Optimisation of the search for new high mass Higgs bosons in the four-lepton channel with the CMS experiment

Mandarić, Marko

Master's thesis / Diplomski rad

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Split, Faculty of Science / Sveučilište u Splitu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:166:735715

University of Split
Faculty of Science

OPTIMISATION OF THE SEARCH FOR NEW HIGH MASS HIGGS BOSONS IN THE FOUR-LEPTON CHANNEL WITH THE CMS EXPERIMENT

Master thesis

Marko Mandarić

Split, September 2022.

The work for this thesis was done at the LLR laboratory (Laboratoire Leprince-Ringuet) in Palaiseau, France, as a part of an Erasmus+ project. I would like to thank everyone there for their welcome, help and support; especially Christophe Ochando and Axel Buchot, with whom I worked more closely.

Also, I would like to thank my professor and mentor Toni Šćulac for always providing advice, guidance and help during my student years, as well as mentoring both my bachelor and master thesis. On top of that, he made my trip to France possible. I am truly grateful.

Temeljna dokumentacijska kartica

Sveučilište u Splitu
Prirodoslovno – matematički fakultet
Odjel za fiziku
Ruđera Boškovića 33, 21000 Split, Hrvatska

Diplomski rad

OPTIMIZACIJA POTRAGE ZA NOVIM HIGGSOVIM BOZONIMA NA VISOKIM MASAMA U KANALU RASPADA NA ČETIRI LEPTONA S CMS EKSPERIMENTOM

Marko Mandarić

Sveučilišni diplomski studij Fizika, smjer Astrofizika i fizika elementarnih čestica

Sažetak:

U ovom radu prezentiraju se rezultati istraživanja mogućih poboljšanja u analizi raspada Higgsovog bozona visoke mase na 4 leptona. Korišteni su podaci CMS eksperimenta iz Velikog hadronskog sudarača u CERN-u, skupljeni od 2015. do 2018. godine, tzv. Run 2. Izvodi se u studija u sklopu povećanja efikasnosti u potrazi za novim skalarima, tj. Higgsovim bozonima visokih masa. Provedena su tri tipa analize. Za početak, prikazana je usporedba dvaju uzoraka podataka. Jedan od njih nastao je prije, a drugi nakon kalibracije detektora u 2019. godini. Usporedba je ranije provedena za Higgsov bozon iz Standardnog Modela, onaj mase 125 GeV, a ovo je prva provjera u analizi visoke mase. Nadalje, proučava se potencijalno poboljšanje analize popuštanjem nekih zahtjeva u njoj. Uspoređuje se eventualni benefit relaksacije granične vrijednosti sa količinom dodatnih pozadinskih smetnji koje mogu otežati analizu. Konačno, istražuju se produkcijski mehanizmi Higgsovog bozona i kategorizacija samih produkcijskih događaja u analizi visoke mase promatrajući ponašanje vrijednosti određenih diskriminacijskih varijabli. Za svaku je temu objašnjena metoda i pokazani rezultati istraživanja.

Ključne riječi: CMS, Higgsov bozon, analiza visoke mase, optimizacija

Rad sadrži: 62 stranice, 33 slike, 42 tablice, 16 literaturnih navoda. Izvornik je na engleskom jeziku.

Mentor: doc. dr. sc. Toni Šćulac

Neposredni voditelj: dr. sc. Christophe Ochando

Ocjenjivači: doc. dr. sc. Toni Šćulac,
doc. dr. sc. Marko Kovač,
mag.phys. Andro Petković

Rad prihvaćen: 30. rujna 2022.

Rad je pohranjen u Knjižnici Prirodoslovno – matematičkog fakulteta, Sveučilišta u Splitu.

Basic documentation card

University of Split
Faculty of Science
Department of Physics
Ruđera Boškovića 33, 21000 Split, Croatia

Master thesis

OPTIMISATION OF THE SEARCH FOR NEW HIGH MASS HIGGS BOSONS IN THE FOUR-LEPTON CHANNEL WITH THE CMS EXPERIMENT

Marko Mandarić

University graduate study programme Physics, orientation Astrophysics and elementary particle physics

Abstract:
Research of some possible optimisations of the high-mass Higgs boson decaying to 4 leptons analysis is presented. The full Run 2 data (2015-2018) collected by the CMS experiment in the Large Hadron Collider is used. The study is carried out as part of the effort to make the search for new scalars (heavy Higgs bosons) more efficient. There are three parts of the analysis conducted. The first is comparing two data samples, before and after an updated calibration of the detector. Next, a study is performed to try to determine whether it is possible to loosen some restrictions in the analysis with a goal of improving the results, while also controlling the "explosion" of noise events. Finally, categorisation of the production mode of the events (how the Higgs boson is created) is explored, taking into consideration the values of the discriminating variables. For each topic, the methods of work are explained and results shown.

Keywords: CMS, Higgs boson, high-mass analysis, optimisation

Thesis consists of: 62 pages, 33 figures, 42 tables, 16 references. Original language: English.

Supervisor: Assist. Prof. Dr. Toni Šćulac

Leader: Dr. Christophe Ochando

Reviewers: Assist. Prof. Dr. Toni Šćulac,
Assist. Prof. Dr. Marko Kovač,
mag.phys. Andro Petković

Thesis accepted: September 30, 2022.

Thesis is deposited in the library of the Faculty of Science, University of Split.

Contents

1 Introduction

CERN, the European Organization for Nuclear Research, is one of the biggest scientific organisations in the world. Thousands of people are involved in this collaboration, their work mostly related to high energy particle physics. There are many cooperating institutes, laboratories etc. all around Europe, but when you say CERN, most people will think of the LHC accelerator on the Swiss-French border.

The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator. The LHC consists of a 27-kilometre ring of superconducting magnets with a number of accelerating structures to boost the energy of the particles along the way. The beams inside the LHC are made to collide at four locations around the accelerator ring, corresponding to the positions of four particle detectors – ATLAS, CMS, ALICE and LHCb [1]. The reason the LHC is so big is because the particles need to reach very high speeds and energies for new particles to be created and possible new physics discovered. The detectors are also impressive technical and technological miracles with great precision and efficiency.

Possibly the most well-known achievement made in CERN was the recent (2012.) discovery of the Higgs boson particle. It was a great success as it represents the confirmation of the theoretical mechanism that explains how all particles gain mass - through an interaction with the Higgs field. The mass of the Higgs boson itself has been determined and measured with rising accuracy; in 2019. the CMS collaboration has announced the most precise measurement of this property so far : $125.35GeV$ with a precision of $0.15GeV$, or 0.12% [2].

The particle that was found and studied extensively is compatible in its properties to the one predicted by the Standard Model (SM). However, the Standard Model is known to have problems in explaining some phenomena such as gravity, dark matter, neutrino mass, which are explained by other theories, or not yet explained at all. So, beyond-the-SM (BSM) physics researches and theories are brewing, looking for answers in other ways, using different strategies. One of such being the search for additional heavy scalars, that would prove the presence a non-minimal Higgs sector. The existence of such a sibling Higgs boson is motivated in many BSM scenarios, so the search for additional scalar resonances in the full mass range accessible at the LHC remains one of the main objectives of the experimental community [3].

The Standard Model Higgs boson has been studied a lot, the analysis framework updated and perfected for searching and learning as much as possible about this specific particle. At the same time, people are looking for a possible high mass Higgs boson. To an untrained eye, a "simple" change in the mass window in which we are trying to find a particle may seem easy enough. However, one needs to keep in mind that the methods perfected for a certain research may not be as good for an another one, since a lot of variables are in play here. I will explain in some examples my contribution in trying to optimise the analysis for a high mass Higgs, working in the $H \rightarrow 4l$ channel, meaning that the final products that the Higgs boson is reconstructed from

are 4 leptons (electrons or muons). I worked on 3 separate topics:

1) Comparing two sets of reconstructed data produced for analysis; in order to check for possible agreements or disagreements

2) Studying optimisation possibilities regarding some analysis restrictions, so-called cuts on certain variables

3) Categorisation of events with respect to the Higgs production modes

All this was done before for SM Higgs, but not as of yet for high mass - here lies my contribution. I will explain in detail what all of the points I presented are, what methods are used, and show my results.

2 CMS Experiment at CERN

The work for this thesis was done within the CMS experiment, using the Run 2 data from the Large Hadron Collider (LHC). I will give a short overview of CERN, CMS and present a little bit about the physics of the Higgs boson, "the main character" in this study.

2.1 History

CERN officially came to be on September 29, 1954. The full name in French is Organisation Européenne pour la Recherche Nucléaire, (European Organization for Nuclear Research), but the 12 founding countries first called it a council in 1952. (Conseil Européen pour la Recherche Nucléaire), hence the "C" in the acronym. On a similar note, the lab is mostly devoted to high energy physics, but at the beginning it studied atomic nuclei, hence the "N".

CERN provides particle accelerators and other technology necessary for elementary particles research to be done by international collaborations. The main site, LHC (Large Hadron Collider) is located at the French-Swiss border, close to Geneva. Four big particle detectors – ATLAS, CMS, ALICE and LHCb are located at certain points of the 27-km ring where the beams of particles collide and interesting processes happen. There is also an impressive computing facility for storing and analysing data, which can be accessed remotely - CERN is historically known as the first place where the World Wibe Web (WWW) was developed [4].

Some of the most important achievements made on CERN include:

-1973: The discovery of neutral currents in the Gargamelle bubble chamber

-1983: The discovery of W and Z bosons

-1999: The discovery of direct CP violation

-2012: The Higgs boson discovery [4]

The last one is probably the most well known by the general public. The Higgs mechanism that theoretically explains how particles gain mass was confirmed with this great discovery, a major breakthrough in modern science.

2.2 CMS

How exactly are particles detected in the LHC? The accelerator is located 100 meters underground. This way a lot of interference is removed, be it human activity, cosmic rays or any other source.

The LHC accelerates proton beams in both directions to energies of around 7 TeV, which is 14 TeV of center-of-mass energy. As seen in Figure 1, there are additional smaller accelerators

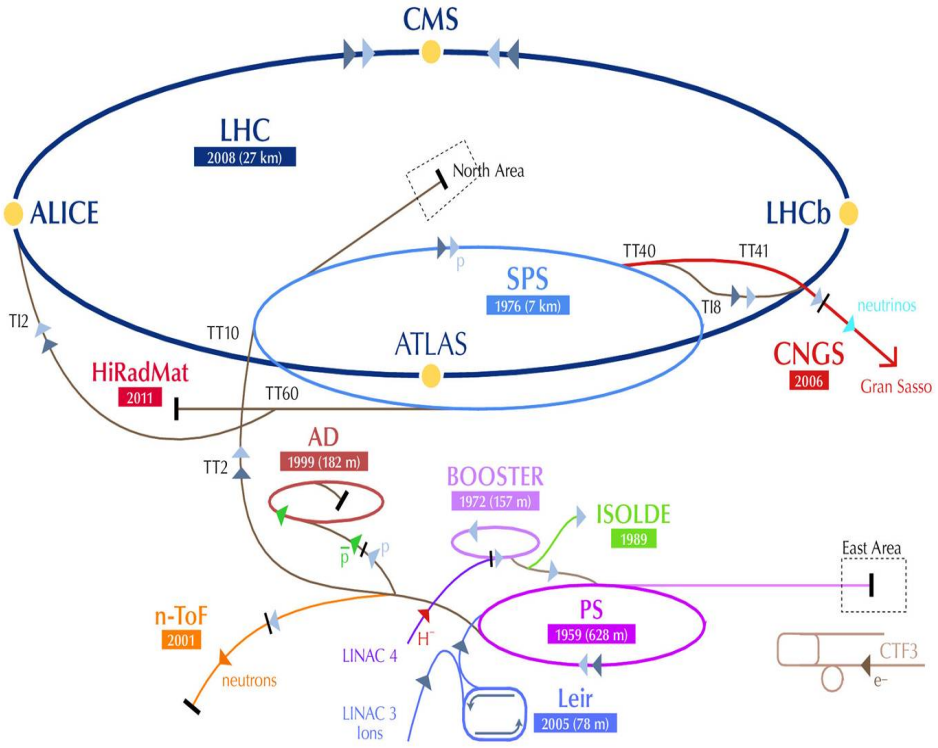


Figure 1: *The complex of accelerators and detectors at CERN. Taken from [4]*

before the particles enter the big ring. The beams are kept stable at their track by magnets, which require an advanced cryogenic system to be kept at working temperature.

