Rediscovering the Higgs boson at the CMS experiment

L. Borek, N. Boudjema, J. Chen, D. De Azevedo Beleza, S. Juks, A. Khanna, J. Revesz, and C. Zhang

Supervised by Prof. Matthew Wing

Department of Physics and Astronomy, University College London

Gower Street, London, United Kingdom

E-mail: leon.borek.19@ucl.ac.uk, noor-ines.boudjema.19@ucl.ac.uk, jiajun.chen.19@ucl.ac.uk, dinis.beleza.19@ucl.ac.uk, stefania.juks.19@ucl.ac.uk, ashuit.khanna.18@ucl.ac.uk, janos.revesz.19@ucl.ac.uk, c.zhang.19@ucl.ac.uk

Abstract: This project focuses on the reconstruction of Higgs events at the CMS detector in the $H \to \tau\tau$ and in the $H \to ZZ$ channels using resources and tools available on the CERN Open Data portal. Those tools were thoroughly assessed to determine how accessible this project is to a person with a primary background in physics.

Contents

Acknowledgments

We wish to thank Prof. Matthew Wing for his guidance and for offering us invaluable help and advice through the course of this project. We also wish to thank the UCL High Energy Physics Group for providing us with access to the UCL HEP Cluster and allowing us to utilise it to run the reconstructed data analysis of this project. In particular, we would like to thank Dr. Edward Edmondson for his assistance in setting up the CMS environment on the Cluster and for providing us with all the necessary knowledge to smoothly run our analysis.

1 Executive Summary

The Higgs boson is an elementary particle whose existence was first postulated by Francois Englert, Peter Higgs and Robert Brout in 1964. It allowed for a unified description of the electromagnetic and weak interactions as gauge theories. The prescription describes how particles obtain mass. The existence of the Higgs was confirmed on the 4th of July 2012 at the European organisation for nuclear research (CERN) by both the ATLAS and CMS experiments at the LHC and is often viewed as one of the scientific breakthroughs of the century [2]. This scientific pursuit involved a 40-year effort and the involvement of tens of thousands of physicists and engineers, but the majority of the experimental data was restricted to the few people taking part in the search for the Higgs, the last missing piece of the electroweak theory. On the 11th of December 2020, CERN announced it would endorse a new open data policy for scientific experiments at the LHC to make scientific research more reproducible and accessible. This led to an array of CERN data being freely available on the open data website [3].

Our project aimed to reproduce CMS results for the Higgs analysis, in the τ final state, using the resources made available from this policy and open data tools to see how accessible the portal is. This involved a thorough testing of the Open Data resources and tools. In addition to this, we aimed to gain a deeper understanding of the Higgs discovery and Higgs physics through LHC published results.

We found that the qualitative data and tools made available by CMS allowed for a clear analysis; the only prerequisites were basic programming knowledge and an elementary understanding of histograms. However, reconstructing the published data was computationally expensive and hard to follow. It also required an advanced level of computing and programming. We were fortunate to have been granted access to the UCL HEP cluster where most of our analysis was run. Despite this, running the necessary code took too long to yield complete results.

2 Subgroups and task allocation

2.1 Conducting the project

- **Literature review subgroup:** Responsible for the research involved in preparation of writing the section "Background physics". Also provided the other subgroups with explanations on the theory around the Higgs boson when required.
  - **Leon Borek:** Responsible for understanding the research around the development of the Higgs Theory and about the Higgs search. Was tasked with understanding Higgs production mechanisms and the nature of Higgs decays, specifically $H \to b\bar{b}$, $H \to \gamma\gamma$, $H \to ZZ$, $H \to WW$, $H \to \tau\tau$. In depth background research was performed on the methods by which particles are: detected, identified, and discovered at the LHC. Further efforts were put towards understanding event reconstruction, simulation, and selection. Particle interactions and previous attempts made to discover the Higgs at the LEP and Tevatron were also looked into, including the Higgs Strahlung process. Additionally, was assigned the minute-taker role of the group.
  - **Dinis De Azevedo Beleza:** Tasked with writing about the experimental facilities involved in the search and finding of the Higgs. Furthermore investigated how the unification of the electromagnetic and weak theory was done in order to help writing that section. Additionally responsible for writing an introduction on Monte-Carlo Methods. Also covered the minute-taker role for certain meetings.
  - **Chenyu Zhang:** Tasked with researching information around the Standard Model and the search for the Higgs boson.
- **Level 2 data analysis subgroup:** Responsible for exploring and analysing the more accessible content of the CMS Open Data. The main purpose of this content was for education and outreach and building an understanding of the methods used for data analysis in particle physics. The various resources investigated include: visualisers, online and offline notebooks (mainly Python) and other possible resources.

- **Stefania Juks:** Became familiar with the CMS event display and its components to visualise, understand, and investigate different decays. Online, ran through the CERN' provided Binder Notebooks to dive deeper into the invariant mass histograms. Offline, recreated the histograms in Python with Jupyter notebooks to capture different effects: the influence of the number of data points, the importance of selecting relevant data, pseudorapidity and more. All of these were documented in the corresponding sections. Explored different educational content available through online resources. Additionally, researched and wrote on the unification of the electromagnetic and weak interactions section. Was assigned the communications officer role of the group.
- **Ashuit Khanna:** Became familiar with plotting histograms using Jupyter notebooks and Excel to identify the most relevant plots and datasets. Captured images from the histogram visualiser. Provided a section on histograms with images and explanation on how to use them for educational purposes. Researched and wrote about different other educational resources available online such as the Particle Physics Playground. Aditionally tasked to eplore the Higgs Mechanism's and symmetry breaking.

- **Level 3 data analysis subgroup:** Conducted the level 3 analysis part of this project. This involved using reconstructed CMS collision and simulated data provided by CERN Open Data to attempt to rediscover the Higgs boson. In addition, the subgroup was tasked with evaluating how open the portal was by investigating how accessible it would be to a regular user (i.e someone with a primary background in physics, not associated with CERN) by familiarizing themselves with the data processing pipelines and the softwares used within the CMS experiment.
  - **Noor-Inès Boudjema:** Researched and compared different computing options, such as VM and Docker. Used the UCL HEP Cluster to conduct the Higgs Awesome Workshop Analysis. Investigated CVMFS and the data pipeline/processing chain at CERN. Also tasked with researching information about the Standard Model and the Higgs Mechanism. Additionally, as the chair, she set up weekly agendas and circulated them prior to the meetings, and ensured the

project ran smoothly by regularly checking up with each subgroup and member. She also set up the Overleaf project used to compile this report.

- **Jiajun Chen:** Set up Virtual Box and Docker, recorded errors encountered and made comparisons between them. Used Docker to conduct fundamental Higgs analysis from Level 1 to Level 3. Set up the CMS environment on the HEP cluster with singularity and solved issues encountered. Used UCL HEP cluster to run Higgs analysis from Level 1 to Level 4. To be more specific, ran Level 4 Higgs analysis with 2012 datasets on the cluster parallelly.
- **János Révész:** Familiarised himself with the CMSSW and ROOT software frameworks, including the EDM, data pipeline and cmsRun syntax, root files and data structures. Set up and used the VirtualBox environment to run the Level 1-3 advanced Higgs analysis. Used the UCL HEP cluster to run the Level 4 analysis on the 2011 MC, 2012 MC and 2011 detector datasets, and wrote a python control script to automate the remote running of these analyses on multiple computers.

2.2 Writing the report

- **Leon Borek:** 3.2.3, 3.3.1, 3.3.2, 3.3.4, 3.3.5, 3.3.6
- **Noor-Ines Boudjema:** 1, 3.1.1, 3.1.2, 3.1.4, 4.1.1, 6.3.3, 6.3.4, 6.4.2, 6.5.1, 7, A
- **Jiajun Chen:** 6.4.1, 6.4.2, 6.5.2, References, A
- **Dinis De Azevedo Beleza:** 3.1.3 - d, e, 3.2, 3.3.3, Figures
- **Stefania-Alexandra Juks:** 3.1.3 - a, b, c, 4.1.2, 5.1, 5.2.3, 5.3.2, 5.4
- **Ashuit Khanna:** 5.2.1, 5.2.2, 5.2.3 - b, 5.3.1, 5.4
- **János Révész:** 6.1, 6.2, 6.3.1, 6.3.2, 6.5.2, 7
- **Chenyu Zhang:** N/A

3 Background

3.1 The role of the Higgs boson in the Standard Model

The Standard Model of particle physics describes with high precision our current understanding of the building blocks of matter and the forces governing them. In this section, we aim to provide a thorough explanation of the organisation of the Standard Model and the different elementary particles it describes. In addition, we will provide the reader with an overview of the different discoveries that led to constructing the Standard Model as we know it today and explain how the Higgs boson fits into it.

3.1.1 The evolution of the Standard Model since the 1950s

The Standard Model is a theory that describes the interactions of all known elementary particles: the three families of leptons and quarks and the gauge bosons responsible for the electromagnetic interaction, weak interaction, and strong interaction [4]. Missing in this description is Gravity as there is as of yet, no consistent framework for a theory of quantum gravity.

At the beginning of the 20th century, scientists discovered the structure of the atom through Rutherford's experiment. They fired alpha particles at a gold foil and noticed that some particles bound back: they had just discovered the electron. This technique of firing particles at each other is still a widely used method in particle physics experiments and constitutes the leading role of most experiments at CERN.

However, scientists met challenges when studying interactions between subatomic particles. In 1954, physicists Zhenning Yang and Robert Mills developed the framework of Yang-Mills gauge theories [5] extending the concept of gauge theory that was known for electromagnetism. This theory sets the ground for the description of charged, self-interacting gauge bosons and would be used later to describe the gluons in quantum chromodynamics, QCD, and the W, Z gauge bosons. Nonetheless, this theoretical description requires *massless* gauge bosons, like the photon and the gluons. The gauge principle predicts that each interaction is described by a unique coupling constant. For instance, the electromag-

netic coupling, as mediated by the photon, is the same between electrons, muons or any electrically charged particle.

Two years later, Jianxiong Wu proved that parity [6] is not conserved in weak interactions processes, contrary to electromagnetic processes. Another difference is that the electromagnetic force is long range, while the weak interaction is short range which suggests a heavy mediator.

In 1961, Sheldon Glashow theorised the unification of two of the four fundamental interactions: the weak and the electromagnetic interactions, into the electroweak force. This led to the development of the theory of the Higgs field, which was first suggested by Peter Higgs, Robert Brout, and Francois Englert in 1964. The Higgs mechanism would provide an explanation for the origin of mass within gauge bosons and a description within a Yang-Mills formulation.

3.1.2 Particle classification

Particle physics is governed by quantum mechanics. Most of the theory is bound by an array of quantum numbers, such as the lepton number, baryon number, spin and charge conservation. They set a foundation for the allowed decays and interactions. Every known subatomic particle may be classified into two primary groups. They are divided into matter particles, called fermions and into force carriers or mediators known as bosons.

Fermions possess a half integer-spin, and according to Fermi-Dirac statistics, they follow the Pauli-exclusion principle. This principle states that fermions cannot occupy the same state [7]. They may be divided into quarks and leptons, which are arranged into three families with increasing mass. Interactions between fermions are governed by the exchange of particles known as bosons.

Bosons are responsible for three of the four known forces of nature: the weak, the electromagnetic and the strong interactions. $W^{\pm}$ and Z bosons are exchanged in the weak interaction, photons (γ) mediate the electromagnetic interaction, while gluons (g) are responsible for the strong interaction. These force mediators are called gauge bosons. This name comes from their essential role in gauge theory and the fact that they follow Bose-Einstein statistics, which dictate the allowed quantum arrangements particles with an integer spin (such as bosons) may take [8]. All of the known elementary bosons, but one, have

spin-1. An important exception is the Higgs boson which has spin-0. As stated above, the quantum Standard Model does not describe gravity [9]. Many physicists, however, believe that gravity is mediated via a hypothetical particle called the graviton with spin 2. Gravity is very weak, making the experimental confirmation of such a particle nearly impossible with today's technology. The particle classification within the standard model is presented in Figure 1.

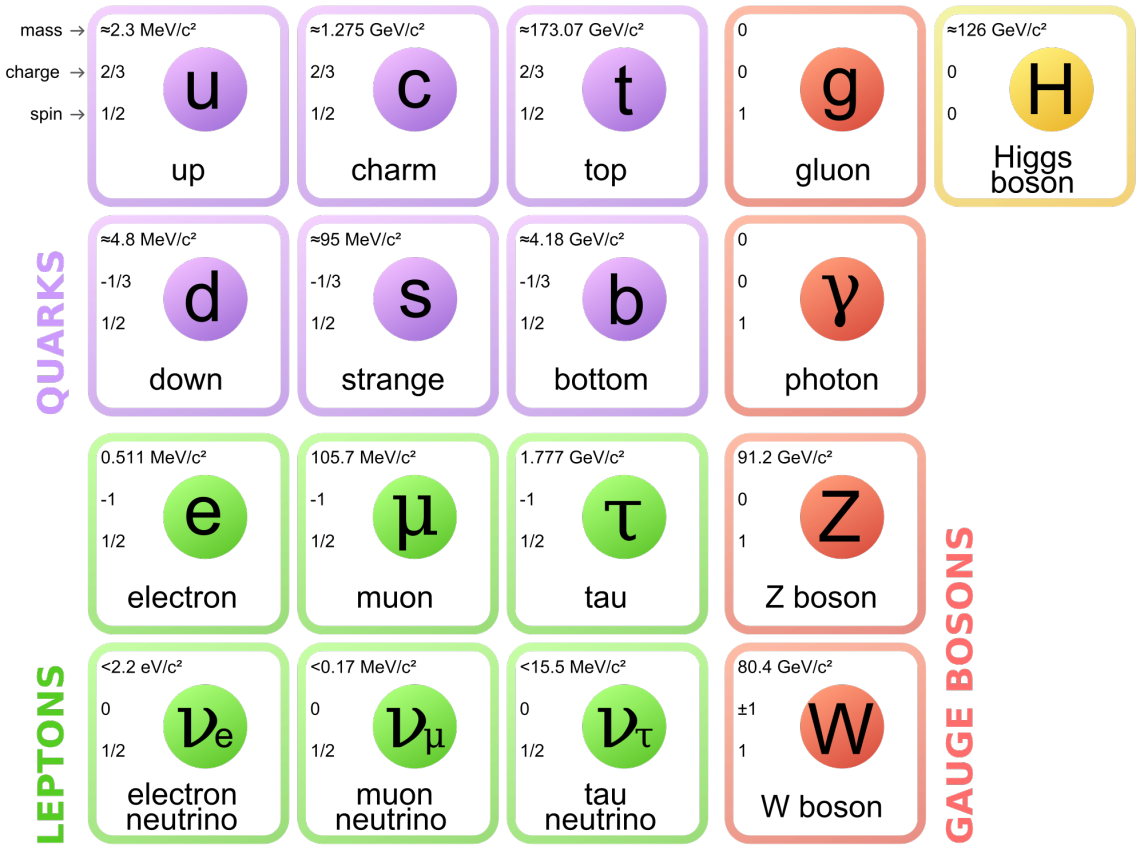


Figure 1. The particle classification within the standard model.

Quarks are strongly interacting fundamental fermions. There are six different quarks: top, bottom, strange, charm, up and down, which come in three different colour charges. Quarks are never found in isolation but are bound into colourless states known as hadrons which are divided into mesons and baryons. Mesons are made up of a quark and an antiquark, while baryons are formed from the association of three quarks [4].

Leptons are fermions that are not affected by the strong nuclear force. Just like quarks,

there are six different leptons. However, they are divided into three charged particles (electrons e^-, muons μ and tau τ leptons) and into three associated electrically neutral particles, which are known as neutrinos, ν_e, ν_μ, ν_τ, [4]. Each neutrino matches to one of the negatively-charged particles with the same name. Neutrinos are nearly massless and interact very weakly with matter, making them extremely hard to detect.

In addition, every particle is associated with an antiparticle that has the same mass but an opposite charge. For example, electrons possess an antiparticle known as the positron, which has the same mass but a positive electric charge rather than negative [10]. Mathematically, antiparticles are described as regular particles travelling backwards in time. However, this concept is disputed from a physical point of view as it violates the principles of causality. The positron, which was the first antiparticle ever observed, was discovered in 1932 by Carl D. Anderson during his study of cosmic rays, soon after its prediction by Paul Dirac [11]. The photon and the Z boson are their own antiparticle.

3.1.3 The unification of the electromagnetic and the weak interactions

a Introduction

The electromagnetic and the weak interactions were first thought to have nothing in common. A striking difference is that electromagnetism does not allow for unstable elementary particles while weak interactions are responsible for decays of heavy particles into lighter ones. In the late 1960s, Weinberg, Salam and Glasgow described how the electromagnetic and weak interaction could be treated as two different aspects of a single unified electroweak interaction. This model has been verified experimentally in the past three decades, with a high level of precision for many observables [12].

b Cross-sections

Cross-sections represent the appropriate observable measurements for scattering events. We will use, as an example, a beam of particles colliding into a fixed target to introduce this concept [13]. After the collision, various new particles are produced with specific rates of production. These rates are proportional to the number of particles that hit the target and the flux, defined as the number of particles per unit area and unit time [14]. Thus, the rate at which a particular reaction r appears in an experiment, W_r, can be written as:

$W_r = JN\sigma$ [13] where J represents the flux, N is the number of incoming particles on the target, and σ_r the partial cross-section as it is for only one reaction. To determine the total cross-section, σ_{tot}, we require a summation over all the reactions produced: $\sigma_{tot} \equiv \sum_r \sigma_r$ [5]. In practice, experiments require the measurement of a differential cross-section, which, for a solid angle, $d\Omega = d\cos\theta d\phi$ in the (θ, ϕ) direction, is defined as: $dW_r \equiv JN\frac{d\sigma(\theta,\phi)}{d\Omega}d\Omega$ [13]. The geometry of the differential cross-section is displayed in Figure 2. Other differential cross-sections can be defined if it depends on other variables than the scattering angles θ, ϕ.

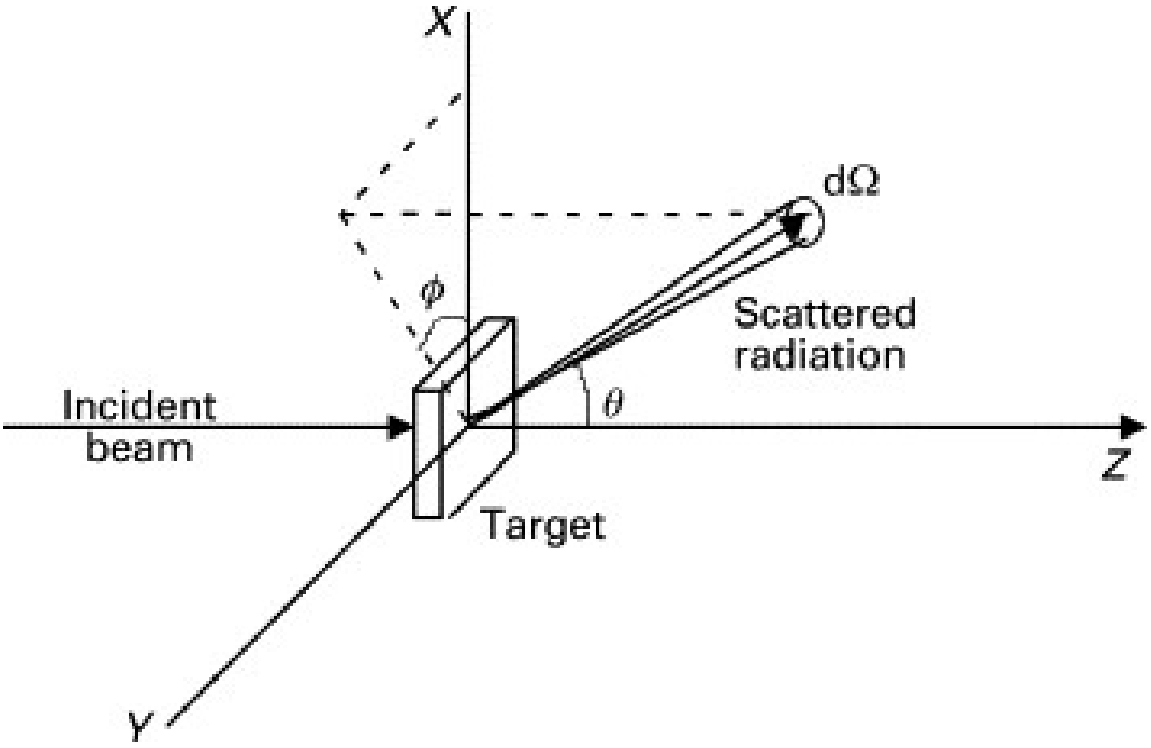


Figure 2. Diagram displaying the geometry of the differential cross-section. Figure from [13].

To determine the total cross-section, we integrate over all solid angles [12]:

$$\sigma_r = \int_0^{2\pi} d\phi \int_{-1}^{1} d\cos\theta \frac{d\sigma_r(\theta,\phi)}{d\Omega} d\Omega$$

.

c Early problems with the weak interaction

Early theories of the weak interaction behave well at low energies as it was treated as a four-point interaction with one vertex.

We will take, for example, a four-fermion interaction. The strength of the interaction is described by the Fermi coupling constant G_F. The problem is that with Fermi's four-fermion theory, the cross-section, which is a *probability* of production, would grow with energy, E, quadratically (as the amplitude of the cross section = $G_F E^2$). A quadratic increase with energy of the cross section leads to a violation of the unitarity limit; the scattered intensity cannot exceed the incident intensity in any partial wave [12].

At first, the W boson was introduced as a propagator, which solved the problem when one boson was involved. However, a unified theory was required due to the issues encountered when more than one W boson is exchanged. Therefore, this description leads to cross-sections with unacceptable values at high energy. This introduced the need for a mechanism to correct the unacceptable high energy behaviour.

d Electroweak unification

The Electroweak (EW) theory was proposed in the 1960s to give a unified description of the weak and electromagnetic interactions. This process of unification was similar to the one Maxwell achieved a century earlier when he unified the magnetic and electric interactions. Many successful predictions were made with the early version of electroweak theory. The theory predicted, alongside the photon associated to the electric current, the existence of a new neutral weak current in addition to the already known charged current reactions. Similar to the photon, the neutral weak current is mediated by a Z boson, while the mediator of the charged weak current is mediated by the $W^{\pm}$. The mass of the W and the Z were extracted from a combination of many observables. As mentioned earlier, the necessity of the gauge bosons being massive contradicts the principle of gauge invariance imposed by the Yang-Mills construction. The Higgs mechanism will provide a solution that exploits spontaneous symmetry breaking. One of the most important predictions that came from this mechanism was the existence of the Higgs boson, which is a spin-0 boson [12].

Before this theory was fully developed, higher order calculations of Feynman diagrams where several W bosons are involved, problems were encountered. The contribution of a

diagram at higher order was to be infinite. This was partially solved with the presence of the Z bosons and photons in the theory. The individual contributions for the diagrams involving these particles would still be infinite. However, when adding all the diagrams that contribute to the same process, the infinite contributions would cancel, and finite results were obtained. This cancellation follows from two important relations from the EW theory, which are the unification condition and the anomaly condition [13].

We emphasize that we can consider the electromagnetic and the weak interactions as separate forces to describe the interactions between particles at low energy. However, the full EW theory is needed at higher energies.

e The Weinberg-Salam Model

The EW theory was proposed in the late 60s by Weinberg, Glashow and Salam. Four massless mediating bosons were postulated. They were then divided in a triplet state and a singlet state that could be distinguished by the value of 'weak isospin', I. The triplet state had a value I = 1 and the bosons were denoted $W_\mu^{(1)}, W_\mu^{(2)}, W_\mu^{(3)}$. $W_\mu^{(3)}$ is electrically neutral gauge boson. On the other hand, the singlet state was denoted by a boson B_μ which is an isoscalar, I = 0. It is also electrically neutral. This was the basis for the Weinberg-Salam model. The triplet of W is described by a Yang-Mills gauge theory based on the non-abelian group $SU(2)$, while the singlet B field is described by an Abelian $U(1)$ gauge lagrangian much like the theory of electromagnetism, or QED: Quantum Electrodynamics. This is why the electroweak theory of Glashow, Salam and Weinberg is known as the $SU(2) \times U(1)$ theory. As stated before, gauge symmetry prohibits mass terms for the gauge bosons. We know however that there are three massive bosons, $W_\mu^+, W_\mu^-, Z_\mu^0$ and only one massless neutral gauge boson, the photon A_μ. Masses, as we will see next, can be introduced through the Higgs mechanism while maintaining the gauge symmetry of the system. The first two charged massive bosons correspond to W_μ^+, W_μ^-, while the neutral Z_μ^0 and A_μ are combinations of the states $W_\mu^{(3)}$ and B_μ, they mix once masses are introduced. Because of this mixing, the masses of the $W^\pm$ and the Z are not the same [12].

3.1.4 How the Higgs boson fits in the Standard Model

The Higgs mechanism is based on the idea of "Spontaneous Symmetry Breaking" or hidden symmetry. This mathematical theory was developed by Yoichiro Nambu in the 1960s for

which he received the 2008 Nobel prize, nearly five decades later. The idea is that a system is described by an equation, the Lagrangian, which is fully symmetric much the same as the function $f(x) = x^2 + a$ is symmetric in $x \leftrightarrow -x$. If $a > 0$, $f(x) = 0$ admits two solutions, $x = a$ or $x = -a$ but picking up only one of the two solutions makes the system asymmetric. To provide a better understanding of this key notion, we will use an example that was first formulated by T.W.B. Kibble [15]. The equation is provided by a scalar potential that is introduced in the theory. Higgs, Englert and Brout considered a toy model to give mass to one gauge boson, like the photon described by a $U(1)$ gauge symmetry. We can understand the mechanism of symmetry breaking with the aid of figure 3. In the first instance described by a bowl, the marble would just sit at the origin. This is the most symmetric point and also happens to be a stable point of minimum energy. The configuration is completely symmetric: the potential is symmetric around the origin and the marble sits at the origin. Consider now the case of a potential that is still symmetric around the origin but has a concave bottom (just like a wine bottle). If the white marble is at the top of the concave surface, it is at a symmetric position. However, that position does not corresponds to the minimum energy solution. The white ball at the origin is highly unstable. The ball will roll down, either to the left or to the right. The minimum energy solution, (stable configuration) can be found at one of the two lowest points of that concave surface. But when the ball has reached one of the two bottoms, the symmetry of the system is broken, as the third figure shows. The blue ball is at stable position but the symmetry is lost.

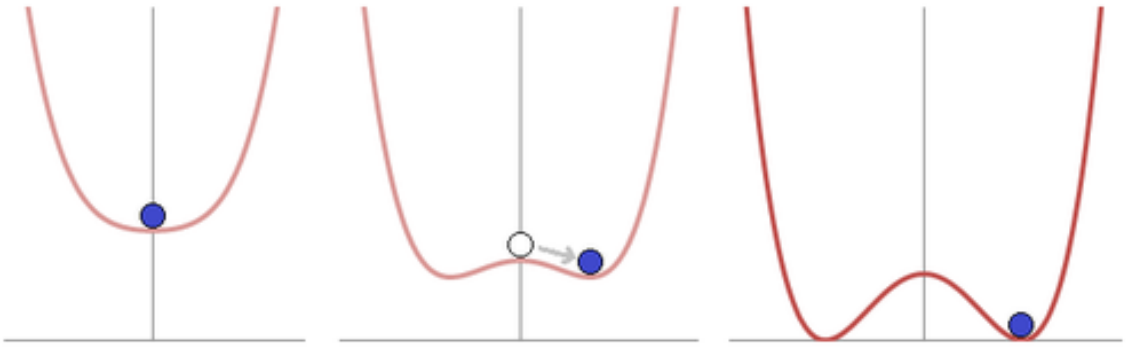


Figure 3. How symmetry breaking can occur. Image from [16].

This simple illustration borrows the key element from the potential used by Higgs

which he himself borrowed from the famous potential of Goldstone, called the Mexican hat or sombrero potential illustrated in figure 4). Here the most stable position is any position in the rim. We can see here that there is an infinite, continuous set of configurations, choosing any one solution breaks the symmetry.

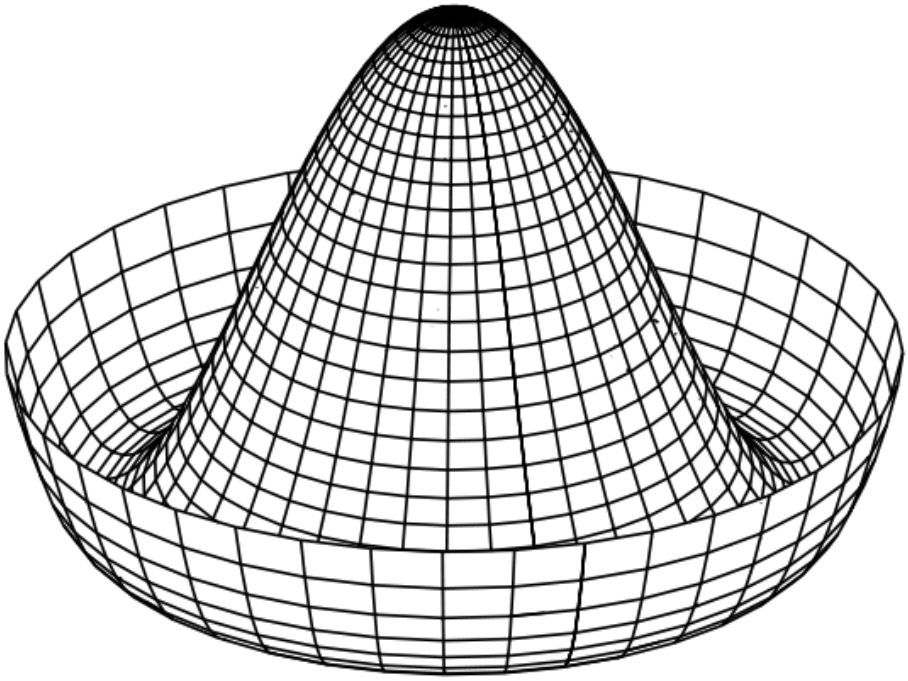


Figure 4. A graph for GoldStone's "sombrero" potential. Image from [15].

Spontaneous symmetry breaking, when applied to this situation, predicts the existence of a massless boson which is given the name of "Nambu-Goldstone" boson. The system prefers the lower energy state, away from the origin, at a position, v called the vacuum expectation value, because it is stable configuration. $v \neq 0$ introduces a mass scale. When the gauge interaction is made to couple to the scalar potential, the massless gauge boson and the massless Goldstone boson combine to give a massive gauge boson with a mass proportional to the vacuum expectation value, v.

Although frowned upon at first, it became evident a few years later that the Higgs mechanism was crucial in broadening the understanding around particle physics and was key in the development of the Standard Model.

This important mechanism was applied to the electroweak interaction based on $SU(2)\times$

$U(1)$. An appropriate Mexican hat potential was needed, so that three gauge bosons could acquire mass. Therefore, three Goldstone bosons were needed. The Higgs mechanism in the standard model exploits a scalar doublet field for the electroweak theory. It consists of three Goldstone bosons with two charged scalars and one neutral together with the electrically neutral physical Higgs particle. The three Goldstone bosons are "eaten up" by the W^+, W^- and Z so that they acquire mass, and gives them a longitudinal polarisation. This mechanism introduces a mass scale, the vacuum expectation value of the Higgs field, the value that minimises the potential, with $v = 246GeV$. The electroweak gauge symmetry $SU(2)_W$ (weak-isospin) symmetry and a $U(1)_Y$ weak-hypercharge gauge symmetry[7] mix to give the electromagnetic gauge interaction and the weak interaction with the three gauge bosons W^+, W^-, Z. The gluons are described by an $SU(3)$ gauge symmetry (with $3^2 - 1 = 8$ gluons) which is not subject to the Higgs mechanism, so the gluons remain massless.

The mechanism also gives mass to fermions. An important property of the Higgs mechanism is that the interaction of the Higgs boson, H, with the fermions and gauge bosons is proportional to the mass of the particle the Higgs couples to. In this respect, the largest couplings of the Higgs are to the top, Htt, the W and Z boson, HZZ, HWW. This important characteristic explains the properties of the Higgs boson: its decay and mode of production.

3.2 The Experimental apparatus

3.2.1 A history of CERN

The first ideas to create a physics research facility arose around 1940 during World War II [17]. Following the example of other organisations, a group consisting of some of the most well-known scientists such as Louis de Broglie and Niels Bohr came up with the idea of a European Atomic Physics Laboratory. Since this area of physics is very costly, the laboratory would enable different scientists and countries to split up the costs of the facility. In December 1951, the first resolution regarding the creation of a European Council for Nuclear Research was adopted. Two months later, eleven countries (Belgium, Denmark, France, the Federal Republic of Germany, Greece, Italy, the Netherlands, Norway, Sweden, Switzerland and Yugoslavia [18]) signed the resolution and the acronym CERN was born.

When all the details were decided, the draft convention was finished and made available to sign by the original member states plus the UK [18]. It was accepted and signed by all of them. It described how the financial contributions to the organisation were carried out. To this day, CERN has increased the number of member states from the 12 founding states to 23 states [18]. However, CERN has agreements with many other countries to exchange information and to do co-operations.

Since its beginnings, CERN has built and replaced several of its accelerators [17]. The synchrocyclotron (1957) was CERN's first accelerator and mainly focused on nuclear physics. The Proton Synchrotron was built in 1959 and was the first accelerator to focus on particle physics. In its collider (ISR – Intersecting Storage Rings), protons collide with each other resulting in the creation of an array of particles. The information retrieved thanks to the collider provided physicists with important information to build the basis of the Large Hadron Collider.

In 1976, the Super Proton Synchrotron was built. The SPS had a 7km tunnel used to accelerate proton beams, which led to higher velocities and collision energy than seen previously. Three years later, the SPS was changed to a proton-antiproton collider and from these collisions the W and Z bosons were discovered. The SPS was changed again one more time into a heavy-ion collider before the Large Electron-Positron collider (1988) (LEP) tunnel was built. The LEP tunnel was a 27km ring that was used again to make particles collide and study the new particles created from those collisions. In 2000, the LEP

was shut down to build new detectors to be used in the new collider, the Large Hadron Collider (LHC) (2008). The LHC uses the LEP tunnel to collide hadrons and study the particles created from those collisions. The main discovery done using the LHC was the Higgs boson in 2012. Nowadays, the LHC is still running, however there are already plans to create a new tunnel of 100km ring. This would significantly increase the ranges of energies to be studied in the collisions. CERN's accelerator complex can be seen in Figure 5.

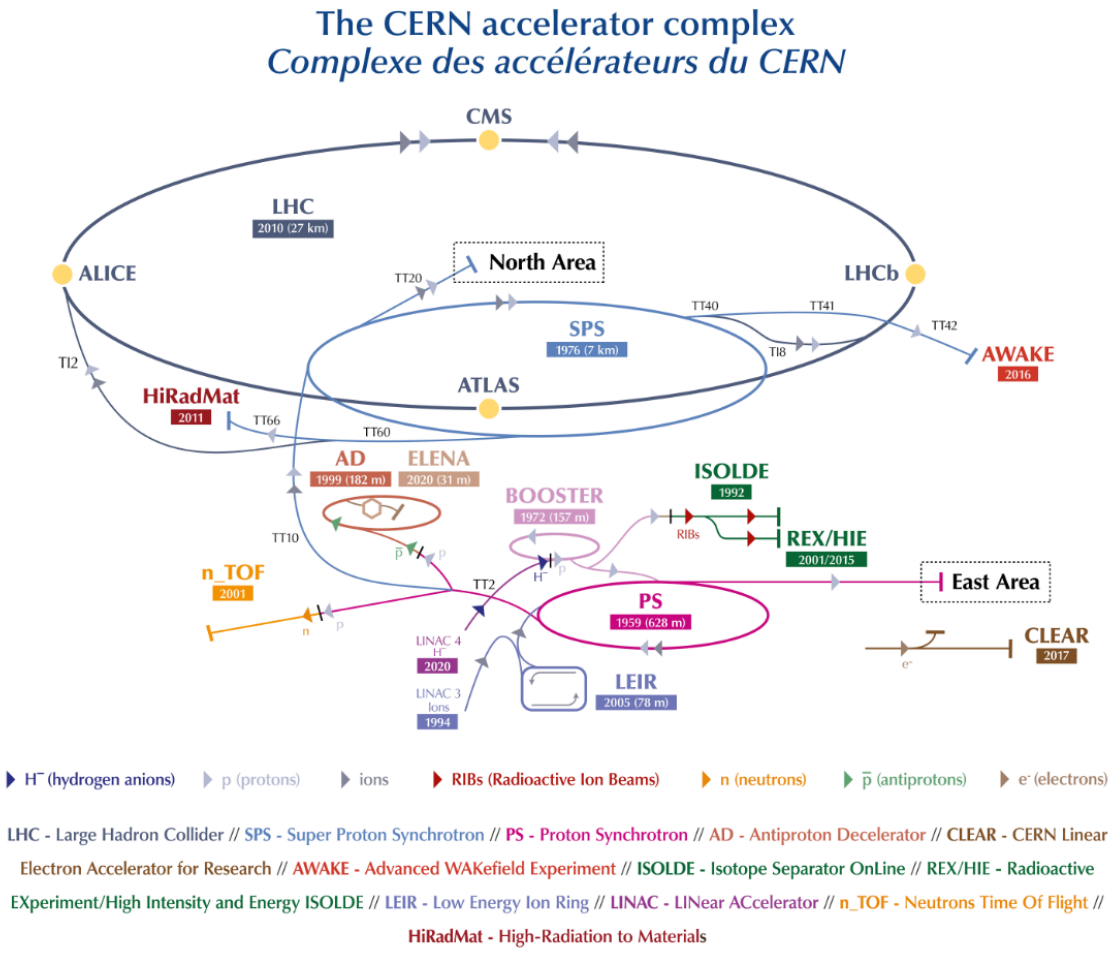


Figure 5. Picture showing the particle accelerators and colliders in CERN. Figure from [19].

3.2.2 The Large Hadron Collider (LHC)

Accelerators, such as the LHC, are used to study the properties of particles [20]. They accelerate charged particles, like electrons and protons (the LHC studies hadron collisions like protons), with velocities close to the speed of light and collide them onto a target or against other particles moving in the opposite direction.

The accelerators steer and accelerate particles using electromagnetic fields. The radiofrequency cavities increase the velocity of the particles by giving them an electrical impulse. A diagram of a radiofrequency cavity is available in Figure 6. There are also two types of magnets involved in this process [21]. The lattice magnets keep the beams stable and aligned by bending and tightening their trajectory. The insertion magnets are located right outside of the detectors and are responsible for squeezing the particles closer together so that they collide with an incoming beam travelling in the opposite direction inside the detector.

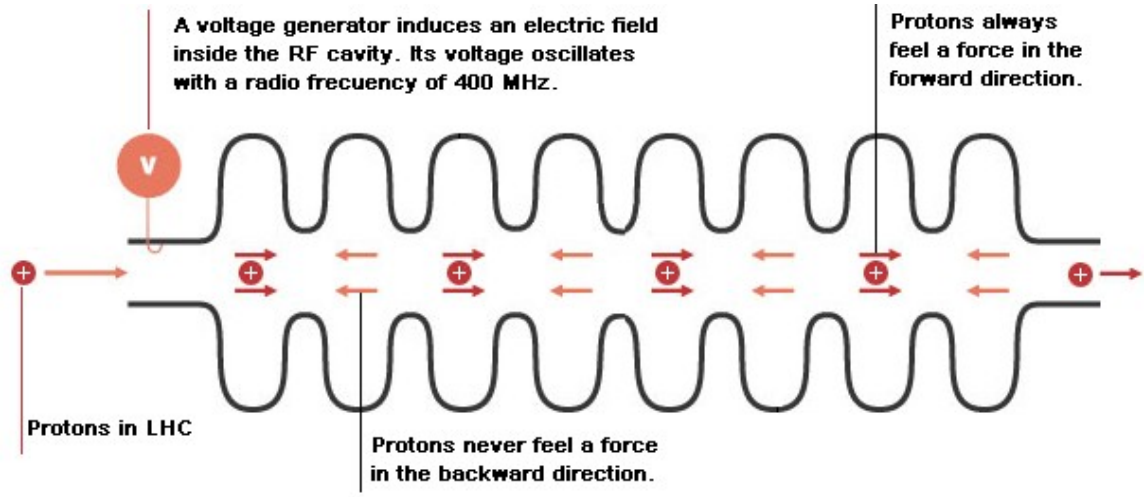


Figure 6. Diagram representing how particles interact with the electric field inside radiofrequency cavities. Figure from [22].

Nowadays, The LHC [23] is the world's largest and most powerful particle accelerator. The first beam of protons was steered in the LHC on the 10th of September 2008. The LHC has a ring size of 27km. This accelerator was built with the intent of answering [24] many questions such as: what gives matter its mass? Why is there more matter than antimatter in the universe? How have the universe components evolved?

Since the LHC was founded, it has been consistently breaking world records for collision and beam energy . This is due to the fact that, unlike other accelerators, it uses protons. Prior to the LHC, there used to be another accelerator at CERN (LEP) that used electrons and positrons, rather than protons. Protons are much more massive than those two particles (938MeV compared to 0.5110MeV), and because of a phenomena known as the Synchrotron radiation, the beam and collision energy created at the LHC are much greater than at the LEP.

Synchrotron radiation takes place in circular accelerators, such as the LHC and the LEP. It refers to the electromagnetic radiation emitted when charged particles travel in curved paths. Lighter particles emit more power due to this phenomena. As a result, protons retain more energy than electrons in collisions, with electrons radiating Synchrotron radiation at approximately 10^{13} the rate of protons: a significant energy loss. This explains high differences in energy outputs between the LHC and the LEP.

In 2009, CERN recorded collisions of 2.36TeV (1.18TeV per beam) in the LHC, in 2010, collisions at 7TeV (3.5TeV per beam) and finally, in 2012, collisions at 8TeV (4TeV per beam). It was also in 2012 that the first experimental evidence for the Higgs Boson was found. As said in section 3.2.1, particle physics is a very costly experimental area, this can be checked by the cost of the LHC. Its construction cost 6510 million CHF (Swiss Franc) and its maintenance and upgrades cost an extra 100MCHF. After the first shutdown to do maintenance and upgrades, CERN recorded collisions of 13TeV (6.5TeV per beam) in the LHC detectors.

It should be noted that particles are not only accelerated within the LHC. The protons originate from a hydrogen bottle and travel through the Duoplasmatron proton source. From there, they are accelerated a first time in the LINAC linear accelerator before accelerating further at the Proton Synchrotron Booster. They undergo the same procedure at the Proton Synchrotron and the Super Proton Synchrotron. Only after these procedures do the protons enter the LHC.

The LHC has four detectors within its ring where particles collide. These detectors are known as ATLAS, CMS, Alice and the LHCb. The first were the two main detectors involved in the discovery of the Higgs boson. It should be noted that in order to get extra reliability, scientific findings from each experiment were kept secret from each other. Both

detectors however discovered the Higgs and found that its mass was at around 125MeV.

3.2.3 Prior experiments that contributed to the finding of the Higgs boson

a The LEP - the Large Electron-Positron Collider:

As mentioned previously, the LEP was the first collider built in the tunnel where the LHC lies today. This means that by the time it was built, it was and actually still is the largest electron-positron collider [25]. It started being excavated in February 1985 and was completed in three years. When it started operating, the LEP had 128 accelerating cavities and 5176 magnets. The experiments were observed in four different detectors: ALEPH, DELPHI, OPAL and L3. On 14 July 1989, the first beam circulated in the LEP. Its initial energy was 91GeV, which corresponds to the mass of the Z boson with the aim of enabling physicists to create and investigate the particle. W and Z bosons, which are responsible for the weak force, were discovered in 1983 at CERN. The LEP operated for 7 years at about 100GeV and produced around 17 million Z bosons.

In the following years, the LEP was upgraded by the addition of superconducting accelerating cavities, with the aim of studying W bosons. The collider's energy reached 209GeV in 2000. In principle, this is sufficient to produce the Higgs boson but requires the Z boson to be off-shell. This means that the Z boson does not satisfy the equations of particle motion. This is demonstrated by the fact that in every collision, 91GeV are used in the decay of the Z boson. This means that we only have access to 118GeV, which is lower than the mass of the Higgs boson at 125GeV. As a consequence, the portion of Higgs bosons produced at the LEP is extremely small; e^+e^- colliders do not have the means to produce a high amount of Higgs bosons. However, while the LEP was unable to produce and detect Higgs bosons to a sufficient extent, it allowed scientists to conclude that its mass was superior to 115GeV, narrowing the search for the Higgs Boson. It should be noted that the collider could not create an energy higher than 209GeV because of the Synchrotron radiation. Positrons and electrons may collide and produce a Z boson which may effectively radiate: this is known as the Higgs Strahlung process. Once the accelerator centre of mass energy reaches the combined masses of the Z and Higgs, this process becomes a preferential occurrence with an extremely low background. This process is hence very attractive and provides a promising reason to create higher energy electron positron colliders in the future. Thus one can conclude that at the LEP, any potential

Higgs production must have occurred through vector boson fusion, just not enough to be clearly distinguished from the background. As stated, this signal is not observable due to how low the cross section is for this Higgs production mechanism in accompaniment with LEP energies. Figures 7 and 8 demonstrate these potential Higgs production routes:

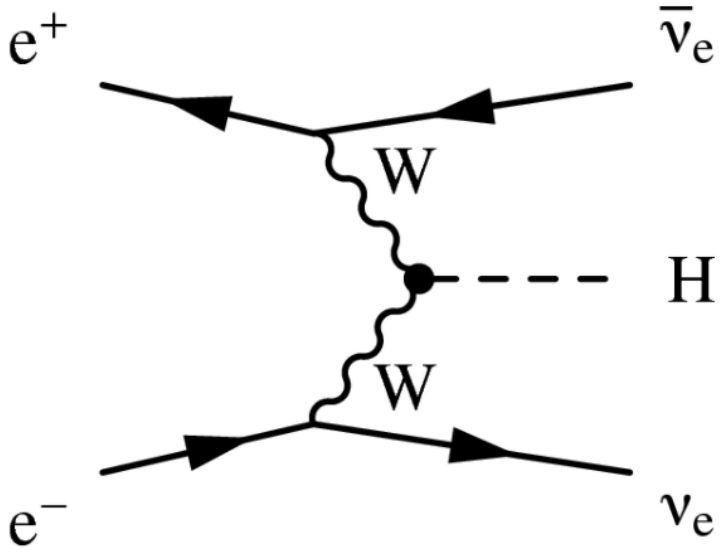


Figure 7. Higgs production via vector boson fusion. Figure from [26].

