control regions

The control regions specific selection criteria for the 2-lepton channels are listed in table 6.10. The selection criteria are common for the electrons and muons.

- The $t\bar{t}$ control region is defined by inverting the dilepton invariant mass cut.

- The Z+LF control region is defined by inverting the b-tagging requirement of the signal region, to enhance the light-flavor jets contribution.
- The Z+HF control region is the most similar to the signal region. An inverted m(jj) selection is applied to remove any overlap with the signal region.

Variable / CR	$t\bar{t}$	Z + LF	Z + HF
m(jj) (GeV)	–	$[90, 150]$	$\notin [90, 150]$
b-tag$_{j1}$, b-tag$_{j2}$	Tight+Loose	<Loose+Loose	Tight+Loose
$m_{\ell\ell}$ (GeV)	$\notin [0, 10]$, $\notin [75, 120]$	$[75, 105]$	$[85, 97]$
p_T^{miss} (GeV)	–	–	< 60
$\Delta\phi$(V,jj) (rad)	–	> 2.5	> 2.5

TABLE 6.10: Definition of control regions for the 2-lepton channel. Only the selections that are different from the signal region ones are reported.

DNN for signal versus background discrimination

The goal of the analysis is to perform a fit of a multivariate discriminator able to distinguish the signal from the backgrounds. In all the previous iterations of the analysis BDTs were trained by channel. The input features were optimized using iterative procedures.

DNNs were employed for the first time in 2017 data analysis: as for the BDTs, the DNNs for signal are trained separately for each channel using simulated samples for signal and all background processes. The set of input variables was chosen based on the previous BDT optimizations. The main changes in the inputs were due to the introduction of the kinematic fit in the 2-lepton channel. The input variables are considered after the regression and FSR recovery in all channels.

Among the most discriminating variables, the p_T and the b-tag discriminators of the two jets used to build the Higgs boson candidate are used in all the channels. The dijet system variables are also important: the m(jj), p_T(jj) and the angular separation in η among the two jets are used in all channels. In the two lepton channel the kinematic fit is applied. The kinematic fit returns also the resolution on the dijet invariant mass which is added to the DNN input variables.

The vector boson transverse momentum, p_T^V, is used in all channels; the dilepton mass is added to the input variables in the 2-lepton channel. Other discriminating variables rely on the fact that the vector boson and the Higgs boson are back-to-back in the hard scattering rest frame: the transverse momentum balance and the $\Delta\phi$ between the Higgs boson and vector boson candidate are used.

The extra jets, which describe the event topology, are exploited in all channels. Additionally, the track jets (or soft-activity jets, see 3) counter is used. For the VH analysis the track jets with an overlap with the selected leptons or jets are not counted. The track jets provide a clean estimate of the extra hadronic activity in the event, which is generally larger in background events compared to VH events.

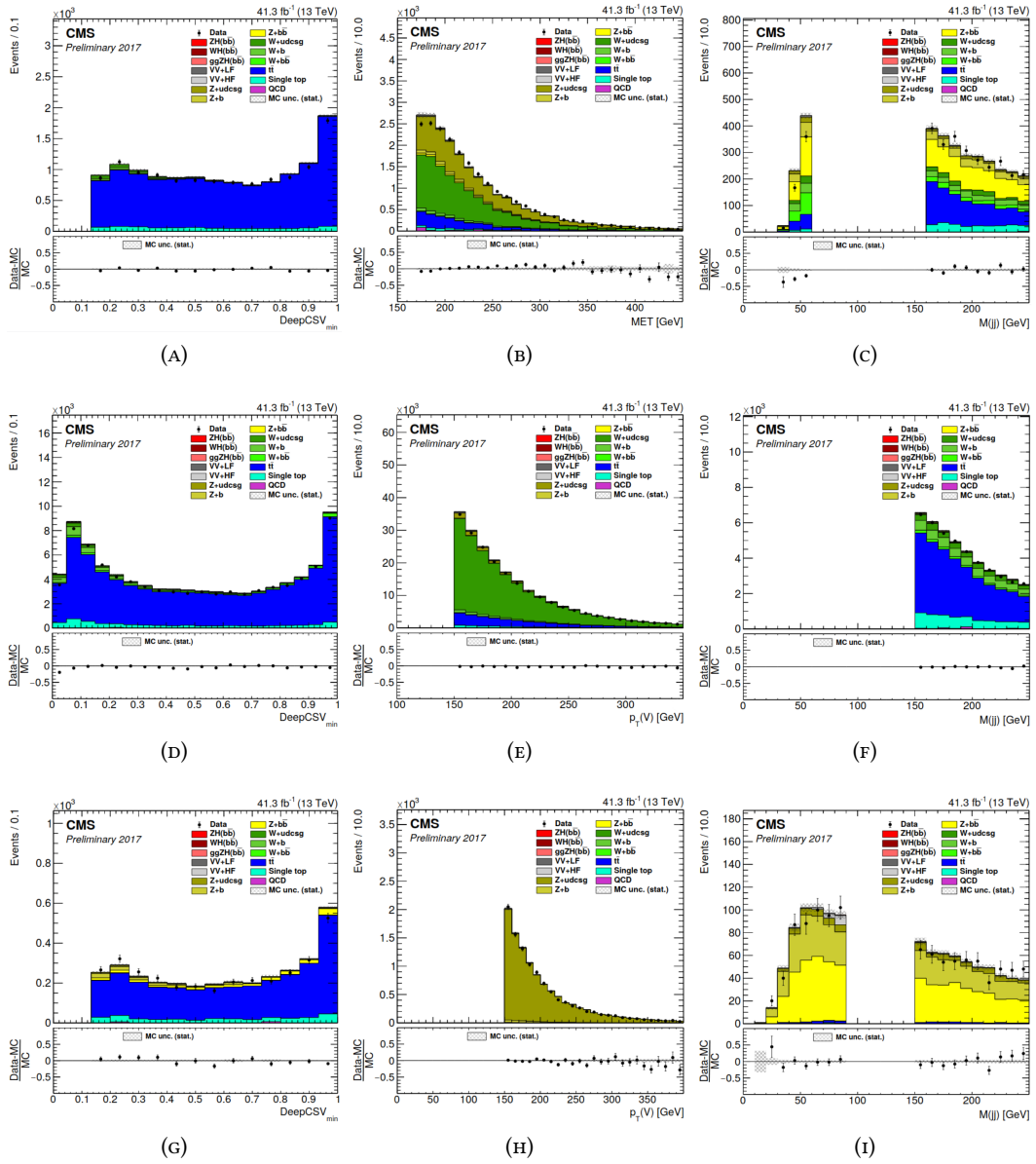


FIGURE 6.15: Control regions distributions of some chosen variables: here reported are the the b-tag$_{j2}$ for the $t\bar{t}$ control regions,the p_T^V for the V+LF regions and the dijet mass for the V+HF control regions. The plots are shown for the 0-lepton channel control regions (A,B,C), the 1-lepton channel control regions (D,E,F) and the 2-lepton channel high-p_T^V control regions (G,H,I), respectively. The scale factors derived from the a global control region and signal region fit are applied to the distributions.

Some variables are highly discriminating in the 1-lepton signal region, which is dominated by the $t\bar{t}$ background. Among those the ϕ angle between the lepton and the p_T^{miss}, the W boson transverse mass and the reconstructed top mass (m_{Top}). The full list of DNN inputs per channel is listed in the table 6.11.

The DNN is trained using half of the simulated sample, so that the remaining half is used for the validation and the application stages. In this case a proper test sample is not used, but the compatibility between the output in the training and the validation samples ensures there is no overtraining. The total number of training events is of about 100 000 per channel

0-lepton	1-lepton	2-lepton
$p_{\mathrm{T}}^{\mathrm{j1}}$, $p_{\mathrm{T}}^{\mathrm{j2}}$	$p_{\mathrm{T}}^{\mathrm{j1}}$, $p_{\mathrm{T}}^{\mathrm{j2}}$	$p_{\mathrm{T}}^{\mathrm{j1}}$, $p_{\mathrm{T}}^{\mathrm{j2}}$
b-tag$_{\mathrm{j1}}$, b-tag$_{\mathrm{j2}}$	b-tag$_{\mathrm{j1}}$, b-tag$_{\mathrm{j2}}$	b-tag$_{\mathrm{j1}}$, b-tag$_{\mathrm{j2}}$
m(jj)	m(jj)	m(jj) (kin. fit)
-	-	$\sigma_{m(\mathrm{jj})}$ (kin. fit)
p_{T}(jj)	p_{T}(jj)	p_{T}(jj) (kin. fit)
$\Delta\eta$(jj)	$\Delta\eta$(jj)	$\Delta\eta$(jj)
$\Delta\phi$(jj)	-	-
-	-	$m_{\ell\ell}$
$p_{\mathrm{T}}^{\mathrm{V}}$	$p_{\mathrm{T}}^{\mathrm{V}}$	$p_{\mathrm{T}}^{\mathrm{V}}$ (i.e. $p_{\mathrm{T}}^{\ell\ell}$)
$\Delta\phi$(V,jj)	-	$\Delta\phi$(V,jj) (kin. fit)
-	p_{T}(jj)/$p_{\mathrm{T}}^{\mathrm{V}}$	p_{T}(jj) (kin. fit)/$p_{\mathrm{T}}^{\mathrm{V}}$
$\max_{\text{extra jets}}$ b-tag$_{\mathrm{j}}$	-	-
$\max_{\text{extra jets}}$ $p_{\mathrm{T}}^{\mathrm{j}}$	-	-
$\min_{\text{extra jets}}$ $\Delta\phi(p_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j})$	-	-
-	N. extra jets (1-lep)	N. extra jets (2-lep)
-	-	N. recoil jets (kin. fit)
N. track jets (p_{T}>5 GeV)	N. track jets (p_{T}>5 GeV)	N. track jets (p_{T}>5 GeV)
-	m_{Top}	-
-	$\Delta\phi(p_{\mathrm{T}}^{\mathrm{miss}}, \ell)$	-
-	$\mathrm{M_T}$	-
-	$p_{\mathrm{T}}^{\mathrm{miss}}$	$p_{\mathrm{T}}^{\mathrm{miss}}$

Table 6.11: List of input variables used in the training of the multivariate discriminators, and their use in the different lepton category.

and category.

Two categories are used at training time: background, including all background reweighted according to their cross sections, and signal. All the input variables are standardized subtracting the mean and dividing by the standard deviation of the original distributions.

Several attempts were performed to optimize the DNN architecture and loss function, both

in Keras package [112] + TensorFlow [113] and in pure TensorFlow. In particular two loss function were tested: the standard binary cross-entropy and a customized loss function that aimed at maximizing the S/B ratio.

The final optimized architecture is a feed-forward network, consisting of 5 hidden layers with 32 nodes for each layer. A 10 or 20% dropout and batch normalization are applied after each layer to ensure regularization (see figure 6.16). The leaky ReLU [111] is used as activation function (see figure 4.3), and the binary cross-entropy as loss function (see 4.4).

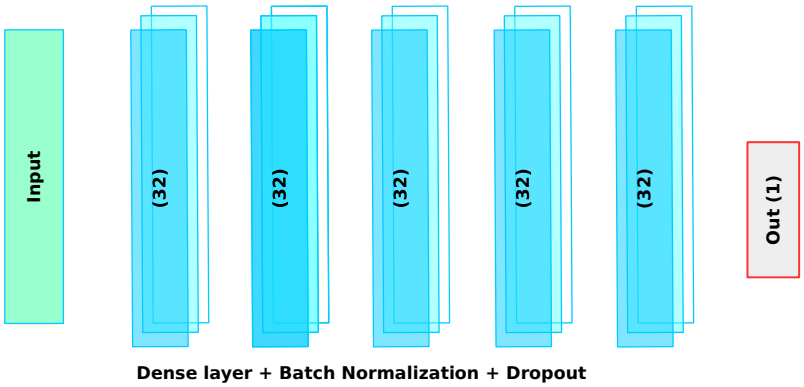


Figure 6.16: Schematic representation of the optimized DNN architecture.

For the final fit the DNN score was transformed using the mapping

$$x \rightarrow (\sqrt{x} + x^{12})/2,$$

and 15 equidistant bins between 0 and 1 were used. The mapping ensures that the signal is better sampled in the most sensitive bins of the distribution.

Deep learning for background modeling

The DNN score is used as the fitted variable in each signal region, while different strategies are used in the control regions. For the $t\bar{t}$ and V+LF control regions, which have very good purity in the target background process, only the yields of these processes are considered in the fit.

For the V+HF control region, in the 2-lepton channel the main components are V+jets and V+b jets. The minimum b-tag discriminator among the two jets is used to fit the scale factors in 2 bins: one bin has a larger component of V+light-flavor jets, while the second bin is more enriched in V+b jets.

In the 1 and 0-lepton channels the V+HF control region is enriched in V+b jets, but contains several background processes (see table 6.12), including $t\bar{t}$ and single top production.
A dedicated DNN (DNNHF) is therefore trained to distinguish among the background components. The DNNHF uses the same variables as the signal region DNN, but is trained to individually distinguish the $t\bar{t}$, single top and V+jets in 3 categories: 0b, 1b, 2b. The same architecture of the signal region DNN is also used, with the only difference being the last layer, which has 5 output nodes instead of one, and the softmax activation instead of the

Channel /Process	0-lepton	1-lepton (e)	1-lepton (μ)
W+jets, 0b	5.1	7.2	6.9
W+jets, 1b	4.3	6.5	6.5
W+jets, 2b	3.5	2.9	2.8
Z+jets, 0b	3.8	0.4	0.5
Z+jets, 1b	12.7	1.4	1.5
Z+jets, 2b	29.1	0.6	0.6
single top	9.8	13.7	13.6
$t\bar{t}$	31.2	66.7	67.0
VV (Heavy Flavor)	0.5	0.4	0.4
VV (Light Flavor)	<0.1	0.1	0.1

TABLE 6.12: Control region background percentages for the V+HF control regions of the 0 and 1-lepton channels.

sigmoid. The loss function used or the DNNHF is the categorical cross-entropy.

The DNN outputs five probabilities: one for each background process. The distribution used in the final fit is made of five bins: a bin is filled if the DNN output for the event has the maximum probability in the corresponding category. The three DNNHF distribution for the 0-lepton and 1-lepton (μ and e) channels are shown in figure 6.17. The discrepancies between data and simulation are adjusted by scale factors (and to a lesser extent by the nuisance parameters) obtained from the simultaneous control region and signal region fit. The post-fit distribution are shown in section 6.2.8.

Statistical treatment

The significance of the observed excess of events in the fit is computed using the profile likelihood asymptotic approximation. The test statistic used to compute the significance and the treatment of the nuisance parameters follow the recommendations in [133].

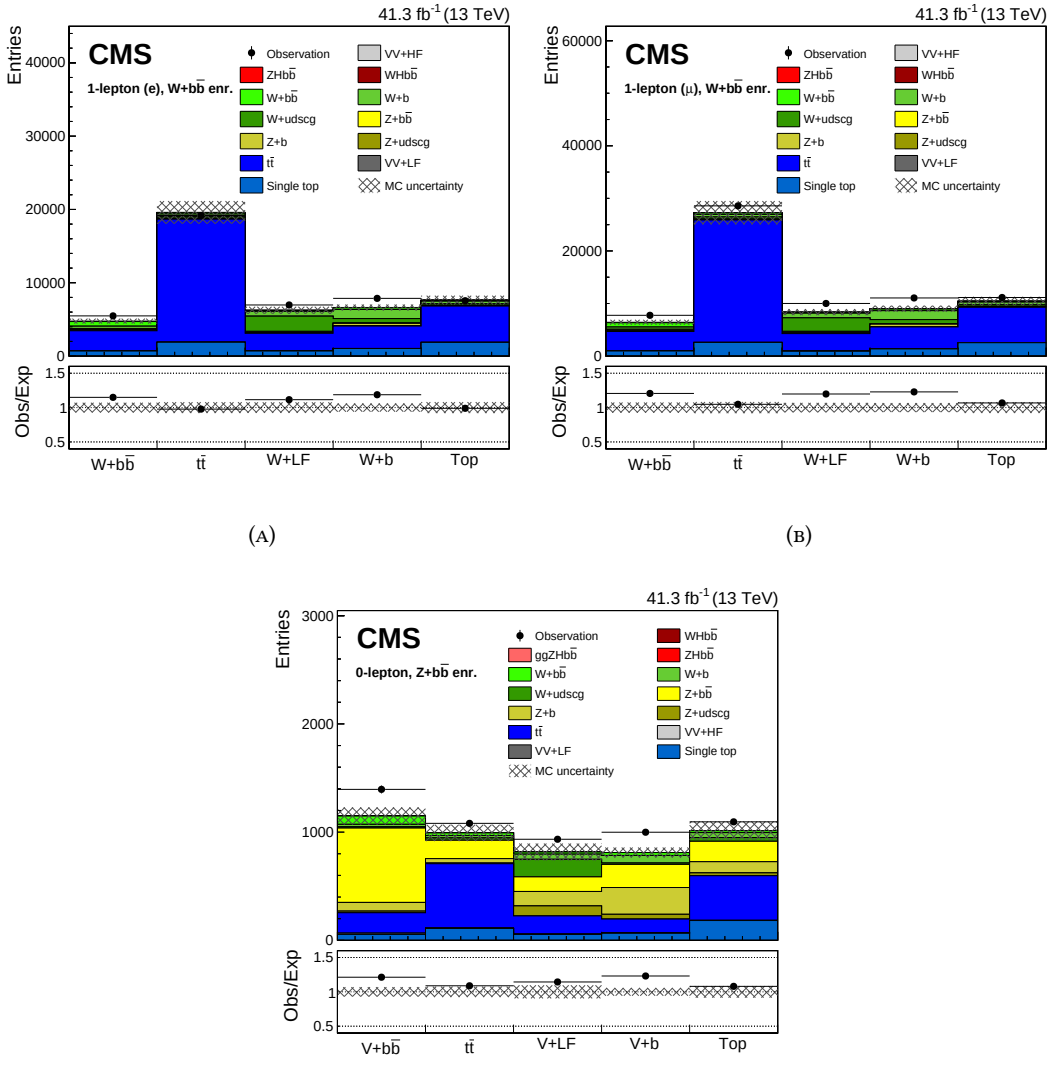
Each systematic uncertainty is described using a nuisance parameter θ_i. The probability density functions (pdfs) of the nuisance parameters, $\rho(\theta_i \,|\, \tilde{\theta}_i)$, are interpreted as posterior probabilities with $\tilde{\theta}_i$ as initial estimate of the nuisance parameter, and are expressed as

$$\rho(\theta_i \,|\, \tilde{\theta}_i) = p(\theta_i \,|\, \tilde{\theta}_i) \times \text{prior}(\theta_i).$$

The prior is kept flat, while the function $p(\tilde{\theta}_i \,|\, \theta_i)$ is used to build the global likelihood function. The expected signal s and background b models are described depending on the parameters θ_i.

The systematic uncertainties are modeled with log-normal pdfs, or propagating the uncertainty to obtain up and down variation histograms. In the latter case the histograms, together with the central one, are used to model $s(\theta_i)$ and $b(\theta_i)$ and a Gaussian pdf with $\mu = 0$ and $\sigma = 1$ models the nuisance. The floating background normalizations are parametrized with nuisance parameters with a flat $p_i(\tilde{\theta}_i \,|\, \theta_i)$.

The global likelihood function is expressed as the product of likelihoods of the data in each DNN histogram for signal regions and the chosen control region histogram, multiplied by the nuisance pdfs:



(A) (B)

(C)

FIGURE 6.17: Pre-fit distributions of the DNNHF variable for 2017 analysis in the 1-lepton channel for muon (A) and electron (B) control regions, and for the 0-lepton channel (C). The post-fit distributions are shown in figure 6.20 in the results section.

$$L(\,\text{data}\,|\,\mu,\theta\,) = \prod_{n=1}^{N} L_n(\,\text{data}\,|\,\mu,\theta\,) \times \prod_{i=1}^{N_\theta} p_i(\,\tilde{\theta}_i\,|\,\theta_i\,)\,,$$

where N is the number of fitted histograms N_θ is the number of nuisance parameters, and μ is a multiplying factor to the signal normalization. In each histogram, the individual likelihood is defined as:

$$L(\,\text{data}\,|\,\mu,\theta\,) = \prod_{k=1}^{N_B} \text{Poisson}(\,\text{n}_k\,|\,\mu \cdot s(\theta)_k + b(\theta)_k\,)\,,$$

where s_k and s_k are the expected signal and background yield in bin k, n_k is the observed yield, and N_B is the number of bins.

The significance of the observed excess of events in the fit is computed using the test statistic:

$$q_0 = \frac{L(\,\text{data}\,|\,0,\hat{\theta}_0)}{L(\,\text{data}\,|\,\hat{\mu},\hat{\theta}\,)} \qquad \text{with} \qquad \hat{\mu} \geq 0\,,$$

where $\hat{\theta}_0$ is the θ value maximizing the likelihood at the numerator in the background-only hypothesis, and $\hat{\theta}$ and $\hat{\mu}$ are the values maximizing the likelihood at the denominator. The constraint $\hat{\mu}$ is due to the fact that we are not interested in interpreting a downward fluctuation below the expected background.

More specifically, in this iteration of the VH($b\bar{b}$) analysis, the fit uses 28 distributions coming from independent regions in total. The signal regions are 7 (2-lepton μ and e both high and low-p_T, 1-lepton μ and e, 0-lepton). A binned DNN distribution with 15 bins is used for each signal region.
For each signal region, 3 control region plots are used, making it 21 control distributions and 28 in total. Among those, for 7 regions actual distributions are used: 5 bins in the 0 and 1-lepton V+HF control regions and 2 bins in the 2-leptons V+HF control regions. The remaining 14 control region distributions are used just to extract the normalization: a 1 bin histogram is used for all the V+LF and $t\bar{t}$ regions.
The nuisance parameters are 386 in total, when the MC statistical uncertainty is included in the count, otherwise about 250. Among those, 16 independent floating normalization parameters (or scale factors) are used: 4 $t\bar{t}$ scale factors (0-lepton, 1-lepton, 2-lepton high and low-p_T regions), 3 scale factor for the Z($\nu\nu$)+jets process (0b, 1b, 2b), 3 scale factor for the W($\ell\nu$)+jets process (0b, 1b, 2b) and 6 scale factor for the Z($\ell\ell$)+jets process (0b, 1b, 2b, both high and low-p_T).

Three fits are performed: one with the global VH($b\bar{b}$) signal strength as parameter of interest, one fitting separately the signal strengths in the 1,2,0-lepton channels and one where the WH and ZH signal strengths are extracted separately.

6.2.7 Systematic uncertainties

The systematic uncertainties used in the fit are propagated consistently to the DNN output in the signal regions and the variables used for the fit in the control regions. The systematic uncertainties can affect both the final significance and the signal strength uncertainty.

The systematic uncertainties are implemented as normalization variations of a process, shape variations of the fitted distribution, or both. If a shape uncertainty is used, the uncertainty is either applied to the relevant objects and propagated through the event selection and the DNN computation, or treated as an event weight when building the variations about the nominal distribution.

The full list of systematic uncertainties considered for the analysis is listed below.

- **Experimental uncertainties**
  - Luminosity: a normalization uncertainty of 2.3% is used, as measured for 2017 luminosity collected by CMS.
  - Pileup: a shape uncertainty is obtained by varying the minimum bias cross section used in the pileup reweighting applied to simulated events.
  - Lepton efficiency: the muon and electron trigger, reconstruction, and identification efficiencies are determined centrally for the CMS analyses using the tag-and-probe method in Z events. The corresponding scale factors are applied to the Monte Carlo samples. The systematic uncertainty on the scale factor is evaluated and a 2% normalization uncertainty is used.
  - $p_{\mathrm{T}}^{\mathrm{miss}}$+jets trigger: the variations of the parameters describing the trigger efficiency curve correction were used to assess a normalization uncertainty. A total uncertainty of 1% is estimated.
  - Jet energy scale: the energy scale for each jet is varied up and down within one standard deviation, individually for each source of uncertainty as recommended centrally for CMS data analysis. In total, 27 sources of uncertainty are considered. The uncertainties are propagated through the selection and the DNN computation, resulting in a shape and normalization uncertainty.
  - Jet energy resolution: the jet energy is varied for each b jet used to build the Higgs boson candidates using the 10% resolution scale factor measured in the regression validation (see chapter 5). A 10% uncertainty about the nominal smearing is used. The uncertainty propagation is analogous to the jet energy scale one.
  - Jet energy resolution for additional jets: jets which are not b-tagged, thus not needing the regression, but are used as FSR jets or in the event selection, require the standard jet energy resolution smearing, derived centrally for the CMS collaboration. The corresponding uncertainties are used. The uncertainty propagation is analogous to the jet energy scale one.
  - Unclustered $p_{\mathrm{T}}^{\mathrm{miss}}$: the uncertainty on the $p_{\mathrm{T}}^{\mathrm{miss}}$ is mostly covered by the jet uncertainties. For the unclustered component a 3% normalization uncertainty is used.
  - Jet b-tagging: the b tagging scale factors computed centrally for the CMS collaboration have variations implemented via jet weights, which are multiplied to get event weights. The "up" and "down" event weights are applied to remake the DNN distribution, and the resulting DNN histogram is used in the fit as a shape and normalization uncertainty. The systematic uncertainties for the b-tagging, are decorrelated in five p_{T} and three η bins.
  - $\Delta\eta$(jj) Monte Carlo reweighting: the V+jets samples are corrected as a function of $\Delta\eta$(jj). The entire reweighting applied to LO simulation is used as shape systematic uncertainty.

  - W boson and $t\bar{t}$ transverse momentum reweighting: the systematic uncertainties on the p_T^V corrections are taken from the uncertainties on the fitted correction and implemented as a shape systematic uncertainty.

The jet energy scale and jet energy resolution are the only uncertainties that affect the signal region acceptance and the mutual signal and control region acceptances, besides the shapes and normalizations of the fitted distributions, and are propagated through the analysis keeping the selections consistent with the up and down variations. Other experimental uncertainties affecting the shape of the distributions (pileup, jet b-tagging, $\Delta\eta$(jj) reweighting, $t\bar{t}$ transverse momentum reweighting) are implemented and propagated using event weights.

The latter two uncertainties are placed by choice into the "experimental" group as they are meant to correct for effects known from experimental measurements. They are applied to the relevant samples only, while the rest of the systematics is applied to both signal and background simulation.

- **Theoretical uncertainties**
  - $H \to b\bar{b}$ decay branching ratio: three independent normalization uncertainties, as recommended in [17], are used.
  - Signal cross section: the total signal cross section has been calculated at NNLO accuracy, and the normalization uncertainty is about 2% for the WH and qq→ZH processes, and 25% for gg→ZH process.
  - Theoretical signal p_T spectrum. A mismodeling of the p_T spectrum of the vector boson could lead to acceptance differences. The calculations available (electroweak and QCD) are applied as a central value correction. The normalization uncertainty estimated for the electroweak correction is 2%, while the QCD correction bring an uncertainty of 5%.
  - Background Estimate: an uncertainty of 15% is assumed for single top and diboson processes (approximately the uncertainty on the measured cross sections). The other background models are verified in data, with the associated uncertainties from the control regions.
  - PDF uncertainties: the uncertainty is encoded in a set of PDF replicas. For each process, the RMS of all the variations is checked in each bin of the DNN distribution and the largest variation among the bins is used as normalization uncertainty for the process.
  - Perturbative QCD scale variations: the perturbative QCD renormalization and factorization scale variations of 1/2 and 2 times the nominal values are considered separately for each process and taken as uncorrelated sources of systematic uncertainties, affecting both shape and normalization.

Among the theoretical uncertainties, the only ones implemented as shape and normalization uncertainties are the perturbative QCD scale variations. All the other systematic uncertainties are implemented as normalization only uncertainties.

- **Monte Carlo simulation size**
  - The shape of the DNN is allowed to vary within the bin-by-bin statistical uncertainties from the MC samples. The bin-by-bin uncertainties are used together with the normalization uncertainty for the total sample size, which is treated independently. The implementation follows the Barlow-Beeston method [134].

6.2.8 VH($b\bar{b}$) results with 2017 data

For the 2017 data, the observed significance is 3.3σ above the background-only hypothesis, while an excess of 3.1σ is expected for the SM Higgs boson. The corresponding measured signal strength is $\mu = 1.08 \pm 0.34$, where the uncertainty combines both statistical and systematic components. The largest sources of uncertainty and their observed impact on μ from the fit are listed in table 6.14. The dominant sources of uncertainty, other than the purely statistical one, which depends only on the dataset size, arise from the background normalizations, the simulated sample size, the b-tagging efficiency and misidentification rates, and the modeling of the V+jets background.

The distributions of the DNN in the 2-lepton signal region are shown in figure 6.18. The high-p_T signal region DNNs are shown in (A) for the dimuon channel and in (B) for the dielectron channel. The corresponding discriminators in the low-p_T signal region are shown in (C) and (D). The DNN distributions for the 1-lepton and 0-lepton channels are shown in figure 6.19.

The scale factors obtained for the full fit (signal region + control regions) for the 2017 analysis are reported in table 6.13. Post-fit plots for the DNNHF distributions are shown in figure 6.20. The post-fit distribution of the variable used in the Z+HF control region for the 2-lepton channel, the lower b-tag score among the two jets, in 2 bins, is shown also for the 2-lepton(ee) high-p_T^V selection.

Process / Channel	0-lepton	1-lepton	2-lepton low-p_T	2-lepton high-p_T
W+jets, 0b	1.04 ± 0.07	1.04 ± 0.07	–	–
W+jets, 1b	2.09 ± 0.16	2.09 ± 0.16	–	–
W+jets, 2b	1.74 ± 0.21	1.74 ± 0.21	–	–
Z+jets, 0b	0.95 ± 0.09	–	0.89 ± 0.06	0.81 ± 0.05
Z+jets, 1b	1.02 ± 0.17	–	0.94 ± 0.12	1.17 ± 0.10
Z+jets 2b	1.20 ± 0.11	–	0.81 ± 0.07	0.88 ± 0.08
$t\bar{t}$	0.99 ± 0.07	0.93 ± 0.07	0.89 ± 0.07	0.91 ± 0.07

TABLE 6.13: Data/MC scale factors for the 2017 analysis in the 0-, 1- and 2-lepton channels from SR+CRs fit. The errors include both statistical and systematic uncertainties. Compatible fitted values are obtained from the CR-only fit.

