CERN Summer Student Programme

Project Report

Search for VH(cc) in the fully hadronic channel

Isadora Luiza Bozza Galvão

Supervisors:

Fabio Monti

Beatriz Ribeiro Lopes

Switzerland

2024

Contents

1 Introduction

The Standard Model (SM) is a well-established theory of elementary particle physics that predicts the compounds of matter, as well as their interactions. This model provides a description of the electromagnetic interaction, mediated by the photons; of the weak interaction mediated by the W and Z bosons, and the strong interaction, by the gluons. The Higgs boson gives origin to the other particles masses through electroweak symmetry breaking, for the W and Z bosons, and through the Yukawa interaction in the case of the fermions.

The discovery of the Higgs boson (H) in 2012 by the Compact Muon Solenoid (CMS) and A Toroidal LHC Apparatus (ATLAS) Collaborations at the CERN Large Hadron Collider (LHC) marked a new chapter in particle physics, completing the SM. Exploring the interactions of the Higgs boson with the gauge bosons and fermions are among the primary LHC goals, once the measured interactions and properties are essential to test the SM predictions. The currently observed Yukawa couplings between H and the fermions are compatible with the SM predictions. These are important measurements, considering that such predictions could be significantly modified by beyond the standard model physics. The current results by the CMS Collaboration are shown in the figure below. Brief overviews on the status of H boson measurements are presented in Refs. [4, 12].

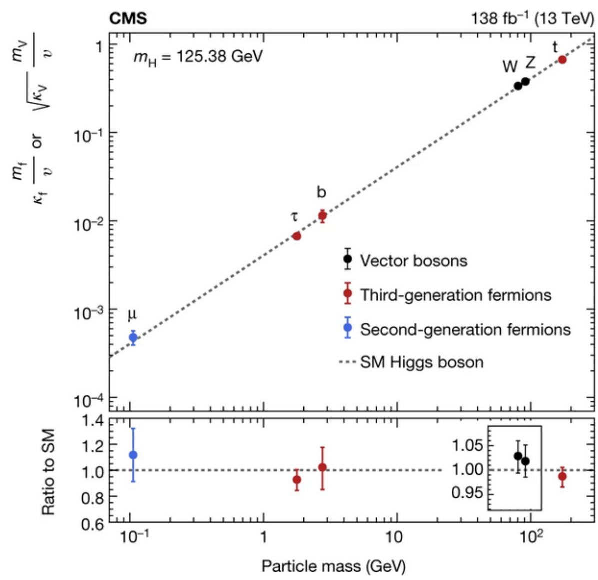


Figure 1: The measured coupling modifiers for vector bosons and fermions in CMS [4].

As can be noticed in Fig. (1), measurements for the charm quark are still missing. Accessing the H to charm coupling is very challenging, as it has one of the lowest branching fractions. For this reason, it was previously thought that this measurement was out of reach for the LHC, but the recent results [3, 5, 6] suggest it may be possible to achieve good sensitivity already with the available data. The channels showing the highest sensitivity are the same as the case of H to b-quark [2, 8], in particular the VH production process, where a H boson is produced in association with a vector boson (W or Z).