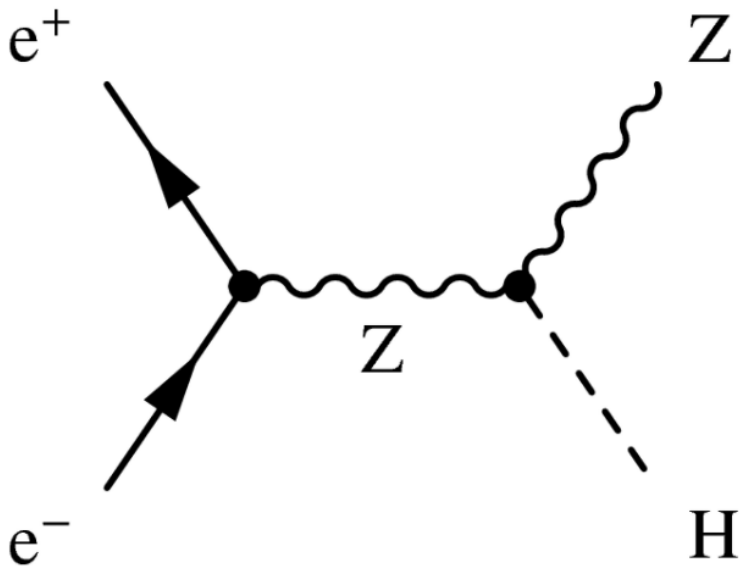


Figure 8. Higgs Strahlung process. Figure from [26].

On the 2nd of November 2000, LEP was closed and the construction for the LHC started. An important finding by the LEP was the confirmation of the fact that there are only three generations of fermions (matter particles).

b The Tevatron:

The Tevatron is another collider that participated in the finding of the Higgs Boson. It was located in Illinois, USA, and is part of the Fermilab facilities. It was the second most powerful particle accelerator in the world by the time of its shutdown in September 2011, just after the LHC [27]. The detector was able to achieve a maximum energy of 1TeV (almost five times higher than the maximum energy output of the LEP at 209GeV) The Tevatron was built in a 4-mile circumference tunnel buried 25 feet below ground level and had more than 1000 superconducting magnets, which is much smaller than the one used for the LEP (27km). However, collisions mainly took place between protons and antiprotons, which are much more massive than the electrons and positrons used in the LEP (938MeV compared to 0.5110MeV). This resulted in a much higher energy output in the Tevatron than the LEP due to the Synchrotron radiation (also called Magneto-Bremsstrahlung). The Tevatron used two detectors to study collisions: the CDF and the DZERO. By the time the Tevatron was built, it was already known that the Higgs boson had a mass superior to 115GeV. As a result, the Tevatron primarily focused on finding a range for the mass of the Higgs boson. It found in the 2010s that the Higgs mass in fact laid between 115GeV and 140GeV. This was confirmed by the LHC in 2012.

The production mechanisms at the Tevatron are the same as those at the LHC, even though there are less gluons compared to the LHC. The Tevatron was able to obtain a 3.1σ excess for the Higgs [28], but this was small considering how many years the data had been taken over. The answer as to why this excess wasn't larger is two-fold: Firstly, compared to the LHC, the instantaneous luminosity of the Tevatron was relatively small [26] (i.e. a much less number of events would occur within the same time).

The second reason, which is also the general answer to why finding the Higgs was so difficult, lies in the way the Higgs decays. A detailed chart-pie of the branching ratios for a 125GeV Higgs is displayed in Figure 9:

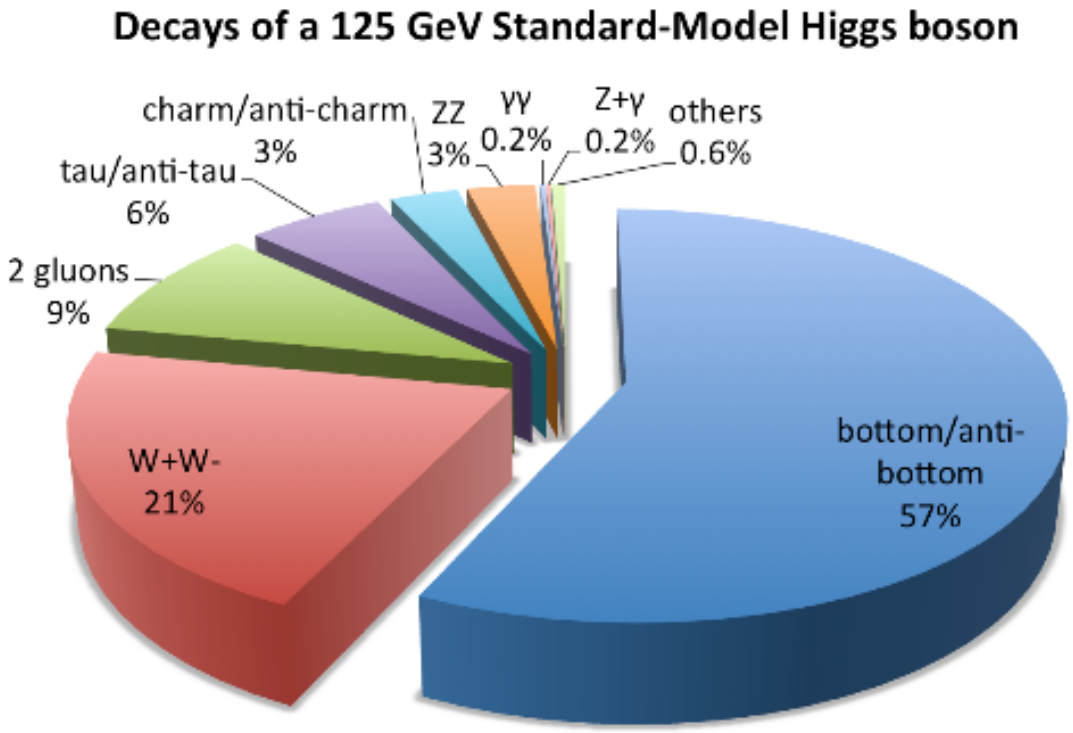


Figure 9. A pie chart of the Higgs boson's decay channels, as well as their branching ratios. Figure from [29].

The cleaner Higgs signatures which are more sensitive with higher mass resolution at a Higgs mass of 125GeV are $H \to ZZ$ and $H \to \gamma\gamma$; their branching ratios however are significantly smaller than other decay channels, as displayed in Figure 9. While these would have been the ideal channels to investigate in order to obtain a larger excess as the background is much smaller, a number of events beyond the scope of data storage at that time would have been required to obtain enough data given how low the branching fractions for those channels are [26].

It is noteworthy that if the Higgs were to be lighter, it would have been produced copiously at LEP and hence acceptably discovered far earlier. If it were heavier, the branching fraction to the b quark pair would have been considerably smaller and thus it would have been observed at the Tevatron. The mass of 125GeV simply turned out to be a signal which was hard to distinguish from the background, being the reason why it took

a long time to discover the Higgs [26]. Some of the Tevatron achievements [30] include the discovery of the existence of the top quark and five baryons which helped to test and refine the Standard Model. The Fermilab's accelerator complex can be seen in Figure 10.

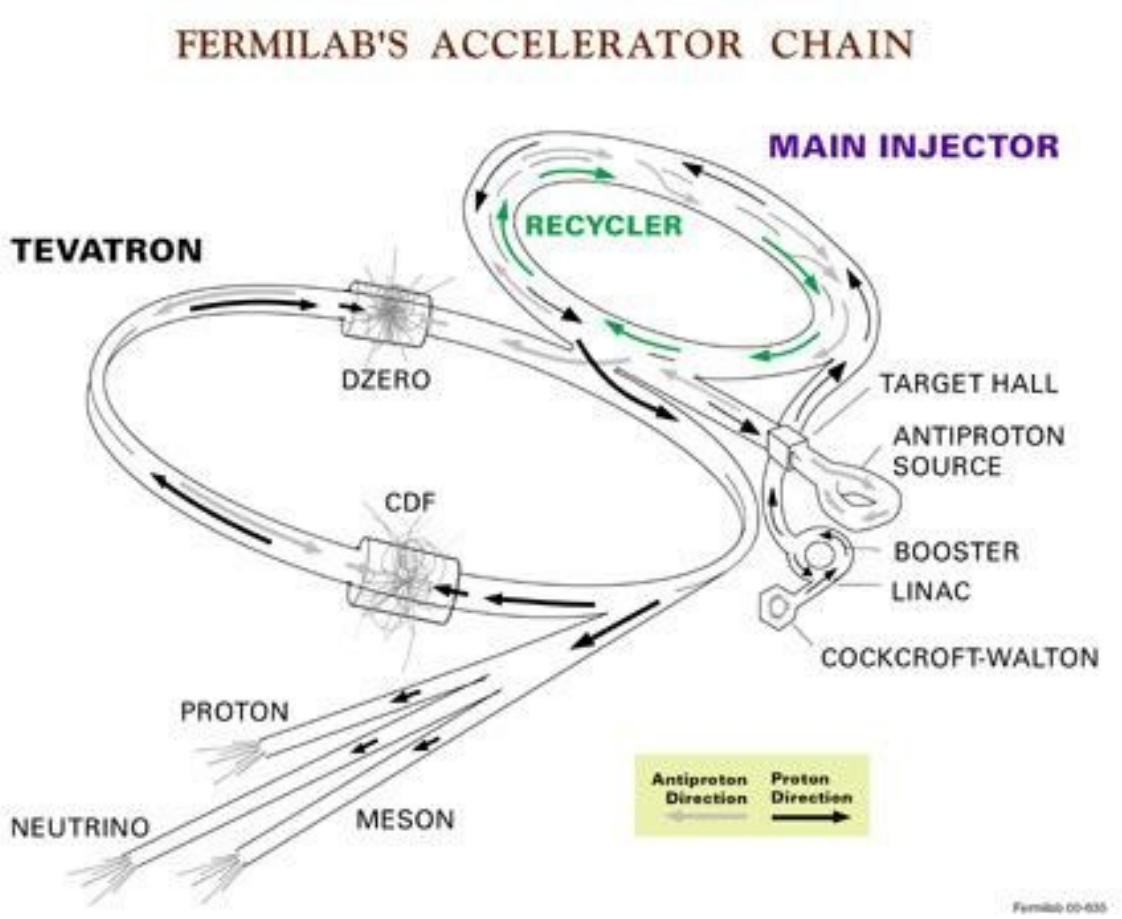


Figure 10. Fermilab's accelerator chain. Figure from [30].

3.2.4 ATLAS – A Toroidal LHC Apparatus vs CMS – Compact Muon Solenoid

ATLAS and CMS are the two LHC detectors which were involved in the discovery of the Higgs boson in 2012. Both are general purpose detectors which means that they possess the same functionalities and components (which only vary in dimensions). They however use different technical solutions and magnet systems. A diagram of a general-purpose detector can be seen in Figure 11.

A general purpose detector contains 6 layers. Particles travel from the innermost to the outermost layers, which include:

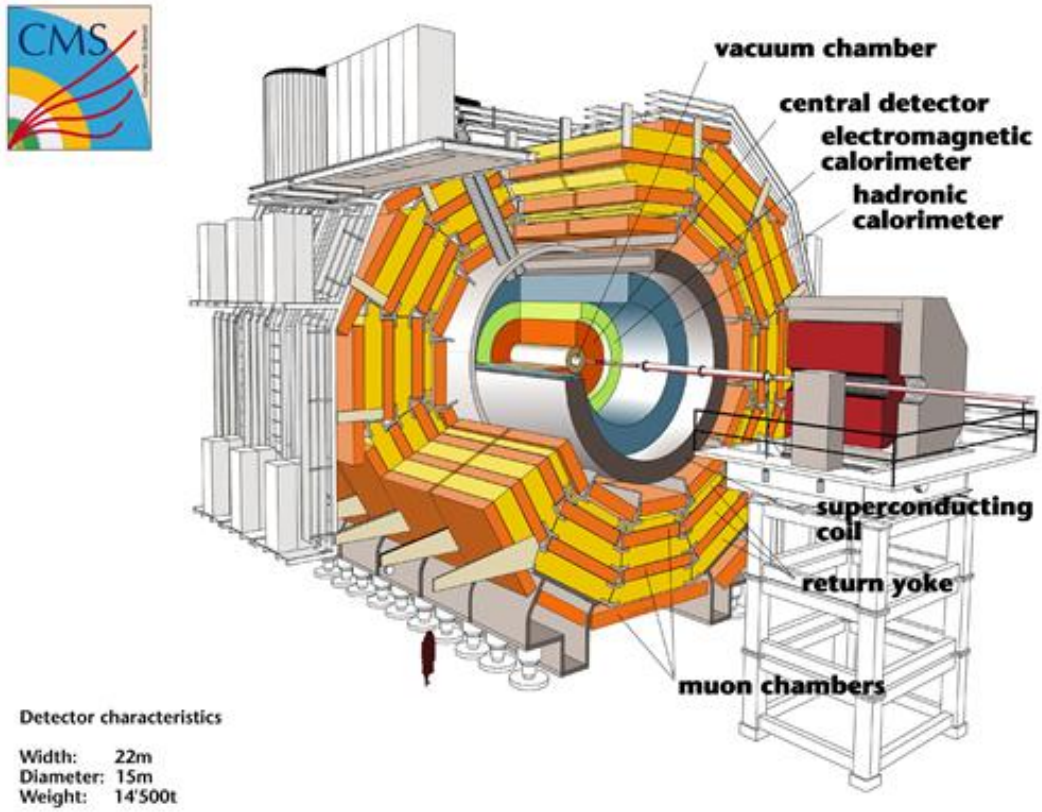


Figure 11. Cutaway view showing the outer four layers for detecting muons (interleaved with three layers of iron), the central calorimeters and the inner tracking system. Figure from [31].

- Vacuum Chamber: this is where the particles collide
- Central detector/silicon tracker: The charged particles interact with the silicon which creates a charge, known as a hit. This hit is recorded in one of the 75 million electronic sensors. A particle creates several hits as it travels through the tracker which are then used to determine the particle's path.
- Calorimeters: Calorimeters measure the energy of the particles which pass through them, detectors contain two: an electronic calorimeter (ECAL) and a hadronic calorimeter. The ECAL measures the energy of electrons and photons by stopping them completely. Hadrons, neutrinos and muons, however, only lose a small amount of energy due to ionisation when going through the ECAL and continue their path until they reach the hadronic calorimeter. There, hadrons are stopped completely and their energy is measured. Calorimeters can stop most known particles except muons and neutrinos; muon chambers responsible for muon detection usually make up the outermost layer of a detector due to the fact that muons interact very little with matter and so easily pass through inner subdetectors [32].
- Superconducting coil [33]: This magnet is used to bend the trajectory of charged particles. This is useful for two reasons: firstly, charged particles bend in opposite ways if they have opposite charges, allowing us to assess whether they are positive or negative; secondly, this allows scientists to measure the momentum of the particle since high-momenta particles bend less than low-momenta particles within the same magnetic field.
- Muon chambers: Here, the muons are finally stopped, allowing the measurement of their energy.

It is important to note that neutrinos pass through all these layers without being detected.

CMS:

The CMS is built around a large solenoid magnet [34]. This magnet takes the form of a

cylindrical coil of superconducting cable and produces a magnetic field of around 4 Tesla. For comparison, the magnetic field of the Earth is about 100 000 times smaller. This field is confined by a steel "yoke" that forms the bulk of the detector's 14,000-tonne weight.

The CMS, even though involved greatly in studying the Standard Model, also has other purposes such as searching for the evidence of other dimensions or finding particles that could make up dark matter.

The CMS is 21 metres long, 15 metres wide and 15 metres high. One curiosity about CMS is that it was the only detector from the LHC collider that was not built in the place where it is today. Indeed, its components were built in 15 different places and then transported to the cave where it is today to be reassembled. An image of the CMS can be seen in Figure 12.

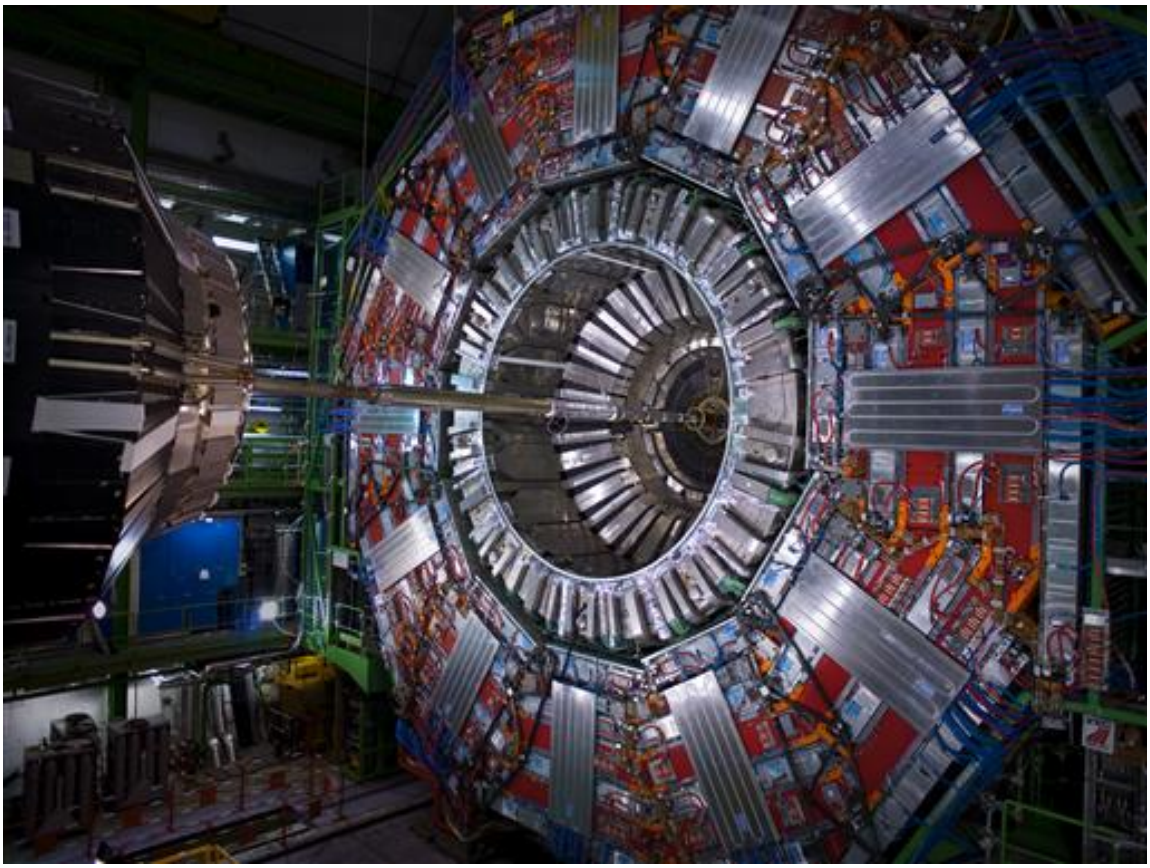


Figure 12. The CMS detector. Figure from [31].

ATLAS:

The magnet system of the ATLAS detector differs from CMS' [35]. ATLAS contains a central solenoid magnet that creates a magnetic field of 2 Tesla. ATLAS also has 6 barrel toroid magnets and two end-caps toroid magnets which reach a magnetic field of 4 Tesla. ATLAS is 46 metres long, 25 metres high and 25 metres wide, making it the largest particle detector in volume in the world. It weighs 7000 tons. An image of the ATLAS can be seen in Figure 13.

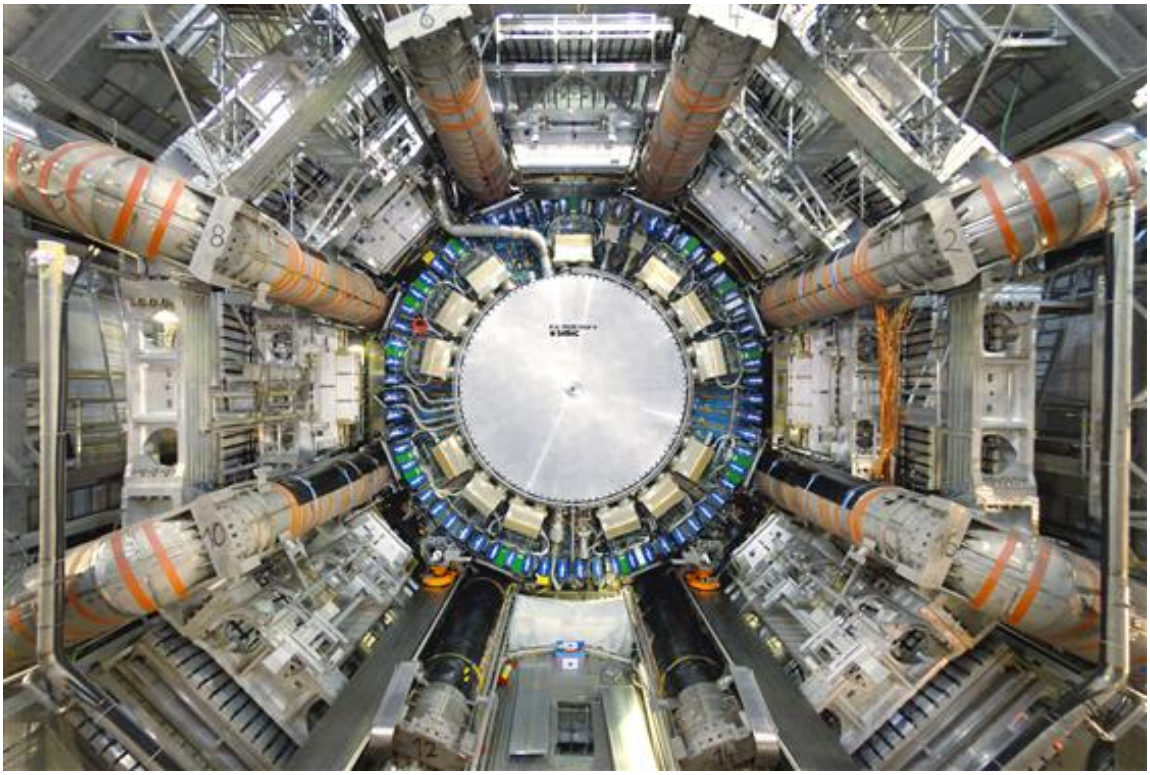


Figure 13. The ATLAS detector. Figure from [36].

3.3 Finding the Higgs Boson

3.3.1 Higgs production mechanisms at the LHC

Cross sections quantify the likelihood of a given particle's creation or decay. According to the standard model, the Higgs boson may be produced through different processes. These include, in decreasing order of cross sections: the gluon-gluon fusion (ggF), the vector boson fusion (VBF), the vector boson associated production (VH) and the top quark associated production ($tt \to H$) [37]. Figure 14 presents the Feynman diagrams of these processes. We shall elaborate on each production mechanism below:

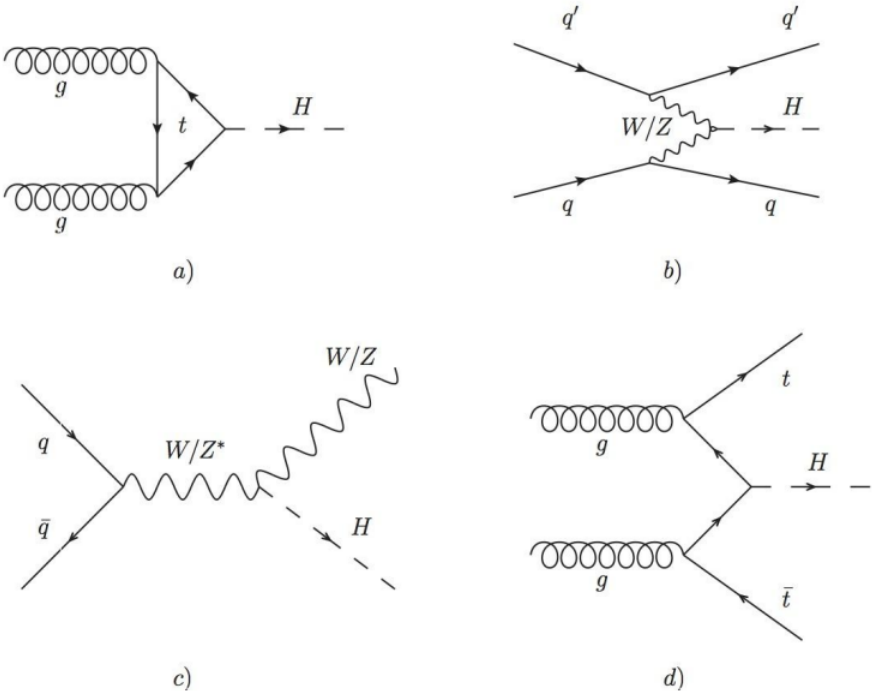


Figure 14. 4 most prominent Higgs production mechanisms in high energy pp collisions. (a) ggF (b) VBF (c) VH (d) $tt \to H$. Figure from [38].

The most dominant Higgs production mechanism is the gluon-gluon fusion (ggF) process. When the accelerated beams collide, there is a high likelihood of two gluons interacting through a virtual loop of heavy coloured top quarks that produces a Higgs boson; this production mode has the highest cross section for Higgs production at the LHC [38].

Following this in prominence is the vector boson fusion (VBF). This occurs when each of the incoming quark emits a W or Z boson. The fusion of these two vector bosons

produces the Higgs, $W^+W^- \to H$ or $ZZ \to H$. The Higgs boson is accompanied by two energetic jets in opposite directions [38].

During vector boson associated production (VH), the incoming quark pair inside the protons produces a virtual vector boson which splits into a Higgs and an associated vector boson [38].

Out of the four main Higgs production mechanisms in high energy hadron collisions, the top quark associated production, $t\bar{t}H$ or top-anti top fusion [38], is the least contributive as shown in Figure 15.

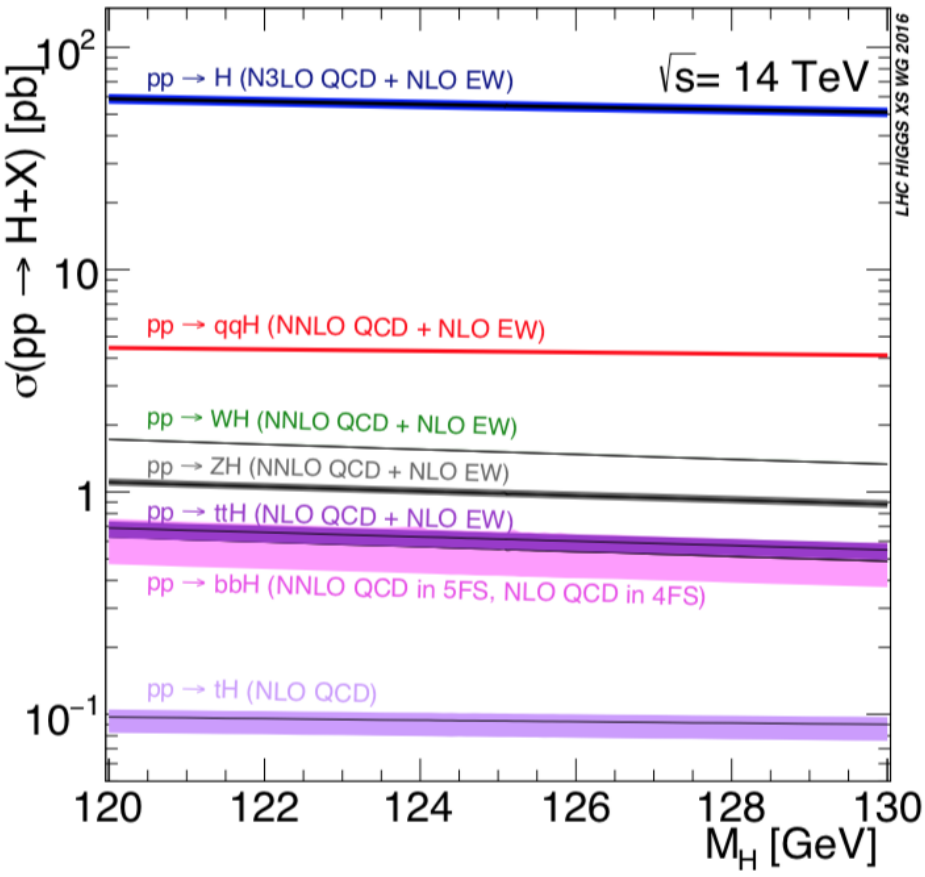


Figure 15. The cross-sections of the main 4Higgs production mechanisms at the LHC. Figure from [38].

3.3.2 Particle Identification

Quantitative characteristics of particles, such as their speed, mass, and charge, can be extracted from the particle detectors and thus allow for the identification of these particles. The information extracted can be implemented into a computational simulator that can reconstruct events. The most commonly used method for this is the Monte Carlo technique which, with respect to particle physics, produces a range of possible origins for a detected particle. It can also take into account intermediary pre-detection encounters that might affect the particles' properties, for example energy losses through collisions with the matter in the detector. This highlights the necessity for calorimeters to stop or 'absorb' particles produced from collisions; this forces them to deposit all of their energy within the detector so that the calorimeters can measure their full energy.

Once the properties and paths of these particles have been calculated and simulated, they can be evaluated as decay products and the invariant masses of these particles can be determined in order to identify a new parent particle. The invariant mass is the component of a system's mass that is independent of motion, otherwise known as its rest mass.

Using the example of a Higgs decay to a dimuon pair channel [39], the invariant mass of two detected muons is taken over a large number of total particles detected to create a plot of invariant mass to determine whether there is a statistical tendency for the invariant mass of said pairs to group around certain mass values. These mass values indicate the mass of a particle that has decayed to this muon pair. A dimuon histogram that illustrates an invariant mass distribution can be seen in Figure 16.

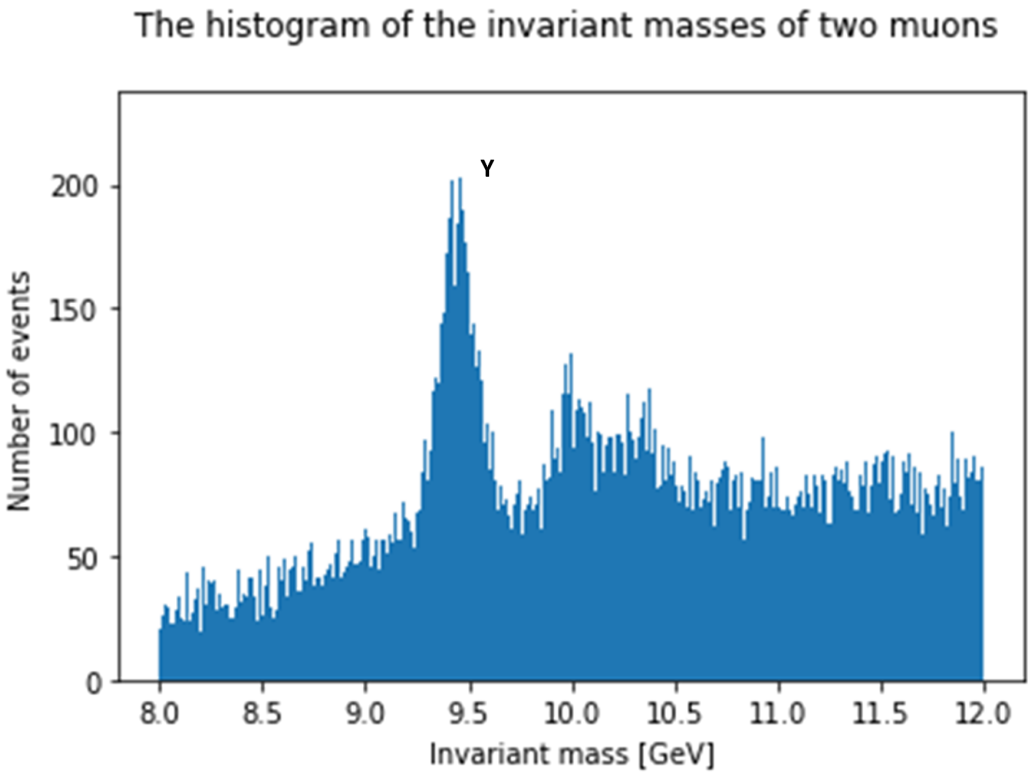


Figure 16. Dimuon invariant mass histogram that displays the Upsilon Meson. Figure from [40].

Most of the detected muons do not originate from a Higgs decay but rather from an array of different particle interactions. It is common for detectors to misidentify muons and incorrectly label them as another particle (for example, muons are hard to distinguish from pions). All of these factors contribute to what is known as a background, however these various misidentifications of $H \rightarrow \mu\mu$ decays become statistically insignificant as an increasing number of invariant masses are calculated. The signals that are remnants of a particle decay filter through the random contributions of this background to form resonances around certain mass values; these indicate the presence of a particle at the origin of said decay sequence. These resonances can be held to some significance test by statistically testing the invariant mass signals and calculating the deviation of this signal from the expected distribution. The extent of these excess events that lie above the background correspond to a probability value that gives a measure of the likelihood of such signals being obtained due to random fluctuations. The smaller this p-value, the more likely the signal is indicative of a new particle, not some fluctuation. These statistical

significance can be measured across multiple decay channels and then combined as was done by CMS, displayed in Table 1 and illustrated in Figure 17.

Decay mode/combination	Expected (σ)	Observed (σ)
$\gamma\gamma$	2.8	4.1
ZZ	3.8	3.2
$\tau\tau$ + bb	2.4	0.5
$\gamma\gamma$ + ZZ	4.7	5.0
$\gamma\gamma$ + ZZ + WW	5.2	5.1
$\gamma\gamma$ + ZZ + WW + $\tau\tau$ + bb	5.8	5.0

Table 1. Summary of the subchannels, or categories, used in the analysis of each decay mode. Table from [41].

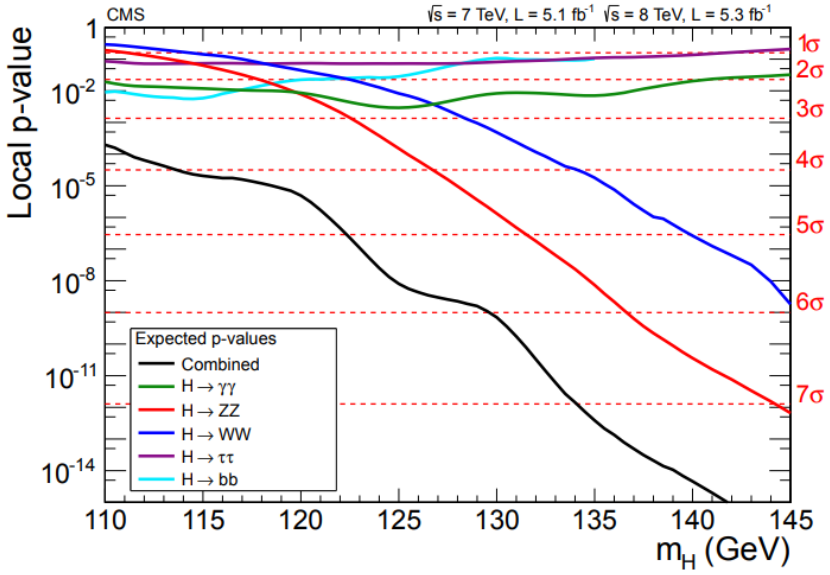


Figure 17. p-values that would be expected for a Standard Model Higgs boson as a function of mH, for the decay modes $\gamma\gamma$, ZZ, WW, $\tau\tau$, bb and their combination. Figure from [41].

3.3.3 Monte Carlo simulations

Monte Carlo methods, also known as Monte Carlo simulations, are a set of mathematical and data-analysis techniques for the prediction of the outcomes of an uncertain event [42]. They were developed during WWII by John Von Neumann and Stanislaw Ulam to aid decision-making in highly uncertain situations [43]. These methods have a variety of real-world applications, including stock price forecasting and project management. Monte

Carlo simulations are particularly popular in high-energy physics because they allow for the accurate estimation of important quantities in problems that do not have easy analytical solutions.

These methods use a probability distribution for each variable with an associated uncertainty to generate a model that depicts probable outcomes. Then, calculations are repeated multiple times with various sets of random integers between the provided minimum and maximum values. The calculations in a typical Monte Carlo experiment are performed thousands of times to generate a high number of possible outcomes [42]. Because of their precision, these methods are useful for long-term forecasting. With a large number of inputs, the number of forecasts correspondingly grows, allowing for a more accurate prediction of outcomes. At the end, we obtain a number that describes the likelihood of each outcome.

The completion of Monte Carlo simulations involves three steps. The predictive model is initially built up, with the dependent variable we want to forecast and the independent variables that will drive the model identified. The independent variable probability distributions are then specified. This is commonly done by defining the range of plausible values and assigning probability weights using past data [42]. Finally, the calculations are performed a number of times to obtain a representative sample of the number of alternative outcomes and their probability.

The CMS experiment includes a high precision testing of the Standard Model and thoroughly searches for new physics. This requires the modelling of expected theoretical behaviour via data simulation and a detailed comparison with experiment data. At CERN, these simulated events are generated through Monte Carlo event generators. Monte Carlo methods are widely used in particle physics as they provide accurate estimates to problems without simple analytical solutions, and determine key quantities with a high degree of precision.

3.3.4 Event Reconstruction

It is unrealistic to attempt to reconstruct every origin of every particle; it is also undesirable as most of the products of high energy proton collisions are low energy particles that are of little interest. In order to limit the number of selected particles to investigate, requirements

must be set. These can often take the form of minimum energies or requirements based on pseudo-rapidity, transverse momentum, and velocity. The pseudo-rapidity, η, is the spatial coordinate used to determine the angle of a particle relative to the beam axis. It is defined as [38]:

$$\eta = -ln[tan(\theta/2)]$$

Where θ is the polar angle. When θ is 90°, the pseudo-rapidity is zero and tends to infinity as θ approaches zero [38]. Polar coordinates (r, ϕ) are used in the transverse plane, however, where r is the distance to the beam pipe and ϕ is the azimuthal angle around the pipe.

An an example, physics analysis for the case of a standard model Higgs decay to two W bosons in the two-lepton final state provided by ATLAS [44] proves as useful in showing event selection criteria that give insight into what these requirements might look like [44]:

- Single-electron or single-muon trigger satisfied;
- Exactly two isolated, different-flavour opposite-sign leptons (electrons or muons) with $p_T > 22$ and 15GeV, respectively;
- Missing transverse momentum EmissT larger than 30GeV;
- Exactly zero or at most one jet with $p_T > 30$GeV, and exactly zero b-tagged jets (MV2c10 @ 85% WP) with $p_T > 20$GeV;
- Azimuthal angle between E_T^{miss} and the dilepton system $\Delta\phi(\ell\ell, E_T) > \pi/2$
- Transverse momentum of the dilepton system $p_T > 30$GeV;
- The invariant mass of the two leptons m$\ell\ell$ must satisfy: 10GeV $< m\ell\ell <$ 55GeV;
- Azimuthal angle between the two leptons $\Delta\phi(\ell, \ell) < 1.8$.

After these requirements have been set, the data can be compared to the Monte Carlo prediction for the distribution of, for example, the dilepton transverse mass [44]. This would appear as follows in Figure 18.

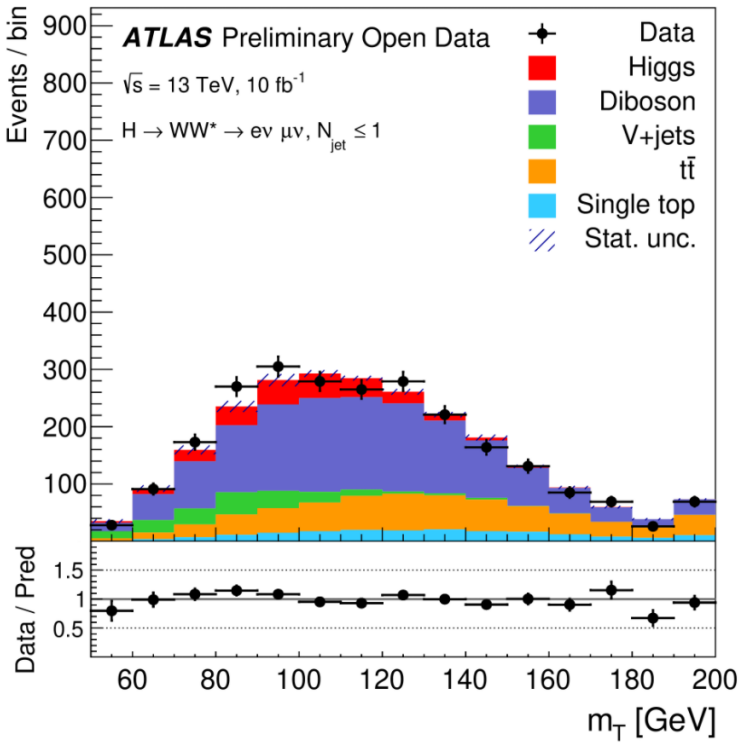


Figure 18. Plot of transverse mass against events for the $H \to WW$ channel with two-lepton final state. Figure from [44].

At the end, one is able to compare data and MC prediction for the distribution of e.g. the dilepton transverse mass, as seen below. A small excess in data is observed which corresponds to the production of the Higgs boson.

In relation to CMS, their 2012 Higgs discovery paper stated that an event description algorithm was used, called 'particle-flow' [45] [46], to reconstruct and identify each selected particle with an optimised combination of all subdetector information. [41]. The determination of the particle momentum is governed by the identification of the particle type (e.g. muon, electron, neutral hadron, charged hadron, photon) in this process. Per bunch (proton group) crossing at the LHC, there is an estimated average number of 9 and 19 proton-proton (pp) interactions in the 7TeV and 8TeV data sets respectively. An event-by-event evaluator, 'FASTJET technique' [47], is used to subtract the energy from overlapping pp interactions from the underlying events. This is based on the calculation of the pseudorapidity dependent transverse momentum density per event.

Another instance of CMS requirements is through using the isolation of leptons and

photons extensively as a means of selection criteria [41]. Of the particles that were reconstructed, requirements on the scalar sum of their transverse momentum were set in a confined region that is defined as $\sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ where $\Delta\eta$ and $\Delta\phi$ are the pseudo-rapidity and azimuthal angle differences between the particle and reconstructed particle directions [41].

3.3.5 The Higgs boson decay channels

The Higgs is very unstable and decays rapidly. Its decays constitute its signature. As we have remarked, the Higgs coupling is proportional to the mass of the particle it couples to. The largest coupling of the Higgs is therefore to the top quark, followed by the W/Z bosons, the b quark. Among the leptons, its coupling is largest for the tau lepton (τ). Since the Higgs has a mass of 125GeV, it cannot decay into a pair of top quarks (which would require a mass of 2*174GeV). Figure 19 shows the branching ratio of the Higgs boson depending on its mass. The branching ratio is defined as the probability for a particle to decay into certain particles. Note that there is a small branching ratio into photons and gluons despite the fact that these particles do not have mass. In fact these decays occur through a virtual loop of top quarks that merge to give two photons since the top have both colour and electric charge. In the case of the decay to two photons there is also a virtually charged W loop.

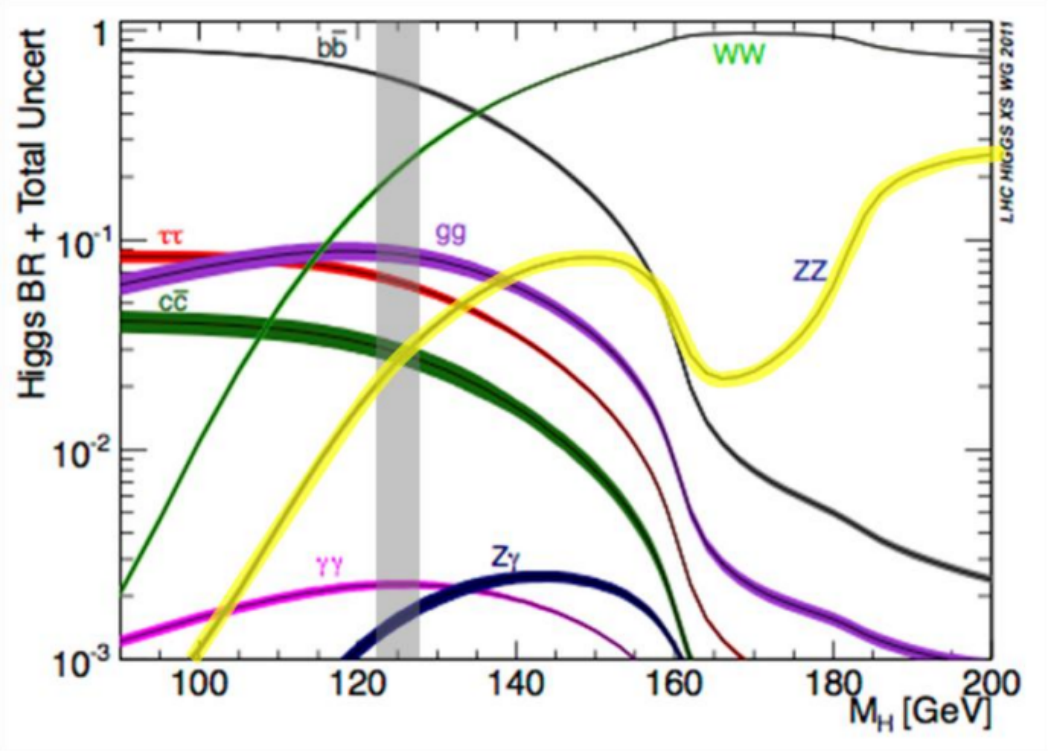


Figure 19. The Standard Model Higgs branching ratios that correspond to different decay channels as a function of the Higgs boson mass. Figure from [38].