As mentioned before, there are four main detectors (more smaller ones). The biggest two are CMS and ATLAS, multi-purpose detectors on the opposite sides of the ring with similar tasks but different technologies implemented inside them. Since I worked with the CMS group, I will focus on describing the CMS detector.

CMS stands for Compact Muon Solenoid. "Compact", because of the ratio of its dimensions and mass (21 meters long, 15 meters wide and 15 meters high; weighing 14000 tons). "Muon", because of an advanced muon detection system with amazing efficiency. "Solenoid", because of the solenoidal magnet inside of it which creates a strong magnetic field. I will try to explain, in short, al subsystems of the detector.

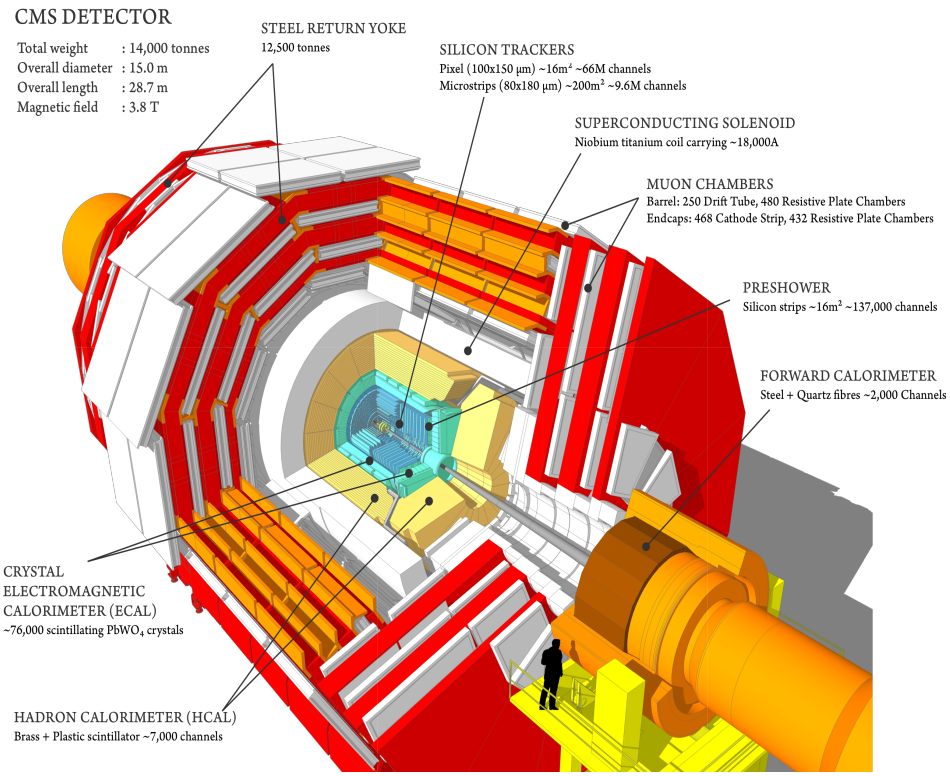


Figure 2: *A view of the CMS design with a human for scale. Taken from [4]*

The important thing to know about the CMS detector is that it should be able to detect pretty much all kind of particles. This is why it is called a multi-purpose detector. Sensitivity in all ranges and to all kinds of particles is not easy to achieve. Different parts of the detector have a purpose of detecting different particles. The structure of the detector is often described as onion-like, with each layer having a special design.

The first layer, closest to the particle collisions is the tracker. Its job is to reconstruct the paths of the charged particles. Since it is placed in a magnetic field, it is possible to differentiate the charge of particles, positive, negative or neutral, by the curvature (or lack thereof) in their tracks. Also, from the curvature, their momentum can be calculated. The sensors are made from silicon with two different techniques: silicon pixels (100x150 μm^2) in the inner layer and silicon strips in the outer layer of the tracker. The main challenge here is to achieve high granularity to be able to separate nearby particles. Also, each pixel must have its own readout channel, so the electronics play a big role.

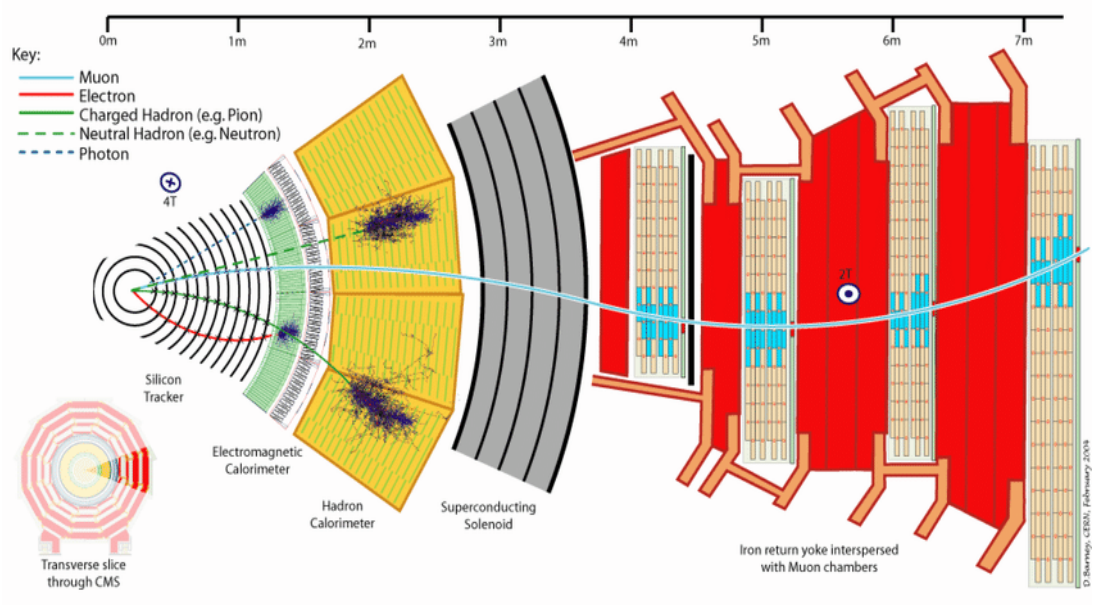


Figure 3: *A view of the CMS onion-like structure of sub-detectors. Taken from [4]*

Zooming out from the collision spot, next up is the electromagnetic calorimeter (ECAL). It is made of a cylindrical inner part (barrel) and two endcap disks on the ends. As every calorimeter, it is used for measuring energy, and electromagnetic suggests that it measures it for the particles that interact via the EM force - electrons and photons. CMS uses lead tungstate ($PbWO_4$) crystals as a way of measuring energy. The principle is that the crystals scintillate when electrons and photons pass through it and interact with it. The crystal must be transparent for the light that is produced in this process so that it can be collected by the photodetectors in the back of the crystal, the signal amplified and analysed. So basically, it collects the photons and, since the light produced is proportional to the energy of the incoming particle, there is a way to calculate it.

After the ECAL, there is HCAL - the hadronic calorimeter. This means that it measures the energy of hadrons, particles that do not interact with the ECAL because they are made of quarks and gluons. It is built as alternating layers of a dense absorber and plastic scintillator tiles. When a hadron hits the absorber, a shower of particles is produced. The scintillating material produces light that is, similarly to ECAL, absorbed. The optic fibres send the light to photodetectors, signal is amplified and read as a measure of the incoming particle's energy.

A crucial part of CMS is the solenoid magnet. Its task is to bend the trajectories of charged particles by producing a strong magnetic field of 4 Tesla. An iron yoke around the magnet is the part of CMS that weights the most, a staggering 10000 tons, and it is responsible for the fields homogeneity.

Muons are particles that have a much smaller chance of interaction than electrons, so they are not caught in the ECAL, nor do they show themselves in the HCAL. This is why big muon chambers at the furthermost layer of the CMS are required to efficiently detect muons, their

momentum and charge. A complex sub-detector system is built of Drift tube chambers, Cathode strip chambers and Restistive plate chambers. The principle remains similar, after an interaction by the muon with the detector, resulting energy is collected, summed up and read out.

There are particles that the CMS cannot detect - neutrinos. The probability of interaction for a neutrino is so small that it simply passes through all the layers basically unnoticed. The good thing is that, if all other particles in the experiment are measured, neutrinos can be indirectly accounted for as they carry the "missing" energy and momentum. Neutrino detection requires entire huge detectors with their own technology and is still an ongoing work.

One more thing to mention when talking about the detector is the trigger system. With so many readouts happening very fast - a collision happens every 25 nanoseconds, that is a very large amount of data in a very short time - it is impossible to save all these information. There is a trigger system developed to discard great amounts of data, while still finding the events interesting to save - where possible new physics can be discovered. The trigger is split into 3 levels. L1 is hardware based, L2 and L3 are the software, high level trigger. Initial 40 MHz of incoming data need to be reduced down to manageable 100 kHZ. Still, the amount of data produced every year is measured in hundreds of petabytes.

3 Higgs boson

3.1 Standard Model

CMS (as ATLAS) has been primarily built to search for the Higgs boson. It would have been a triumph for the Standard Model , a quantum field theory describing three of the four known fundamental forces in the universe, as well as classifying all elementary particles. The problem it had for a long time was that all the particles described were to be massless. More precisely, the Lagrangian (the function that characterizes the state of a physical system) of the theory is said to have a symmetry if it is invariant under certain transformations. Gauge bosons having mass broke the gauge symmetry and fermions' mass violated the symmetry of the QCD part (quantum chromodynamics, theory studying the strong interaction) . As a solution, the Higgs mechanism was introduced in 1964 by three independent groups of physicists - a spontaneous symmetry breaking mechanism which predicts a scalar field that allows other elementary particles in the SM to acquire mass, and a particle (boson) is sought for in order to confirm the existence of this field. As we know, it was accomplished in 2012, but how did the scientists know where to look and what kind of detectors to build? Well, the theory had predictions about the production mechanisms and decay modes of the Higgs boson.