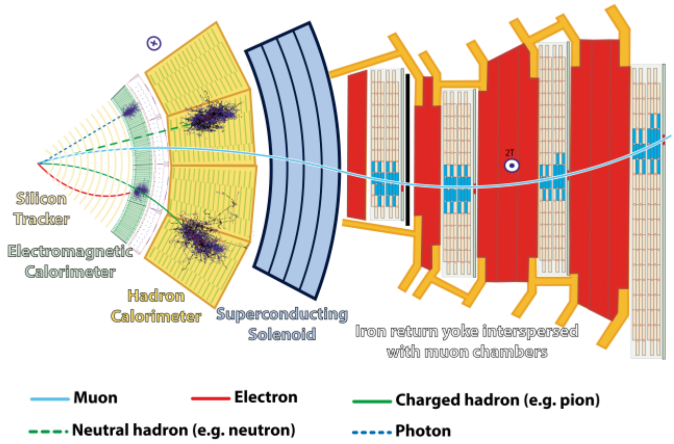
In this project, the hadronic channel is chosen, and the production process is VHcc, with the Z or W decaying to quarks and the Higgs to a charm anti-charm pair. A fundamental part of this analysis is the jet flavor tagging, which is done using the Particle Transformer (ParT) algorithm [10]. In order to suppress backgrounds, the merged-jet topology was used. In this topology, the H and V bosons each decay into two jets which have an overlap between them, and can thus be reconstructed as one single large-radius jets. The ParT algorithm aims at associating each jet with their probable origin, as described in Section 3.1.

Besides the jet identification, this channel presents an overwhelming amount of background events, when compared with the leptonic one. Two strategies were adopted to improve the sensitivity and the discrimination between signal and background events: on the one hand, the optimization of the trigger choice, in order to increase the number of signal events, and on the other hand, a multivariate analysis technique, known as Boosted Decision Trees (BDT), to separate between signal and backgrounds in order to obtain a clearer signature.

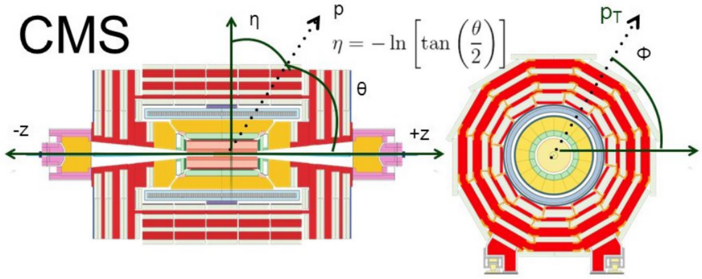
2 CMS Detector

The CMS is one of the four main experiments at the LHC and one of the largest, both in size and amount of personnel. Its main feature is a 6-meter-long magnet solenoid, providing a magnetic field of 3.8 T. Apart from the solenoid, CMS is composed of different layers, as seen in Fig. 2a: the innermost one is the tracker, responsible for recording the trajectory of charged particles; the second and third layers are the calorimeters, the first calorimeter is the electromagnetic calorimeter (ECAL), which measures the energy deposits of photons, positrons and electrons, while the other one is the hadronic calorimeter (HCAL), where the energy deposited by hadrons is measured. The muon chamber is a special feature to measure the muon trajectories and it is the outermost layer on the detector.

The CMS coordinate system, Fig. 2b, is right handed, with the origin in the center of the detector. In Cartesian coordinates, the z-axis is along the beam pipe direction, the x-axis points to the direction of the center of the LHC ring, and the y-axis, perpendicular to both x and z-axis, pointing upwards. In cylindrical coordinates, the azimuthal angle (ϕ) is defined in the xy plane, as $\phi = \arctan(y/x)$; the radius is $r = \sqrt{x^2 + y^2}$ since the polar (scattering) angle, θ, is not Lorentz invariant, it is useful to define the pseudorapidity: $\eta = -\ln[\tan(\theta/2)]$ as one of the kinematic variables.



(a) Sliced view of the CMS Detector layers [1].



(b) CMS Detector coordinate system, extracted from [11].

Figure 2: Overview of the CMS detector.

3 Analysis Strategy

3.1 Merged-Jet Topology

Following the strategy from previous analysis [3, 5, 8], the jets used for the VH(cc) analysis are the so-called "merged jets", with a larger radius than commonly used ($R = 0.4$). These jets are reconstructed with a radius $R = 1.5$, where $R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$. For this topology, $p_T^H \approx p_T^V \geq 200$ GeV is required, so that the H boson Lorentz boosted. The jets are also required to be back-to-back, $2.2 < \Delta\phi$ (ak15, ak15) $< \pi$. Fig.3 shows that the number of events is well distributed in the chosen range, with higher number of events for $\Delta\phi$ (ak15, ak15) approaching π.