The largest branching ratio is into b quarks followed by decays into WW. In this case the W bosons cannot both be on-shell, meaning one of them does not satisfy equations of particle motion; this signature relies on the further decay of the W bosons. In the environment of a hadronic machine such as the LHC, decays into hadrons (including b quarks) are challenging since the strong interaction produces a significant background. The decays of Higgs bosons into τ leptons constitute 6% of the total number of decays. The produced leptons also decay, often into hadrons and neutrinos which detectors can't detect, leading to missing energy in the mass of the Higgs boson. This channel is thus very hard to exploit.

The $H \to ZZ$ and $H \to \gamma\gamma$ decays are much more sensitive with a higher mass resolution at $m_H = 125$GeV. Their branching ratios are however significantly smaller than other decay channels. While these would have been the ideal channels to investigate in 2012 (as they contain a large excess for a small background), a large number of events would have been required to obtain enough data given how low the branching fractions for

those channels are. The technology at the time was however unable to store such a high amount of information [26].

3.3.6 CMS Higgs Decay Channels Investigated

Before the Higgs had been officially discovered, CMS took up investigation over the following five Higgs decays: $H \to b\bar{b}$, $H \to \gamma\gamma$, $H \to ZZ$, $H \to WW$, $H \to \tau\tau$. In the 2012 CMS Higgs discovery paper [41] , it was stated that above an expected background, an excess was observed with a local significance of 5.0 standard deviations (σ) around a mass close to 125GeV. This signalled the production of an undiscovered particle, the Higgs boson. The data used corresponded to integrated luminosities of up to 5.1 fb^{-1} at 7TeV and 5.3 fb^{-1} at 8TeV [48]. The standard deviation for a standard model Higgs boson of 125GeV was 5.8σ. The excess was found to be most significant in the two decay modes: $\gamma\gamma$ and ZZ; they possessed the highest mass resolution of the chosen channels. When a fit was performed to these signals, a mass of 125.3 $\pm$ 0.4 (statistical) $\pm$ 0.5 (systematic)GeV was obtained. Based on the diphoton decay, it was indicated that the new particle was a spin-less boson.

The search sensitivity of a decay channel, for a given value of m_H, is dependent on: the cross section, the branching ratio of the decay into the specified final state, the efficiency of signal selection, the background level [48], and the mass resolution. Higgs decays to ZZ or to $\gamma\gamma$ are particularly clean channels, but the branching percentage for $H \to ZZ$ is 2.67 %, and Hß$\gamma\gamma$ is only 0.228 %; a large dataset is thus needed to be able to observe these decays. Such a dataset was not available at the time of the Tevatron. It is important to understand why these low cross-section channels provided the best evidence for the Higgs' initial discovery. While the branching ratios for these channels are small, there is a much smaller level of background events that Higgs specific decays must be discerned from.

a $H \to b\bar{b}$

From Figure 19, it can be seen that at a mass of 125GeV, the Higgs mostly decays to a $b\bar{b}$ quark-antiquark pair. These quarks hadronise. Because of the strong interaction, quarks are unable to exist in isolation; free quarks spontaneously generate more quarks until they are all bound to a pair or triplet (hadrons). Out of the four fundamental forces, the attractive strong interaction (which governs quark behaviour and is described by quantum

chromodynamics (QCD)) is the only one that yields an increase in force for an increase in distance, eventually levelling off to a constant value. The energy required for a quark to stay isolated is actually higher than the energy required to create new quarks [49]. Once a certain quark separation distance is reached, the potential energy becomes high enough to create a quark-antiquark pair which is capable of binding with another separated quark and hence hadrons are formed. In the case of a highly energetic quark or gluon present in an LHC collision for example, this process can occur several times, resulting in a jet. Jets are not a particularly clean signature; they make it difficult to extract information about the primary interaction because it is difficult to discern which particles originated from each quark or gluon and which ones come from other debris. This is a source of uncertainty in the energy of the original b quarks. In addition, there is an overwhelming number of bottom quarks produced by QCD in hadron colliders. This is why, despite being the most frequent Higgs decay, the $b\bar{b}$ decay has been troublesome in producing a significant excess.

An observation of 5.4σ significance of the Higgs in this channel was obtained in 2018 by ATLAS, 6 years after the first official Higgs observation. [50] It was iterated by the experiment how large backgrounds from multijet production make it very difficult to search for the Higgs in its dominant gluon–gluon fusion production mode. Due to a smaller background, the production modes that were most sensitive to detecting $H \to b\bar{b}$ decays are associate productions of a Higgs boson in accompaniment with a W or Z boson. When the vector boson decays to leptons, efficient triggering is enabled which significantly reduces multi-jet background.

CMS noted the dominating backgrounds in this channel to be: vector bosons in association with jets, top quarks produced in pair or single, and production of dibosons (WW, WZ, ZZ) where one of the bosons decays hadronically. Requirements of a large transverse momentum for the dijet in addition to a minimal jet activity were set in order to significantly reject background events [41].

b $H \to \gamma\gamma$

In the CMS group's analysis of the $H \to \gamma\gamma$ channel, a diphoton invariant mass distribution range of 110–150GeV was evaluated on an irreducible background from two photons produced from QCD. This channel is particularly 'clean'; its background is typically re-

duced by enhancing the signal to background ratio through various means. An aspect of the background that is reducible is the reconstructed photon candidates that originate from misidentified jet fragments. In the LHC, photons can be produced [51]: directly in pairs from a hard partonic interaction (i.e. partons refer to quark or gluon interactions), singly from non-perturbative fragmentation of a hard parton, or from the decay of mesons.

In the CMS experiment [41], the analysis sensitivity was enhanced by separating candidate diphoton events into mutually exclusive categories with differing expected signal-to-background ratios. These are based on criteria satisfaction of dijets [52] and the properties that the reconstructed photons possess; this is aimed at selecting VBF Higgs boson production events. Satisfaction of transverse momentum (p_T) requirements for the two photons was necessary for event selection; both photons had to be reconstructed inside a fiducial region of $|\eta| < 2.5$. Specific criteria was was applied to dijets that possessed the largest transverse momentum in events within $|\eta| < 4.7$. The dijets were held to p_T thresholds of 30 and 20GeV with a required minimum pseudorapidity of 3.5. Their invariant mass was required to be larger than 350 and 250GeV for the two data sets, 7 and 8TeV, respectively [41]. Figures 20 and 21 display two means of observing the Higgs excess obtained from this channel, through analysing p-values and calculating the invariant mass of the diphotons:

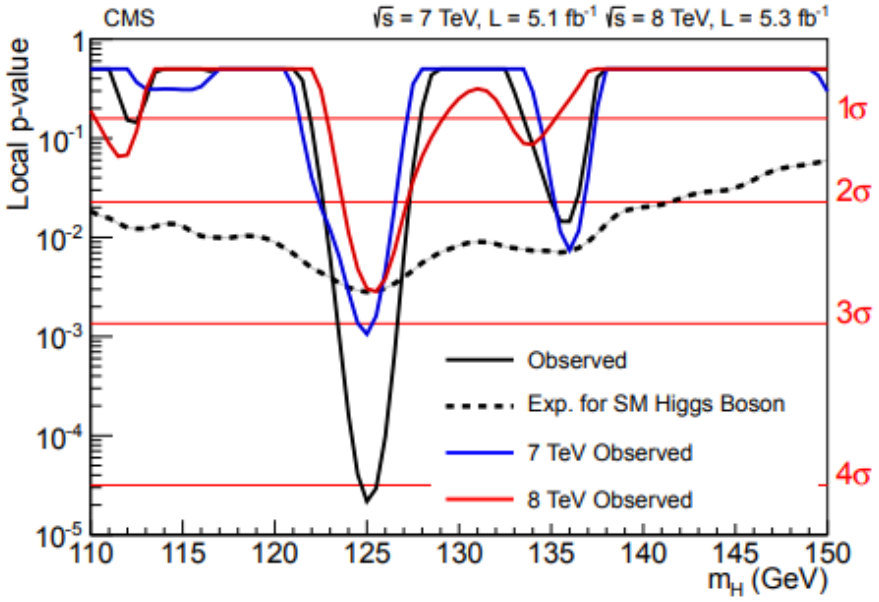


Figure 20. p-value as a function of mH in the diphoton channel using data sets of 7 and 8TeV and their combination. For a SM Higgs boson with mass 125GeV, the expected local p-value from this combined data set is represented by the dashed line. Figure from [41].

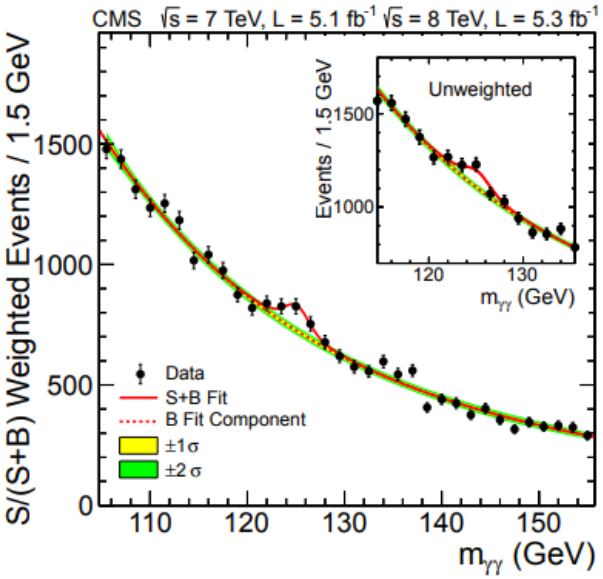


Figure 21. Distribution of diphoton invariant mass with $S/(S+B)$ weights per category where S and B respectively represent the number of signal and background events. This is calculated from the signal-plus-background fit that is applied to all diphoton categories simultaneously. Figure from [41].

c $H \to ZZ$

Known as the 4 lepton decay, this decay channel has a final state signature of four isolated leptons taking the form of two pairs of the same flavour with opposite charges that originate from a shared primary vertex [38]. When selecting these lepton pairs, the criteria is that their total mass must be closest to the mass of the Z boson. However, since the Higgs boson and the Z boson have approximately masses of 125GeV and 91GeV respectively; it is required that the second Z boson is off-shell (since $2m_Z > m_H$). There are four distinct final states for this decay: $\mu^+\mu^-\mu^-\mu^+$ (4μ), $\mu^+\mu^-e^+e^-$ ($2\mu 2e$), $e^+e^-\mu^+\mu^-$ ($2e2\mu$), and $e^+e^-e^+e^-$ ($4e$). The 2μ2e and the 2e2μ final states are different in that the flavour of the lepton pair with mass closest to that of the Z boson is flipped [38].

In 2012, CMS only reported the search range of 110-160GeV for this channel. Each decay final state was analysed separately due to differences in the mass resolutions and reducible backgrounds.

Irreducible background for this channel consists of events where two Z bosons are produced by the annihilation of a quark-antiquark pair or through a gluon-gluon process. The leptonic decay of these Z bosons will result in a four leptons final state; though the intermediary Higgs boson will be missing. The contributions of this process to the background can reliably be estimated using Monte Carlo simulations [38].

Reducible background in this case is primarily dominated by Z^+jets events. A Z boson is produced in association with jets that may be misidentified as leptons and could interfere with the four lepton final state of the Higgs signal. If these jets are misidentified as leptons, this may lead to the same final four leptons state as the signal. Other sources of reducible background include $\tau^-\tau^+$ and WZ^+jets events. Jets do not behave as leptons do and thus have different properties which can be exploited to reduce their contributions [38]. Figure 22 shows Feynman diagram examples of reducible and irreducible background events respectively.

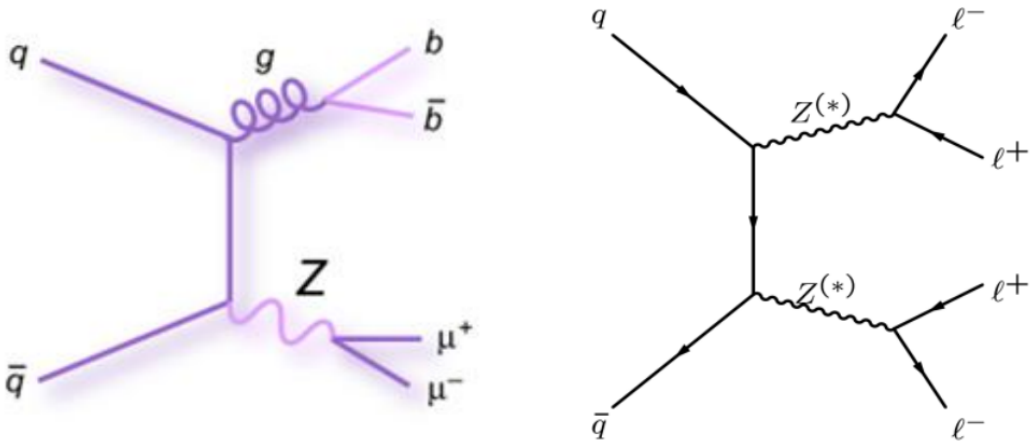


Figure 22. Left: Example of a reducible background event where the jets formed by the b quarks would be misidentified as electrons. Right: Two Z bosons formed by two quarks decay into four leptons producing an irreducible background event. Figure from [38].

Corresponding electron and muon requirements were made by CMS such as pT $>$ 7, 5GeV and $|\eta| < 2.5$, 2.4 respectively; potential electron or muon pairs produced by means of a decaying Z boson were required to have origins of a shared primary vertex [41]. Invariant mass and p-value data plots for this channel that was obtained by CMS and presented in 2012 is displayed in Figures 23 and 24:

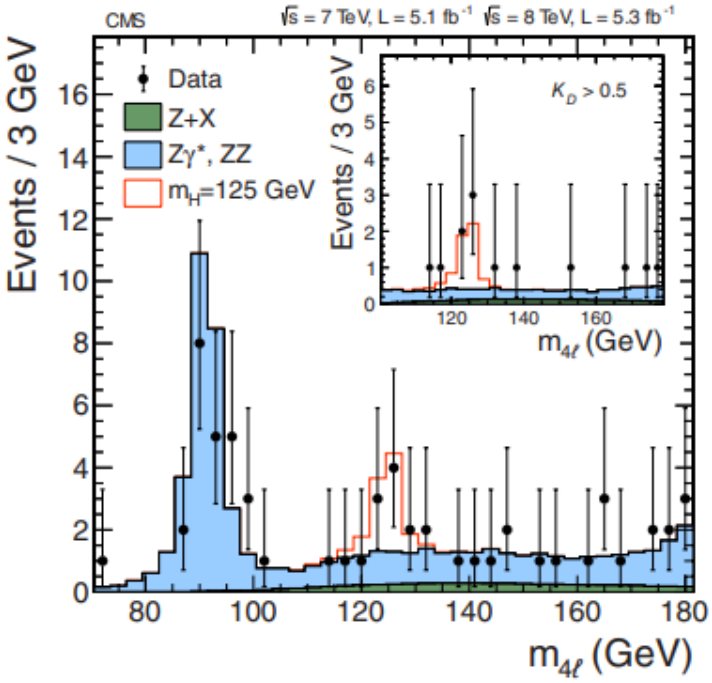


Figure 23. Four-lepton invariant mass distribution corresponding to the ZZ $\rightarrow 4\ell$ channel analysis. Filled histograms display the background; open histograms display the expected Higgs signal for a Higgs boson of mass 125GeV in addition to the expected background. The points represent the data, the filled histograms represent the background, and the open histogram shows the signal expectation for a Higgs boson of mass 125GeV, added to the background expectation. Figure from [41].

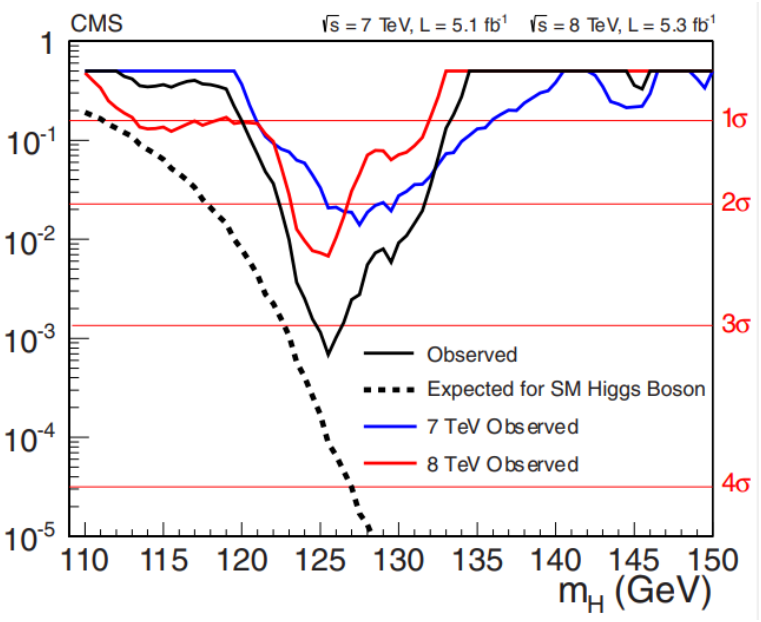


Figure 24. p-value as a function of mH in the ZZ channel using data sets of 7 and 8TeV and their combination. For a SM Higgs boson with mass 125GeV, the expected local p-value from this combined data set is represented by the dashed line. Figure from [41].

d $H \to WW$

CMS noted that around a mass range close to a W diboson threshold of 160GeV, the $H \to WW$ decay has high sensitivity to a SM Higgs boson. Due to developments in lepton identification tools and methods of optimising missing transverse energies (which come about as a result of trouble detecting neutrinos) for LHC pileup, this sensitivity was extended to 120 GeV.

In 2014, ATLAS obtained a 6.1σ excess above a background for this decay channel [53]. The two W bosons in this channel produce a sensitive experimental signature when they decay sequentially to '$\ell\nu\ell\nu$' where 'ℓ' is an electron or muon and 'v' is a neutrino. This typically takes a final state form of two oppositely charged and isolated leptons since the neutrinos are so hard to detect. The dominant background for this channel originates from non-resonant WW diboson production. Other contributors are: tt pairs, single top quarks, W^+jets where the jet is misidentified as a lepton, and non-resonant WZ and ZZ processes [44]. Figure 25 displays the sequential means of which two W bosons are produced via an intermediary Higgs boson while Figure 26 displays an example of a diboson production background event.

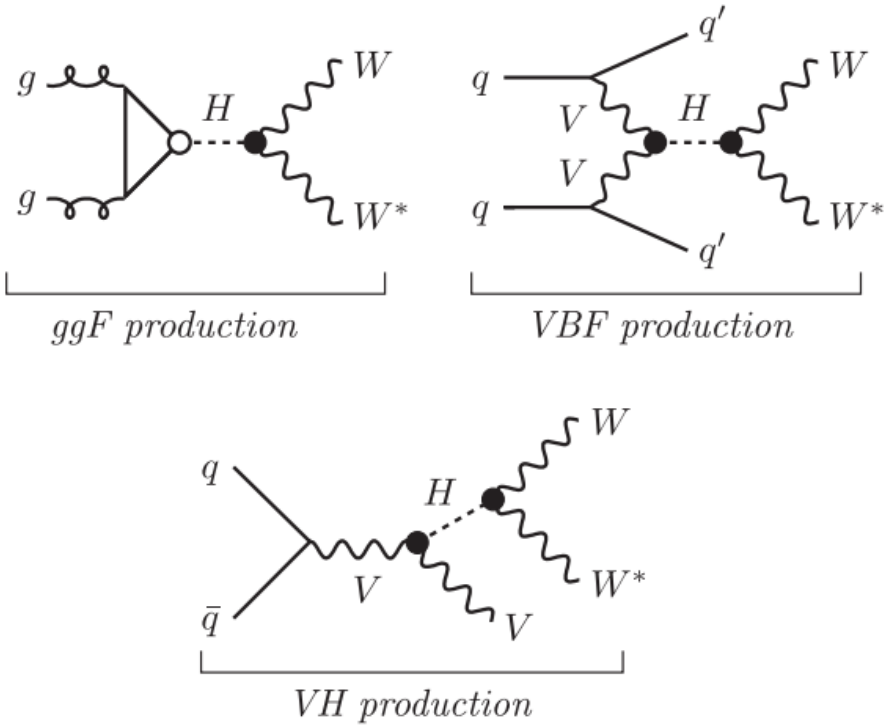


Figure 25. W bosons are produced via an intermediary Higgs boson. Figure from [54].

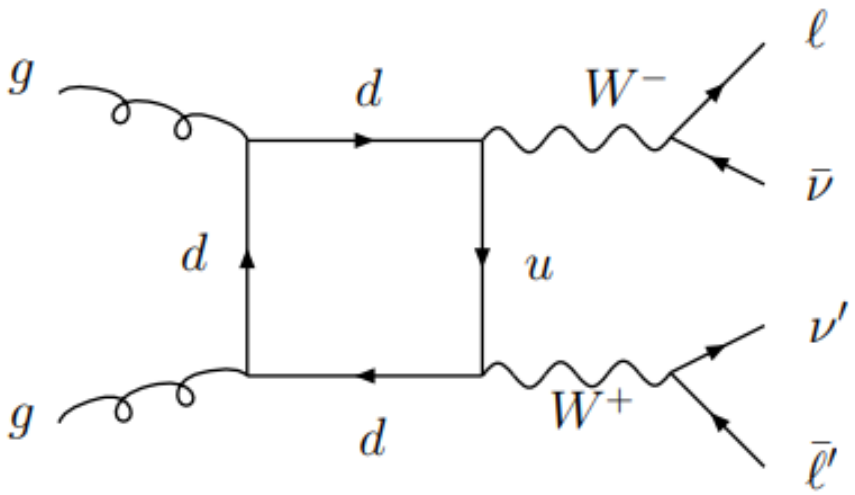


Figure 26. gluon-induced WW diboson production background. Figure from [55].

e $H \to \tau\tau$

Note that this signature depends on the type of τ decay one looks for. An example of τ

decays is show in Figure 27 below.

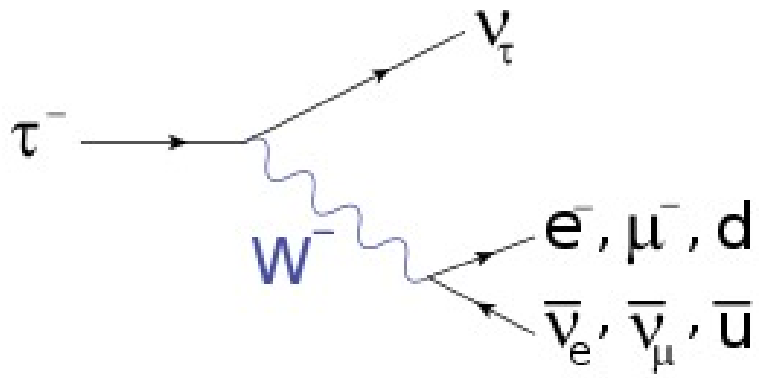


Figure 27. Final states in τ decays. The final state in quarks u, d is the hadronic final state [55].

CMS searched for the $H \to \tau\tau$ decay mode in four exclusive subchannels which corresponded to four decay routes of the tau pair [41]. These decay products were eμ, μμ, $e\tau_h$, and $\mu\tau_h$; muons and electrons are the result of leptonic tau decays and τ_h is used to denote tau decays that are hadronic. A mass range of 110-145GeV was chosen as the search region in which the signal was expected to take the shape of a broad excess on an invariant mass distribution of the tau pair. A 2017 CMS paper presented a 5.5σ excess for this decay [56]. This channel has a smaller background contribution than that of the $b\bar{b}$ decay channel, with a main background source being the $Z \to \tau\tau$ decay. Neutrinos are very difficult to detect, this is problematic since a large amount of energy can be transferred to the neutrinos produced in the tau decays. This limits the separation power of the invariant mass estimator that would help discern the $H \to \tau\tau$ from the large and irreducible $Z \to \tau\tau$, $\ell\ell$ [56]. Figure 28 illustrates a comparison of $H \to \tau\tau$ signals and dominant background channels.

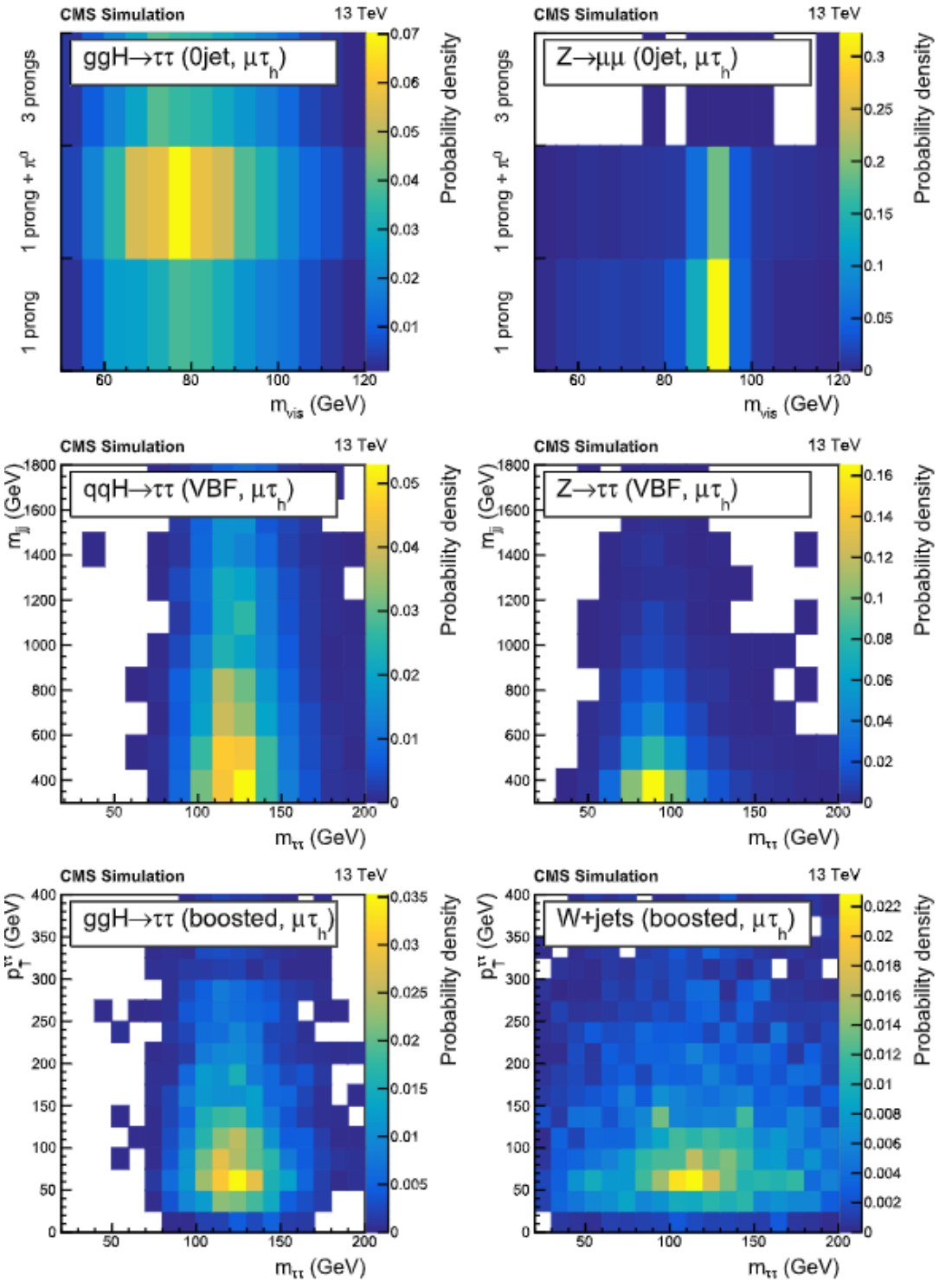


Figure 28. CMS simulation: Left-side graphs display distributions for the Higgs signal; the right-side displays some dominant background processes. "prong" refers to the number of charged particles in the final state of the tau decay. Figure from [56].

4 The CMS Open Data

4.1 Reasons behind the CMS Open Data

4.1.1 Introduction

The CERN Open Data portal is the access point to a growing range of data produced through the research performed at CERN [57]. It includes over two petabytes of data alongside accompanying software and documentation to understand and analyse it. This wealth of information is made public for educational and research purposes.

The options are aimed at different levels meant to build an understanding of particle physics gradually. Data produced by the LHC experiments are usually categorised in terms of difficulty in four different levels:

1. **Published Results (Level 1) Policy:** Peer-reviewed publications represent the primary scientific output from the experiments [58].

2. **Outreach and Education (Level 2) Policy:** Dedicated subsets of data are used for education and outreach. Data is carefully selected and formatted to make it easily accessible [58].

3. **Reconstructed Data (Level 3) Policy:** Reconstructed collision data and simulated data allow to perform a complete scientific analysis.

4. **Basic raw data (Level 4) Policy:** This allows one to reconstruct the events and simulation software, allowing the production of new simulated signals.

The Open Data portal focuses mainly on releasing event data from levels 2 and 3. In this analysis, precise and detailed instructions on the Open Data website were followed as a way to reproduce data for the Higgs boson.

Firstly, we start with Level 2 tools aimed at beginners and intermediates. At a Beginner level, there are options available for visualising and understanding collisions. The intermediate level begins the analytical understanding using short coding tutorials accessible online or offline programmes to calculate invariant masses, understand and extract the information available in physics histograms.

In addition, we will also dive into Level 3 tools by reconstructing actual events and simulating data using Monte Carlo event generators and raw ROOT files. This thorough analysis was conducted not only to produce results, but also to test how accessible the CMS Open Data portal is and assess whether a user with a basic knowledge in physics could reconstruct and analyse the available data.

While access to such a wide range of data allows detailed inspection, recognising the relevant data sources can come as a challenge, especially for someone with little to no experience in particle physics and programming. We focus on providing a proper insight into the accessibility of these resources and we will make notes of suggestions or improvements that would ease the analysis of such data. Further discussions will be noted in section 7.

4.1.2 The Data

The Open Data CSV Database offers valuable resources for outreach and education. Besides the documentation, we note the availability of multiple data files with collision events available in an online GitHub repository [59]. Data from the CERN OpenData portal, CMS doc database and cms-opendata-education GitHub organisation [60] are all available in a single file, 'csvDatabase.csv' [61].

'csvDatabase.csv' [59] contains multiple .csv files of various decays (four leptons, two electrons - two muons, ...) along with their online path to CERN for use in Binder/Google Collab. These can also be downloaded and used for Python, R or ROOT analysis.

Moreover, a specific Jupyter Notebook, 'csvFileDocumentation.ipynb' [60], browses the database to provide information on the files. These can be filtered through several criteria such as filename, number of events, invariant mass, momentum, energy, number of decay products and many more.

5 Level 2 Analysis

5.1 Beginner Level

5.1.1 The CMS event display (iSpy)

a Description of its functionalities

The CMS Event Display (Figure 29) [62] is an online tool designed for the visualisation of real particle collisions from the CMS experiment at the LHC. CERN provides a brief guide on displaying collision events, including those that lead to the discovery of the Higgs boson [63]. The Higgs candidates' events available are three runs for 4lepton decays and ten for diphoton decays.

Its usage is reasonably intuitive, and information can be found on the 'Need Help' button located on the top right corner. 'Settings' provides further options for the display. For instance, the 'Show information dialogs' enables a short description of the options on the left-hand side (see Figure 30). Further to the right, the 'About' button contains links to the CMS experiment on the CERN website and any questions, comments or problems can be brought up at ispy-developers@cern.ch. The collision events can be saved as .JPG images (however, we must note that this functionality is only available using Mozilla Firefox).

The CMS Event Display has a Stereo View built-in, which, with a simple VR Viewer (Google Cardboard), allows a virtual 3D tour of an actual collision. This option has not been tested.

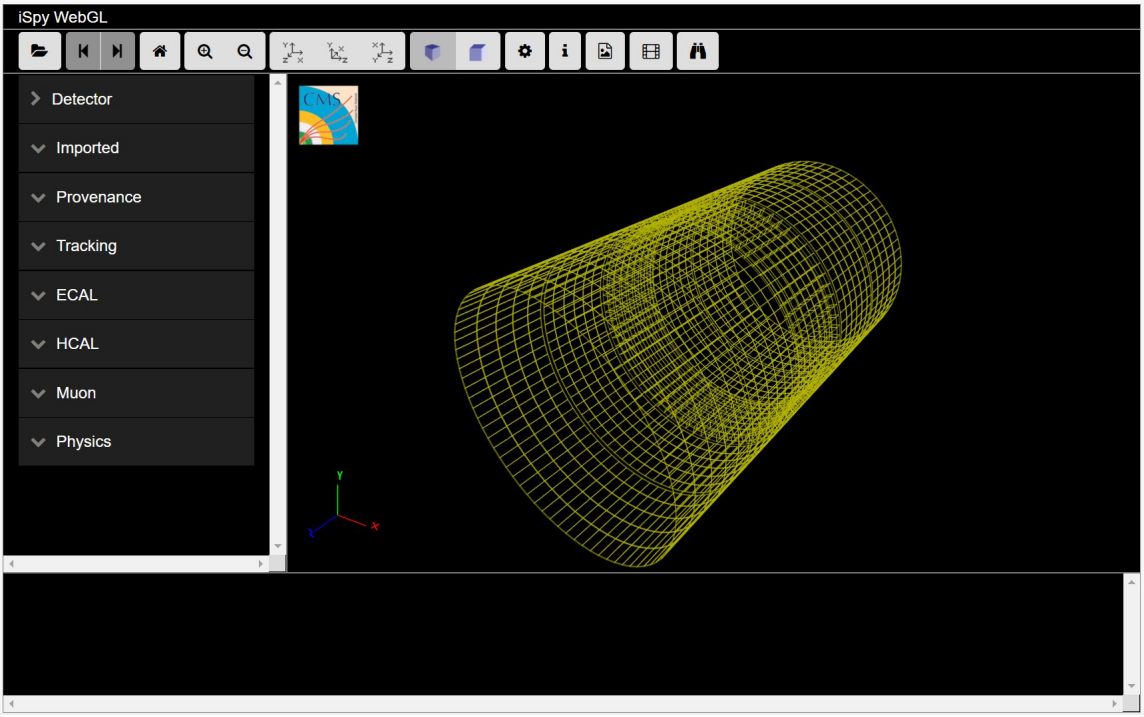


Figure 29. CMS event display overview. [62]

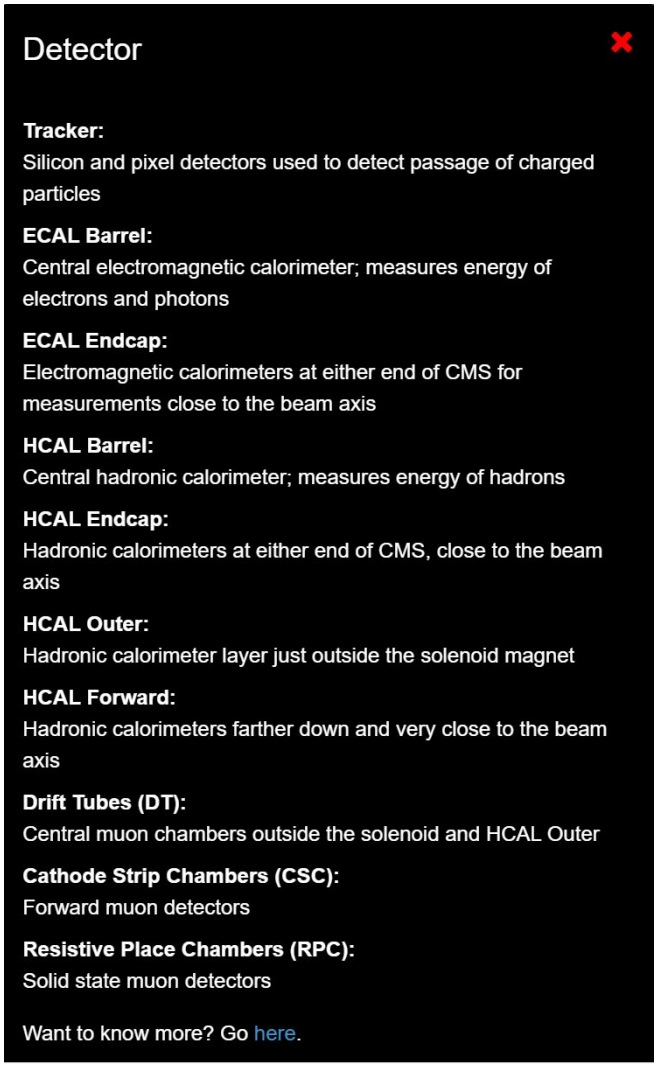


Figure 30. Contents table providing brief notes on the detector parts available. [62]

b Visualisation options in parallel to the real detector

The CMS detector acts as a 3D camera that captures unstable particles' positions as they decay (with a rate of up to 40 million times per second [64]). A depiction of the visualiser can be observed in figure 29. Its design involves several concentric layers of material that enable the reconstruction of particle tracks as well as the calculation of their energy and momentum. The CMS detector was built following a set of desired features [65], such as a high-quality central tracking system for exact momentum measurement, a high-resolution electromagnetic calorimeter (ECAL), hermetic hadron calorimeter (HCAL) and a high-performance muon detector.

The Silicon Tracker, which is the inner layer of the detector, produces hits when charged particles pass through as they interact via the electromagnetic force. The particle tracks are obtained by joining together these particle hits. The resulting charged particle trajectories will bend by the CMS's powerful magnet. Particles bend into different directions depending on their electric charge.

This is done to observe the particle's charge as depending on whether positive or negative, the particle will bend in different directions in the same magnetic field. Moreover, it helps calculate the momentum which is related to the amount by which particles bend.

Calorimeters constitute the outer layers of the detector. The electromagnetic calorimeter will enable the measurement of the energy of electrons and photons by stopping them. Hadrons will stop once they reach the Hadron Calorimeter, while muons require special sub-detectors. Tracking devices and muon chambers allow for the muon's momentum measurement.

The options available on the left-hand side of figure 29 are, in order, Detector, Imported, Provenance, Tracking, ECAL, HCAL, Muon, Physics related to the actual components of the CMS described above and in section **3.2.4**. The CMS event display reproduces in 3D all of these components accurately, allowing the users to explore each layer of the detector and understand what happens to different particles. For example, preshower recoil hits can be observed by ticking the corresponding option to observe the particle hits in the fine detector located in front of the ECAL.

Figure 31 displays the signatures of an electron, photon, charged hadron and muon. Different signatures allow scientists to identify various particles. We can see where a specific type of particle is stopped as well as the path it leaves in the detector.

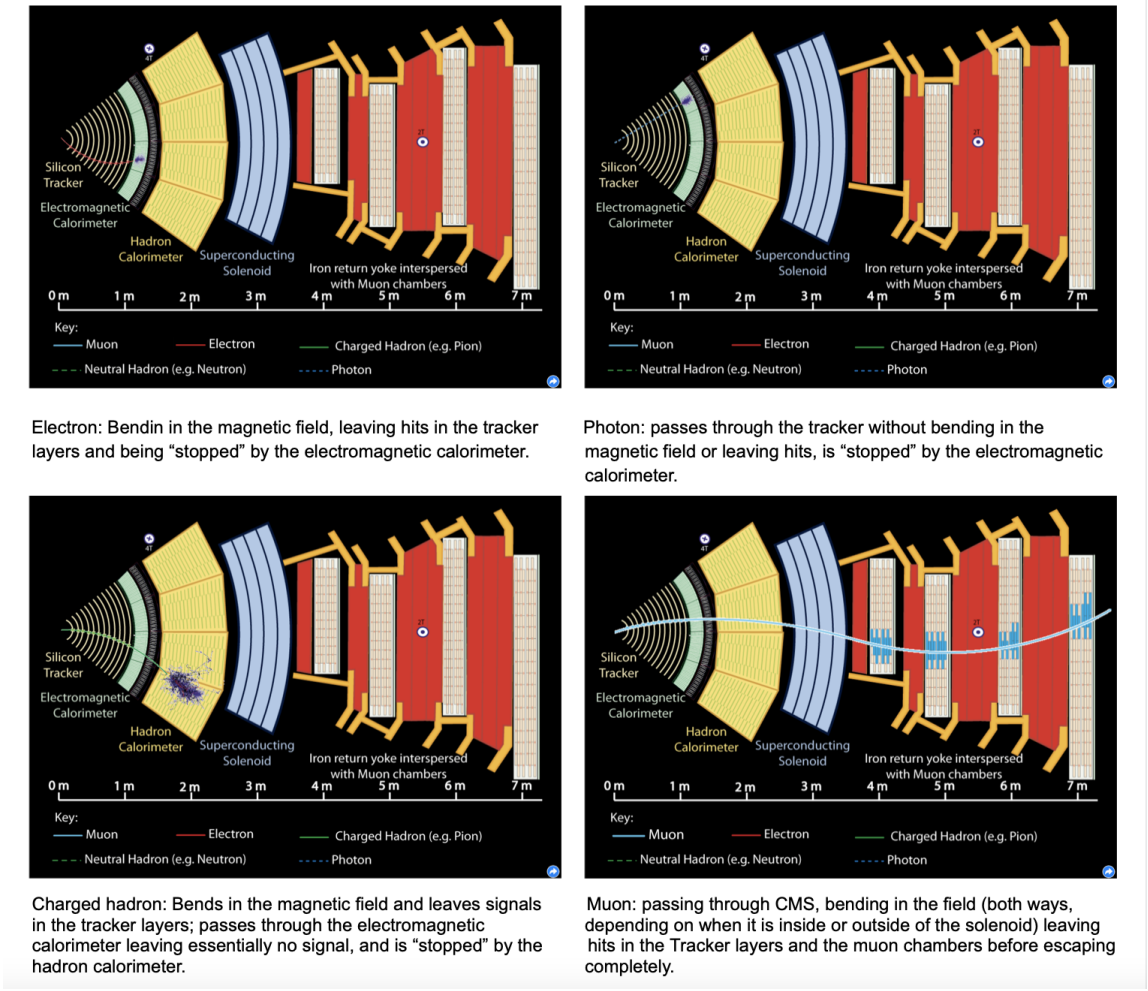


Figure 31. CMS events for four different types of particles and their signatures. [66]

The CMS event display allows for a 3D view of the particle's track to understand their nature and gives insight into various decay modes. We will investigate two different decay modes for the Higgs boson in the following sub-section.

c The Higgs Candidate events

When two protons collide, many particles are created (in figure 32, we can observe all the yellow tracks that lead to the collision point). The CMS event display allows an interactive view of these tracks. All the detector layers can be shown or hidden using the left-hand side buttons to visualise different collision points and understand which particles collide with each other. The tracks of the particles are also made visible. Figure 32 shows the

4lepton decay specific to a Higgs candidate event. The four red distinguishable tracks represent four muons (hence 4lepton decay) that pass through the inner tracker, ECAL, HCAL, through which they leave tracks until they reach the muon chambers.

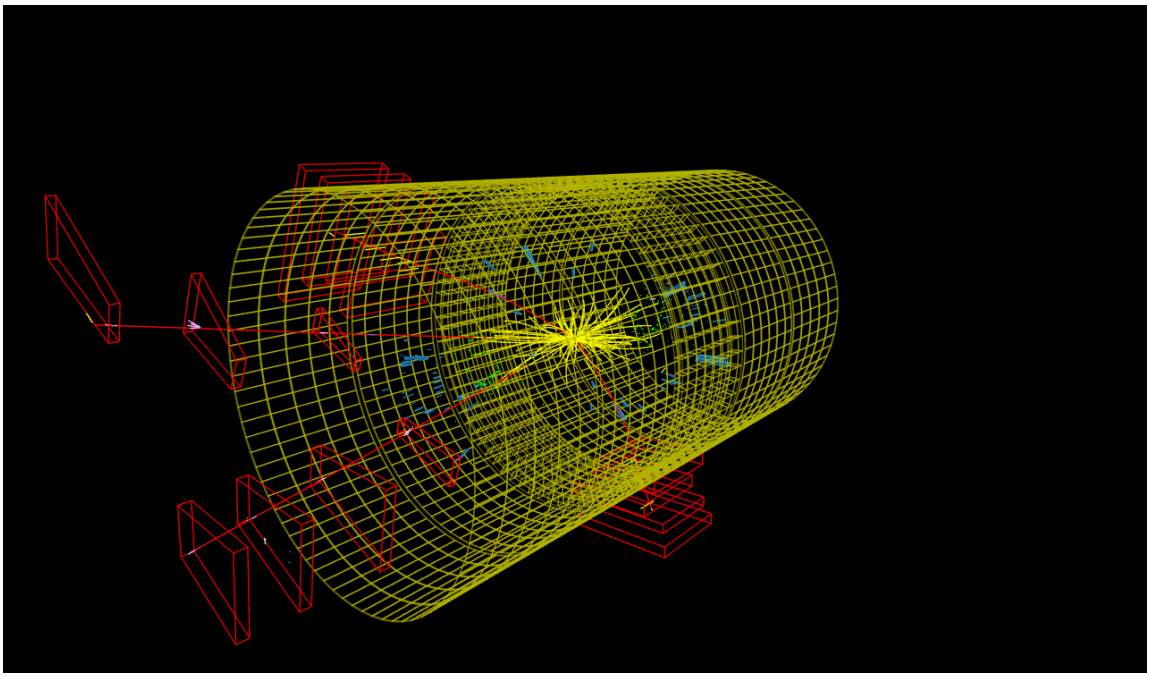


Figure 32. 4lepton decay.[67]

Diphoton events constitute another option for the visualisation of Higgs candidates via the Open Data website. Below, we display four different runs of the same decay (Figure 33). The HCAL detector was disabled to allow a better view of the two-photon tracks (in green).