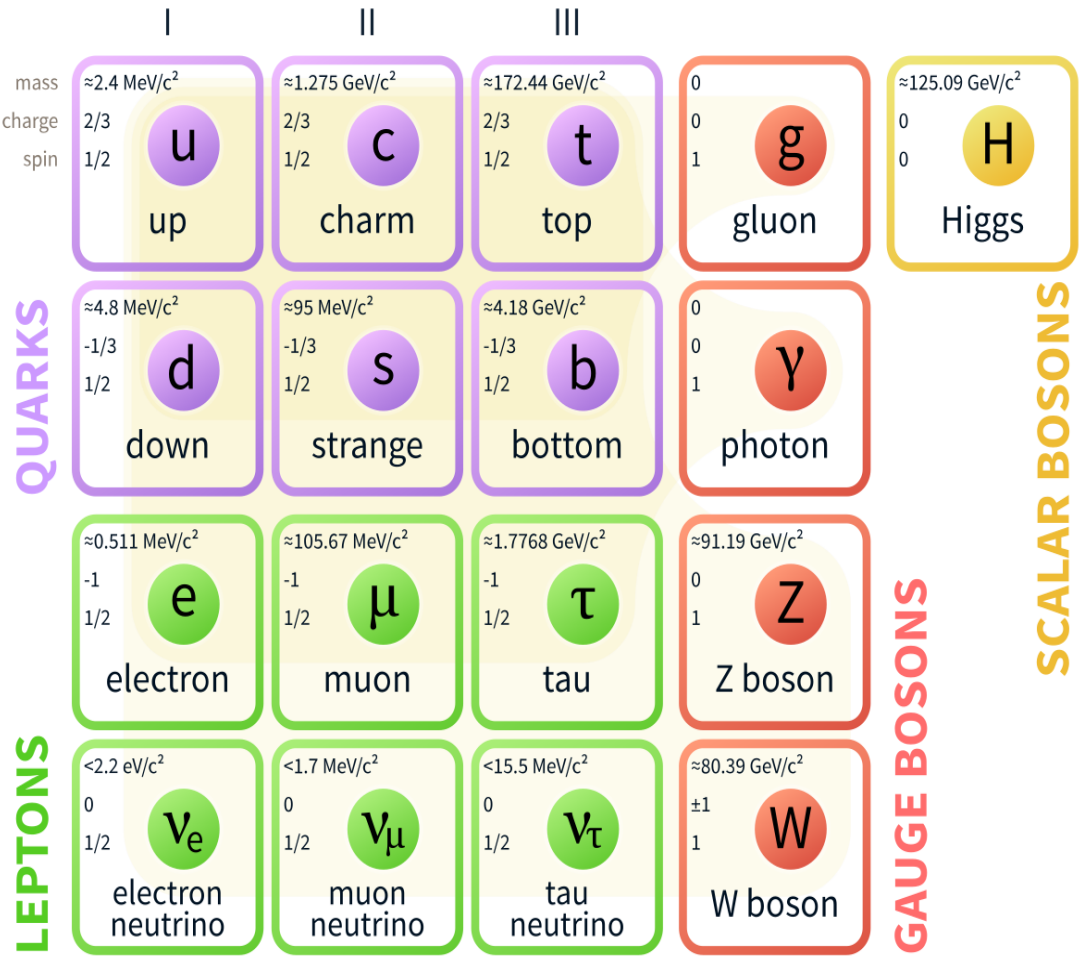


Figure 4: *The Standard Model of particle physics. Taken from [5]*

There are a lot of ways that Higgs can be produced according to the SM. However, there is a difference in the cross section and some modes are dominating, while some don't occur as often. Focus will be put on the main ones which were also actually used in the later study of the high mass Higgs beyond the SM. This includes gluon fusion and vector boson fusion. Feynman diagrams of these processes are shown below.

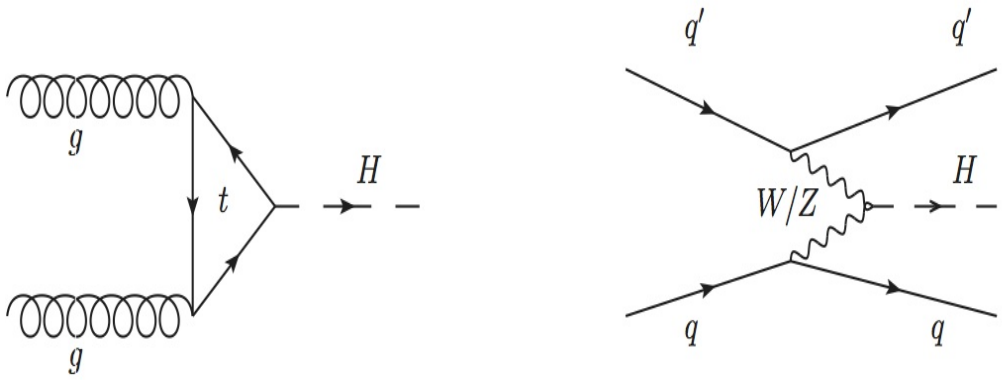


Figure 5: *Two main production modes of the Higgs boson - gluon fusion (left) and vector boson fusion (right). Taken from [4]*

The gluon fusion, noted ggH, where two gluons fuse via an intermediate loop of virtual quarks, has the largest cross section dominating other production modes by more than one order of magnitude. This is because the gluon luminosity is very large in pp collisions at the high centre-of-mass energies provided by the LHC [4].

Vector boson fusion, noted VBF, is the second production mechanism at the LHC with a cross section roughly an order of magnitude smaller than that of ggH. It occurs when two fermions exchange virtual W or Z (vector) bosons, which immediately fuse into the Higgs boson. This production mechanism is very important and interesting because it has a clear signature with forward and backward jets with high invariant mass [4]. This will be used in the Categorisation section of this thesis.

From the thesis' title it can be seen that I am working in the $H \rightarrow 4l$ analysis; the decay of the Higgs boson into 4 leptons. Quantum mechanics teaches us that heavier particles tend to decay to lighter ones, if possible. The SM has predictions about the mean lifetime of the Higgs boson, which for 125 GeV Higgs is $1.6 \times 10^{-22}s$. This is way to short for Higgs to reach the detector, which means that we have to detect the decay products and reconstruct it backwards. Similar to the production mechanisms, there is a lot of decay modes with different probabilities of it happening. Here, the mode with the highest branching ratio (highest probability) is not necessarily the best option to search for Higgs. The reason is that the decay products are hard to differentiate from background events (not from Higgs decay). This is why the problem of the lower branching ratio can, especially at high luminosity of events, be circumvented if the decay channel benefits from complete and effective reconstruction of the final state. In simple terms, there may not be so many events in a decay mode, but we can precisely detect all particles in it. This is the case in reality; the "Golden channel" name was given to the $H \rightarrow ZZ^* \rightarrow 4l$ decay. Complete reconstruction of the final objects, very good momentum resolution and great signal to background ratio are all reasons why.

Of course, finding 4 leptons that could come from a Higgs boson is not enough to proclaim it Higgs signal - there are a lot of processes with the same decay products, but without ever creating Higgs. There are several of these predicted by the SM. These processes are what is called irreducible background - meaning they do not come from our desired event, but all the same particles are reconstructed as the decay product (in our case, 4 leptons are found, they may have come from 2 Z bosons, but Higgs was not created in the first place). Irreducible refers to the fact that we can not reduce the number of these background events, which points to the existence of reducible background.

In a perfect world, perfect people with perfect detectors would never make a mistake in recognising a particle or measuring some property. In reality, some do get miss-reconstructed and selected as matching the targeted event and classified as signal. As this can be reduced by a number of ways, it is called reducible background. A main tool in this is doing selection cuts designed to reject the background events and keep the signal ones. All reducible background in

the 4 lepton channel is denoted by Z+X, where X stands for a Z boson reconstructed from two unrelated leptons [4]. This needs to be carefully accounted for, and it will also play a role in my high mass analysis.

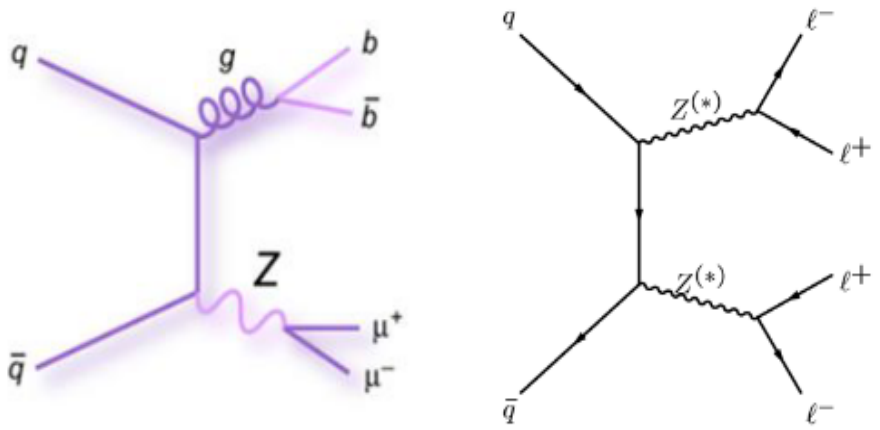


Figure 6: *Left: If the jets by b quarks were to be misidentified as electrons, it is an example of a reducible background event. Right: Two Z bosons decay into four leptons but the Higgs boson was not produced at all, an example of an irreducible background event. Taken from [7]*

In order to obtain good estimates of the background in the signal region, control regions orthogonal to the signal region (i.e. that do not contain any signal) and enhanced in a specific background are defined. The fake rate or efficiency in that control region, that is the ratio of events that are wrongly identified as electrons (or muons), is then used to extrapolate how many events are misidentified in the signal region [7].

One more thing to note; 4 leptons as final objects have 3 possible combinations: 4 electrons ($4e$), 4 muons (4μ) and 2 electrons 2 muons ($2e2\mu$).

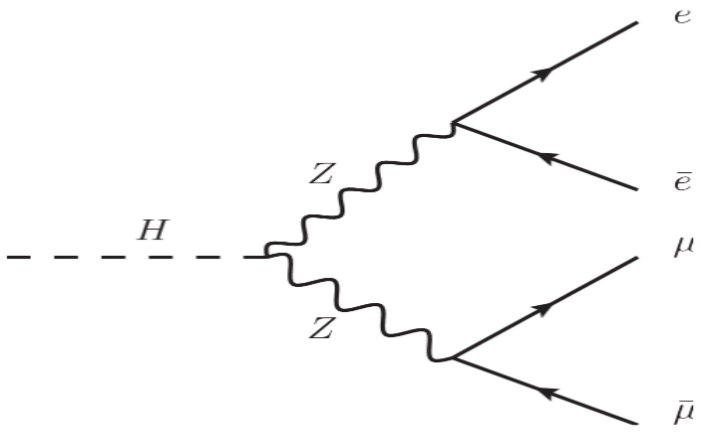


Figure 7: *Feynman diagram of the decay of the Higgs boson to four leptons ($2e2\mu$ case). Taken from [8]*

3.2 High mass

I mentioned in the previous section that the SM predictions about the lifetime of the Higgs boson is known for 125 GeV. However, the mass of the Higgs boson is a free parameter, meaning it was not predicted beforehand by the theory. Think about this for a second - when planning and building the detectors, people weren't sure what is the mass of the particle they were searching for. When it was finally found, with more than 5σ certainty (meaning 99.99994% confidence), the excess of events was observed around 125 GeV. The mass was later measured with rising accuracy over the years.

