Vector boson fusion tagging with graph neural networks

Study of the Stability of Tagging Algorithms Applied to the Analysis of Higgs-to-Invisible Decays

Internship report
Summer Student Project - STAG-F017
Supervised by *Andrea Malara* and *Santeri Laurila*

Antoine Dierckx

ULB, Belgium
Summer 2024

Contents

Abstract

The goal of this internship is to explore the feasibility of using new methods based on Graph Neural Network (*GNN*) to distinguish between two production modes of the Higgs boson.

Traditionally, production modes classification (for instance, the Higgs boson) uses high-level kinematic variables like the energy or the angle between particle jets. However, these variables are derived from a more fundamental set of low-level data (for instance, the energy or the (η, ϕ) coordinate of individual particles).
GNNs can directly process low-level data, therefore maximizing the information used, by treating the particles as nodes in a graph.

Our main objectives are:

1. Compare the efficiency between GNNs and traditional variables for Higgs boson production mode classification.
2. Study the sensitivity of the GNNs to changes in the simulations.
3. Understand and quantify the variations of both high and low-level variables under the changes in the simulations.

To do so, we will study samples from simulations to understand how much they change under variations of the generators and the tunes of the latter, and how the GNN responds to these changes. We want to quantify the impact of such variations on the physical variables as well as on the GNN output, and on the classification performance associated.

We show that the GNN output fluctuates within $10 \sim 20$ %, while its classification performance remain stable within 1 %.

Cover and end illustration by Mette Ilene Holmriis.

> "The curiosity of the human mind is essential if you want citizens who think rather than accept the first nonsense they come to."
> *François Englert*[1]

1 Overview of the internship

This internship took place at CERN in Geneva, Switzerland, as part of the *STAG-F-017* course, figuring the program of Master in Physics - research-oriented. It ran from July 1 to August 2 on a part-time basis, then from August 2 to August 31 on a full-time basis.

CERN Summer Student Program

This internship echoes the structure of the CERN Summer Student Program.
The program was launched in 1962[2, 3] under the impetus of then General Manager Victor Weisskopf . It is made in two parts: the lectures and the internship.

The lectures cover all the main aspects of physics at CERN, with 26 different topics addressed within only five weeks.[4] They are part of four main branches: particle physics, computing, accelerators, and statistics. The lectures take place in the main auditorium, every morning of July. They offer an opportunity to meet the other students while the internship typically is an individual project (with the supervisor) or is conducted in small groups.

		Monday	Tuesday	Wednesday	Thursday	Friday
		1/7	2/7	3/7	4/7	5/7
Week 1	09h15-10h10		Introduction	Particle World	Raw Data to Physics Results	Detectors
	10h25-11h20		Particle World	Detectors	Particle World	Raw Data to Physics Results
	11h35-12h30		Detectors	Raw Data to Physics Results	Detectors	Particle World Q&A
		8/7	9/7	10/7	11/7	12/7
Week 2	09h15-10h10	Detectors	Accelerators & Beam Dynamics	Statistics	Nuclear Physics	Theoretical Particle Physics
	10h25-11h20	Accelerators & Beam Dynamics	Magnet superconductivity	Accelerators & Beam Dynamics	Theoretical Particle Physics	Statistics
	11h35-12h30	Magnet superconductivity	Statistics	Theoretical Particle Physics	Statistics	Nuclear Physics
		15/7	16/7	17/7	18/7	19/7
Week 3	09h15-10h10	Theoretical Particle Physics	Theoretical Particle Physics	Future High Energy Colliders	Astroparticle Physics	Cosmology
	10h25-11h20	Standard Model	Physics & Medical Applications	Standard Model	Cosmology	Future High Energy Colliders
	11h35-12h30	Physics & Medical Applications	Standard Model	Astroparticle Physics	Standard Model	Cosmology
		22/7	23/7	24/7	25/7	26/7
Week 4	09h15-10h10	RF superconductivity	Electronics, DAQ and Triggers	Heavy Ions	Neutrino Physics	Physics at Hadron Colliders
	10h25-11h20	Predictions at Hadron Colliders	RF superconductivity	Electronics, DAQ and Triggers	Physics at Hadron Colliders	Heavy Ions
	11h35-12h30	Electronics, DAQ and Triggers	Predictions at Hadron Colliders	Physics at Hadron Colliders	Heavy Ions	Physics at Hadron Colliders
		29/7	30/7	31/7	1/8	2/8
Week 5	09h15-10h10	Quantum Gravity	Beyond the Standard Model	Antimatter in the Lab	Beyond the Standard Model	Flavour Physics
	10h25-11h20	Physics at Lepton Colliders	Accelerator Operation & Design	Flavour Physics	Antimatter in the Lab	Beyond the Standard Model
	11h35-12h30	Accelerator Operation & Design	Physics at Lepton Colliders	Beyond the Standard Model	Flavour Physics	Close out

Figure 1: Lectures program for the CERN Summer Student Program 2024 [2]

The CERN Summer Student program gathers around 300 students (from member and non-member states) pursuing bachelor's or master's degrees in physics, computing, engineering and math. Students come for between 8 to 13 weeks. At the end of their stay, the students are invited to present their internship in the Student Session, where they have 10 minutes to summarize their internship.

1.1 Presentation of CERN

CERN is an international particle physics laboratory. Its acronym stands for European Organization for Nuclear Research (*Conseil Européen pour la Recherche Nucléaire*) and was founded in 1952 by 12 European countries[5]. It has 23 member states and 11 associate member states.
CERN is also an official United Nations General Assembly observer[6]. The CERN convention, signed in 1953, specifies:

> The Organization shall have no concern with work for military requirements and the results of its experimental and theoretical work shall be published or otherwise made generally available.

In 1954, the construction began close to Geneva with the construction of the first buildings and the Proton Synchrotron, and the convention was ratified by the 12 founding countries, including Belgium.

Figure 2: Aerial view of the site, June 15th, 1955[7]

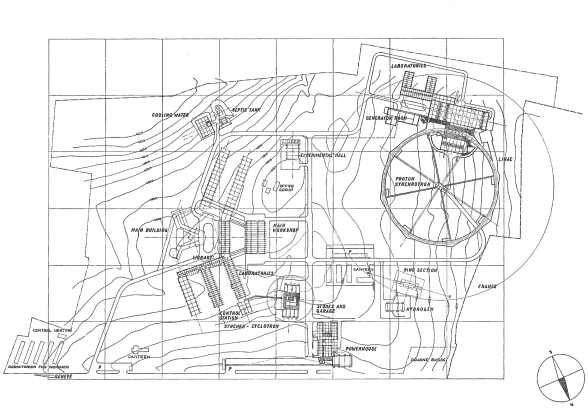


Figure 3: Plan of the site from Dr. Steiger, the chief architect, and his collaborators. [7]

During the Cold War, CERN was one of the few places in the world where Western and Eastern scientists worked side by side[8].
Today, CERN is the world's largest particle physics laboratory, with over a billion euros budget and more than 100 nationalities. Many people are working at CERN, including[9]:

- $\sim$ 2 700 staff
- $\sim$ 12 000 scientific users from $\sim$ 70 countries, including 129 from Belgium.
- $\sim$ 800 fellows
- $\sim$ 400 students
- $\sim$ 300 summer students
- $\sim$ 3000 contractors

The people working at CERN are either employed (*MPE*) or Associated (*MPA*). Staff members and Fellows constitute the MPE, while International collaboration (*MPAc*; users and other associates), exchange of scientists (*MPAx*), and training (*MPAt*; students, trainees, apprentices) constitute the MPA. The vast majority of the members of personnel are Users ($\sim$ 70%).

The current Director General is Fabiola Gianotti.

Objectives

The main goal of CERN is to test our current understanding of particle physics and to explore physics beyond the Standard Model. More precisely, the near-future objectives are: [10, 11]

1. "Deliver world-class scientific results and knowledge":
    - Fully exploit the potential of the Large Hadron Collider during its third and final run, including for the High-luminosity LHC (*HL-LHC*) project (new of the dipoles and quadrupoles, detector updates of Atlas and CMS, ...).

- Upgrade the injector complex to allow new possibilities for experiments like *ISOLDE* and *AWAKE*.
- Support the development of neutrino physics by collaborating with the *DUNE* experiment in constructing a second cryostat.
- Continue to support the theoretical particle physics.

2. Study the options for a future collider:
   Study the technical and financial feasibility of different types of collider, including:
   - The Future Circular Collider (*FCC*).
   - The Compact Linear Collider (*CLIC*).
   - Muon colliders.
3. Increase the return to the Member and Associate Member States, including with the industry.
4. Strengthen CERN's impact on society

Budget

According to the CERN annual report of 2023[9], their total expenses are 1 305,6 MCHF (about the same in euros). The repartition of the budget is illustrated in Figure 4.
In the previous years, the budget was stable: 1224.9 MCHF in 2022[12], 1228.4 MCHF in 2021[13], 1157.4 MCHF in 2020[14] and 1259.7 MCHF in 2019[15].

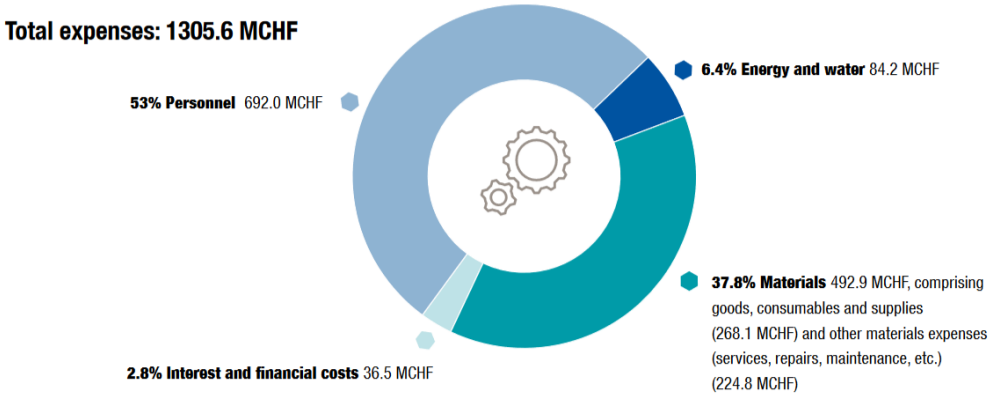


Figure 4: CERN expenses [9]

The Compact Muon Solenoid program alone represent 13 070 kCHF for 2024, separated in 8340 kCHF for staff expenses and 4730 for material expenses[16]. It is a 16% decrease from last year[17].

Each Member State must contribute to CERN funding according to its GDP, in a way decided by the CERN Council[1]. In 2024[18], Belgium contributed for $\sim 2,8$ % of the Member States' contributions with around 34 million euros. Some Associate Member States also contributed $\sim$ 40 million euros, while the total Member States' contribution is around 1 220 million euros.

Diversity & Inclusion Program

Promotion of diversity and inclusion at CERN dates back to the 90s[19]. In 1993, an advisory group published recommendations for equal opportunities for women. The role of Equal Opportunities Officer (*EOO*) was established in 1996, followed by the creation of the Equal Opportunities Advisory Panel (*EOAP*) in 1998.
By 2010-2011, a broader Diversity Program was created, and the *EOO* and *EOAP* were dissolved.