The choice for the V and H candidates is a fundamental ingredient for the result of the analysis, and an option would be to select by the kinematic variables. However, both requirements on the jets p_T and on the $\Delta\phi$ cause the jets to have similar transverse momentum distributions, as it can be seen in Fig. 4. The lack of difference between the two plots suggests that using only the ak15 jets p_T is not the optimal way to do it, once that only by kinematic features of the process is not possible to distinguish between V and H candidates.

After the two large jets are selected, the jet tagging is done using the ParT [10], which has the goal to classify the particles originating the jet, returning outputs with

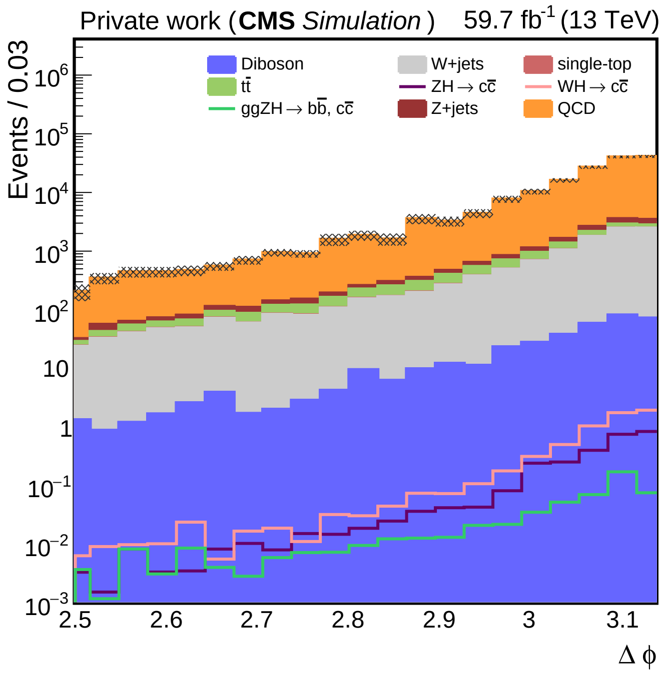


Figure 3: Distribution of the number of events for $\Delta\phi$ (ak15, ak15).

probabilities for the jet to originate from $H \to c\bar{c}$, $H \to b\bar{b}$ and $H \to q\bar{q}$. Besides the outputs for Higgs, it also returns scores for QCD multijets.

The dominant backgrounds in the analysis are from QCD multijets, V + jets process, $t\bar{t}$ and single top production. The jets from QCD and from the V decays can mimic the H signature, making it harder to distinguish them. This motivated the next steps: the trigger studies and the BDT training, which will be discussed in the following subsections.

3.2 Trigger Studies

The CMS detector has two trigger systems to filter the collision events. The first one is the L1 and it filters the events based on hardware processors, using information provided by the calorimeters and muons detectors. The second is the High-Level Trigger (HLT) and it processes the events using the event reconstruction software. The HLT returns trigger paths, which consists of steps done to reconstruct and apply selections on physics objects based on specific requirements. A common selection in HLT is defined by a threshold on Ht (the scalar sum of jet p_T momenta) and the standard trigger path is for Ht$>$ 1050 GeV [7].