The results of the fit are summarized in figure 6.21. The left plot (A) shows the distribution of events of all channels as a function of the post-fit value of $\log_{10}$(S/B) for the 2017 data. The signal (S) and background (B) yields are determined from the DNNs used in each analysis (figures 6.18, 6.19). The lower panel shows the ratio between the data and the background expectation. As the ratio departs from unity, the background-only hypothesis is disfavored, while the prediction for the signal is compatible with the observation.

The right plot shows the signal strength by process (WH) or (ZH), and by analysis channel: 2, 1 or 0-leptons. The per-channel signal strengths are compared to the global signal strength (green band). The per-channel signal strengths are found to be compatible with the global signal strength fit with a probability of 96.9%.

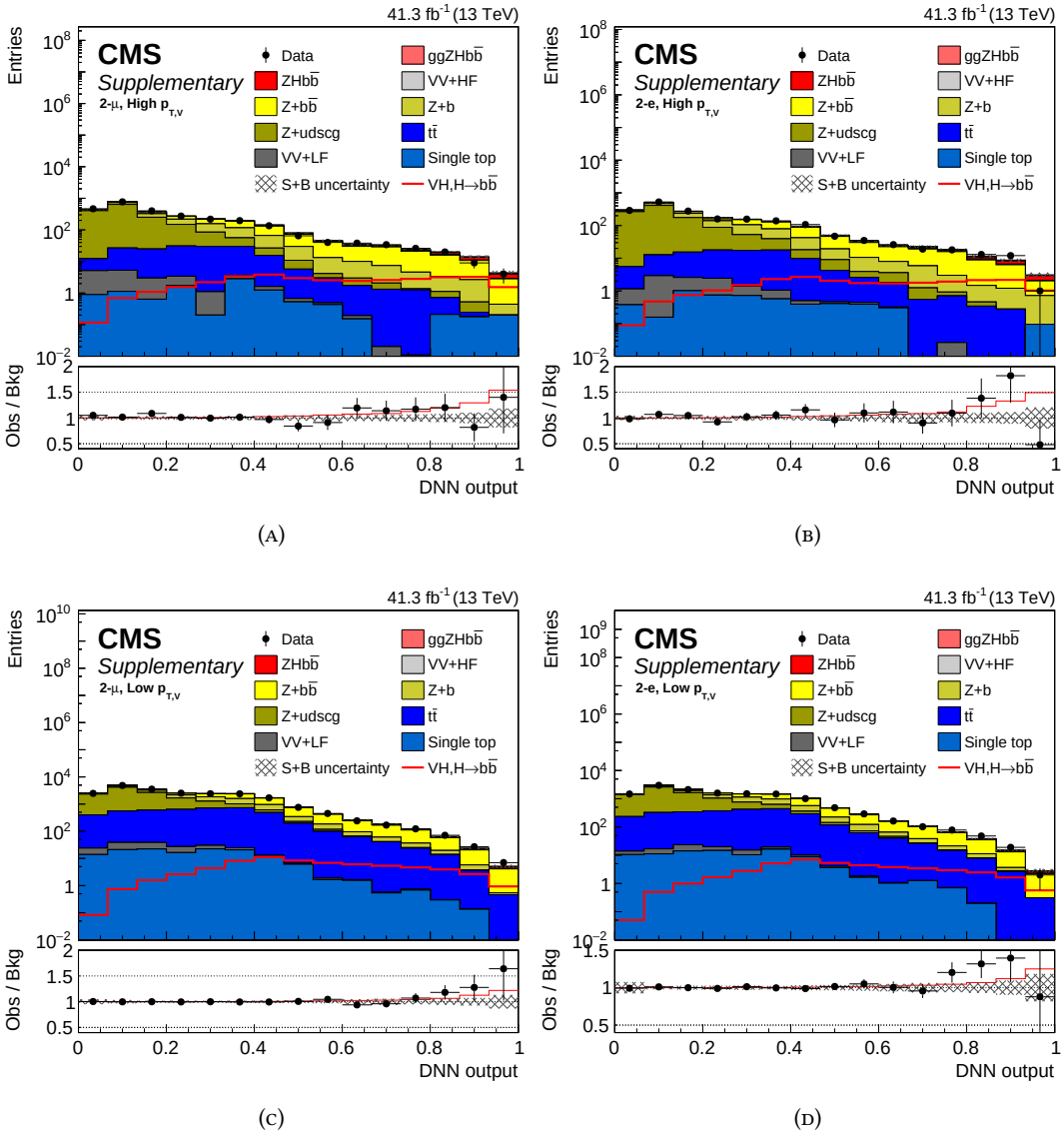
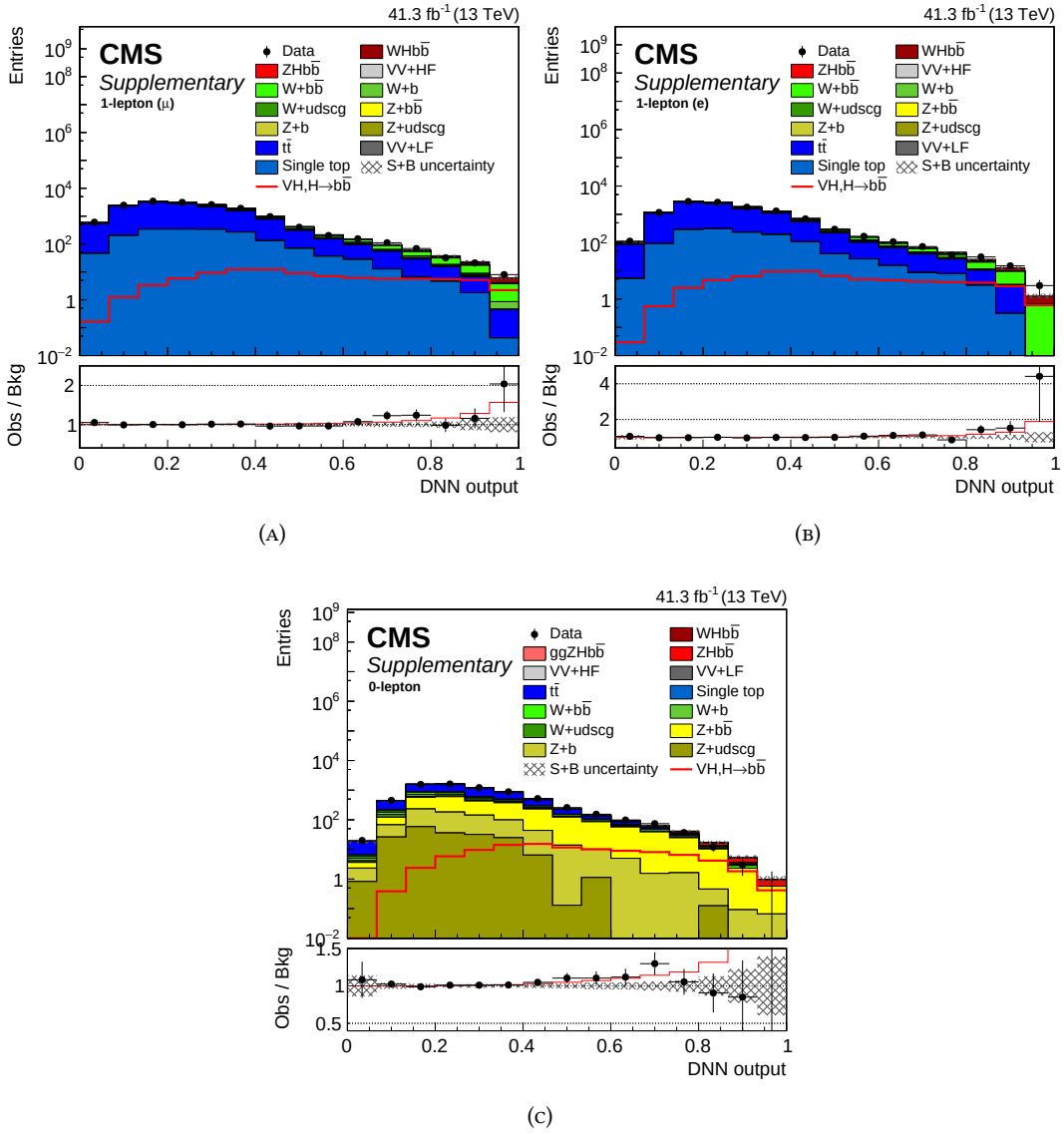


FIGURE 6.18: Post-fit distributions of multivariate discriminator output channels for 2017 analysis, after all signal region pre-selection criteria have been applied. 2-lepton muon (A) and electron (B) channel for high p_T^V region, in the second row the low p_T^V is shown.

A more detailed uncertainty breakdown is shown in figure 6.22. The two rows show the pull of the nuisance parameters, $(\theta_{postfit} - \theta_{initial})/\sigma_{initial}$ or the ratio with respect to the initial value in case the parameters are normalization parameters (scale factors), and the impact of the nuisance on the parameter of interest. The impact plot is reported for the global signal strength fit. Several checks were performed before the fit: first, a fit was performed using an Asimov dataset, then the data were used in the fit only for control regions (21 regions). Finally, the full signal+control region fit was performed (28 regions). The nuisances showed a similar impact in all the fits. The goodness of fit was also checked using the Kolmogorov-Smirnov test in 14 regions, the signal and the V+HF ones, yielding a p-value between 0.05 and 0.9 for all the regions except for one with p-value 0.01.



(A) (B)

(C)

FIGURE 6.19: Post-fit distributions of multivariate discriminator output channels for 2017 analysis, after all signal region pre-selection criteria have been applied. Third row: 1-lepton muon (A) and electron (B) channel, 0-lepton channel (C).

6.2.9 VZ($b\bar{b}$) cross-check

The VZ diboson process with $Z \to b\bar{b}$, having an identical final state as the VH process with $H(b\bar{b})$ process, except for the dijet mass, and a dijet mass very close to signal one, is used to validate the analysis method. The VZ($b\bar{b}$) cross section is about $\sim$ 5-10 times larger than the Higgs one, therefore the analysis, even if tuned for the VH($b\bar{b}$) signal, is sensitive to this process.
The VZ($b\bar{b}$) cross check has been performed also in the previous versions of the analysis by the CMS collaboration [119, 103, 35].

To extract this diboson signal, the DNNs are trained using the simulated samples for this process as signal. All the other processes, including VH production, are treated as background. The only modification made to the analysis is the requirement that the signal region is in the dijet mass window [60, 160] GeV for all channels.

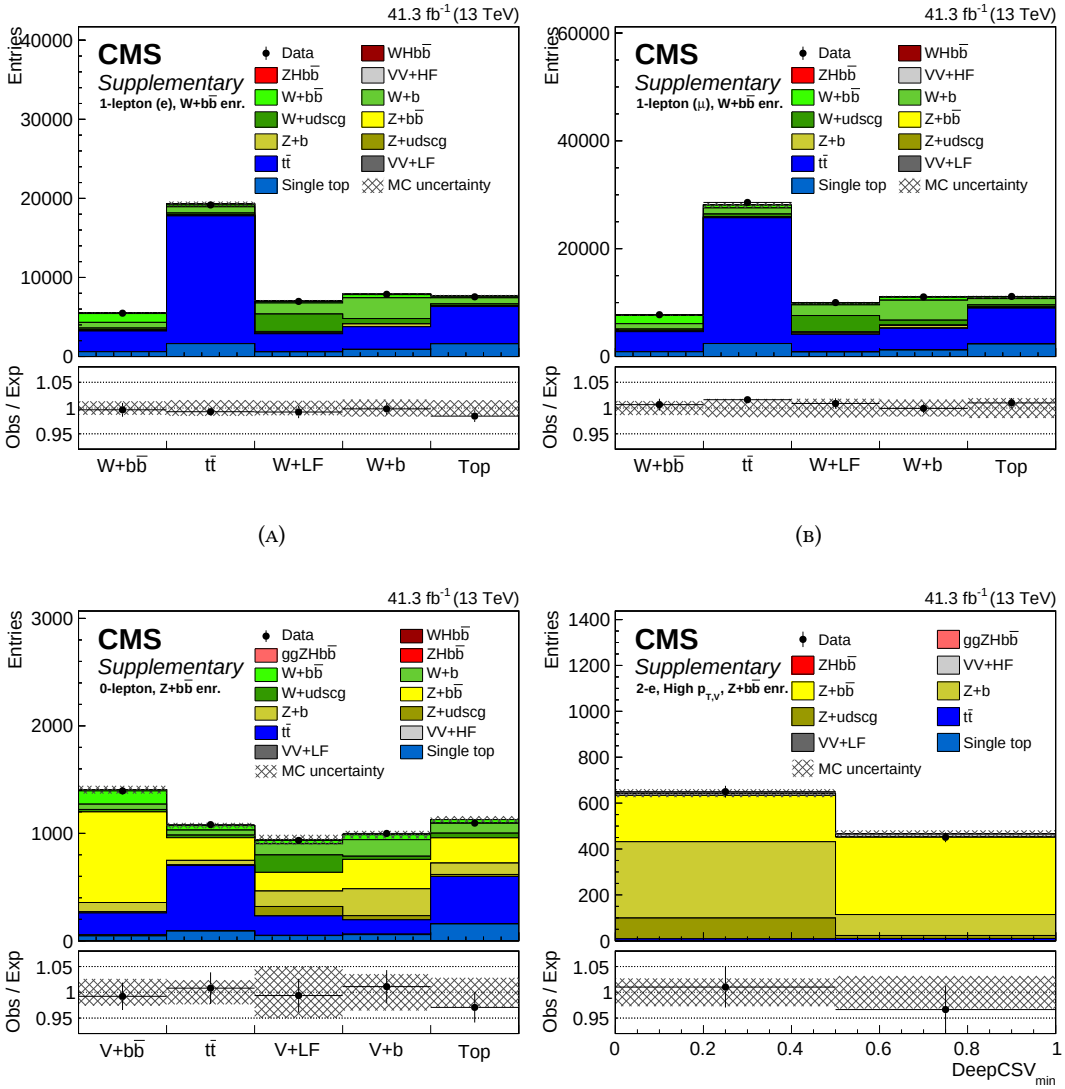
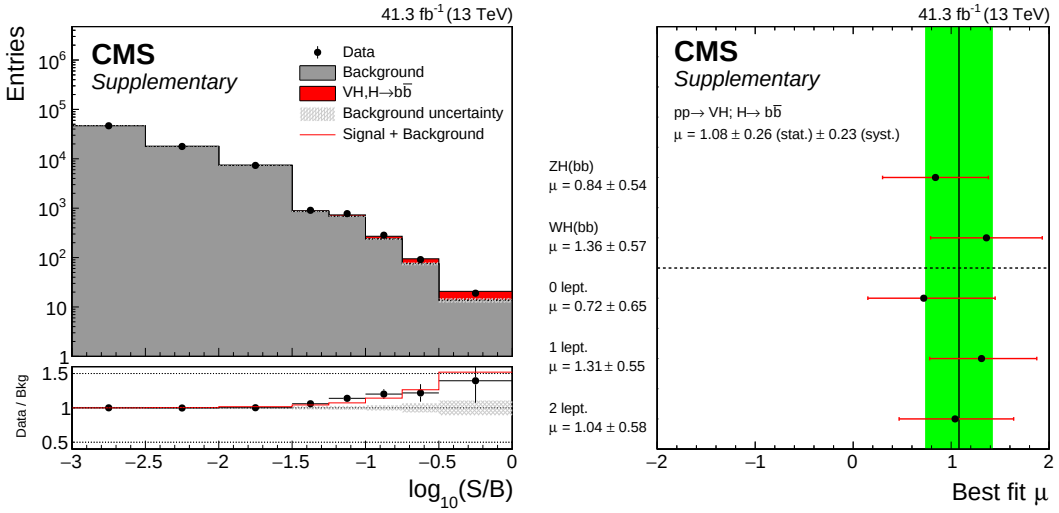


FIGURE 6.20: Post-fit distributions of the DNNHF variable for 2017 analysis in the 1-lepton channel for muon (A) and electron (B) control regions, and for the 0-lepton channel (C). The post-fit distribution of the lower b-tag score among the two jets, in 2 bins, is also shown for the 2-lepton(*ee*) high-$p_{\mathrm{T}}^{\mathrm{V}}$ selection in (D).

For the 2017 dataset, the combined WZ and ZZ production processes have an observed significance of 5.2σ compared to the background-only hypothesis, with an expected significance of 5.0σ. The observed signal strength of the VZ($b\bar{b}$) process is $\mu = 1.05 \pm 0.22$.

Figure 6.23 summarizes the diboson results obtained with 2017 data. The left plot (A) shows the distribution of events of all channels sorted according to the post-fit value of $\log_{10}$(S/B) for the 2017 data. The lower panel exhibits a good compatibility between the data and the expected total event yield. The right plot shows the signal strength by analysis channel: 2, 1 or 0-leptons. The per-channel signal strengths are compared to the global signal strength (green band) and are found to be compatible with the single signal strength fit with a probability of 64.2%.



(A) (B)

FIGURE 6.21: Distributions of signal, background, and data event yields sorted into bins of similar signal-to-background ratio, as given by the result of the fit to their corresponding multivariate discriminant (A). All events in the 2017 VH($b\bar{b}$) signal regions are included. The red histogram indicates the Higgs boson signal contribution, while the grey histogram is the sum of all background yields. The best-fit signal strength and uncertainty per-channel and for the WH and ZH processes, extracted from a simultaneous fit of all channels for the 2017 analysis (B).

Uncertainty source	$\Delta\mu$
Statistical	**+0.26 -0.26**
Normalization of backgrounds	+0.12 -0.12
Experimental	**+0.16 -0.15**
b-tagging efficiency and misid.	+0.09 -0.08
V+jets modeling	+0.08 -0.07
Jet energy scale and resolution	+0.05 -0.05
Lepton identification	+0.02 -0.01
Luminosity	+0.03 -0.03
Other experimental uncertainties	+0.06 -0.05
MC sample size	**+0.12 -0.12**
Theory	**+0.11 -0.09**
Background modeling	+0.08 -0.08
Signal modeling	+0.07 -0.04
Total	**+0.35 -0.33**

TABLE 6.14: Major sources of uncertainty in the measurement of the signal strength μ, and their observed impact ($\Delta\mu$) from a fit to the 2017 data set, are listed. The total uncertainty is separated into four components: statistical (including data yields), experimental, MC sample size, and theory. Detailed decompositions of the statistical, experimental, and theory components are specified. The impact of each uncertainty is evaluated considering only that source. Because of correlations in the combined fit between nuisance parameters in different sources, the sum in quadrature for each source does not in general equal the total uncertainty of each component.

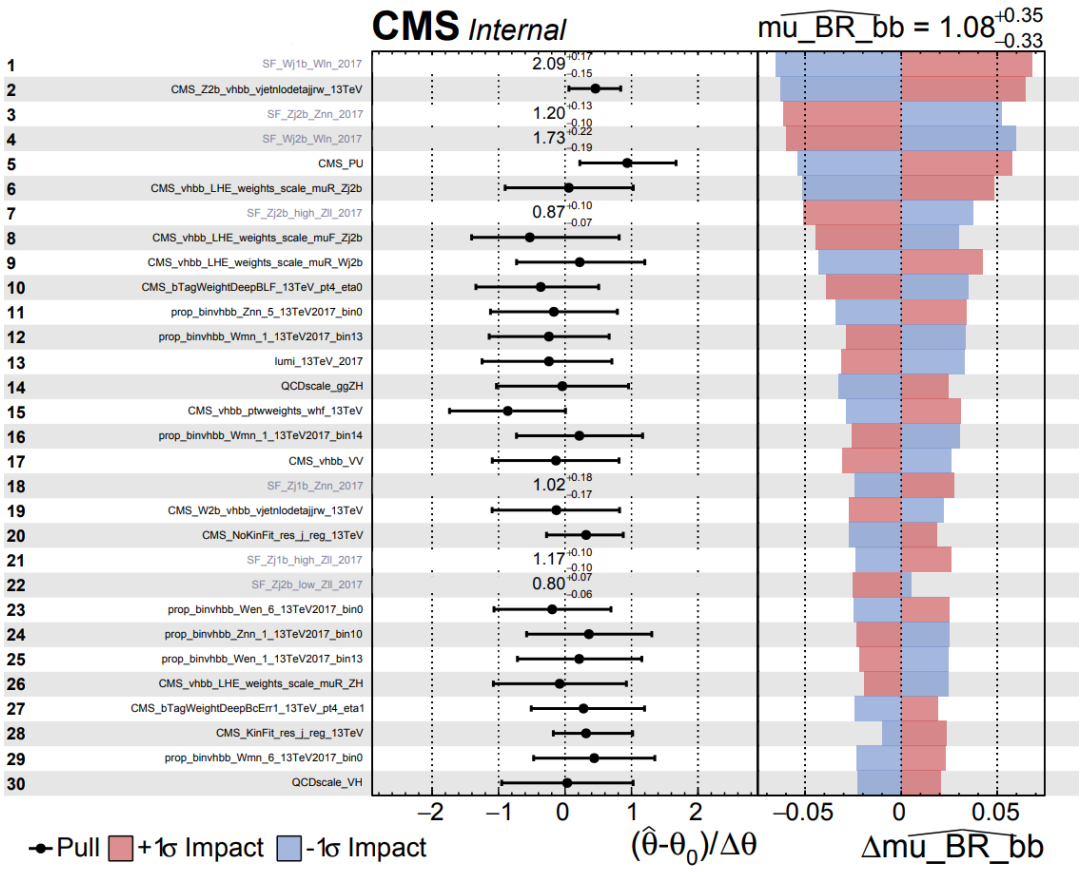
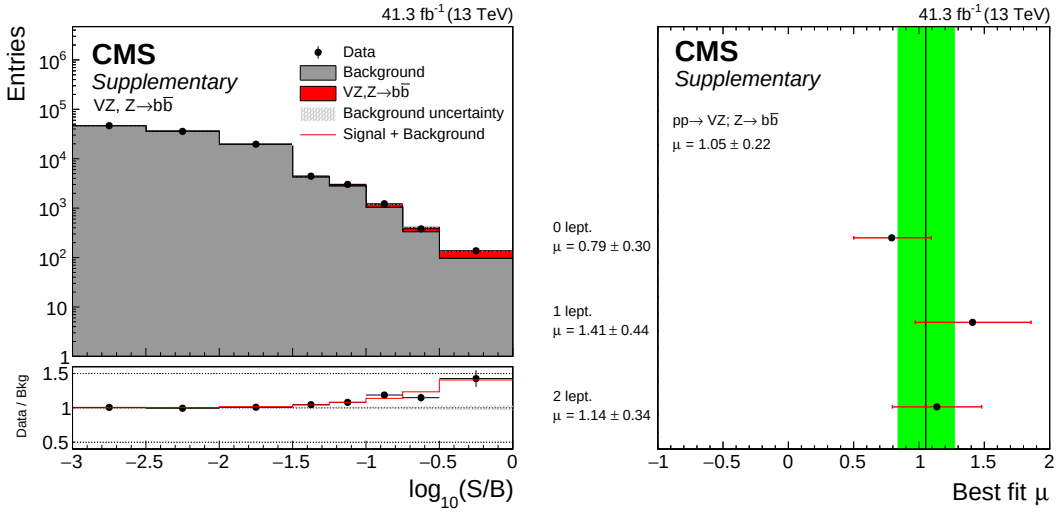


FIGURE 6.22: Post-fit impact plot: the left row shows the pull of the nuisance parameters, $(\theta_{postfit} - \theta_{initial})/\sigma_{initial}$, while the right one shows the impact of the nuisances on the parameter of interest. The nuisances are ranked by the impact: only the 30 most impacting nuisances are reported here. Among the nuisances that have the largest impact, one can find the ones affecting the backgrounds in the most sensitive bins of the DNN (V+HF scale factors, LO to NLO reweighting, V+HF theoretical uncertainties).



(A) (B)

FIGURE 6.23: Distributions of signal, background, and data event yields sorted into bins of similar signal-to-background ratio, as given by the result of the fit to their corresponding multivariate discriminant. All events in the 2017 VZ($b\bar{b}$) signal regions are included. The red histogram indicates the VZ($b\bar{b}$) signal contribution, while the grey histogram is the sum of all background yields. The best-fit and uncertainty per-channel signal strengths extracted from a simultaneous fit of all channels for the 2017 analysis in the VZ($b\bar{b}$) validation analysis.

6.3 Combined $H \to b\bar{b}$ results

The VH, H($b\bar{b}$) analysis results were combined with previous results for the $H \to b\bar{b}$ decay both in the same production mode and in other production modes. The results came from the analysis of 2016 and Run 1 data.

6.3.1 VH, H($b\bar{b}$) combination

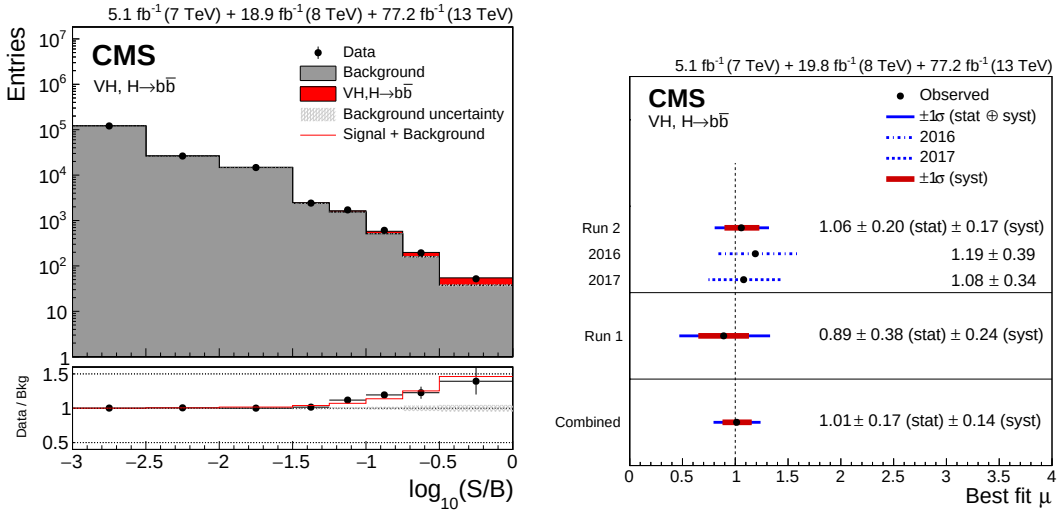
The result was first combined with the previous Run 2 VH($b\bar{b}$) result, obtained using 35.9 fb^{-1} of data collected in 2016. The combined result yields an observed signal significance of 4.4σ, with 4.2σ expected, and a signal strength of $\mu = 1.06 \pm 0.26$. All systematic uncertainties are assumed to be uncorrelated in this fit, except for theory uncertainties and the dominant uncertainties in the measurement of the jet energy scale, which are assumed to be fully correlated.

The VH($b\bar{b}$) results from Run 2 are combined with the results of the corresponding CMS analysis of the Run 1 data using collisions at $\sqrt{s}$ = 7 and 8 TeV, with data samples corresponding to integrated luminosities of up to 5.1 and 18.9 fb^{-1}, respectively. Systematic uncertainties in this fit are assumed to be uncorrelated for separate collision energies, except for the theory uncertainties.

The combination gives an observed signal significance of 4.8σ, with 4.9σ expected. The measured signal strength is $\mu = 1.01 \pm 0.22 = 1.01 \pm 0.17$ (stat.) $\pm$ 0.09 (exp.) $\pm$ 0.06 (MC stat.) $\pm$ 0.08 (theory).

The combined results for VH($b\bar{b}$) are summarized in figure 6.24. The left plot (A) shows again the distribution of events of all channels sorted according to the post-fit value of $\log_{10}$(S/B). The inputs of the plot are either the DNN or BDTs for each channel depending on the technique used in each year.

The signal strengths VH production, with $H \to b\bar{b}$ are shown in the right plot (B). The Run 1 and Run 2 (2016+2017) are reported separately and in combination. The significances and signal strengths are also reported in table 6.15. The 2017 results are quoted both globally and split by channel, together with the combined ones.



(A) (B)

FIGURE 6.24: Distributions of signal, background, and data event yields sorted into bins of similar signal-to-background ratio, as given by the DNN fit result (A). The events in the VH($b\bar{b}$) signal regions of the combined Run 1 and Run 2 data sets are included. The red histogram shows the Higgs boson signal contribution, while the grey histogram is the sum of all background yields. The bottom panel shows the data to background ratio, with the total uncertainty indicated by the grey band. The red line shows the sum of signal plus background divided by the background. Best-fit value of the signal strength μ, for the fit of all VH($b\bar{b}$) channels in the Run 1 and Run 2 data sets (B). The results of the 2016 and 2017 measurements, the Run 2 combination, and the Run 1 result are also shown. The error bars indicate the 1σ systematic (red) and 1σ total (blue) uncertainties, and the vertical line the SM expectation.

Data set	Expected significance	Observed significance	Signal strength
2017			
0-lepton	1.9σ	1.3σ	0.73 ± 0.65
1-lepton	1.8σ	2.6σ	1.32 ± 0.55
2-lepton	1.9σ	1.9σ	1.05 ± 0.59
Combined	3.1σ	3.3σ	1.08 ± 0.34
Run 2	4.2σ	4.4σ	1.06 ± 0.26
Run 1 + Run 2	4.9σ	4.8σ	1.01 ± 0.23

TABLE 6.15: Expected and observed significances, in σ, and observed signal strengths for the VH($b\bar{b}$). The results are shown for the 2017 data, the combined Run 2 (2016 and 2017) data, and for the combination of the Run 1 and Run 2 data sets. For the 2017 analysis, the results are shown also separately by channel and for a combined simultaneous fit to all channels. (m_{H} = 125.09 GeV for all the results).