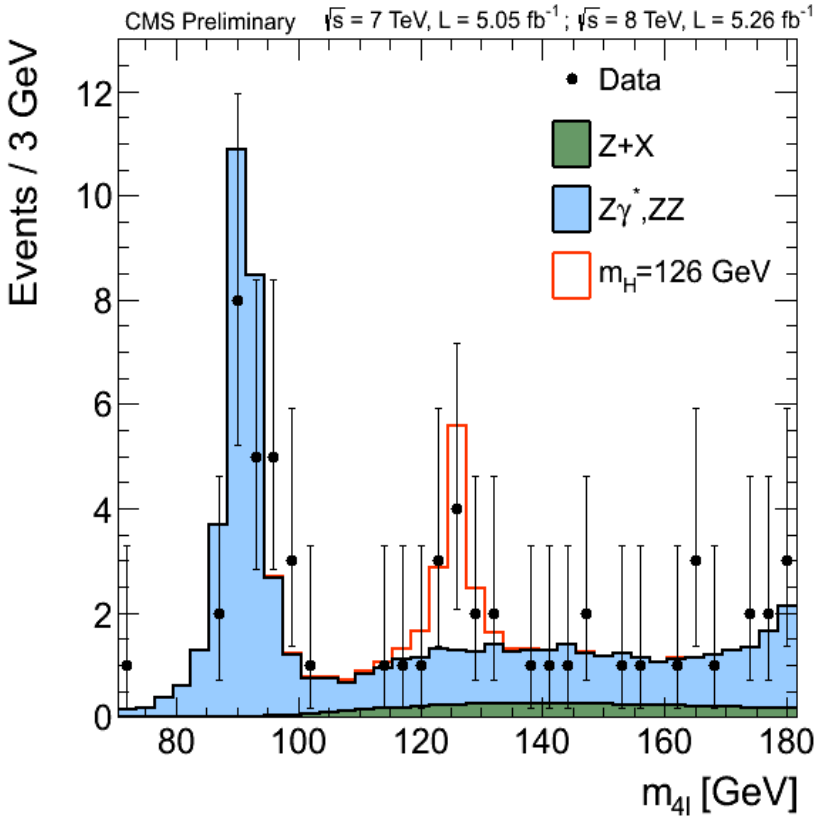


Figure 8: *The distribution of the four lepton invariant mass - the reconstructed Higgs mass - for the $ZZ \rightarrow 4l$ analysis of CMS. Taken from [3]*

The Standard Model, albeit the best theory we have about the elementary particles, does not provide a complete picture. There are still open questions: gravity, dark matter and the neutrino mixing and mass are some of the examples.

Gravity is one of the four fundamental interactions but it is not described in the SM. It is so different from the three other forces - establishing a theory that includes all of them proved to be incredibly hard. Gravity is described in the Einstein's General Relativity theory (GR). An attempt at combining the SM with the GR is made with a new field associated to gravity as mediator: a particle called graviton is theorised, with no experimental evidence of its existence [3].

Furthermore, we know from astronomical observations that only 5% of the matter and energy content of our universe is formed by the ordinary matter (hadrons and leptons), the other 95% is composed of dark matter (25%) and dark energy (70%). The SM does not offer good candidates or explanations for the dark matter and dark energy problems [3].

Concerning the neutrinos, in the SM they were assumed, as other particles, massless. However, flavour oscillation implies that they must have non-zero mass. It is not clear if the small neutrino masses can arise from the same electroweak symmetry breaking mechanism that is true for the other SM particles [3].

These difficulties that the SM faces explain the presence of other, beyond-the-SM (BSM) theories. Some of these scenarios involve having to "bring into existence" other particles that have not yet been found to exist, so the searches of additional heavy scalars are performed. The existence of a sibling Higgs boson is motivated in many BSM scenarios. This particle is thought to be heavier than the SM one, so the research in the full mass range accessible at colliders remains one of the main objectives [3].

Bear in mind that the desired confidence level to be proclaimed a discovery was achieved for 125 GeV Higgs boson, but that doesn't mean that there are no other particles in other mass ranges. It also doesn't mean that there are - it would be useful to know, with 5σ certainty, that there isn't a high mass Higgs boson - so this type of analysis is well worth doing. I will show a study that I performed with the goal to improve it.

4 Data samples comparison

Monte Carlo simulation methods refer to computational algorithms which use randomness (random sampling) to solve numerical problems (even deterministic). They are mostly used in calculations with a probabilistic interpretation, also numerical integration (with complicated boundary conditions etc.) and simulating systems with many degrees of freedom, too difficult to simulate deterministically. Monte Carlo simulations are extensively used for various purposes in modern high-energy physics experiments [6].

Due to the complexity of hadron-hadron collision at LHC, Monte Carlo generators are fundamental to simulate the result of such collisions. It is impossible to predict what happens event-by-event: in fact, in quantum mechanics we can only calculate the probability of having a certain result [3].

For these simulations to be as good as possible, one needs to take into consideration a lot of effects, the physics of the processes and the detector itself. We need to understand each step of the interaction very well, so reconstructing the events is very challenging.

The detectors are not perfect, even though people go at great lengths to try to make them so. In the LHC, the huge amount of data taken per second, in conditions as extreme as they can get (high energies), the material used for the detectors needs to be very carefully and cleverly designed. The amount of radiation that some parts need to endure is very high and, of course, they deteriorate over time. Some subdetectors wear out faster than others. This is why the LHC does not run non-stop. It is divided in data taking periods called Runs, and off periods called long shotdowns. Run 1 took place from 2009 to 2013, then Long shutdown 1 was from 2013 to 2015, Run 2 went on from 2015 to 2018, before Long shutdown 2. Finally, this year, 2022 (just while I did the work for this thesis using Run 2 data), Run 3 has started. During the periods that data is not being taken and the LHC is off, not only are parts being replaced, but improvements are made and certain parts either replaced or calibrated. As I said, a lot of effort is put into optimising the accelerator and the detectors.

During the data taking periods, Runs, the detector tends to lose some of its characteristics over time. For example, the ECAL crystals used for measuring energy lose their transparency.

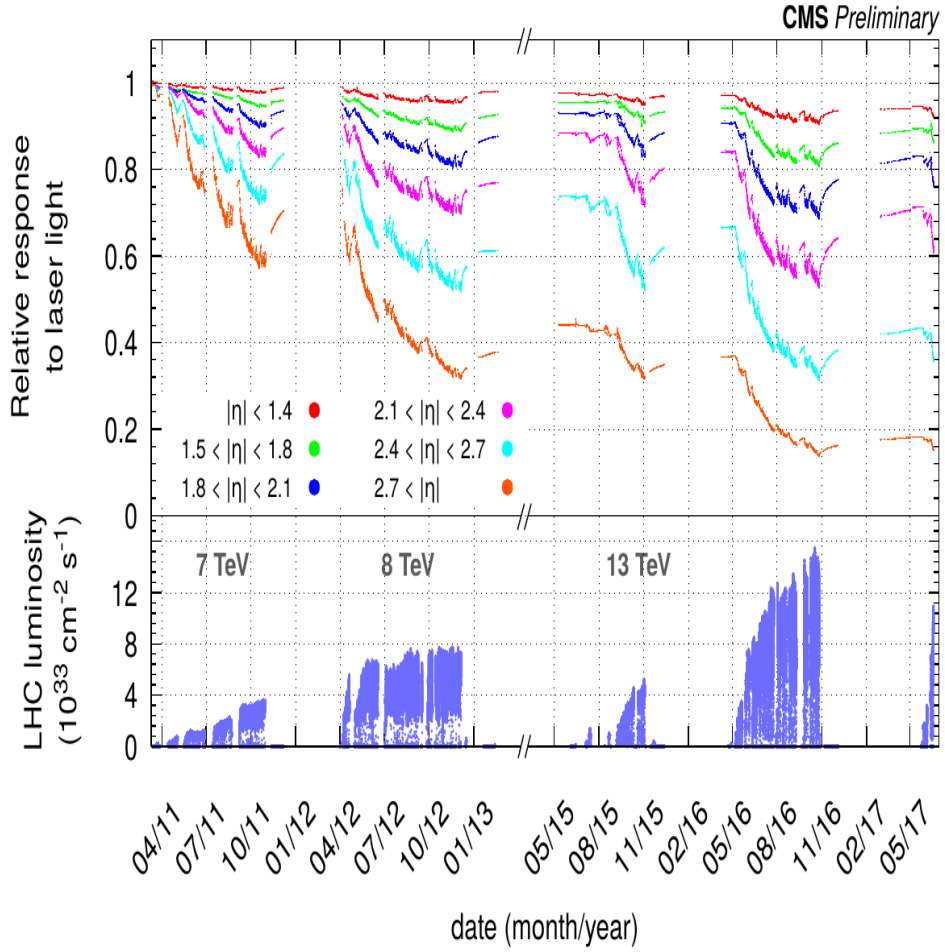


Figure 9: *Time evolution of the ECAL response monitoring laser. The reduction in response during data-taking periods is caused by radiation damage to ECAL crystals [4].*

Figure 8 shows that the response is getting generally worse with time, even though some corrections are made. Still, by the end of the data taking period, the efficiency drops significantly. This absolutely needs to be taken into account when reconstructing the data. One needs to be aware of the fact that the detector is not collecting all the light because the crystal transparency is disturbed, hence the energy calculation needs to be modified.

When a simulation is done, information is stored in large files, namely root files. Here, everything known about the reconstructed particles of the events is stored - mass, momentum, energy, scattering angle... Also, I should mention jets - experimental signatures of quarks and gluons produced in high-energy processes. As quarks and gluons have a net colour charge and cannot exist freely due to colour-confinement, they are not directly observed in nature. Instead, they come together to form colour-neutral hadrons, a process called hadronisation that leads to a collimated spray of hadrons called a jet [10].

To better understand some of the variables, I will shortly describe the reference frame for the measurements.

The standardised coordinate system has an origin at the nominal interaction point, meaning

in the very center of the detector. The x-axis is chosen to point to the center of the accelerator, the y-axis points up, and the z-axis is the proton beam direction. It is a right-handed coordinate system.

It is often more convenient to use the cylindrical coordinates. The transverse plane is the x-y plane, azimuthal angle Φ measured from the x-axis in the x-y plane taking values from -π to π, and the polar angle Θ from the z-axis and it goes from 0 to π.

The particle trajectories are often described in the transverse plane because the activity in this plane is interesting when searching for new phenomena. The transverse momentum is the projection of any momentum onto the x-y plane and often used to denote the magnitude of this projected vector [4].

Pseudorapidity is calculated as $\eta = \frac{1}{2} ln \left(\frac{p+p_z}{p-p_z} \right) = -ln \left(tan \frac{\Theta}{2} \right)$

It is a variable containing information about a place in the detector where a particle was observed. Due to the shape of the detector, we divide it in two elements: central one called barrel, and two opposite endcaps.