Recently, CMS has been adopting a new approach for trigger selection, the data scouting, which reduces the Ht threshold to Ht$>$ 410 GeV. The scouting is suggested as a strategy to improve the signal sensitivity, motivating these trigger studies. A number of different trigger paths were used to evaluate the trigger efficiency for signal and for background, in order to estimate how the sensitivity would change according to the trigger

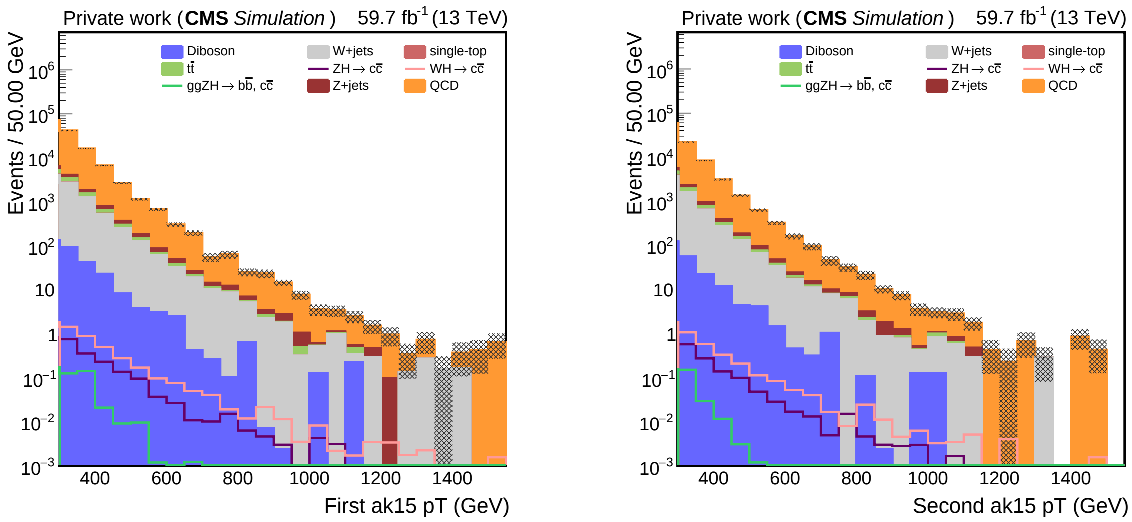


Figure 4: Control plots for the p_T for each of the ak15 jets.

selection.

The samples were divided in categories defined by the ParT scores for V and H candidates and also required to be orthogonal, such that there is no overlap between categories. The categories made were the following ones:

- bbcc: one jet with higher bb score and the other with higher cc score.
- bbbb: both jets with high bb score ($H \rightarrow b\bar{b}$ and $V \rightarrow b\bar{b}$).
- bblight: one jet, most likely Higgs candidate, decaying to bb and the V candidate to quarks.
- cclight: same case as the previous one, but with the H candidate decaying to cc.

the one used for this analysis was the cclight category, which allows better event selection for the chosen process.

After the categories have been defined, and the cuts applied to the samples, it was possible to extract the number of events for the ZHcc and WHcc signals passing the listed triggers paths, as well as for background, where only QCD and $t\bar{t}$ were considered, as they are the dominant ones.

Table 1: VHcc process in cclight category

path	# sig. events	sig. eff.	# bkg. events (10^3)	bkg. eff.	sensitivity
no trigger	12.918	100%	928.484	100%	0.0134
HLT_PFHT400	11.050	85.54%	779.30	83.93%	0.0125
HLT_PFHT1050	0.478	3.70%	5.847	0.63%	0.0062
1050 + others	2.942	2.942%	90.974	9.80%	0.0097

No significant difference was found between the standard trigger path (Ht 1050 GeV) and the scouting scenario when comparing the trigger efficiency for signal and background.

However, when comparing the number of events between the three paths, the scouting option has more than one order of magnitude more events than the 1050 path and the 1050 path combined with other trigger paths [1], thus suggesting scouting would be a good choice for increasing the number of events, but an additional method to reduce the number of events for background is also needed.

3.3 BDT Training

As shown by the control plots in Fig. (3) and (4), and by the trigger efficiency (table 1), the background dominates over the number of events for signal in more than 2 orders of magnitude. Therefore, to use the scouting trigger keeping the high number of signal events and reducing the background, a boosted decision tree (BDT) algorithm was trained. The BDT was designed using the Toolkit for Multivariate Data Analysis (TMVA) package for ROOT [9].