[1] The CERN Council is the supreme decision-making authority of the Organization, composed by delegates of all its twenty-three Member States.

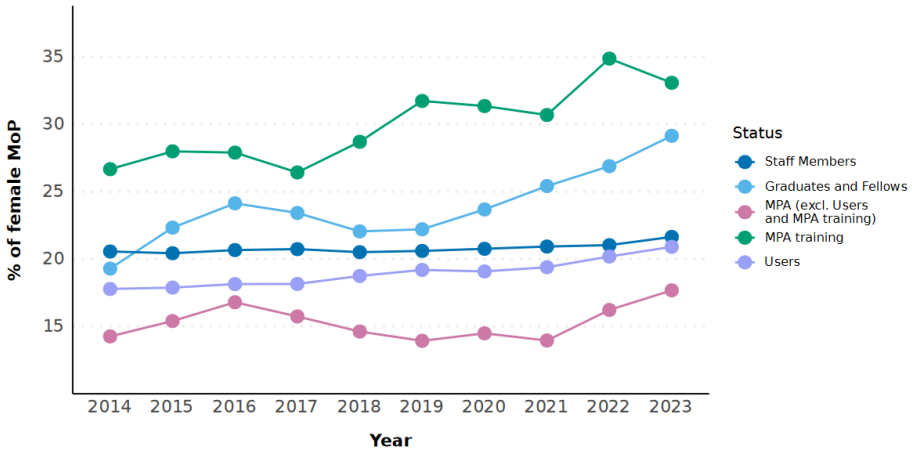


Figure 5: Proportion of Female Members of the Personnel over the last 10 Years (2014 – 2023)[20]

Today, there is still a huge amount of progress to be made, even if approaches promoting inclusiveness and diversity are increasingly present (for example, with the "25 by '25" project).

Student Opportunities

Besides the Summer Student Program, CERN offers many possibilities to pursue a career for different types of profiles and qualifications. I'll limit myself here to opportunities for PhD students. There are three main ways of joining CERN for a PhD:[21, 22]

1. The Doctoral Student Program:
   For interests in Applied Physics, Engineering, or Computing, one can apply to this program to spend up to 36 months at CERN. Candidates must be from a CERN Member or Associate Member State. CERN provides a contract, monthly allowance, travel support, and health insurance.
   CERN typically selects around 40-50 doctoral students at each selection committee. The last one was in February 2024.
2. The Marie-Curie PhD Position:
   This program is funded by the European Union. Candidates must have a master's degree. This program offers similar conditions that the Doctoral Student Program, with a shorter period (6 months).[2]
3. CERN Collaboration with a university:
   Any university or Institute (for example the Inter-University Institute For High Energies *IIHE*) collaborating with CERN might offer PhD positions on subjects closely related to the research at CERN.
   In general, this means spending a few weeks to a few months a year on site (at CERN), and the rest of the time at the university or institute.

1.2 Particle accelerators at CERN

The accelerator complex at CERN is a succession of particle accelerators with increasingly high energies. Each machine injects the beam into the next one, the *LHC* being the last one.
The biggest and most powerful particle accelerator is currently the *LHC* (*Large Hadron Collider*) where particle beams are accelerated to a peak energy of 6,5 TeV. The *LHC* is a 27-kilometer ring of superconducting magnets, which bend the trajectory of the proton beams, along with several accelerating structures that boost the protons to speeds approaching the speed of light.
Within the accelerator, two high-energy particle beams travel in opposite directions and cross each other at 4 locations. Those beams move in two tubes maintained in a vacuum. They are guided around the accelerator ring by a magnetic field generated by superconducting electromagnets. To keep these electromagnets in a superconducting state, the required temperature is around $-271.3\,C$, a temperature lower than that of outer space.

Additionally, many of the intermediate accelerators have their own experimental facilities, enabling research at lower energy scales.

[2] This opportunity was part of an EU project (INTENSE) for the year 2022.

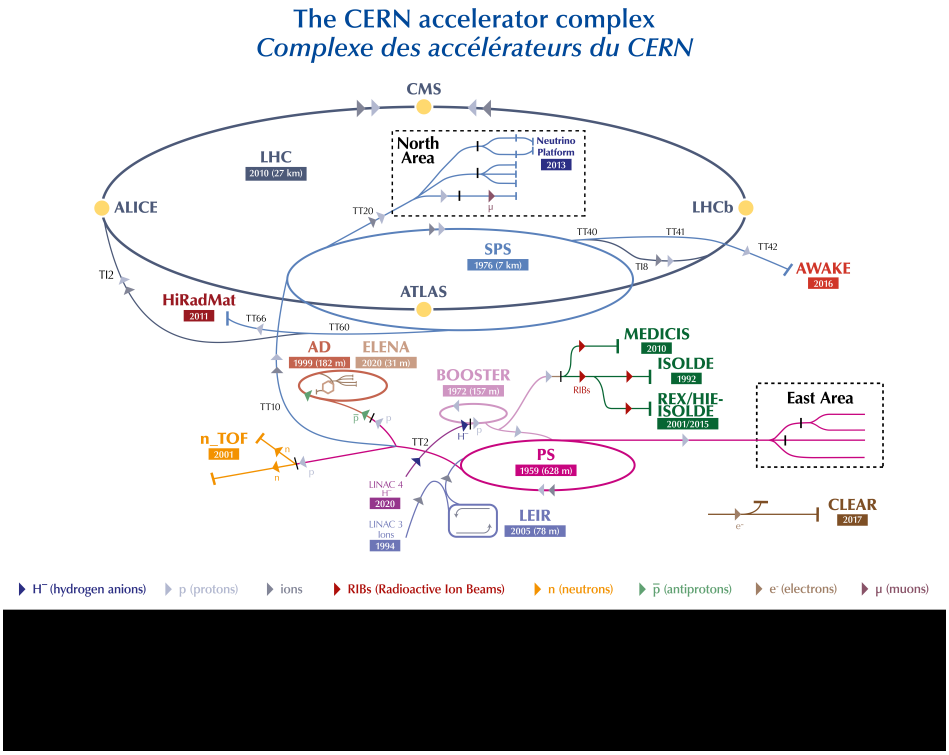


Figure 6: The CERN accelerator complex layout in 2022[23]

The process of accelerating protons through CERN's accelerator complex is as follows:

1. Before 2020, the source of the proton beam was hydrogen atoms taken from a simple bottle containing hydrogen. The atoms were ionized, and only the protons were kept.
   Since 2020, the Linear accelerator 4 (*Linac4*) is the source of proton for the accelerator complex. It accelerates the negative hydrogen ions to 160 MeV.[24]
2. From Linac4, the beam goes to the Proton Synchrotron Booster (*PSB*), where the ions are stripped of their electrons, leaving only protons. The PSB accelerates the protons to 2 GeV, after which they are transferred to the Proton Synchrotron (PS), where they are further accelerated to 26 GeV.
3. The beam is then sent to the Super Proton Synchrotron (SPS), where the protons reach 450 GeV.
4. Finally, the protons are transferred into the *LHC*, where they are injected into two beam pipes, one circulating clockwise and the other anticlockwise.
   It takes 4 minutes and 20 seconds to fill each ring of the *LHC*, and 20 minutes to accelerate the protons to their maximum energy of 6.5 TeV. Under normal conditions, these beams can circulate for many hours.

In addition to protons, the *LHC* also accelerates lead ions. These ions are produced by vaporizing a purified lead sample and injecting it into Linac3. The lead ions are collected and pre-accelerated in the Low Energy Ion Ring (*LEIR*) before following the same route as protons through the *PS*, *SPS*, and *LHC*. The lead ions ultimately reach a maximum energy of 2,76 TeV per nucleon in the *LHC*.

The products of these collisions are recorded by the *ALICE*, *ATLAS*, *CMS*, *LHCb*, *LHCf*, *MoEDAL*, *TOTEM* experiments, *FASER* and *SND@LHC*.
This year is the third year of the RUN 3 of the *LHC*, which operates at an energy of 13.6 TeV.

1.3 The CMS detector

As Belgium is one of the institutes involved in the CMS collaboration, my internship is part of this. Therefore, I will be focusing on the CMS experience and the associated collaboration.

Overview

The CMS detector is one of the two large detectors at the LHC (the other one being ATLAS). CMS stands for *Compact Muon Solenoid*. The CMS detector is one of the largest detectors globally, measuring 21 meters in length, 15 meters in diameter, and weighing 14,000 tonnes.

The detector

The CMS detector aims to reconstruct the type of particles, their position, energy, and momentum from collisions at the LHC. CMS has a cylindrical geometry, layered around the collision point. This cylinder is blocked on either side by ''caps''. This particular geometry means that we have to work with an adapted coordinate system.

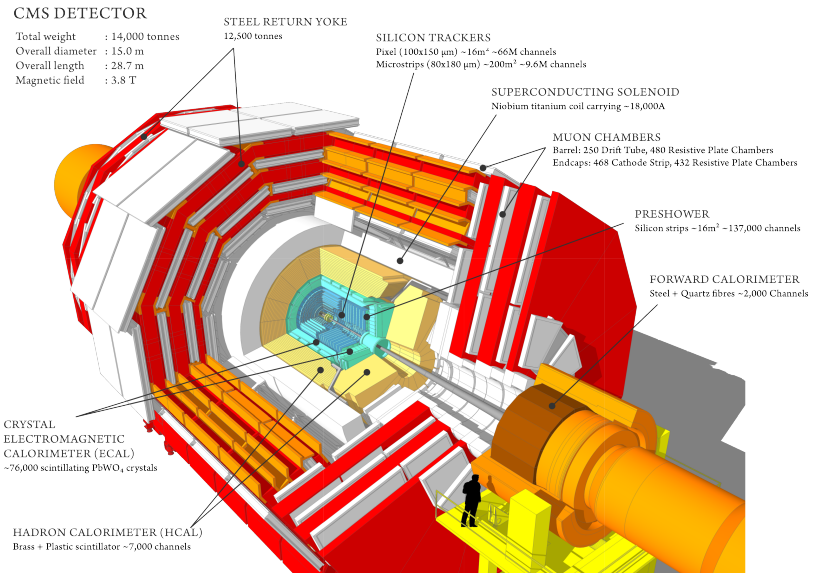


Figure 7: Sectional view of the CMS detector. [25].

The following coordinate system is adopted:

- We take the z axis to be the axis of the beam. At the same time, we define the plane perpendicular to this axis, the transverse plane. Subsequently, we will consider certain quantities restricted to this plane, for example, the transverse momentum p_T, which CMS measures.
- The angle θ is defined as the polar angle with respect to the z axis.
- The ϕ angle is defined as the azimuthal angle, included in the transverse plane.