Uncertainty source	$\Delta\mu$ (2017)	$\Delta\mu$ (Run2)	$\Delta\mu$ (Run1 + Run2)
Statistical	**+0.26 -0.26**	**+ 0.20 -0.20**	**+0.18 -0.17**
Normalization of backgrounds	+0.12 -0.12	+0.10 -0.09	+0.08 -0.07
Experimental	**+0.16 -0.15**	**+0.11 -0.10**	**+0.10 -0.09**
b-tagging efficiency and misid.	+0.09 -0.08	+0.07 -0.06	+0.05 -0.05
Jet energy scale and resolution	+0.05 -0.05	+0.03 -0.03	+0.03 -0.03
Lepton identification	+0.02 -0.01	+0.01 -0.01	+0.01 -0.01
Luminosity	+0.03 -0.03	+0.04 -0.02	+0.03 -0.02
Other experimental uncertainties	+0.10 -0.09	+0.05 -0.05	+0.06 -0.05
MC sample size	**+0.12 -0.12**	**+0.08 -0.08**	**+0.06 -0.06**
Theory	**+0.11 -0.09**	**+0.11 -0.10**	**+0.09 -0.08**
Background modeling	+0.08 -0.08	+0.09 -0.09	+0.07 -0.07
Signal modeling	+0.07 -0.04	+0.07 -0.05	+0.05 -0.03
Total	**+0.35 -0.33**	**+0.26 -0.25**	**+0.23 -0.22**

TABLE 6.16: Uncertainty breakdown for the measurement of the signal strength μ. The sources of uncertainty as in table 6.14. The Run 2 and Run1 + Run2 combination fits are shown (together with the 2017 one, reported also here for comparison).

Invariant mass plot

The results obtained with a fit of the DNN distributions are accompanied by a mass-independent analysis, where both the VZ($b\bar{b}$) and the VH($b\bar{b}$) excesses are visible over the backgrounds. As in the VZ analysis, the signal region is defined to be in the interval [60, 160] GeV in m(jj).

The mass-independent version of the analysis was performed using the same DNN discriminator of the main analysis for 2017 data, but fixing the values of some of the mass-correlated input features when running the inference. For the 2016 dataset, dedicated DNNs were trained with the same procedure as for 2017 and the same variables were fixed.

The variables whose values were fixed at their mean value in the background are listed by channel in table 6.17.

0-lepton	$p_{\mathrm{T}}^{\mathrm{j1}}$, $p_{\mathrm{T}}^{\mathrm{j2}}$, m(jj), p_{T}(jj), $\Delta\eta$(jj), $p_{\mathrm{T}}^{\mathrm{V}}$
1-lepton	$p_{\mathrm{T}}^{\mathrm{j1}}$, $p_{\mathrm{T}}^{\mathrm{j2}}$, m(jj), $\Delta\eta$(jj)
2-lepton	$p_{\mathrm{T}}^{\mathrm{j1}}$, $p_{\mathrm{T}}^{\mathrm{j2}}$, m(jj) (kin. fit), $\sigma_{m(\mathrm{jj})}$ (kin. fit), $\Delta\eta$(jj)

Table 6.17: List of input variables fixed at their mean value in the mass-decorrelated evaluation

Fixing a few of the variables highly correlated to the dijet mass, including the dijet mass itself, has the effect of moderately reducing the sensitivity of the analysis. However, the DNN distribution is not highly correlated to the dijet mass, therefore it is possible to use the DNN to select events and then fit the VH($b\bar{b}$) and VZ($b\bar{b}$) signals in the dijet mass distribution. The effect of such a procedure is shown in the sketch in figure 6.25: the top row shows the effect on the dijet mass distribution of a cut on the DNN; the bottom row shows the effect of the same cut on the DNN evaluated with fixed values of the dijet mass correlated inputs.

After the DNN evaluation, the events are categorized into four bins of increasing S/B ratio according to the score of their corresponding discriminant. The resulting four m(jj) distributions in each data set are fit together with the same distributions used in the control regions, to extract signal and background yields. The observed (expected) significance of this fit is 2.7 (3.0) σ, with a signal strength $\mu = 0.91^{+035}_{-0.34}$, which makes it less sensitive compared to the 4.4 (4.2) σ of the Run 2 results.
The fitted m(jj) distributions are combined and weighted by S/(S+B), where S and B are computed from the Higgs boson signal yield and the sum of all background yields for each category considering their fitted normalizations, respectively.

The combined m(jj) distribution, is shown in figure 6.26, before (A) and after (B) background subtraction. The VZ($b\bar{b}$) and the VH($b\bar{b}$) contributions are separately visible in grey and red, respectively.

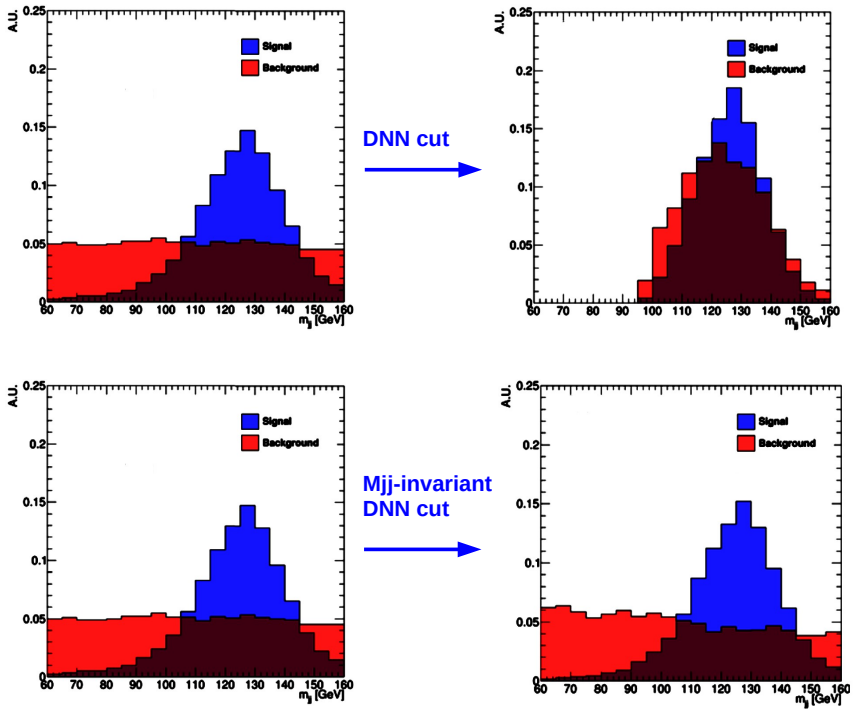
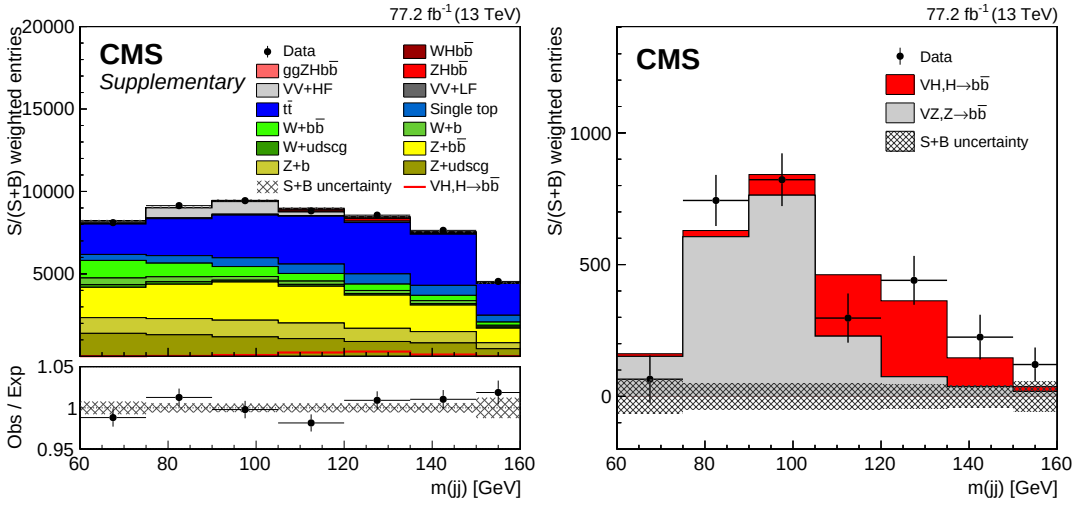


FIGURE 6.25: Sketch of the expected behavior of the m(jj) distribution after applying a cut on the standard DNN (top row). In case the DNN is decorrelated from the m(jj) little background sculpting is expected (bottom row). The signal and background histograms, with the background including the VZ($b\bar{b}$) process, are normalized to unity.



(A) (B)

FIGURE 6.26: Distribution of m(jj) for events weighted by S/(S+B) in all channels combining the 2016 and 2017 data sets (A). The weights are derived from a fit to the m(jj) distribution. Distribution of m(jj) after background subtraction (B): the data (points) and the fitted VH signal (red) and VZ background (grey) distributions are shown, with all other fitted background processes subtracted. The error bar for each bin represents the pre-subtraction 1σ statistical uncertainty on the data, while the grey hatching indicates the 1σ total uncertainty on the signal and all background components.

6.3.2 H($b\bar{b}$) Observation

The global VH($b\bar{b}$) combination has an expected significance of 4.9σ and a corresponding observed significance of 4.8σ. In order to reach the 5σ threshold a global combination of H($b\bar{b}$) measurements using CMS data was performed.

The included analyses targeted the following production processes: VH (reported above), gluon fusion (ggF) [118], vector boson fusion (VBF) [102], and associated production with top quarks ($t\bar{t}H$) [32, 135, 136]. All the available channels are used. These analyses use data collected at 7, 8, and 13 TeV, depending on the process. In this fit, most sources of systematic uncertainty are treated as uncorrelated. The dominant jet energy scale uncertainties are treated as correlated between processes at the same collision energy, while the theory uncertainties are correlated between all processes and data sets.

The observed (expected) signal significance is 5.6 (5.5) σ, and the measured signal strength is $\mu = 1.04 \pm 0.20$. In addition to the overall signal strength for the H→$b\bar{b}$ decay, the signal strengths for the individual production processes are also determined in this combination, where contributions from a single production process to multiple channels are properly accounted for in the fit. All the results are summarized in figure 6.27.

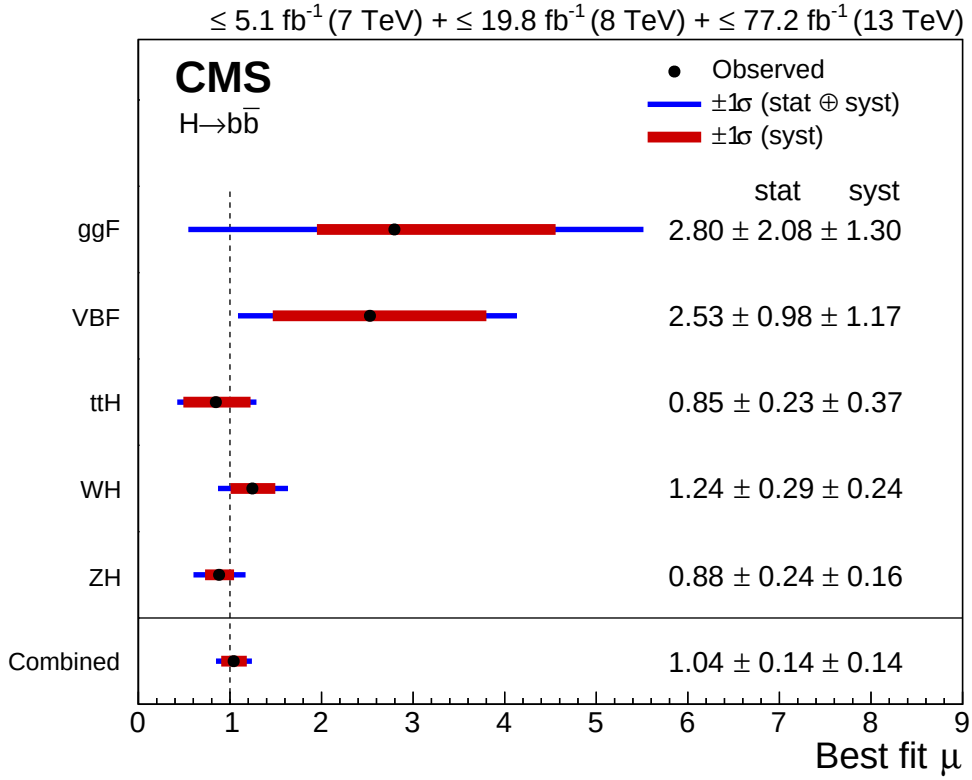


FIGURE 6.27: Best-fit value of the H → $b\bar{b}$ signal strength with 1σ systematic (red) and total (blue) uncertainties by production mode, together with the combination. The vertical line shows the SM expectation. The results are extracted from a single fit combining all the analyses (at m_H = 125.09 GeV).

Chapter 7

Conclusions and outlook

The analysis presented establishes the observation for the Higgs decay to bottom quarks with CMS data. An observation of the $b\bar{b}$ decay of the Higgs boson by the ATLAS Collaboration [36] was published at the same time. The two analyses use similar strategies, but were performed independently. Yet, both analyses reach about the same sensitivity and the results are in agreement.

This observation closes a big chapter in the experimental history of the Higgs boson. The 5 most sensitive decay channels at the LHC are finally all firmly established. The final states with bosons were the ones contributing to the discovery and first to be observed. Fermionic final states followed in Run 2: the decay to tau pairs was observed with 2016 data and at the same time the coupling to top quarks was measured. Adding the 2017 data, the observation of $H(b\bar{b})$ completed the picture. A significant amount of data was collected in 2018, but other final states cannot be measured with similar precision at this point.

The near future of $H\rightarrow b\bar{b}$ measurements includes the transition to differential measurements. This is already possible thanks to full Run 2 luminosity, so a Run 2 analysis with differential measurements is now being developed. The ATLAS collaboration already published an analysis using the same integrated luminosity as the "Observation" one, but focusing on differential measurements [137]. The STXS framework, briefly described in paragraph 7.1, is used to define the phase-space categories where the signal is extracted.
The VH, $H\rightarrow b\bar{b}$ analysis is also limited by experimental systematic uncertainties, so new strategies are being tested in order to reduce the systematic uncertainties.

Rarer decay modes are also a current target of the LHC experiments, in order to access the coupling to the second generation of fermions: the $H\rightarrow \mu\mu$ decay is the most promising decay channel for this purpose. Machine learning techniques used for $H\rightarrow b\bar{b}$, and in particular Deep Learning, turn out to be useful also for this analysis. During the last year of my Ph.D., I contributed to the VBF production dedicated analysis with the optimization of a DNN. A brief description of the analysis and of the expected results are given in paragraph 7.2, while the details of the training procedure are in appendix D.

Deep Learning was a fundamental tool in my work, both in the $VH(b\bar{b})$ analysis and to develop the DeepVertex b-tagger and the b jet energy regression. It is desirable in the coming years to deploy the combined tagger of DeepVertex and DeepJet, and adopt it at the analysis level. Deep Learning techniques will likely also be used for other tasks, among which the reconstruction the secondary vertex itself. Parameter sharing techniques, as LSTM nodes or conv 1×1 layers, turned out to be useful for processing the jet data, and can be applied to other tasks. On the other hand, new techniques specific to other kinds of data formats will be developed, making Deep Learning a central tool for LHC analysis and data reconstruction.

7.1 H→bb̄: Simplified template cross sections framework

The next Higgs analyses at the LHC, at least in the main channels, are targeting differential measurements. A common framework has been developed, and is now being adopted by most of the analysis groups. This is the "Simplified Template Cross Sections" (STXS) framework: it has been developed with the goal of minimizing the dependency form theory, but to maximize the sensitivity and be flexible for different interpretations. The differential measurements are performed in fiducial phase-space regions motivated by theory, but independently defined based on the reconstructed objects.

A schematic overview of the simplified template cross section framework is shown in figure 7.1.

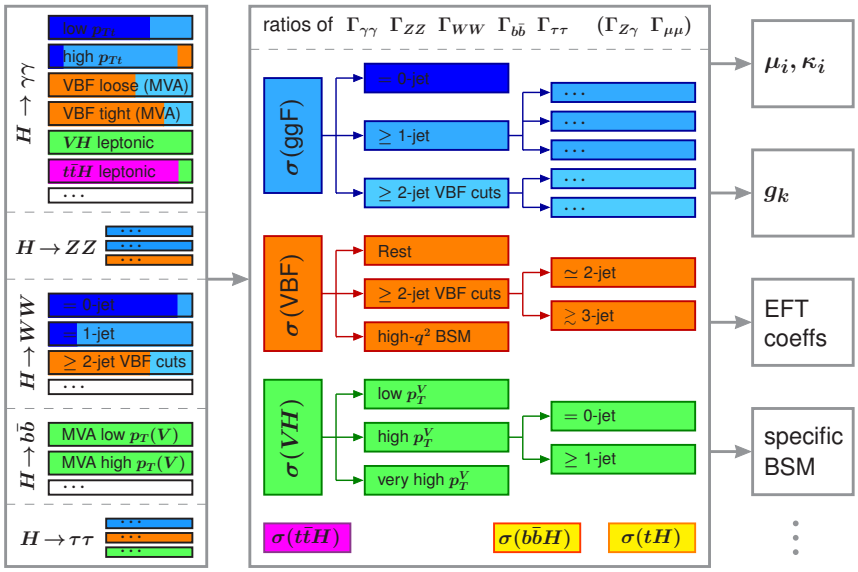


FIGURE 7.1: Schematic overview of the simplified template cross section framework [17]

On the left the experimental analyses are divided by decay channel, production mode and other experimental categories. In the center there is a sketch of the simplified template cross sections, which are determined from the experimental categories by a global fit They are cross sections per production mode, split possibly into mutually exclusive kinematic. The different Higgs boson decays are treated by fitting also the partial decay widths.
The measured simplified template cross sections together with the partial decay widths can then be used as input for different interpretations (figure 7.1 right). The signal strengths and couplings can be extracted, but also specific models of new physics and deviations from the SM parametrized in the effective field theory framework can be tested.

The theory dependence is aimed to be shifted to the interpretation: theoretical predictions and their uncertainties wouldn't be used for the actual measurement. The SM expected rates and kinematics will be used only as a guideline to build the analyses.
The fiducial regions or bins are built based on well-measured experimental quantities, and have a correspondence to truth bins. There will be residual theoretical uncertainties due to the experimental acceptances for each truth bin. In order to facilitate the interpretation the same definitions should be used by all the analyses, however, some will have limited sensitivity in some bins. For this reason, several stages with an increasing number of bins are defined.

In the future, as a larger amount of data will be analyzed, a finer binning will be adopted. In figure 7.2, a possible binning for the VH production mode, for which the H→b$\bar{\text{b}}$ channel contributes majorly, is shown.

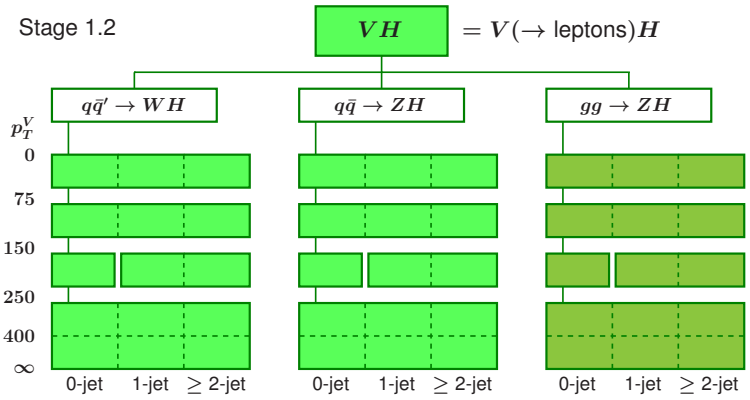


FIGURE 7.2: Proposed stage 1.2 STXS binning for for VH [17]

The categories are built by production process, separating the qq→WH, the qq→ZH and the gg→ZH processes. Events are further categorized by the transverse momentum of the vector boson p_T^V and if possible by the number of extra jets, where a common definition of extra-jet is adopted by all the analyses.

7.2 H→ μμ with CMS Run 2 data

The search for the H→ μμ is currently the most sensitive among the rare decay channels. The H(μμ) appears as a ∼3-4 GeV wide peak above a large background of mainly Drell-Yan events. In the full Run 2, about 1000 signal events are expected, while the DY to dimuon events are about 85000/GeV in the relevant region.

The most sensitive result published by CMS uses data collected in 2016 [138] in combination with the analogous Run 1 result. Data are found to be compatible with the predicted background. The expected upper limit is 2.2 times the standard model value with an expected significance of 1.0 standard deviations. The corresponding observed upper limit is 2.9 with an observed significance of 0.9 standard deviations. This corresponds to an observed signal strength of 1.0 ± 1.0 (stat) $\pm$ 0.1 (syst).
In the 2016 analysis events are classified into categories using variables that are largely uncorrelated with $m_{\mu\mu}$ in order to enhance the sensitivity to the Higgs boson signal. The primary Higgs boson production mechanisms targeted by this analysis are VBF and ggH. A fit of the invariant mass spectrum is then performed in each category to extract the signal.

A full Run 2 analysis is currently being finalized. Instead of a single analysis with multiple categories, dedicated production searches by production mode have been implemented. In particular a dedicated VBF channel analysis has been studied extensively. Dedicated t$\bar{\text{t}}$H and VH searches have also been added.

The dedicated search for VBF H→ μμ exploits the clean final state and excellent mass resolution due to the decay into muons of the Higgs boson, and the peculiar VBF topology, with two forward jets. Exploiting the VBF helps to reduce the background greatly, with respect to the ggH production. The main disadvantages are the lower VBF cross section (10% of the

total) and the very low branching ratio for H$\to \mu\mu$ ($2{\cdot}10^{-4}$). Therefore, only about a hundredth of events are expected in the full Run 2. After the event selection, which has a signal efficiency of $\sim$20% about 20 signal events are left, compared to a few thousand background events. The main background sources are the DY+jets and VBF production of a Z boson, as a small component of the Z resonance tail overlaps with the Higgs peak. Minor background sources are the $t\bar{t}$, single top and diboson production.

Similarly to the VH analysis, multivariate techniques that use the full event topology are employed to obtain the best possible discriminating power. Actually, the mass resolution is much better in this channel, and this makes in principle the search of a peak over a smooth background more sensitive. However, the total number of signal events expected is much lower, making the presence of a peak of about 20 events over a background of a few hundreds less significant. Multivariate techniques remove backgrounds optimally and isolate a handful of signal like events, thus maximizing the signal significance.

The best result in the multivariate analysis were given by a DNN, for which I tested several training options and setups. The DNN output in the best performing setup for the 2016 signal region is shown in figure 7.3. More details about the DNN training procedure can be found in appendix D.

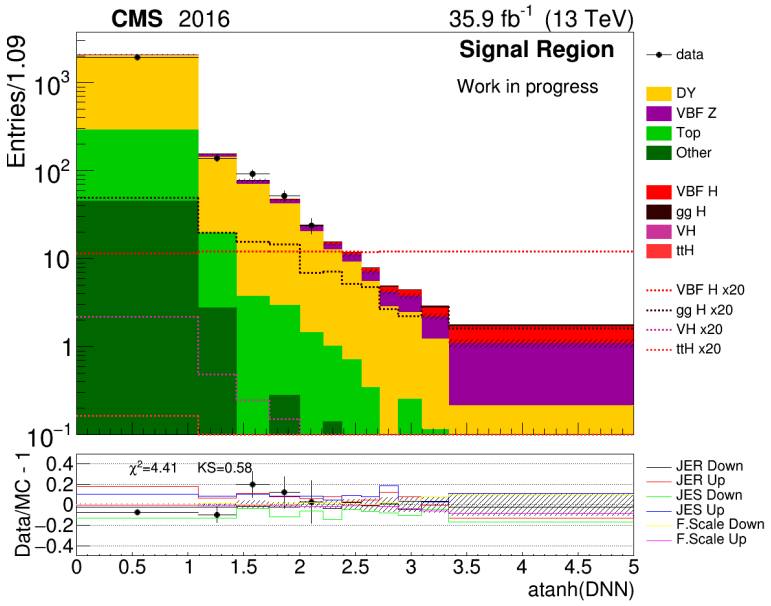


FIGURE 7.3: DNN distribution for 2016 data in the signal region. The DNN distribution in the signal region together with the DNN in the sideband are currently planned to be used in the final fit.

Appendix A

DeepVertex inputs

In this appendix reference plots for the input variables used in the DeepVertex training are reported.

A.1 DeepVertex Inputs

The jets used as input for the DeepVertex training have flavor composition as shown in figure A.1. The p_T and η distributions are built to be uniform in each flavor. The global jet features fed to the DNN, p_T, η, *phi* and m, are shown in figure A.2. The actual input distributions, after undergoing standardization and in case of p_T also a transformation, are shown in figure A.3.

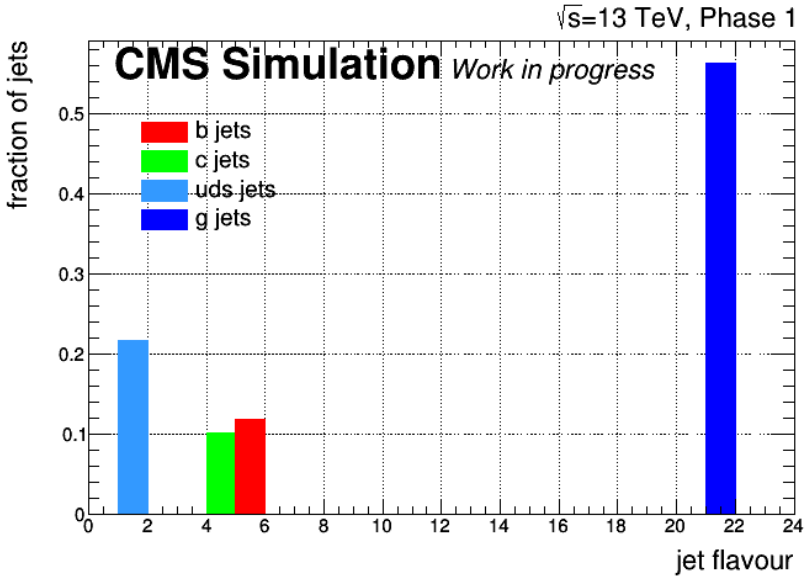
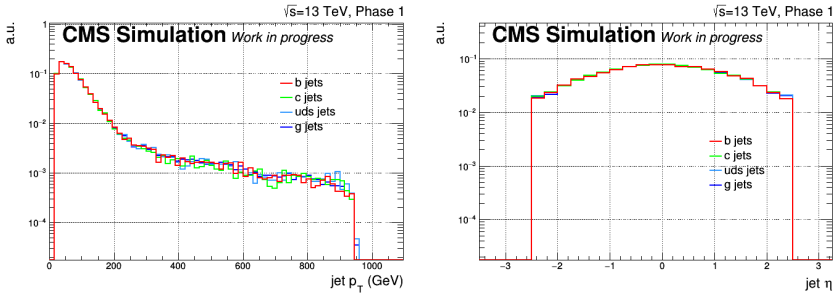
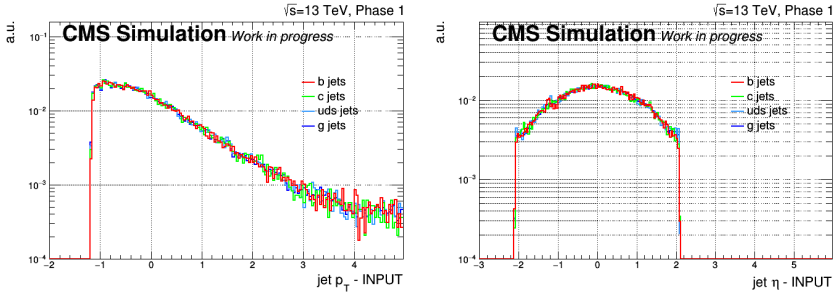


FIGURE A.1: flavor composition of the sample used for the training of DeepVertex. The four bins are filled with the fraction of jets per flavor. The bins are chosen according the convention: b quark jet - label "5", c quark jet - label "4", uds quark jet - label "1", gluon jet - label "21".



(A) (B) (C) (D)

FIGURE A.2: DNN input features - the jet 4 vector variables by jet flavor. All the distributions are normalized to unity.