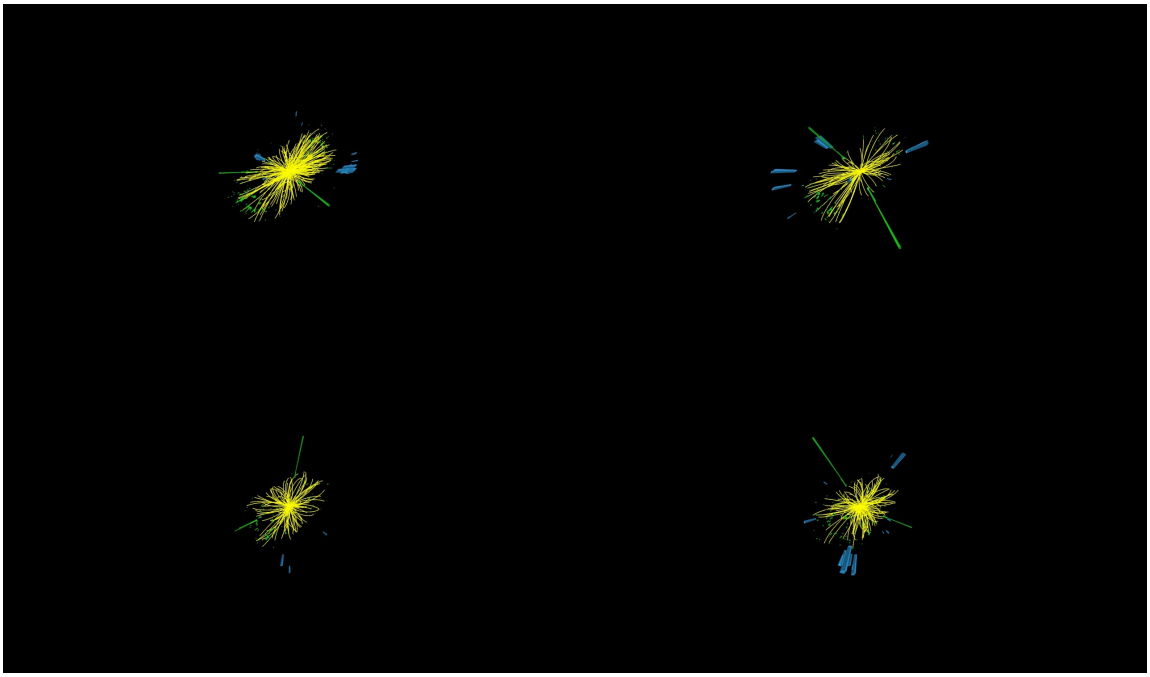


Figure 33. Comparison between different runs of the same event – diphoton decay. [68]

5.2 Intermediate Level

5.2.1 The CMS histogram visualiser

Another available resource on the CMS Open Data website is the histogram visualiser. It allows one to plot histograms for different particle events, including dimuon events, dielectron events, and W boson decays. It is possible to visualise various types of histograms based on characteristics such as the total energy (E), transverse momentum (p_T), pseudorapidity (η) and invariant mass (M). Figure 34 illustrates examples for two of these options [69]. The visualiser also allows the user to change between logarithmic and linear scales very easily.

Figure 34a shows the pseudorapidity, η, of the first muon in dielectron events, ranging from $\eta = -2$ to $\eta = 2$, with most events having an absolute pseudorapidity inferior to 1. As mentioned above, pseudorapidity is the spatial coordinate used to determine the angle of a particle relative to the beam axis. Pseudorapidity affects the resolution of the measurement of momenta. Particles with low absolute pseudorapidity can be measured more accurately than those with a high absolute pseudorapidity.

Figure 34b shows the transverse momentum for the first muon in a dimuon decay in GeV for different events with a peak at roughly 1 GeV and 13 GeV. The transverse momentum is the momentum that is perpendicular to the momentum transfer between the beam and the particle [70].

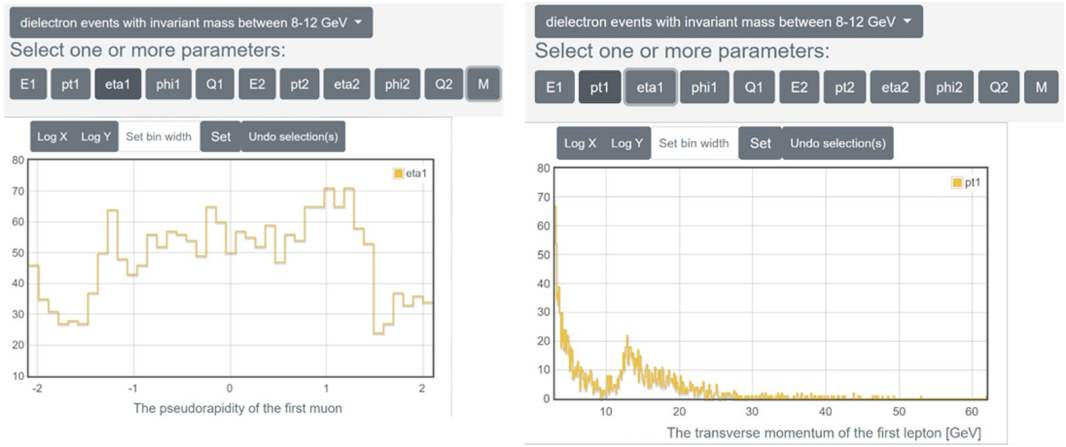


Figure 34. Histograms using visualiser, (a) showing the pseudorapidity of the first muon for dielectron events, (b) showing the transverse momentum of the lepton for W bosons decaying to a muon and neutrino [70] .

5.2.2 Spreadsheet programs histogram analysis

Excel

One of the tools available on the CMS Open Data site is a guide with instructions on how to plot histograms on Excel. This is available as a PDF file which is also directly found on the GitHub website [71]. The guide is very thorough and straightforward. It allows users to achieve relevant results using CSV files which can be downloaded from the Open Data site. Figure 35 shows such a plot which was produced through Excel. The peak corresponds to the approximate value of the mass of a particle. In this case, the peak is at around 91.2 GeV, which is roughly the mass of a Z boson [72].

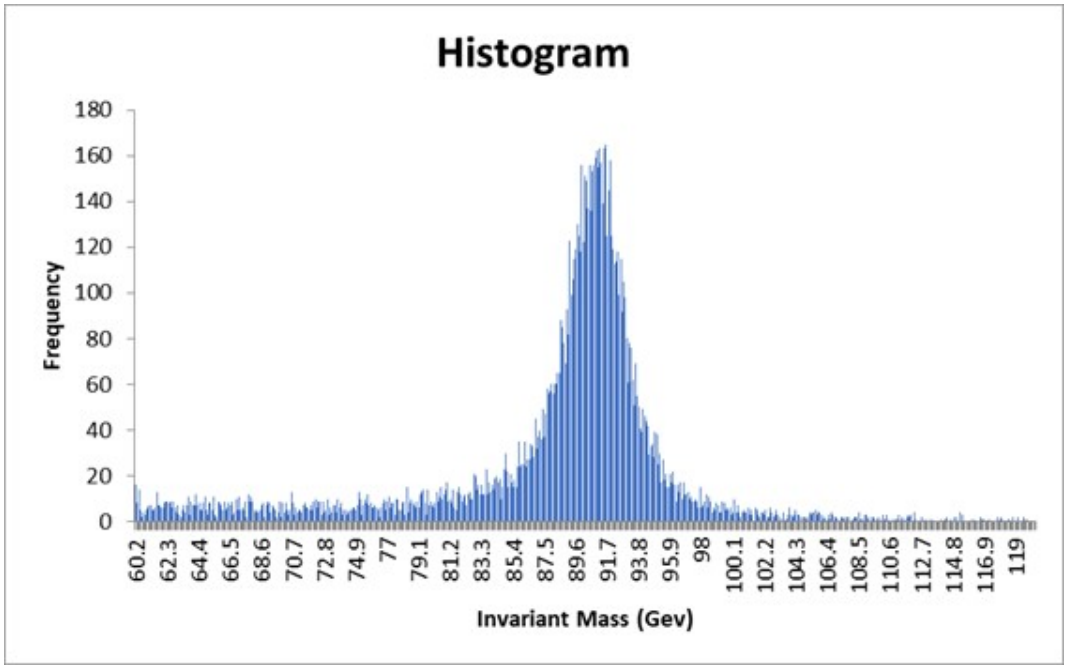


Figure 35. Histogram plotted using Excel, using Zmumu_Run2011A_masses.csv downloaded from the Open Data site. It shows thousands of muon-muon events from proton-proton collisions [71].

5.2.3 First step to programming using physics data

The first step towards data analysis can be easily achieved online without the need to download any specific software. Binder constitutes such an option [73], which we will look into with more detail and mention other possibilities. The following section is all browser-based; no installations are required. This is a convenient way for anyone interested in the topic to begin their particle physics analysis journey.

a Binder

Binder is an online service for the building and the sharing of reproducible and interactive computational environments from online repositories.

CERN provides six notebooks for Binder use.

One of them is a 'Guide to using Python' [74]. It represents a valuable resource for someone without prior computing knowledge. Some of the topics approached and required for other notebooks we will look into are data types, modules, basic calculus, random data generation, plots and common issues. The guide contains all the required information and possibly more for someone to complete the rest of the notebooks on their own whilst

understanding the code. Thus, it represents an excellent starting point for the CMS data analysis. Moreover, all the other notebooks start with very brief python explanations.

The 'Quick start to CMS Open Data' [75] notebook looks into plotting the invariant mass histogram for two muons (Figure 36).

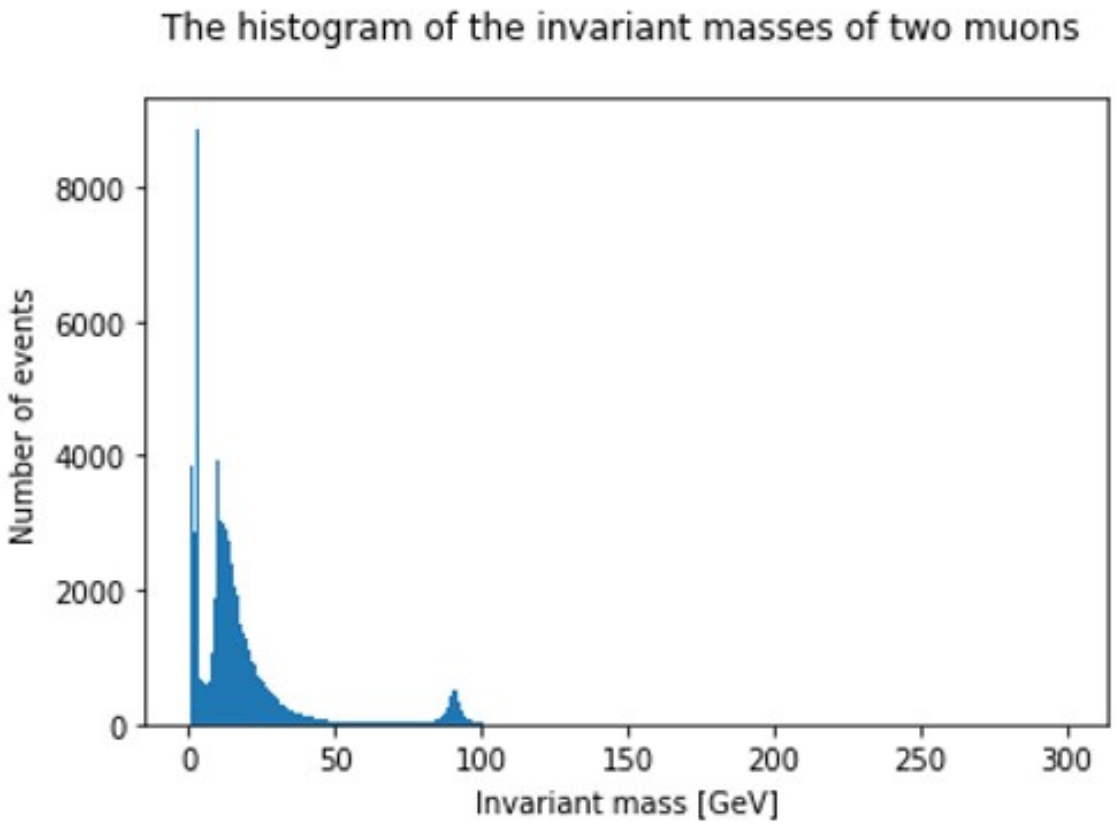


Figure 36. Invariant mass histogram for two muons with peak at 90GeV corresponding to the Z boson [75].

By comparing the peaks visible on the graph with a dimuon decay particle list, we can understand how different peaks correspond to specific particles, depending on their invariant mass. For instance, the last visible peak from Figure 36 located around 90 GeV corresponds to the invariant mass of the Z boson [70]. To better identify particles, one can zoom in on specific parts of the graph to study the energy of their particle of choice.

The exercises provided dive deeper into the basic principles of invariant mass reconstruction, plotting histograms and more. We begin to understand the calculation of the

mass and produce the values for further plotting.

The Invariant mass does not represent a physical mass. It is a mathematical concept. It is used to investigate the existence of a particle whose decay products are known. The invariant masses of the decay products must be equal to the physical mass of the original particle.

The calculation of the invariant mass: $M = \sqrt{(E_1 + E_2)^2 - (\vec{p_1} + \vec{p_2})^2}$, where $(\vec{p_1}+\vec{p_2})^2$ is the square of the vector norm, calculated with the use of: $(\vec{p_1} + \vec{p_2})^2 = (px1 + p_{x2})^2 + (p_{y1} + p_{y2})^2 + (p_{z1} + p_{z2})^2$.

In the 'Open Data with CMS - outreach and education', we look into the relation between the particle's energy and the resulted peak and, following the dimuon example previously stated, we reproduce Figure 36 (the one above) for both low and high-energy muons, zooming in on the Z boson peak (see Figure 37).

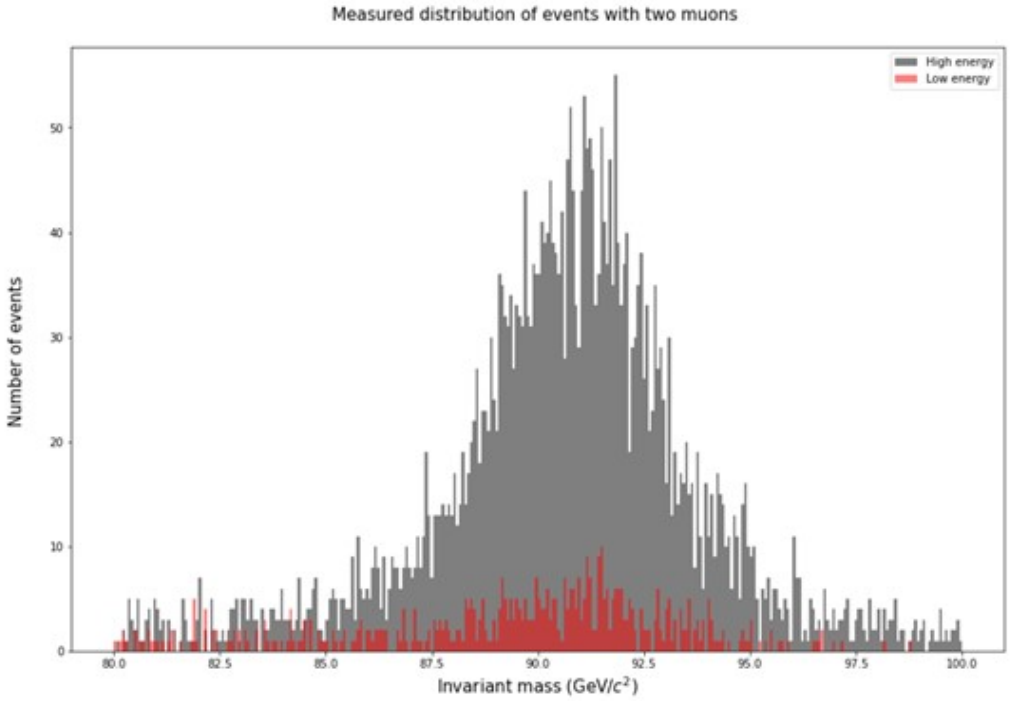


Figure 37. The Z boson peak presented on the measured ditribution of events for muons with both low and high energy. [76].

In Figure 38, we can see from the code how the momentum (p_T) (consequently the en-

ergy) for both muons was set to a threshold of $30\text{GeV}/c^2$ to create the different distribution for an observation of the importance of these selection criteria.

```
threshold = 30

highE = bump[(bump.pt1 >= threshold) & (bump.pt2 >= threshold)]
lowE = bump[(bump.pt1 < threshold) & (bump.pt2 < threshold)]

fig = plt.figure(figsize=(15, 10))

plt.hist(highE.M, 300, range = (80,100), alpha = 0.5 , color = 'black', label = 'High energy')
plt.hist(lowE.M, 300, range = (80,100), alpha = 0.5 , color = 'red', label = 'Low energy')

plt.xlabel('Invariant mass (GeV/$c^2$)', fontsize = 15)
plt.ylabel('Number of events \n', fontsize = 15)
plt.title('Measured distribution of events with two muons \n', fontsize = 15)
plt.legend()

plt.show()
```

Figure 38. Code used for plotting the muon histograms for two different energy regime: low and high. [76].

In the 'Open Data with CMS: making animations' notebook [77], we can observe an animation of the plot (Figure 37) created with 25000 data points. The x-axis is in units of energy (GeV) while the y-axis represents the number of events.

Due to the high number of data points, this animation can last around 8 minutes to be produced on a low-grade laptop. It provides helpful insight on how the number of points involved in the plot increases the accuracy and helps define the peaks. Below, we created a timelapse of the plots (Figure 39) and displayed the final histogram (the last plot in the timelapse).

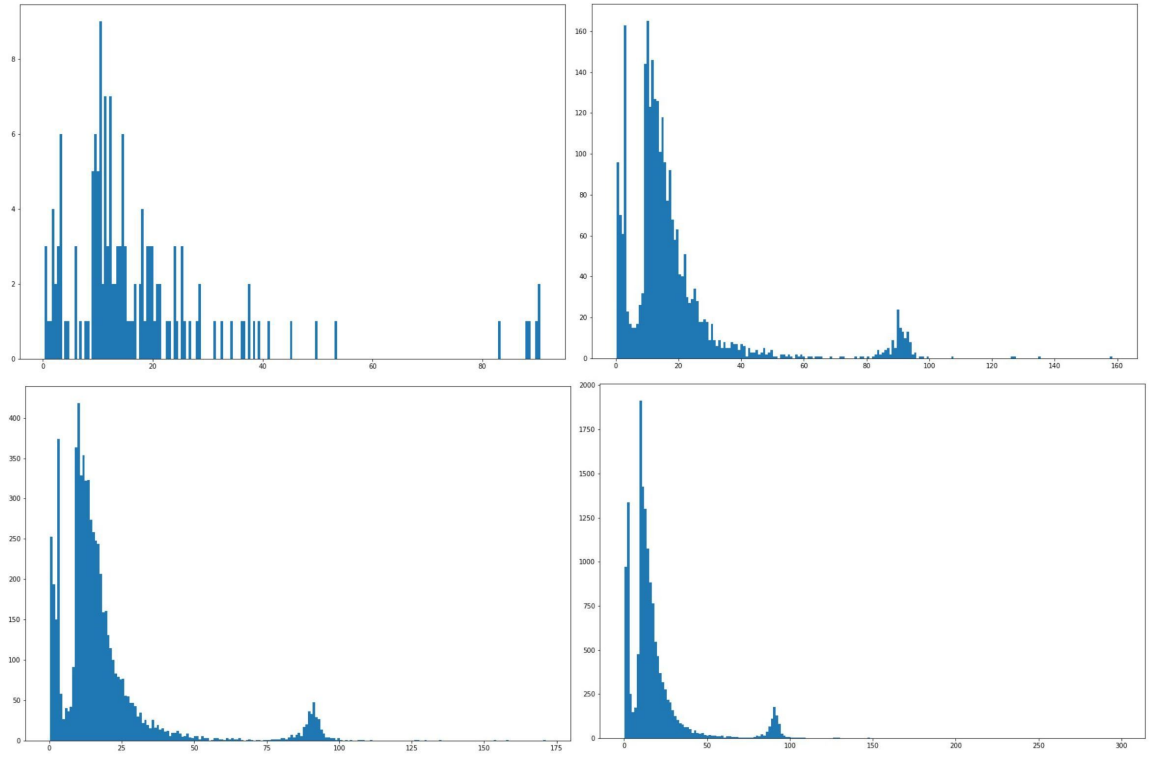


Figure 39. Snapshot of four different stages of the animation: an initial plot with a small number of events, two intermediate plots and the final one obtained after the addition of all the events available in the data set. [77].

Two other workshop notebooks are available to understand how to calculate the invariant mass and perform such an analysis on different data sets provided. We can plot various histograms (Figure 40) and compare them with Figure 41 to identify the particles that resulted in the decay. An identification of several of these peaks will be discussed in the next section.

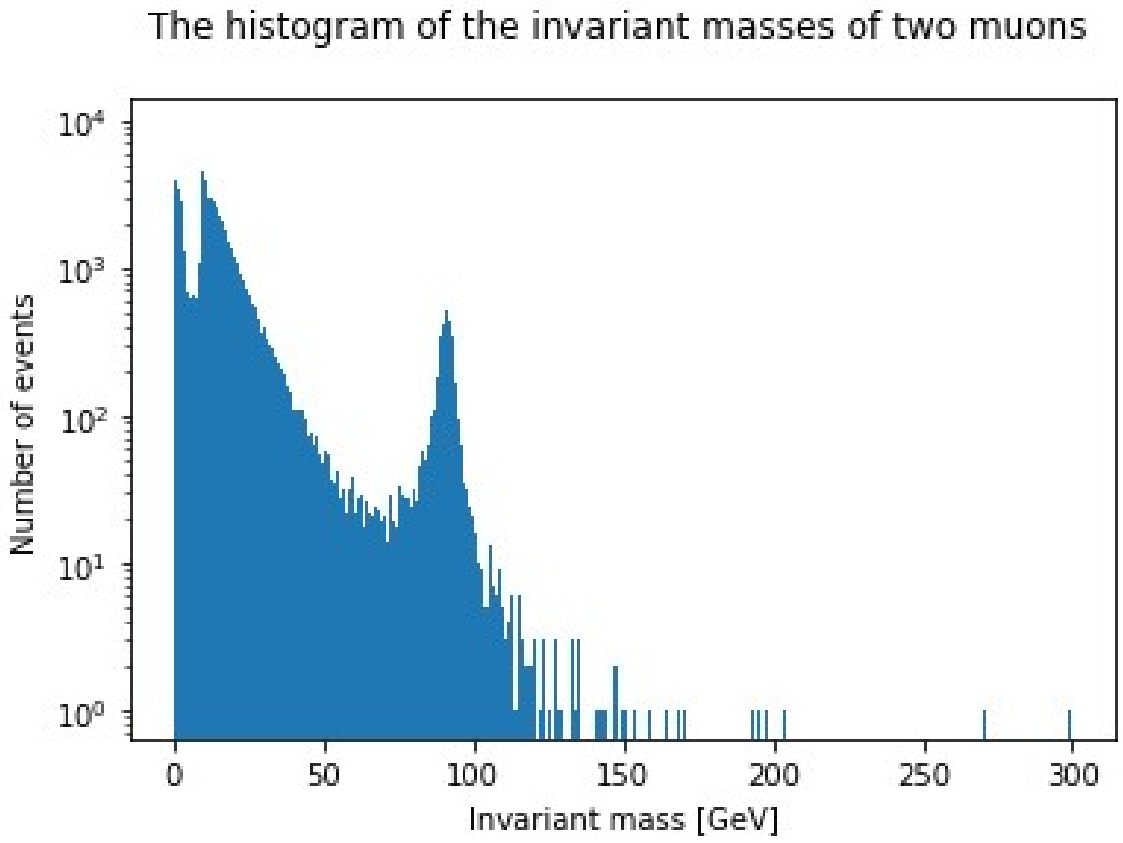


Figure 40. The invariant mass histogram plot obtained from one of the workshop activities. [40].

Particle	Mass [GeV]
η (eta)	0.548
ρ, (rho)	0.775
ω (omega)	0.782
ϕ (phi)	1.019
J/ψ (J/psi)	3.097
ψ' (psi)	3.686
Υ (ypsilon)	9.460
Z-boson	91.188

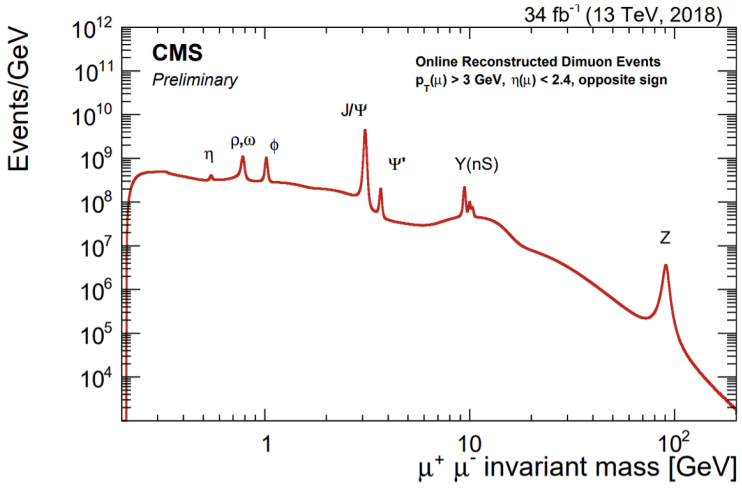


Figure 41. figure on the left - List of various particles' masses; plot on the right - CMS provided histogram from reconstructed dimuon events, displaying the corresponding peaks of the particles mentioned in the figure on the left. [40].

A disadvantage of this method, which relies on an online hub, is that when too many people are logged on, an error message appears. This means that we may have to wait before gaining access to the hub. The waiting time is usually relatively short (up to five minutes).

Analysing these datasets can also be done through Google Collab [78]. Another option is VISPA [79], an interactive scientific data analysis tool that can be accessed directly in the browser. It allows the code to run in Python, R or even ROOT, and is very useful for smaller data sets.

b In Jupyter using Python

A zip file contains several CSV files and Python notebooks, introductory and on various measurement calculations. These data sets are pre-selected to only include physically interesting events extracted from different decays for educational purposes [80].

There is a dedicated folder for notebooks explaining basic Python, the Jupyter notebooks and working with data. These are available along with short exercises to practise one's understanding of the topic. These notebooks are similar to the Excel sheets as they use CSV files to make plots using the data provided. The plots and data produced a match to various particles, mesons and bosons and confirmed our knowledge of the standard model. These notebooks are pretty self-intuitive, and very minimal coding knowledge is required to use them.

The 'Exercises with Open Data' folder dives deeper into the basic principles of invariant mass reconstruction, plotting histograms and more. Figure 42 was produced using the notebook provided by Open Data shows a histogram plotted using over 30000 events where exactly two muons were detected in a proton-proton collision. The peak on the plot represents a particle with an invariant mass of roughly 3.1 GeV, which corresponds to J/Psi Meson [72]. This plot is a zoomed-in version (option possible to be done within the notebook) of Figure 36.

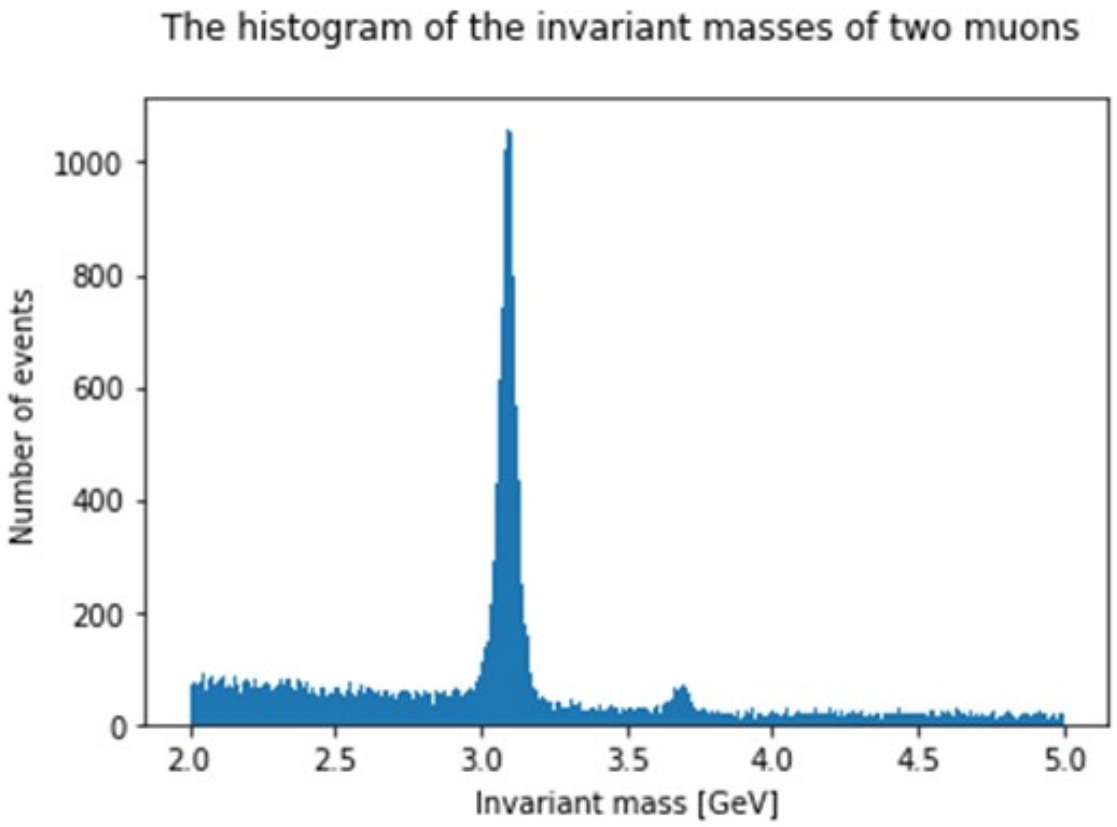


Figure 42. Histogram plotted on Jupyter notebooks using a dimuon dataset. Notice a peak at 3.1 GeV [80].

The J/Psi meson consists of a charm quark and a charm antiquark ($c, \bar{c}$. Its mass is roughly 3.5 times larger than a proton. The discovery of the J/Psi meson was made by two American groups that were working independently of each other in 1974. Its discovery helped improve the understanding of quarks and their interaction and supported the theory of the existence of the charm quark [81].

This process of discovering particles has been used for many years in particle physics and is relevant to the Higgs as the same methodology is used through finding its invariant mass in particle collisions. It's an example of how a vast amount of collision data recorded is used to find new particles and provide evidence for the standard model.

Figure 43 shows a histogram that used ten diphoton events, with an invariant mass between 120-130 GeV. The peak is at roughly 125.7 GeV. This corresponds to the mass of

the Higgs boson [72]. It was the only notebook which contained datasets directly relevant to the Higgs boson, highlighting the fact that selecting relevant data sets within CMS Open Data is a difficult task.

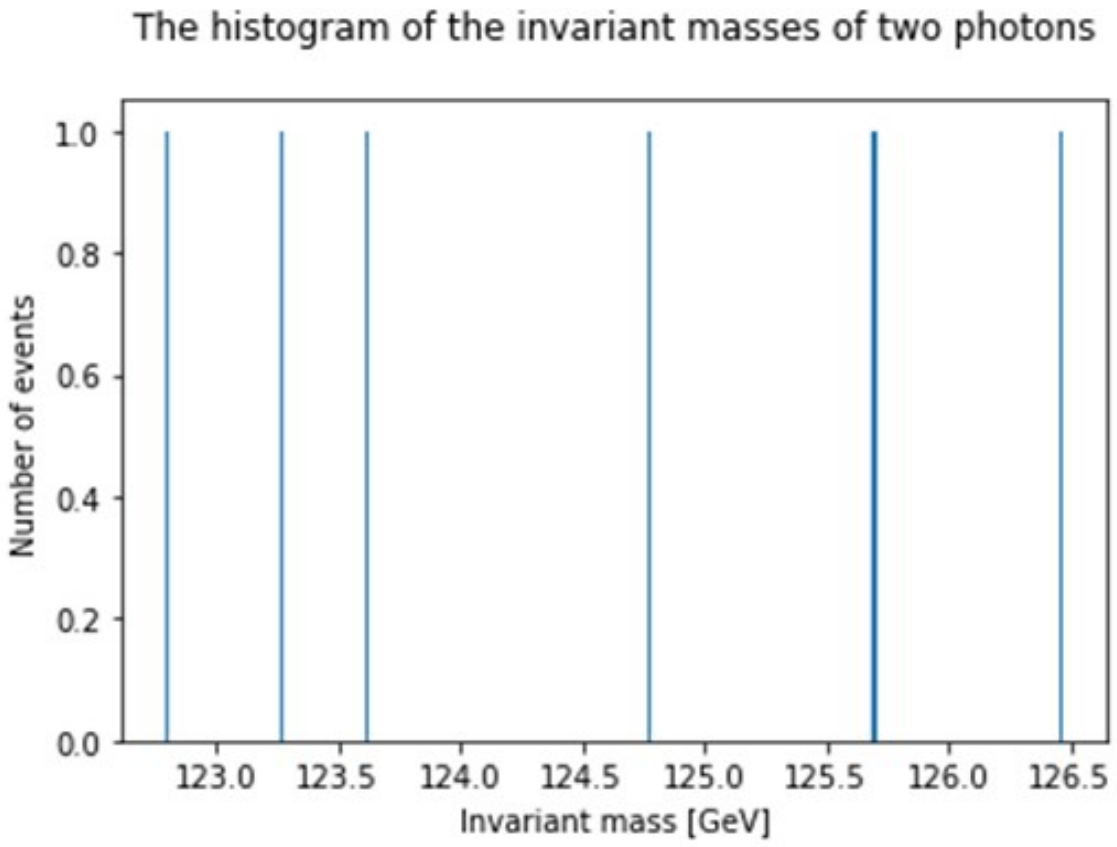


Figure 43. Histogram plotted on Jupyter notebooks using a diphoton dataset using 10 events [80].

In the following plot (Figure 44), we label various particles from the histogram obtained after running the code provided. There are 8 particle peaks noticeable, mentioned in Figure 45.

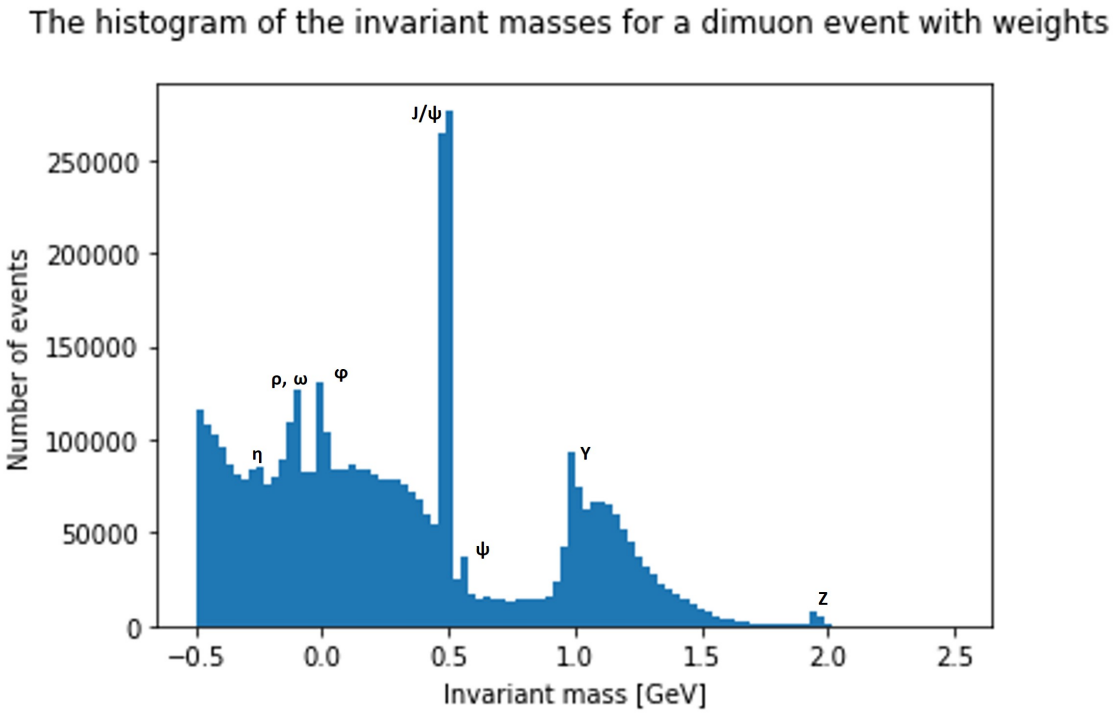


Figure 44. Dimuon histogram created with weighted data. [80].

Particle symbol	Particle
η	Eta meson
ρ	Rho meson
ω	Omega meson
φ	Phi meson
J/ψ	J/Psi meson
ψ	Psi meson
Υ	Upsilon meson
Z	Z boson

Figure 45. List of particle names. [40].

So far, the data we have worked with was carefully selected, containing only relevant events for the decays. In reality, the data has a tremendous background. Scientists perform clever cuts to minimise the background while keeping the signals. An important histogram that highlights such a procedure can be created with the 'Invariant-mass-histogram-weights'[80].

The next two figures (46 and 47) illustrate the importance of data selection. The Weighted plot allows for a clearer view of the peaks in comparison the unweighted one where no cuts were made.

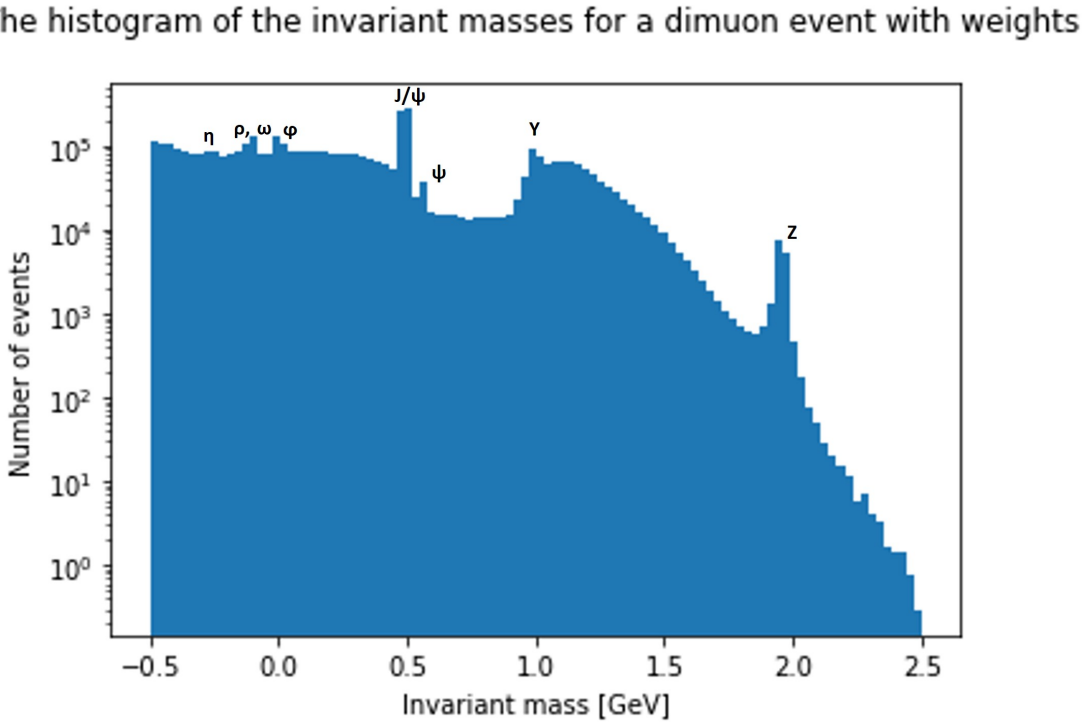


Figure 46. Dimuon histogram created with weighted data. [80].

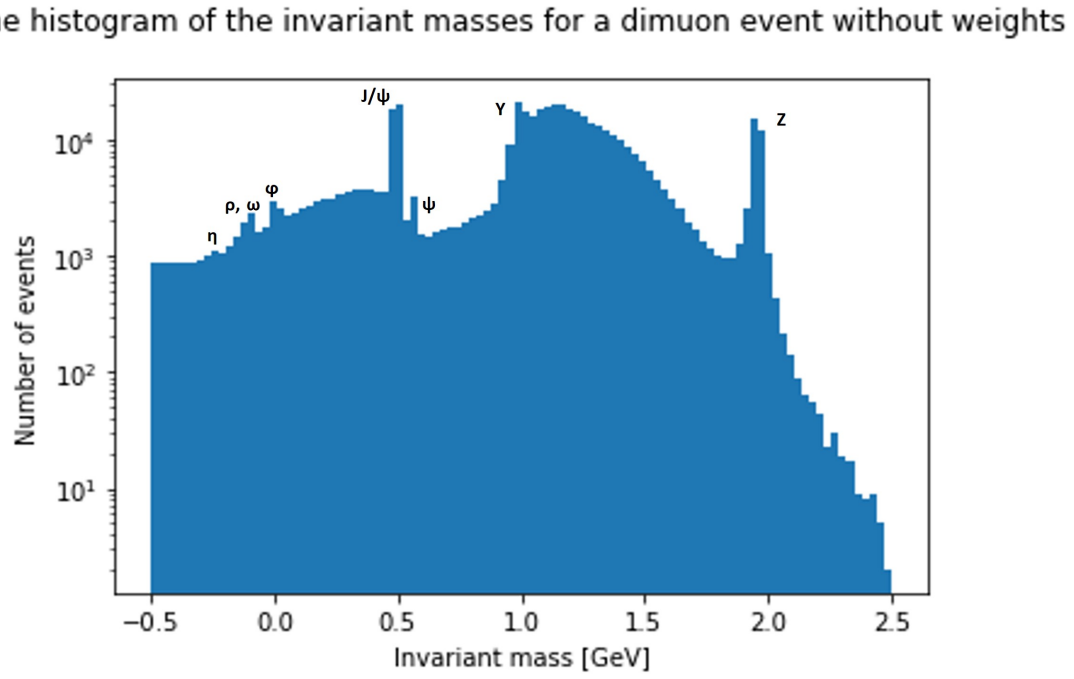


Figure 47. Dimuon histogram created with unweighted data, containing background. [80].

c In Jupyter using R

CERN provides the same resources for the same analysis to be completed in R [82]. An R tutorial is provided to understand the basic use of the R syntax as well as a complete example overlooking the calculation of the invariant mass.

5.3 Other Online Resources

5.3.1 The Particle Physics Playground

The Particle Physics Playground [83] is an online outreach programme which provides students with simplified particle physics data and python tools to interact with it. These are provided through notebooks that can be opened directly through Google Colab (recommended) or through Jupyter. Datasets are provided in HDF5 file format. Along with the CMS experiment, the playground also includes activities relating to other experiment, such as the CLEO experiment and the BaBar experiment. Although students are expected to have prior Python knowledge, this is an excellent way to learn about the standard model and how particles are discovered. Detailed youtube tutorials are provided to teach students how to use Google Collab as well as an FAQ section. After completing the activities, students are able to obtain an answer-key to their notebooks via email.

5.3.2 Computing tutorials for particle phyisics analysis

The Open Data website includes two other courses which aim to introduce undergraduate students to computing techniques used in Particle Physics. These follow up on the previously discussed resources available, acting as a starting point for the conduction of more advanced analysis.

The Computing Methods in High-Energy Physics course

Offered by the University of Helsinki, this course explores three different programming languages: FORTRAN, C++ and python. ROOT is introduced through a series of lectures which focus on data analysis and visualisation. The course introduces different softwares for the calculation of cross-sections and branching ratios. It also includes an event generation simulation and regroups the basics of CMSSW (to be discussed later in our report) as well as some notions on grid computing. Prior computer science skills are a prerequisite for the course. [84].

The CMS HEP Tutorial

Previously part of a one-week course, this tutorial is aimed at undergraduate students with knowledge of particle physics but no prior experience in data analysis. The materials are

available online and can easily be followed as an introduction to analysis tools in high-energy physics using data sets from the released CMS data. C++ and ROOT experience is advantageous, but an introduction is provided for the ROOT framework. Some of the topics approached include Data sets and Monte Carlo Simulations with practice questions available [85] .

5.4 Open Data Level 2 discussion

The Open Data website provides a significant amount of resources for the Beginner and Intermediate analysis of Level 2. If the guide is followed step-by-step, it builds one's understanding of the discovery of the Higgs boson and particle physics in general. The visualiser could benefit from a more thorough documentation with details on its physical aspects. Moreover, it would be interesting to observe other decays than just 4leptons and diphotons to investigate the effect of the extra background.

Concerning the histogram visualiser, it is quite straightforward to use and navigate through the different plots. It would however be interesting to see more plots directly related to the Higgs. It is also hard to find which datasets are relevant to a specific analysis as the portal contains so many.

The Binder analysis is a good starting point for further investigation of the practices used in particle physics. It is very convenient as the analysis is solely performed online, which could attract an audience with little computing experience. The guides are straightforward and challenge one's understanding through exercises while discussing the physics behind the plots. These would provide valuable information to students intrigued by high energy physics. Using Binder is however restricted at times due to a high number of users. From our experience, the longest waiting time to access the notebooks was about five minutes which does not significantly hinder the access to the plots.

Moving to the Jupyter analysis requires the installation of Python, though it does not demand more computing abilities than Binder. The notebooks provided are more extensive in terms of the physics measurements observed. There are notes on diverse particle physics quantities. We believe this resource achieves its objective - building knowledge at an intermediate level. The same analysis can also be reproduced in R.

The two tutorials discussed in section **5.3.2** would level someone's computing skills, preparing them for the advanced analysis.

6 Level 3 Analysis

CERN Open Data [86] offers data in many different file formats, from .csv or .txt files, provided for lighter analysis in excel and python to data provided in .root files using the event data model (EDM) for research level analysis with the full CMS Software (CMSSW) Framework [1]. The next sections will introduce the main parts of CMSSW, and the popular particle physics library ROOT [87] along with the cmsRun executable and python configuration file that are generally used for data manipulation and analysis. Following on that, we will explain the different elements which compose the data pipeline at CMS. Finally, we will discuss our experience setting up and working with these software frameworks through two example analyses provided by the CERN Open Data Portal, one using ROOT and the other using the CMSSW Framework.