Definition 1.1 (Pseudorapidity). The pseudorapidity η, a function of θ, is defined as follows:

$$\eta \equiv -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] \tag{1.1}$$

Property 1.1. *The differences in pseudorapidity are Lorentz invariants for boosts along the z axis.*

From ϕ, η and p_T, one can reconstruct the three components of the momentum (p_x, p_y, p_z):

$$p_x = p_T \cos\phi \tag{1.2}$$
$$p_y = p_T \sin\phi \tag{1.3}$$
$$p_z = p_T \sinh\eta \tag{1.4}$$
$$\tag{1.5}$$

1.4 The CMS collaboration

The CMS collaboration gathers 6288 people [26, 27] from 57 countries and regions.
"The CMS Management Board, chaired by the CMS Spokesperson, is responsible for directing the CMS experiment following the policies agreed by the CMS Collaboration Board. The Spokesperson represents the Collaboration in dealing with other organizations and committees, while the CMS Collaboration Board is the governing body of the experiment that makes all major decisions."[28, 29]

To participate in the collaboration, they are three types of membership[30]:

1. Full Membership:
   The full membership includes full rights and obligations, including the right to vote in the Collaboration Board, the ability to sign all CMS publications, and eligibility for leadership roles. Financial and operational contributions are expected.
2. Cooperating Institute:
   This type of membership concerns institutions aiming for full membership, and is limited to about five years. Members contribute to specific projects, sign relevant publications, and participate in Collaboration Board discussions without voting rights.
3. Associated Institute:
   An associated institute would focus on technical contributions in areas like engineering or computing, without involvement in physics analyses or financial obligations.

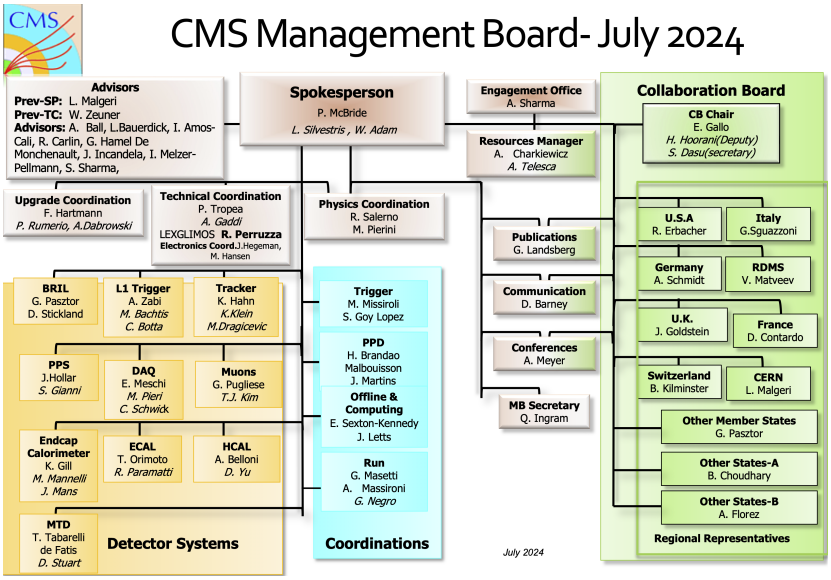


Figure 8: CMS Management Board Organigram[31]

The CMS collaboration is organized in different Coordination Areas, each specialized in some aspect of the detector.
The different Coordination Areas are:
Offline Software and Computing (O&*C*), Physics Coordination (*PC*), Physics and Performance Datasets (*PPD*), Run Coordination (*RC*), Technical Coordination (*TC*), Trigger-HLT (*TSG*), Upgrades (*UC*).

Within the Physics Coordination, there are two types of working groups: the Physics Object Groups (*POGs*) and Physics Analysis Groups (*PAGS*)[32]. For each group, a *Twiki* page is available to access any useful information and to make collaboration between groups easier.

The different Physics Object Groups are:
BTV: B-tagging and Vertexing, TRK (Tracking), EGM (Electron and Photons), JME (Jet and Missing Energy), MUO (Muons), TAU (Taus), LUM (Luminosity), PRO (Protons (in PPS)).

The different Physics Analysis Groups are:
BPH (B Physics and Quarkonia), SMP (Standard Model Physics), TOP (Top Physics), HIG (Higgs Physics), SUS (Searches for new physics in final states with Unbalanced pT and Standard objects), EXO (Searches for Exotica), B2G (Searches for Beyond SM particles decaying to top quarks and Higgs and Gauge bosons), HIN (Heavy-Ion Physics).

Some tasks being useful for different Coordination Area, a third type of groups, the Shared Groups, is constituted. The different Shared Groups are:
Generator and MC production (shared with *O&C*), Machine Learning (shared with *O&C*), Particle Flow (shared with *PPD*).

The analysis in which this internship is taking place is part of the physical coordination within the HIG group.

1.5 Work environment

I worked in a shared office with my supervisor, Dr. Andrea Malara. The office is located in the Belgian section of building 40, on the Meyrin site. Being close to my supervisor facilitated the interactions and allowed me to ask (many) questions when needed. The office is equipped with multiple whiteboards, which are used both for personal use and to explain various concepts.
Next to the office is the main restaurant, R1, which is open throughout the day. It is an important place because coffee breaks allow us to connect with other researchers and network for future projects.

During my internship, I worked alongside another summer student who initially focused on a similar subject. In the first two weeks, we followed the same introduction to coding, allowing us to assist one another. However, our projects quickly diverged and we began to work individually.

Part of the internship is to participate in regular meetings with members of the same working group. These meetings, typically held weekly, often included participants from different universities, such as collaborators from Boston and the Université Libre de Bruxelles (ULB). The main purpose of such meetings is to present updates on ongoing research and get feedback and comments to help plan the next steps of the work. Throughout my internship, I made 4 presentations that allowed me to work on my communication skills.

2 Vector Boson Fusion and invisible decay of the Higgs

2.1 Higgs physics

The Higgs boson is part of the Standard Model of Particles. It was discovered for the first time at CERN in 2012, in CMS and ATLAS.
In ATLAS, it was discovered through the $h \to \gamma\gamma$ (at loop level) and the $h \to 4l$ channel. In CMS, it was discovered through the $h \to 4l$ channel.
Since then, Higgs physics has been used to probe the Standard Model and to explore the physics beyond it.

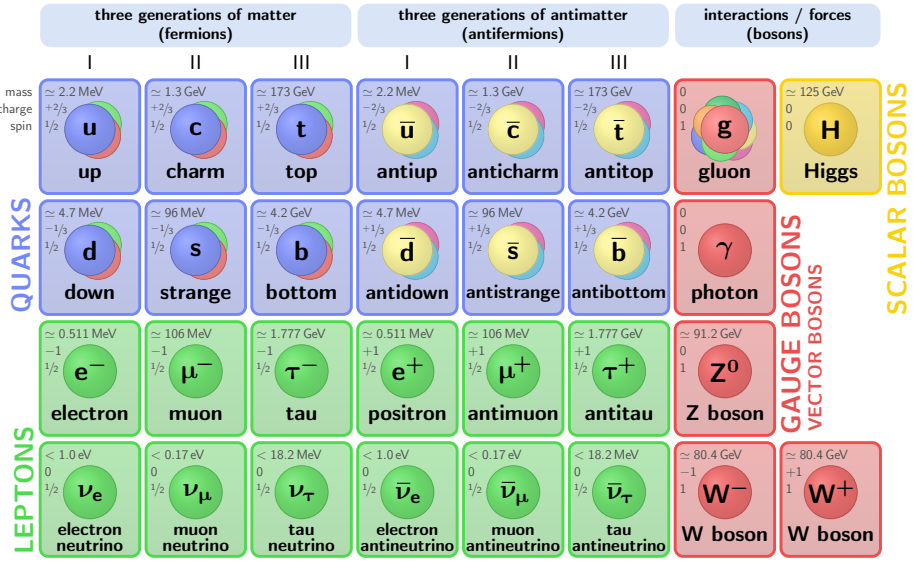


Figure 9: Summary of the elementary particles of the Standard Model [33]

2.1.1 Higgs production

The Higgs production depends on the characteristics of the accelerator: what is being collided (proton-proton, electron-electron, heavy ions, etc.), what is the energy in the center of mass, ...

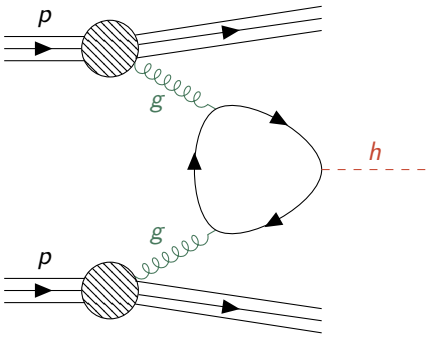


Figure 10: Feynman diagram associated to the ggH production mode

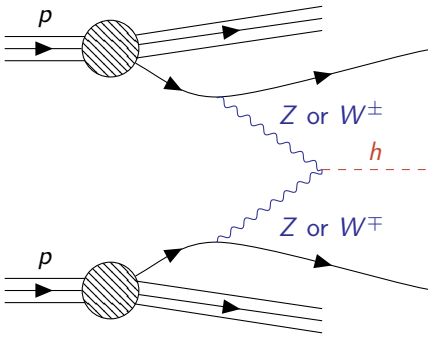


Figure 11: Feynman diagram associated to the VBF production mode

ggF

At the LHC, the main production mode is the gluon-gluon fusion (*ggF*): the emission of two gluons from the protons, that go into a quark loop (dominated by the contribution of the top quark [3]) and produce a Higgs boson.

It's relative branching ratio at $\sqrt{s} = 13$ TeV is $\sigma/\sigma_{\text{tot}}|_{\sqrt{s}} \approx 87\%$[34].

[3] the Higgs boson couples approximately 35 times more strongly to the top quark than to the next heaviest quark, the bottom quark, resulting in the bottom quark's contribution being suppressed by a factor of 35^2

VBF

The second most important channel is the vector boson fusion (VBF): the emission of two vector bosons (either two Z or a W^+ and a W^-) that fuse to form a Higgs boson.
Each initial-state quark emitting a vector boson remains roughly along its initial direction, staying close to the beam. They will hadronize and produce two forward jets in the final state. If we denote the η coordinate of each jet η_1 and η_2, this means that

$$\Delta\eta \equiv |\eta_2 - \eta_1| > 1 \tag{2.1}$$

It's relative branching ratio at $\sqrt{s} = 13$ TeV is $\sigma/\sigma_{\text{tot}}|_{\sqrt{s}} \approx 7\%$[34].

2.1.2 Higgs decay

Unlike the Higgs production, the Higgs decay is independent of the energy in the center of mass and of the characteristic of the accelerator.