For the training, the `Factory` class is used, which receives samples from signal and background as input. The signal samples used were from ZHcc and WHcc, while for background, only QCD was considered. Besides the input samples, it is necessary to choose variables to be used in the training, which were the kinematics features, such as $\Delta\phi$ between the two large-jets, but also between each of the ak15 jets and the additional ak4 jet with the highest p_T, as well as the ParT scores for bb, cc, qq and QCD. The TMVA displays a graphical user interface (GUI) with results from the training and testing of the BDTs, such as the correlation between each variable, a check for overtraining and the ROC curve. These were used to evaluate the training and to guide optimization. The ranking of the variables with the best signal and background separation is also shown in the terminal. For the training used in this analysis, the rank of the 10 first variables is the following:

1. ParT QCD score for ak15_2
2. ParT QCD score for ak15
3. ParT cc score for ak15_2
4. $\Delta\eta$ (ak15, ak4)
5. ParT qq score for ak15
6. ParT cc score for ak15
7. ParT qq score for ak15_2
8. Pseudorapidity of ak15_2
9. Pseudorapidity of ak15
10. $\Delta\phi$ (ak15_2, ak4)

For each event is assigned a BDT score, which will return whether it is more likely to be signal or background. The score closer to 1 means a signal like event, while closer to 0, background like. In this project, the BDT score was plotted with all the signal

[1] HLT_PFHT330PT30_QuadPFJet_75_60_45_40_TriplePFBTagDeepCSV_4p5, HLT_AK8PFJet400_TrimMass30, HLT_AK8PFJet330_TrimMass30_PFAK8BoostedDoubleB_np4, HLT_AK8PFJet330_TrimMass30_PFAK8BoostedDoubleB_np2

and background samples, so a cut could be chosen in such a way as to have the optimal number of events for background and signal, as seen in Fig. (5).

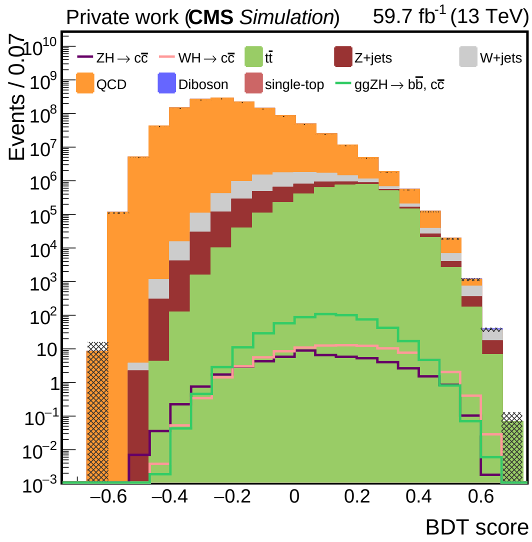


Figure 5: Output of the BDT Training.

Looking at the BDT output, a different behavior can be noticed for QCD, and partially also for Z+jets and W+jets when compared to the signal. On this project, the main focus on background was for QCD, for which the BDT output was found to perform well enough. However, to obtain a better result for other backgrounds, a new training should be performed adding other backgrounds as input samples, considering also new variables and training parameters.

A reasonable cut on the BDT score was found to be > 0.4, where a significant number of events from QCD can be eliminated while maintaining enough events for signal.

4 Results and Conclusions

The analysis was performed in the cclight category. Selected events were required to satisfy Ht> 400 GeV (to work on the scouting scenario), and a cut on $n_{\mathrm{ak4jets}} < 4$ was added, to reduce $t\bar{t}$ background. The pseudorapidity region was $|\eta| < 2.4$ and the missing transverse energy, MET < 150 GeV.

Subsequently, the Higgs candidate was chosen to be the ak15 jet with the highest ParT Hcc score and the highest regressed mass. After this selection, the regressed mass for the V and H candidates was plotted.