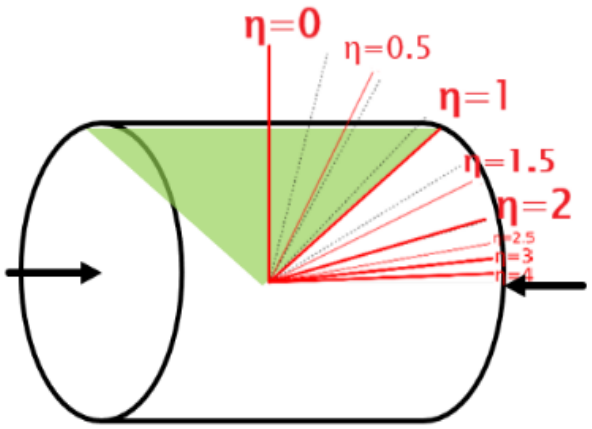


Figure 10: *Pseudorapidity values shown on a cylindrical detector. Taken from [11]*

A scientist that wants to do an analysis has access to numerous variables and can do calculations with them, plot histograms, compute efficiencies and so on. This is what I did - in this chapter I will show a comparison between two data sets via comparing some of these variables.

An improved ECAL calibration for the full Run 2 dataset (2016-2018) was performed during 2019 to achieve optimal performance. The CMS experiment has used these calibration sets for the full Run2 data reprocessing, which has been carried out during Long shutdown 2. This reprocessing constitutes the “Ultra Legacy” data set, which will be used for analyses requiring optimal energy resolution and will be preserved for future analyses on Run 2 data [9].

So the detector’s response changes over time and people need to take care of this. This is

done in phases, where the events are reconstructed again and again. A table below shows this specifically.

Table 1: *Phases of data reconstruction with the appropriate names.*

PHASE	NAME	DETECTOR CONDITIONS
Prompt Reconstruction	PromptReco	Initial detector conditions
Re-Reconstruction	ReReco	Updated detector conditions applied
Re-Re-Reconstruction	UltraLegacy (UL)	Final detector conditions applied, imperfections corrected

So the Ultra Legacy samples are the newest and are expected to be the ones that are the best and are used for the future analyses on the full data. They have been produced both for standard model Higgs boson and high mass, offshell Higgs boson, as well as for background events. The comparison that I made was between ReReco and UL; it has been carried out and proclaimed satisfactory for 125 GeV Higgs; I made the first comparison for high masses. This is important because if we are to use Ultra Legacy samples to do analyses on the Run 2 data, we need to be sure there are no significant differences and unexpected behaviour in the samples due to the calibration. Of course, it is not expected for them to be identical, but an approximate agreement in histogram shapes and efficiencies is what we are going for.

4.1 Method

What does it mean to compare two data sets? One aspect of it is plotting the distributions of various variables in order to visualize them. There are many to choose from, here is the list of the ones I used:

ZZMass (The reconstructed mass of the 4 leptons that came from the two Z bosons. I also plotted this distribution in different channels, with respect to the type of leptons in the final state)

nCleanedJetsPt30 (number of jets with transversal momentum higher than 30 GeV)

JetPt (The transversal momentum, p_{Tjet})

JetEta (The pseudorapidity, η_{jet})

DiJetMass (The mass of the leading 2 jets)

DeltaEta (A constructed variable, the difference in Eta between the two leading jets, $\Delta\eta$)

DeltaPhi (A constructed variable, the difference in Phi, the scattering angle, between the two leading jets, $\Delta\Phi$)

It is also useful to calculate and compare the selection efficiency for UL and ReReco. The relevant information here is, first, the number of events generated per different category. I used the number of generated events in the lepton acceptance for different channels of final states, namely the 4 muons, 4 electrons and 2 muons 2 electrons final states. I am making a certain selection of the reconstructed events, such as putting a threshold on the leptons' momentum, and counting the second relevant quantity, the selected events. Finally, the selection efficiency is defined as the number of selected events over the number of generated events and is calculated for each final state, for each dataset.

4.2 Results

In this section I will show, in figures and tables, the results of my work. I specified the contents of each figure for clarity; whether those are signal or background data and what mass it is done for.

Background events:

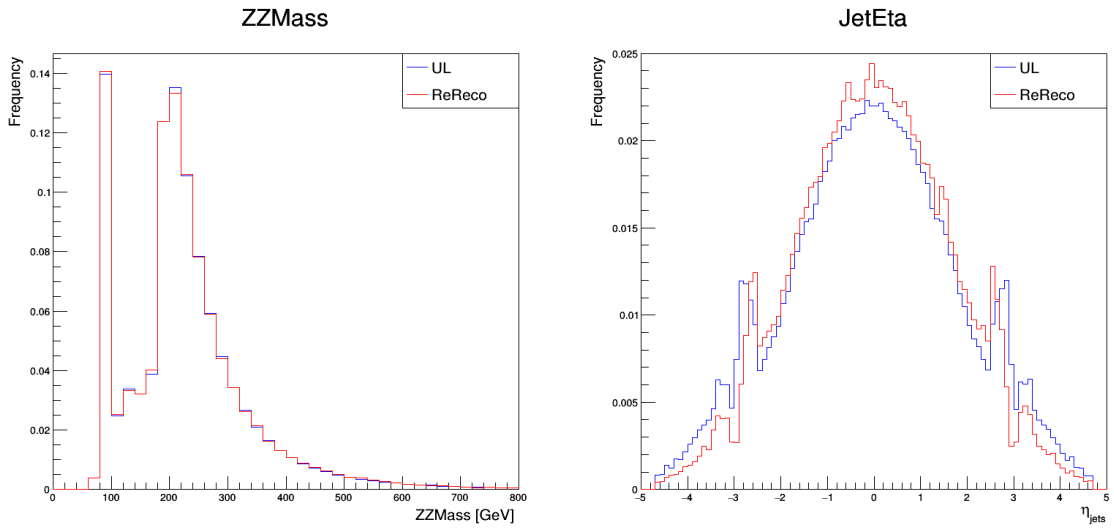


Figure 11: *Results of the UL vs ReReco comparison for backgorund events. Above the histograms are the names of variables, shown also on the x axis (left ZZMass, right JetEta). Every histogram is normalised to 1 (y axis, frequency). Plots of all of the variables can be found in the Appendix A.*

Signal events, ggH mode, mass 1000GeV:

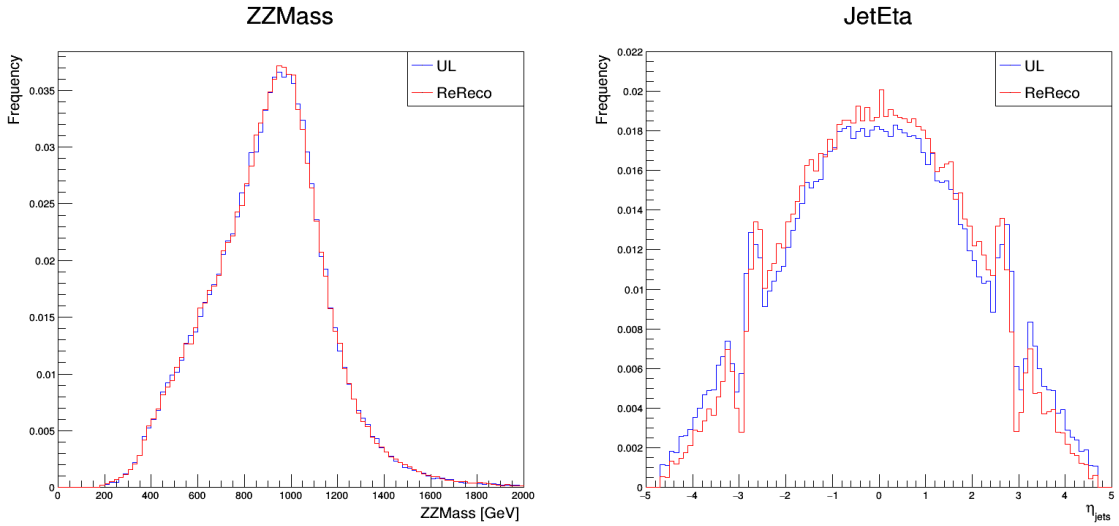


Figure 12: *Results of the UL vs ReReco comparison for signal events, ggH mode, mass 1000 GeV. Above the histograms are the names of variables, shown also on the x axis (left ZZMass, right JetEta). Every histogram is normalised to 1 (y axis, frequency). Plots of all of the variables can be found in the Appendix A.*

Signal events, VBF mode, mass 1000GeV:

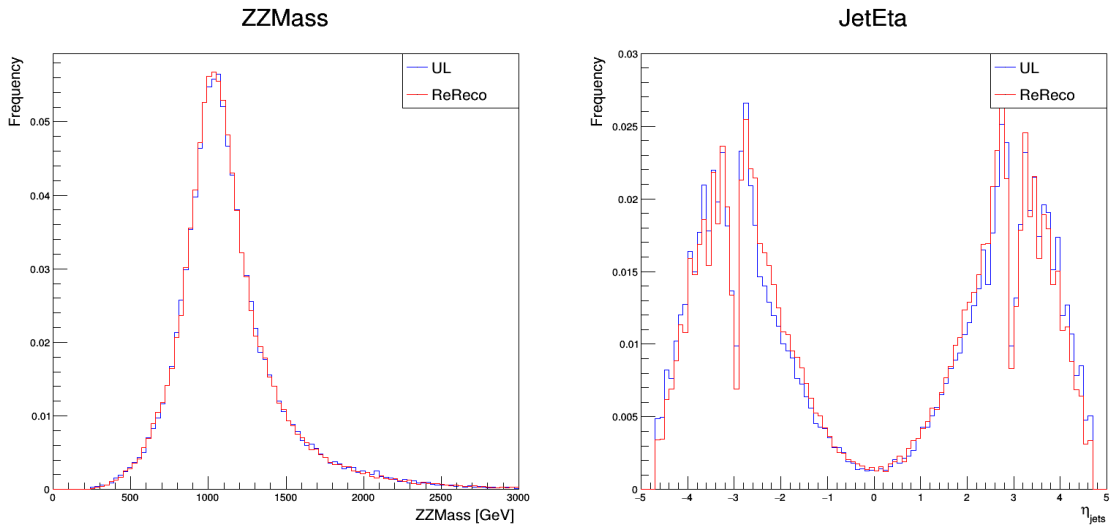


Figure 13: *Results of the UL vs ReReco comparison for signal events, VBF mode, mass 1000 GeV. Above the histograms are the names of variables, shown also on the x axis (left ZZMass, right JetEta). Every histogram is normalised to 1 (y axis, frequency). Plots of all of the variables can be found in the Appendix A.*

Signal events, VBF mode, mass 1500GeV:

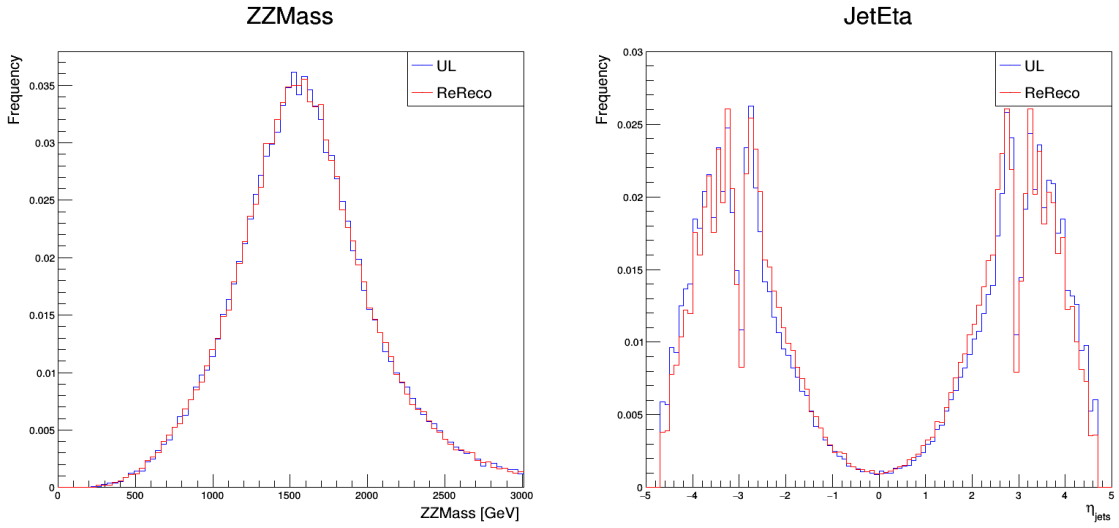


Figure 14: *Results of the UL vs ReReco comparison for signal events, VBF mode, mass 1500 GeV. Above the histograms are the names of variables, shown also on the x axis (left ZZMass, right JetEta). Every histogram is normalised to 1 (y axis, frequency). Plots of all of the variables can be found in the Appendix A.*

Final note regarding the plots: additional checks were made on Standard Model Higgs files as insurance. First I did the comparison for ggH 1000 GeV, then changed the production mode, and finally changed the mass. The results in each case indicated a very good agreement between Ultra Legacy and ReReco.

Next I will show the efficiencies I calculated for each dataset.

Table 2: *Selection efficiency for background files. It is expected that Ultra Legacy has a slightly lower efficiency here.*

Background	UL	ReReco
Efficiency 4e channel	8.5%	10.2%
Efficiency 4μ channel	12.8%	16.4%
Efficiency 2e2μ channel	9.7%	11.6%

Table 3: *Selection efficiency for ggH 1000 GeV files. It is expected that Ultra Legacy has approximately the same efficiency as ReReco.*

ggH 1000 GeV	UL	ReReco
Efficiency 4e channel	86.5%	86.7%
Efficiency 4μ channel	59.0%	59.1%
Efficiency 2e2μ channel	72.0%	71.8%

Table 4: *Selection efficiency for VBFH 1000 GeV files. It is expected that Ultra Legacy has approximately the same efficiency as ReReco.*

VBFH 1000 GeV	UL	ReReco
Efficiency 4e channel	85.9%	85.6%
Efficiency 4μ channel	60.3%	59.1%
Efficiency 2e2μ channel	71.4%	71.7%

Table 5: *Selection efficiency for VBFH 1500 GeV files. It is expected that Ultra Legacy has approximately the same efficiency as ReReco.*

VBFH 1500 GeV	UL	ReReco
Efficiency 4e channel	79.7%	79.7%
Efficiency 4μ channel	58.2%	58.3%
Efficiency 2e2μ channel	68.3%	68.2%

These results were expected and are of big significance. They represent a confirmation that the Ultra Legacy samples are ready to be used in future high mass Higgs analyses. As previously mentioned, my contribution lies in the fact that this was the first such comparison for high mass search. I would like to stress that the fact that everything went well for Standard Model Higgs did not necessarily imply the same would happen for my study, so it was a success.

5 "Significance of Impact Parameter" cut analysis

When any experiment in particle physics is being done, we want to understand processes which can be described as quite extreme to our human standards in terms of speed, energy, lifetime of particles etc. Some of them are extremely high, some extremely low. In either case, we need to have a way to very precisely detect all particles and their properties, which is not an easy task. The detectors at the LHC are very sophisticated and made better continuously, however, they can not be perfect. There is a lot of thorough tests and conditions applied subsequently to the detector data with a goal of being as sure as possible that, for example, an electron is truly an electron, and a photon is truly a photon, as well as where do they come from. The main challenge is the presence of a sizable amount of fakes, i.e. other objects that pass the reconstruction procedure and are thus considered as lepton candidates. In order to deal with these, one has to implement a set of requirements to reduce the amount of fakes while losing as few as possible real particles [4].

This procedure involves, but is not limited to:

-Setting a limit to the transversal momentum of particles, since for low-p_T ones it can be hard to reliably determine the track and momentum

-To account for detector acceptance, a cut on the pseudorapidity of the particles is applied [4]

-The particles are required to satisfy the primary vertex constrains, where the absolute values of the impact parameter with respect to the primary collision vertex in the transverse plane and in the longitudinal direction must be limited. The 3D impact parameter between the candidate and the primary vertex is defined as the minimal Euclidean distance between the two, marked IP_{3D}. A more robust observable 3D impact parameter significance is constructed using the tracking uncertainty on the impact parameter. For a Standard Model Higgs boson a selection is made:

$$SIP_{3D} = \frac{|IP_{3D}|}{\sigma_{IP_{3D}}} < 4 \tag{5.1}$$

[4]

There is more to it, reconstruction and "cleaning" is a long process, but I will not get deep into it. My focus is on the last point - SIP. The cut that is made, < 4, is a result of optimising it for the analysis of the 125 GeV Higgs boson. Search for High mass Higgs opens a question: could it be beneficiary to change the cut? Allowing particles more distanced from the primary vertex will definitely increase our signal, but can we control our background? This is what I try to answer.

Before getting into the method of my study, let me present a bit more details about what changing a SIP cut means. To understand this, let's go back to the basics - in the LHC, protons

are accelerated and then made to collide at speeds near the speed of light. Every time two bunches of protons pass each other, some of the protons will collide at very high energy to create other particles. The exact places where this happens are called primary vertices. At maximum luminosity more than twenty primary vertices are expected [12]. One can think of a vertex as a spot where an interaction happens. After these primary vertices secondary vertices are created and so on, depending on the particle and whether it decays further into other particles. So at every point that there is a certain interaction, there is a vertex.

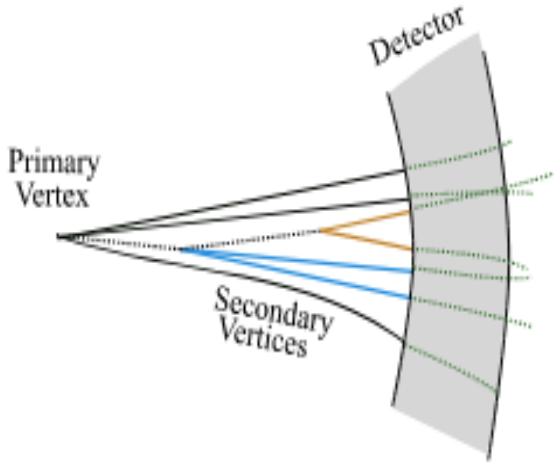


Figure 15: *Primary and secondary vertices. Taken from [13]*

It is not irrelevant how far from a vertex, in the transverse plane and in the longitudinal direction, a particle is found. A 3D impact parameter is defined as the shortest Euclidian distance between the primary vertex and a particle candidate. This is of more importance when a lot of particles are produced and a lot of background events are present as it might become harder to distinguish them. Now, a cut is applied not just to this 3D impact parameter, but to a more complex variable constructed using also the tracking uncertainty, SIP, defined at (5.1), but the idea is the following: if we reduce the cut we get more signal to enter our analysis since we are simply allowing in more distant, less selected events. However, with this, we also get more background events into consideration.

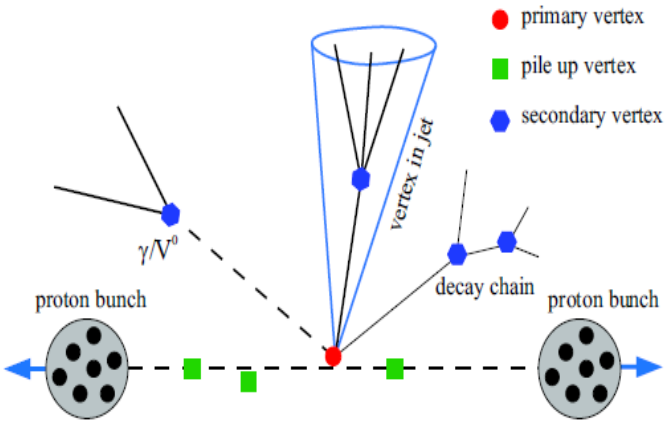


Figure 16: *Aftermath of a proton-proton collision with multiple vertices. Taken from [12]*

For high mass Higgs analysis it becomes a question of finding a balance - is it worth it to reduce the cut to improve the signal, since maybe it would get much better than the cut used for the SM Higgs;or, would the background events simply explode with no visible improvements in the signal region? Note that just having more background, which is inevitable to happen, does not necessarily mean a defeat - maybe this background can be differentiated and subsequently accounted for in the analysis. That said, I will now present the work I did on the matter.

5.1 Method

Here we are dealing with the so-called reducible background (it can be reduced by applying some cuts because it is composed of "fake leptons"), also named "Z+X" (Z is the Z boson and X represents something else that is wrongly reconstructed). I talked about this previously in section 3. So, the important thing that one should differentiate is background leptons that are real, but not interacting in our targeted process and fake leptons that are not real but are detected as such, that can maybe be recognised and rejected afterwards-they are reducible.

How do these events even enter our selection? For instance, if we have in our jet a charged pion (π_+) and a neutral pion (π_0) nearby => π_+ will leave a track, π_0 decays to two photons => ECAL energy deposit => the two combined are recognised as a lepton. We apply several cuts in our analysis to kill the fakes but we don't cut too hard because our signal cross section is low so we try to have a reasonable compromise. Of course, it would be ideal if there was some

This kind of background cannot be estimated by Monte Carlo samples mainly because:

- the probability to fake leptons is very rare so it would require a lot of MC samples to get reasonable statistics at the end (it's rare, but since the cross section is much larger than the signal we are considering, at the end, it still matters)

- the physics behind faking leptons may not be well reproduced

=> at the end, we prefer using the data directly to measure this background ("data-driven

techniques").

The work I did can be summarised in a few steps:

1) Apply different SIP cuts on the leptons

-I changed the SIP cut from < 4 to < 5, < 6, < 7, < 8, < 9 and < 10 respectively, so there are 6 new sets of results.

2) Plot the distributions

3) Compute the selection efficiencies

4) After signal, do the same for Z+X background

-Find the expected number of events.