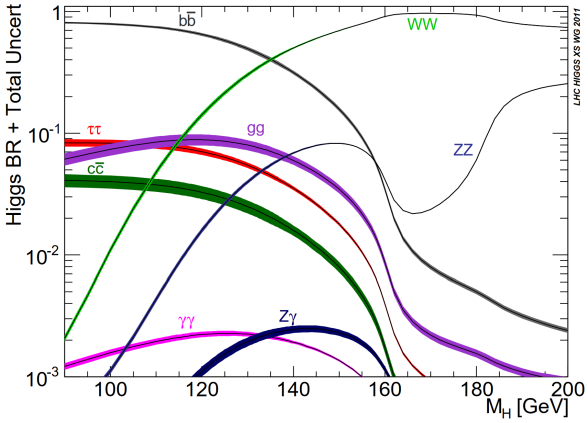


Figure 12: Higgs branching ratios and their uncertainties for the low mass range.[35]

Background

The Higgs boson mainly decays into a bottom quark and antiquark ($b\bar{b}$), with a branching ratio of $\mathcal{B}_{H\to b\bar{b}} \approx 58\%$. However, investigating this decay mode poses significant experimental challenges. In the dominant gluon-gluon fusion (ggF) production mode, the $H \to b\bar{b}$ signal is often overwhelmed by backgrounds generated through quantum chromodynamics (QCD) multijet events [36].
To date, the LHC has successfully observed all the main production modes and most of the key decay channels of the Higgs boson, including decays into $b\bar{b}$, WW, $\tau\bar{\tau}$, ZZ, and $\gamma\gamma$.

Signal

One of the rarer decay channels of the Higgs boson involves its decay into invisible particles. Within the Standard Model, the only invisible decay of the Higgs boson occurs via $H \to ZZ \to 4\nu$, with a branching ratio of $\mathcal{B}_{H\to ZZ\to 4\nu} \approx 0.1\%$.
However, various theoretical models propose that the Higgs boson could act as a portal between the Standard Model and a dark sector. In such scenarios, the Higgs boson might decay into a pair of dark matter (DM) particles, which would not interact with the detector material, thereby contributing to the branching ratio. As a result, the direct search for these invisible decays of the Higgs boson is a crucial approach for investigating dark matter production and Beyond Standard Model physics.

2.2 Overview of the internship

The goal of this internship is to contribute to a better understanding of the decay of the Higgs boson into invisible particles. We will focus on the vector fusion production (VBF) and the Higgs to invisible decay ($Hinv$).
Observed and expected upper limits on $(\sigma_{\text{VBF}}/\sigma_{\text{SM}}) \times \mathcal{B}(H \to \text{inv})$ at 95% C.L. for 2012–2018 data are presented in Figure 13. The combination of data collected in 2012, 2015, 2016, 2017, and 2018 sets an observed (resp. expected) upper limit on the invisible decay branching ratio of the Higgs boson, $\mathcal{B}(H \to \text{inv})$ at less than 18% (resp. 10%) with 95% confidence. This is the most precise constraint on $\mathcal{B}(H \to \text{inv})$ achieved so far[37].

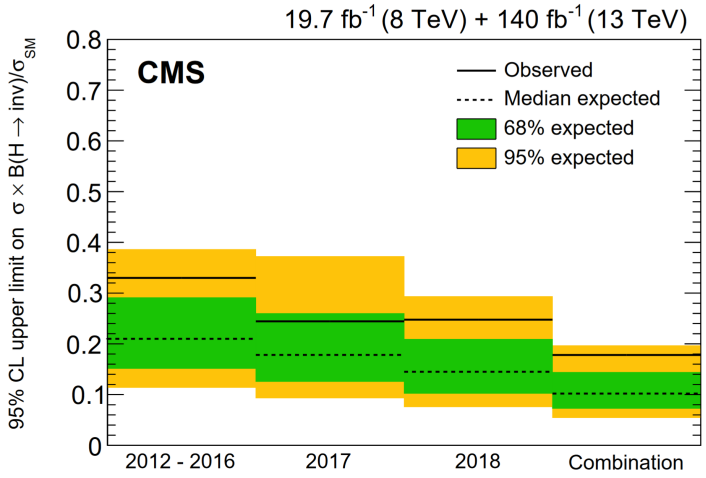


Figure 13: Observed and expected 95% CL upper limits on $(\sigma_{VBF}/\sigma_{SM}) \times \mathcal{B}(H \to \text{inv})$ for all data-taking years considered, as well as their combination, assuming an SM Higgs boson with a mass of 125.38 GeV.[37]

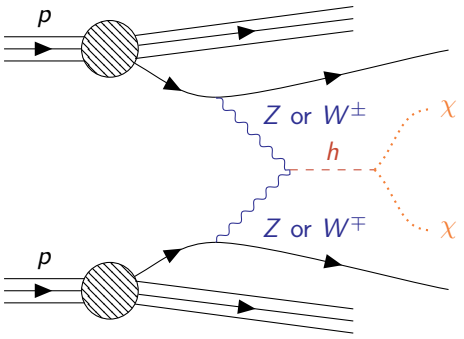


Figure 14: Feynman diagram associated with the VBF Higgs production decaying into unknown particles χ

This summer project is part of the effort of developing a Neural Network capable of distinguishing between *ggF* from *VBF* production modes. The traditional discrimination method uses *high-level* kinematic variables, while the new approach using a graph Neural Network is trained with *low-level* variables.

Definition 2.1. We distinguish two types of variable:

1. low-level variables: measurements provided by the detectors (ex: $\vec{p}$, E_T, η, ... for each individual particle)
2. high-level variables: non-linear combinations of low-level variables that capture useful high-level information (p^μ of reconstructed jet, number of particles in each jet, ...)

[38]

The discrimination power of these different methods between ggF and VBF production modes is compared using simulated event samples.

Signature

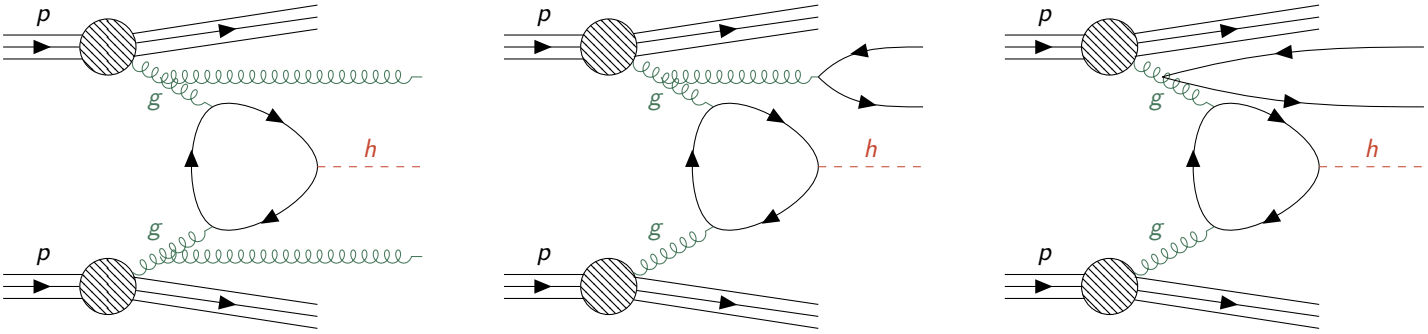
To classify the events, we look at their *signature*: number of reconstructed jets, number of muons, angles between the jets, and so on. This allows us to make *selections* to select particular productions and decay modes. In the VBF-Hinv case, we expect:

- Two jets coming from the hadronization of the two quarks the emitted the vector bosons
- Large angular separation between those two jets
- Large invariant mass of the reconstructed jets
- Missing transverse energy

Definition 2.2. The invariant mass of a physical system is defined as:

$$\mathcal{M} \widehat{=} p^{\mu} p_{\mu} = \sqrt{E^2 - p_x^2 - p_y^2 - p_z^2} \tag{2.2}$$

Interesting variables to discriminate the *signal* (VBF-Hinv) from the *background* (everything else) could therefore be the difference of η between the two jets, and the invariant mass of the two jets.
Despite the cuts one could apply to the events, one can expect background from the ggH production mode. We show a few examples of higher-order processes that could contribute to the background by mimicking the signature:



Machine Learning

Instead of using the usual variables, one could be tempted to use a combination of those (and/ or a combination of low-level ones) in order to maximize the event tagging. A way to find such a combination is to use machine learning.

2.3 Internship Objectives

Since the variable generated by the *Deep Neural Network* is trained on simulation, one has to make sure that the use of different generators or tunes does not change significantly the results of the tagging.
Let's first introduce some concepts in order to define precisely the goal of the analysis.

2.3.1 Useful concepts

Generators and Tunes

In particle physics, an "*event*" describes the outcome of a particle collision or decay process, where the final-state particles must conserve the energy, momentum, and quantum numbers of the initial state. Due to the random nature of quantum processes, these events vary, with the number and properties of outgoing particles changing each time. The probability distributions governing these variations can be inferred from experimental data or predicted by theoretical models.

An *"event generator"* is a tool used to simulate these events based on the theory. These simulations are then compared with data from the detectors. `PYTHIA` is one of the most widely used event generators in particle physics. It models the complex interactions and decays that occur during collisions (see Figure 15), producing a detailed account of the resulting particles and their properties.

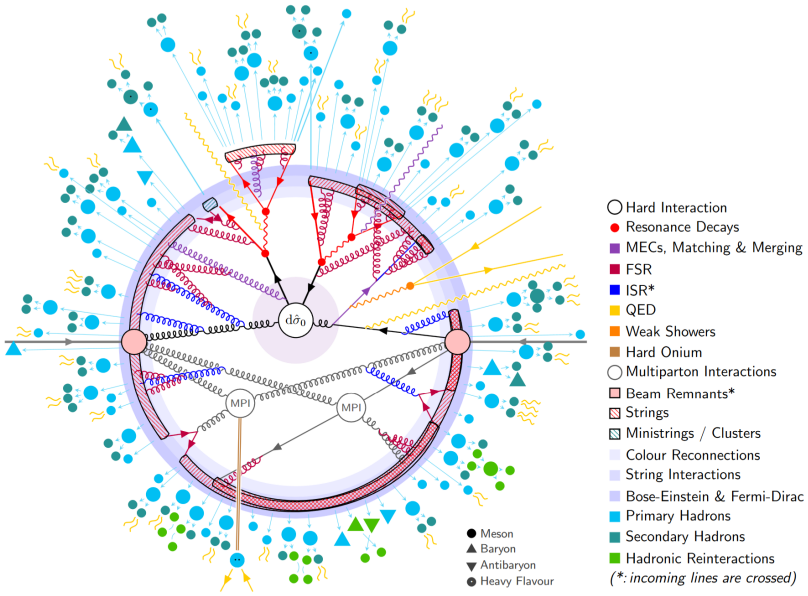


Figure 15: Schematic of the structure of a $pp \to tt$ event, as modelled by PYTHIA[39]

The predictions of PYTHIA depend on some free parameters that must be tuned to match experimental data. Thanks to the factorization theorem, the generation process can be split into several steps. Let's list some of them:

1. Hard Scattering: Two partons from the incoming hadrons undergo a hard collision, producing outgoing particles based on parton distribution functions and matrix elements from perturbation theory.
2. Resonance Decays: Short-lived resonances, such as W, Z bosons, or top quarks, decay into stable particles.
3. Radiative Corrections: Matrix-element corrections and *parton showers* account for *initial and final-state radiation*, adding more particles to the event.
4. Multiple Parton Interactions (MPI): Additional parton scatterings occur, adding complexity to the event, distinct from "pileup" from multiple collisions.
5. Color Reconnection and String Formation: Partons are confined into strings, which fragment into hadrons. Color reconnection may alter the string configuration before fragmentation.
6. Hadronization: The strings fragment into hadrons, which may experience Bose-Einstein or Fermi-Dirac effects.
7. Hadron Decays: Unstable hadrons decay into stable particles.
8. Final-State Interactions: In densely populated regions, particles may rescatter or recombine before the event concludes.