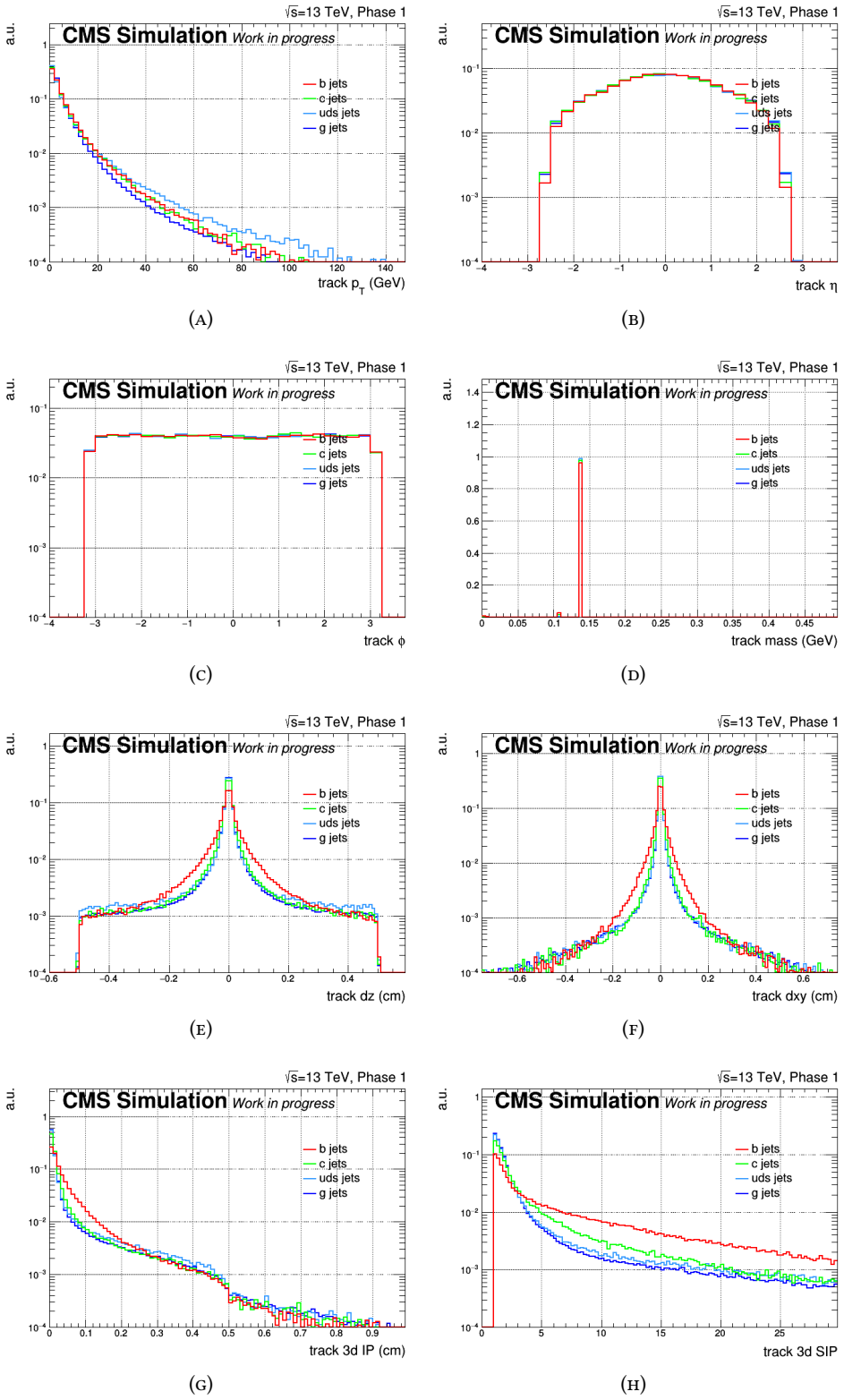


(A) (B) (C) (D)

FIGURE A.3: DNN input features - the jet 4 vector variables by jet flavor. The inputs are transformed and standardized. All the distributions are normalized to unity.

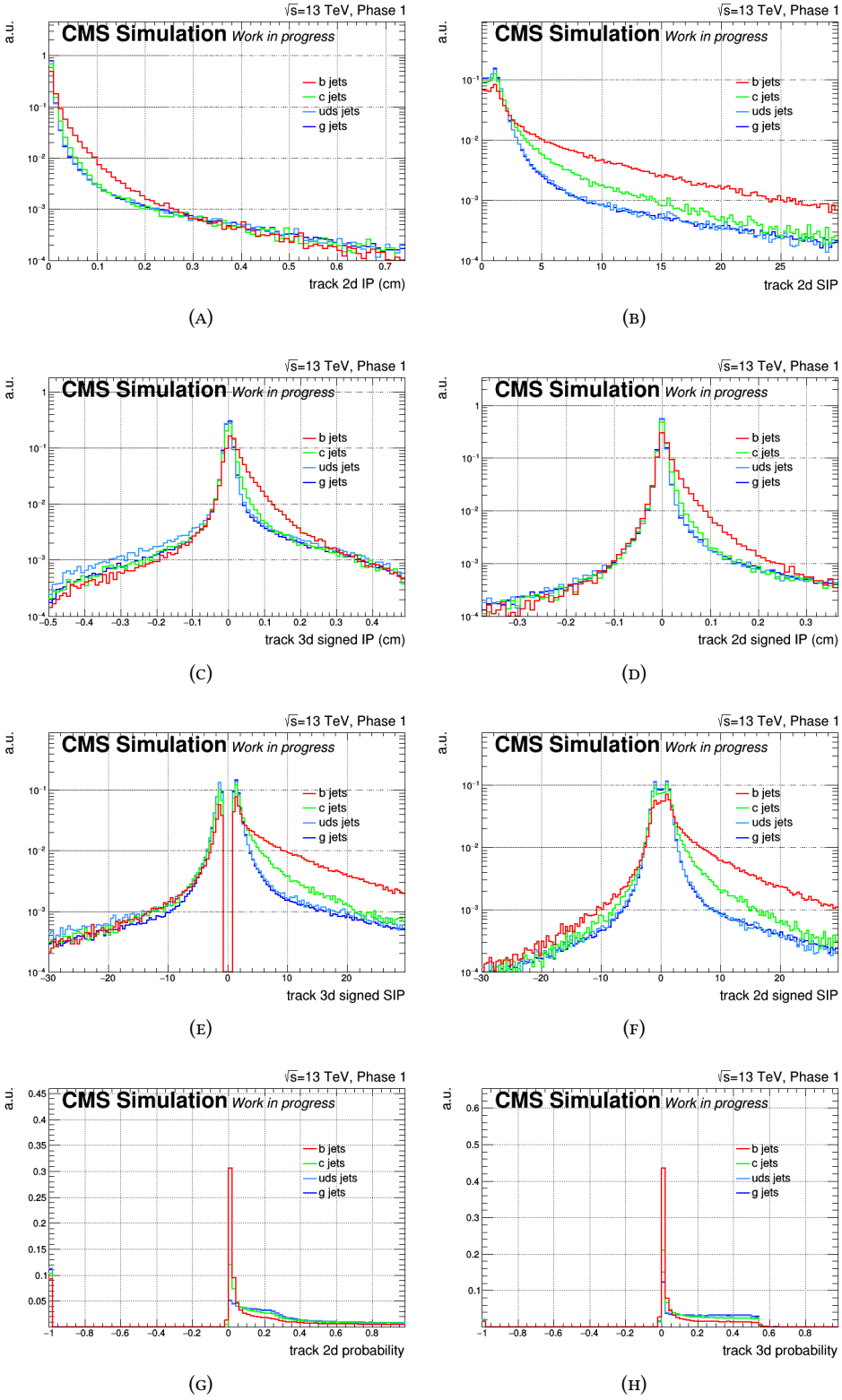
The seeding tracks features used as input for the DNN are :

- the 4 vector componests: p_T, η, *phi* and m of the track (figure A.4 (A-D))
- the transverse and longitudinal impact parameters: *dxy* and *dz* (figure A.4 (E), (F))
- the 2d (transverse) and 3d impact parameters and their significances: 3D IP, IP significance, 2D IP, IP significance (figure A.4 (G), (H), figure A.5 (A), (B))
- the 2d (transverse) and 3d impact parameters and their significances with the jet realtive sign: signed 3D IP, IP significance, 2D IP, IP significance (figure A.5 (C-F))
- the 2d and 2d track probabilities (figure A.5 (G), (H))
- track quality features: the reduced χ^2, the number of hits and pixel hits (figure A.6 (A-C))
- the distance from the jet axis and the distance of the point of closest approach to the jet axis from the primary vertex (figure A.6 (D), (E))



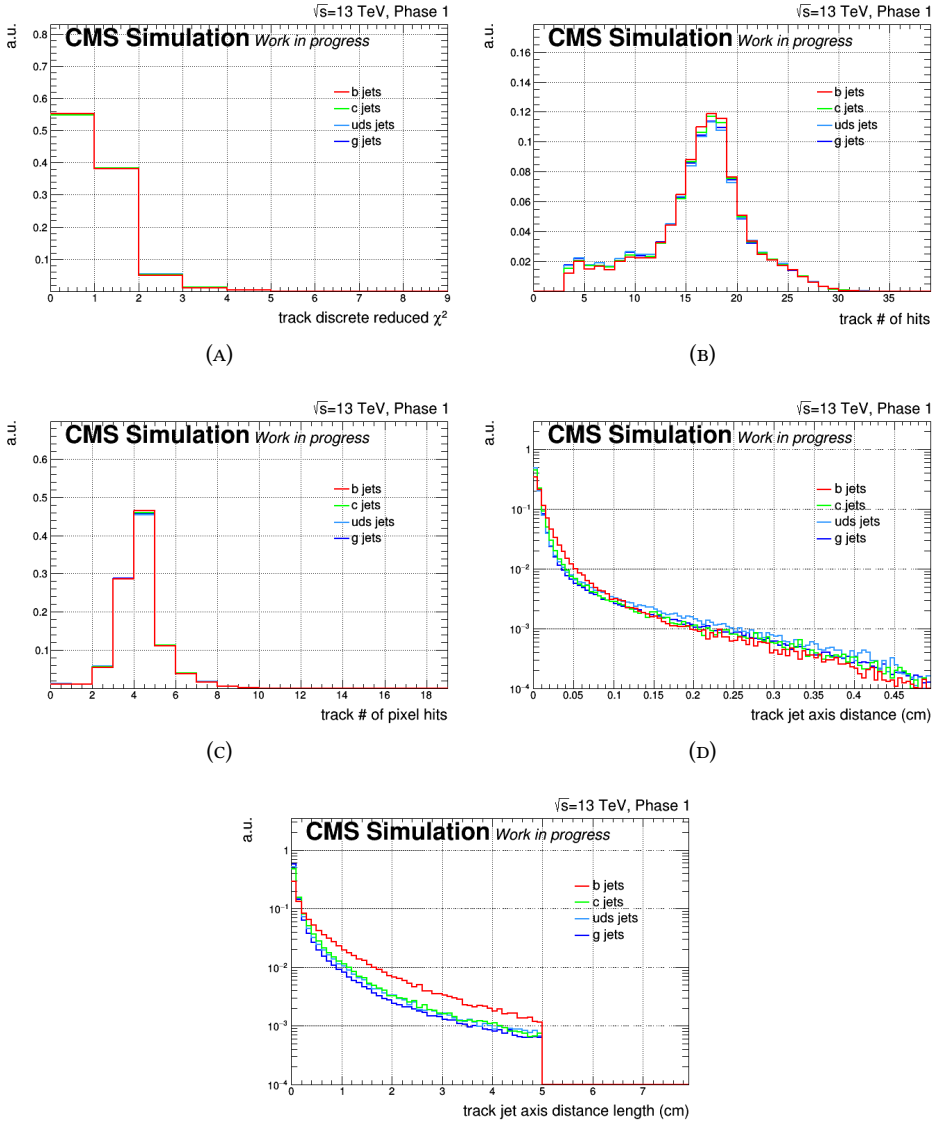
(A) (B) (C) (D) (E) (F) (G) (H)

FIGURE A.4: DNN input features - the inputs relative to the seeding tracks are shown. All the seeding tracks of a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity.



(A) (B) (C) (D) (E) (F) (G) (H)

FIGURE A.5: DNN input features - the inputs relative to the seeding tracks are shown. All the seeding tracks of a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity.



(A) (B) (C) (D) (E)

FIGURE A.6: DNN input features - the inputs relative to the seeding tracks are shown. All the seeding tracks of a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity.

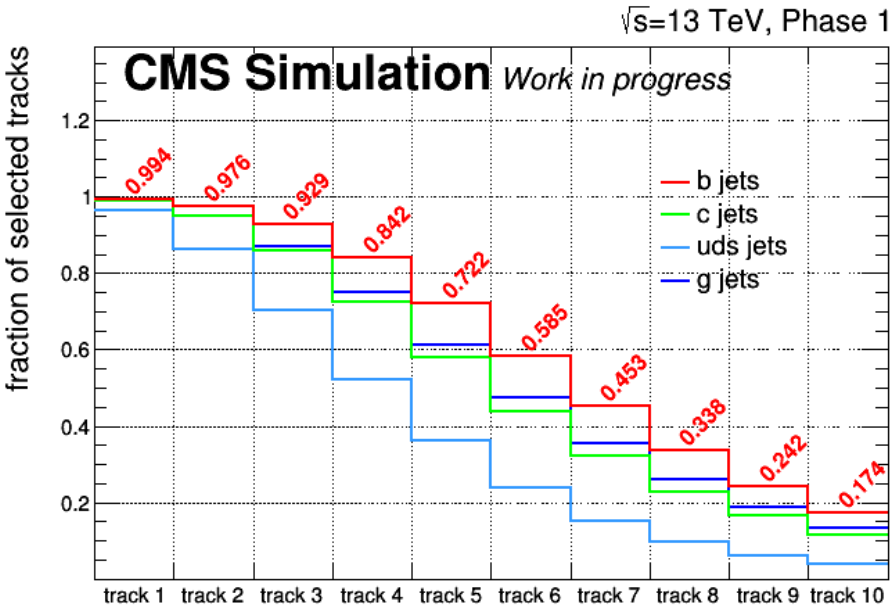


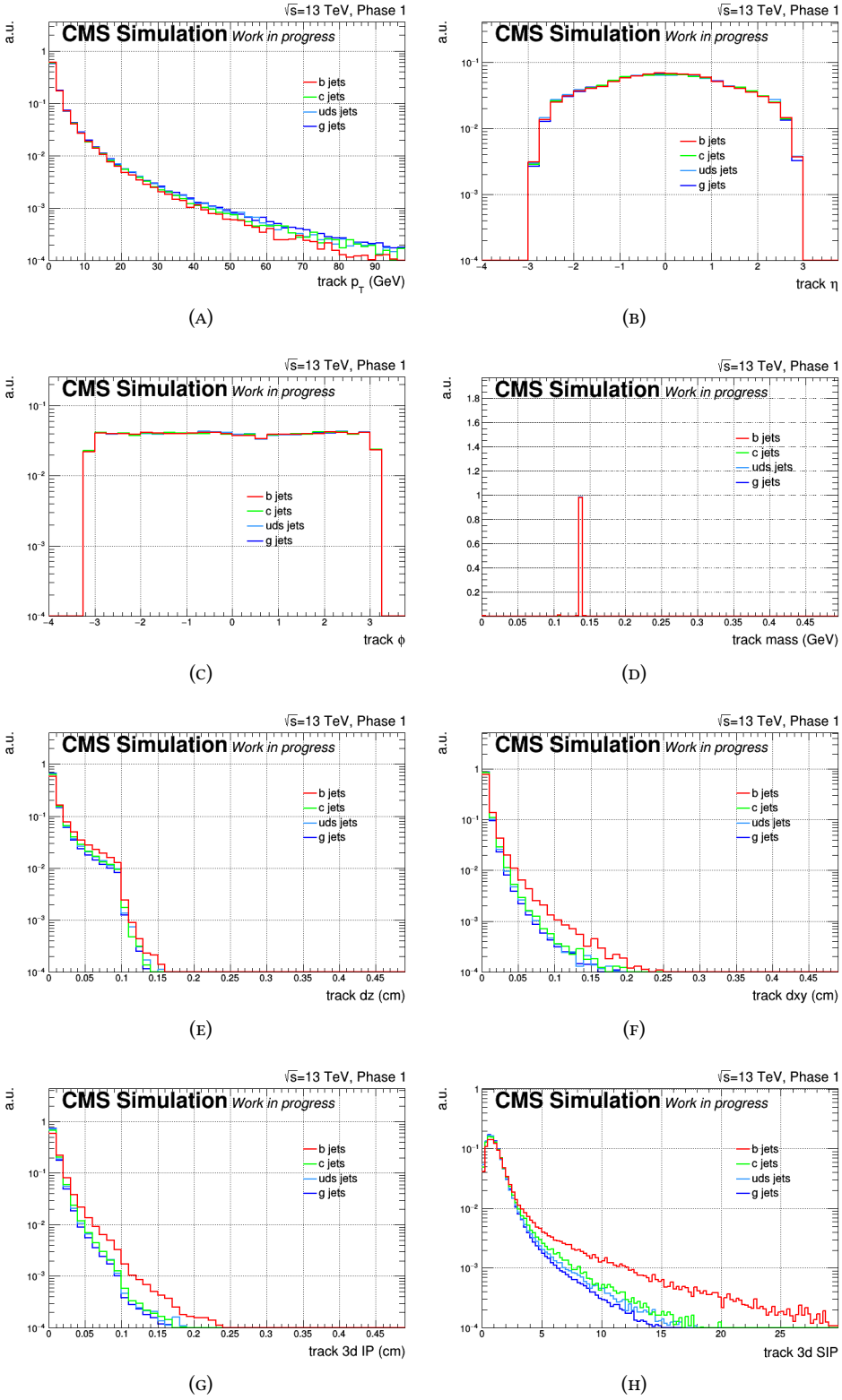
FIGURE A.7: Fraction of tracks selected as seeding tracks by flavor by track sorting position. The tracks are sorted by signed impact parameter significance. The fractions reported on top of the bins histograms are the ones of b jets, which have the maximum seeding efficiency.

The neighbor tracks features used as input for the DNN are :

- the 4 vector componests: p_T, η, *phi* and m of the track (figure A.8 (A-D))
- the transverse and longitudinal impact parameters: *dxy* and *dz* (figure A.8 (E), (F))
- the 2d (transverse) and 3d impact parameters and their significances: 3D IP, IP significance, 2D IP, IP significance (figure A.8 (G), (H), figure A.9 (A), (B))

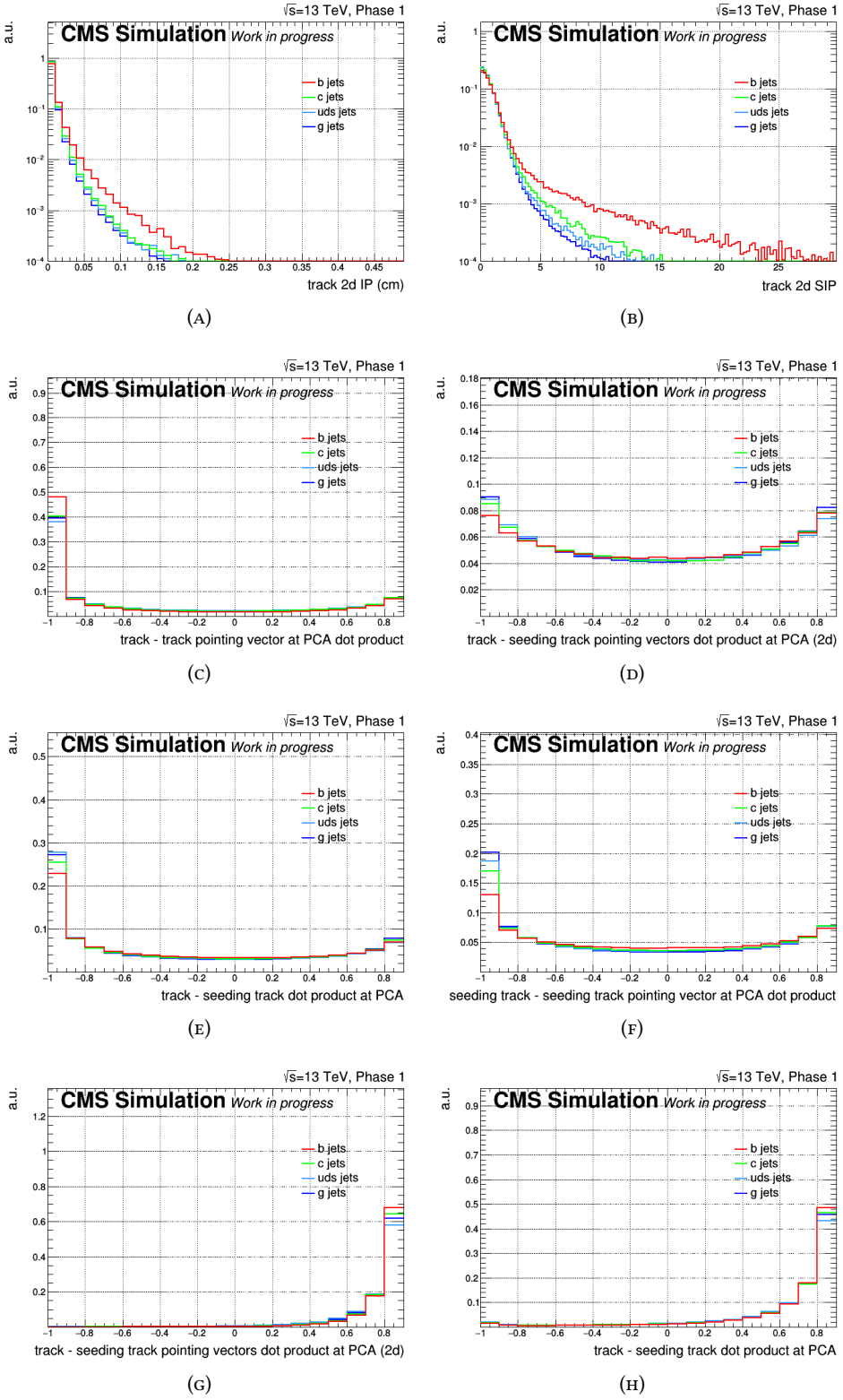
The other variables are relative to the point of closest approach between the seeding track and its neighbour. These are:

- the distance at point of closest approach, or PCA, and its significance (figure A.10 (A), (B))
- the PCA (x, y, z) coordinates both on the seeding track and on the neighbour track, an the uncertainties (Δx, Δy, Δz) for both points (figure A.11, A.12)
- the scalar product between the track and the PCA direction both for the neighbour track and the seeding track (figure A.9 (C), (F))
- the scalar product between the seeding track and the neighbour track (both in 3D and in the transverse plane) (figure A.9 (D), (E))
- the scalar product between the PCA directions on the seeding and neighbour track (both in 3D and in the transverse plane) (figure A.9 (G), (H))
- the PCA distance from the primary vertex, both for the PCA on the seeding track and the neighbour track (figure A.10 (G), (H))
- jet relative variables: the distance pf the PCA (the central one) from the jet axis, the $\Delta\eta$ and $\Delta\phi$ of the PCA direction from the jet axis direction, the scalar product between the jet direction and the direction given by the momentum 4-vector sum of the two tracks. (figure A.10 (C-F))



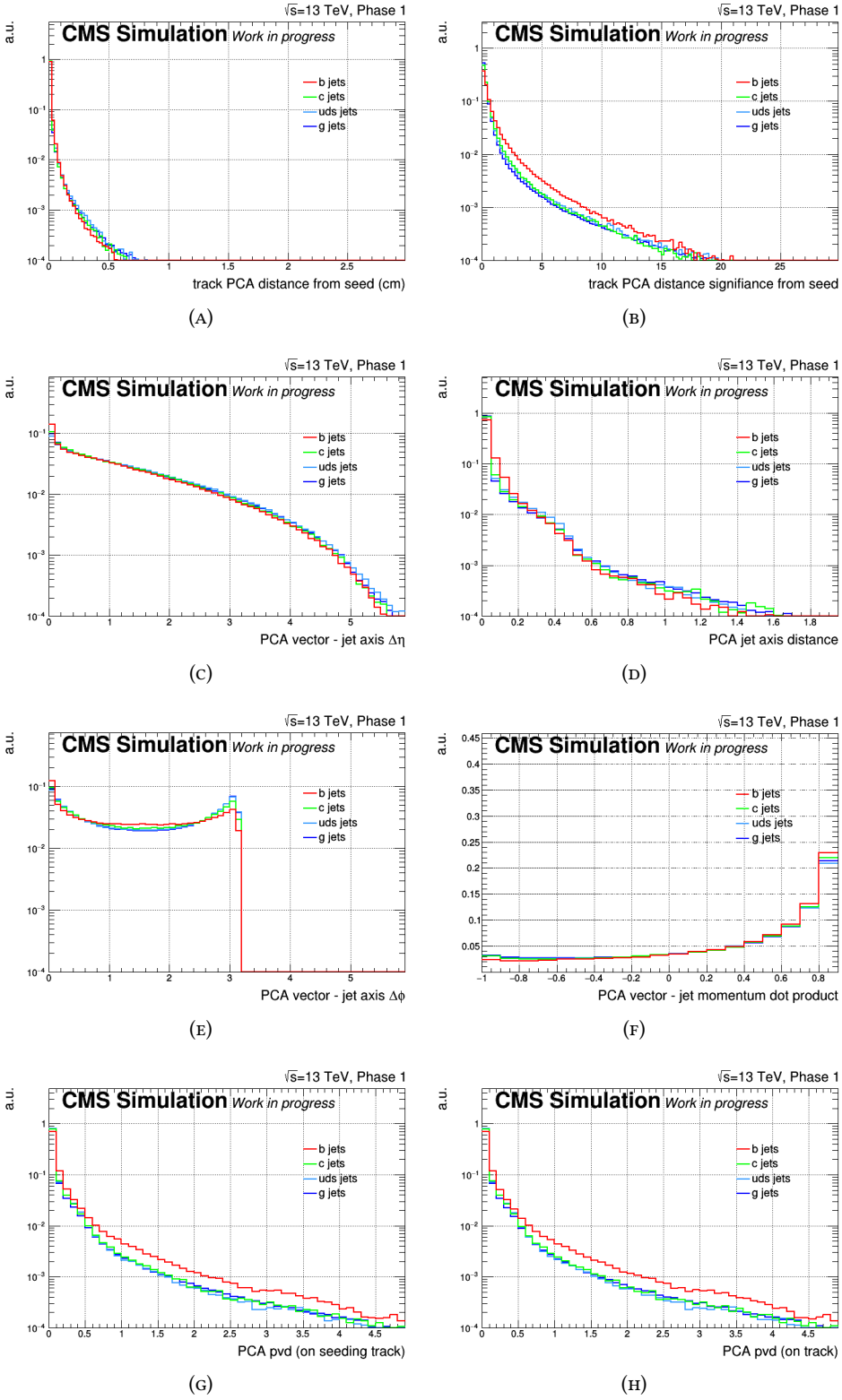
(A) (B) (C) (D) (E) (F) (G) (H)

FIGURE A.8: DNN input features - the inputs relative to the neighbour tracks are shown. All the neighbour tracks for each seeding track in a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity.



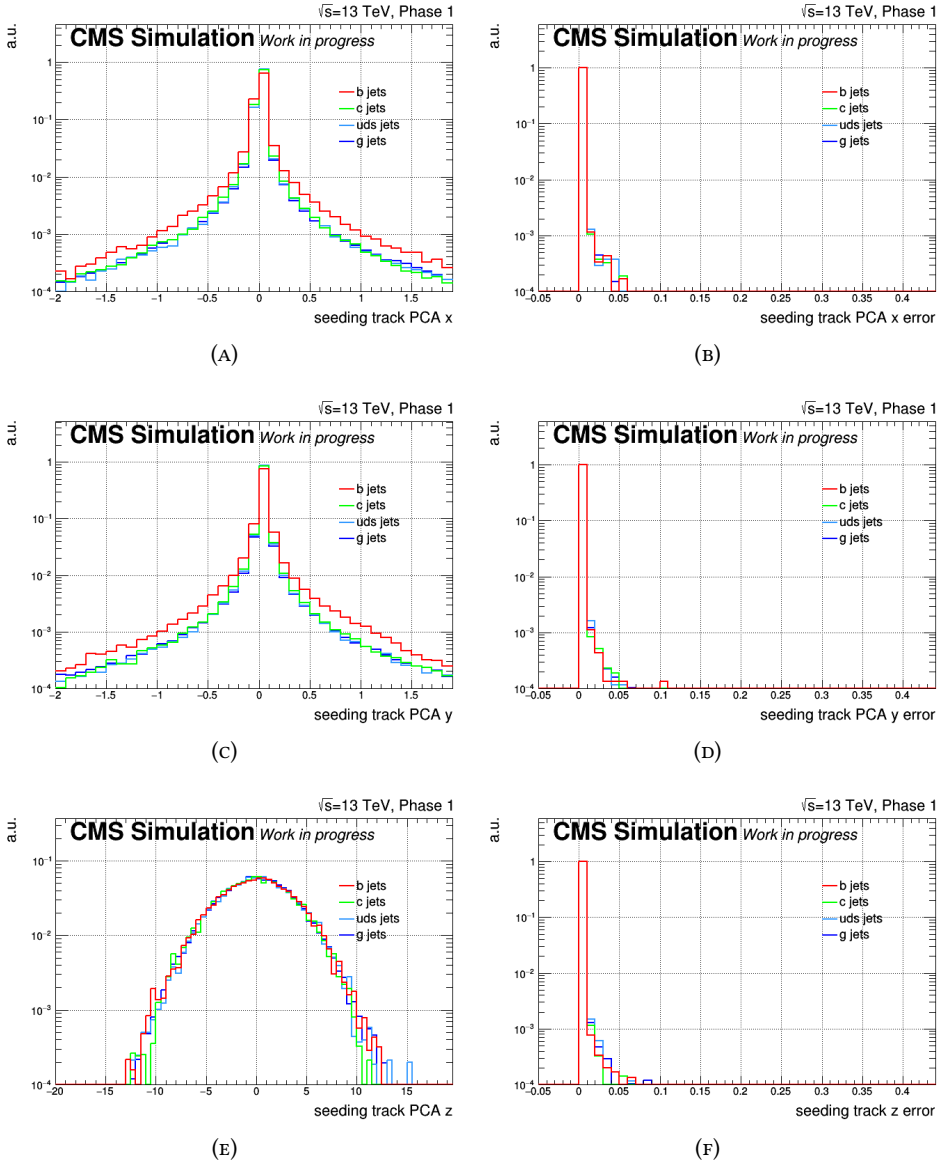
(A) (B) (C) (D) (E) (F) (G) (H)

FIGURE A.9: DNN input features - the inputs relative to the neighbour tracks are shown. All the neighbour tracks for each seeding track in a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity.