5) Compare the variation of the signal and the background

-First do a Fit in order to get an interval in which to compare (for a range of masses from 135 to 1500 GeV; use an interval of [mass-σ_i, mass+σ_i], where σ_i is obtained from the fit), since the optimal cut may be mass-dependent. In my study I explored a mass range from 135 to 1500 GeV

-The easiest thing to do would be, for a given mass, plot the ROC curve - on one axis the signal efficiency, on the other axis the background efficiency (or rejection). Here, it is not possible because of the nature of the background data. But we can do something else:

- take the $SIP < 4$ as a reference

- compute the variation (in percentage) of signal and background for every other cut - $\frac{NumberOfEvents(SIP_i) - NumberOfEvents(SIP_4)}{NumberOfEvents(SIP_4)}$, where $i = 5, 6, 7, 8, 9, 10$

- plot the variation of signal vs variation of background

- do this for each mass point (several "low mass" points but just a few high mass since, starting from a certain mass, Z+X should not have much of an effect)

- should be able to see if for some other SIP cut we get a better signal without also increasing the background (finding the best ratio)

6) Calculate also the significance and plot the variance of this significance

-An another way to represent this is to see how the significance evolves. The easiest way to compute it is simply $\frac{S}{\sqrt{B}}$ where S is the number of signal events and B is the number of background events. However, I used a bit more complex one, $\sqrt{2[(S+B)\ln(1+\frac{S}{B}) - S]}$, which, for $S << B$, reduces to $\frac{S}{\sqrt{B}}$ [14].

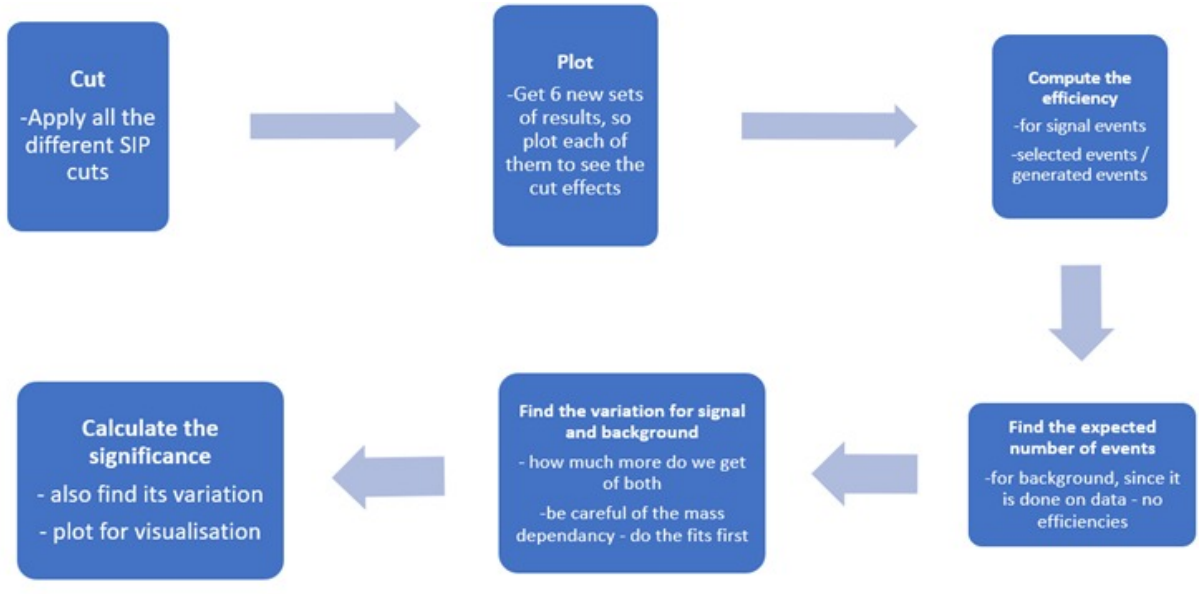


Figure 17: *A graph of the steps of my method for easier tracking.*

5.2 Results

Following the steps in the previous section, here I will show all that I did. Just a note, I am setting the cut on higher numbers, but one is to say that the cut is being reduced, since the condition is being relaxed.

The first result I'll show is the plots. The variables used are the same as the ones mentioned in the previous section. I will show a couple where there is a visible distinction between lines for the SIP cut made on 4,5,...,10.

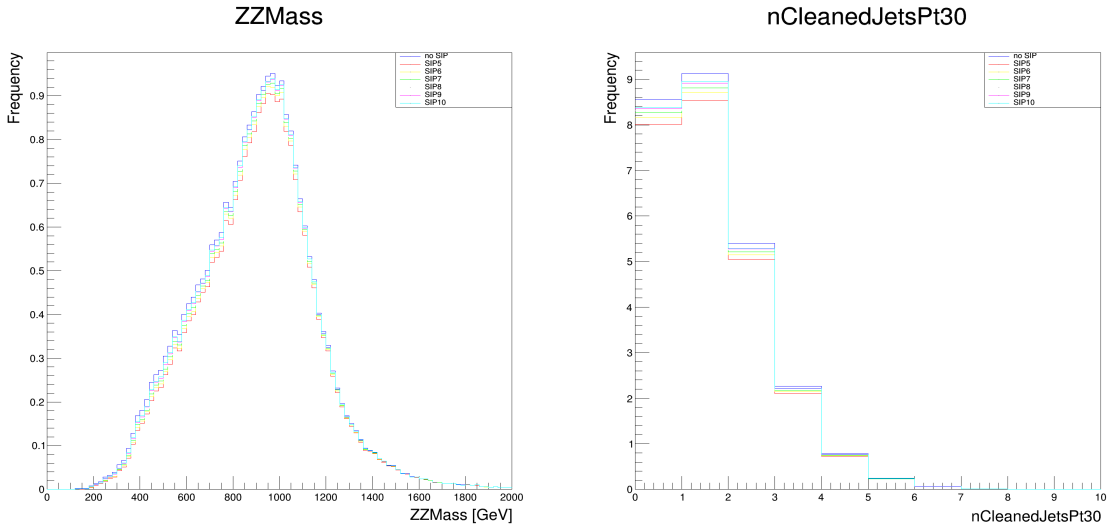


Figure 18: *Comparison when different SIP cuts are applied. The ZZMass (left) is normalised to 1, but the nCleanedJetsPt30 (right) isn't so the increase for reduced cuts can be seen better. These are made for ggH_1000GeV.*

Now, it is important to estimate the Z+X background well. The code for Z+X does several things:

- compute the "fake rate", i.e. the probability that a "loose" lepton passes the "tight" criteria - this is computed in a control region made of Z + exactly 1 loose lepton (CRZL)

- build the control region, made from Z + at least 2 loose leptons, where the fake rate will be applied - the *ZZMass* distribution of this control region (CRZLL) is stored

- apply the fake rate to 2 loose leptons in CRZLL to estimate the Z+X contribution - look at the evolution of these numbers as a function of the SIP cut

More detailed description of this can be found in [4], section 4.5.2.

I mentioned that I performed a Fit in order to get an interval in which I calculated the variation. I used RooFit, instructions I used can be found here [15]. The output of a fit is, among other things, a value of σ.

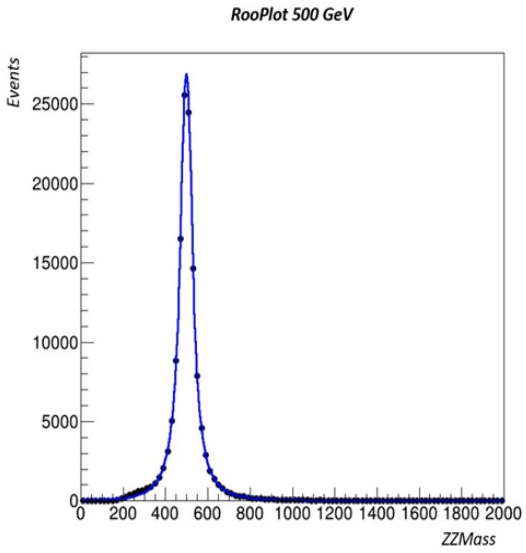


Figure 19: *Example of a RooFit performed for 500 GeV. For higher masses the fit was a bit less precise.*

As I described, I used an interval of $[mass - \sigma, mass + \sigma]$ to find the number of events and calculate the variation. The results are shown in the tables below:

Table 6: *Variation of signal and background for 135 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 135 GeV	0.0356	0.0527	0.0623	0.0682	0.0721	0.0747
Z+X bkg	0.0469	0.0977	0.1407	0.1790	0.2149	0.2436

Table 7: *Variation of signal and background for 1000 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 1000 GeV	0.0395	0.0597	0.0721	0.0800	0.0860	0.0899
Z+X bkg	0.0461	0.0833	0.1150	0.1427	0.1715	0.1914

Other tables for all the mass points can be found in Appendix B.

I made a plot of all these variations, signal vs background. Since the data was in tables, I used Excel, which is why the plots are a bit different than so far.

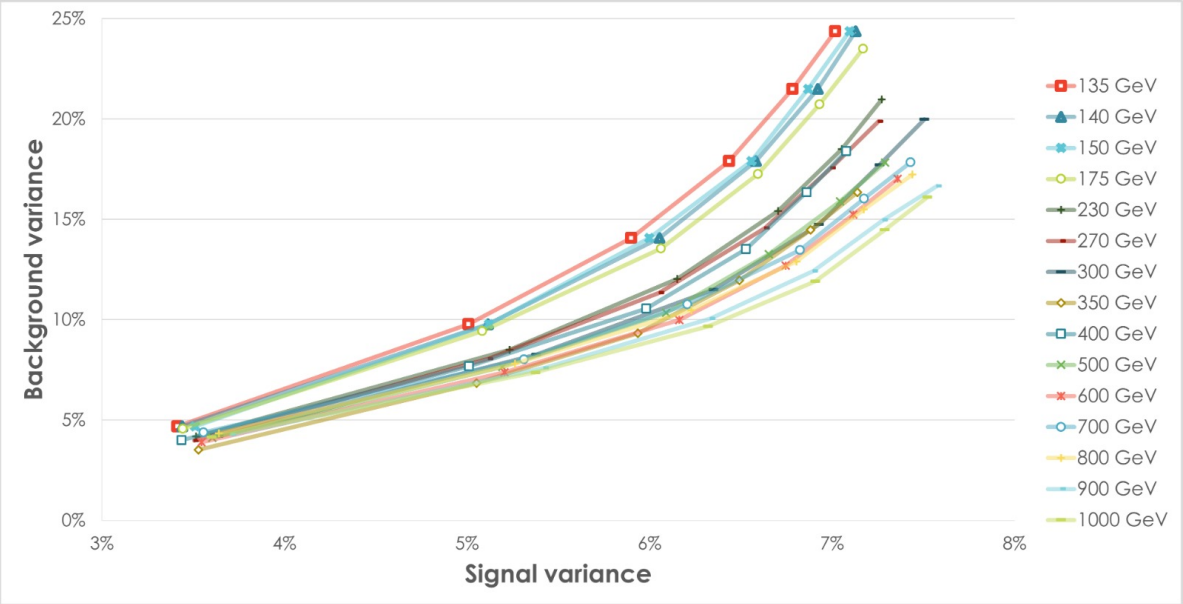


Figure 20: *Signal vs background variation, in percentages.*

Reading the data from the tables or taking a look at the plot, one can easily conclude that with each step reducing the SIP cut we get more increase in background than in signal. However, there are other techniques that can help us dismiss some of this background. That goes to say that this is not enough information to answer the question of the cut.