6.1 An Overview of the CMS data analysis pipeline

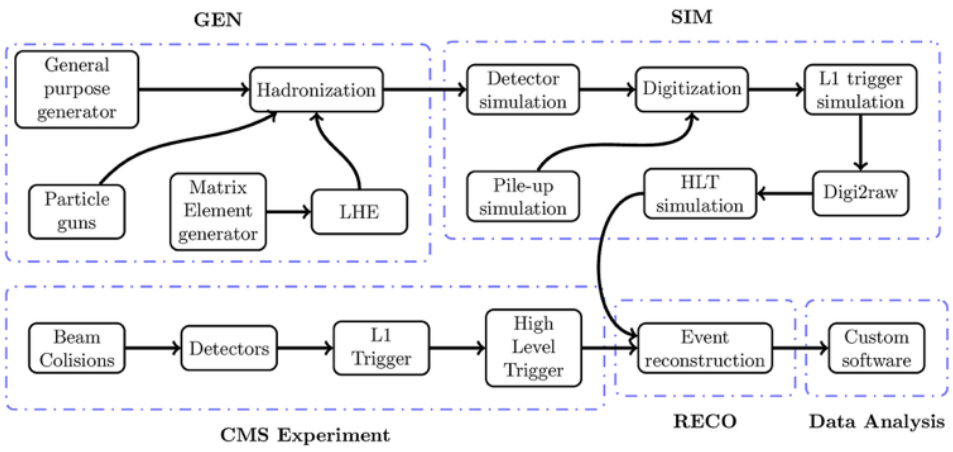


Figure 48. Data analysis pipeline. Figure from [88].

The CMS experiment consists of a complicated data analysis pipeline that can roughly be separated into four large sections (Figure 48) . The CMS experiment section takes raw data from the proton beam collisions using the L1 and High Level Triggers (HLT). In parallel to these steps, data is also generated from theoretical models (GEN) which

undergo a detector simulation (SIM). Event reconstruction (RECO) is then performed on both datasets to create higher level physics objects ready for analysis. After these steps, both reconstructed datasets undergo the same analysis and the results are compared. We will delve into the softwares and modules used to perform these steps by investigating CMSSW and ROOT.

6.2 CMSSW and ROOT

Note

To understand the inner workings CMSSW and ROOT, a basic understanding of object oriented programming is required. We found [89] to be a valuable resource.

6.2.1 Event Data Model (EDM)

The CMS Software Framework (CMSSW) is a custom C++ library [90] that consists of over a thousand subpackages [91], written by members of the CMS Collaboration. It contains everything needed to perform data analysis provided by the CMS detector including, simulation, generation and reconstructing algorithms and class definitions for data storage. The framework heavily utilises object oriented programming for data storage, as shown in Figure 49.

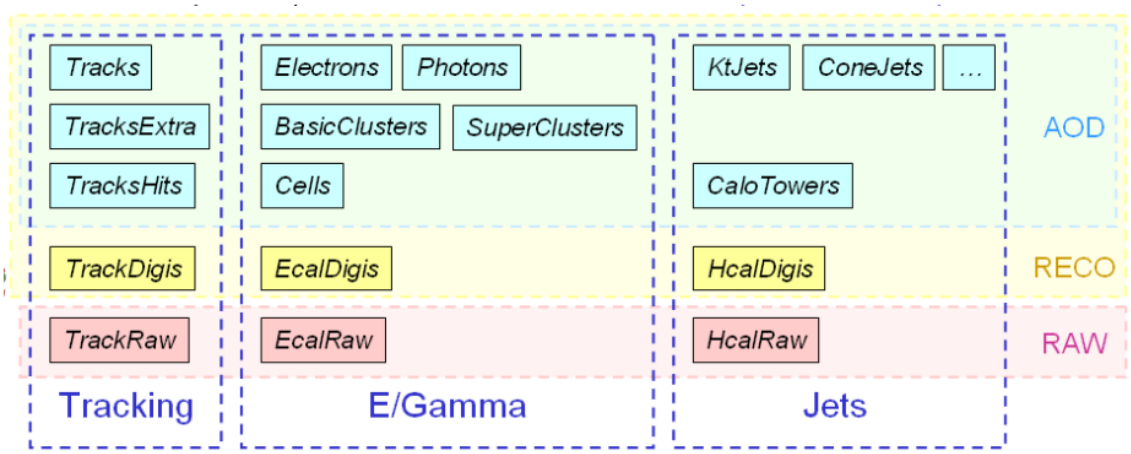


Figure 49. Event data container. Figure from [91]

Collision data is always stored in C++ objects (TrackRaw, KtJets,...) which are related to particular collisions, are then stored in the Event data container (hence the name Event Data Model) using the tree data structure (more about this in the ROOT section). CMS classifies objects stored in the Event object into three levels: RAW, RECO and AOD, based on how much processing is needed to extract them from the original detector signals.

RAW

RAW englobes C++ objects that store the very basic information directly read from the detector. This includes objects like TrackRaw for particle positions or EcalRaw and HcalRaw for energy related data.

RECO/AOD

To run an analysis on the collected RAW data, it first has to be reconstructed to higher level objects, like TrackDigits or EcalDigits. This simply means creating higher level objects from the low level RAW objects. Most of the time, this data is still too detailed for analysis so further reconstruction is performed to create high level physics objects like Electrons or Tracks called AOD (Analysis Object Data) data that is a subset of RECO. The properties of these particles/collisions are then analysed to compare it to theoretical predictions.

To achieve reproducibility, the Event object also stores information about the processes undergone by the original RAW data which is called Provenance Tracking.

Note that although the Event object forms the basis of the framework, CMS members aim to store objects which belong to different data levels in separate files. This often means that the RAW, RECO and AOD data objects that belong to the same Event are stored in separate .root files. Much of this is possible thanks to the flexibility of the tree data structure. All of this is discussed in detail below.

6.2.2 ROOT

ROOT is a C++ software framework [87], explicitly made for particle physics data processing and analysis. Although the CMSSW Framework works independently of ROOT, it implements many of its functionalities using the same source code. It also heavily utilises

the tree data structure [92] and .root files for data storage.

a Tree data structure

The tree data structure is optimised for storing large quantities of same-class data in so-called branches and leafs. A detailed picture can be observed in Figure 50. Many ROOT objects use the tree data structure in their implementation. A good example would be the RDataFrame which saves tabular datasets formed by rows and columns and is optimised to read selected rows and columns without loading the whole RDataFrame into memory. Another example is the TNTuple object, which uses the tree data structure by saving a column of floating-point numbers in branches. Data for histograms is often stored in TNTuples.

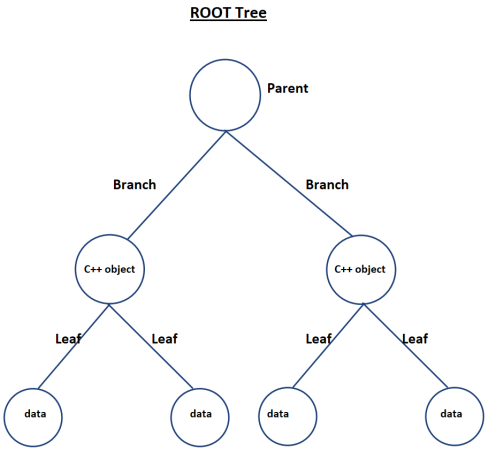


Figure 50. Tree data structure.

b .root files

Another essential ability of ROOT that the CMSSW software uses is that any user-defined C++ object can be stored in .root files in compressed binary format. The ROOT software can read these objects without the original class definitions by saving the necessary meta-data from the class definitions into the .root files together with the saved objects. ROOT also offers a graphical user interface to inspect these .root files.

Figure 51 depicts a tree data structure saved into a .root file, opened via the ROOT graphical user interface. The small tree, branch, and leaf icons correspond to parts of the tree

that help navigate the file.

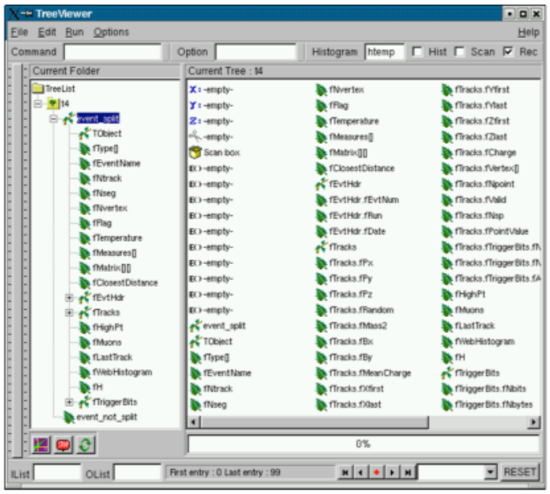


Figure 51. .root file containing a ROOT tree, opened via the ROOT GUI. Figure from [92]

Figure 52 shows a .root file which contains histograms.

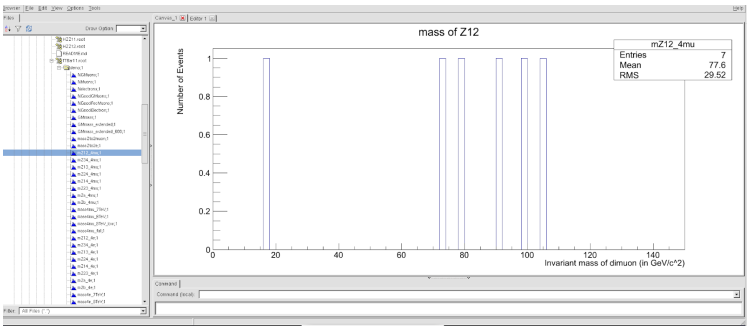


Figure 52. .root file containing a ROOT histogram, opened via the ROOT GUI.

6.2.3 cmsRun

The CMSSW Framework provides a lightweight modular system for performing processes

on data files. A python configuration file configures the cmsRun executable by specifying what data to use, which modules to execute on the data, the parameter setting for each module and what order to execute the modules in. All the code that interacts with the Event data is hidden within modules, for example, the algorithms for skimming or reconstruction, where skimming just means selecting the physically interesting events from a large dataset for further analysis. Altogether, six different modules can be loaded through the configuration file, and they each have a specific job. This modular nature ensures ease of development, distribution and reproduction.

a The six different module types [1]

Source

Source can read Events from a .root file or the global DAQ and create empty events that are later filled up.

EDProducer

The producer module can access data from an Event and produce and save new data to the same Event. An example use-case would be accessing the RAW data stored in an Event container, running reconstruction algorithms on this data and creating new RECO and AOD objects, which are then stored as a part of the same Event.

EDFilter

EDFilter can filter data based on predefined conditions. It reads data from an Event and returns a boolean value that is used to determine whether the Event should be processed or not.

EDAnalyzer EDAnalyzer can read data from an event, but it cannot add data to that Event, unlike the EDProducer. It typically writes output data to newly created .root files, e.g. histograms.

EDLooper

Controls looping over an input source's data.

OutputModule

Once all Events have been iterated, the OutputModule saves the edited Events into a new .root file.

Figure 53 shows an example where the Source("POOL Source") module loads an event from a .root file, two EDProducers("Digitizer", "Tracker") reconstruct the RAW data to higher level objects, then the EDFilter("NTrackFilter") skims the Events based on some conditions and finally the OutputModule("POOL Output") writes these Events into a new .root file. In this case the path specified that each Event is first accessed by the EDProducer("Digitizer") then the EDProducer("Tracker") and finally the EDFilter("NTrackFilter").

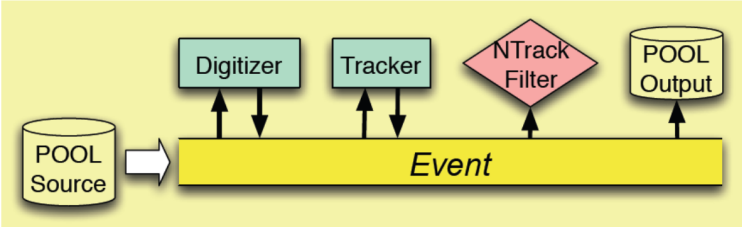


Figure 53. Modules accessing an Event object. Figure from [1].

Figure 54 provides an overview of how different data types, like RAW or RECO objects such as RecHit are stored in an Event container. These events are then sequentially stored in an EventCollection. To do an analysis, the config file configures the cmsRun executable by loading the predefined Producer. The cmsRun command then iterates through the EventCollection by opening an Event, extracting the AOD data, running the analysis on the AOD data and then writing the products back to the Event. It then moves on to the next Event, until this is executed on the whole EventCollection.

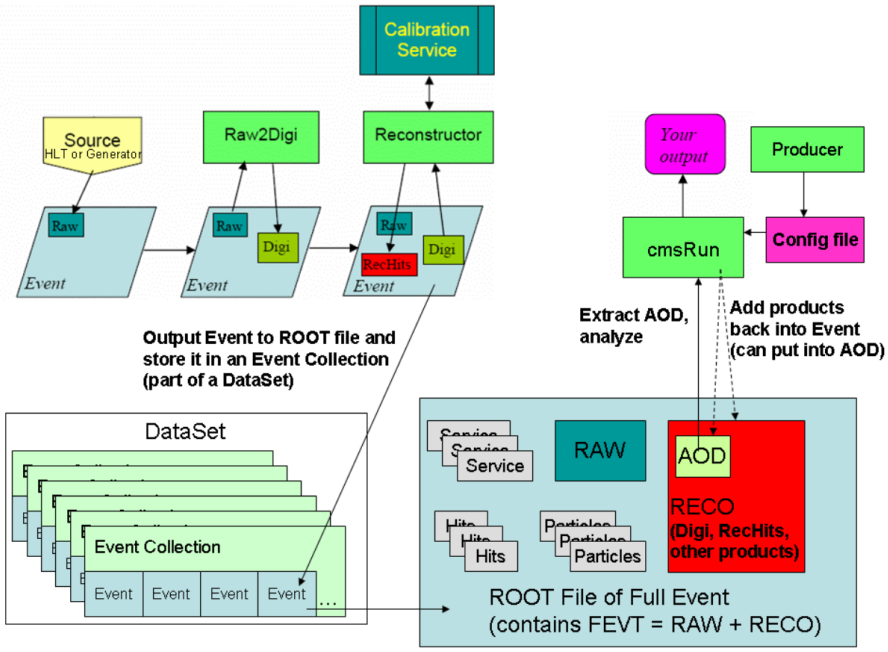


Figure 54. Event Data Model and cmsRun. Figure from [1]

A large part of research goes into writing these modules to find the valuable events and remove the background processes to reconstruct a final state particle's invariant mass. The analysis could be set up using a 50 line python config, while the HiggsAnalyzer module we used was almost 2500 lines long. This just shows that the challenging part of an analysis is writing the modules while running the cmsRun exe file is just the very last step of the whole process.

FWLite

The CMSSW Framework also has a lighter version called FWLite (Framework Lite) that allows quick editing and testing. It uses plain ROOT along with packages necessary to interact with CMS specific objects. The first example analysis we followed used a similar setting via plain ROOT software.

6.3 The data pipeline

6.3.1 Triggers

When the LHC is running at full capacity, roughly 1 billion protons collide every second. Each of these collisions leave a trace in the detector, but today's technology only allows one to save data for a few hundred collisions. This selection is done by the Trigger and Data Acquisition System (TriDAS), first by using a Level 1 Trigger that selects the 100,000 most interesting collisions. Then, the High Level Triggers (HLT) further select the 100 most interesting collisions using sophisticated and processing intensive algorithms. This cannot be done in real time, so data is stored in a short-term memory, called buffer until the partial reconstruction and selection is done by the HLT. After the final cut is made, the resulting RAW data and higher level objects from the HLT processing are stored in a C++ Event object.

6.3.2 Storing and Skimming data (CMS Computing Model)

The low level information of particle tracks (TrackRaw) or energy (EcalRaw), saved as RAW data (Figure 55), has to be reconstructed to higher level physics objects, like Electrons, TracksHits or ConeJets before analysis. This requires an immense computing power, with limited funding. To do this, the CMS experiment uses a grid system called the CMS Computing Model. The idea is that although data is collected at CERN, other steps, such as reconstruction and simulation, are distributed amongst other CMS Collaborators around the world, and uses their local computing resources. The grid system is separated into tiers based on functionality.

Tier-0

Tier-0 covers the one-site computing done at CERN. RAW data is received from the CMS Online Data Acquisition and Trigger System. It is repacked and archived as RAW data files and distributed into 10 datasets. These datasets then undergo reconstruction, producing RECO, AOD and miniAOD files which are then distributed amongst the Tier-1 centers.

Tier-1

There are 7 Tier-1 centers around the world, in the USA, France, Spain, Taiwan, the UK,

Germany and Italy. These sites tape archive the received RAW data then perform reconstruction and further skimming on it. The resulting RECO and AOD data is redistributed to CERN and other Tier-1 centers. Tier-1 centers also provide storage and redistribution for MC events generated at Tier-2 centers.

Tier-2

These centers perform Monte Carlo generations and simulations then distribute the resulting data back to the Tier-1 centers. The Tier-2 centers are responsible for providing data to the local physics groups, and in general their activities are driven by the needs of the CMS Collaborators located in their area.

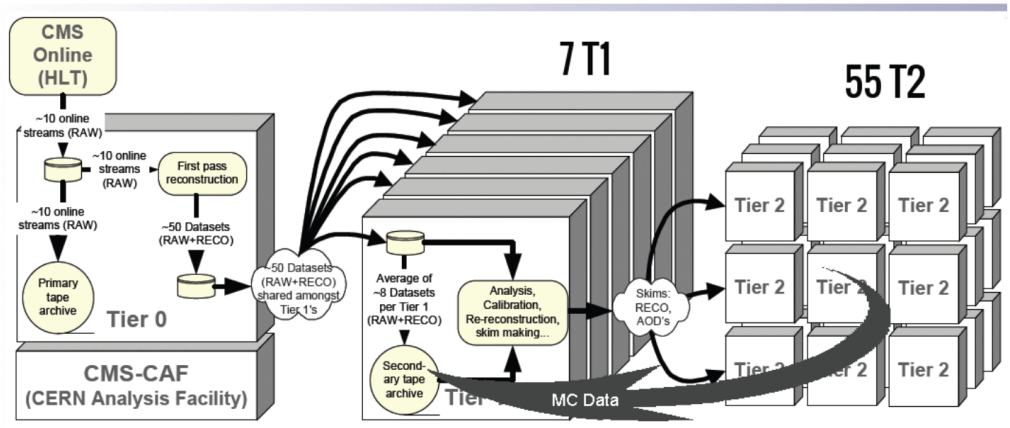


Figure 55. CMS Computing Model. Figure from [93]

6.3.3 Generating and simulating data

The CMS experiment requires the modelling of expected theoretical behaviour [94] via data simulation and a detailed comparison with experimental data. Simulated events are generated via computational models and in particular through Monte Carlo Event Generators, which consist of a series of algorithms which aim to calculate an approximate result by using probabilistic techniques, such as random sampling [95]. These methods are widely used in High Energy Physics as they allow the estimation of key quantities such as the sensibility of a detector or the form of a signal. Monte Carlo Event generators are interfaced onto the CMSSW software and form the key part of the simulation as they are used to generate HEP events. Most simulations at CMS use the PYTHIA 8 program, a general-purpose Monte Carlo event generator for the production of events in high-energy collisions [96]. It should be noted that simulated events must undergo several steps before they are in the required format for analysis (see figure 56) [97].

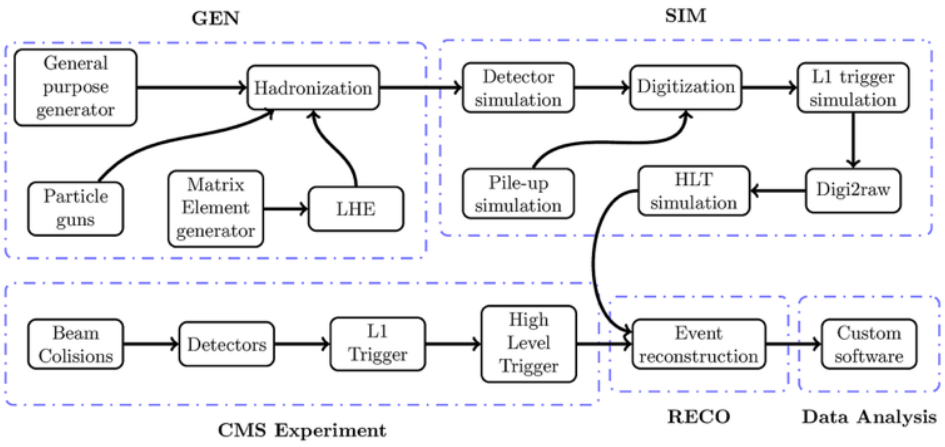


Figure 56. Diagram of the steps CMS data files go through before they can be analysed [98].

- **Event generation (GEN)**: Initial and final states are generated through particle collisions (primarily proton-proton) via the Monte Carlo Event generators [97]. Most of the files are written in the LHE (Les Houches Event) file format, which is primarily used to easily store processes across different programming languages [99]. Matrix elements (ME) allow us to easily stock information about our systems of study.

- **Simulation (SIM):** The interaction between particles inside the detector and the

detector (in our case CMS) is simulated. The detector response is obtained from the particle interactions [97]. The Geant4 detector simulation toolkit is at the core of most HEP simulations.

- **Reconstruction (RECO):** Particle trajectories and energies from the detector are reconstructed, allowing a full data analysis [97].

Simulations and detectors create an array of events which cannot all be recorded. Triggers (such as HLT and L1) are introduced to determine which particle collisions within the detector should be stored and which ones should be discarded. These triggers are simulated in the SIM step of the data pipeline and significantly reduce the size of the theoretical datasets. It should also be noted that within the detector, the effect of multiple interactions per beam and events overlay. This creates a high luminosity at the LHC, leading to a non-negligible level of background to the response of the detector system [99]. This effect is known as pile-up and is emulated via a detailed simulation of the detector, by selecting specific events. Simulated events are then stored in ROOT files where they can be analysed and processed, just like experimental data.

6.3.4 Monte Carlo event generators, the example of Pythia

Pythia 8 is a parton shower Monte Carlo event generator. This kind of computer program is designed to simulate high energy collisions down to their final stable state. The aim is to generate a large number of simulated collisions and track their properties down to their respective stable particles [94]. Generating such a high number of collisions is achieved through parton (quarks + gluons) showers.
Event generators make use of Monte Carlo methods which utilise pseudo-random numbers (generated via a computer) to simulate the event-to-event fluctuations one would observe in a real-life experiment.
The steps that come together in a complete event generator can be subdivided into 3 stages (see Figure 57):

- **ProcessLevel:** This is where the nature of the event is decided. This is often a hard process, where the strong coupling is perturbative to the system [95]. Put more simply, the coupling constant leads to quarks being unbounded to a particle.

- **PartonLevel:** This is where the parton shower happens. By the end of this step, a realistic partonic structure is obtained.
- **HadronLevel:** As the partons grow further apart from each other, the strong interaction coupling rises. This triggers the hadronisation process during which the partons are bounded into colourless hadrons [95].

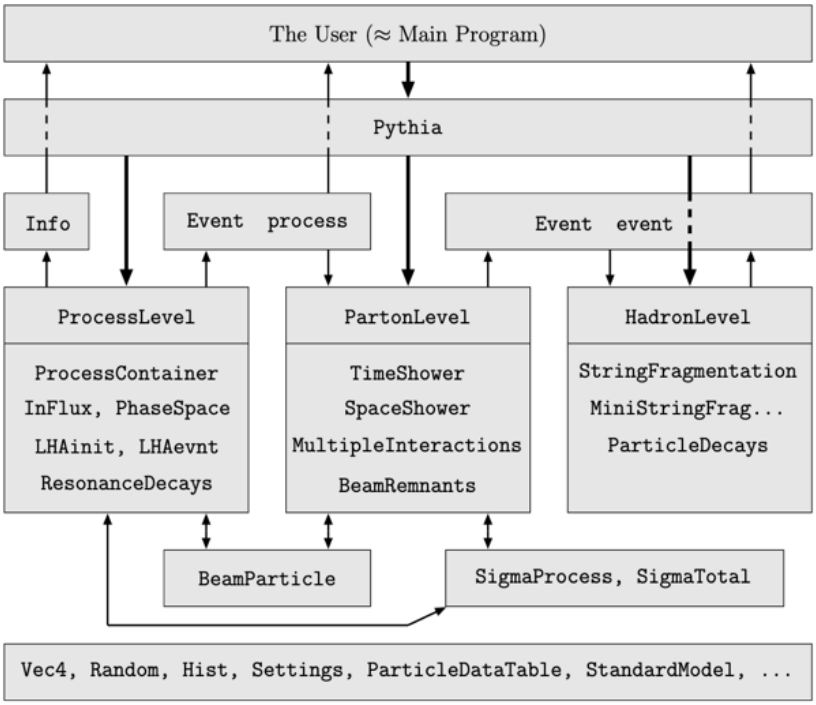


Figure 57. The relationship between the main classes in Pythia 8.

Information between these 3 levels is stored through the **Event** class (through **Event event** and **Event process**), which tracks the evolution of the partons between each step. A smaller Info class keeps track of useful properties of the system, such as kinematical variables like momentum and energy [95]. **BeamParticle** tracks the partonic content left in the beam after a certain number of interactions [95]. These generated events and classes are then simulated via the SIM step of the data pipeline.

6.4 Environment Setup

6.4.1 Docker vs Virtual Machine

To conduct the analysis, a specific environment had to be set up. We followed the instructions available on the CMS Open Data guide to set up the 2011-2012 CMS datasets [100]. We also ran demo-analyzers to check the validity of the CMS environment, making sure the analysis could be conducted smoothly. The CMS Open Data guide proposes two options to install this environment: through a virtual machine (VirtualBox) or a Docker container. The tutorial is relatively easy to follow, provided the user has a basic knowledge of Linux command lines[101][102]. However, whilst following these instructions, we ran through a series of errors/problems, which are all detailed in the Appendix.

Docker was more complicated to use than Virtual Box, as extra commands and application installations needed to be downloaded without any instructions provided by the CMS Open Data guide. Specifically, different programs needed to be downloaded separately to view the graphical interface related to ROOT. For a macOS/Linux user, XQuartz had to be downloaded, while Windows users required a VNC program.

Extra commands and installations were also needed for non-Linux users to interact with/ set up the environment properly. Again, the guide does not provide any information concerning these extra commands, making the setup of Docker extremely time consuming and complex. The VirtualBox option is much more self-contained as we only require a straightforward installation and a direct download of the CMS CVM 2015 image for the environment. In addition, the VirtualBox guide also contains a section that addresses frequent problems that occur in installing and setting up VirtualBox. One must however note that these two softwares only work on a desktop environment with sufficient GUI, and with a very stable network.

In addition, Docker needs about 40 Gbs to be successfully installed with the CMS 2015 image which is very large. VirtualBox only needs about 10 Gbs for comparison. One should however note that VirtualBox lags a lot and there is usually a lot of delays when

using this method to assess data, and freezes very frequently. Which application should be used depends primarily on the user's aims. VirtualBox is usually more convenient to use than the Docker container. However, if the user is running a high amount of data and needs a faster response, Docker may be a better choice but only if the environment has been set up correctly by the user, a task which is very difficult to achieve for people with little knowledge regarding Docker containers.

Also, one thing to note is that there are clashes between VM and Docker. The VM on the computer crashed and could not be used properly after the installation of Docker. The reasons for this crash are still uncovered, and no effective solutions have been found online.

6.4.2 The UCL HEP Cluster

In addition to the two applications discussed above, we also considered the UCL HEP Cluster[103]. Essentially, a computing cluster is composed of a high number of nodes (computers) which combine to provide the user with more computational power[104]. Specifically, the user may be able to use different computers in parallel through a batch farm and the nohup command[105]. This setting was essential in obtaining some of our results as it significantly reduced data processing time.

In Particular, we used the UCL High Energy Physics Cluster (dubbed as UCL HEP Cluster). This is a Linux Cluster made up of machines running CentOS 7 and Scientific Linux 6. It allowed the group to collaborate more easily on the project through a shared environment. It is very important to note that while the HEP Cluster already had general CERN tools and frameworks like CVMFS and ROOT installed, it did not have any CMS specific frameworks or tools, such as the CMSSW environment, which had to be sourced through the Singularity docker to conduct the $H ß ZZ ß 4l$ analysis. Singularity also enabled us to package the entire libraries and data available through CernVM, and enabled members to easily share and locally download files[106]. The only insufficiency was the fact that the Cluster does not provide any guidance regarding the set up of CMSSW environment, which is expected as it is not aimed at the general public.

One should also note that the HEP Cluster did not require any specific instructions to be set up and was much more powerful and responsive than the Virtual Machine which could only be locally run on our computers. In addition, the Cluster did not require any specific instructions to be properly set up as it was already ready to use. The Cluster was accessed through `ssh` with primary directory navigation and text manipulation, `ssh -Y` was also used to label the machine as a trusted client and open Root files through Xquartz in MacOS. For windows users, winscp and Putty were mainly used to interact with the GUI which was essential to visualise and work within the ROOT browser.

6.5 Advanced Level Analysis

In this section, we review two example analyses conducted by the group. The first one studies the decay of the Higgs boson into two tau leptons, and mainly makes use of the ROOT framework. The second analysis focuses on the decay of the Higgs boson into two Z bosons and uses the full CMSSW framework. We evaluate how reproducible these analyses are by the general public and how much insight they give into particle physics. In addition, we discuss possible improvements.

6.5.1 Awesome Workshop: $H \to \tau\tau$

a Background/Introduction

The Awesome Workshop example analysis is an educational version of a Higgs to two tau lepton analysis based on NanoAOD samples from the CERN Open Data portal. It uses real data and simulated events from the CMS experiment in 2012 and aims to study the decay of a Higgs boson into two tau leptons. This analysis very loosely follows the same setup as the official CMS analysis published in 2014, which used events recorded in 2011 and 2012 [107]. The CMS 2014 analysis established the existence of the Higgs boson's decay into two tau leptons. It required a full consideration of all systematic uncertainties, which is an extremely complicated and computationally expensive task. To ease our investigation, the analysis is reduced to a qualitative measure of the kinematic properties of the event, without conducting a statistical analysis [107].

The Higgs $H \to \tau\tau$ decay was discovered very recently as the tau leptons decay very quickly and the decay products always contain neutrinos (which are typically undetectable) and often hadrons [4]. Thus, the invariant mass cannot be fully reconstructed. A variable as close as possible to the invariant mass is plotted, "the visible mass", but this often has very poor resolution, leading to a peak at 91 GeV rather than 125 GeV [4]. Thus, the aim of this analysis is mainly to test whether we can recreate the visible mass plot shown below and assess the quality of the workshop instructions. The main background for this process is the $Z \to \tau\tau$ event. The Z boson may decay directly into two τ leptons, just like the Higgs boson, making this process very hard to distinguish from the signal. Other processes such as $Z \to ll$ and top anti top pair ($t\bar{t}$) production may also be easily misinterpreted as a pair

of tau particles. To enhance the Higgs event over these background processes, the weak boson fusion ($b\bar{b}$) process is simulated as it leads to additional identifying features (such as two high-energy jets from the final state quarks) [4], as well as the gluon fusion ($g\bar{g} \rightarrow H$). The data used in the analysis is split into two sets:

- **Monte Carlo:** Most processes for this analysis are generated via Monte Carlo simulations (theoretical data).
- **Real Data:** QCD multijet background decay occurs very often at the LHC. A proper simulation of this process is extremely complex and requires a lot of computational power. It is thus easier to use real experimental data rather than Monte Carlo simulations. These primary datasets are extracted from the physical detector data in AOD format[107].

The analysis relies on the following steps:

- **Skimming:** NanoAOD files containing the data and simulated events are pre-processed as a means to significantly reduce the size of the datasets. In addition, a pair selection is performed to identify the muon and tau collections the pair is most likely to have originated from.
- **Producing histograms:** Each quantity from the skimmed datasets is plotted as a histogram. Because of the data driven QCD estimations, histograms have to be produced with a same-charge tau lepton pair selection.
- **Making the final plot:** The histograms are combined to form the final plots, showing the data taken with the CMS detector along with a comparison with the expectation from the background estimations [107]

We expect to be able to locate the Higgs boson on the final plot, with an excess of events at 91GeV/c^2. Essentially, our goal is to recreate the following plot (Figure 58):

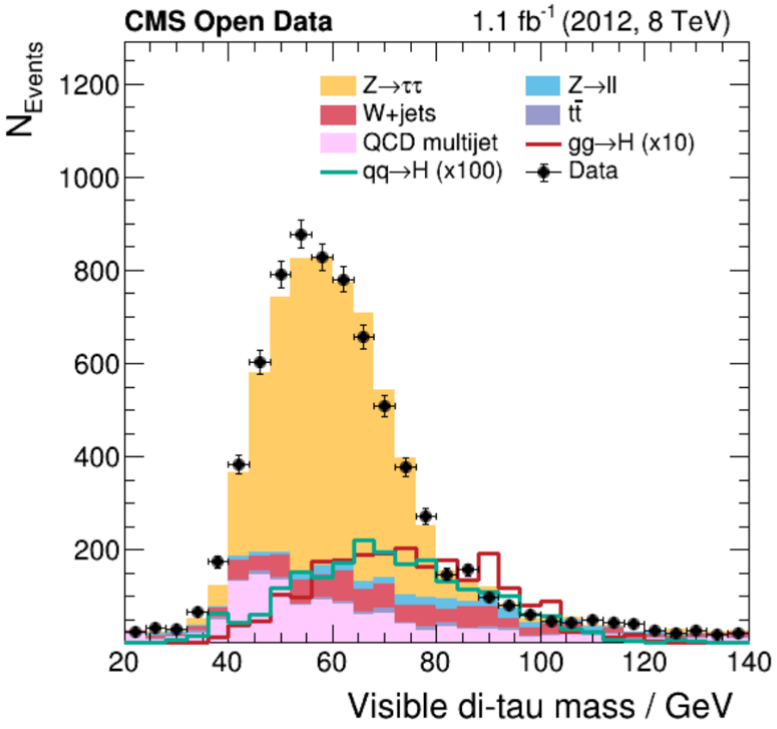


Figure 58. Visible ditau mass for the Higgs boson in the $H \to \tau\tau$ decay and background [107].

b Conducting the analysis

The analysis requires ROOT version 6.16 or above and a recent C++ compiler. The Workshop provides the user with a guide which outlines the steps to setting up the environment and running the analysis, providing them with the option of using either Docker or the CERN LxPlus service, which requires a CERN account. As we did not have a CERN account, the CERN LxPlus option had to be discarded. The instructions to set up the environment via Docker were not carved out for a macOS Mojave (version 10.14.6), from which the analysis was conducted. As a result, most of the investigation was run through the UCL HEP Cluster and we followed the CVMFS instructions for the environment set up. This was possible as general CERN tools like CVMFS and ROOT were already installed on the Cluster. The guide contains detailed instructions and accompanying videos for each step of the analysis. This made the investigation quite straightforward. The analysis could be conducted by someone with basic knowledge in programming and physics. However, not much insight into the computing aspect of the analysis is provided and there is very

little input asked from the user, other than changing the selection of python histogram files. The process however involves many steps, making it unintuitive. We compile an explanation of the computational steps and relevant files to provide the reader with a better understanding of the investigation.

Step 1: Skimming the events

- `download.sh`: NanoAOD files are downloaded from the CERN server.
- `skim.cxx`: Performs a selection on the minimal requirements of an event, using different selection rules for each process.
- `skim.sh`: compiles the skim.cxx file into an output executable script: `skim`. Executes the skim script on every sample and outputs a skim.csv file containing all the skimmed data.

Step 2: Producing histograms (Figure 59)

- `histograms.py`: produces histograms for each variable in the dataset and for each process resulting in the final state.
- `histograms.sh`: produces the histograms from skimmed samples by running `histograms.py`. It merges the histograms into a single root file: histograms.root. These may be opened using RootBrowser.

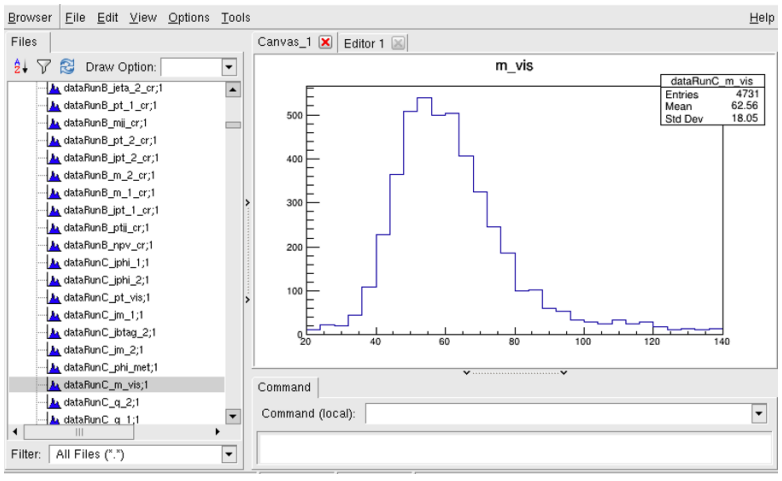


Figure 59. Invariant mass plotted against number of events for a Monte Carlo $H \to \tau\tau$ 2012 CMS root file. Opened with Root Browser

Step 3: Making the plots

- `plot.py`: implements the plotting step of the analysis. It combines the histograms into plots. More particularly, it combines the histograms to the QCD estimations: it takes the data from the control region in the histograms and subtracts all known processes defined in the simulation. The remainder is defined as QCD processes. The shape is then extrapolated into the signal region with a scaling factor.

The computational analysis took about an hour in total to run all the files, which is a feasible amount of time for most computers. We produced plots for 34 different parameters such as the transverse momentum and the angle of impact. More particularly, we obtain a plot of the visible mass (in GeV) (Figure 60):

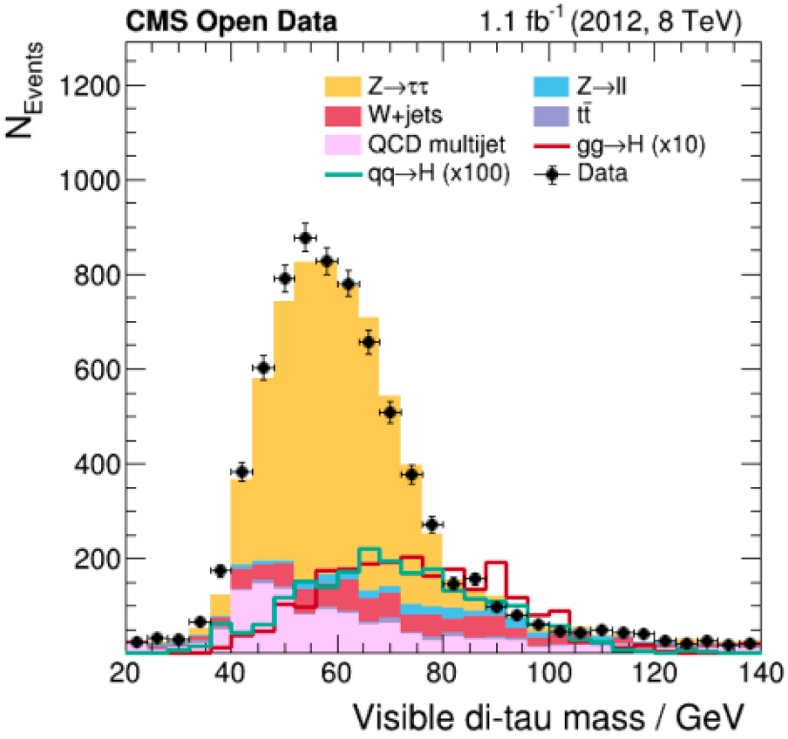


Figure 60. Final plot of the Awesome Workshop analysis, displaying the visible mass of the analysed processes and QCD background, opened with okular. We can see that the mass of the Higgs and the Z boson both peak below their expected mass of 125 GeV and 91 GeV.

As expected, the mass of the Higgs boson peaks at around 91 GeV, instead of 125 GeV. This is because tau leptons decay almost instantaneously into neutrinos which have

zero mass and are thus often undetected. We call this the visible di-tau mass, rather than the invariant mass, as the mass we find was only reconstructed from the products of the decay which could be recorded.

c Discussion

It is very clear that both the Monte Carlo simulation results and the graph presented on the Awesome Workshop website appear to be very similar. We can conclude that the instructions and the open data are successful in replicating the analysis. However, not much independent computational work is required from the user during the investigation. In theory, every file is customizable and open to use but in real life, this requires the user to have a pre-existing familiarity with the file system and relevant tools (such as ROOT and CVMFS). This would necessitate a guide to provide a proper insight into the computational tools the analysis uses, which it fails to do. Certain key files such as skim.cxx and histograms.py have a very large file size and contain very few comments. It is thus nearly impossible for a user to produce a significantly different analysis.

It should also be noted that as stated above, the analysis requires an environment which may be setup through Docker or through CVMFS. The group was able to run the analysis through CVMFS via the HEP Cluster. However, most people do not have access to such a resource making the setup only feasible through Docker. The instructions to set up the environment via Docker are not tailored to macOS users and certain Windows versions: this makes the analysis impossible to conduct for an important number of users.

It is important to note that this example analysis does not reflect a full rediscovery of the Higgs boson. Indeed, only the visible mass (91 GeV) is produced here as a reflection of the almost instantaneous subsequent decay of the tau leptons. This led the group to look at other Higgs decays through different Open Data files. In particular, an analysis of the $H \rightarrow 4l$ decay channel was conducted as a means to properly reconstruct the Higgs boson mass and highlight its existence.

6.5.2 CMS Higgs Analysis: $H \rightarrow ZZ \rightarrow 4\ell$

The CMS Open Data website provides an opportunity [86] to partially reproduce the analysis that led to the discovery of the Higgs boson in 2012 [16]. It is a simplified implementation of the original CMS Higgs to four lepton analysis, only uses about 50% of the original dataset, and merely focuses on the $H \rightarrow$ ZZ$\rightarrow$ $4l$ channel. Because of enhanced calibrations, the example employs legacy copies of the original CMS data sets in AOD format, which differs slightly from those used for the publication. It also employs legacy versions of the related Monte Carlo simulations, similar but not identical to those in the original paper. These simplifications lead to a reduced significance compared to the results broadcasted in the Higgs paper. The analysis is also divided into four levels, each going a few steps deeper into the process that in the end should result in a plot similar to the one in Figure 61, found in the 2012 Higgs paper.

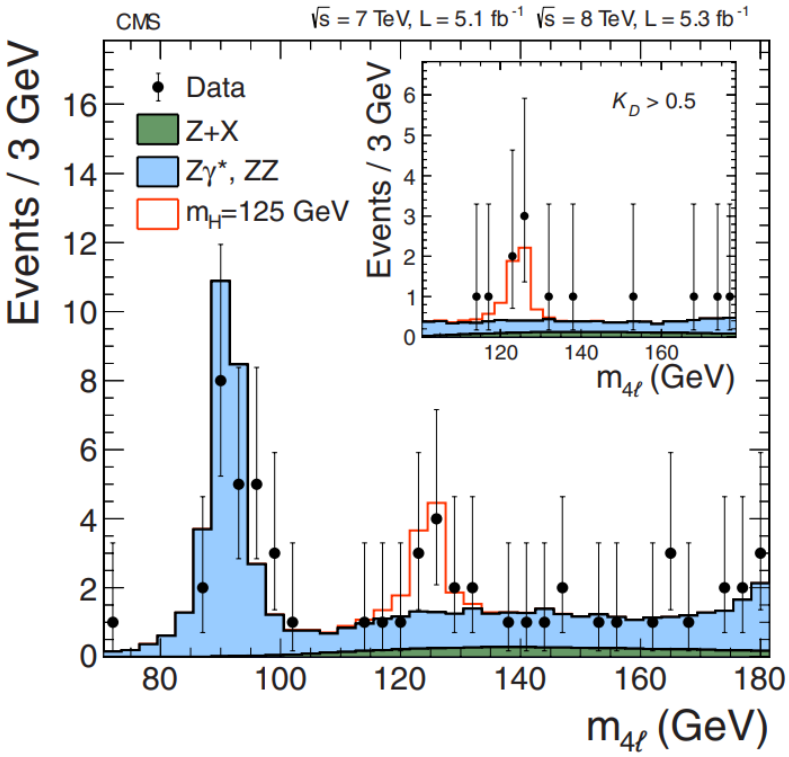


Figure 61. 4l invariant mass histogram from the 2012 Higgs paper [15].

a Physics Background

The only decay channel used in this analysis is the so-called "golden channel" $H \to ZZ \to 4l$ decay. The final leptons are either two muon pairs, two electron pairs or one pair each, particles that can be directly detected in the detector. The goal is to identify the four lepton decay products, used to first reconstruct the invariant masses of the two Z bosons. Then, these are used to reconstruct the invariant mass of the decaying particle, which should manifest as a resonant peak in the ZZ invariant mass histogram at around 125 GeV (the mass of the Higgs boson). The most prominent background is the quark to two Z boson annihilation, resulting in the same final states. However, the given data comes in a heavily skimmed AOD format containing only the most promising ZZ Higs candidates.

b Analysis Process

This whole process follows the general steps outlined in the cmsRun section. The necessary modules are loaded through a python configuration file. Then, the cmsRun exe file analyses the specified .root data files and produces new histogram .root files (one per each data file). Afterwards, the histograms are combined and a final plot is produced using the plotter.cc ROOT script, displaying the invariant mass of the decay products similar to the original plot (Figure 61).

In total, there are four main subsets of the whole dataset. Detector data from 2011 and 2012 and Monte Carlo generated data from 2011 and 2012. The analysis is generally the same on all subsets, except that detector data also needs a validation.txt file specified in the cmsRun config file, for data validation purposes.

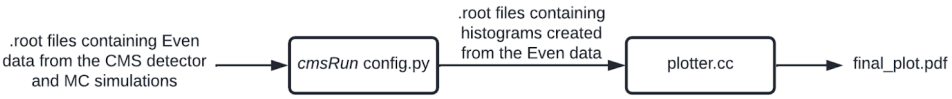


Figure 62. Analysis process.