To generate the matrix element, the POWHEG (for *Positive Weight Hardest Emission Generator*) method is used. It is designed to interface next-to-leading order (NLO) QCD calculations with parton shower simulations, to simulate the behavior of particles after a collision more precisely.
Indeed, PYTHIA only simulates particle emissions at the leading logarithmic level, which is not accurate enough for precise measurements. To improve the accuracy, NLO calculations need to be incorporated. When the matrix element has been generated, PYTHIA is used for radiative corrections.

The generator's behavior can be modified by changing its internal parameters or by applying different models for specific processes like parton showers. These modifications are done by tuning parameters to match the experimental data better. To get accurate simulation of "*underlying event*" (everything which is not coming from the primary hard scattering process, nor pile-up), a specific tuning of PYTHIA 8, based on the data from the CMS experiment, is used: CP5 (standing for *CMS PYTHIA 8 Tunes*).
More detailed information about CP tunes can be found in the following reference: [40].
CP5 tune can be set on 3 modes: *Central*, *Up* and *Down*.

Weight

Weights are factors applied to individual events or processes in a Monte Carlo simulation to account for various uncertainties or to implement corrections based on different theoretical assumptions. Common examples include *ISR* (Initial-State Radiation) and *FSR* (Final-State Radiation) weights, which adjust how much radiation is emitted

before or after the main scattering event. Such weights can be set on 3 modes: *Central*, *Up* and *Down*[41]

Unlike the generator settings or tunes, which must be configured before the simulation is run, weights can be applied and adjusted after the events have already been generated.

Particle Flow Candidates (PFC)

A particle flow candidate is the output of the particle flow algorithm, which represents a stable particle, such as a charged hadron, neutral hadron, photon, electron, or muon, as reconstructed from the detector signals.
These candidates are derived by associating and integrating data from various subsystems of the detector, including calorimeter clusters (from *ECAL* and *HCAL*), tracks from the tracker, and hits in the muon system.
Each Particle Flow Candidate is characterized by its momentum components, energy, charge, and type, etc. Such variables are referred as low-level variables.

The PFCs present in this analysis are: charged hadrons, neutral hadrons, photons, muons, hadrons in HF (charged and neutral hadrons detected in the *Hadronic Forward* calorimeters), and electromagnetic in HF (charged and neutral non-hadronic matter detected in the *Hadronic Forward* calorimeters).

2.3.2 Classification with graph neural networks

Classifier and ROC curve

As mentioned in the section 2.2, we would like to classify the events to discriminate *ggF* from *VBF*.
Such an algorithm is called a *classifier*.

Definition 2.3. We define the True Positive Rate TPR and the False Positive Rate FPR as:

$$\mathrm{TPR} \equiv \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}} \tag{2.3}$$

$$\mathrm{FPR} \equiv \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = \frac{\mathrm{FP}}{\mathrm{N}} \tag{2.4}$$

where TP stands for True Positive and FN for False Negative.

The TPR represents the proportion of correctly identified signal events relative to the total number of true signal events. The FPR, on the other hand, indicates the proportion of background events that are erroneously classified as signal.

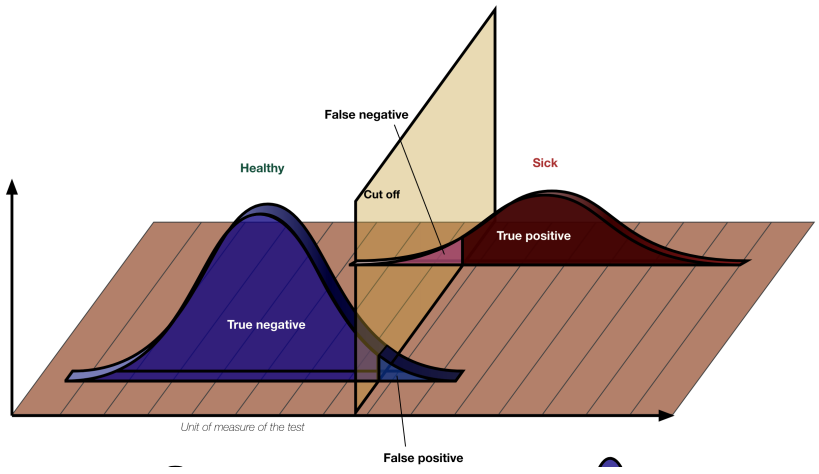


Figure 16: Illustration of the True Positive Rate and the False Positive Rate[42].

The performance of the classifiers is depicted using a certain type of curves. Selecting a threshold determines classification: events to the right of the threshold are labeled as signal, while those to the left are labeled as background. As the threshold is varied incrementally, corresponding TPR and FPR values are obtained, which are then plotted to generate the ROC (*Receiver Operating Characteristic)* curve. Hence, a ROC curve is the set of coordinates

$$\Big\{\big(\mathrm{FPR}(\mathrm{cut}), \mathrm{TPR}(\mathrm{cut})\big) \in [0,1]\times[0,1] \quad \forall\ \mathrm{cut} \in \mathbb{R}\Big\} \tag{2.5}$$

The model's classification performance is quantified by the *Area Under the Curve* (AUC), which measures the model's ability to distinguish between classes. A higher AUC reflects superior model performance.

ParticleNet

ParticleNet is a type of graph neural network (GNN) architecture, based on the Dynamic Graph Convolutional Neural Network (DGCNN).
ParticleNet originates from jet tagging, a process aimed at identifying the particles that initiate a jet. This model treats a jet as an unordered permutation-invariant set of particles, each carrying a feature vector, much like a *point cloud*. Using this technique, ParticleNet has shown significant improvements in jet tagging accuracy compared to previous methods.

A key component of ParticleNet is the edge convolution (EdgeConv) operation, which starts by representing a point cloud as a graph. Each vertex corresponds to a point, and edges connect each point to its k nearest neighbors. In this way, a local patch needed for convolution is defined for each point as the k nearest neighboring points connected to it.

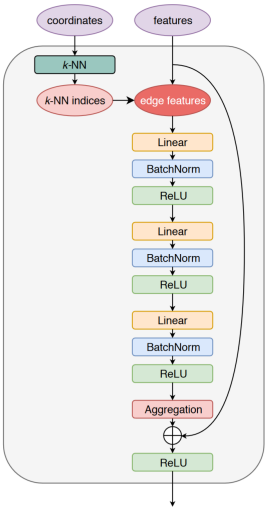


Figure 17: The structure of the EdgeConv block.[43]

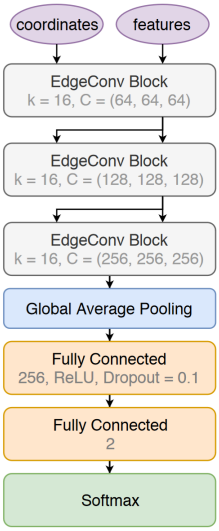


Figure 18: The architectures of the ParticleNet [43]

Figure 17 illustrates that EdgeConv needs two classes of features as input: the "coordinates", which include the relative values of η and ϕ for each PFC, and "features"

$$\left(p_T,\ \log_{10}(p_T),\ \eta,\ \phi,\ \text{energy},\ \log_{10}(\text{energy}),\ \textit{pdgId}^{4},\ \text{charge},\ \text{puppiWeight}^{5}\right)$$

The EdgeConv block begins by identifying the k nearest neighbors for each particle, using the input coordinates to calculate distances. The "edge features" for the EdgeConv operation are then created from the "features" input, based on the indices of these nearest neighbors.
The EdgeConv operation uses a three-layer Multilayer Perceptron (MLP), where each layer consists of a linear transformation, batch normalization, and ReLU activation. It includes a shortcut connection inspired by *ResNet*, allowing input features to bypass the transformation layers.

On the other hand, ParticleNet architecture is composed of three EdgeConv blocks, where the first block uses particle coordinates in $\eta - \phi$ space to compute distances, and the subsequent blocks use learned feature vectors.
Each block considers 16 nearest neighbors, with increasing channel sizes across the blocks. After the EdgeConv blocks, global average pooling aggregates the features, followed by a fully connected layer with ReLU activation and dropout to prevent overfitting.
The final output for binary classification is produced by another fully connected layer, followed by a softmax function.

[4] "PDG Identifiers are digital object identifiers assigned and used by PDG to reference items of PDG data such as particles, particle properties, decay modes and review articles. "[44]

[5] The PileUp Per Particle Identification weights encode the probability that each particle originates from pileup rather than from the primary vertex.[45]

Regarding the classification with graph neural network, the performance of the ParticleNet architecture is evaluated on all events. The training of the ParticleNet model was done by the collaborators in the analysis team. Its output, the DNN score, is a number between 0 and 1 and represents the probability that the input belongs to a particular class.

2.3.3 Objectives

Now that we introduced the various concepts, we are ready to precisely define the objectives of the analysis.

> We want to quantify the impact of tunes and weights variations on high and low-level variables, as well as on the *DNN score*.
>
> 1. We'll study variations of parton shower (*PS*) weights (both on ggH and VBF):
>   *ISR* and *FSR*, ↑ and ↓
> 2. We'll study `pythia8` with tunes (on VBF only):
>   *CP5* (nominal), *CP5*↑ and *CP5*↓
>
> We are analyzing deviations from the nominal, focusing on significant fluctuations (beyond statistical noise), consistent offsets where results remain systematically above or below the expected values, and shape effects where deviations occur only in specific regions of the range.

Remark. We restrict the tune analysis to the signal, as samples of the *CP5*↑ and *CP5*↓ variations for *ggH* were not yet available at the time of the internship.

Remark. The number of events in each sample varies a lot. Indeed, for *VBF*, *CP5*↑ and *CP5*↓ only contain ~ 3000 events, while *CP5* (nominal) has ~ 6000 events. For *ggH* on the contrary, the number of events is ~ 40000, which lowers significantly the fluctuations due to the lack of statistics.

2.4 Analysis

To perform the analysis, we first need to clean the Particle Flow Candidates. Then, we feed the DNN with the 100 first PFCs (the ones with the greatest momentum). The DNN gives back the *DNN score* both for the signal and the background, which allows us to plot the ROC curve associated.

2.4.1 PFCs cleaning

Final cleaning on PFCs

We impose

1. $\min p_T = 0.2\ GeV$ and $\max p_T = 10\ TeV$ and $|\eta| < 5.2$
2. $p_T \geq 1\,GeV$ if the PFC is a photon
3. $p_T \geq 3\,GeV$ if the PFC is a hadron in HF, an electromagnetic in HF, a neutral hadron or a muon.
4. PFC from primary vertex

Updates to the cleaning algorithm

The training of the DNN was made with a bug in the code. Namely, the neutral mask used was of the form:

$$\text{old neutral mask} \equiv \left(\text{id} \neq i \mid \text{id} \neq j \mid \dots\right) \mid p_T > 3\,GeV$$

And was not affecting the muons. The new mask is of the form:

$$\text{new neutral mask} \equiv \left(\text{id} \neq i \;\&\; \text{id} \neq j \;\&\; \dots\right) \mid p_T > 3\,GeV$$

If adding the muon to the mask did not affect significantly the performance of the DNN, swapping the *or* conditions by the *and* ones did.