The mass peaks are visible in Fig. 6 when analyzing the signal processes, however they are rather broad in the range of $120 - 160$ GeV for the Higgs candidate, $70 - 100$ GeV for the W candidate and $90 - 110$ GeV for the Z candidate. Moreover, the background still dominates and for that reason, a cut was applied for BDT score > 0.4. The primary goal of reducing the QCD background was achieved, as can be seen in both plots in Fig. 7.

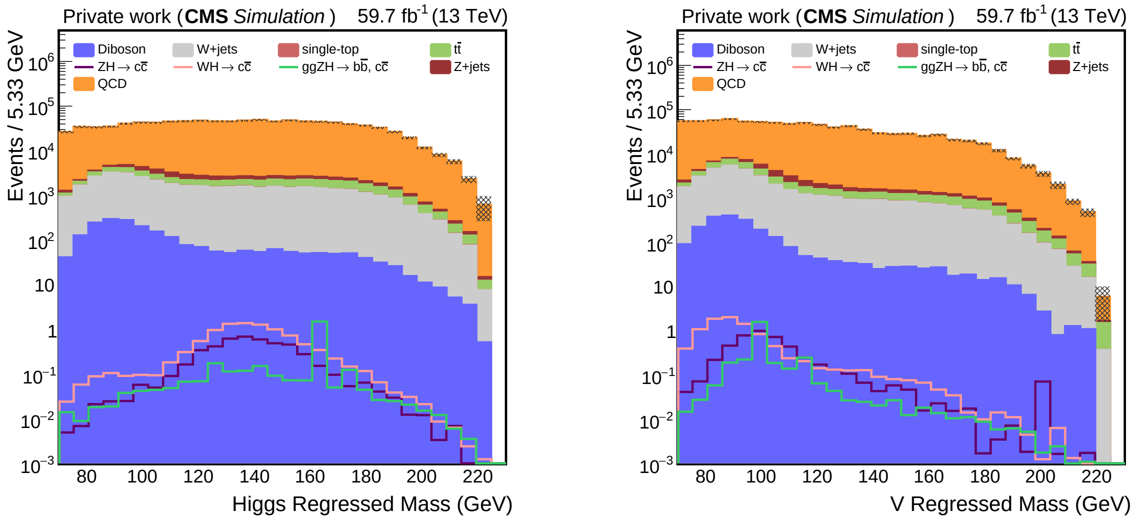


Figure 6: Regressed Mass for V and H candidates before BDT cut.

In addition, the other backgrounds, such as $t\bar{t}$ and diboson, though with less significant difference, were also reduced. For QCD, the number of events went from $50 \cdot 10^4$ in the peak region to 10^3, while for the signal processes the same order of magnitude was maintained.