`validation.txt`

The validation files are used in the python configuration file (config.py) to validate the real

datasets used for the analysis. There is one for 2011 and one for the 2012 detector data.

`indexfile.txt`

The index file contains URL links to the CERN online data servers. These files are used in the configuration file as the data source. All data files are accessed through these URLs, so the data doesn't have to be downloaded manually, which would require hundreds of gigabytes of space. For the Level 4 analysis, the user has to run the analysis on 70 index files.

`histogram.root`

These are the .root files that store the histograms generated from the data files during the analysis. There is one generated per index file. The histogram.root files are first combined into the final 15 histogram.root files using the hadd command. Finally, a plot can be produced running the ROOT script plotter.cc., which will output the final histogram in .pdf format.

`config.py`

This is the python configuration file that is passed as a parameter to the cmsRun executable. The configuration file follows the previously mentioned syntax, and it only requires some small amount of editing. The source has to be specified one by one for each index file. For the detector data, the validation.txt file also has to be provided.

`plotter.cc`

This ROOT script combines the histogram.root files to create a final plot shown in Figure 63.

c Step-by-step analysis

1. Download the necessary index and validation files from the CMS open data website.
2. Insert the index file as a data source into the config.py configuration file, and insert the necessary certificate.txt file as the validation file when running on detector data.
3. Run the analysis using the cmsRun config.py command from the CMSSW environment; this can last from a few hours to a few days, depending on the number of Events in the .root file. Upon completion, the analysis generates a new histogram.root file, one per index file.

4. Repeat the first three steps on every index file.

5. Combine the resulted histrogram.root files into 15 final ones, specified in the plotter.cc using the hadd command.

6. Run the plotter.cc ROOT script to combine these 15 final histogram.root files into a final plot showing the reconstructed invariant mass of the 4 lepton decay products.

7. Compare the final plot with Figure 63 from the original paper.

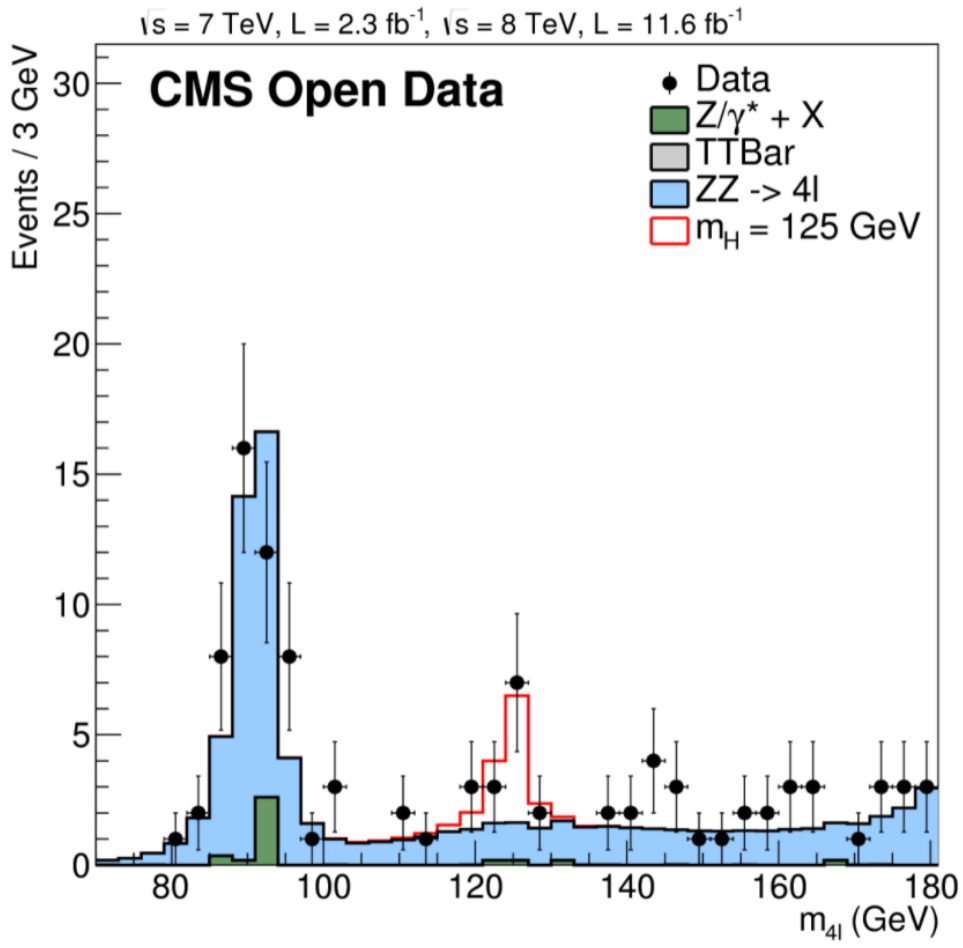


Figure 63. The 4l invariant mass plot provided for the Level 1 analysis. Figure from [108]

.

d CMS Open Data levels

The process is divided into four levels differing in difficulty, aimed at a broad audience. The analysis provides a conveniently organised folder with subfolders Level 1-4 where all of the necessary files are provided until Level 3.

Level 1

Level 1 does not require any processing, as it follow Step 7 from our Analysis step-by-step. It solely requires opening the provided histogram (Figure 63) and comparing it to the original one (Figure 61).

Level 2

The Level 2 analysis follows the last two steps of the Analysis step-by-step (6 and 7). All of the histogram.root files are provided in the folder, so no further computing is needed.

Level 3

The Level 3 example is the first level that requires the full CMSSW software and a working CMSSW environment, following steps 3 to 7. It is simplified in that the user will run the analysis on only two index files, which are provided along with the two configurations. One index file contains the URLs to data files of Monte Carlo Simulations from 2011, and the other contains the URLs to the actual experimental data file from 2011. A validation file is also provided for the latter to ensure the proper run of the analysis.

Level 4

The Level 4 analysis follows the complete step-by-step guide.

e Our experience and results from the Level 1-4 analysis

Level 1

Already at this stage, background knowledge in particle physics is required. The description of the analysis lacks information on the physics concepts involved, a critical part for our understanding. Further information should be provided to better interpret the plots obtained, particularly on invariant mass reconstruction.

Level 2

Computing related difficulties already arise at this level since access to a working ROOT environment is necessary to run the plotter.cc script. Guidance is provided on setting up a

Docker container or a virtual machine using VirtualBox, but these were difficult to implement, as we explained in the setup section. Running the plotter.cc file is straightforward otherwise, but this does not offer any insight into the analysis process or the physics in general.

Level 3

This Level 3 demands a similar amount of input from the user as level 2, with the difference that Level 3 also includes running two analysis jobs with the cmsRun executable. The index, validation, and config files are provided in the Level 3 folder, so no code editing is necessary. This will, in the end, produce the same plot as Levels 1 and 2 did, except for one data point added by the user-run analysis (the blue triangle on Figure 64).

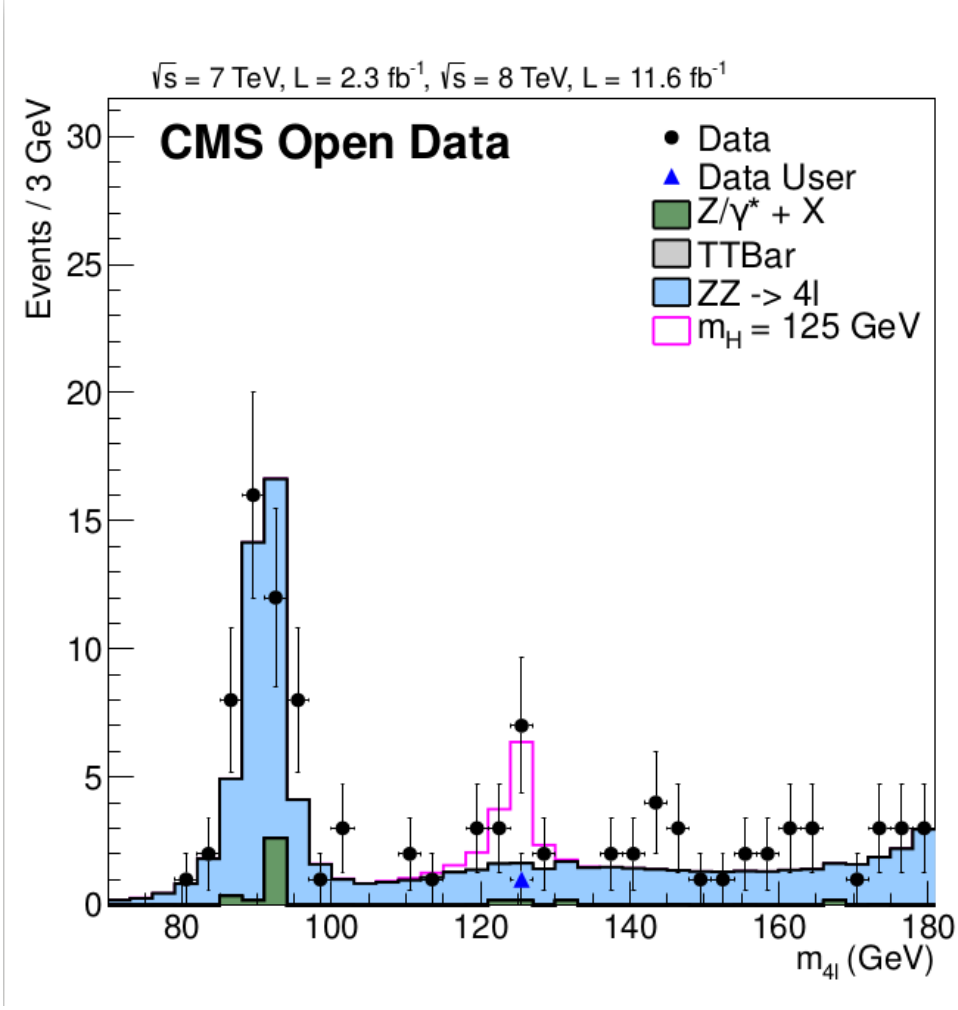


Figure 64. 4l invariant mass plot as a result of our Level 3 analysis.

Further investigating the configuration file contents and how everything is processed requires a read-through of the CMSSW documentation [1]. This experience proved challenging and time-consuming. Understanding the configuration file does not provide insight into the physics, since all of that is hidden inside the analyzer module. It is difficult to understand what is being done by the code without an expert knowledge of C++ , particle physics and thorough research.

Level 4

The overall process turns inefficient for several reasons. Before running the analysis, the index files have to be downloaded one-by-one, requiring around an hour for the process alone. They could be directly provided in the download folder on CMS website. Running these index files one-by-one proves to be even more inconvenient as they can take up to 2-3 days if a stable internet connection is provided for the entirety of the time. In the case of a connection loss, one would need to restart the run.

If one does not have access to a working CMS environment, the code editing must be done in a VirtualBox or Docker container, which was found to be a slow and laggy process. To speed up this process, we used the HEP cluster, with the CMSSW framework set up in singularity, making it possible to run the analysis on multiple index files in parallel. We also wrote a short python script that automated this process to avoid running the analysis on every index file one-by-one. The script can be found in the appendix.

Even after taking all these steps to simplify our workflow, and having access to 8 computers, it took us roughly a week to finish the analysis. Since most people do not have access to multiple computers this would take around 3-4 weeks. Having to do it through Docker, or VirtualBox could slow it down even more, depending on the specifications of the hardware used.

Nevertheless, after finishing the analysis and combining the files, we obtained the final result shown below, in Figure 65.

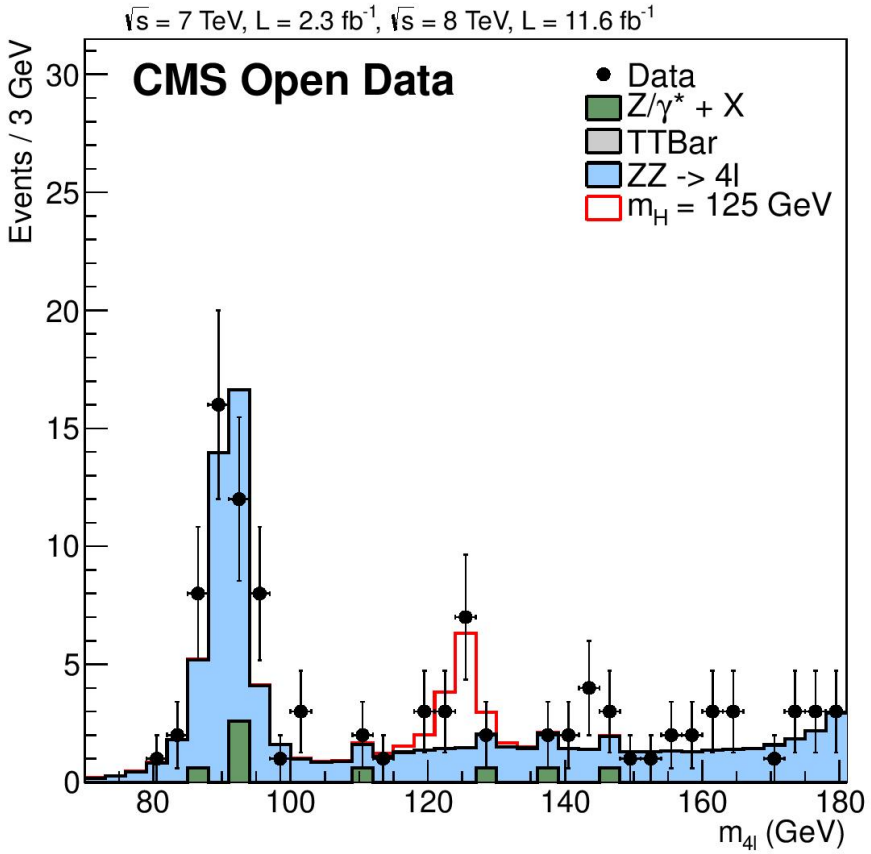


Figure 65. 4l invariant mass plot as a result of our Level 4 analysis.

We believe that some of the Monte Carlo generated data was different from the data used for the level 1-3 plots and this caused the minor deviations in the $Z/\gamma^* + X$ and $ZZ \to 4l$ backgrounds compared to the Level 1-3 plots (Figure 63 and 54). This is also an obvious difference when comparing it to the original plot from the Higgs paper (Figure 63), where that background is prevalent in the entire energy range. Nonetheless the simulated and detector data match and all of the significant features are visible on the plot. The resonant peak around 125 GeV corresponds to the invariant mass of the Higgs while the resonant peak around 91 GeV corresponds to the invariant mass of the Z boson.

We also decided to discover some of the other data in the histogram files, as the Level 4 analysis description suggests. For someone with a university physics background this provided some interesting extra information, but many of the plots used abbreviations that

we could not understand. The analysis could use more of this exploration and explanation approach to give the user more insight both into the software and the physics background.

Figure 66 Shows one of these plots. It is the reconstructed mass of the Z boson from two Muons.

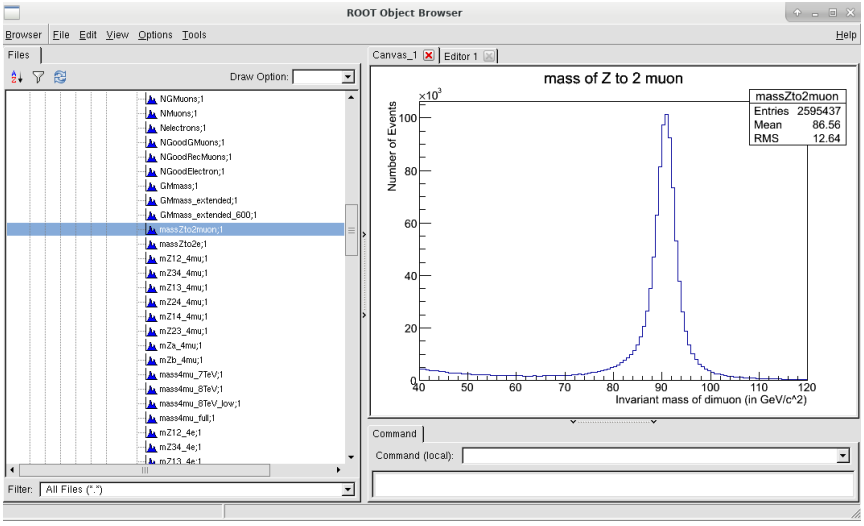


Figure 66. Invariant mass of the Z boson, from one of the histogram.root files generated by our Level 4 analysis.

Figure 67 shows a plot of the transverse momentum of an electron, originating from the Higgs to four lepton decay. This parameter is often used for skimmig as mentioned in section 3.3.4.

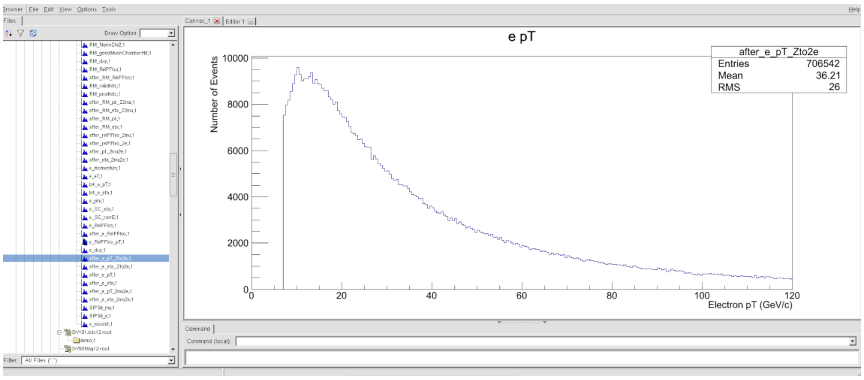


Figure 67. Transverse momentum histogram of the electrons from a Higgs to four lepton decay. This histogram.root file was also generated by our Level 4 analysis.

7 Conclusion

Providing insight into the work done at CERN to audiences without any particle physics knowledge, and with all kinds of background is understandably an extremely challenging task. However, we found that having multiple approaches like visualisers, excel sheets and jupyter notebooks could open up the CMS experiment, and CERN generally, to a wide range of people interested in the topic. These simpler examples also included enough background information to supply a user with a better understanding of both particle physics and data analysis.

The only real problem we faced was finding the right level of examples on the CMS Open Data website, as it does not provide a straightforward and accessible user experience. The search function works well, but someone unfamiliar with the website, or the technical terms may be overwhelmed by the sheer amount of search options and results.

Having an Introduction to CMS Open Data tab on the website could be a good solution to this problem, where example exercises could be organised into a increasingly difficult hierarchical order. This would also make them easy to find and help navigate the webpage. After completing a selected exercise, the user could be encouraged to either try the more difficult ones, or start looking for similar exercises using the search option.

Having access to example analyses using ROOT or the CMSSW framework could fit well into the more difficult end of the provided exercises, however, since these require a considerable amount of field specific knowledge, more detailed description is necessary both on the physics and the software part. This was an issue for both advanced analyses, since they both contained minimal information on CMSSW or ROOT, and even after finding the right resources, learning to use these libraries through their documentation proved to be a prolonged process.

An online lecture series that would build up this knowledge lecture by lecture would be a more suitable solution. It could follow a similar structure our report did, whereafter discussing the physics background, the software parts could be introduced one-by-one, with increasingly difficult example exercises. A intermediate knowledge of C++ and object oriented programming is essential for these exercises, but there are already many adequate resources present on the internet [109], [110], [111] that one could be pointed towards.

The initial difficulties setting up the environments, the lack of resources for debugging, and

the poor user experience could also discourage many users from going further into these analyses. To overcome this, CERN could provide access to a few remote PCs, with the environments already set up, similarly to our use of the HEP Cluster.

References

[1] C. Collaborators. (2022) Cmssw documentation. Accessed on 16/03/2022. [Online]. Available: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSSWFramework

[2] CERN. (2020) The higgs boson. Accessed on 16/03/2022. [Online]. Available: https://home.cern/science/physics/higgs-boson

[3] ——. (2020) Cern announces new open data policy in support of open science. Accessed on 16/03/2022. [Online]. Available: https://home.cern/news/press-release/knowledge-sharing/cern-announces-new-open-data-policy-support-open-science

[4] E. Nurse and A. Korn, *PHAS0040: Nuclear and Particle physics.* University College London, 2020, accessed on 16/03/2022.

[5] R. L. Yang, C. N. Mills, "Conservation of isotopic spin and isotopic gauge invariance," *Physical review*, vol. 96, no. 191, 1957, accessed on 16/03/2022. [Online]. Available: https://journals.aps.org/pr/pdf/10.1103/PhysRev.96.191

[6] S. L. Glashow, "Partial symmetries of weak interactions," *Science Direct*, vol. 10, no. 101, 1961, p. 1, doi: 10.1016/0029-5582(61)90469-2, accessed on 16/03/2022. [Online]. Available: https://sitn.hms.harvard.edu/flash/2020/chien-shiung-wu-a-heroic-experimental-physicist/

[7] W. Pauli, "Exclusion principle and quantum mechanics," *Nobel lecture*, p. 3, 1946, accessed on 12/03/2022. [Online]. Available: https://www.nobelprize.org/uploads/2018/06/pauli-lecture.pdf

[8] T. P. Pearsall, *Quantum Photonics*, 2nd ed. Springer, 2020, iSBN: 9783030473242, accessed on 13/03/2022.

[9] R. Feynman, *Feynman Lectures on Gravitation*, 1995, p.80, iSBN 0-201-62734-5. accessed on 13/03/2022.

[10] D. Griffiths, "Introduction to elementary particles," pp. 65–66, 2018, iSBN: 978-3-527-40601-2. accessed on 13/03/2022.

[11] CERN. (2018) Carl anderson discovers the positron. Accessed on 13/03/2022. [Online]. Available: https://timeline.web.cern.ch/carl-anderson-discovers-positron

[12] D. H. Perkins, *Introduction to High Energy Physics, 4th ed.* Cambridge, 2012, accessed on 17/03/2022. [Online]. Available: http://www.gammaexplorer.com/wp-content/uploads/2014/03/Introduction-to-High-Energy-Physics-4th-ED.pdf

[13] B. R. Martin, *Nuclear and particle physics - an introduction.* Wiley, 2006, accessed on 17/03/2022. [Online]. Available: https://www.fisica.net/nuclear/Martin

[14] G. Altarelli. (2000) The standard electroweak theory and beyond.

[15] S. C. et al. (CMS Collaboration), "Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc," *Phys. Lett.*, vol. B 716, p. 30, 2012, https://inspirehep.net/literature/1124338.

[16] CERN. (2022) The higgs boson. Accessed on 07/03/2022. [Online]. Available: https://home.cern/science/physics/higgs-boson

[17] ——. (2022) Our history. Accessed on 07/03/2022. [Online]. Available: https://home.cern/about/who-we-are/our-history

[18] ——. (2022) Our member states. Accessed on 07/03/2022. [Online]. Available: https://home.cern/about/who-we-are/our-governance/member-states

[19] ——. (2022) The accelerator complex. Accessed on 07/03/2022. [Online]. Available: https://home.cern/science/accelerators/accelerator-complex

[20] ——. (2022) Accelerators. Accessed on 08/03/2022. [Online]. Available: https://home.cern/science/accelerators

[21] ——. (2022) Pulling together: superconducting magnets. Accessed on 01/02/2022. [Online]. Available: https://home.cern/science/engineering/pulling-together-superconducting-electromagnets

[22] X. Vidal and R. Manzano. (2022) Rf cavities. Accessed on 15/03/2022. [Online]. Available: https://www.lhc-closer.es/taking_a_closer_look_at_lhc/0.rf_cavities

[23] CERN. (2022) The large hadron collider. Accessed on 01/02/2022. [Online]. Available: https://home.cern/science/accelerators/large-hadron-collider

[24] ——. (2022) The lhc starts up. Accessed on 01/02/2022. [Online]. Available: https://timeline.web.cern.ch/lhc-starts

[25] ——. (2022) The large electron-positron collider. Accessed on 23/02/2022. [Online]. Available: https://home.cern/science/accelerators/large-electron-positron-collider

[26] Stack-Exchange. (2016) Why did it take so long to find the higgs? Accessed on 16/03/2022. [Online]. Available: https://physics.stackexchange.com/questions/246841/why-did-it-take-so-long-to-find-the-higgs

[27] Fermilab. (2014) Accelerator. Accessed on 28/02/2022. [Online]. Available: https://www.fnal.gov/pub/tevatron/tevatron-accelerator.html

[28] CDF and D0, "Evidence for a particle produced in association with weak bosons and decaying to a bottom-antibottom quark pair in higgs boson searches at the tevatron,"

vol. 2, no. 10, p. 6, 2012, accessed on 16/03/2022. [Online]. Available: https://journals.aps.org/prl/pdf/10.1103/PhysRevLett.109.071804

[29] R. Hyneman, "Measuring higgs boson couplings, including to the top quark, in the diphoton decay channel with run 2 data collected by the atlas detector," *University of Michigan*, vol. 2027, no. 42, p. 17, 2020, accessed on 16/03/2022. [Online]. Available: https://deepblue.lib.umich.edu/handle/2027.42/155211

[30] Fermilab. (2014) The tevatron - 28 years of discovery and innovation. Accessed on 28/02/2022. [Online]. Available: https://www.fnal.gov/pub/tevatron/

[31] CERN. (2022) Cms images gallery. Accessed on 08/02/2022. [Online]. Available: https://home.cern/resources/image/experiments/cms-images-gallery

[32] ——. (2020) How a detector works. Accessed on 16/03/2022. [Online]. Available: https://home.cern/science/experiments/how-detector-works

[33] ——. (2022) Detector. Accessed on 08/02/2022. [Online]. Available: https://cms.cern/detector

[34] CMS. (2022) Cms. Accessed on 08/02/2022. [Online]. Available: https://home.cern/science/experiments/cms

[35] ATLAS. (2022) Atlas fact sheet. Accessed on 08/02/2022. [Online]. Available: https://cds.cern.ch/record/1457044/files/ATLAS

[36] ——. (2022) Atlas images gallery. Accessed on 08/02/2022. [Online]. Available: https://home.cern/resources/image/experiments/atlas-images-gallery

[37] ——. (2019) All together now: adding more pieces to the higgs boson puzzle. Accessed on 16/03/2022. [Online]. Available: https://atlas.cern/updates/briefing/adding-more-pieces-higgs-boson-puzzle

[38] N. Dicaire. (2015) Background estimations in the higgs to four leptons decay channel. Accessed on 16/03/2022. [Online]. Available: https://cds.cern.ch/record/2053731/files/DicaireReport2015_CERN.pdf

[39] W. Clavin. (2020) Extremely rare higgs boson decay process spotted california institute of technology, 2020. Accessed on 16/03/2022. [Online]. Available: https://www.caltech.edu/about/news/extremely-rare-higgs-boson-decay-process-spotted

[40] CERN, "Cms opendata for binder," 2022, accessed on 16/03/2022. [Online]. Available: https://notebooks.gesis.org/binder/jupyter/user/cms-opendata-ed-ooks-for-binder-0iejmy47/notebooks/SummerStudentWS.ipynb

[41] CMS, "Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc," *Physics Letters B*, vol. 10, no. 1016, p. 1, 2013, accessed on 16/03/2022. [Online]. Available: https://arxiv.org/abs/1207.7235

[42] I. C. education. (2020) Monte carlo simulations. Accessed on 07/03/2022. [Online]. Available: https://www.ibm.com/cloud/learn/monte-carlo-simulation

[43] A.Biyani. (2020) What is the monte carlo method. Accessed on 07/03/2022. [Online]. Available: https://careerfoundry.com/en/blog/data-analytics/monte-carlo-method/#monte-carlo-method

[44] CERN. (2016) Example of physics analysis: The case of the sm higgs boson production in the $h \to ww$ decay channel in the two-lepton final state. sm higgs boson production in the $h \to ww$ decay channel in the two-lepton final state. Accessed on 16/03/2022. [Online]. Available: http://opendata.atlas.cern/release/2020/documentation/physics/DL2.html

[45] ——. (2009) Particle-flow event reconstruction in cms and performance for jets, taus, and met. Accessed on 16/03/2022. [Online]. Available: http://cdsweb.cern.ch/record/1194487

[46] ——. (2010) Commissioning of the particle-flow event reconstruction with the first lhc collisions recorded in the cms detector. Accessed on 16/03/2022. [Online]. Available: http://cdsweb.cern.ch/record/1247373

[47] G. P. Matteo Cacciari, "Pileup subtraction using jet areas," *Physics Letters B*, vol. 659, pp. 119–126, 2008, accessed on 16/03/2022. [Online]. Available: https://arxiv.org/abs/0707.1378

[48] R. Salerno, "Higgs searches at cms," *arXiv preprint arXiv:1301.3405*, vol. 10, no. 48550, 2013, doi: 10.48550 , Accessed on 16/03/2022.

[49] ATLAS. (2018) Higgs boson observed decaying to b quarks – at last! Accessed on 16/03/2022. [Online]. Available: https://atlas.cern/updates/briefing/higgs-observed-decaying-b-quarks

[50] ——. (2018) Observation of $h \to b\bar{b}$ and vh production with the atlas detector. Accessed on 16/03/2022. [Online]. Available: https://arxiv.org/abs/1808.08238

[51] T. Binoth, "Two photon background for higgs boson searches at the lhc," vol. 18, no. 25, p. 1, 2000, accessed on 16/03/2022. [Online]. Available: https://inspirehep.net/literature/527546

[52] CMS, "Search for the standard model higgs boson decaying into two photons in pp collisions at s=7tev," *Physics Letters B*, vol. 710, no. 3, p. 1, accessed on 16/03/2022.

[Online]. Available: https://doi.org/10.1016/j.physletb.2012.03.003

[53] ATLAS, "Observation and measurement of higgs boson decays to ww^* with the atlas detector," *Phys. Rev. D*, vol. 92, p. 2, accessed on 16/03/2022. [Online]. Available: https://doi.org/10.1103/PhysRevD.92.012006

[54] ATLAS and I. Caprini, "Observation and measurement of higgs boson decays to ww^* with the atlas detector," *Physical Review D*, vol. 92, p. 2, 12 2014, accessed on 16/03/2022. [Online]. Available: https://journals.aps.org/prd/abstract/10.1103/PhysRevD.92.012006

[55] T. Binoth, M. Ciccolini, N. Kauer, and M. Krämer, "Gluon-induced ww background to higgs boson searches at the lhc," *Journal of High Energy Physics*, vol. 2005, no. 03, pp. 065–065, Mar 2005, accessed on 16/03/2022. [Online]. Available: http://dx.doi.org/10.1088/1126-6708/2005/03/065

[56] A. Sirunyan, "Observation of the higgs boson decay to a pair of leptons with the cms detector," *Physics Letters B*, vol. 779, p. 283–316, Apr 2018, dOI=10.1016/j.physletb.2018.02.004,Accessed on 16/03/2022. [Online]. Available: http://dx.doi.org/10.1016/j.physletb.2018.02.004

[57] CERN. What is cern open data? Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/docs/about

[58] ——. Cern open data policy for the lhc experiments. Accessed on 16/03/2022. [Online]. Available: http://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments

[59] Cms-Opendata-Education. the organization of wide files. Accessed on 16/03/2022. [Online]. Available: https://github.com/cms-opendata-education/cms-opendata-education

[60] ——. Csv file documentation. Accessed on 16/03/2022. [Online]. Available: https://github.com/cms-opendata-education/cms-opendata-education/blob/master/csvFileDocumentation.ipynb

[61] ——. Csv database. Accessed on 16/03/2022. [Online]. Available: https://github.com/cms-opendata-education/cms-opendata-education/blob/master/csvDatabase.csv

[62] CERN. The cms event visualiser. Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/visualise/events/cms

[63] ——. Cms guide to the educational use of cms open data. Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/docs/cms-guide-for-education#visualise-collisions

[64] ——. Cern accelerating science. Accessed on 16/03/2022. [Online]. Available: https://cms.cern/index.php/detector

[65] ——. Cern accelerating science. Accessed on 16/03/2022. [Online]. Available: https://cms.cern/news/cms-detector-design

[66] ——. Cms document. Accessed on 16/03/2022. [Online]. Available: https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=5582

[67] CMS. Visualizing 4lepton events. Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/visualise/events/cms

[68] CERN. Visualizing diphoton events. Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/visualise/events/cms

[69] ——. Visualizing histograms. Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/visualise/histograms/

[70] ——. glossary. Accessed on 16/03/2022. [Online]. Available: https://cms.cern/content/glossary

[71] ——. Cms-opendata-education/cms-spreadsheet-materials-multiple-languages: This repository contains exercises using cms open data for spreadsheet software in multiple languages. Accessed on 16/03/2022. [Online]. Available: https://github.com/cms-opendata-education/cms-spreadsheet-materials-multiple-languages

[72] P. Z. D. Group). (2020) Prog. theor. exp. phys. Accessed on 16/03/2022. [Online]. Available: https://discovery.ucl.ac.uk/id/eprint/10116889/1/ptaa104.pdf

[73] CERN. The binder project. Accessed on 16/03/2022. [Online]. Available: https://mybinder.org/

[74] ——. Github: Cms online notebooks for binder. Accessed on 16/03/2022. [Online]. Available: https://mybinder.org/v2/gh/cms-opendata-education/cms-online-notebooks-for-binder/master?filepath=Guide-to-using-Python.ipynb

[75] ——. (2022) Cms opendata for binder. Accessed on 16/03/2022. [Online]. Available: https://notebooks.gesis.org/binder/jupyter/user/cms-opendata-ed-ooks-for-binder-0f7x4a6n/notebooks/quick-start-to-CMS-open-data.ipynb

[76] ——, "Cms opendata for binder," 2022, accessed on 16/03/2022. [Online]. Available: https://notebooks.gesis.org/binder/jupyter/user/cms-opendata-ed-ooks-for-binder-oc0ueagf/notebooks/Open-Data-with-CMS-outreach-and-education.ipynb

[77] ——, "Cms opendata for binder," 2022, accessed on 16/03/2022. [Online]. Available: https://notebooks.gesis.org/binder/jupyter/user/cms-opendata-ed-ooks-for-binder-8yrxkmid/notebooks/Open-Data-with-CMS-making-animations.ipynb

[78] ——. (2022) Cms opendata for binder. Accessed on 16/03/2022. [Online]. Available: https://notebooks.gesis.org/binder/jupyter/user/cms-opendata-ed-ooks-for-binder-0iejmy47/notebooks/SummerStudentWS.ipynb

[79] CMS. (2022) Vispa. Accessed on 16/03/2022. [Online]. Available: ttps://vispa.physik.rwth-aachen.de/

[80] ——. (2022) Jupyter notebooks using cms open data. Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/record/5101

[81] T. E. of Encyclopaedia Britannica. (2022) Accessed on 16/03/2022. [Online]. Available: https://www.britannica.com/science/J-psi-particle

[82] E. N. Villegas Garcia. (2022) Instructions for use of cms open data in r. Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/record/5102

[83] (2022) P3 particle physics playground. Accessed on 16/03/2022. [Online]. Available: https://particle-physics-playground.github.io/

[84] S. Lehti. (2022) Computing methods in high-energy physics. Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/record/61

[85] S. A. Sander, Christian. (2022) Cms hep tutorial. Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/record/50

[86] C. Collaborators. (2022) Cern open data. Accessed on 16/03/2022. [Online]. Available: https://opendata.cern.ch/

[87] ——. (2022) Root framework. Accessed on 16/03/2022. [Online]. Available: https://root.cern/

[88] ——. (2022) Cmssw documentation. Accessed on 16/03/2022. [Online]. Available: http://opendata.cern.ch/docs/cms-mc-production-overview

[89] R. Data. (2022) W3 school. Accessed on 16/03/2022. [Online]. Available: https://www.w3schools.com/cpp/cpp_classes.asp

[90] C. Collaborators. (2022) Cmssw framework github. Accessed on 16/03/2022. [Online]. Available: https://github.com/cms-sw/cmssw

[91] S. Malik. (2022) Cmssw documentation. Accessed on 16/03/2022. [Online]. Available: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSSWFramework

[92] C. Collaborators. (2022) Root trees. Accessed on 16/03/2022. [Online]. Available: https://root.cern.ch/root/htmldoc/guides/users-guide/Trees.html

[93] ——. (2022) Cmssw documentation. Accessed on 16/03/2022. [Online]. Available: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookComputingModel

[94] M. Borodin, K. De, J. G. Navarro, D. Golubkov, A. Klimentov, T. Maeno, D. South, and A. Vaniachine, "Unified system for processing real and simulated data in the atlas experiment," *arXiv preprint arXiv:1508.07174*, 2015.

[95] T. Brall, M. Dommert, W. Rühm, S. Trinkl, M. Wielunski, and V. Mares, "Monte carlo simulation of the cern-eu high energy reference field (cerf) facility," *Radiation Measurements*, vol. 133, p. 106294, 2020.

[96] PYTHIA. Pythia. Accessed on 16/03/2022. [Online]. Available: https://pythia.org/

[97] A. Collaborators. Atlas open data. Accessed on 16/03/2022. [Online]. Available: http://opendata.atlas.cern/books/current/openatlasdatatools/_book/data_and_simulated_data.html

[98] C. Collaborators. Cms monte carlo production overview. Accessed on 16/03/2022. [Online]. Available: http://opendata.cern.ch/docs/cms-mc-production-overview

[99] J. Alwall, A. Ballestrero, P. Bartalini, S. Belov, E. Boos, A. Buckley, J. M. Butterworth, L. Dudko, S. Frixione, L. Garren *et al.*, "A standard format for les houches event files," *Computer Physics Communications*, vol. 176, no. 4, pp. 300–304, 2007.

[100] Opendata.cern.ch. Cern open data portal. Accessed on 16/03/2022. [Online]. Available: http://opendata.cern.ch/docs/cms-guide-for-research.

[101] ——. Cern open data portal. Accessed on 16/03/2022. [Online]. Available: http://opendata.cern.ch/docs/cms-virtual-machine-2011.

[102] ——. Cern open data portal. Accessed on 16/03/2022. [Online]. Available: http://opendata.cern.ch/docs/cms-guide-docker.

[103] Hep.ucl.ac.uk. (2019) Software/geant4/ucl hep cluster - pbtwiki. Accessed on 16/03/2022. [Online]. Available: https://www.hep.ucl.ac.uk/pbt/wiki/Software/Geant4/UCL_HEP_Cluster.

[104] P. R. Computing. What is a cluster? Accessed on 16/03/2022. [Online]. Available: https://researchcomputing.princeton.edu/faq/what-is-a-cluster.

[105] Hep.ucl.ac.uk. (twiki login) batchfacilities ¡ computing ¡ twiki. Accessed on 16/03/2022. [Online]. Available: https://www.hep.ucl.ac.uk/twiki/bin/view/Computing/BatchFacilities.

[106] sylabs.io. Singularity container. Accessed on 16/03/2022. [Online]. Available: https://sylabs.io/guides/3.5/user-guide/introduction.html.

[107] software carpentry. Awesome workshop analysis. Accessed on 16/03/2022. [Online]. Available: https://awesome-workshop.github.io/awesome-htautau-analysis/

[108] C. Collaborators. (2022) Cmssw framework github. Accessed on 16/03/2022. [Online]. Available: https://github.com/cms-opendata-analyses/HiggsExample20112012

[109] T. Cherno. (2022) C++ by the cerno. Accessed on 16/03/2022. [Online]. Available: https://www.youtube.com/watch?v=18c3MTX0PK0&list=PLlrATfBNZ98dudnM48yfGUldqGD0S4FFb

[110] thenewboston. (2022) C++ programming tutorials playlists. Accessed on 16/03/2022. [Online]. Available: https://www.youtube.com/watch?v=tvC1WCdV1XU&list=PLAE85DE8440AA6B83

[111] D. Banas. (2022) Cmssw framework github. Accessed on 16/03/2022. [Online]. Available: https://www.youtube.com/watch?v=DamuE8TM3xo&list=PLGLfVvz_LVvQ9S8YSV0iDsuEU8v11yP9M

[112] C. O. Data. (2021) Windows firewall issue. Accessed on 16/03/2022. [Online]. Available: https://opendata-forum.cern.ch/t/windows-firewall-issue/68.

A Commands and Code used

This appendix includes commands and code used during the data analysis process.

Virtual Box Setup

The instructions for setting up VirtualBox were fairly easy to follow and straightforward as each step was very detailed[101]. The first step consisted of downloading VirtualBox as well as the CERN Virtual Machine Image. This image gets the CMS software (CMSSW) from /cvmfs/cms.cern.ch and the jobs running on the CMS open data Virtual Machine (VM) (Figure 68). We used the CMS VM Image version 1.5.3 which is recommended for the analysis of 2011 data[101]. It also has a large enough hard disk space storage and cache for full event range for 2012 data. It should be highlighted that the group encountered a few errors when opening the image on virtual box:

- **Error 1:** E_INVALIDARG (0x80070057)
  This error was in fact just a storage error which was easily fixed by freeing up more disk space on our computer. It should however be noted that all members of the Open Data group encountered this problem.

- **Error 2:** Could not start the machine because the following interfaces were not found: vboxnet0 (adapter 2). You can either change the machine's network settings or stop the machine.
  This was quite common within the group with 3/3 people encountering this error. It was fixed by changing the network settings from "host only adapter" to "NAT".

- **Error 3:** NS_Error_Failure (0x80004005)
  This error is quite common within versions of macOS 10.13 and 10.14. To fix the error, VirtualBox had to be uninstalled using the VirtualBox_uninstall.tool within the VirtualBox downloaded dmg and reinstalled using VirtualBox.pkg

Once these issues were fixed, we were able to access the Virtual Machine. One should note that it took around 10 minutes for the VM to load everything upon its first run, so it is important to not close/restart VirtualBox during this time. CMSSW version 5.3.32 was

needed for the analysis of 2011 CMS data. To ensure the correct version of CMSSW is running, we need to execute the following command in the terminal emulator:

```
cmsrel CMSSW_5_3_32
```

Listing 1. Command line

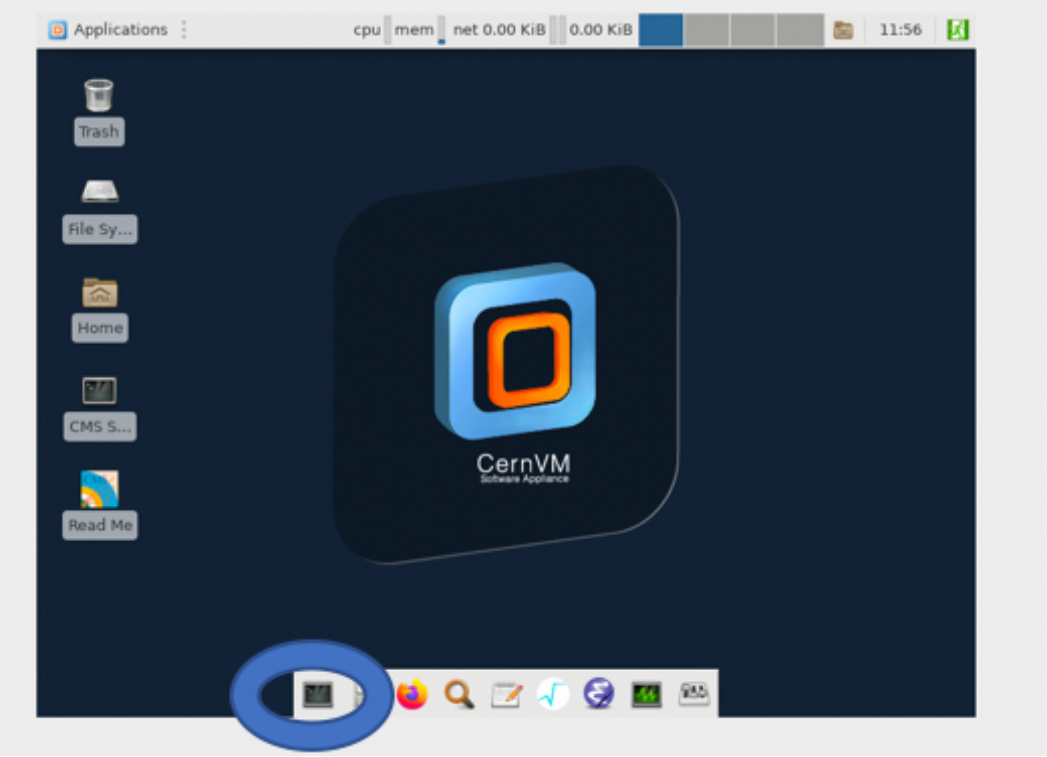


Figure 68. Screenshot of CERN VM home; The circled application is the terminal emulator where the code stated in this section was run

Then every time we run an analysis; we need to make sure we are in the CMSSW_5_3_32/src/ directory by executing the following commands:

```
cd CMSSW_5_3_32/src
cmsenv
```

Listing 2. Command line

The guide also explains how to open a CMS AOD file in Root and how to open the ROOT GUI [101]. We found this particularly helpful, as this constitutes the basis of a lot of CERN analysis, and gives a brief introduction to ROOT.

Overall, we argue that this guide is very helpful in setting up the environment. However, many issues could have been fixed much quicker. Although the Virtual Box set up instructions [101] contain a list of common problems, none of these errors were mentioned.

Docker Setup

To set up the environment on Docker, we followed the steps described in the Cern Open Data guide [102]. It should be noted that, unlike VirtualBox, the instructions contain almost no details and are only tailored to Linux/macOS users. They do not take into account Windows systems users. In addition, it is important to note that the disk space to download docker and the CMS image is extremely large, and takes up nearly 40 GBs. Therefore, one can argue that the guide provides incomplete information to set up the required environment on Docker successfully.

Our group managed to find a solution for the setup of Docker through Windows and MacOS system:

1. Install Docker through the instructions provided on the Open Data guide and the Docker website

2. Download the CMSSW image, following instructions on Cern website.

3. When Docker is run Windows WSL2 , any containers based on CentOS6 face the possibility of failure. This issue can be solved by adding a new .wslconfig file in the $< username >$ folder in users directory[102]:

```
[wsl2]
kernelCommandLine = vsyscall=emulate
```

Listing 3. Adding file

Then writting the following command to end the process command prompt:

```
wsl --shutdown
```

Listing 4. Command line

and restarting the computer to ensure the completeness of the process.