This result is shown in Figure 19. In this figure, we used the following labeling:

- *dNN nominal* for the ROC curve from the DNN output produced with the new neutral mask, affecting the muons. This is the cleaning algorithm used in the rest of this analysis.
- *dNN_1* for the ROC curve from the DNN output produced with the old neutral mask, not affecting the muons.
- *dNN_2* for the ROC curve from the DNN output produced with the old neutral mask, affecting the muons.

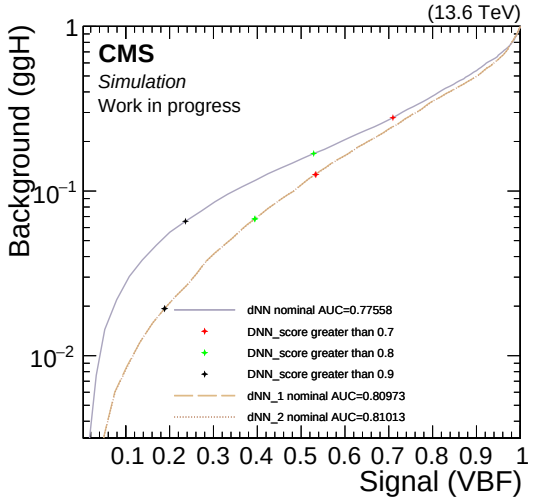


Figure 19: Evolution of the ROC curve under different PFCs cleaning

Discussion on PFC Energy Distribution

The simulation samples present a non-physical distribution of the total energy of the PFCs.
Namely, we observe events until 35 TeV, both for ggH and VBF.

However, this must be put into perspective given the small number of events above 14 TeV.

Remark. Naturally, the weights do not modify the maximum total energy reached by the PFCs.

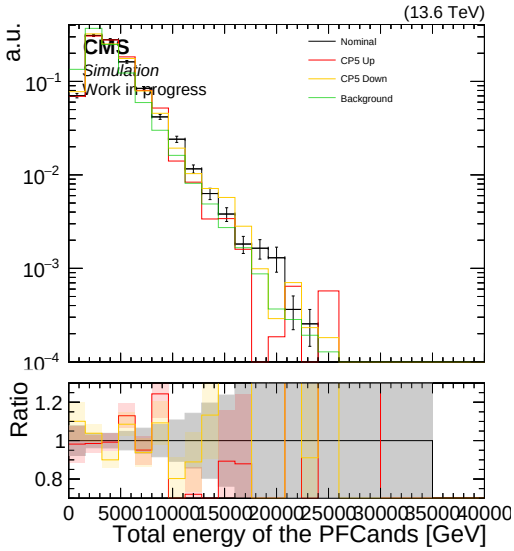


Figure 20: Total energy of the PFCs, varying under CP5 tunes

2.4.2 ROC curves

DNN score

The DNN score for ggH and VBF, under the weights and the tunes variations are:

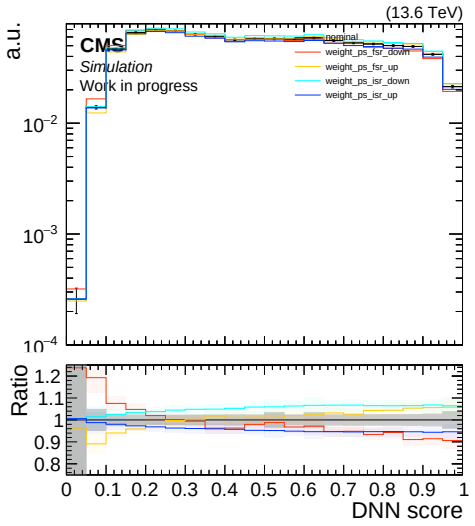


Figure 21: DNN score for ggH, varying under PS weights

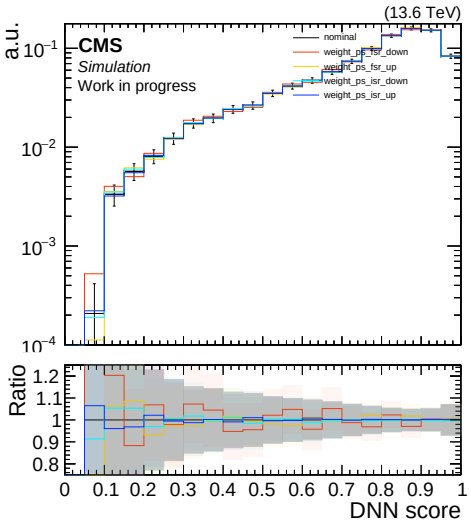


Figure 22: DNN score for VBF, varying under PS weights

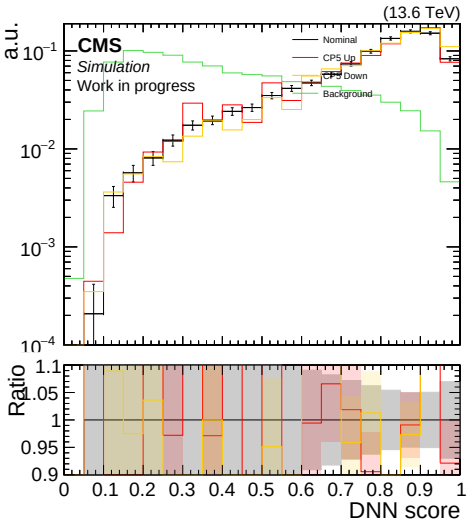


Figure 23: DNN score for ggH and VBF, varying under CP5 tunes

The corresponding ROC curves are:

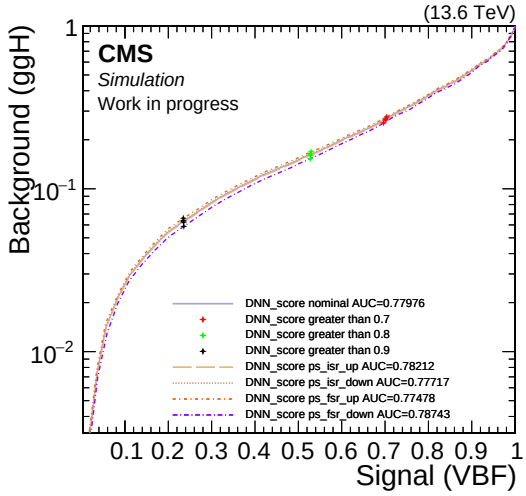


Figure 24: ROC curve of the DNN score, varying under PS weights

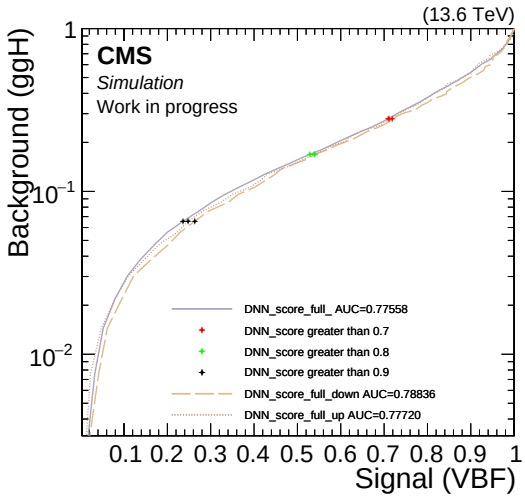


Figure 25: ROC curve of the DNN score, varying under CP5 tunes

We observe that the *DNN score* can vary within $\sim 10-20\%$ when we vary the Parton Shower weights and the CP5 tunes, but the ROC curves stay stable within $\sim 1\%$.

To track the evolution of the cut in the ROC analysis, we plotted points for fixed values of the *DNN score*, namely at *DNN score* $= 0.7$, *DNN score* $= 0.8$ and *DNN score* $= 0.9$. We do not see big changes among the variations.

Remark. As expected, those tracking points are purely horizontal for the CP5 variations, since the background is fixed. To understand where the *DNN score* of ggH variation comes from, we need to study the low-level variable. Before diving into the analysis, let's show the same ROC curves for the invariant mass of the two jets.

Invariant mass of the two jets m_{jj}

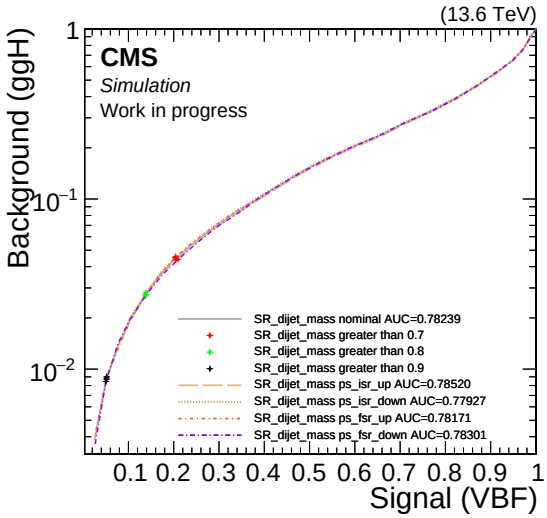


Figure 26: ROC curve of the m_{jj}, varying under PS weights

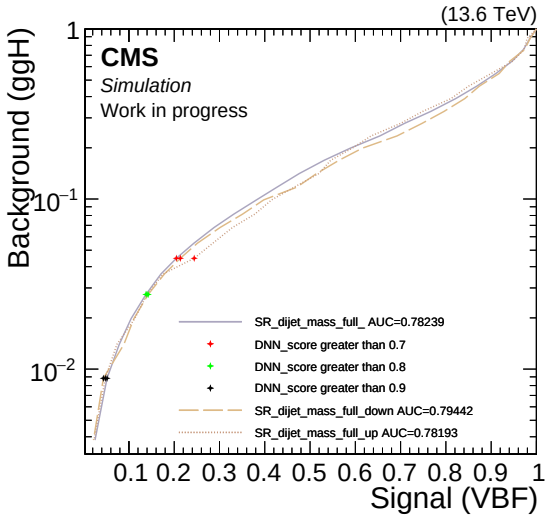


Figure 27: ROC curve of the m_{jj}, varying under CP5 tunes

We see that the variable m_{jj} furnishes a better classifier than the *DNN score* after the change in PFC selection (*higher area under the curve)*. We can suspect that after retraining the DNN on the new selection, the performance of the latter would outperform the classifier power of m_{jj}.

2.4.3 Distributions

The DNN sees only the 100 first Particle Flow Candidates, and is given the following variables for training:

$$\left\{ p_T,\ \log(p_T),\ \eta, \phi,\ \text{energy},\ \log(\text{energy}), \text{pdgId},\ \text{charge},\ \text{puppiWeight},\ \text{fromPV} \right\} \tag{2.6}$$

We then analyze the variation of those variables, limited to the first 100 Particle Flow Candidates. To refine the analysis, we classify the PFCs depending on their flavor and with a jet-related selection.

Remark. We normalize the integral of each distribution against the integral of the nominal.

Jet related selections

The η of each PFC being known, we classify them by their position relative to the jets.