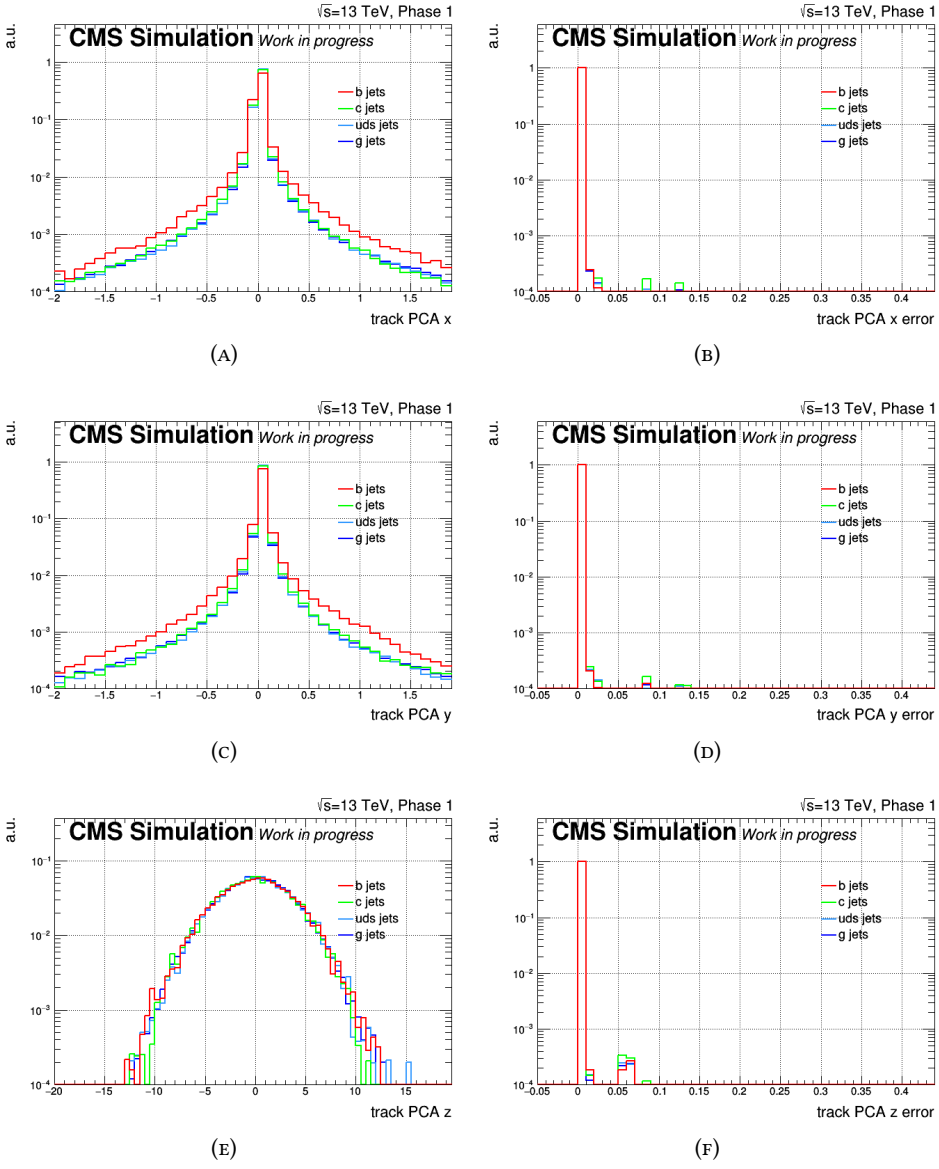
(A) (B) (C) (D) (E) (F) (G) (H)

FIGURE A.10: DNN input features - the inputs relative to the neighbour tracks are shown. All the neighbour tracks for each seeding track in a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity.



(A) (B) (C) (D) (E) (F)

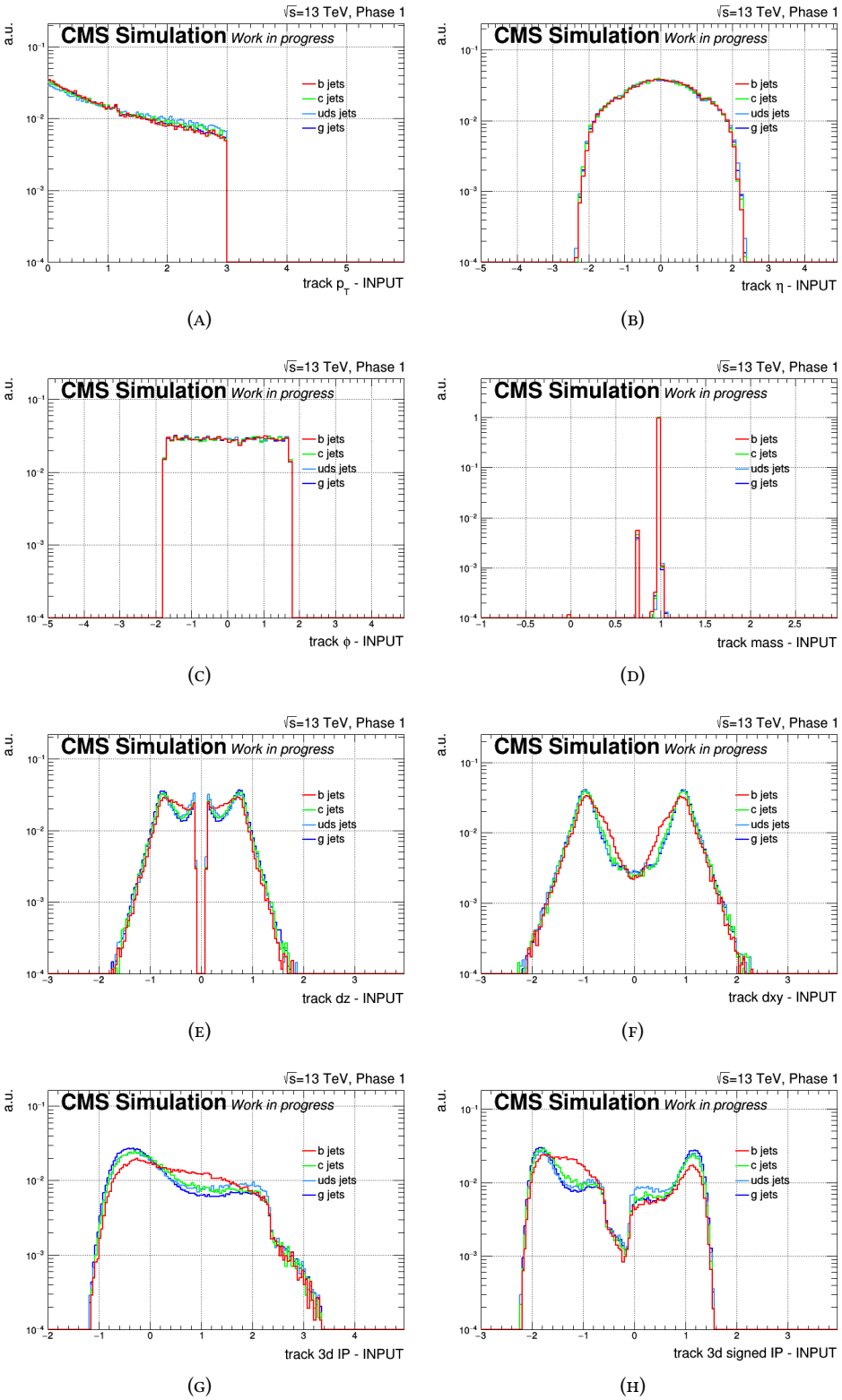
FIGURE A.11: DNN input features - the inputs relative to the neighbour tracks are shown. All the neighbour tracks for each seeding track in a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity.



(A) (B) (C) (D) (E) (F)

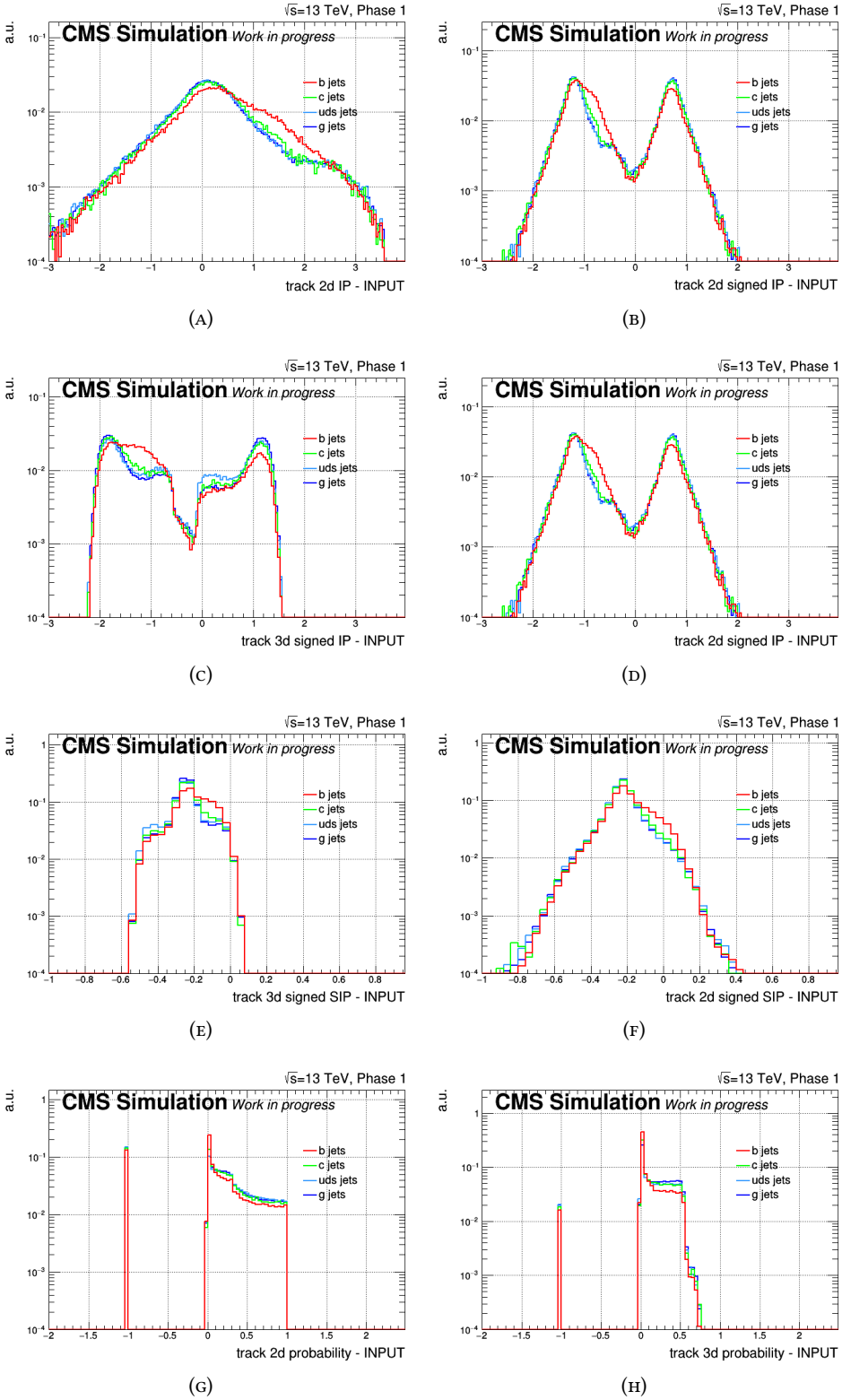
FIGURE A.12: DNN input features - the inputs relative to the neighbour tracks are shown. All the neighbour tracks for each seeding track in a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity.

The actual DNN input distributions are shown in figures A.13, A.14, A.15 for the neighbour tracks features; in figures A.16, A.17, A.18, A.19, A.20 for the neighbour tracks features.



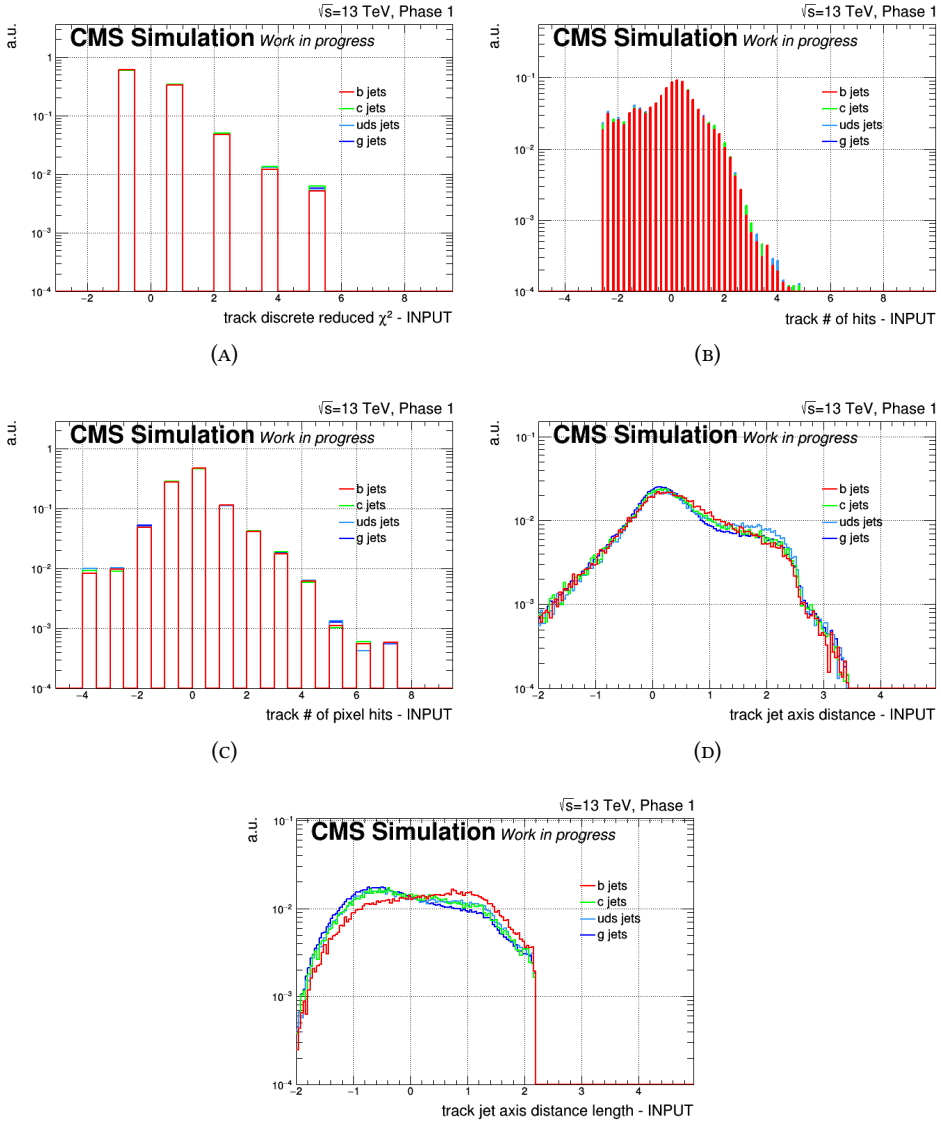
(A) (B) (C) (D) (E) (F) (G) (H)

FIGURE A.13: DNN input features - the inputs relative to the seeding tracks are shown. All the seeding tracks of a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity. The variables are trasformed and standardized.



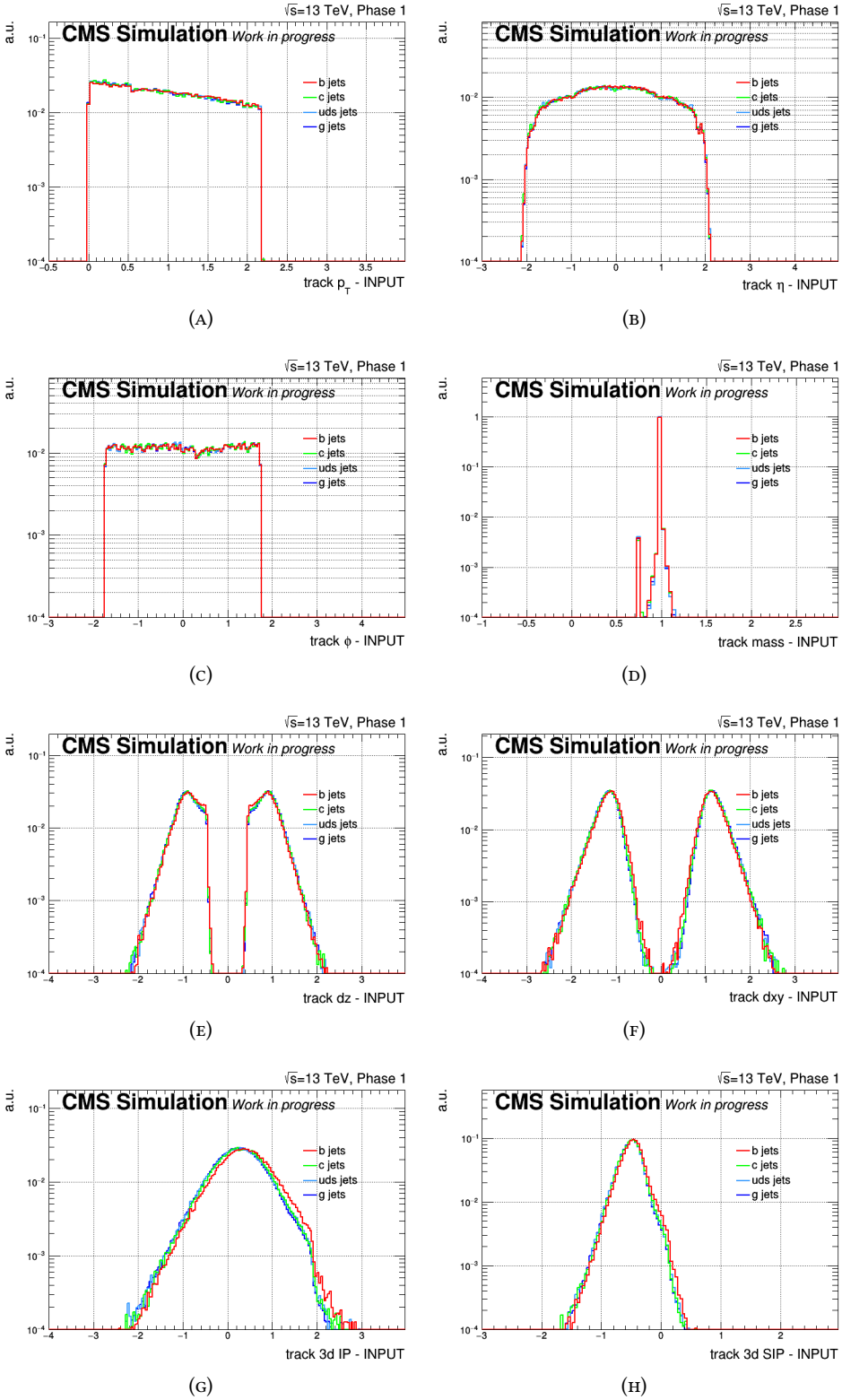
(A) (B) (C) (D) (E) (F) (G) (H)

FIGURE A.14: DNN input features - the inputs relative to the seeding tracks are shown. All the seeding tracks of a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity. The variables are trasformed and standardized.



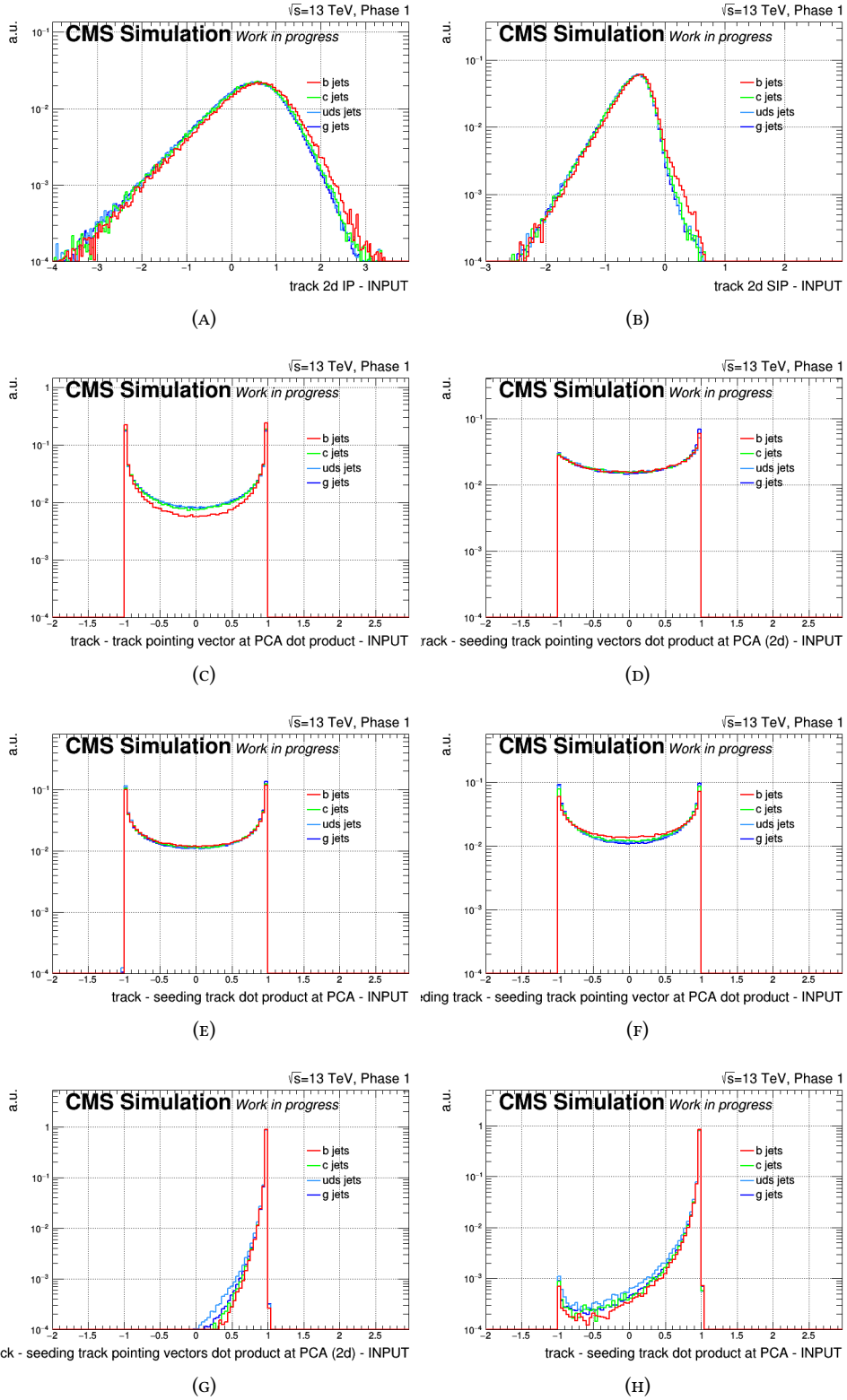
(A) (B) (C) (D) (E)

FIGURE A.15: DNN input features - the inputs relative to the seeding tracks are shown. All the seeding tracks of a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity. The variables are trasformed and standardized.



(A) (B) (C) (D) (E) (F) (G) (H)

FIGURE A.16: DNN input features - the inputs relative to the neighbour tracks are shown. All the neighbour tracks for each seeding track in a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity. The variables are trasformed and standardized.



(A) (B) (C) (D) (E) (F) (G) (H)

FIGURE A.17: DNN input features - the inputs relative to the neighbour tracks are shown. All the neighbour tracks for each seeding track in a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity. The variables are trasformed and standardized.

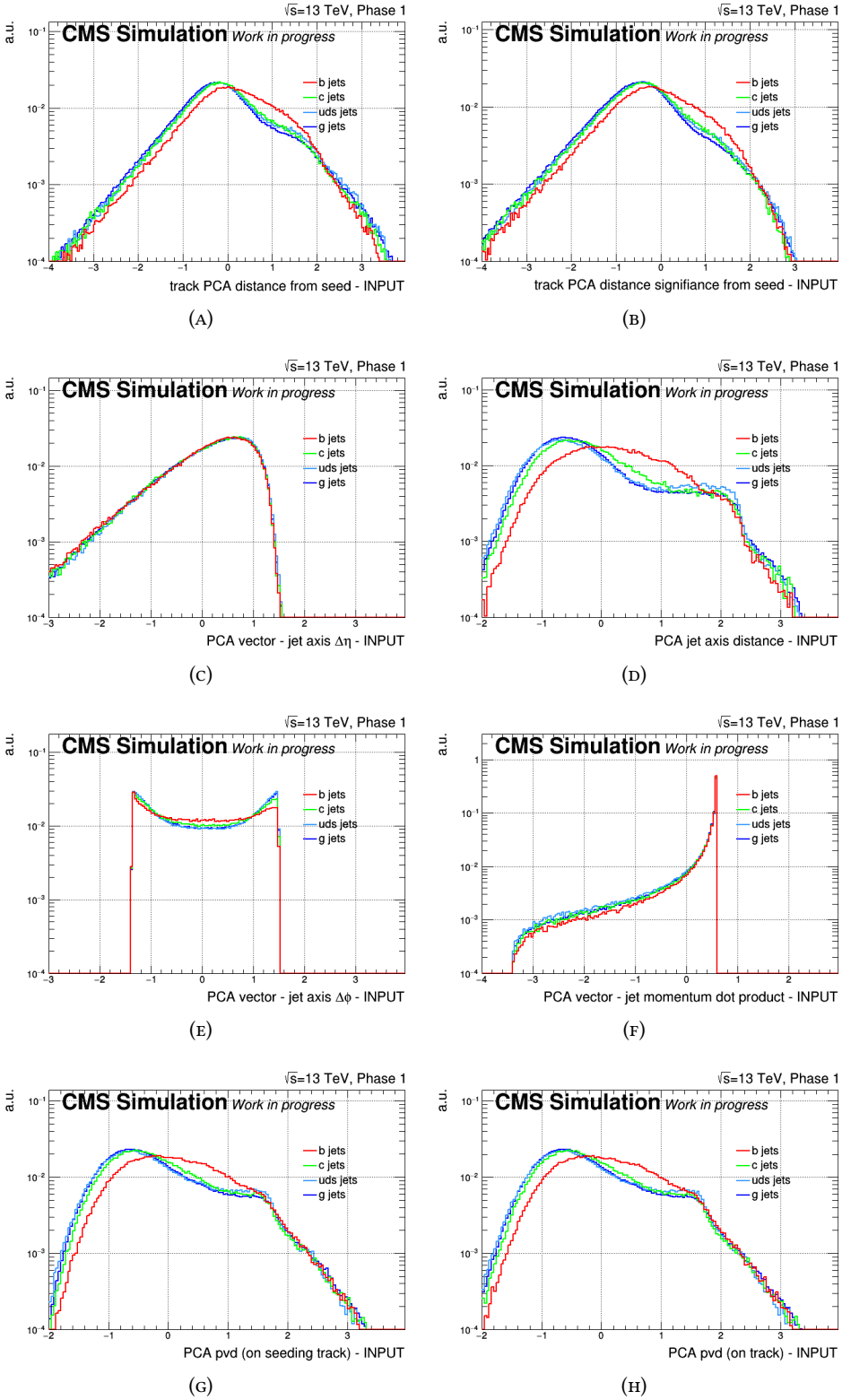
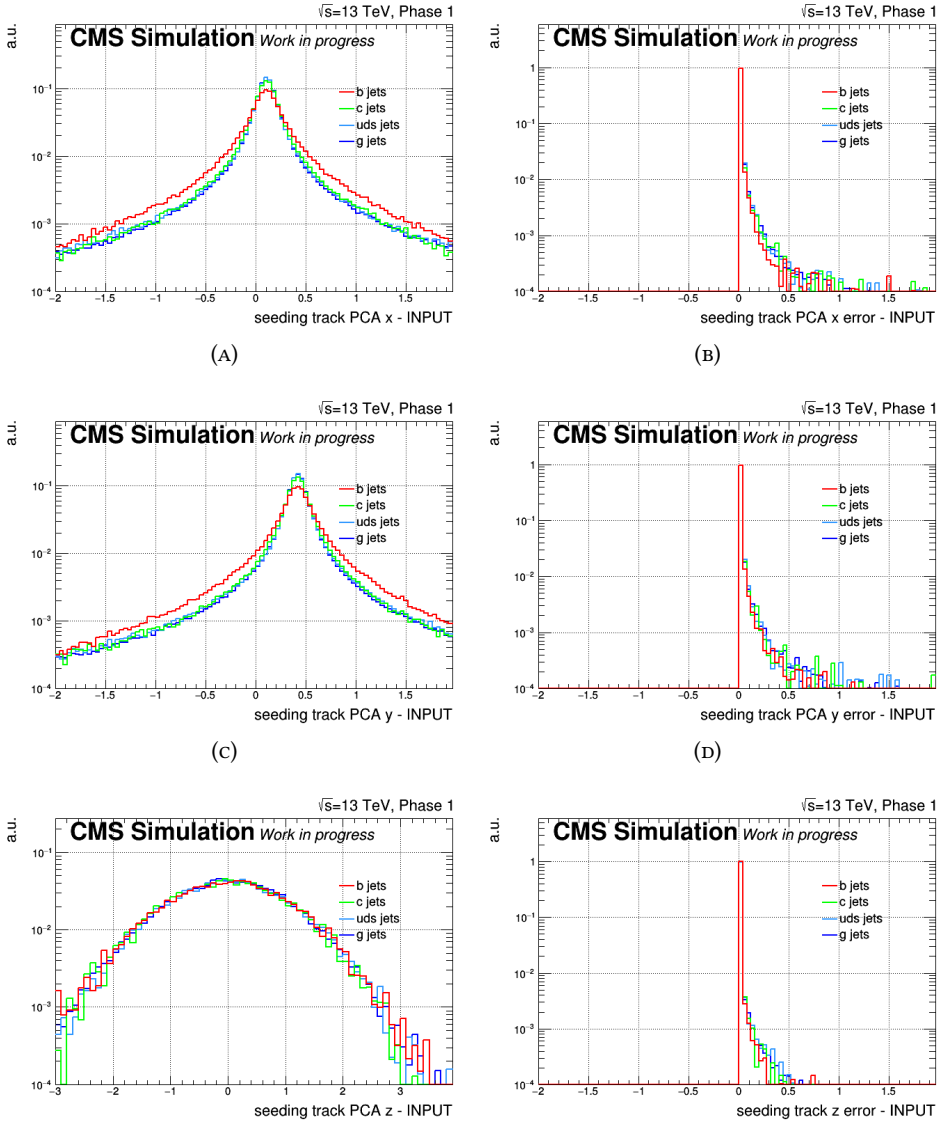
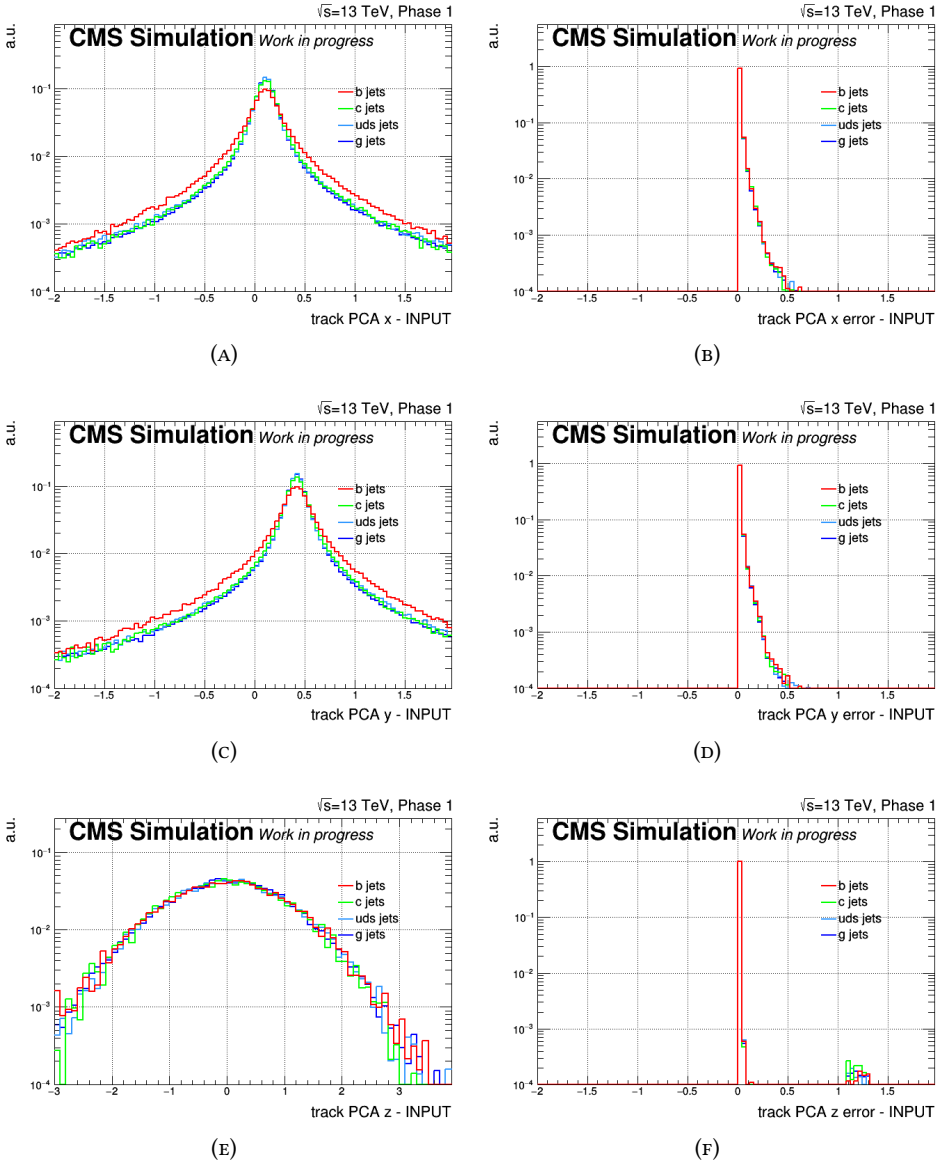


FIGURE A.18: DNN input features - the inputs relative to the neighbour tracks are shown. All the neighbour tracks for each seeding track in a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity. The variables are trasformed and standardized.