4. Download a VNC viewer to the local machine, here we use TigerVNC on windows and XQuartz on MacOS. Then we can access the GUI in TigerVNC / XQuartz with the address given in the startup message and choose our password. It opens with an xterminal of the container. If it does not open, it may be that the Windows firewall is blocking it. To enable them, go to Windows Firewall → Advanced settings → Inbound rules → New rule. Then select ports and type 5900-5902: Go Next → Allow connection → Next, then name the rule. After doing that, go to Windows Firewall → Allow an app through firewall → Change settings and mark the TigerVNC. Then it should be able to view the graphical interface[112].

5. To test, start ROOT (Figure 69) by typing `root` in the container terminal prompt. In the ROOT prompt, type `TBrowser t` to open the ROOT graphical window. If the graphical window opens you are all set and you can exit from ROOT either by choosing the "Quit Root" option from Browser menu of the TBrowser window or by typing .q in the ROOT prompt.

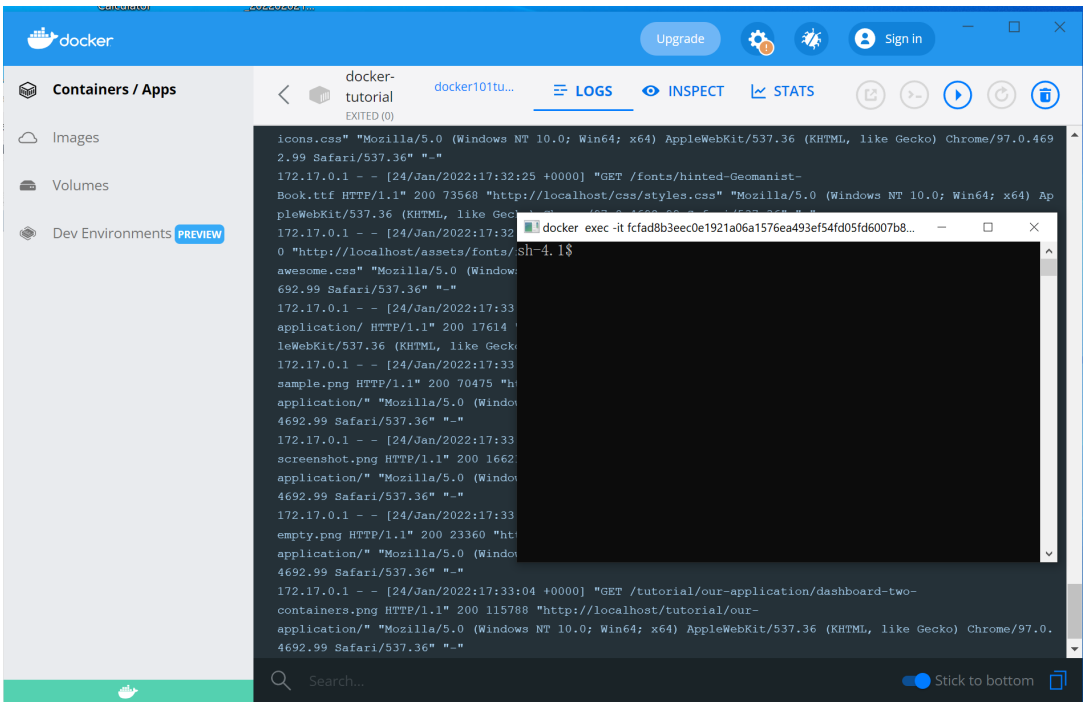


Figure 69. Screenshot of Docker Container with command line within the shell

In Docker, ROOT files can be visualised through XQuartz and the TigerVNC Program depending on different operating systems. However, the included text editor, `Vi`, is not tailored to the observation or the edition of large files. An experienced Docker user might be able to find an easy way to work around this issue and use the text editor of best suited, however, this is not feasible for someone with no prior experience in this software. In addition to this, extra plugging are necessary to ensure that Docker processes and runs data correctly. Installing these pluggings is however very tedious.

It should be noted that this container is relatively much faster and responsive compared to Virtual Machine, especially for a MacOS user. Docker would constitute a good option for a hobbyist if the instructions included in the CMS Open Data guide were more detailed and thorough.

Analysis Work Flow

This section introduces the work flow of the analysis process. We first remotely connected to the UCL HEP Cluster from a local desktop, and then set up the CMSSW image within singularity. A shared folder was also created to allow for collaboration between group members. Finally, the different files composing the analysis may be run in parallel on different computers to maximize the efficiency.

HEP Cluster

The Cluster was accessed through ssh with primary directory navigation and text manipulation. We interacted with root and histograms through Winscp, Xquartz, and Putty. This is how we connected to the cluster:

```
ssh username@plus1.hep.ucl.ac.uk
password
```

Listing 5. Command line

Then we can connect to PCS in the cluster by:

```
ssh pc<number>
```

```
password
```

Listing 6. Command line

and successfully start to work in the clusters.

Singularity

When we first tried to install the CMSSW image in singularity, we kept getting the following error: `disk quota exceeded`. We thus decided to use the pre-installed package, `CMSSW_5_3_32` to build the CMSSW environnment in singularity.

The commands are listed below:

```
cd /unix/pdpwa2/cmssw
source cmssw_5_3_32.sh
```

Listing 7. Command line

(Note: The following command operates inside the Singularity shell)

```
source /etc/profile.d/cmsset_default.sh
cmsrel CMSSW_5_3_32
cd CMSSW_5_3_32/src/
cmsenv
```

Listing 8. Command line

Then we can work on the CMS environment.

A shared folder was created to allow all the group members to access Singularity shell, edit and run the analysis:

```
singularity shell -B / cvmfs :/ cvmfs ,./ shared :/ shared
cmssw_5_3_32_latest.sif
```

Listing 9. Command line

Then we can access the shared folder with CMSSW environment in the singularity shell by:

```
cd /shared
```

Listing 10. Command line

Parallel Running on HEP Cluster

Due to the high number of files that needed to be processed for the level 4 analysis, the `nohup` command was used to shorten the time taken to run everything. It enables the cluster to run the analysis in the background parallel and remotely on different PCs, even when connection is lost from the local computer. The command we used was:

```
nohup cmsRun <filename>.py
```

Listing 11. Command line

config .py Files

There are generally four .config files and four control files for 2011/2012 experimental data and Monte Carlo simulations. The config file contains validation JSON files and input for different index files to be run. The control file includes all index files and calls the config file to run all the index files parallelly. The files below have been customized to run the analysis for 2012 experimental data by using the `cmsRun.py` command on the UCL HEP Cluster. It is an example of what we have done, and the files for the 2011 experimental data have similar approaches.

To start with, the modified config file is:

```
import sys
import FWCore.ParameterSet.Config as cms
from RecoMuon.TrackingTools.MuonServiceProxy_cff import *
import FWCore.PythonUtilities.LumiList as LumiList
import FWCore.ParameterSet.Types as CfgTypes
import FWCore.Utilities.FileUtils as FileUtils

name = sys.argv[-1]
begin = '../datasets/Data2012/'
end_txt = '.txt'
end_root = '.root'
max_events = -1

input = begin + name + end_txt
output = name + end_root
```

```
16
17
18 process = cms.Process("Demo")
19
20 # intialize MessageLogger and output report
21 process.load("FWCore.MessageLogger.MessageLogger_cfi")
22 process.MessageLogger.cerr.threshold = 'INFO'
23 process.MessageLogger.categories.append('Demo')
24 process.MessageLogger.cerr.INFO = cms.untracked.PSet(
25         limit = cms.untracked.int32(-1)
26         )
27 process.options = cms.untracked.PSet( wantSummary = cms.untracked.bool(True
       ) )
28
29 # set the maximum number of events to be processed
30 process.maxEvents = cms.untracked.PSet( input = cms.untracked.int32(
       max_events) )
31
32 # define JSON file for 2012 data
33 goodJSON = '../datasets/Data2012/Cert_190456-208686
       _8TeV_22Jan2013ReReco_Collisions12_JSON.txt'
34 myLumis = LumiList.LumiList(filename = goodJSON).getCMSSWString().split(','
       )
35
36 # use the following if you want to run over a full index file
37 files2012data = FileUtils.loadListFromFile (input)
38 process.source = cms.Source("PoolSource",
39     fileNames = cms.untracked.vstring(*files2012data
40     )
41 )
42
43 # apply JSON file
44 process.source.lumisToProcess = CfgTypes.untracked(CfgTypes.
       VLuminosityBlockRange())
45 process.source.lumisToProcess.extend(myLumis)
46
47 # number of events to be skipped (0 by default)
48 process.source.skipEvents = cms.untracked.uint32(0)
49
```

```
process.demo = cms.EDAnalyzer('HiggsDemoAnalyzerGit')

# output file name
process.TFileService = cms.Service("TFileService",
    fileName = cms.string(output))

process.p = cms.Path(process.demo)
```

Listing 12. Analysis Python code

Then we can run the control file to start the analysis:

```
import os
import sys

fileList = ['CMS_Run2012B_DoubleElectron_AOD_22Jan2013-v1_20000_file_index'
    ,
            'CMS_Run2012B_DoubleElectron_AOD_22Jan2013-v1_20001_file_index'
    ,
            'CMS_Run2012B_DoubleElectron_AOD_22Jan2013-v1_30000_file_index'
    ]

min = int(sys.argv[-2])
max = int(sys.argv[-1])+1

for i in range(min,max):
    name = fileList[i]
    nohup = str(i) + '_Data2012nohup.log'
    command = 'nohup cmsRun 2012_Data.py ' + name + ' > '+ nohup + '&'
    os.system(command)
    print("File " + str(i) + " is now running")
```

Listing 13. Analysis python code

In the HEP Cluster, the control file was run using:

```
python <control file>.py <start position> <end position>
```

Listing 14. Command line

It should be noted that we customised the start and end positions of the control files to run them in parallel on the different PCs of the UCL HEP Cluster.

B Agendas and Minutes

Meeting 0

Agenda

1pm UK, 14/01/2022 (Online)

Meeting Chair: Prof. Matthew Wing

Minute Taker: Leon Borek

Present: Prof Matthew Wing, Leon Borek, Noor-Ines Boudjema, Jiajun Chen, Dinis De Azevedo Beleza, Stefania Juks, Ashuit Khanna, Janos Revesz, Chenyu Zhang

Absent:

Meeting Agenda

The agenda for this meeting was to introduce ourselves to the board member, Prof Matthew Wing, and cover some background context for the project before setting a direction for the group collective to take before the next meeting. The meeting began with a short 'around the room' type introduction where names and courses of each present member were stated. Professor Matthew Wing spoke about his background as a particle physicist who also works in accelerator and plasma physics. It was mentioned with an accompanied slideshow that his main research is associated with the AWAKE experiment at CERN (proton-driven plasma wakefield acceleration) and other novel acceleration schemes. A key contribution of his work is to attempt to make more cost-effective acceleration schemes. He also works on the LUXE experiment investigating QED in strong fields and has previously worked in collider physics.

Project Background (presented by Prof Matthew Wing):

- The Higgs boson's significance is touched on through mentioning how it 'gives' particles their respective mass. It completes the unification of the electromagnetic and nuclear weak forces by explaining why there can exist massless and massive force carriers; it explains that in an electromagnetic force and weak force the force carriers are photons and W / Z bosons respectively.

- The Higgs boson was first postulated in the 1960s and discovered in 2012 at the LHC (it was the only collider to yield collisions of sufficient energies).
- It was problematic to find due to its unknown mass leading to difficulties determining what energies to use. It was proved to be elusive because its mass was unknown; it has high backgrounds in pp collisions which contributes to its elusivity.
- Questions concerning the analysis conducted; can the results published via the CMS data be reproduced, improved, and can any extensions be made? Is there something that can be learned from analysing ATLAS data in comparison to CMS? They are two separate experiments (almost acting as competitors) with independent measurements, which is beneficial; we have no allegiance to either experiment and thus can use the published data freely. Can we use some type of ATLAS analysis technique for CMS data?
- It is a new development for such data to be published. For example, in the past, data from the large electron positron collider (LEP) might be disk stored and only used by the scientists relevant to the experimental undertaking at that time; it might then be discarded and would not certainly be available for public usage or anyone outside of those involved in the experiment.
- The benefits to this new 'public data' development would be that other organisations or relevantly educated members of the public can attempt to reproduce / re-verify the results obtained and check for any possible improvements or extensions. Perhaps even new investigations could be made on this data if any new insights were to arise.
- In this new development, experiments do provide some of their data freely but often it is only a fraction initially, more however may be released in the future.
- Is there sufficient / significant documentation on how easy it is to do the analyses? Is this data worth being made public if extensive educational experience surrounding the subject is required to use the data in any meaningful way i.e. is it realistic for someone in the general public to be able to look for such potential improvements?

- This was related to the framework of the general public financing science and it was thus questioned if they are able to use it; is a PhD in particle physics required to comprehend such data.
- Open CMS data will be used due to its extensive documentation.

Project outline (lead by Prof Matthew Wing):

- Thorough literature review of: CMS detector, Higgs physics (proton-proton collisions), CMS and ATLAS papers, Higgs et al. papers, and LEP search papers.
- Develop an understanding regarding the CMS analysis of Higgs physics / discovery (potentially ATLAS, and even other experiments such as LEP or ILC).
- Test the CERN open data system (via the open data at CERN and accompanying guide) and an efficient / easy method to conduct certain analyses should be identified.
- It should be possible to follow guide instructions and documentation should be kept of what works and what doesn't.
- Reproduce results from the CMS publication and possibly extend this with further data. It should be considered whether this is feasible i.e. is it only possible using computers advanced beyond the average person's finance?

Early approach / issues (lead by Prof Matthew Wing):

- Importance of planning, organising, and setting a timeline was highlighted; suggested looking for interactive web tools that might aid in data analysis and proposed account setups on UCL virtual machines if the group collective isn't in possession of sufficient computing resources.
- Code should be made to run through the CMS data and extract 'important' data to perhaps run through said virtual machines or a container.
- Papers should be read in parallel to experiment with aspects of computing; suggested setting up a computing environment as soon as possible due to the possibility of technical problems arising.

- It was mentioned that knowledge of linux and possibly setting up a linux partition / emulator could be helpful to which Stefania responded that she could be capable of doing this.

- Read all of the documentation (split into groups to look at this?)

Meeting conclusions / questions:

Prof. Wing suggested that some should read a discovery paper from CMS, one from ATLAS, then compare what information and understanding was developed to come to a better culminated understanding.

Janos highlighted that the literature paper contents can be difficult to fathom.

Prof. Wing suggested not to just rely on primary sources, maybe look for Wiki or New Scientist reports to help put things in more layman terms summary of Higgs particle? Reference this, this might help with general understanding and help with a specific primary source which would help us understand it better

Textbooks to help? Lots of textbooks around Higgs might be out of date (Higgs mechanism first proposed in 1960) https://www.hep.phy.cam.ac.uk/ thomson/MPP/ModernParticlePhysics.html - powerpoint slides from textbook Prof. Wing mentioned

Prof Wing also suggested the following group split:

- One person read about Higgs physics, one person read about experimental apparatus, one person read about discovery papers (CMS and ATLAS)

- Couple people look at web-online resources, look at event pictures, obtain visual tools to help with understanding. These people interact with people looking at experimental discovery papers

- Another group, how best to do as high statistics as possible data analysis - set up a virtual machine, can we use our own computers?

- 3-2-3 split

Make a central storage space for everything where everyone can contribute and pin things for others to see - Google drive?

Next meeting with board member (Prof Matthew Wing): 12:00 (midday), 21/01/202 (Meet this time every week)

Board Member Meeting Adjourned: 13:47, 14/01/2022

Group Meeting Carried On:

Group roles decided:

- Noor - Meeting Chair
- Stefania - Communications Officer
- Leon - Minute Taker

It was decided to hold off on assigning group roles concerning the actual project undertaking, the group collectively agreed that every member should partake in some extent of reading through literature papers, reading through the publicised CMS data, and search for other potential online resources that may aid. It was decided that the specific roles concerning the project progression would be decided on Tuesday and that this would most likely follow the group assignments suggested by Wing. Further Board Member meetings planned: 12:00 (midday), 21/01/202 (Meet this time every week) (Online)

Further non-Board Member meetings planned: 12:00 (midday),Tuesday, 18/01/2022 (Online) 15:15, Thursday, 20/01/2022 (Online)

Meeting 1

Meeting Agenda

Board member meeting started: 12:05pm UK, 21/01/2022 (Online)

Present: Prof Matthew Wing, Leon Borek, Noor-Ines Boudjema, Jiajun Chen, Dinis De Azevedo Beleza, Stefania Juks, Ashuit Khanna, Janos Revesz, Chenyu Zhang **Meeting chair:** Noor-Ines Boudjema

Minute taker: Leon Borek

Objective: To reflect on the work accomplished this week and set up goals for the next board meeting. Provide a brief review of last week's work and what it entailed.

Literature review team (Leon, Dinis, and Chenyu):

- Apparatus review and explanation (Dinis)
- Review of Higgs mechanisms and physics papers (Chenyu)
- Review of Higgs mechanisms and physics papers (Chenyu)

Data analysis team (Janos, Jiajun, and Noor):

- Experience setting up VirtualBox to use opendata@CERN
- Review of CERN's instructions
- Setting up the CMS environment and running a demo analyser

Online resources team (Stefania and Ashuit):

- Research of extra resources, possible visualisations of data

Minutes

Literature review:

- Prof. Wing points out that Leon's take on every literature review member needing to collaborate closely is necessary. Someone might go off and read something in a paper that doesn't end up aiding in understanding; emphasising the necessity for 3D visual resources. Suggested that when reading a paper, even if you don't understand all of it, highlight certain issues and bring them to board meetings to discuss.

- Dinis added some subsections to his part of the literature review: first about CERN as an organisation, then about the LHC as a whole, and one about ATLAS and CMS detectors separately.

- Prof. Wing approved these subsection outlines and highlighted that the contrast between ATLAS AND CMS is important. It was suggested to do an overview of both in the final report, going into more detail on CMS because we are using their data for analysis [e.g. explain certain parts of the detector and how it relates to the analysis (use calorimeter to reconstruct particles ... use tracking detectors etc.)].

- Dinis added that it might be good to include a virtual tour on the CERN website in the report.

Data Analysis Team:

- Noor tried to set up a CMS 2011 virtual machine by following CERN's instructions. First step was to install a tool machine and then download a virtual box which is a multi-platform app to run virtual machines. An attempt was made at downloading CERN virtual machine platform; different errors kept occuring (space storage eg. errors which meant downloading more packages from the terminal - perhaps because of mac not using linux). This issue was overcome, it was highlighted that the abbreviations in the documentation are yet to be understood. The laggy nature of the box was spoken about and it was suggested that looking into using a docker might be a better option as suggested by CERN.

- Prof. Wing emphasised that it is important to choose a method early; it is not ideal to go down a certain development route before realising we are limited.

- Noor mentioned the requirement to collaborate with the literature review team - Prof. Wing pointed out that ideally the discovery paper can help you redo the experiment; whether that's true or not is yet to be discovered but in principle that is a goal of writing a paper.

- Jiajun tried to set up a CMS environment where a demo analyser is used to check whether a virtual machine works. A working directory and skeleton for the demo analyser was created and the next steps are uncertain.

- Prof. Wing said that Noor and Jiajun should compare what they did i.e. histogram panel vs data code window.

- Janos emphasised that it is hard to find resources for documentation which leads to difficulties in knowing what to search for. A paper can be found to tell us what to do but what about what it represents?

- Prof. Wing spoke about how a lot of understanding is necessary to reproduce results. A thought experiment was provided of one person attempting to reproduce results with no understanding which would prove very difficult.

Online Resources Team:

- Ashuit initially struggled but found some visualiser tools for visualising particle interactions in the detector on the open data site and aid in the form of Jupyter Notebook(s).

- Stefania talks about looking into how we could do histograms in python and suggested having discussions with the literature review team to gain a more holistic view on how a detector works by combining a paper with visual aid. It was highlighted how the jupyter notebooks could be used to plot histograms compatible with the csv / excel files on CERN website which will be looked into.

- Prof. Wing followed up by saying it would be good to see a pp event producing Higgs particles which would help a lot in understanding the Higgs physics and how the detector works. This could be a big discussion point for the next meeting.

Further Actions / Meeting Conclusions:

- Noor suggested setting goals to which Stefania followed up by suggesting having a collective discussion to try and understand some literature.

- Stefania asked that the literature review team could continuously provide understanding of what they have already and highlight areas that lack in understanding, perhaps in the form of a word document on a platform with group-wide access, which will translate into the data analysis well.

- Noor suggested trying to set up a docker as a goal. Jiajun spoke about comparing a docker to an analogue virtual machine to see which works better e.g. which has lower delay.

- Janos highlights Prof. Wing's idea that one person could try to plough through data analysis on an instructional basis to see what can be accomplished.

- Prof. Wing pointed out that this is an assessment of what analysis works and that a stage might be reached where the virtual box just stops working; there was agreement with Janos' preceding point.

- Prof. Wing requested that the minutes email him the minutes, agenda, and meeting times. (Minutes specifically as this was missed in the previous week). Prof. Wing also voiced the importance of risk analysis even though it might not seem relevant; there are risks of the virtual machine failing, a computing failure ruining a data plot, illness of a group member affecting the group, and possibility of losing data.

- A meeting to be set up in person meeting for 8 group members - Tuesday 12pm midday agreed upon.

Board member meeting adjourned: 12:40pm, 21/01/2022

Meeting 2

Agenda

Board member meeting started: 12:00pm, 28/01/2022 (Online)

Present: Prof Matthew Wing, Leon Borek, Noor-Ines Boudjema, Jiajun Chen, Dinis De Azevedo Beleza, Stefania Juks, Ashuit Khanna, Janos Revesz, Chenyu Zhang **Meeting chair**: Noor-Ines Boudjema

Minute Taker: Leon Borek

Literature review team

- Combined results of searches for the standard mode Higgs boson in pp collisions at root(s) =7TeV – CMS collaboration [21st reference of discovery paper] (Leon)
- Handbook of LHC Higgs cross sections: Inclusive Observables"- LHC Higgs Cross Section Working Group, S. Dittmaier et al. (Leon)
- Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC – The ATLAS Collaboration, G.Aad et al. (Chenyu)
- Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC – The CMS Collaboration, S.Chatrchyan et al. (Chenyu)
- CERN organization and presentation (Dinis)

Data analysis team

- CERN twiki analysis overview (Janos)
- Root documentation review (Janos/Noor)
- CMS HEP tutorial review (Janos) (http://opendata.cern.ch/record/50)
- Presentation of the data extracted from the Higgs example 2011 exercise (Janos)
- CMS Outreach Exercise and obtained results (see https://github.com/cms-opendata-analyses/OutreachExercise2011) (Noor)
- GitHub Open data resources presentation (Noor)+ Explanation on how to access the computer cluster

- Experience using Docker (Noor/Jiajun) + VirtualBox and Docker comparison (Jiajun)

Online Resources

- Experience using the CMS spreadsheet materials (https://github.com/cms-opendata-education/cms-spreadsheet-materials-multiple-lang uages/blob/master/EN-Plotting-a-histogram-with-Excel.pdf) (Ashuit)
- Overview of the extracted histograms (Ashuit)
- CMS visualizer and possible Jupyter notebooks application (Stefania)

Minutes

Literature Review

- Leon mentioned that some of the referenced papers actually seem more concise with information than the actual discovery paper itself. It was asked what is referred to by the terms 'excess' and 'background'; what does the local p-value refer to?
- Wing stated that you can take the 4 vectors of the 2 final particles produced from a Higgs decay which will produce an invariant mass resonance peak around the Higgs' mass.
- Wing went onto speak about excess and background. There will be a Higgs distribution with a resonance curve (ideally) and there will be some background of random processes that are part of the same distribution because it's not certain which particles are produced as a result of a Higgs decay. Background can be produced by anything that might appear as a signal. We are looking for an 'excess' above a background, this background might be understood through theoretical expectations (monte carlo program), perhaps the background is fit via a function and any resonance that doesn't fit is identified. Statistical variation i.e. small 'bumps' in the background shouldn't always be taken as significant as there are statistical probabilities of background producing said peaks, which is why the deviation from said background is a critical point for analysis, generally in particle physics for an event to be considered significant it must lie about 5 standard deviations from the background .

- Leon asked whether the reason for different peaks for different decays is down to how common certain decays are. Wing went on to describe how the probabilities for the Higgs to decay to certain final state couples differ thus result in differing peak significances per decay - depending on the particles we have the measurements can be more accurate e.g. the most common Higgs decay is to a B quark and an Anti-B quark. However, this is a difficult measurement to make because 2 jets must be identified in which both contain a B and Anti-B quark respectively, the jets could be combined to obtain a Higgs - these kinds of jets are produced millions more times than the Higgs in LHC collisions thus we obtain lots of random background that can mimic a Higgs

  particle. This makes reproducing said decay event very difficult. On the other hand, a Higgs decay to 2 photons has a much lower branching ratio (produced far less frequently) but it is much easier to reconstruct and the background is lower, looking for 2 photons is a much cleaner signal in a particle detector since it is electromagnetic, a B quark will end up as a bunch of hadrons in a jet whereas photons don't decay further hence this was seen as the golden channel to obtain the Higgs; the contributors to accuracy of measurement here are ease of reconstruction and branching ratios, in the experiment multiple possible Higgs events were combined to obtain the best results.

- Leon enquired about a table that encompassed background events in the CMS discovery paper. Wing mentioned that because it was so difficult to find, they were using every possible channel that could exist, even if it contributed a tiny amount. This was based on Table 3 of 'Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC' referring to the background events. Wing highlighted that you're looking for a particular signal event; the background from a Higgs to tautau could come from a Z to tautau or QCD for example, a W being produced alongside some jets could also mimic a Higgs signal. It's important to think of all possible physics processes that could mimic Higgs, this is what is displayed. Leon spoke about perhaps writing a summary of what has been read in literature review form, perhaps a final version summarising the discovery paper for the final report.

- Chenyu had similar queries to Leon about the local background events and p-values measured in the experimental procedure within the CMS discovery paper which had been answered by Prof Wing.

- Dinis wrote his first subsection about CERN - why and when it started. The first idea originated in WW2, lots of top scientists were fleeing Europe due to war; creating this facility incentivised said scientists to return. Trying to get countries to unite in creating a facility like this would allow for resource sharing due to how expensive it was to build. Starting a project like this allowed for unfriendly nations to collaborate and break down conflicts by incentivising collaboration. Geneva was picked as the location due to Switzerland being a neutral state in war (eliminating war biases to an extent) and is relatively centralised in Europe allowing for relatively equal ease of access from around Europe. Previous machines built at the facility had been mentioned.

- Prof. Wing suggested that the literature review team should look into the actual theory of the Higgs alongside researching experimental discovery procedure i.e. go back and read the original Higgs paper proposing the existence of said boson.

Data Analysis team:

- Janos primarily looked at online resources such as documentation for 'root' framework introducing 'root' i.e. plotting / histograms / analysis etc. for the code being used. Another resource Janos looked into was the CERN CMS offline workbook, it was described as a slightly less detailed documentation of the whole CMS software package that they use. He did this due to issues encountered when running the GitHub project they were following and working through different levels of said code. The CMS documentation being used in the code however appeared out of date and thus either we have to go back and use the software package they used or modernise the code.

- His main question was about how the events are reconstructed from the data (displayed in other data formats) - questions about monte carlo simulations and how it is utilised i.e. what is done with the data and the simulated data? How does the simulation add to the original data?

- Prof. Wing spoke how when events occur, particles are measured in the detector that measure energy / momentum / position. There is no knowledge however of what occurs between the particle's original position and the point at which it is detected, e.g. a particle's energy measured as 5GeV doesn't account for any potential energy losses through collisions in the detector before the point of detection. Also for example a pion or electron might appear exactly the same based on energy measurements; in order to convert the signals to particle identification, a monte carlo simulation is utilised as it describes what happens when the particles collide. A subsequent simulation of the entire detector is then used that can account for these collisions and how particle energy deposits / measurements might change based on where the particle is identified in the detector. The simulations work to unfold energy deposits / scattering angles and 'work backwards' towards particle identification, they help understand efficiencies and help understand how well the data reconstructs the particles being observed.

- Janos went on to point out that there is an incredible amount of resources, how much depth should this be gone into in the final report based on computing documentation?

- Prof. Wing said to summarise this but to focus on the physical principles used, in the end this is a physics project that discusses physics processes and at the end a data plot will be shown to display the reproduction of Higgs discovery data i.e. we don't want to be describing tons of code. Provide enough to give people an idea, but mainly focus on physical principles used, should also discuss the computing aspects referring to how feasible it is for members of the public to use this data. Include a virtual box / docker discussion in a separate chapter which doesn't disturb flow of physics analysis. Perhaps use a workflow diagram describing the computing aspects, but keep it to its own chapter (use an appendix?), include the changes that have to be made i.e. adapting GitHub code as it is a large part of the process.

- Noor found Jupyter Notebooks from CERN open data, a lot of them have analysis of different Higgs decays / different scenarios the Higgs analysis can be used for. She thought it would be a good idea to go through these and extract data. Her and Janos found a CMS outreach exercise that teaches how to use the CMS environment, there

was a 4 lepton decay analysis, there was trouble running the code, 1 event could be run and data extracted. She tried to access the computer cluster (UCL HEP). Could a docker be compiled into this? She is able to log into the cluster but doesn't know where to go from here.

- Janos added how the computing runs very differently based on the computer used, Noor has had a lot of trouble running the docker etc. (computer crashes).

- Jiajun noted how the docker is more complicated to install compared to the virtual machine since extra commands are needed in addition to online instructions in order to direct the docker. Certain programs need to be downloaded for windows and mac respectively to view the code interface. The virtual box is much more self contained and only one installation is required, it has a very high delay however. While the docker command line and applications can run much faster, the user needs to be able to set up an independent docker environment to get the graphical interface to work and a text editor needs to be set up for use. There are minimal guidelines for this from CMS. He thinks the virtual machine would be much more convenient but the docker may be a better choice for faster analysis, but is this worth the set up time and complications setting up environments? On the UCL HEP cluster he looked at the STP client which interfaces with GUI, this could be helpful for visualising and working with 'root'. He noted there is a singularity docker on the cluster which contains a lot of CERN virtual machine libraries with data; it could be helpful to analyse this data in the future.

- Prof. Wing said to document how Jiajun went about installing a docker on his computer, resources from google etc. Perhaps other Data Analysis people could follow said documentation.

Online Resources Team

- Ashuit was following guides on the CERN website on how to do invariant mass histograms. He could follow some sample data provided and is continuing attempts to do this with different data sets also.

- Stefania found documentation for what all of the buttons mean on the CMS visualiser and found instructions on how to use ECAL / HCAL (calorimeters) for a pp collision,

she found an example using 2 muons and tried to replicate this to adapt it for a pp collision which she is currently working on. She found a lot of CMS documentation which spoke about the visualiser and VR devices where virtual tours of the detectors can be taken. A browser data analyzer was found, recommended by CERN, it was looked into how analysis might be done using this. Jupyter Notebooks were found which describe how to find the invariant mass from collisions and mentioned that if Ashuit can find the relevant data sets, they can be looked at using said resources.

- Prof. Wing said to try and get a visualisation for any events possible as these could be points for discussion in the following meeting.

Further Actions/Meeting Conclusions:

- Leon suggested getting some Final Report subsections laid out from a Literature Review perspective; this prevents mindless reading and sets out a framework for identifying the most important parts of scientific literature review to be used and referenced for the project.

- Stefania mentioned the risk assessment and a timeframe on its completion, Wing said whenever this can be completed it can be reviewed in a Board Meeting.

Board Member Meeting Adjourned: 12:56pm, 28/01/2022

Meeting 3

Agenda

Board member meeting started: 12:00pm, 04/02/2022 (Online)

Present: Prof Matthew Wing (MW), Leon Borek (LB), Noor-Ines Boudjema (NB), Jiajun Chen(JC), Dinis De Azevedo Beleza (DDAB), Stefania Juks (SJ), Ashuit Khanna (AK), Janos Revesz(JR), Chenyu Zhang (CZ) **Meeting chair**: Noor-Ines Boudjema **Minute Taker**: Leon Borek

Objectives: To reflect on the work accomplished this week and set up goals for the next board meeting. To discuss problems encountered and find solutions to fix them.

Literature review team

- Presentation of the LHC (Dinis)
- Review of Short but Sweet – Peter Higgs (Leon)
- Review of The Higgs mechanism – Hasse diagrams for symplectic singularities, first author: A. Bourget (Chenyu)

Data Analysis team:

- Level 2 analysis of the Higgs exercise via the HEP cluster, Level 3 analysis of the Higgs exercise via virtualbox and presentation of results, Level 4 overview and problems encountered (Janos and Jiajun)
- Setting up the CMS environment on the HEP cluster using singularity and problems encountered (Jiajun)
- Presentation of the Awesome Workshop on Higgs example analysis (Noor)

Online Resources Team:

- Presentation of histograms for various datasets on spreadsheet and Jupyter (Ashuit)
- CMS visualisations of different decays, Jupyter notebook measurement's calculations (Stef)

Minutes

Literature Review Team:

- DDAB presented his research on the particle accelerator mechanics, specifically the particle path up until collison and how charged particles are accelerated to velocities close to 'c'; EM fields in the LHC are used to accelerate said particles (in this case Hadrons) for head-on collisions (radiofrequency cavities and insertion magnets were talked about). The increase of the LHC's potential energy capabilities over time were talked about along with the cost of construction, the several detectors used as preliminary accelerators before final collisions, and the 4 detectors. A CERN animation was displayed to illustrate this.

- LB went through a series of plots found on "'Handbook of LHC Higgs cross sections: Inclusive Observables' paper - LHC Higgs Cross Section Working Group, S. Dittmaeier". The "'Short and Sweet' paper - Peter Higgs". MW highlighted the importance of the branching ratio plot as it explains why all the different channels are used; he emphasised researching terms such as: symmetry breaking, standard model Lagrangian, and vacuum expectation values. He recommended sticking to articles from New Scientist or Scientific American for example to become more familiar with the topic which would later make more intellectually challenging resources more accessible.

- CZ spoke about the significance of the standard model and outlined it with reference to its function as a particle classifier. The function of the Higgs particle (providing mass to W and Z bosons) was talked about and compared to the traditional considerations for massless bosons; the reason for the Higgs' classification as a boson was mentioned. Questions were raised about electroweak symmetry breaking; what are the states for symmetrical / unsymmetrical cases and what happens in the symmetry breaking process. MW stated to describe the standard model in the report and similarly recommended reading through simpler explanation of symmetry breaking and then to explain it to the group in the next meeting. Higgs' formulation of the mechanism in the 60s was talked about from a previous meeting with the main takeaway being to read more popularised science.

Data Analysis Team:

- JR spoke about doing the level 2 and 3 Higgs example and showed plots from both of these and spoke about the differences in the data generation despite almost identical plots being obtained. The code for this was shown with file types talked about and many references to the significance of 'root' here being made. The data processing is repeated for simulated and real data. On the histograms, data points represent real data and the histogram bars represent the monte carlo simulated data. Level 2 was able to run on the cluster but this was not the case for level 3 and thus had to run on the virtual box. Difficulties encountered with respect to data storage and processing power for the level 4 simulation were described and thus asked if there was an available CPU / GPU cluster since this couldn't be done on the base HEP cluster. MW suggested trying to use the HEP cluster bath farm to run jobs on and asked whether a small sample of the data could be used; JR stated that there were still errors with this for level 3 relating to FW packages. JC also mentioned trying to use singularity to install relevant level 3 software but even this was out of memory range on the cluster.

- JC spoke about 'Quota limit exceeded' errors that were encountered when downloading the CMS environment from his home directory; PTA was used to create a new directory but this didn't work due to disk space. MW highlighted that this is already installed on PDWA2 as CMS SW and suggested trying this; if this didn't work he suggested contacting himself to check why the quota is exceeded which may be necessary if the software already on the PDWA2 might be out of date.

- JR had a question about how this data is obtained from the events we analyse. MW replied by saying that in general, the distribution possesses events data points. There will be a certain amount of data taken that represents millions of events, these events are narrowed down to isolate ones that look like a Higgs event and hence requirements / conditions have to be made of the events to be considered. Data points are made from combining invariant masses of same events, hence reducing many events to one data point. The requirements are set to make clean data / well measured particles / so that the particles combine to give a particular invariant mass. He mentioned it

could be useful

to talk about data and code correspondences within the report. For example, if a plot shows 4 leptons, we are looking at 4 leptons that we know a Higgs could decay to; there won't be many other things that form a 4 lepton decay invariant mass around this Higgs region. It is not known event by event whether a Higgs is present, only by integrating over a huge number of events can this be reconstructed and a statistical test can be run to determine whether this is significantly different to the background. JR asked whether the code should be skimmed to try and find correspondences to which MW suggested playing with the requirements in the code e.g. a requirement of an electron to be considered might be an energy threshold of 4GeV, what if this was changed to 5? The signal over the background should have as large a difference as possible.

- NIB spoke about a different data analysis route taken using a workshop with training modules. The production of the Higgs is called the signal with main production modes being via gluon fusion and vector boson fusion. The Higgs boson decays into 2 taus and 2 leptons has a very short lifetime and is thus very hard to observe. The most prominent background processes for this are Z ¿ tau + tau; other things also interfere and can be misidentified such as W + Jets and top / anti-top pairs. For all said processes, the workshop equips us with different simulations that simulate different decays all of which are simulated via a computational approach. QCD events are too computationally expensive to conduct simulations for, instead the actual data is looked at. The quota error was also experienced when trying to run this data through setting up an environment. This issue was overcome and the data is currently being skimmed; preprocessing events reduces the size of the data sets significantly by only selecting relevant events. Difficulties are encountered due to the code being written in C++ (memory management sake). File conditions were touched on and how this relates to the conditions for selecting a muon pair / tau pair for example. Problems were encountered because this workshop was set up for people with CERN accounts primarily, this root version had to be downloaded from a link and then ran, HEP runs on the 6.2 version of root which didn't work here. There were explanations on how to use this with docker but this did not work.

When NIB asked about documentation resources for histogram abbreviations, MW highlighted that there is no globally accepted / commonly used documentation for this data, it was mentioned that usually PT represents transverse momentum and JPT means jet transverse momentum. PT is normally transverse momentum. Noor - look for documentation, do you know of any websites with documentation for these abbreviations for histograms. Usually JPT means jet pt.

Online Resources Team:

- AK looked at data sets on the open data websites though it is still difficult to determine which of these are relevant. A few of these were tried using a CERN spreadsheet guide to create sample histograms which relate to the invariant masses. The same was tried for Jupyter notebooks and similar histograms were obtained with peaks representing real particle invariant masses. A number of these were displayed and MW spoke about how

  these related to the method by which new particles are identified through finding invariant masses. Number events against invariant mass, posit. A histogram of the invariant masses of two muons displayed the reproduction of the J-psi and J-prime particle identifications. These principles of particle combinations are exactly the same principles as those used in identifying the Higgs and can be included in the report. Perhaps look for prominent Higgs decays or significant Higgs background events.

- SJ provided a visualisation of a particle collision which was discussed and spoke about how it was possible to display 4 lepton decays or 2 photon decays. It was mentioned how data for individual particles can be extracted by clicking on their detection point and how parts of the visualiser can be removed / included to isolate individual aspects of the collision. There were questions about how to know what should be displayed based on what is necessary, how much detail should be included and what does preshower mean? MW described 'preshower' by stating that it means that there is a big calorimeter and the particle energies are measured in that; the particle has travelled quite a distance thus there is a preshower detector to give some extra information on particles going through the detector and hence more information is provided and there is a better chance of reconstructing the original particles. SJ

mentioned that a description of the interface will be laid out, possibly to include in the report. VISPA was looked into and it was found to be able to run root, but only with reduced data sets. Perhaps some of JR's code in the browser could be run on this. Online CERN related activities were found but this was only possible using MAC, perhaps someone using Mac OS could try this. MW mentioned that there should be a description of the detectors and what parts are utilised in the report. Important information about particle properties and how they interact with matter can be extracted. For example, muons can just reach the muon chamber and go through the rest of the detector without much interaction, electrons can go through the tracking detectors and deposit energy in the calorimeter but will not get through to the muon chambers, photons will only be measured in calorimeters etc., these points are important.

Further Actions / Meeting Conclusions:

- Next board member meeting to be conducted online at 12:00pm, Wednesday 09/02/2022.

Board Member Meeting Adjourned: 13:21, 04/02/2022

Meeting 4

Agenda

Board member meeting started: 12:00pm, 04/02/2022 (Online)

Present: Prof Matthew Wing (MW), Noor-Ines Boudjema (NB), Jiajun Chen(JC), Dinis De Azevedo Beleza (DDAB), Stefania Juks (SJ), Ashuit Khanna (AK), Janos Revesz(JR), Chenyu Zhang (CZ)

Absent: Leon Borek - was ill at the time of the meeting and could not attend

Meeting chair: Noor-Ines Boudjema

Minute Taker: Dinis de Azevedo Beleza

Objectives: To reflect on the work accomplished this week and set up goals for the next board meeting. To discuss problems encountered and find solutions to fix them. To better understand the Higgs symmetry breaking.

Literature review team

- Overview of detectors (Dinis)
- Higgs symmetry breaking (Chenyu)
- Review of Natural electroweak symmetry breaking from scale invariant Higgs mechanism, A. Farzinnia et al. (Chenyu)

Data analysis team

- Presentation of the di-tau graph obtained from the HSF Workshop Higgs analysis and problems encountered (Noor)
- Setting up the level 4 Higgs analysis on VirtualBox (Janos, Jiajun)
- Setting up the CMS environment and running the level 3 analysis on the cluster (Jiajun)

Online resources team

- CMS visualizer report and explanation of the detector (Stefania)

Minutes

Literature Review Team:

- - DDAB presented his research on the ATLAS and CMS detectors involved in the finding of the Higgs boson. Both have the same goals but use different technical solutions and different magnet-system. Then a description on the main components of the detectors was made, particularly a description on how the silicon tracker identifies the path of the particles, a description on how the electromagnetic calorimeter, hadronic calorimeter and muon chambers measure the energy and momenta of the particles. A final description on the importance of the detector's magnet was made. Finally, the formats and intensities of the ATLAS and CMS magnets were mentioned together with the dimensions of the detectors. NIB mentioned that SJ was also starting to write about the detectors so it would be a good idea for both to discuss what was found and figure out what to write so that an overlap of information is not made.

- CZ started mentioning that the elementary particles interact forming a coupling with the Higgs field and acquiring their mass. An image was shown in order to help illustrate and explain this concept. It was explained that the particle will be stable in the lowest order state but we cannot now the exact position until measuring it. When measuring the position, the system will stop being symmetric and symmetry breaking will occur. A description of the Higgs mechanism on how the variation in energy will give particles their mass was also made. Finally, it was explained that this Higgs mechanism is only important for the bosons and not the other particles. MW mentioned that this topic is really hard to explain in such a short time but some very nice points were said in the explanation. Also said that one of the main challenges for the final report will be to summarize this topic and the rest of the background into a limited number of pages.

Data Analysis Team:

- NIB started out explaining that she spent most of this week compiling and reorganizing the histograms shown in the week before. The visible mass histogram was shown to us. It was mentioned that this histogram is interesting since we can see the differences in the visible mass to the real mass expected. A possible reason for this

difference was given to be the smallest particles like neutrinos that cannot be seen so their masses is

not included. It was explained that some problems were encountered when running the code which made this task longer and harder to do. MW asked if the data shown in the histograms was only for the di-taus. NIB answered that this is one of the things that she is still trying to completely understand but as far as she could find in that moment it looked like that was the case. It was also discussed whether this approach would be useful to write in the report as an alternative to the other approach being done by the data analysis team.

- JC started explaining that he tried using the pre-installed package for the CMS visualizer in the UCL data cluster and this attempt was successful. He also mentioned that a lot of time was lost trying solve errors when setting up the environment in this process. The 'quota limit' error from the week before happened again this week so the older package (the pre-installed) was the one used to run the level three test. Run this test consumes around 2-3 minutes. Finally, it was mentioned that an attempt at the level four test was going to be done in the week after this meeting. MW mentioned that it would be important to know the percentage of data used in the three-level test of the full CMS database in order to understand how long it would take to analyze the full data instead of just a small part.

- JR explained that he had been trying to make the singularity container over the CMS environment on the cluster. He highlighted that despite trying for a lot of time to get this, it still is not working for him even though JC has managed to run this part. He mentioned that he will try to get this done over the week after the meeting. Finally, he also mentioned that he was trying to do the level-four test. MW mentioned that it would be a good idea to swap notes between JR and JC since they are doing the same part and can help each other out. NIB added that it would be a good idea to get the different level tests in the final report because they help identify properties of the Higgs.

Online Resources Team:

- SJ said that in the week before the meeting she started writing about the CMS visualizer for the formal report. In parallel, she also started writing about the detector in order to get a better understanding of that section. She mentioned also that the problems encountered in the Jupyter notebooks the week before were solved and she could also run them on Windows. MW highlighted that there is clear overlap between DDAB and SJ detector parts but since one is reading about the detectors in a technical perspective and the other in an animated perspective there is room to combine both or follow on one from the other.

Further Actions / Meeting Conclusions:

- It was mentioned that the Report will probably be written using Latex. Also it was added that most of the people wrote in their midterm review that there was a lack of communication between the group and that it should be dealt with and improved for the second half of the group project.

- Next board member meeting to be conducted online at 12:00pm, Friday 18/02/2022.

Board Member Meeting Adjourned: 12:41, 09/02/2022

Meeting 5

Agenda

Board member meeting started: 12:00pm, 18/02/2022 (Online)

Present: Prof Matthew Wing (MW), Leon Borek (LB), Noor-Ines Boudjema (NB), Jiajun Chen(JC), Dinis De Azevedo Beleza (DDAB), Stefania Juks (SJ), Ashuit Khanna (AK), Janos Revesz(JR), Chenyu Zhang (CZ)

Meeting chair: Noor-Ines Boudjema

Minute Taker: Leon Borek

Objectives: To reflect on the work accomplished this week and set up goals for the next board meeting. To discuss problems encountered and find solutions to fix them. To better understand the Higgs symmetry breaking.