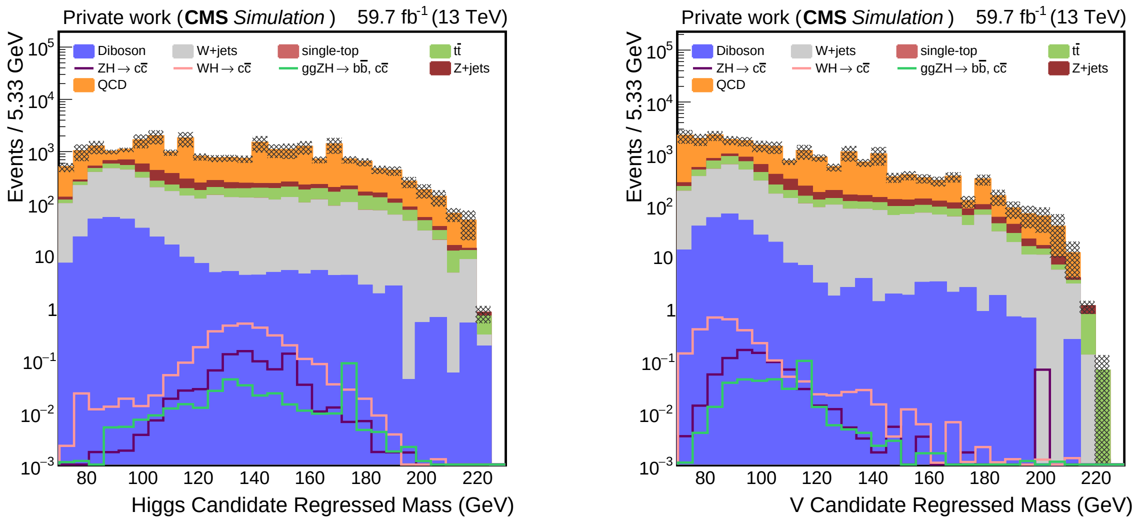


Figure 7: Regressed Mass for V and H candidates after BDT cut.

To estimate the sensitivity (signal/$\sqrt{\text{background}}$), a mass region of $110 - 160$ GeV was defined for Higgs candidate and $70 - 110$ GeV for V candidate and the number of events was extracted for signal and background. This estimation was done without and with the BDT cut to evaluate how much sensitivity is gained by using the BDT. For the backgrounds, only QCD and $t\bar{t}$ were considered.

The sensitivity showed an improvement of two times after the BDT cut, which suggests the BDT training is a suitable option for classification between background and signal.

Table 2: Sensitivity estimation

BDT score	only QCD	QCD and $t\bar{t}$
without	0.023	0.023
> 0.4	0.047	0.045

Although only trained using QCD as background sample, it also showed a relevant separation for the other background processes.

To conclude, this project illustrated the potential gain in signal efficiency using the scouting trigger and one strategy to separate signal and background. The association of both indicates a significant improvement potential for the analysis of VHcc. These improvements are needed, once the process has higher number of background events when compared to signal, and to the other Higgs production processes and decay modes.

References

[1] Barney, D. (2016). CMS Detector Slice. Acessed in Sep 2024. `https://cds.cern.ch/record/2120661`.

[2] CMS Collaboration (2018). Observation of higgs boson decay to bottom quarks. *Phys. Rev. Lett.*, 121:121801.

[3] CMS Collaboration (2020). A search for the standard model higgs boson decaying to charm quarks. *Journal of High Energy Physics*, 2020(3).

[4] CMS Collaboration (2022). A portrait of the higgs boson by the cms experiment ten years after the discovery. *Nature*, 607(7917):60–68.

[5] CMS Collaboration (2023a). Search for higgs boson and observation of z boson through their decay into a charm quark-antiquark pair in boosted topologies in proton-proton collisions at $\sqrt{s} = 13$ tev. *Physical Review Letters*, 131(4).

[6] CMS Collaboration (2023b). Search for higgs boson decay to a charm quark-antiquark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.*, 131:061801.

[7] CMS Collaboration (2024a). Enriching the physics program of the cms experiment via data scouting and data parking.

[8] CMS Collaboration (2024b). Search for bottom quark associated production of the standard model higgs boson in final states with leptons in proton-proton collisions at $\sqrt{s} = 13$ tev.

[9] Hoecker, A., Speckmayer, P., Stelzer, J., Therhaag, J., von Toerne, E., and Voss, H. (2009). TMVA - Toolkit for Multivariate Data Analysis.

[10] Qu, H., Li, C., and Qian, S. (2024). Particle transformer for jet tagging.

[11] Ribeiro Lopes, B. (2023). Differential ratio measurements of $\mathrm{t\bar{t}}\gamma/\mathrm{t\bar{t}}$ and a search for exclusive $\mathrm{t\bar{t}}$ with the CMS experiment. Presented 26 Oct 2023.

[12] Salam, G. P., Wang, L.-T., and Zanderighi, G. (2022). The higgs boson turns ten. *Nature*, 607(7917):41–47.