(A) (B) (C) (D) (E) (F)

FIGURE A.19: DNN input features - the inputs relative to the seeding tracks are shown. All the neighbour tracks for each seeding track in a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity. The variables are trasformed and standardized.



(A) (B) (C) (D) (E) (F)

FIGURE A.20: DNN input features - the inputs relative to the seeding tracks are shown. All the neighbour tracks for each seeding track in a jet are used, and the distributions are shown by flavor of the jet, not by flavor of the originating hadron. All the distributions are normalized to unity. The variables are trasformed and standardized.

A.2 DeepJet input features

The features used in the DeepJet training are listed below for reference.

The jet and global event features included are:

- the jet p_T and η.
- the number of charged PF candidates, of neutral PF candidates, of secondary vertices associated with the jet
- the number of primary vertices in the event
- b-tagging specific features:
  - the ratio of track sum transverse energy over jet energy
  - the ΔR distance between the jet axis and track 4-vector sum
  - the category of secondary vertex (one associated vertex or a "pseudovertex")
  - the 3D and 2D signed impact parameter and significances of first track lifting mass above charm
  - the number of tracks passing the two different b-tagging selection criteria.

The features (12) of the charged PF candidates included are:

- the p_T of the secondary vertex (1)
- the ΔR relative to the jet axis (1)
- the invariant mass of the vertex (1)
- the number of associated tracks (1)
- the χ^2 and normalized χ^2 of the vertex fit (1)
- the 2d (transverse) and 3d impact parameters and their significances: signed 3D IP, IP significance, 2D IP, IP significance (4)
- the vertex/jet energy ratio (1)
- the cosine of the angle between the vertex and the jet axis directions (1)

The features (16) of the charged PF candidates included are:

- the p_T, η relative to the jet axis (2)
- the ΔR relative to the jet axis (1)
- the momentum component parallel to the jet axis and its fraction relative to the track momentum (2)
- the 2d (transverse) and 3d impact parameters and their significances with the jet relative sign: signed 3D IP, IP significance, 2D IP, IP significance (4)
- the distance from the jet axis (1)
- the p_T over jet p_T ratio (1)
- the minimum ΔR from a secondary vertex, if present (1)

- the track quality of the association to the primary vertex (2)
- the probability of being a pileup track, in the form of a weight assigned the PUPPI pileup removal algorithm (1)
- the track χ^2 and a quality flag (1)

The features (6) of the neutral PF candidates included are:

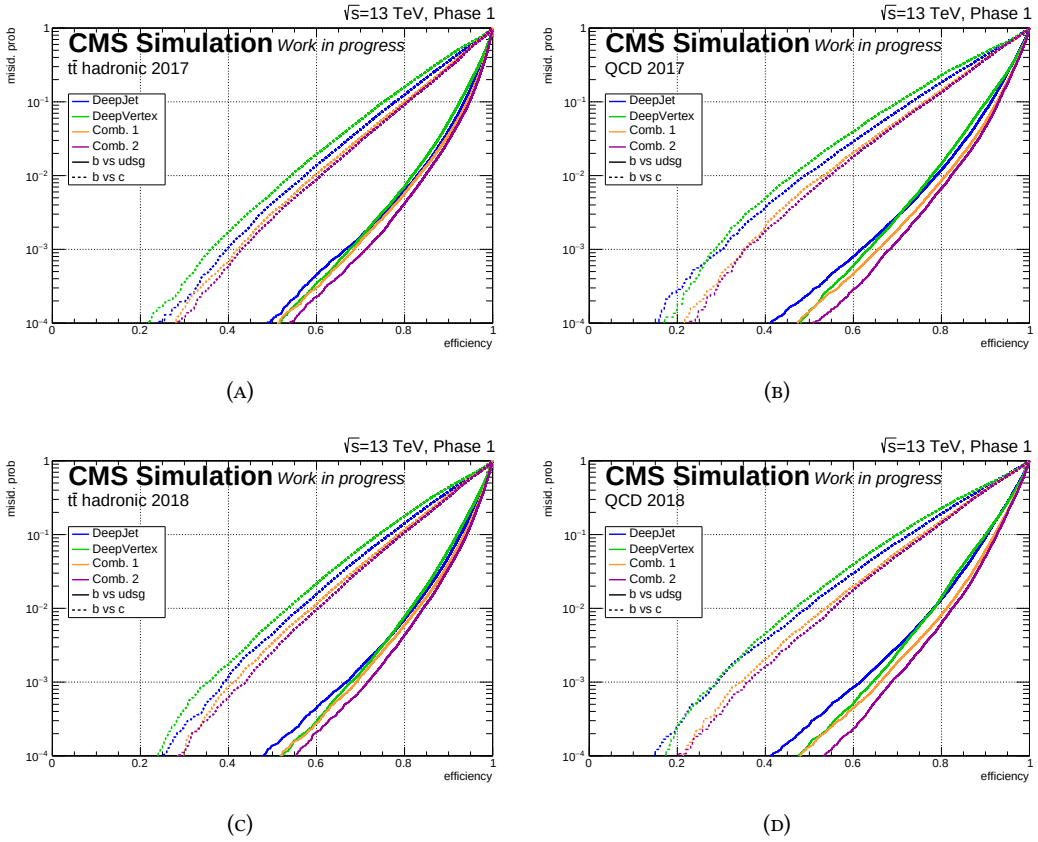
- the p_T relative to the jet axis (1)
- the ΔR relative to the jet axis (1)
- the photon ID flag (1)
- the fraction of energy in the hadron calorimeter (1)
- the probability of being a pileup energy deposit, in the form of a weight assigned the PUPPI pileup removal algorithm (1)
- the minimum ΔR from a secondary vertex if present (1)

Appendix B

DeepVertex results

This appendix contains an extended version of the DeepVertex results presented in chapter 5. The ROC curves by year (2017 and 2018 simulation) and in p_T bins are reported for both $t\bar{t}$ and QCD simulation.

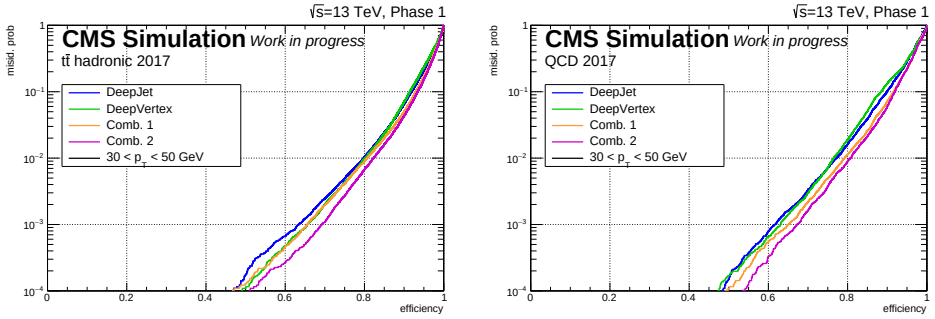
B.1 ROC curves



(A) (B)

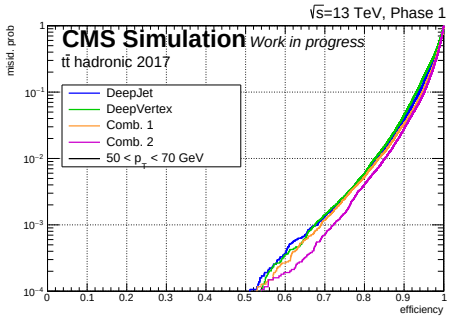
(C) (D)

FIGURE B.1: DNN results for both 2017 and 2018 simulated samples - Inclusive jet spectra for the all hadronic $t\bar{t}$ and QCD are used.

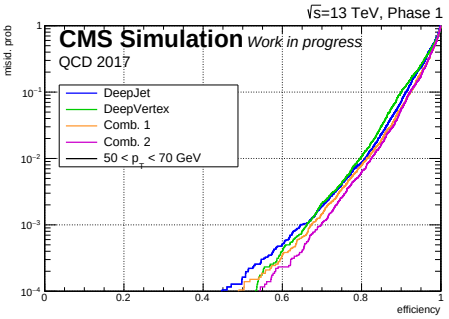


(A)

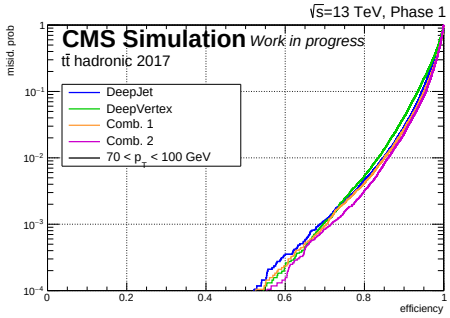
(B)



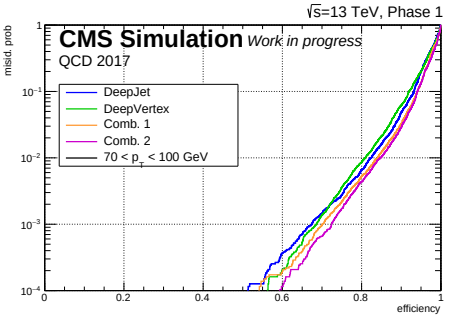
(C)



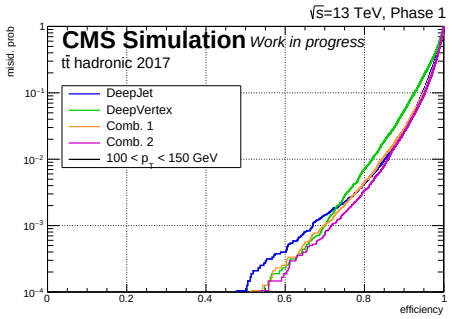
(D)



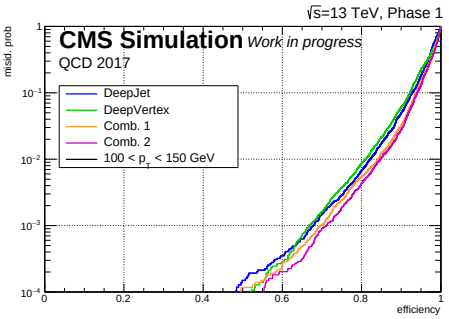
(E)



(F)

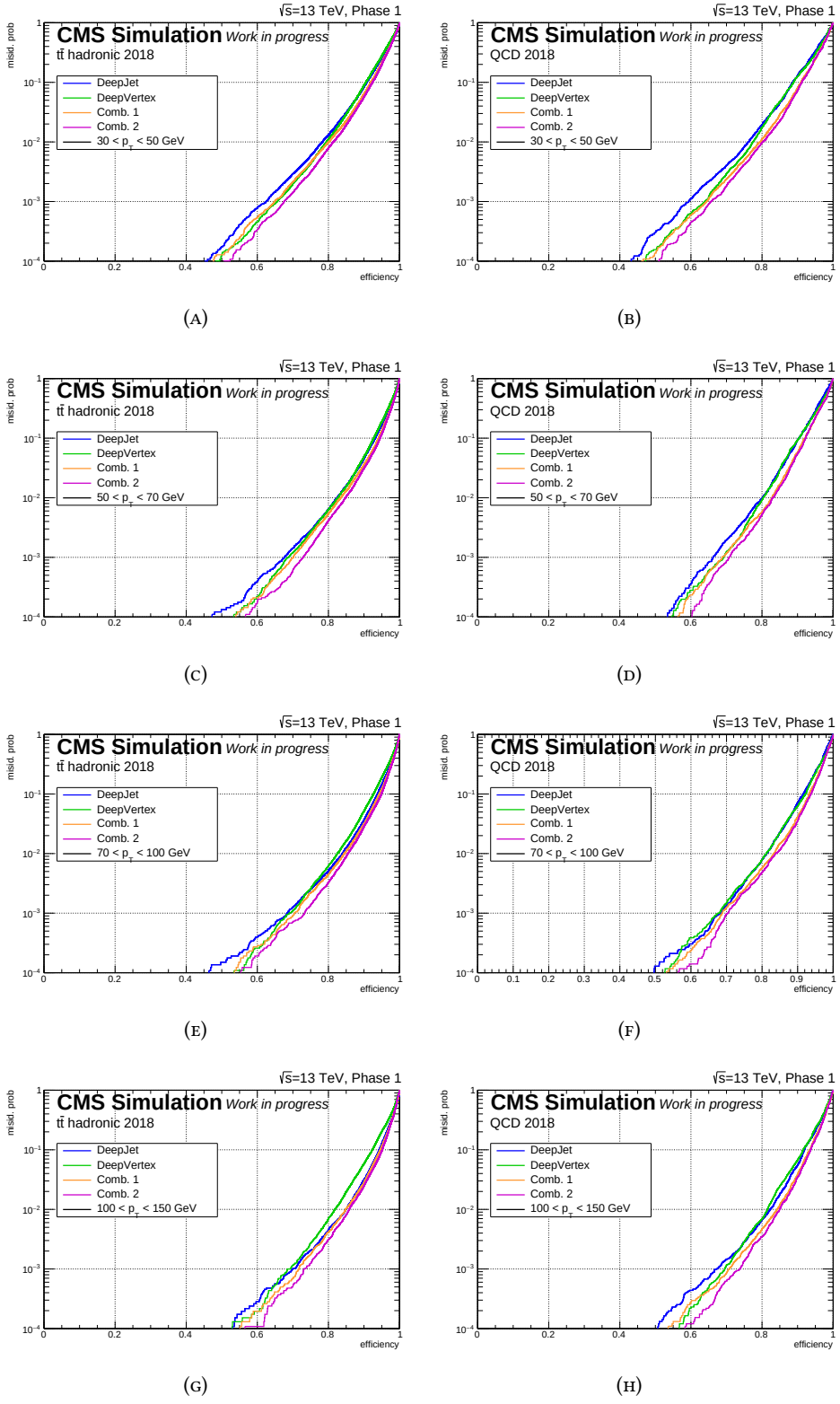


(G)



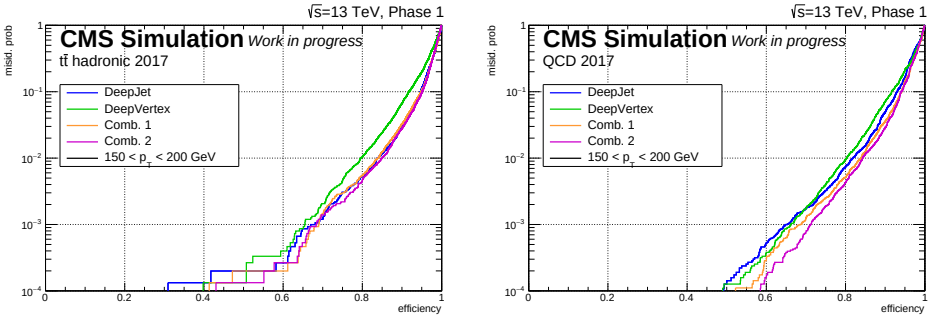
(H)

FIGURE B.2: DNN Roc curves in p_T bins (2017 simulation).



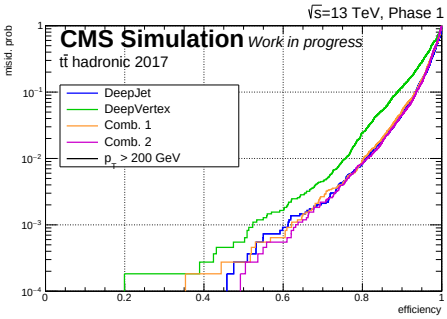
(A) (B) (C) (D) (E) (F) (G) (H)

FIGURE B.3: DNN Roc curves in p_T bins (2018 simulation).

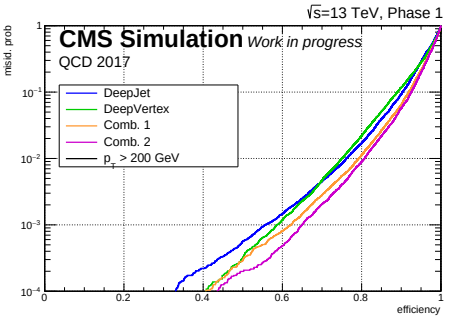


(A)

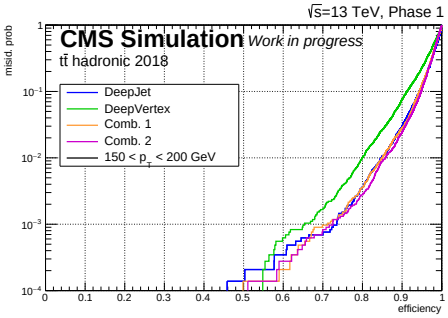
(B)



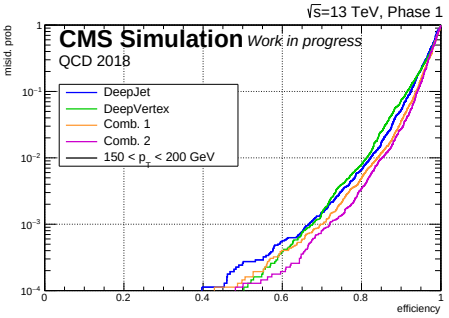
(C)



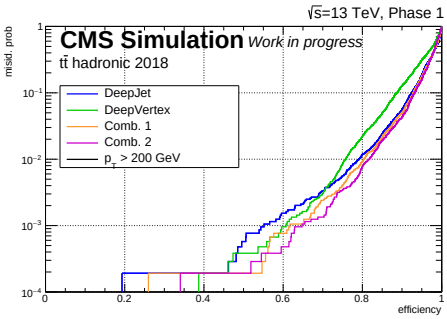
(D)



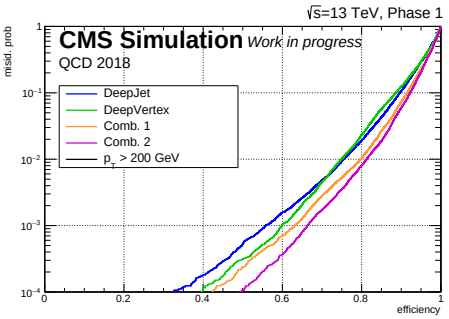
(E)



(F)



(G)



(H)

FIGURE B.4: DNN Roc curves for the high p_T bins (2017 and 2018 simulation).

Appendix C

Regression comparisons

The appendix contains comparison plots between the BDT regression and the new DNN based regression performed using Z($\ell\ell$)H($b\bar{b}$) signal events. No FSR recovery is applied. The net effect of the FSR recovery is also shown for the Z($\ell\ell$) channel.

C.1 Z($\ell\ell$)H($b\bar{b}$) regression comparisons

The DNN based regression is compared to the BDT based one used in [35] and to the baseline dijet invariant mass (figure C.1). No FSR recovery is applied, while the selection is the 2-lepton signal region, except for the data/MC full re-weighting. The high and low $p_{\mathrm{T}}^{\mathrm{V}}$ cases are both tested. The regression effects are tested in the inclusive case, in case one b jet contains a lepton and in case no lepton is reconstructed inside the jet cone.

The peak value is shifted to values close to 125 in all cases, both for the BDT and DNN based regression. The DNN based regression is also found to provide the best invariant mass resolution, measured as the gaussian width of a Bukin function [131], in all cases, as expected from single jet performance studies. The mean and standard deviations obtained from the fits are also reported in table C.1.

Variable / CR	Baseline		BDT regression		DNN regression	
	μ	σ	μ	σ	μ	σ
low $p_{\mathrm{T}}^{\mathrm{V}}$	110.6	19.8	120.1	19.2	122.4	17.9
low $p_{\mathrm{T}}^{\mathrm{V}}$ semileptonic	108.6	19.4	117.9	19.4	122.8	18.3
low $p_{\mathrm{T}}^{\mathrm{V}}$ no leptons	113.7	19.6	122.4	18.5	122.1	17.6
high $p_{\mathrm{T}}^{\mathrm{V}}$	116.5	17.9	123.1	17.4	123.8	15.5
high $p_{\mathrm{T}}^{\mathrm{V}}$ semileptonic	113.0	18.0	120.0	18.2	122.9	16.5
high $p_{\mathrm{T}}^{\mathrm{V}}$ no leptons	120.5	16.6	126.1	15.7	124.7	14.3

TABLE C.1: $\mathrm{Z}(\ell\ell)\mathrm{H}(\mathrm{b\bar{b}})$ regression comparisons. The low $p_{\mathrm{T}}^{\mathrm{V}}$ and the high $p_{\mathrm{T}}^{\mathrm{V}}$ cases are analyzed separately. The μ and σ obtained from are reported in the inclusive case, in case one b jet contains a lepton and in case no lepton is reconstructed inside the jet.

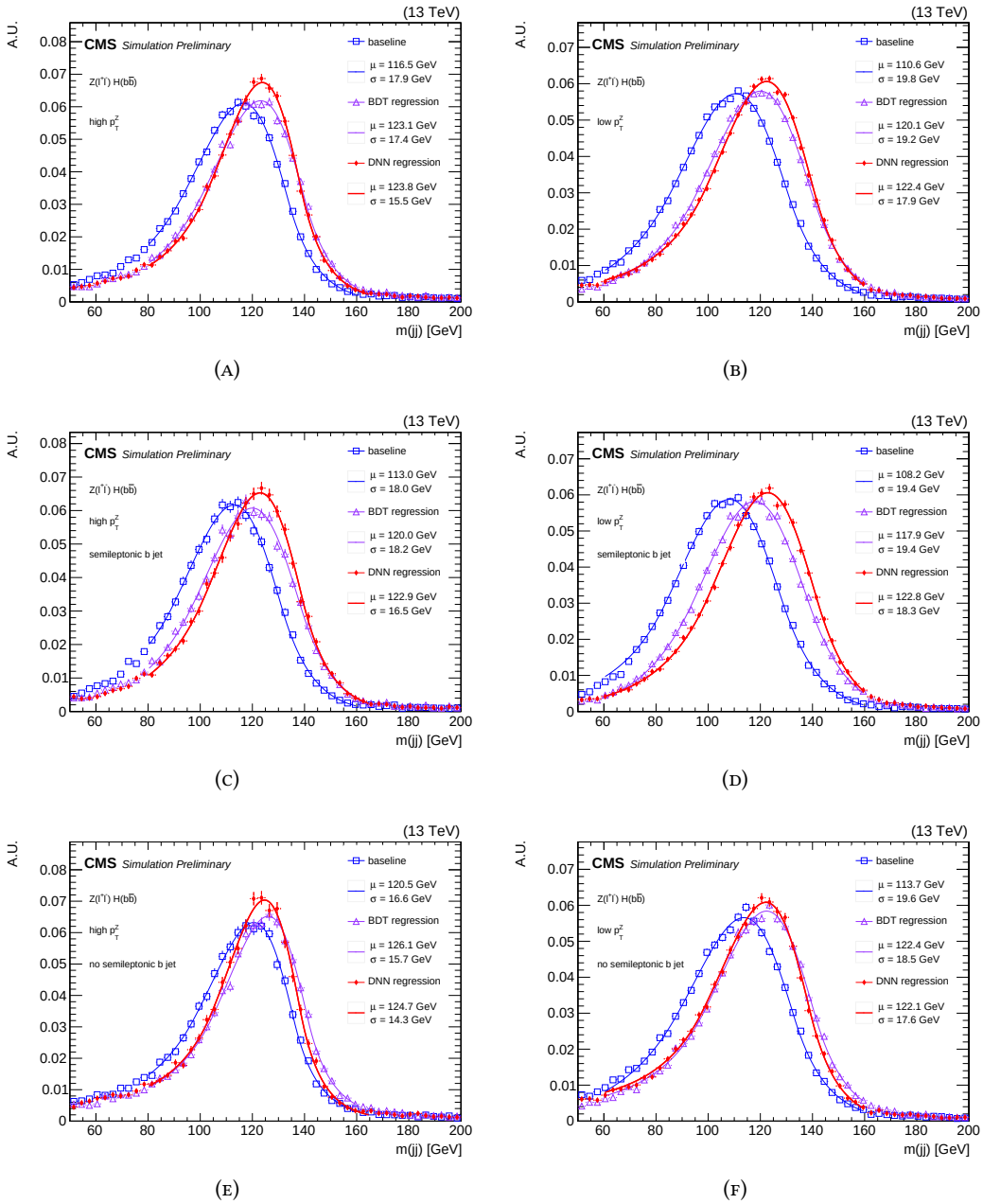
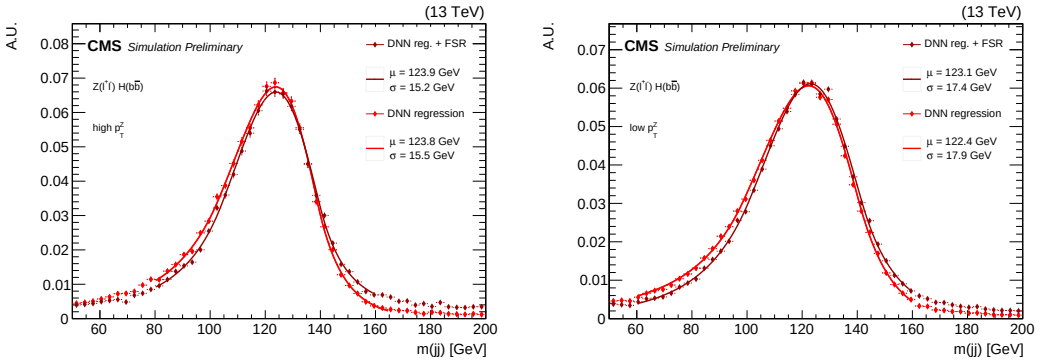


FIGURE C.1: DNN based regression applied to dijet mass of the $Z(\ell\ell)H(b\bar{b})$ events after the 2 lepton channel selection. The regression is also tested separately for semileptonic (C,D) and no semilpetonic (E,F) decays inside the b jets.

C.2 Z($\ell\ell$)H($b\bar{b}$) FSR recovery

Additional jets selected among those within $\Delta R < 0.8$ of either Higgs candidate b jet and passing the $p_T > 20$ GeV and $|\eta| < 3.0$ selection cuts are added as FSR to the invariant mass computation. The net effect is shown in figure C.2.



(A) (B)

FIGURE C.2: FSR recovery applied to the dijet mass of in Z($\ell\ell$)H($b\bar{b}$) events after the 2 lepton channel selection.

Appendix D

DNN training for VBF H $\rightarrow \mu\mu$

This appendix describes the training procedure of the DNN used for the H $\rightarrow \mu\mu$ search in the VBF channel. Similarly to the VH analysis, multivariate techniques are employed to obtain the best possible discriminating power. In this analysis the mass resolution is much better ($\sim$3%), as the Higgs boson decays into muons and not to quarks. However, the total number of events expected is much lower, making a peak search over a smooth background less sensitive. Multivariate techniques allow the optimal usage of the full event topology and isolate a handful signal like events, thus maximizing the signal significance.

The training of the DNN was performed in parallel with the training of a Bosted Decision Tree (BDT). The DNN exploits feed-forward architechtures and is trained with the KERAS package [112] and Tensorflow backend [113].