Definition 2.4. The angular distance ΔR between two objects is defined by

$$\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{2.7}$$

The following selections are defined:

1. `in_jet_1`:

$$\Delta R_{\text{PFC}-\text{jet}_1} < 0.4 \tag{2.8}$$

2. `in_jet_2`:

$$\Delta R_{\text{PFC}-\text{jet}_2} < 0.4 \tag{2.9}$$

3. `in_any_jet`:

$$\texttt{in\_jet\_1} \mid \texttt{in\_jet\_2} \tag{2.10}$$

4. `not_in_any_jet`:

$$\neg\texttt{in\_jet\_1}\ \&\ \neg\texttt{in\_jet\_2} \tag{2.11}$$

5. `within_jets`:

$$\Big((\eta_{\text{jet}_1} < \eta < \eta_{\text{jet}_2}) \mid (\eta_{\text{jet}_2} < \eta < \eta_{\text{jet}_1})\Big)\ \&\ \texttt{not\_in\_any\_jet} \tag{2.12}$$

6. `outside_jets`:

$$\Big((\eta < \eta_{\text{jet}_1}\ \&\ \eta_{\text{jet}_2} < \eta) \mid (\eta < \eta_{\text{jet}_2}\ \&\ \eta_{\text{jet}_1} < \eta)\Big)\ \&\ \texttt{not\_in\_any\_jet} \tag{2.13}$$

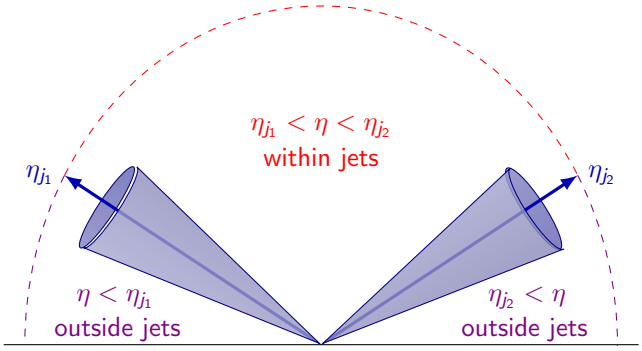


Figure 28: Illustration of the two regions defined by the jets

The PFCs `not_in_any_jet` (that includes `within_jets` and `outside_jets`) can be interpreted as coming from *underlying events*.

2.5 Results

The general results are the following:

1. The signal VBF is generally much more stable than the background ggH under the change of weights, with fluctuations typically of the order of the statistical uncertainty of the nominal. Therefore, the VBF weight variations will not be displayed in this report.
2. The parton shower weights *Initial State Radiation* tend to give the greatest offset from the nominal, by the order of $\sim 5-10\%$. Those effects virtually disappear when the distributions are normalized to a fixed value.
3. The parton shower weights *Final State Radiation* tents to give shape effects of the order of $\lesssim 5\%$. Those effects are unaffected by normalization.
4. Due to the lack of statistics, results on CP5 variations are more difficult to interpret. We can nevertheless see that there is no dramatic difference between the nominal and *CP5*↑ and *CP5*↓.
5. The subregion where the greatest variations are observed is usually:
   - `within_jets`
   - restricted to charged hadrons

   We can link those variations to the presence of **underlying events**.

2.5.1 Energy and transverse momentum

We observe fluctuation at high and low p_T for ggH. At low p_T, the most important variations are observed for charged hadron `within_jets`. The same behavior is observed for the photons, the charged hadrons in HF, and the neutral hadrons.

At high p_T, most of the events are logically coming from `in_any_jet`.

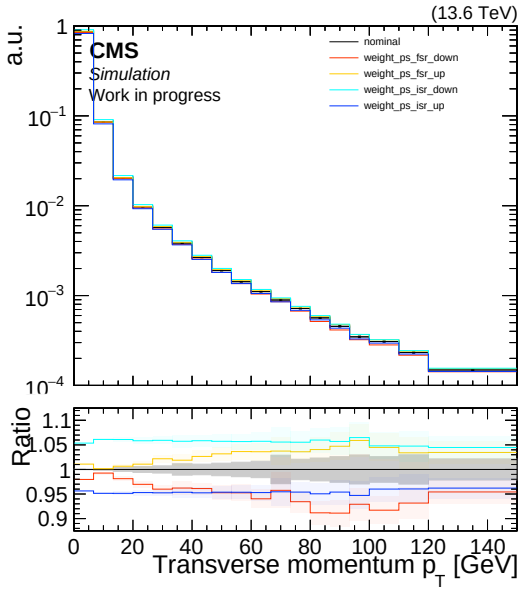


Figure 29: p_T of the first 100 PFCs for ggH, varying under PS weights

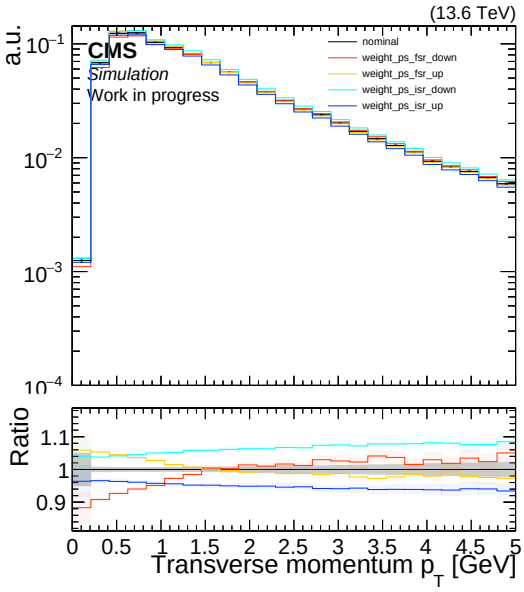


Figure 30: p_T of charged hadrons (restricted to the first 100 PFCs) for ggH, `within_jets`, varying under PS weights

The CP5 tunes variations seem to fluctuate around the nominal, without clear off-set of shape effect. This behavior can be explained by considering the low number of events.

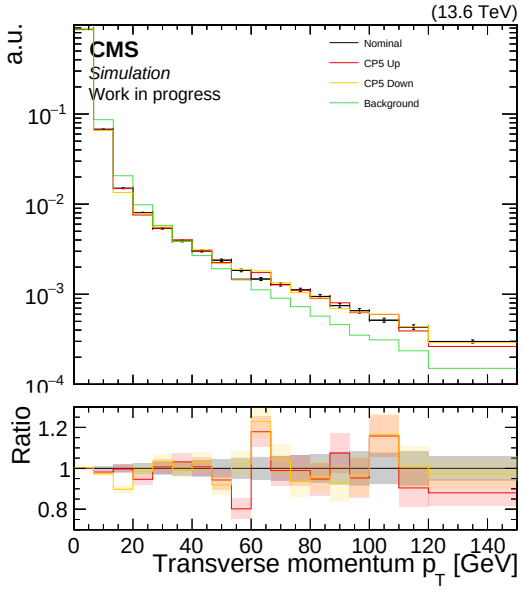


Figure 31: p_T of the first 100 PFCs, varying under CP5 tunes

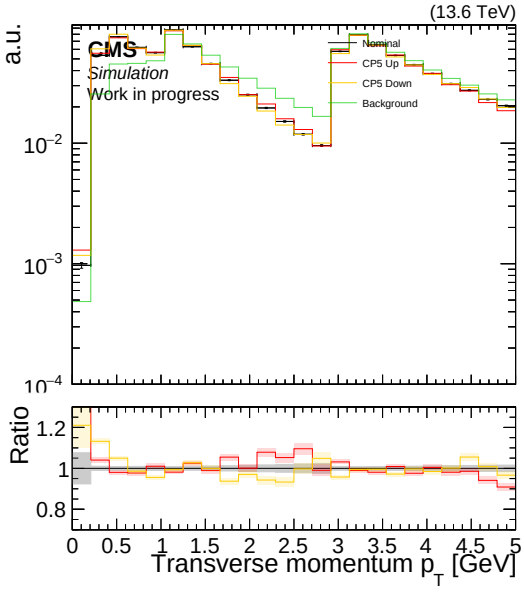


Figure 32: p_T of the first 100 PFCs, varying under CP5 tunes

Concerning the energy, the greatest variation appears for charged hadrons and photons `within_jets` for ggH under the PS weights variation.
The deviation is of order $\sim 5 - 10\%$. We observe deviations of order $\sim$ 5% for the neutral hadron `within_jets` too.

Under the CP5 tunes variation, we observe fluctuations around nominal that can reach up to 40 % of the signal in some subregion (for instance, charged hadrons `within_jets`), but there is no clear deviation.

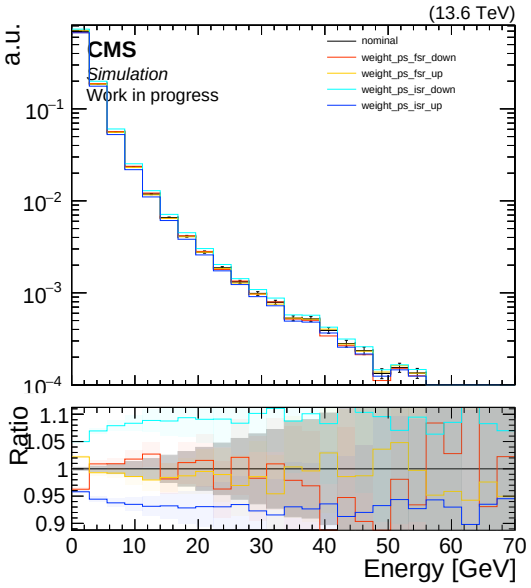


Figure 33: Energy of charged hadrons (restricted to the first 100 PFCs) for ggH, `within_jets`, varying under PS weights

2.5.2 Charge

We observe a typical variation of order $\sim$ 5% for the ggH under PS weights variation.

The subregion where the variation is the most important is `within_jets`. Similar behavior is found for charged hadrons, neutral hadrons, and photons.

The fluctuations for VBF under CP5 tunes variations are less important, of order of $\lesssim$ 5%.

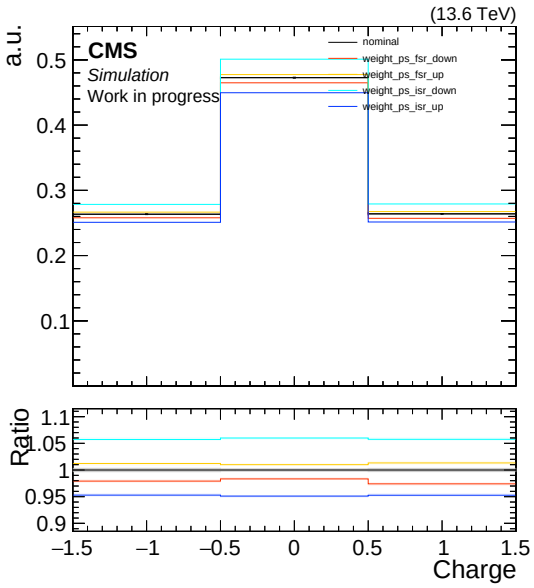


Figure 34: Charge of the first 100 PFCs in ggH, `within_jets`, varying under PS weights

2.5.3 Eta

Unlike the other variables, we observe a typical variation of order $\sim 5\%$ both `in_any_jet` and `not_in_any_jet` (under the weights' variation, for ggH), with no specific behavior depending on the type of particle.