Brief introduction to the proposed plan for the report (Noor) **Literature review team**

- Overview of the proposed subsections on the detectors (Dinis)
- Discuss new possible subsections for the experimental apparatus
- Discuss sections regarding the Higgs discovery and Higgs physics (Leon and Chenyu)
- The Standard Model: How far can it go and how can we tell? First author: J.M. Butterworth (Chenyu)

Data Analysis team:

- Running level 4 tests of the Higgs analysis on the UCL cluster (Jiajun)
- Partial results from the level 4 analysis on virtualbox (Janos)
- Making a measurement and determining uncertainties on the Example analysis for the awesome workshop (Noor)
- Discuss proposed subsections on data analysis for the report (Jiajun, Noor, Janos)

Online resources team:

- Discussion on the CMS visualizer section of the report (Stefania)
- Python Jupyter notebook analysis (Stefania)

- Questions to Prof. Wing on the CMS open data (Stefania)
- Presentation of histograms obtained via a diphoton dataset (Ashuit)

Minutes

Literature Review Team:

- LB planned to talk about CMS specifically and wanted to discuss the ordering of certain discussions in the theory introduction / explanation of particle accelerator facilities and experimental methods.
- DDAB suggested first including the section describing the apparatus and then moving onto the actual discussion of how CMS discovered the Higgs once all of the relevant terminology has been covered in said previous section.
- MW made a point of mentioning previous searches for Higgs and why they couldn't yield a Higgs boson due to energy incapabilities.
- LB spoke about putting the Higgs physics section before discussion of the CERN and particle detector facilities; discussion should be had between CZ and LB about organising the compartments of said section .

Online Resources Team:

- NIB talked about these sections to be followed by CMS open data, small section explaining what it is and who it is for, then go through more accessible CMS data analysis e.g. Jupyter Notebooks / Visualiser.
- SJ stated to start with visualiser, use it to compare with what was done in the actual experiment, describe everything that's been done in comparison to the experiment and what is being addressed and to then go through Jupyter Notebooks provided by CERN (including a discussion of invariant masses). Notebooks highlighting the importance of utilised data points could be discussed and how much data can be used in a histogram to display the invariant masses. Attempts are still being made to get the 'physics playground' working and 'VISPA' still needs to be looked into further.

- SJ has encountered an issue with exporting the visualiser images but is sure that this will be sorted soon; Jupyter Notebooks from CERN about pseudorapidity are being worked on for use. How was pseudorapidity considered in experiments, were non-centred collisions weighted less? MW spoke about how events can be accepted if they are of good quality and well measured, even if the collision does not occur at the centre of the beam pipe. If the collision is off centre it is more likely to be poorly measured as some information will be lost. Events can occur all along the beam pipe, the best quality measurements will occur at the centre however because the detector is

  created symmetrically about this point.

- AK displayed a 2 photon invariant mass histogram, there were questions about how to use more data points as there didn't seem to be much data displayed. SJ went on to show some invariant mass plots for Z ¿ 2 muons. All of these notebooks should be discussed about what can be done with these notebooks, it would be ideal to see if more data could be used for the Higgs plot. MW said to include plots of the dimuon mass over a large range; these are very important to include as it shows how the mass reconstruction works through identifying resonant peaks. The x-axis shows the logarithm of the mass. The plot displays Zo at 2, B mesons at 1, and Jpsi / Jprime mesons at 0.5. The background comes from identifying particles that appear similar to muons or from identifying muons from random sources that are not the decays displayed. There was a 5 sigma significance when finding the Higgs, meaning that it is very certain or very unlikely to be a random fluctuation; there have however been particles discovered with 7 sigma and these experiments couldn't be reproduced, for example with pentaquarks. With the Higgs, this significance has only been strengthened through re-testing i.e. through increasing the channels used to increase the significance (branching ratios).

Data Analysis Team:

- NIB talked about introducing packages and software that has been used (e.g. CMSSW) with how they all function and fit together in analysing the CMS open data. Talk about 'ROOT' in relation to CERN data analysis, there will be a discussion of how

data is stored and generated at CERN (discuss skimming / monte carlo simulations). Explain how the environment was set up through a docker or virtual machine; how much detail should be gone into for this given how much code there was to set this up, how much code should be included in the written section?

- MW stated that it would be a good idea to reserve a specific section to discuss technical aspects and coding, don't mix coding with written sections. This would make it easier for report readers to identify specific sections in which are of particular interest for them and they do not have to read every section to 'fish' for information relevant to them. Always make it clear what each section is trying to do. Might be too much code to include all of it in the appendix; LB suggested that perhaps uploading to GitHub and then including a link would be more sensible. MW agreed with LB that monte carlo simulations should be first introduced in the first sections alongside descriptions of detector facilities.

- JC is going to compare virtual machine and docker.

- NIB will go into some detail on the 'awesome workshop' analysis that was done, perhaps it would be useful to include a recap of some Higgs or background decays that have been analysed through this. MW talked about that section 3 will turn out to be quite extensive with some physics discussion with physics results of a plot showing the invariant mass of the Higgs for example.

- MW highlighted the importance of suggesting improvements for the open data system and including critique; the project isn't about whether a PhD particle physicist can conduct this analysis, but whether a general, most likely reasonably well versed in physics, member of the public can conduct it. It was mentioned that the length of the report shouldn't be of large concern unless it is arbitrarily very long.

- JR is trying to run the first index of double electron 2011 A; the running was taking longer than 18 hours to run this first index without being completed, JC asked a question about whether or not this was normal. JR suggested limiting the number of events or running on multiple PCs; MW spoke about using a 'nohup' command that will prevent the running from halting after logging out.

- NIB presented a plot displaying a logarithmic Z ¿ Tau Tau decay, an important Higgs background decay from the ‘awesome workshop’.

Further Actions/meeting conclusions:

- Schedule the next meeting next friday at the same time
- MW suggested asking more questions throughout each meeting.

Board Member Meeting Adjourned: 13:53, 18/02/2022

Meeting 6

Agenda

Board member meeting started: 12:00pm, 25/02/2022 (Online)

Present: Prof Matthew Wing (MW), Leon Borek (LB), Noor-Ines Boudjema (NB), Jiajun Chen(JC), Dinis De Azevedo Beleza (DDAB), Stefania Juks (SJ), Janos Revesz(JR), Chenyu Zhang (CZ)

Absent: Ashuit Khanna

Meeting chair: Noor-Ines Boudjema

Minute Taker: Leon Borek

Objectives: To reflect on the work accomplished this week and set up goals for the next board meeting. To discuss problems encountered and find solutions to fix them. To better understand the Higgs symmetry breaking.

Meeting Objectives: To check up on the progress made with the report. To solve arising problems within the OpenData analysis groups.

Literature review team:

- Prior experiments overview (Dinis)
- Review of the Standard Model section of the report
- Overview of the report section on Higgs decays (Leon)
- Review of the Standard Model section of the report and review of The Standard Model: How far can it go and how can we tell?, J.M. Butterworth (Chenyu)

Data Analysis team

- Running level 4 tests of the Higgs analysis on the UCL cluster (Janos, Jiajun)
- Discussing results (Janos, Jiajun)
- Monte Carlo simulations and data analysis at CERN (Noor)

Online resources team,

- Presentation of a possible tutorial on CMS OpenData which could be included to the report

Minutes

Literature review team:

- DDAB spoke about the infrastructure of the LEP, which was in place of where the LHC now is and how the construction of the LEP was the largest civil engineering project before the channel started. The initial LEP energy was 91GeV so that Z bosons could be produced and investigated; it operated for 7 years around 100GeV and produced approximately 17 million Z bosons. The LEP was eventually upgraded with superconducting cavities to investigate W bosons; it reached energies of 209GeV in 2000. This was not enough to find the Higgs given that 91GeV were taken for the production of a Z boson which only leaves approximately 118GeV for the Higgs production which is not sufficient. He went on to mention that one of the most important findings from the LEP was that there are 3 and only 3 generations of matter.

- CZ gave a background overview of the Standard Model with the help of a slideshow accompaniment, explaining what it is and how it is related to the Higgs boson. It is a theory to describe 3 out of 4 fundamental forces, namely: strong, weak, and electromagnetic interactions, excluding gravity. Gravity governs mass interactions,

  electromagnetism governs charge interactions, weak interactions are responsible for things such as radioactive decay and nuclear fission / fusion, strong interactions are responsible for describing how quarks combine. It was highlighted that the 2 branches of elementary particles are fermions, with half integer spin, and bosons, with integer spin. Bosons are classified as force carrier particles.

- CZ displayed a timeline of notable progressions for the Standard Model, in particular: extending gauge theory to non-abelian groups (to describe strong interactions), the demonstration that parity wasn't conserved in weak interactions, the emergence of electroweak theory, the incorporation of the Higgs mechanism into the electroweak interaction, the discovery of neutral weak currents caused by Z boson exchange at CERN, experimental confirmation that hadrons are composed of fractionally charged quarks, and the first coining of the term 'Standard Model'. A chart displaying differ-

ent particle discoveries for different years was then shown with the exception of the gluon which was discovered in 1979.

- LB spoke about his write-up on how particles are detected; all that can be done is isolate as many particles of interest, determined by which decay channels are being investigated, and take an invariant mass combination of all possible decay products that include said particles. These particles are arbitrarily chosen based on some requirements and thus there is no guarantee that they originate from the same interaction or decay. As more data is gathered, a 'background' level of invariant masses will be produced as a result of particles that have been falsely assigned as contributors to the same decay i.e. they will not tend to some invariant mass value of a pre-decay particle. There will be a statistical tendency for particles that have been correctly attributed to the same decay to yield an invariant mass of a particle such as a Higgs or Z boson. 'Particle flow' event algorithm was the event reconstruction method used in the CMS. A Higgs branching ratio pie chart displaying different cross sections was displayed which was followed by a discussion of why certain decay channels are investigated over others. Resources on the 4 Higgs production mechanisms were also shown with brief descriptions of each.

- SJ had a query about the nature of gauge theory to which MW explained that in a gauge theory, a transformation such as a movement of a reference frame can be changed without a change of physics. For example, there might be angular symmetry, gauge theories deal with symmetries or invariance under transformations.

Data Analysis Team:

- JC and JR are running background tests, 2 tests have been run thus far with more currently happening in the background. JR specifically tried to run the analyses on multiple PCs for the HEP cluster using 'nohup', but the output files are too large to store. Could a different location to store these files be found? This couldn't be done with a batch file but it is being attempted using python. There has been trouble with some index

  files; sometimes when connecting to servers to download data files it does not work. Their section of the report and perhaps the level 4 analysis could be prospectively

finished by next week. MW suggested trying to reduce the size of the output files since there is going to be a lot of information that won't be used; is there a way that it can be compressed? If it is already very efficiently stored then zipping the file might not save much, but it is worth trying.

- NIB has spent the week investigating how CERN generates data and how monte carlo simulations work. A diagram showing event reconstruction and simulation with respect to particle physics was displayed. Initial and final states are generated through pp collisions, they rely on theoretical calculations and experimental inputs. The file format used to store information easily; it is adaptable for various programming languages. The detector simulation shows the interaction between the generative particles inside the particles, the detector response will be obtained from the particle interactions (GN4). Simulated event reconstructions are then compared to the real data and then analysed in a custom software. There was some uncertainty as to what was meant by a 'trigger' with reference to this diagram and the simulation process. MW brought up a web page that showed a trigger challenge. Depending on the CoM energy on the x-axis, there is a certain cross section for certain event categories. Most often in pp collisions there will just be some production of low energy particles; exotic particles may be produced such as a bottom/antibottom pair but the cross sections of these are significantly higher than that of the Higgs production. There is too much data to investigate every single pp event. Requirements must thus be set to limit these for example by keeping all high energy leptons or photons for example as these can be indicative of a Higgs event. These act as triggers that keep certain classes of interesting events and reduce the number of events for investigation by several orders of magnitude.

- MW highlighted that NIB's essential overview of how an analysis is done should be introduced at the start of the report. LB and NIB discussed how this might be implemented, for example perhaps a short overview of how these simulations relate to the particle physics process and 'cookie-cutter' explaining how it works with more detail being gone into later when analysing data?

- JR had the idea of perhaps adding a small tutorial at the end of the report on how to

use the CMS open data, but thought that this also depends on the group's collective conclusion. In his personal opinion the CMS open data analysis process is not easy to follow; if the rest of the group feels the same then a tutorial could be very useful to provide a better overview of what was done to analyse the data. This could be left to the end however since this draws on any conclusions we might have. This tutorial could combine how the online resources and data analysis teams went about finding the information and means of data display that they found. MW added that it could be a good exercise to create this tutorial and have a literature review team member go through it to see if it can be utilised by someone else. A big problem with the CMS analysis is that there is not an easy step-by-step flow to the instructions. JR went on to

add that some of the guides might have to be redone due to the ones used not containing enough detail, though this could be very time consuming. SJ suggested that to save time, referencing could be made use of.

Further Actions / Meeting Conclusions:

- NIB brought up the issue of setting a draft deadline; LB went on to add that writing up the sections might not take as much time as actually ensuring that the report flows well from section to section and that there aren't too many unnecessary repeats of certain things that might be brought up in multiple sections. A report template has been put up that group members can add their sections to.
- Next board member meeting to be conducted online at 12:00pm, Friday 04/03/2022.

Meeting Adjourned: 12:52pm, 25/02/2022

Meeting 7

Agenda

Board member meeting started: 12:00pm, 25/02/2022 (Online)

Present: Prof Matthew Wing (MW), Leon Borek (LB), Noor-Ines Boudjema (NB), Jiajun Chen(JC), Dinis De Azevedo Beleza (DDAB), Stefania Juks (SJ), Janos Revesz(JR), Chenyu Zhang (CZ), Ashuit Khanna (AK)

Meeting chair: Noor-Ines Boudjema

Minute Taker: Leon Borek

Meeting Objectives: To check up on the progress made with the report. To discuss how much further the analysis should be conducted

Literature review team

- Progress made with the report (all of literature review)
- Discussion about particle identification (Leon)
- Discussion around the electroweak theory and review of the paper Unanswered Questions in the Electroweak theory, Chris Quigg

Data analysis team

- Progress made with the report (all of data analysis)
- Discussion about the CMSSW and Pipeline part of the plan and whether it should be changed (Janos, Noor)
- A review on Pythia 8 and parton shower Monte Carlo generators (Noor)
- Report on the progress of the level 4 Higgs analysis

Online resources team

- Progress made with the report (all of online ressources)
- Plots and calculations in Jupyter notebook overview (Stefania)
- Discussion around the risk assessment form

Minutes

Literature review team

- DDAB - Tevatron is the second most powerful particle accelerator. only topped by the LHC. Maximum energy output of around 1 TeV. Notable accomplishments of the Tevatron are its discovery of the top quark and 5 Baryons which helped refine the standard model. MW pointed out that the centre of mass energy was 2 TeV and pointed out that the reason that higher energies can be reached with protons than electrons even though the LEP and LHC have the same sized tunnels, is due to synchrotron radiation.

- LB spoke about how despite the LEP possessing a 209 GeV centre of mass energy, which is theoretically enough to produce a Higgs, electron positron colliders don't have the highest Higgs production cross section (Higgs Strahlung pointed out by MW) at this energy. The Higgs only couples very lightly to light particles; heavier particles must be produced in order to produce Higgs in a gluon-gluon fusion, huge data storage would have been required to observe the Higgs through this vector boson fusion at the LEP. Decays to a Z and Higgs boson couldn't occur because the LEP centre of mass energy wasn't high enough.

- LB - a Higgs excess signal was observed at the Tevatron but only 3.1 standard deviations. The instantaneous luminosity of the Tevatron compared to the LHC was very small because the production of antiprotons is much slower that these luminosities can't be as easily reached. Also, since the most common Higgs decay to a bottom-antibottom quark pair has such a high background and unclean signature, it made the Higgs very hard to detect since the cleaner signatures such as diphotons or two Z bosons have much lower branching ratios and thus huge data storage is required in order to obtain enough events. MW highlighted the importance of the Higgs Strahlung process in which a Z radiates a Higgs, if the centre of mass energy reaches the combined masses of the Z and Higgs, this process occurs preferentially and there is very low background. Mw went on to speak about how the regions of search for Higgs mass were decided through fitting data to standard model parameters (blue band plot); there is no fundamental prediction stating what the Higgs mass

should be. A talk through of the most prominent Higgs production mechanisms at the LHC were provided by MW from LB's document.

- CZ explained that the weak force is responsible for radioactive decay in subatomic particles and electromagnetism is the force that governs the interactions of charged particles. According to electroweak theory, when a system reaches an energy higher than 246 GeV, this reaches the vacuum expectation value of the Higgs field and these two forces combine to the electroweak force. The Higgs boson governs mass provision of Z and W bosons.

Data Analysis Team:

- NIB talked about finishing her section about data generation / virtual machine and wrote a small section about Monte Carlo and Pythia 8. The steps that the simulated events go through at CMS was written about alongside descriptions of triggers and pileup. The stages of event simulation and how it relates to parton showers was discussed. The nature of the strong field between partons at high energies was mentioned which triggers hadronisation. It's been understood that event processing records partons and the characteristics of the changes that they experience through sublevels of event processing. Event event follows the evolution of the beams from initial to final states (hadronisation process). The info class keeps track of key information (momentum / energy). MW spoke about how the Pythia 8 processing diagram represents a general purpose Monte Carlo that can simulate many kinds of processes (Higgs / EM / QCD). Parton showers simulate the fact that not everything can be calculated (i.e. there are many interactions that could occur between 2 gluons, not just a Higgs production) from first principles to all orders. It was spoken about that triggers don't simply cut data, the triggers have to be simulated because in reality it is much more complex.

- JC talked about how his writing of the report was ongoing and there was confidence that a way had been found to run the level 4 data analysis which was still running.

- JR had questions about how the detector distinguishes between pp collisions between at the detector. MW spoke about proton bunch crossings and emphasis on time precision of detectors to discern which collisions occurred from different bunch crossings

was emphasised. Proton pileup occurs when more than one proton in a crossing collides, not all protons will collide at the centre and thus trackers must be competent at tracing despite the detectors not being symmetrical about the non-centre points of the detector. High level triggers assumingly are the most sophisticated triggers to keep the most interesting events where more complex algorithms can be applied. The vertex refers to the point of origin. MW stated that small overlap shouldn't be of concern as long as each section clearly addresses a particular distinct thing. He mentioned that 'diary-like' content should be left for the appendix and referenced in the report if needed.

- JR mentioned that he would write about cms config files that are involved in the data generation / simulation / configuration which could be mentioned in the appendix.

Online Resources Team:

- SJ presented the structure of her report section (beginner level / intermediate level / online / offline / other online resources). Could the visualiser be utilised in DDAB's section and perhaps talk about how changing the values of magnetic fields could affect results for example? MW said that this was down to us to decide whether it fitted the flow of the report. There were also queries about the display shown in the visualiser, particularly about muons and electron paths. The reason why muons might not always

  show up in the visualiser could be due to the fact that muons don't interact much with matter and so there won't be much of a signal there. For example, neutrons too can pass through a HCAL which might not leave a track. Diphoton invariant mass histograms were displayed.

- AK asked about whether he should go into detail on the J psi / prime plots and MW replied that this is dependent on the rest of the report flow.

- The risk assessment was discussed

Further actions/meeting conclusions:

- Meet next friday at the same time

Meeting Adjourned: 1:30pm, 04/03/2022

Meeting 8

Agenda

Board member meeting started: 12:00pm, 11/03/2022 (Online)

Present: Prof Matthew Wing (MW), Noor-Ines Boudjema (NB), Jiajun Chen(JC), Dinis De Azevedo Beleza (DDAB), Stefania Juks (SJ), Janos Revesz(JR), Chenyu Zhang (CZ), Ashuit Khanna (AK)

Absent: Leon Borek - ill at the time of the meeting and could not attend

Meeting chair: Janos Revesz

Minute Taker: Dinis de Azevedo Beleza

Meeting Objectives: To check up on the progress made with the report

Literature review team

- Progress made with the report
- Higgs mechanism

Online resources team:

- Progress made with the report (all of online resources)
- The visualizer (Stefania)

Minutes

Literature Review Team

- DDAB presented his progress during this week. He mentioned that he spent this week mainly organizing the sections that he had written in the weeks before to make sure that they would fit together and that they would be ready for the report. Additionally, he mentioned he wrote a paragraph about the Synchrotron radiation as discussed last week and that he also wrote a small section on Monte Carlo Methods to be used as reference for the data analysis team.
- CZ presented what he has been writing for the past two weeks. He showed those sections while making a brief description on what they were about and their order. He added that he would be adding the sections to the drive after the meeting. MW

stated that the sections presented by the literature review team were good and that it was only a matter of knowing how to put them together and making sure that no repetitions were made.

Data Analysis team

- JC mentioned that the analysis is still running and that he believes that in a day or two it would finish running so that they can get plots. Additionally, he said that he and JR are finishing writing the final report sections on this analysis.

- NIB started by mentioning that she was basically finished with her sections. Then, she showed a plot from one of the weeks before to explain that she figured out the reason why the mass of the Higgs boson in that decay was shown as a smaller value than the 125 GeV that we know as being its actual mass. Additionally, she gave an overview of her written sections. MW reiterated that the sections look good however it will always depend on the way that the report is organized and if they make sense to be together in that order. He also added that a good way of presenting this in a report is to show that in a specific decay is not very easy to discover the Higgs but then in others with different backgrounds and rates it is easier and make that comparison. In the end, NIB just asked if the template on Overleaf looked fine and MW stated that using Overleaf or Latex would work.

- The deadline for the report was discussed and MW stated that the group needs to be careful because joining different sections is not just about the technical part but also understanding if those sections fit well together.

- JR explained what he has been writing for the report and how the data analysis team has organized their part. He also mentioned that they will make some comparisons between different types of analysis and will explain which are simpler and faster to do. He followed by explaining an experience that he had during a Monte Carlo simulation in which he made an error in it and because of that error the analysis had to be started over and that lost him some days that could be useful. He followed by saying that if another problem does not arise the analysis will be finished in a couple of days. MW stated that it is common in this area of physics to do that kind of mistakes that end up in a lost of some days of work and that even him has already

done some. After JR asked if a comparison between two different analysis was a good idea, MW mentioned that it is a great idea and that in High energy physics it is common to have some comparisons, mainly for important experiments like this one.

Online Resources Team:

- SJ showed her report parts and gave an overview about what they were about. Then she showed different plots and questioned if all of them should be shown in the report or just some of them. MW stated that she should consider if the plots have some relevant information to be taken from them and only choose the ones where there is a clear physics meaning and people can learn from them. SJ then followed by saying that she found an animation that would be helpful but she did not know if she should add it since it is a report. MW agreed that the animation would be very helpful and she should try to take 4 snapshots in order to have the different effects in a report. SJ finished by saying that a person should be in charge of the references and making sure that they have the same format.

- AK mentioned what his section was about and gave an overview of it. SJ followed by showing a specific histogram questioning whether it had more options. AK stated that it does but this one was just an example, however, he could add more after the meeting.

General Discussions:

- JR mentioned that the tutorial that was planned to be in the report will not be there since the other parts of the report took much longer than was expected. He followed by asking how the poster should be organised. MW stated that the poster should be approximately divided in equal parts (one for each team) and show the most important parts of the report. JR asked if MW had any idea how the poster session was organized since there is not many available information about it. MW mentioned that he was not sure.

Further Actions / Meeting Conclusions:

- Meet Wednesday at 1pm

Meeting Adjourned: 12:42, 11/03/2022

C Finances

The allocated budget of 250 pounds was not utilised for the completion of our project. All restricted papers were easily accessed through our UCL accounts.

D Certificates

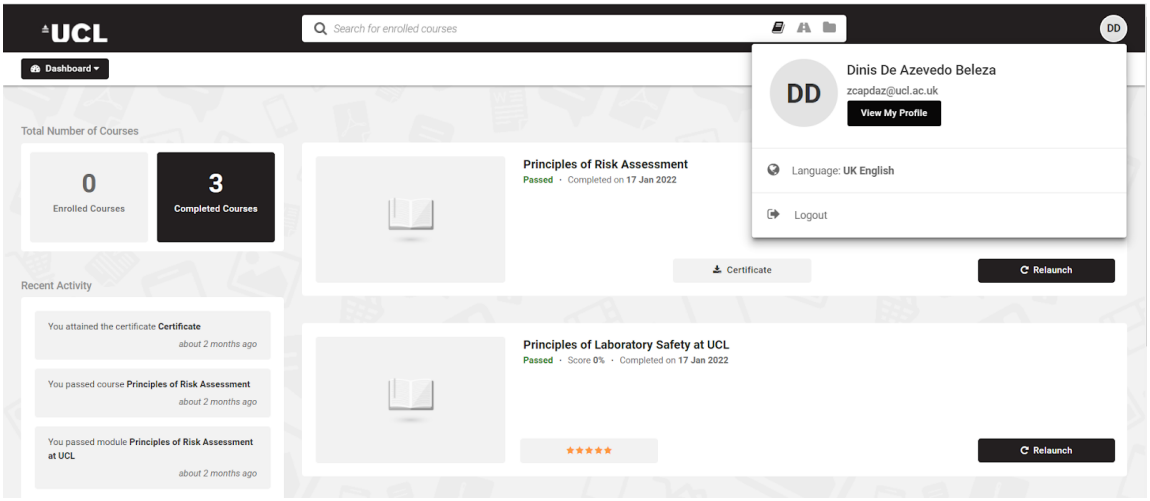


Figure 70. Dinis de Azevedo Beleza's Dashboard page, showing the completion of the required health and safety courses.

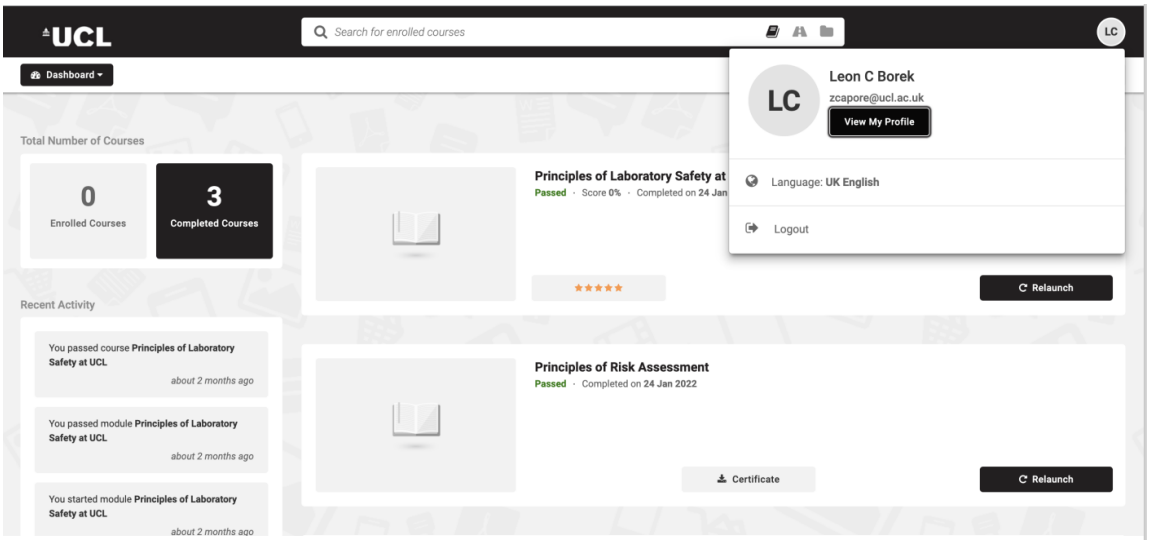


Figure 71. Leon Borek's Dashboard page, showing the completion of the required health and safety courses.

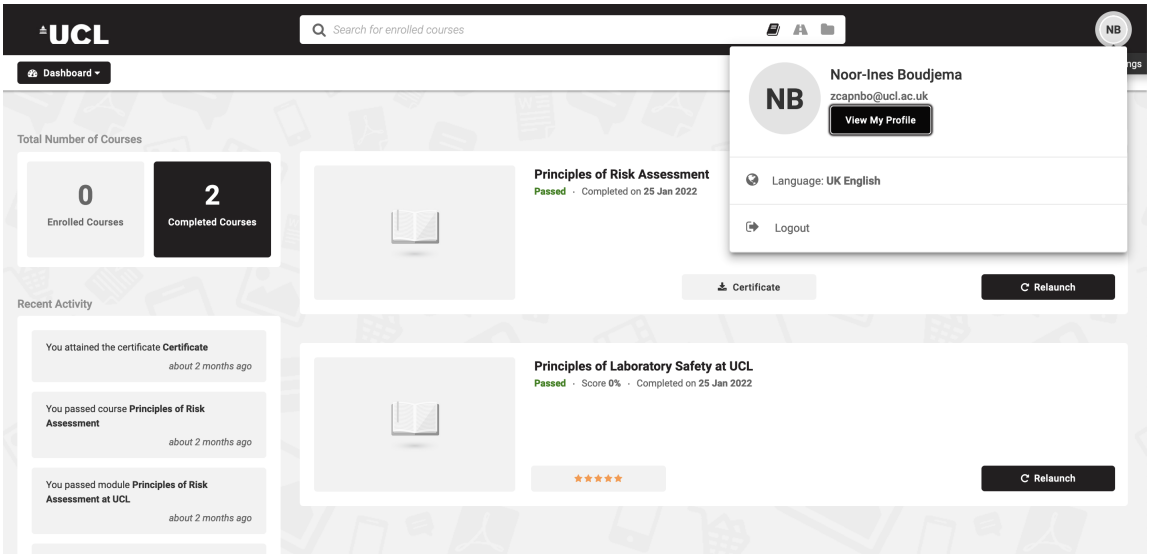


Figure 72. Noor-Ines Boudjema's Dashboard page, showing the completion of the required health and safety courses.

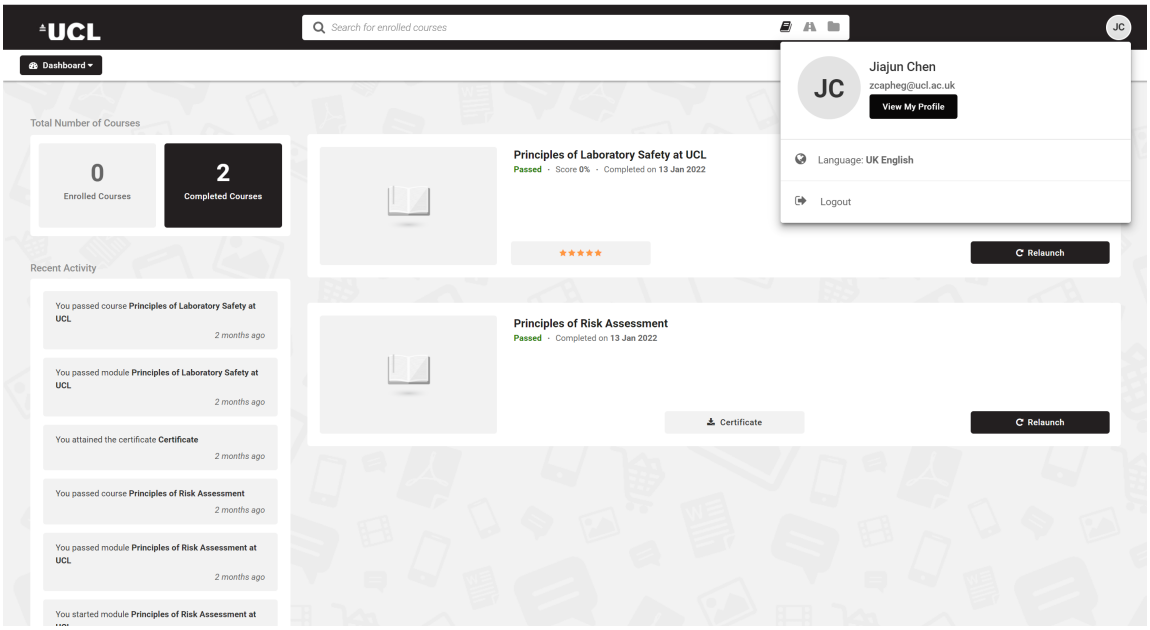


Figure 73. Jiajun Chen's Dashboard page, showing the completion of the required health and safety courses.

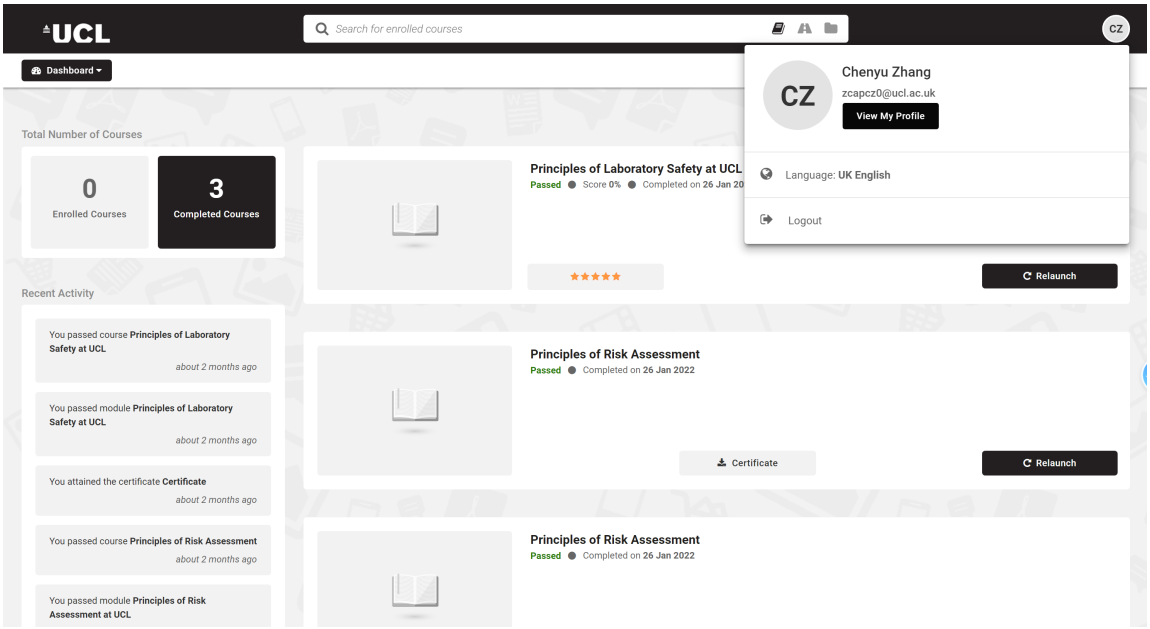


Figure 74. Chenyu Zhang's Dashboard page, showing the completion of the required health and safety courses.

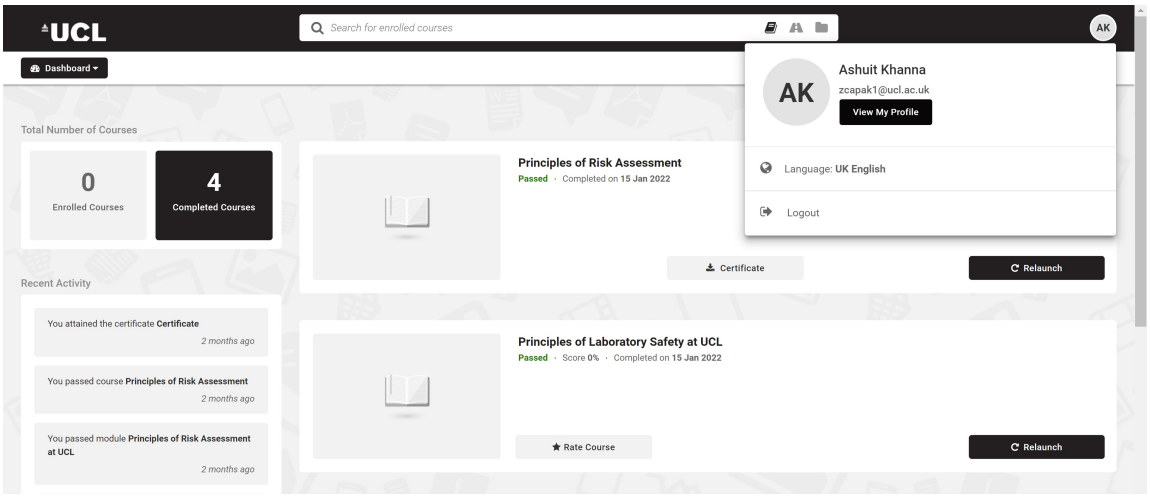


Figure 75. Ashuit Khanna's Dashboard page, showing the completion of the required health and safety courses.

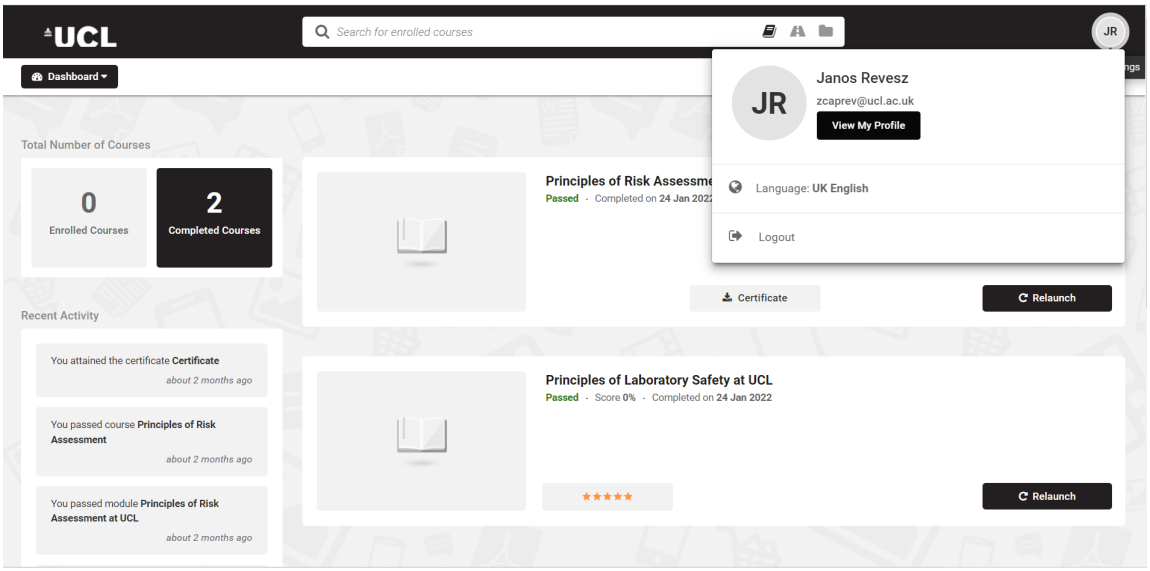


Figure 76. Janos Revesz's Dashboard page, showing the completion of the required health and safety courses.

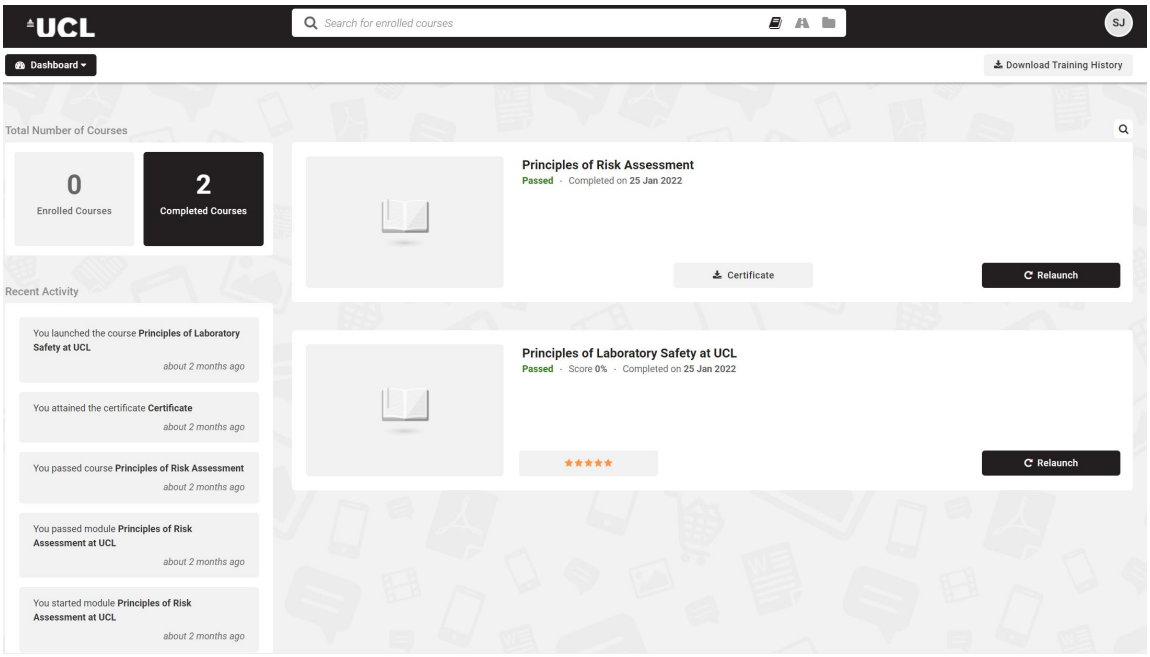


Figure 77. Stefania Juks's Dashboard page, showing the completion of the required health and safety courses.

E Risk assessment form

Project Risk Assessment Form

Project Title	Rediscovering the Higgs boson at the CMS experiment
Location of Experiment	Remote
Description of Experiment	Literature review, online research and data analysis
Lead Experimenter	Noor-Ines Boudjema
Other Persons Involved	Leon Borek, Jiajun Chen, Dinis De Azevedo Beleza, Stefania Juks, Ashuit Khanna, Janos Revesz, Chenyu Zhang
Supervisor / Board Member	Professor Matthew Wing

Hazard Identification *(state the hazards involved in the work. Consider **Chemicals, Radiation, LASERS** (an additional assessment will also be needed), the **environment, equipment, manual handling, electrical equipment, fire and explosion, disposal of waste**)*

1. Remote working on a single device.
2. Running large data sets for longer periods.
3. Network failure while running code remotely through a cluster.
4. Prolonged exposure to computer light.

Risk Assessment *(assess the risks involved in the work and state high, medium or low risk)*

riskNET Incident 6 x 5 risk matrix

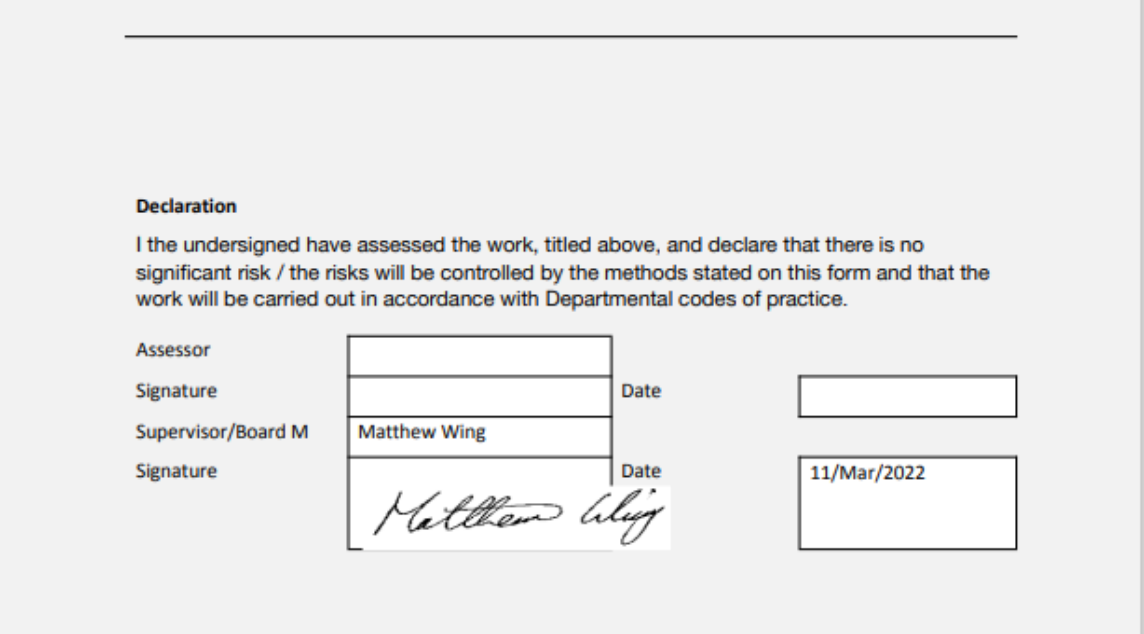
Severity	Likelihood: Remote	Unlikely	Possible	Likely	Certain
Non injury	A	A	A	A	A
Minor injury	A	A	B	B	C
Lost time injury, temporary disability or illness	A	B	C	C	D
Permanent disability or major injury	B	C	C	D	D
Fatality, multiple serious injuries/illnesses	C	C	D	D	D
Multiple fatalities	C	D	D	D	D

A	Very Low Risk - Initial Assessment
B	Low Risk - Local Investigation
C	Medium Risk - Local/Full Investigation
D	High/Very High Risk - Full Investigation

1. Data loss (medium risk).
2. Hardware damage through overheating/overloading (low risk).
3. Compilation failure leading to a loss of time (low/medium risk).
4. Eye strain (low risk)

Control Measures (say how you will reduce the risk to an acceptable level)

1. Complete work in a shared drive, constantly saving updates.
2. Use devices within their allowed specifications - use a cluster for analysis.
3. Attempt large data sets compilations only when connected to a stable network.
4. Take constant breaks from looking at the monitor.



Declaration

I the undersigned have assessed the work, titled above, and declare that there is no significant risk / the risks will be controlled by the methods stated on this form and that the work will be carried out in accordance with Departmental codes of practice.

Assessor			
Signature		Date	
Supervisor/Board M	Matthew Wing		
Signature		Date	11/Mar/2022

Figure 78. Risk Assessment Form.