The signal region event selection, which defines the phase space where the multivariate analysis is performed, can be summarized as follows. Single muon triggers are employed. The offline selection requires two opposite charge isolated muons with p_T > 30,20 GeV and $|\eta|$ < 2.4. Two p_T-leading jets, with $|\eta_j|$<4.7 are selected. The selected jets are required to have p_T > 35,25 GeV, a rapidity gap $\Delta\eta_{jj}$ > 2.5. The signal region is defined applying the condition $|m_{\mu\mu}$-125| < 10 GeV.

Two control regions (Z region and sideband) orthogonal to the signal region are also defined by selections on $m_{\mu\mu}$: the sideband contains events with 110< $m_{\mu\mu}$<115 GeV or 135< $m_{\mu\mu}$<150 GeV; events satisfying the condition $|m_{\mu\mu}$-91|<15 GeV belong to Z region.

D.1 Training setup

In the VBF H$\rightarrow \mu\mu$ analysis a Deep Neural Network (DNN) has been optimized in parallel with the BDT. The inputs of the BDT were chosen performing a scan over a large number of input features.
The DNN inputs were chosen based on the BDT optimization, but using a DNN allowed us to seek extra improvement by adding inputs based on previous analysis experience. Moreover, several training setups and architectures were tested in order to leverage on the flexibility of the DNNs compared to the BDTs.

The training was performed using simulated samples for the three data taking years, 2016, 2017 and 2018, all mixed together. This allows us to exploits the higher statistical power of the simulation, compared to dedicated models for each year of the data taking. The variable "year" is added to the training and it serves as a flag, so that possible discrepancies due to the simulation of different data taking conditions can be taken into account.

The simulated samples used in the training are:

- the background samples:
  - DY+jets, Z($\ell\ell$), MadGraph + pythia, 105 > $m_{\ell\ell}$ > 160 GeV
  - DY+jets, Z($\ell\ell$), MadGraph + pythia, 105 > $m_{\ell\ell}$ > 160 GeV, m_{jj} > 350 GeV
  - VBF Z, Z($\ell\ell$), MadGraph + Hergwig++, 105 > $m_{\ell\ell}$ > 160 GeV
  - $t\bar{t}$, powheg + pythia

- the signal sample:
  - VBF H$\to \mu\mu$, m_H = 125 GeV, powheg + pythia.

The DY+jets samples used in the training are generated at LO precision, and replaced with a NLO sample in the final evaluation. The signal and $t\bar{t}$ samples are generated at NLO precision, while the VBF Z process is generated at LO precision.

The signal and the VBF Z samples are used both in the training and in the final evaluation of the DNN. In order not to lose half of the statical power of the sample in the final evaluation, a 4-fold procedure is used at training time.
Half of the training samples are used for the training. A quarter is used for the validation, i.e. to choose the best performing model. The model is evaluated at test time, i.e. for the final fit and extraction of the significance, on the remaining quarter of the events.

A schematic representation of the k-fold procedure with 4 folds is shown if figure D.1.

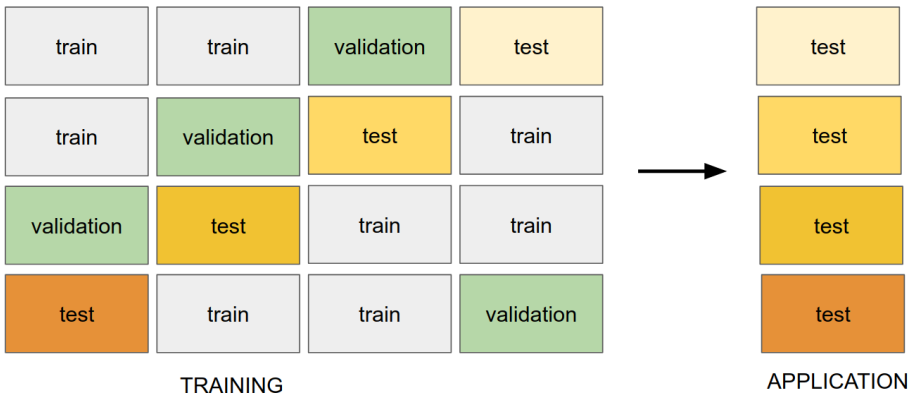


Figure D.1: Scheme of the 4-fold training, validation and evaluation procedure

The procedure is performed for all the samples used in the training, including the ones which are not used as test samples. This allows the partial regularization of the final result, as the fluctuations of single subsets of events are reduced by a factor 2. For the evaluation on data and on the simulated samples not used at training time, the 4 DNNs are computed and the final discriminator is built using the average DNN histogram for each sample.

The events are weighted at training time by applying the event weights used in the analysis. The weights are then divided by the average weight of each class (signal and background), in order to obtain weights of order 1 and the relative weights are consistent in the background class. The top events ($t\bar{t}$, powheg + pythia) are down-weighted, as they have larger

weights compared to the other background simulations, and they are a relatively easy background for the final discriminator.

In the best performing setup, the training is performed in multiple steps and with a suitable architecture.
Four networks are first optimized independently with different inputs and for different tasks. The outputs of the last hidden layer nodes, are then merged and combined to solve the actual classification problem. The final stage of the training consists of fine-tuning the model by unfreezing some of the upstream layers and training them together with the downstream ones.
The loss function used at each step is the binary cross-entropy or "log-loss", which results in the maximum likelihood estimator for binary classification problems.

A schematic representation of the DNN architecture is shown in figure D.2. The grey block indicates the DNNs optimized for single tasks, with their output in blue. The last hidden layer outputs for the 4 networks are merged in as a single vector and used as input for a combination, whose output is shown in red. The preliminary tasks of the training are aimed at optimizing single backgrounds rejection and exploiting the event topology independently of the mass. The four preliminary steps are:

- (1) signal -vs- VBF Z
- (2) signal -vs- DY
- (3) mass independent signal -vs- background
- (4) mass + mass resolution (a quick pre-training is performed just as input to the combination)

The combination step (5.a) uses all the information coming from the networks (1,2,3), which are frozen. The network (4) weights are left unfrozen at this stage. A fine-tuning (5.b) is then performed, where the weights of the network (3) are also unfrozen. All the stages use a minibatch size of 1024 events, while in the final step (5.b) a few epochs have the same minibatch size, and the very final epochs have a 10240 events minibatch size.
Each network in grey is made of 2 to 3 hidden layers with a few tenths of nodes in each hidden layer and has a pyramidal architecture. A 20% dropout is used after each hidden layer in order to regularize the model. The learning rate is also gradually decreased at training time based on the validation loss, thus preventing overfitting.

The input variables used are:

- $m_{\mu\mu}$, $\Delta m_{\mu\mu}$, $\Delta m_{\mu\mu}$ relative - the dimuon mass and the absolute and relative mass resolutions
- m_{jj}, $\log m_{\mathrm{jj}}$ - the dijet mass and its logarithm for the VBF tagging jets
- $R(p_{\mathrm{T}})$, defined as
$$\frac{|\vec{p_{\mathrm{T}}}^{\mathrm{jj}} + \vec{p_{\mathrm{T}}}^{\mu\mu}|}{|\vec{p_{\mathrm{T}}}^{\mathrm{j_1}}| + |\vec{p_{\mathrm{T}}}^{\mathrm{j_2}}| + |\vec{p_{\mathrm{T}}}^{\mu\mu}|},$$
where the p_{T} vectors of the dimuon system, of the dijet system and of the two VBF jets $\mathrm{j_1}$ and $\mathrm{j_2}$ are used

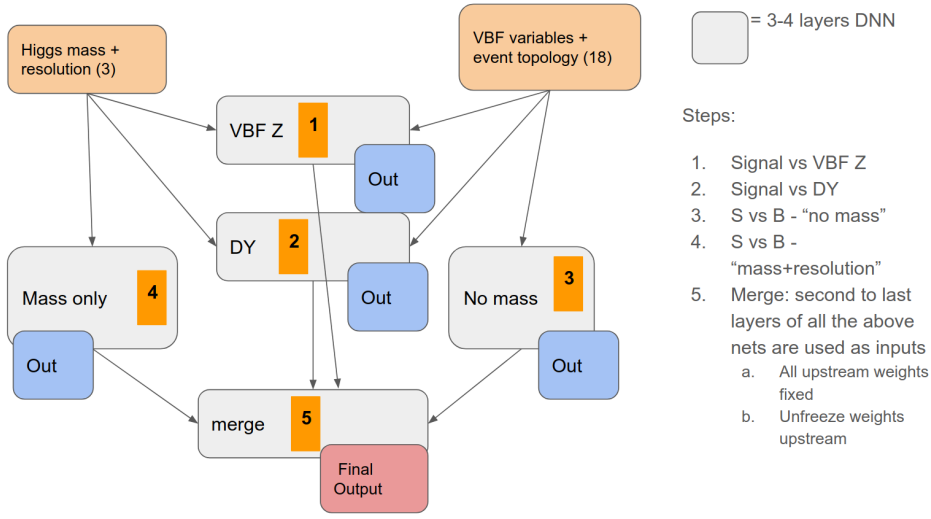


FIGURE D.2: Schematic representation of the DNN architecture: the training procedure consists in optimizing for single tasks, combining the outputs and fine-tuning the model by unfreezing upstream weights with appropriate learning rate.

- the Zeppenfeld variable $z^\star$, defined as

$$z^\star = \frac{y^\star}{|\, y_{j_1} - y_{j_2} \,|},$$

  where

$$y^\star = y_{\mu\mu} - \frac{y_{j_1} + y_{j_2}}{2}$$

  and $y_{\mu\mu}$, y_{j_1} and y_{j_2} are the rapidities of the dimuon system and the two VBF jets
- $\Delta\eta_{jj}$ - the pseudorapidity difference between the 2 VBF jets
- N_5^{soft} - counter of soft activity jets with $p_T > 5$ GeV
- H_T^{soft} - H_T for the above mentioned soft activity jets
- $\min_j \Delta\eta(\mu\mu, j)$ - the minimum η difference between a VBF jet and the dimuon system
- $p_T^{\mu\mu}$, $\log p_T^{\mu\mu}$, $\eta_{\mu\mu}$ - the dimuon 4-vector components
- $p_T^{j_1}$, $p_T^{j_2}$, η_{j_1}, η_{j_2}, ϕ_{j_1}, ϕ_{j_2} - the VBF jets' 4-vectors components
- QGL_{j_1}, QGL_{j_1} - the the quark-gluon likelihood discriminators for the VBF jets.
- the cosine of the θ angle and ϕ angle in the Collins-Soper reference frame.

The Collins-Soper reference frame is the rest frame of the dimuon pair: θ is the angle between one muons and the bisector of the proton minus the other proton direction, while ϕ is the angle between the dimuon plane and the the plane spanned by the protons' directions.

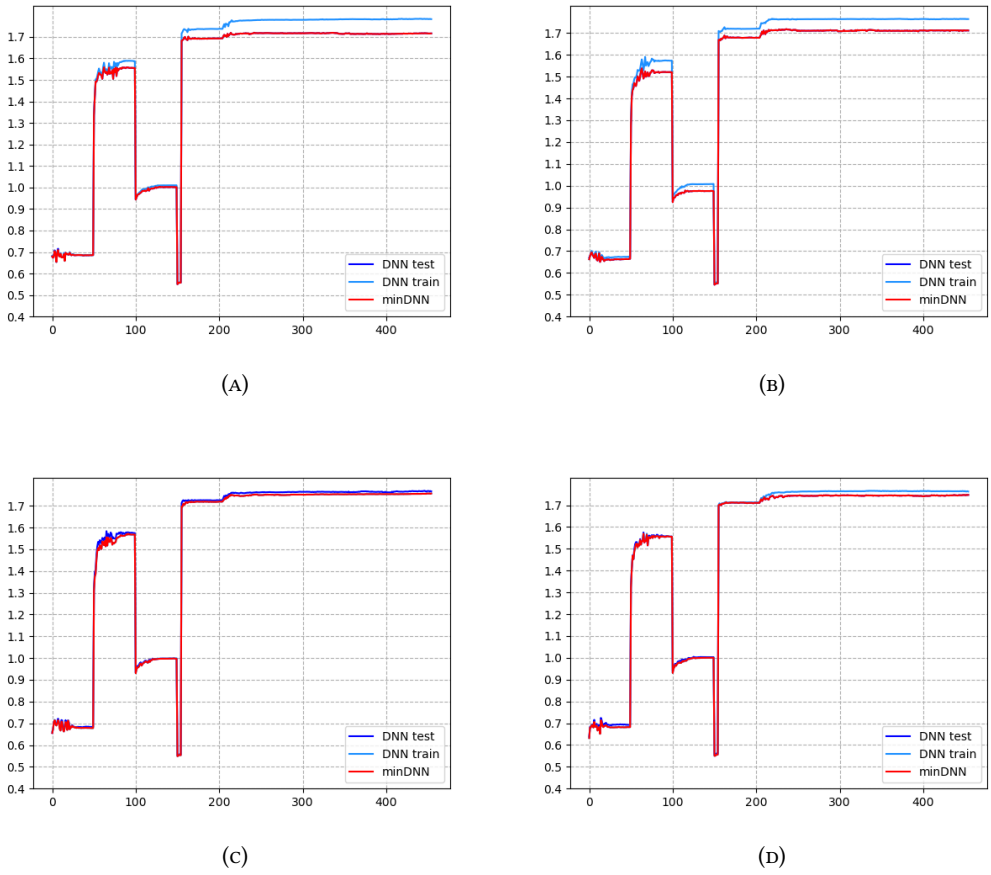
The first three variables are the ones used by the network (4) as shown in figure D.2, while the remaining 22 together with the year are used in the stage (3). All the variables are employed when training against single backgrounds (1,2), and in the for the output classifier (5.a and 5.b).

Before being fed to the DNN all the inputs are standardized. The sample mean is subtracted from each value and the result is divided by the sample standard deviation: as a result, the new input distributions have mean 0 and standard deviation 1.

The best trained model is chosen using the estimated significance in simulation. The significance is computed both for the training and validation fold, and the minimum significance is used to pick the best model. The significance is evaluated using the "asimov" significance

$$\sqrt{2\left((s+b)\cdot\log(1+s/b)-s\right)},$$

where s and b are the signal and background yields in bins containing 0.5 expected signal events. The significances of single bins are then summed in quadrature. The significances for each epoch are shown in figure D.3: the training (light blue) and validation (blue) significance per training epoch are shown. The minimum significance (red) is the estimator used to pick the best performing model. The steps in the plots are due to the different performances in the 5 training steps.



(A) (B)

(C) (D)

FIGURE D.3: Plot of the significance estimator versus training epoch for the 4 folds used in the training. The simulated samples of 2016, 2017 and 2018 are used all together. The significance is shown for training samples (light blue) and for validation samples (blue). The minimum significance (red) is the estimator used to pick the best performing model.

D.2 Training results

A comparison of the DNN performances with respect to the BDT ones are shown by the ROC curves in figure D.4. The test fold is used here. The DNN is found to provide a $\sim 5\%$ better efficiency in the most sensitive analysis phase space, i.e. for background efficiency between 1 and 0.1%.

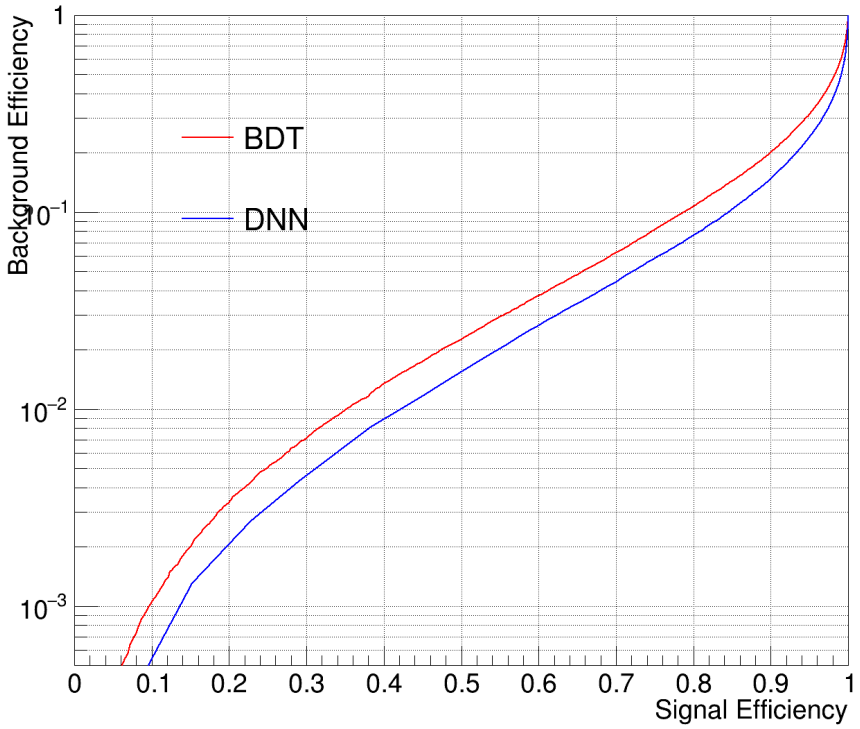


FIGURE D.4: ROC curves comparing the signal versus background efficiency for the BDT, the DNN trained with 2018 samples only and several versions of the DNN trained on the 2016+17+18 simulation.

A comparison of the DNN output in data and simulation is shown in figure D.5. The discriminator output is shown in the signal region, which is centered around $m_{\mu\mu} = 125$ GeV, and in two control regions: the mass sideband and region centered around the Z boson mass. The distributions are shown by data taking year. The DNN output in the signal region together with the sideband is currently planned to be used in the final fit.
Notably, in the signal region the test fold is used at this point, after the training and validation folds were used for training and choosing the best model.

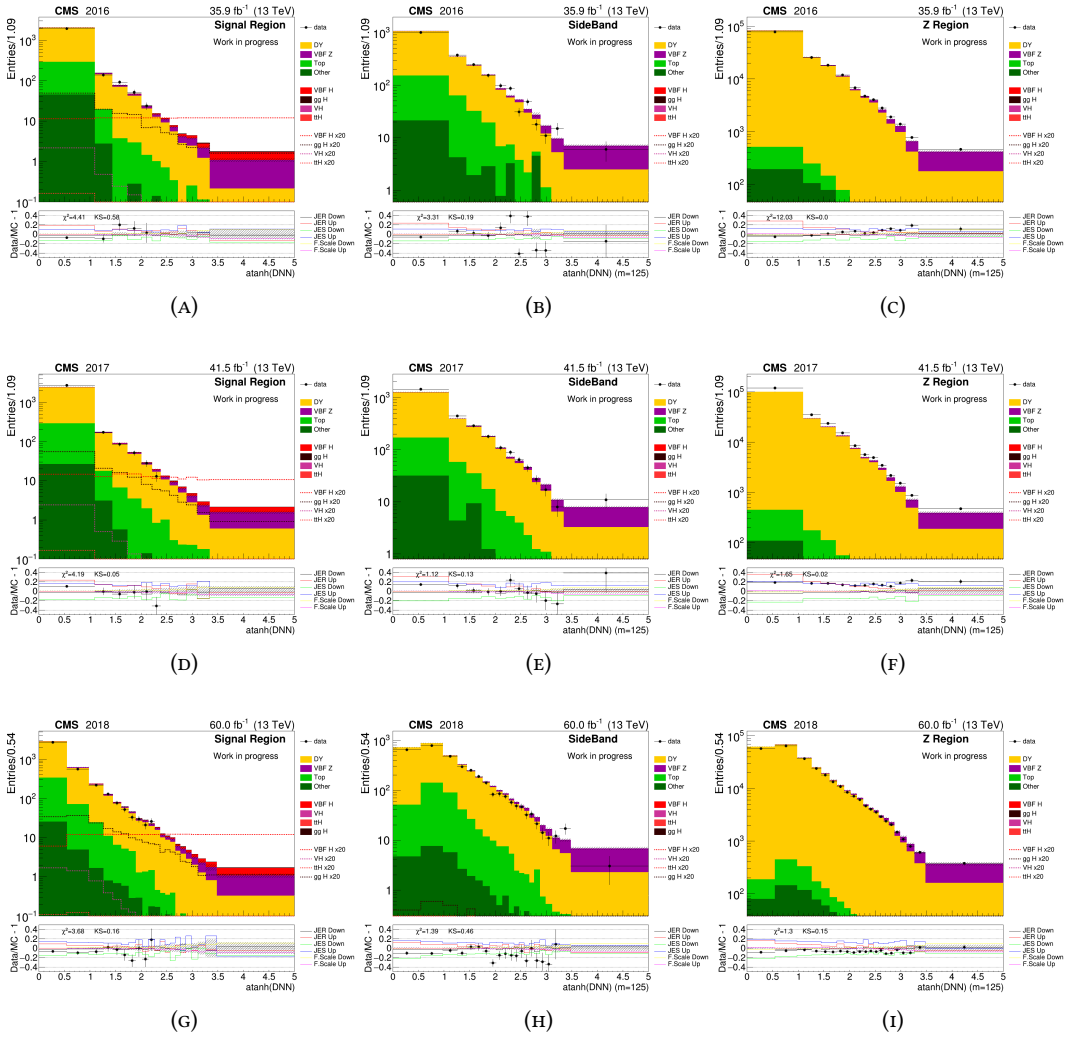


FIGURE D.5: DNN distribution for each data taking year in the signal region (A,D,G), in the dimuon mass sideband region (B,E,H) and in the dimuon mass region around the Z peak (C,F,I). The DNN score was computed with the dimuon mass fixed at 125 in the sideband and in the Z peak regions. The DNN distribution in the signal region together with the DNN in the sideband are currently planned to be used in the final fit.

Appendix E

Higgs boson physics perspectives

In this appendix some of the projected results for the Higgs boson properties are summarized. The appendix is related to the results described in chapter 1 and has a similar structure.

E.1 Future perspectives

The measurement of the couplings is a target of the LHC Run 3 and beyond. The statistical uncertainties on the current measurements will be reduced thanks to the large amount of data expected to be collected in Run3 (300 fb^{-1}) and at the High Luminosity LHC (HL-LHC) (3000 fb^{-1}). The systematic uncertainties that are already limiting some of the current measurements are both theoretical and experimental. Improvements are expected by the end of the HL-LHC Run, however, the systematic uncertainties are expected to be the main limiting factor at that point.

The large amount of data will also allow the measurement of differential cross sections with good precision. The measurement of the Higgs boson differential cross sections can provide constraints on physical parameters that have a small effect on inclusive quantities, but cause larger deviations in specific phase space regions.

A very general framework that accommodates the currently known SM Physics and possible deviations due to new physics, is the so called Effective Field Theory (EFT) framework. In that model the SM is seen as the low energy approximation of a more fundamental theory, which would become apparent at a mass scale Λ beyond the energy ranges currently investigated. The SM Lagrangian contains only terms that have dimension $D \leq 4$ in energy units. In EFT higher-dimension operators are introduced.

The EFT Lagrangian takes the form:

$$\mathcal{L} = \mathcal{L}_{\text{SM}} + \sum_i \frac{c_i^{(5)}}{\Lambda} \mathcal{O}_i^{(5)} + \sum_i \frac{c_i^{(6)}}{\Lambda^2} \mathcal{O}_i^{(6)} + \sum_i \frac{c_i^{(7)}}{\Lambda^3} \mathcal{O}_i^{(7)} + \ldots$$

The higher-dimensional operators are organized in a systematic expansion in D, where each consecutive term is suppressed by a larger power of Λ. where each $\mathcal{O}_i^{(D)}$ is a gauge invariant operator of dimension D and the parameters $c_i^{(D)}$ multiplying the operators in the Lagrangian are called the Wilson coefficients. The effects of the new physics are therefore expected to be suppressed by a factor $c^{(D)}/\Lambda^{D-4}$. For each power D a complete set of gauge invariant operators can be built, providing a general parametrization of the new physics effects which can be probed as deviations from the SM predictions.

This framework can be used also to combine in a global fit the electroweak precision measurements and the Higgs results. Global combined fits are performed in the high energy physics community by the GFitter and HEPFit groups [139, 140].

E.1.1 Precision measurements outlook

The LHC coupling measurements will reach a precision of 1 – 5%. Experimental uncertainties can be reduced to a certain extent, but the precision is going to be affected also by the theoretical predictions.

In order to have an idea of the precision achievable by LHC experiments in Run 3 and beyond one can take as a reference [141]. Results based on data collected at 13 TeV are projected to the Run 3 and to the HL-LHC luminosities.
Two different scenarios are considered for the systematic uncertainties:

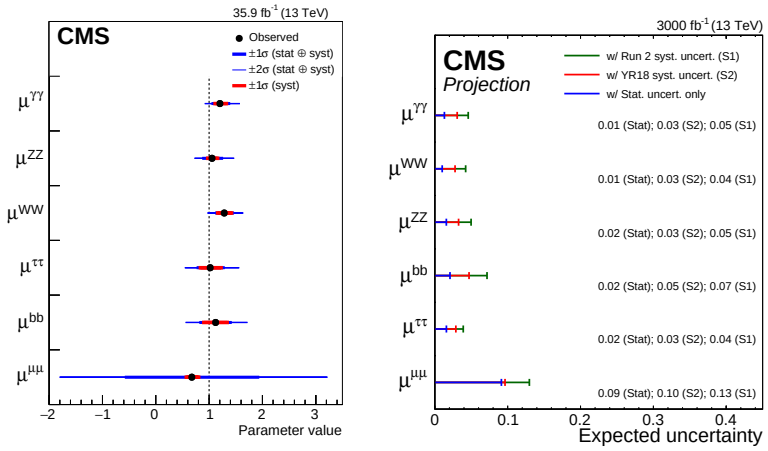
- Scenario 1 (S1), where systematic uncertainties are kept constant with integrated luminosity. The performance of the CMS detector is assumed to be unchanged with respect to the reference analysis;
- Scenario 2 (S2): Theoretical uncertainties are scaled by a factor 1/2, while experimental systematic uncertainties are scaled with $1/\sqrt{L}$ with a lower boundary due to the expected detector performance.

The statistical uncertainty due to the size of the simulated samples is ignored, the detector performance is assumed not to deteriorate with respect to Run 2 in both scenarios.
Figures E.1 and E.2 show the input analyses combined results with 2016 data on the left and the expected projections with the uncertainties in both scenarios. The total uncertainties are decreased for each decay channel and final state component.

It should be however highlighted that all the HL-LHC measurements performed with the current methods and knowledge are going to be mainly affected by systematic uncertainties, both theoretical and experimental. The statistical only component is also plotted in the right plots for comparison. A reduction of the uncertainties below ∼1% will be possible only at e^+e^- colliders, where model independent measurements can be performed. This is in fact a very large physics program and one of the possible future directions for high energy physics, but out of the scope of this thesis.

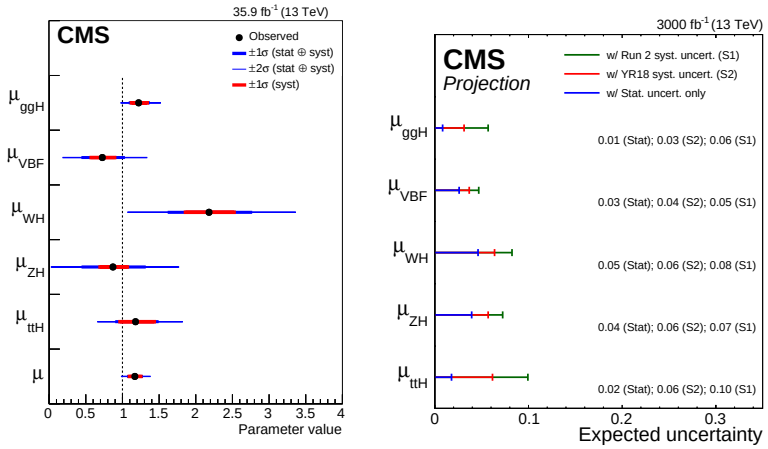
Together with inclusive measurements, differential cross section will be probed. Both the ATLAS and CMS Collaborations have already performed differential measurements of the Higgs boson production cross sections at $\sqrt{s}$ = 8 and 13 TeV (see for example reference [142]). Differential cross sections are measured at 13 TeV in the following variables: $p_{\mathrm{T}}^{\mathrm{H}}$, the transverse momentum of the Higgs boson, its rapidity $|y_{\mathrm{H}}|$, the number of hadronic jets N_{jets} , and the transverse momentum of the leading hadronic jet. The decay channels used are mainly the $\mathrm{H} \to \gamma\gamma$ and $\mathrm{ZZ}^{(*)} \to 4\ell$ decay channels, but the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay channel can also be exploited to probe the very high transverse momentum tail of the Higgs production spectrum. Projections in the differential cross sections measurements can also be found in [141].

The projections for the width using the method as in [40] with the full HL-LHC luminosity predict Γ_H to be measured in the interval [2.0-6.0] MeV at 95% confidence level in the most optimistic scenario S2, as shown in figure E.3 (B). The projection can be compared with the most sensitive measurement of the width with the current amount of data, which uses the 4 lepton channel with 2017 and 2016 integrated luminosity E.3 (A).



(A) (B)

FIGURE E.1: Summary plot showing the signal measured signal streght per decay channel at $\sqrt{s}$ = 13TeV (A) and the projected uncertainty with 3000 fb^{-1} (B). In the projections the $\pm 1\sigma$ uncertainties in two scenarios are shown. The statistical-only component of the uncertainty is also shown.