For the CP5 tune variations, we observe little fluctuations for the subregion `not_in_any_jet`, but greater ones (up to 30 % of the nominal) in `in_any_jet`. Nevertheless, there is still no clear offset of shape effect, and those fluctuations are likely due to few statistics.

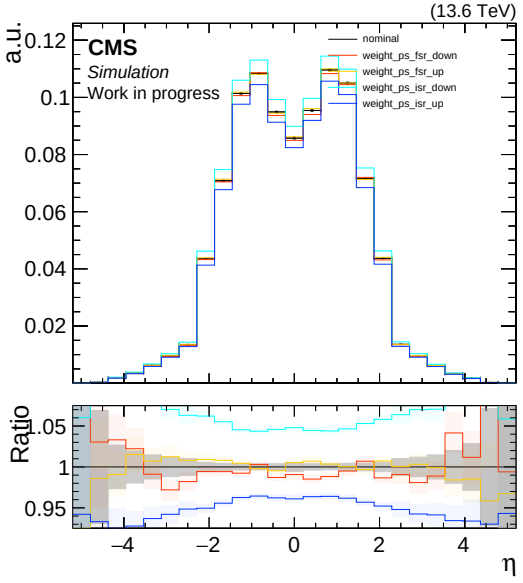


Figure 35: *eta* of the first 100 PFCs in ggH, `in_any_jet`, varying under PS weights

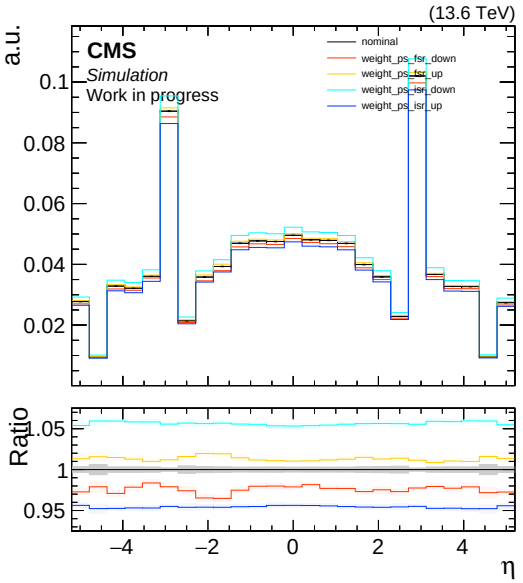


Figure 36: *eta* of the first 100 PFCs in ggH, `not_in_any_jet`, varying under PS weights

3 Conclusion

During this internship, we tried to probe the feasibility of using Graph Neural Networks (GNNs) to classify vector boson fusion from gluon-gluon fusion production modes of the Higgs boson. The study focused on quantifying the stability of Monte Carlo simulations, analyzing the impact of variations in parton shower weights and CP5 tunes on classification performance.

The results show that while the DNN score fluctuates within a 10-20% range under these variations, the ROC curves remain stable within 1%. The main deviations observed come from underlying events and are particularly visible when looking at the transverse momentum of the particle flow candidates.

Future work could focus on retraining the DNN with updated particle selection criteria to verify the result of this study and apply the same type of analysis on the CP5 tunes variations of the gluon-gluon fusion.

Acknowledgment

I'd like to thank Dr. Andrea Malara for his support throughout the internship, and above all for moving heaven and earth to enable me to complete it in the best possible conditions.
I'd also like to thank Dr. Santeri Laurila, Prof. Pascal Vanlaer and Prof. Stephane Goriely for their help and availability, even when the schedule was tight. Finally, I'd like to thank the CERN's Summer Student program team and participants for this amazing summer. Special thanks to Elio for our fascinating discussions on histograms.

References

[1] Griggs J. Birth of a boson: the Higgs has more than one daddy — newscientist.com. https://www.newscientist.com/article/dn22010-birth-of-a-boson-the-higgs-has-more-than-one-daddy/. 2012.

[2] Sicking E. CERN Summer Student Lecture Program 2024. https://indico.cern.ch/event/1347523/contributions/5672503/attachments/2887566/5062396/SSLP%20introduction%20-%202024.pdf. 2024.

[3] cern. CERN's summer student programme turns 40 – CERN Courier. *CERN Courier*. 2019;.
URL https://cerncourier.com/a/cerns-summer-student-programme-turns-40

[4] McCullough M. CERN Summer Student Lecture Program 2024. https://indico.cern.ch/event/1347523/contributions/5672503/attachments/2887566/5061208/SSLP.pdf. 2024.

[5] Le CERN | CERN. 2024.
URL https://home.cern/fr/about

[6] United Nations. Intergovernmental and Other Organizations | United Nations. 2024.
URL https://www.un.org/en/about-us/intergovernmental-and-other-organizations

[7] CERN Annual report 1955. *Tech. rep.*, CERN, Geneva. 1956.
URL https://cds.cern.ch/record/1475694

[8] Our History | CERN. 2024.
URL https://home.cern/about/who-we-are/our-history

[9] CERN. CERN Annual report 2023. *Tech. rep.*, CERN, Geneva. 2024.
URL http://cds.cern.ch/record/2897082

[10] CERN's main objectives for 2021–2025 | CERN. 2024.
URL https://home.cern/resources/brochure/cern/cerns-main-objectives-2021-2025

[11] Video-meeting: Restricted Council - Two-Hundred-and-Second Session. 2024.
URL https://indico.cern.ch/event/1010296/#54-cerns-main-objectives-for-t

[12] CERN. CERN Annual report 2022. *Tech. rep.*, CERN, Geneva. 2023.
URL https://cds.cern.ch/record/2857560

[13] CERN. CERN Annual report 2021. *Tech. rep.*, CERN, Geneva. 2022.
URL https://cds.cern.ch/record/2807619

[14] CERN. CERN Annual report 2020. *Tech. rep.*, CERN, Geneva. 2021.
URL https://cds.cern.ch/record/2771424

[15] CERN. CERN Annual report 2019. *Tech. rep.*, CERN, Geneva. 2019.
URL https://cds.cern.ch/record/2723123

[16] Final Budget of the Organization for the seventieth financial year 2024. Finance Committee - Three-Hundred-and-Eighty Nine Meeting. *Tech. rep.* 2023.
URL https://cds.cern.ch/record/2888205/files/English.pdf

[17] Final Budget of the Organization for the sixty-ninth financial year 2023. Finance Committee - Three-Hundred-and-Eighty Third Meeting. *Tech. rep.* 2022.
URL https://cds.cern.ch/record/2847387

[18] 2024 Annual Contributions to CERN budget | Finance and Administrative Processes Department. 2020.
URL https://fap-dep.web.cern.ch/rpc/2024-annual-contributions-cern-budget

[19] Our History | Diversity & Inclusion Programme. 2024.
URL https://diversity-and-inclusion.web.cern.ch/about/our-history

[20] CERN Annual Personnel Statistics 2023. 2023;CERN Annual Personnel Statistics 2023.
URL https://cds.cern.ch/record/2897705

[21] Careers at CERN. 2024.
URL https://careers.smartrecruiters.com/CERN/students

[22] Doctoral Networks. 2024.
URL https://marie-sklodowska-curie-actions.ec.europa.eu/actions/doctoral-networks?

[23] Landua F. The CERN accelerator complex layout in 2022. Complexe des accélérateurs du CERN en janvier 2022. 2022;General Photo.
URL https://cds.cern.ch/record/2813716

[24] The accelerator complex | CERN. 2024.
URL https://home.cern/science/accelerators/accelerator-complex

[25] CERN accelerating science. https://cmsexperiment.web.cern.ch/news/cms-detector-design, journal=CMS Detector Design | CMS Experiment.

[26] People Statistics | CMS Experiment. 2024.
URL https://cms.cern/collaboration/people-statistics

[27] icms-statistics. 2023.
URL https://icms.cern.ch/statistics/overview

[28] Organisation | CMS Experiment. 2024.
URL https://cms.cern/collaboration/organisation

[29] CMS-doc-3035-v23: CMS Constitution. 2024.
URL https://cms-docdb.cern.ch/cgi-bin/DocDB/ShowDocument?docid=3035

[30] How to Join CMS | CMS Experiment. 2024.
URL https://cms.cern/index.php/collaboration/how-join-cms

[31] CMS-doc-2597-v117: CMS Management Board Organigram. 2024.
URL https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=2597

[32] WebHome < CMS/DCS < TWiki. 2024.
URL https://twiki.cern.ch/twiki/bin/view/CMS/DCS/WebHome

[33] Neutelings I. Standard Model, tikz.net. https://tikz.net/sm_particles/.

[34] CERNYellowReportPageBR < LHCPhysics < TWiki. 2024.
URL https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageBR

[35] Denner A, Heinemeyer S, Puljak I, Rebuzzi D, Spira M. Standard model Higgs-boson branching ratios with uncertainties. *The European Physical Journal C*. 2011;71(9).

[36] Hayrapetyan A, al. Measurement of the Higgs boson production via vector boson fusion and its decay into bottom quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Journal of High Energy Physics*. 2024;2024(1).

[37] Tumasyan A, al. Search for invisible decays of the Higgs boson produced via vector boson fusion in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Physical Review D*. 2022;105(9).

[38] Baldi P, Sadowski P, Whiteson D. Enhanced Higgs Boson to $\tau^+\tau^-$ Search with Deep Learning. *Physical Review Letters*. 2015;114(11).

[39] Bierlich C, Chakraborty S, Desai N, Gellersen L, Helenius I, Ilten P, Lönnblad L, Mrenna S, Prestel S, Preuss CT, Sjöstrand T, Skands P, Utheim M, Verheyen R. A comprehensive guide to the physics and usage of PYTHIA 8.3. https://arxiv.org/abs/2203.11601. 2022.

[40] Sirunyan AM, al. Extraction and validation of a new set of CMS pythia8 tunes from underlying-event measurements. *The European Physical Journal C*. 2020;80(1).

[41] https://cms-nanoaod-integration.web.cern.ch/autoDoc/NanoAODv12/2022/2023/doc_DYJetsToLL_M-50_TuneCP5_13p6TeV-madgraphMLM-pythia8_Run3Summer22NanoAODv12-130X_mcRun3_2022_realistic_v5-v2.html#PSWeight.

[42] Maria LA. File:PPV, NPV, Sensitivity and Specificity.svg - Wikimedia Commons — commons.wikimedia.org. https://commons.wikimedia.org/wiki/File:PPV,_NPV,_Sensitivity_and_Specificity.svg.

[43] Qu H, Gouskos L. Jet tagging via particle clouds. *Physical Review D*. 2020;101(5).
URL http://dx.doi.org/10.1103/PhysRevD.101.056019

[44] Workman RL, Others. Review of Particle Physics. *PTEP*. 2022;2022:083C01.

[45] How CMS weeds out particles that pile up | CMS Experiment. 2024.
URL https://cms.cern/news/how-cms-weeds-out-particles-pile