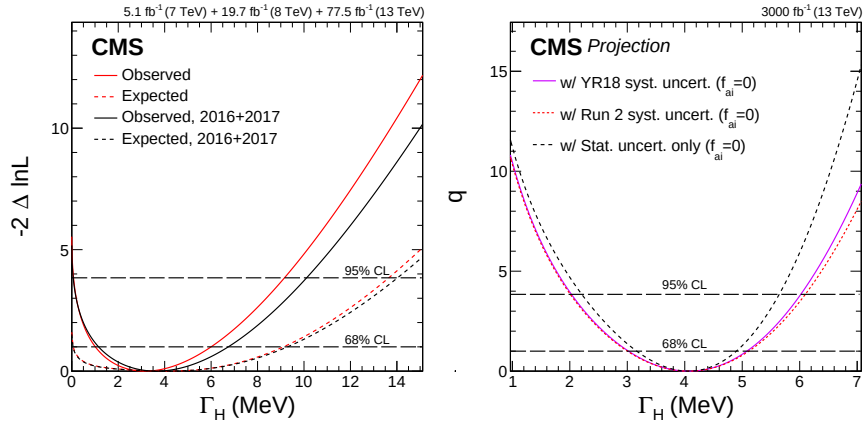
(A) (B)

FIGURE E.2: Summary plot showing the signal measured signal streght per production mode at $\sqrt{s}$ = 13TeV (A) and the projected uncertainty with 3000 fb^{-1} (B) . In the projections the $\pm 1\sigma$ uncertainties in two scenarios are shown. The statistical-only component of the uncertainty is also shown.

E.1.2 Self-coupling

Another important measurement in the context of Higgs Physics is the self-coupling, which is predicted in the SM by the Higgs potential $V(\Phi)$. The self-coupling depends on the parameter λ, which can be predicted once the Higgs mass and the vacuum expectation value v are known. However, a direct measurement could provide insight on the Higgs potential, which could differ from the simple hypothesis in several well-motivated scenarios of new physics.

The self-coupling can be measured in Higgs pair production, but it's currently out of the reach at the LHC due to the lower cross section. After the HL-LHC, with a combination of both ATLAS and CMS data the projections indicate a 4σ expected significance[143]. Several improvements are expected in the meantime, but a precise measurement of the Higgs pair production will be a very hard task for LHC experiments. For this measurement, a e^+e^-



(A) (B)

FIGURE E.3: Likelihood scans Γ_H. The left plot (A) presents results both combining Run 1 and Run2 data (in red) and usign Run 2 data only (black). For The projection (B) to 3000 fb^{-1} two scenarios are considered for the systematic uncertainties.

machine wouldn't give large gains, again due to the very low cross sections, but a higher energy hadron collider or a muon collider would be more suitable.

Bibliography

[1] *Standard Model (Wikipedia, the free encyclopedia).* URL: https://en.wikipedia.org/wiki/Standard_Model#/media/File:Standard_Model_of_Elementary_Particles.svg.

[2] S. L. Glashow. "Partial Symmetries of Weak Interactions". In: *Nucl. Phys.* 22 (1961), pp. 579–588. DOI: 10.1016/0029-5582(61)90469-2.

[3] Steven Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

[4] Abdus Salam. "Weak and Electromagnetic Interactions". In: *Conf. Proc.* C680519 (1968), pp. 367–377.

[5] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Phys. Rev. Lett.* 13 (1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[6] Peter W. Higgs. "Broken symmetries, massless particles and gauge fields". In: *Phys. Lett.* 12 (1964), pp. 132–133. DOI: 10.1016/0031-9163(64)91136-9.

[7] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Phys. Rev. Lett.* 13 (1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[8] Peter W. Higgs. "Spontaneous Symmetry Breakdown without Massless Bosons". In: *Phys. Rev.* 145 (1966), pp. 1156–1163. DOI: 10.1103/PhysRev.145.1156.

[9] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. "Global Conservation Laws and Massless Particles". In: *Phys. Rev. Lett.* 13 (1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585.

[10] T. W. B. Kibble. "Symmetry breaking in nonAbelian gauge theories". In: *Phys. Rev.* 155 (1967), pp. 1554–1561. DOI: 10.1103/PhysRev.155.1554.

[11] Yoichiro Nambu and G. Jona-Lasinio. "Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. 1." In: *Phys. Rev.* 122 (1961), pp. 345–358. DOI: 10.1103/PhysRev.122.345.

[12] Murray Gell-Mann and M Levy. "The axial vector current in beta decay". In: *Nuovo Cim.* 16 (1960), p. 705. DOI: 10.1007/BF02859738.

[13] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Phys. Lett.* B716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[14] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Phys. Lett.* B716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[15] CMS Collaboration. "Observation of a New Boson with Mass Near 125 GeV in pp Collisions at $\sqrt{s}$ = 7 and 8 TeV". In: *JHEP* 06 (2013), p. 081. DOI: 10.1007/JHEP06(2013)081. arXiv: 1303.4571 [hep-ex].

[16] LHC Higgs Cross Section Working Group et al. "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties". In: *CERN-2013-004* (CERN, Geneva, 2013). arXiv: 1307.1347 [hep-ph].

[17] D. de Florian et al. "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". In: (2016). DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph].

[18] CMS Collaboration. "Observation of Higgs boson decay to bottom quarks". In: *Phys. Rev. Lett.* 121.12 (2018), p. 121801. DOI: 10.1103/PhysRevLett.121.121801. arXiv: 1808.08242 [hep-ex].

[19] ATLAS Collaboration. "Measurements of Higgs boson production and couplings in the four-lepton channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector". In: *Phys. Rev.* D91.1 (2015), p. 012006. DOI: 10.1103/PhysRevD.91.012006. arXiv: 1408.5191 [hep-ex].

[20] CMS Collaboration. "Measurement of the properties of a Higgs boson in the four-lepton final state". In: *Phys. Rev.* D89.9 (2014), p. 092007. DOI: 10.1103/PhysRevD.89.092007. arXiv: 1312.5353 [hep-ex].

[21] ATLAS Collaboration. "Observation and measurement of Higgs boson decays to WW* with the ATLAS detector". In: *Phys. Rev.* D92.1 (2015), p. 012006. DOI: 10.1103/PhysRevD.92.012006. arXiv: 1412.2641 [hep-ex].

[22] CMS Collaboration. "Measurement of Higgs boson production and properties in the WW decay channel with leptonic final states". In: *JHEP* 01 (2014), p. 096. DOI: 10.1007/JHEP01(2014)096. arXiv: 1312.1129 [hep-ex].

[23] CMS Collaboration. "Observation of the diphoton decay of the Higgs boson and measurement of its properties". In: *Eur. Phys. J.* C74.10 (2014), p. 3076. DOI: 10.1140/epjc/s10052-014-3076-z. arXiv: 1407.0558 [hep-ex].

[24] ATLAS Collaboration. "Measurement of Higgs boson production in the diphoton decay channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector". In: *Phys. Rev.* D90.11 (2014), p. 112015. DOI: 10.1103/PhysRevD.90.112015. arXiv: 1408.7084 [hep-ex].

[25] CMS Collaboration. "Study of the Mass and Spin-Parity of the Higgs Boson Candidate Via Its Decays to Z Boson Pairs". In: *Phys. Rev. Lett.* 110.8 (2013), p. 081803. DOI: 10.1103/PhysRevLett.110.081803. arXiv: 1212.6639 [hep-ex].

[26] ATLAS Collaboration. "Evidence for the spin-0 nature of the Higgs boson using ATLAS data". In: *Phys. Lett.* B726 (2013), pp. 120–144. DOI: 10.1016/j.physletb.2013.08.026. arXiv: 1307.1432 [hep-ex].

[27] CMS Collaboration. "Constraints on anomalous HVV interactions using H to 4l decays". In: CMS-PAS-HIG-14-014 (2014). URL: https://cds.cern.ch/record/1728251.

[28] ATLAS Collaboration. "Combined Measurement of the Higgs Boson Mass in *pp* Collisions at $\sqrt{s}$ = 7 and 8 TeV with the ATLAS and CMS Experiments". In: *Phys. Rev. Lett.* 114 (2015), p. 191803. DOI: 10.1103/PhysRevLett.114.191803. arXiv: 1503.07589 [hep-ex].

[29] ATLAS Collaboration. "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC *pp* collision data at $\sqrt{s}$ = 7 and 8 TeV". In: (2016). arXiv: 1606.02266 [hep-ex].

[30] ATLAS Collaboration. "Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector". In: *Phys. Rev.* D99 (2019), p. 072001. DOI: 10.1103/PhysRevD.99.072001. arXiv: 1811.08856 [hep-ex].

[31] CMS Collaboration. "Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector". In: *Phys. Lett.* B779 (2018), pp. 283–316. DOI: 10.1016/j.physletb.2018.02.004. arXiv: 1708.00373 [hep-ex].

[32] CMS Collaboration. "Observation of $t\bar{t}H$ production". In: *Phys. Rev. Lett.* 120.23 (2018), p. 231801. DOI: 10.1103/PhysRevLett.120.231801. arXiv: 1804.02610 [hep-ex].

[33] ATLAS Collaboration. "Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector". In: *Phys. Lett.* B784 (2018), pp. 173–191. DOI: 10.1016/j.physletb.2018.07.035. arXiv: 1806.00425 [hep-ex].

[34] ATLAS Collaboration. "Evidence for the $H \to b\bar{b}$ decay with the ATLAS detector". In: *JHEP* 12 (2017), p. 024. DOI: 10.1007/JHEP12(2017)024. arXiv: 1708.03299 [hep-ex].

[35] CMS Collaboration. "Evidence for the Higgs boson decay to a bottom quark–antiquark pair". In: *Phys. Lett.* B780 (2018), pp. 501–532. DOI: 10.1016/j.physletb.2018.02.050. arXiv: 1709.07497 [hep-ex].

[36] ATLAS Collaboration. "Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector". In: *Phys. Lett.* B786 (2018), pp. 59–86. DOI: 10.1016/j.physletb.2018.09.013. arXiv: 1808.08238 [hep-ex].

[37] ATLAS Collaboration. "Measurement of the Higgs boson mass in the $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ channels with $\sqrt{s}$ = 13 TeV pp collisions using the ATLAS detector". In: *Phys. Lett.* B784 (2018), pp. 345–366. DOI: 10.1016/j.physletb.2018.07.050. arXiv: 1806.00242 [hep-ex].

[38] CMS Collaboration. "Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s}$ = 13 TeV". In: *JHEP* 11 (2017), p. 047. DOI: 10.1007/JHEP11(2017)047. arXiv: 1706.09936 [hep-ex].

[39] CMS Collaboration. "Combined measurements of Higgs boson couplings in proton–proton collisions at $\sqrt{s}$ = 13 TeV". In: *Eur. Phys. J.* C79.5 (2019), p. 421. DOI: 10.1140/epjc/s10052-019-6909-y. arXiv: 1809.10733 [hep-ex].

[40] CMS Collaboration. "Measurements of the Higgs boson width and anomalous HVV couplings from on-shell and off-shell production in the four-lepton final state". In: *Phys. Rev.* D99.11 (2019), p. 112003. DOI: 10.1103/PhysRevD.99.112003. arXiv: 1901.00174 [hep-ex].

[41] Lyndon Evans and Philip Bryant. "LHC Machine". In: *JINST* 3 (2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[42] Jean-Luc Caron. "Cross section of LHC dipole.. Dipole LHC: coupe transversale." AC Collection. Legacy of AC. Pictures from 1992 to 2002. 1998. URL: https://cds.cern.ch/record/841539.

[43] Fabienne Marcastel. "CERN's Accelerator Complex. La chaîne des accélérateurs du CERN". In: (2013). General Photo. URL: https://cds.cern.ch/record/1621583.

[44] *CMS Luminosity public plots*. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[45] CMS Collaboration. "CMS Physics". In: (2006). URL: https://cds.cern.ch/record/922757.

[46] CMS Collaboration. "The CMS Experiment at the CERN LHC". In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[47] J Butler et al. *CMS Phase II Upgrade Scope Document*. Tech. rep. CERN-LHCC-2015-019. LHCC-G-165. Geneva: CERN, 2015. URL: https://cds.cern.ch/record/2055167.

[48] CMS Collaboration. *CMS, tracker technical design report*. 1998. URL: http://weblib.cern.ch/abstract?CERN-LHCC-98-6.

[49] CMS Collaboration. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". In: *JINST* 9.10 (2014), P10009. DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569 [physics.ins-det].

[50] David Aaron Matzner Dominguez et al. "CMS Technical Design Report for the Pixel Detector Upgrade". In: (2012). DOI: 10.2172/1151650.

[51] CMS Collaboration. "CMS: The electromagnetic calorimeter. Technical design report". In: (1997). URL: https://cds.cern.ch/record/349375.

[52] CMS Collaboration. "CMS: The hadron calorimeter technical design report". In: (1997). URL: http://cds.cern.ch/record/357153.

[53] J. Mans et al. "CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter". In: (2012). DOI: 10.2172/1151651.

[54] CMS Collaboration. "CMS, the Compact Muon Solenoid. Muon technical design report". In: (1997). URL: https://cds.cern.ch/record/343814.

[55] D. Abbaneo et al. "A GEM Detector System for an Upgrade of the High-eta Muon Endcap Stations GE1/1 + ME1/1 in CMS". In: (2012). arXiv: 1211.1494 [physics.ins-det].

[56] S. Dasu et al. "CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems". In: Technical Design Report CMS (). URL: https://cds.cern.ch/record/706847.

[57] W. Adam et al. "The CMS high level trigger". In: *Eur. Phys. J.* C46 (2006), pp. 605–667. DOI: 10.1140/epjc/s2006-02495-8. arXiv: hep-ex/0512077 [hep-ex].

[58] CMS Collaboration. "2017 tracking performance plots". In: (2017). URL: https://cds.cern.ch/record/2290524.

[59] R. Fruhwirth. "Application of Kalman filtering to track and vertex fitting". In: *Nucl. Instrum. Meth.* A262 (1987), pp. 444–450. DOI: 10.1016/0168-9002(87)90887-4.

[60] Thomas Speer et al. *Vertex Fitting in the CMS Tracker*. Tech. rep. CMS-NOTE-2006-032. Geneva: CERN, 2006. URL: https://cds.cern.ch/record/927395.

[61] K. Rose. "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems". In: *IEEE Proc.* 86.11 (1998), pp. 2210–2239. DOI: 10.1109/5.726788.

[62] R. Fruhwirth, W. Waltenberger, and P. Vanlaer. "Adaptive vertex fitting". In: *J. Phys.* G34 (2007), N343. DOI: 10.1088/0954-3899/34/12/N01.

[63] CMS Collaboration. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det].

[64] W. Adam et al. "Reconstruction of electrons with the Gaussian sum filter in the CMS tracker at LHC". In: *eConf* C0303241 (2003). [J. Phys.G31,N9(2005)], TULT009. DOI: 10.1088/0954-3899/31/9/N01. arXiv: physics/0306087 [physics.data-an].

[65] Peter Z. Skands. "QCD for Collider Physics". In: *Proceedings, High-energy Physics. Proceedings, 18th European School (ESHEP 2010): Raseborg, Finland, June 20 - July 3, 2010*. 2011. arXiv: 1104.2863 [hep-ph].

[66] URL: https://theory.slac.stanford.edu/our-research/simulations.

[67] S. Agostinelli et al. "GEANT4: A Simulation toolkit". In: *Nucl. Instrum. Meth.* A506 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[68] CMS Collaboration. "Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at $\sqrt{s}$ = = 8 TeV". In: *JINST* 10.06 (2015), P06005. DOI: 10.1088/1748-0221/10/06/P06005. arXiv: 1502.02701 [physics.ins-det].

[69] "Electron and Photon performance in CMS with the full 2017 data sample and additional 2016 highlights for the CALOR 2018 Conference". In: (2018). URL: https://cds.cern.ch/record/2320638.

[70] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[71] Matteo Cacciari and Gavin P. Salam. "Dispelling the N^3 myth for the k_t jet-finder". In: *Phys. Lett.* B641 (2006), pp. 57–61. DOI: 10.1016/j.physletb.2006.08.037. arXiv: hep-ph/0512210 [hep-ph].

[72] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "FastJet User Manual". In: *Eur. Phys. J.* C72 (2012), p. 1896. DOI: 10.1140/epjc/s10052-012-1896-2. arXiv: 1111.6097 [hep-ph].

[73] Stephen D. Ellis and Davison E. Soper. "Successive combination jet algorithm for hadron collisions". In: *Phys. Rev.* D48 (1993), pp. 3160–3166. DOI: 10.1103/PhysRevD.48.3160. arXiv: hep-ph/9305266 [hep-ph].

[74] M. Wobisch and T. Wengler. "Hadronization corrections to jet cross-sections in deep inelastic scattering". In: *Monte Carlo generators for HERA physics. Proceedings, Workshop, Hamburg, Germany, 1998-1999*. 1998, pp. 270–279. arXiv: hep-ph/9907280 [hep-ph].

[75] CMS Collaboration. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *JINST* 12.02 (2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014. arXiv: 1607.03663 [hep-ex].

[76] Matteo Cacciari and Gavin P. Salam. "Pileup subtraction using jet areas". In: *Phys. Lett. B* 659 (2008), p. 119. DOI: 10.1016/j.physletb.2007.09.077. arXiv: 0707.1378 [hep-ph].

[77] CMS Collaboration. *Performance of Jet Reconstruction with Charged Tracks only*. Tech. rep. CMS-PAS-JME-08-001. Geneva: CERN, 2009. URL: http://cds.cern.ch/record/1198681.

[78] CMS Collaboration. *Commissioning of TrackJets in pp Collisions at 7 TeV*. Tech. rep. CMS-PAS-JME-10-006. Geneva: CERN, 2010. URL: http://cds.cern.ch/record/1275133.

[79] CMS Collaboration. "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s}$ = 13 TeV using the CMS detector". In: *JINST* 14.07 (2019), P07004. DOI: 10.1088/1748-0221/14/07/P07004. arXiv: 1903.06078 [hep-ex].

[80] Wolfgang Waltenberger. *Adaptive Vertex Reconstruction*. Tech. rep. CMS-NOTE-2008-033. Geneva: CERN, 2008. URL: https://cds.cern.ch/record/1166320.

[81] CMS Collaboration. "Measurement of $B\bar{B}$ Angular Correlations based on Secondary Vertex Reconstruction at $\sqrt{s}$ = 7 TeV". In: *JHEP* 03 (2011), p. 136. DOI: 10.1007/JHEP03(2011)136. arXiv: 1102.3194 [hep-ex].

[82] CMS Collaboration. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *JINST* 13.05 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158 [physics.ins-det].

[83] CMS Collaboration. "Identification of b-Quark Jets with the CMS Experiment". In: *JINST* 8 (2013), P04013. DOI: 10.1088/1748-0221/8/04/P04013. arXiv: 1211.4462 [hep-ex].

[84] CMS Collaboration. *Pileup mitigation at CMS in 13 TeV data*. 2020. arXiv: 2003.00503 [hep-ex].

[85] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[86] Tom M. Mitchell. *Machine learning*. McGraw Hill series in computer science. McGraw-Hill, 1997. ISBN: 978-0-07-042807-2. URL: http://www.worldcat.org/oclc/61321007.

[87] K. Hornik, M. Stinchcombe, and H. White. "Multilayer Feedforward Networks Are Universal Approximators". In: *Neural Netw.* 2.5 (July 1989), pp. 359–366. ISSN: 0893-6080. DOI: 10.1016/0893-6080(89)90020-8. URL: http://dx.doi.org/10.1016/0893-6080(89)90020-8.

[88] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals, and Systems (MCSS)* 2.4 (Dec. 1989), pp. 303–314. ISSN: 0932-4194. DOI: 10.1007/BF02551274. URL: http://dx.doi.org/10.1007/BF02551274.

[89] *Borel function (Encyclopedia of Math)*. URL: https://www.encyclopediaofmath.org/index.php/Borel_function.

[90] Moshe Leshno et al. "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function." In: *Neural Networks* 6.6 (1993), pp. 861–867. URL: http://dblp.uni-trier.de/db/journals/nn/nn6.html#LeshnoLPS93.

[91] Guido F Montufar et al. "On the Number of Linear Regions of Deep Neural Networks". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2924–2932. URL: http://papers.nips.cc/paper/5422-on-the-number-of-linear-regions-of-deep-neural-networks.pdf.

[92] David H. Wolpert and William G. Macready. "No free lunch theorems for optimization". In: *IEEE Transactions on Evolutionary Computation* 1.1 (1997), pp. 67–82.

[93] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning Representations by Back-propagating Errors". In: *Nature* 323.6088 (1986), pp. 533–536. DOI: 10.1038/323533a0. URL: http://www.nature.com/articles/323533a0.

[94] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. 2014. URL: http://arxiv.org/abs/1412.6980.

[95] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, pp. 448–456. URL: http://dl.acm.org/citation.cfm?id=3045118.3045167.

[96] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *J. Mach. Learn. Res.* 15.1 (Jan. 2014), pp. 1929–1958. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2627435.2670313.

[97] Y. LeCun et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1 (1989), pp. 541–551.

[98] Min Lin, Qiang Chen, and Shuicheng Yan. *Network In Network*. 2013. arXiv: 1312.4400 [cs.NE].

[99] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[100] Kenji Abe et al. "Precise measurement of the b quark fragmentation function in Z0 boson decays". In: *Phys. Rev. Lett.* 84 (2000), pp. 4300–4304. DOI: 10.1103/PhysRevLett.84.4300. arXiv: hep-ex/9912058 [hep-ex].

[101] Tuija Aaltonen et al. "Search for the standard model Higgs boson decaying to a $b\bar{b}$ pair in events with one charged lepton and large missing transverse energy using the full CDF data set". In: *Phys. Rev. Lett.* 109 (2012), p. 111804. DOI: 10.1103/PhysRevLett.109.111804. arXiv: 1207.1703 [hep-ex].

[102] CMS Collaboration. "Search for the standard model Higgs boson produced through vector boson fusion and decaying to $b\bar{b}$". In: *Phys. Rev. D* 92 (2015), p. 032008. DOI: 10.1103/PhysRevD.92.032008. arXiv: 1506.01010 [hep-ex].

[103] CMS Collaboration. "Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks". In: *Phys. Rev.* D89.1 (2014), p. 012003. DOI: 10.1103/PhysRevD.89.012003. arXiv: 1310.3687 [hep-ex].

[104] Peter J. Huber. "Robust estimation of a location parameter". In: *Ann. Math. Statist.* 35 (1964), p. 731. DOI: 10.1214/aoms/1177703732.

[105] Roger W Koenker and Gilbert Bassett. "Regression Quantiles". In: *Econometrica* 46 (1978), p. 33. URL: https://EconPapers.repec.org/RePEc:ecm:emetrp:v:46:y:1978:i:1:p:33-50.

[106] John M. Campbell et al. "Top-Pair production and decay at NLO matched with parton showers". In: *JHEP* 04 (2015), p. 114. DOI: 10.1007/JHEP04(2015)114. arXiv: 1412.1828 [hep-ph].

[107] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[108] Torbjörn Sjöstrand et al. "An Introduction to PYTHIA 8.2". In: *Comput. Phys. Commun.* 191 (2015), p. 159. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph].

[109] CMS Collaboration. "Event generator tunes obtained from underlying event and multiparton scattering measurements". In: *Eur. Phys. J. C* 76 (2016), p. 155. DOI: 10.1140/epjc/s10052-016-3988-x. arXiv: 1512.00815 [hep-ex].

[110] Sergey Ioffe and Christian Szegedy. "Batch normalization: accelerating deep network training by reducing internal covariate shift". In: *Proceedings of the 32nd International Conference on Machine Learning* 37 (2015), p. 448. arXiv: 1502.03167.

[111] Andrew L. Maas et al. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". 2013.

[112] François Chollet et al. *Keras*. https://keras.io. 2015.

[113] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: http://tensorflow.org/.

[114] Jerome Friedman Trevor Hastie Robert Tibshirani. *The Elements of Statistical Learning*. 2nd. Springer-Verlag New York, 2009. DOI: 10.1007/978-0-387-84858-7.

[115] CMS Collaboration. *A deep neural network for simultaneous estimation of b jet energy and resolution*. 2019. arXiv: 1912.06046 [hep-ex].

[116] CMS Collaboration. "Determination of jet energy calibration and transverse momentum resolution in CMS". In: *JINST* 6 (2011), P11002. DOI: 10.1088/1748-0221/6/11/P11002. arXiv: 1107.4277 [physics.ins-det].

[117] "Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13TeV with Phase 1 CMS detector". In: (2018). URL: http://cds.cern.ch/record/2646773.

[118] CMS Collaboration. "Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair". In: *Phys. Rev. Lett.* 120.7 (2018), p. 071802. DOI: 10.1103/PhysRevLett.120.071802. arXiv: 1709.05543 [hep-ex].

[119] CMS Collaboration. "Search for the standard model Higgs boson decaying to bottom quarks in pp collisions at $\sqrt{s}$ = 7 TeV". In: *Phys. Lett.* B710 (2012), pp. 284–306. DOI: 10.1016/j.physletb.2012.02.085. arXiv: 1202.4195 [hep-ex].

[120] Keith Hamilton, Paolo Nason, and Giulia Zanderighi. "MINLO: Multi-Scale Improved NLO". In: *JHEP* 10 (2012), p. 155. DOI: 10.1007/JHEP10(2012)155. arXiv: 1206.3572 [hep-ph].

[121] Gionata Luisoni et al. "$HW^{\pm}$/HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO". In: *JHEP* 10 (2013), p. 083. DOI: 10.1007/JHEP10(2013)083. arXiv: 1306.2542 [hep-ph].

[122] Rikkert Frederix and Stefano Frixione. "Merging meets matching in MC@NLO". In: *JHEP* 12 (2012), p. 061. DOI: 10.1007/JHEP12(2012)061. arXiv: 1209.6215 [hep-ph].

[123] Michelangelo L. Mangano et al. "Matching matrix elements and shower evolution for top-quark production in hadronic collisions". In: *JHEP* 01 (2007), p. 013. DOI: 10.1088/1126-6708/2007/01/013. arXiv: hep-ph/0611129 [hep-ph].

[124] Richard D. Ball et al. "Parton distributions for the LHC Run II". In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

[125] Michal Czakon and Alexander Mitov. "Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders". In: *Comput. Phys. Commun.* 185 (2014), p. 2930. DOI: 10.1016/j.cpc.2014.06.021. arXiv: 1112.5675 [hep-ph].

[126] Ye Li and Frank Petriello. "Combining QCD and electroweak corrections to dilepton production in FEWZ". In: *Phys. Rev.* D86 (2012), p. 094034. DOI: 10.1103/PhysRevD.86.094034. arXiv: 1208.5967 [hep-ph].

[127] CMS collaboration. "Measurement of differential cross sections for top quark pair production using the lepton+jets final state in proton-proton collisions at 13 TeV". In: *Phys. Rev. D* 95.9 (2017), p. 092001. DOI: 10.1103/PhysRevD.95.092001. arXiv: 1610.04191 [hep-ex].

[128] Jorgen D'Hondt et al. *Fitting of Event Topologies with External Kinematic Constraints in CMS*. Tech. rep. CMS-NOTE-2006-023. Geneva: CERN, 2006. URL: https://cds.cern.ch/record/926540.

[129] Holger Ritter. "Improvement of the Higgs mass resolution in the ZH L+L- b b channel using a Kinematic Fit with constraints". Presented on Dec 2014. 2014. URL: https://cds.cern.ch/record/2318401.

[130] CMS Collaboration. *Search for resonant pair production of Higgs bosons decaying to two bottom quark-antiquark pairs in proton-proton collisions at 13 TeV*. Tech. rep. CMS-PAS-HIG-16-002. Geneva: CERN, 2016. URL: https://cds.cern.ch/record/2141024.

[131] A. D. Bukin. *Fitting function for asymmetric peaks*. 2007. arXiv: 0711.4449 [physics.data-an].

[132] CMS Collaboration. *Supplementary material of CMS-HIG-18-016*. URL: http://cms-results.web.cern.ch/cms-results/public-results/publications/HIG-18-016/index.html.

[133] LHC Higgs Combination Group ATLAS CMS. *Procedure for the LHC Higgs boson search combination in Summer 2011*. Tech. rep. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11. Geneva: CERN, 2011. URL: https://cds.cern.ch/record/1379837.

[134] Roger J. Barlow and Christine Beeston. "Fitting using finite Monte Carlo samples". In: *Comput. Phys. Commun.* 77 (1993), pp. 219–228. DOI: 10.1016/0010-4655(93)90005-W.

[135] CMS Collaboration. "Search for the Standard Model Higgs Boson Produced in Association with a Top-Quark Pair in *pp* Collisions at the LHC". In: *JHEP* 05 (2013), p. 145. DOI: 10.1007/JHEP05(2013)145. arXiv: 1303.0763 [hep-ex].

[136] CMS Collaboration. "Search for the associated production of the Higgs boson with a top-quark pair". In: *JHEP* 09 (2014). [Erratum: JHEP10,106(2014)], p. 087. DOI: 10.1007/JHEP09(2014)087, 10.1007/JHEP10(2014)106. arXiv: 1408.1682 [hep-ex].

[137] ATLAS Collaboration. "Measurement of VH, $H \to b\bar{b}$ production as a function of the vector-boson transverse momentum in 13 TeV pp collisions with the ATLAS detector". In: *JHEP* 05 (2019), p. 141. DOI: 10.1007/JHEP05(2019)141. arXiv: 1903.04618 [hep-ex].

[138] CMS Collaboration. "Search for the Higgs boson decaying to two muons in proton-proton collisions at $\sqrt{s}$ = 13 TeV". In: *Phys. Rev. Lett.* 122.2 (2019), p. 021801. DOI: 10.1103/PhysRevLett.122.021801. arXiv: 1807.06325 [hep-ex].

[139] *A Generic Fitter Project for HEP Model Testing*. URL: http://project-gfitter.web.cern.ch/project-gfitter/.

[140] *HEPfit : a Code for the Combination of Indirect and Direct Constraints on High Energy Physics Models*. URL: https://hepfit.roma1.infn.it/.

[141] CMS Collaboration. *Sensitivity projections for Higgs boson properties measurements at the HL-LHC*. Tech. rep. CMS-PAS-FTR-18-011. Geneva: CERN, 2018. URL: https://cds.cern.ch/record/2647699.

[142] CMS Collaboration. "Measurement and interpretation of differential cross sections for Higgs boson production at $\sqrt{s}$ = 13 TeV". In: *Phys. Lett.* B792 (2019), pp. 369–396. DOI: 10.1016/j.physletb.2019.03.059. arXiv: 1812.06504 [hep-ex].

[143] M. Cepeda et al. "Higgs Physics at the HL-LHC and HE-LHC". In: (2019). arXiv: 1902.00134 [hep-ph].