This is why I calculated the significance, with the aforementioned formula, $\sqrt{2[(S+B)\ln(1+\frac{S}{B})-S]}$. After doing that for all the cuts and all the masses, I also plotted the variation of the significance, with the similar logic to the one explained before. Here is the result:

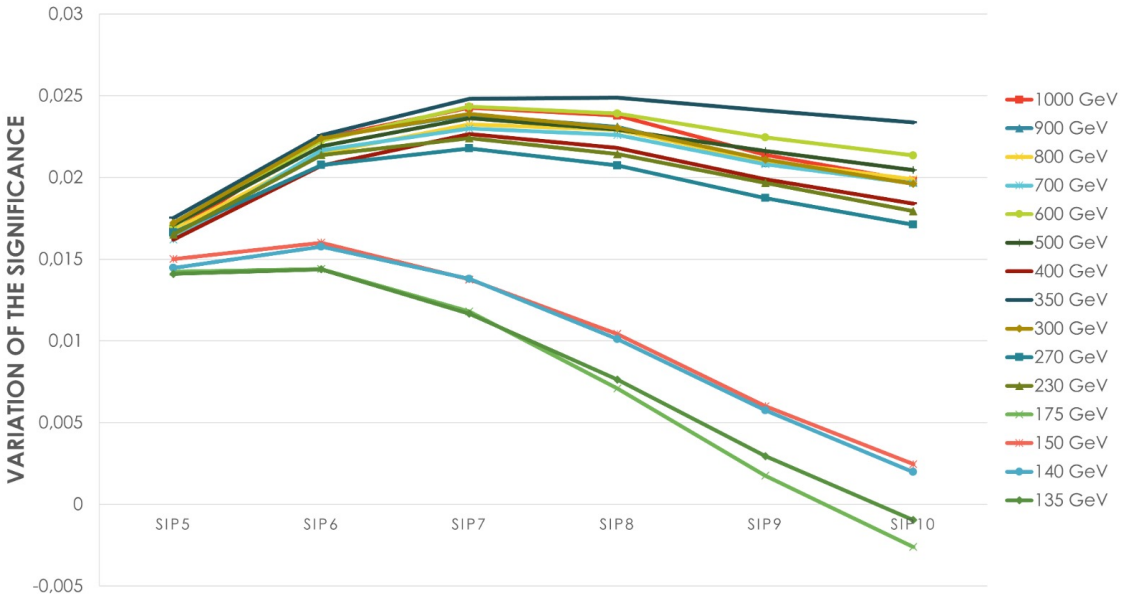


Figure 21: *Variation of the significance.*

The main takeaway from this plot is that there are clearly two "populations", one at the top, the other at the bottom. The four lines at the bottom are for 135, 140, 150 and 175 GeV. This is an indication that the assumption of the mass-dependency of the "ideal" SIP cut from the beginning was right, and the "regular" cut at < 4 may be good for lower masses (meaning higher than 125 GeV, but still lower in comparison to the others in analysis), and for higher masses it might be worth it to reduce the cut.

6 Categorisation

So far, everything I wrote about concerned Higgs' decay. Now, in the last section, I turn to the production of it. The categorisation in point here is referring to determining not the final products of Higgs' interaction, but the way it came to be. There are a lot of production modes, but I will only be talking about the two dominant ones. I have already mentioned them before, ggH and VBF. I will now explain what that means.

ggH basically means gluon + gluon = Higgs. That's really what it is - two gluons fuse into a Higgs boson. Here is a Feynman diagram of this process:

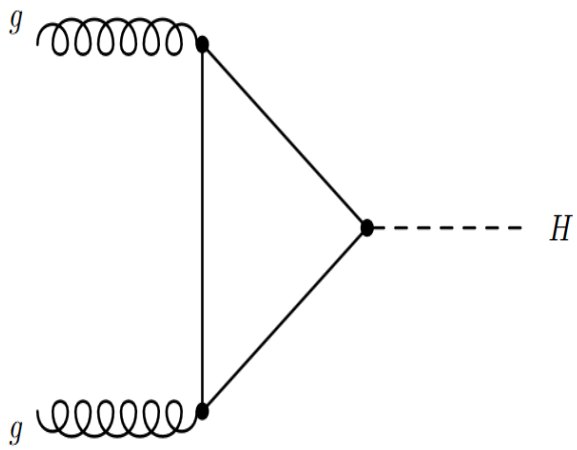


Figure 22: *Gluon-gluon fusion, one of the Higgs boson production modes. Image taken from [16]*

The gluon fusion mode is characterized by a triangle loop of fermions, dominated by heavy quarks (top and, to a lesser extent, b quarks). Measuring ggH helps, if you make some assumptions on the loop content, measuring the coupling of the Higgs boson to the top quark.

Another production mode is VBF which stands for Vector Boson Fusion. It is characterized by two forwards jets and by the fact that the Higgs is produced via the fusion of W or Z bosons; we have thus access to couplings to WW or ZZ.

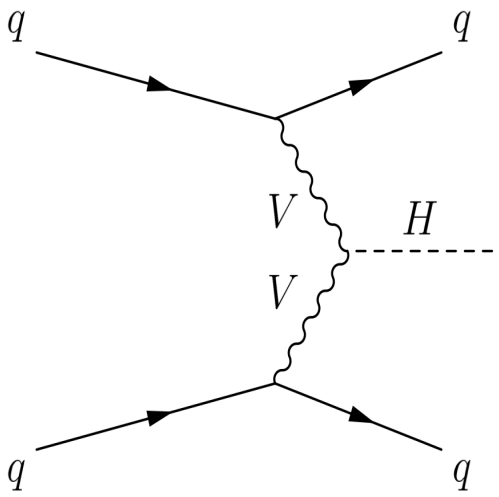


Figure 23: *Vector boson fusion, one of the Higgs boson production modes. Image taken from [16]*

For a Standard Model Higgs boson, both of these are very well analysed and a lot is known.

For the high mass Higgs searches, i.e. the search for new physics resonances, we know we can produce this new particle via ggH or VBF but we don't know the ratio of ggH/VBF (some models will predict only ggH production mode, some only VBF, or a mix of both).

6.1 Method

In the analysis, we focus on selecting the Higgs via its decay products (ZZ $\rightarrow$ 4 leptons). So, we select events where we have any production modes. To separate ggH from VBF, I will use information that is sensitive to the production: mostly the jets. It is visible in figures 20 and 21 that there is a difference in the final state regarding the jets. We expect no jets from a ggH event, and we expect to reconstruct at least one jet from a VBF event. These are the categories that were defined for this study: ggH, VBF1j (1 jet) and VBF2j (2 jets). This is quite a simplified situation, as everything that does not have a jet is simply categorised as ggH-like, and there are cuts on VBF1j and VBF2j events. However, it can give us a rough picture of what is going on.

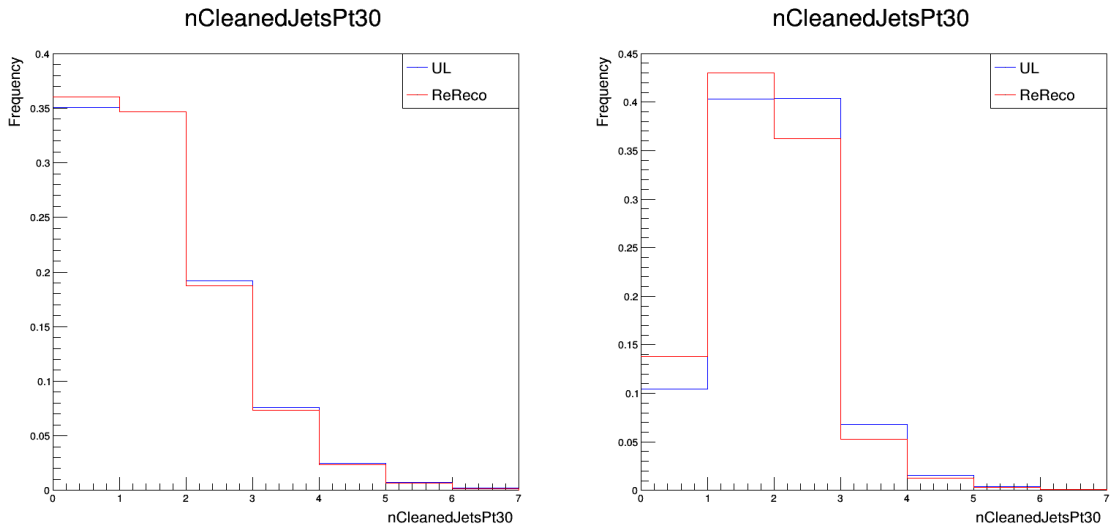


Figure 24: *A difference between ggH (left) and VBF (right) modes regarding the reconstructed jets, both done for 1000 GeV mass point.*

There are variables called "D_VBF1j" and "D_VBF2j" in the framework. They are kinematic discriminants made from Matrix Elements that use the full production information to separate ggH from VBF (combining a lot of variables in a single discriminating one). Instead of forming kinematic discriminants by combining sets of observables with the use of multivariate techniques, an approach based on matrix element calculations was developed. It uses kinematic observables from an event as inputs, and SM Lagrangian to calculate matrix elements that are directly related to the probability of observing an event. This basically means that we are discriminating based on the physical processes themselves, rather than depending on previous computer training [4].

The equations of these discriminants are as follows:

$$D_{VBF2j} = \left[1 + \frac{P_{H+JJ}(\Omega^{H+JJ}|m_{4l})}{P_{VBF}(\Omega^{H+JJ}|m_{4l})}\right]^{-1} \tag{6.1}$$

The denominator is the probability for VBF Higgs production and the numerator is the probability for a ggH + 2 jets production. They are obtained from matrix elements. The Ω denotes kinematics information associated with VBF candidate events described by five angles of the Higgs boson production chain [4].

$$D_{VBF1j} = \left[1 + \frac{P_{H+J}(\Omega^{H+J}|m_{4l})}{\int d_{\eta_J} P_{VBF}(\Omega^{H+JJ}|m_{4l})}\right]^{-1} \tag{6.2}$$

If less than 2 jets are selected, it can be because of jets out of the detector acceptance, not reconstructed or failing the selection requirements. Signal probability can be constructed in events containing exactly one selected jet by simply integrating the probability over the

pseudorapidity of the unobserved jet.

As of now, a single cut is put on these kinematic discriminants. Maybe a cut that depends on the Higgs mass would be optimal?

I took a look at the content of each of the three categories to see what is the percentage of ggH and VBF (1j and 2j) signal, and search how do these percentages evolve with mass. I applied a cut on the discriminant to see if it is a VBF1j or a VBF2j event, and everything else was categorised into ggH. I counted the events and divided, for each, with the total number of events.

Similarly, I will look at the distribution of the discriminating variables and compare the shapes for different masses. Right now, a single cut done for 125 GeV is being used for all of them, and a check is needed to see if there are any problems for higher masses. To be more precise, I evaluated these cuts - the one used for D_VBF1j is made at 0.58442 and for D_VBF2j at 0.46386. I plotted the discriminants for a range of masses from 150 to 3000 GeV.

6.2 Results

Here I will present the percentages of events in each category and show the plots of the discriminants.

The tables and figures below show the evolution (with mass) of the percentage of events in each category calculated for ggH and VBF files respectively.

Table 8: *Percentage of events in each category for 150 GeV. Files are specified on top of columns, categories on the beginning of rows.*

150 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 9: *Percentage of events in each category for 3000 GeV. Files are specified on top of columns, categories on the beginning of rows.*

3000 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

All the other tables can be found in Appendix C.

To make it easier to visualise, I made plots for this:

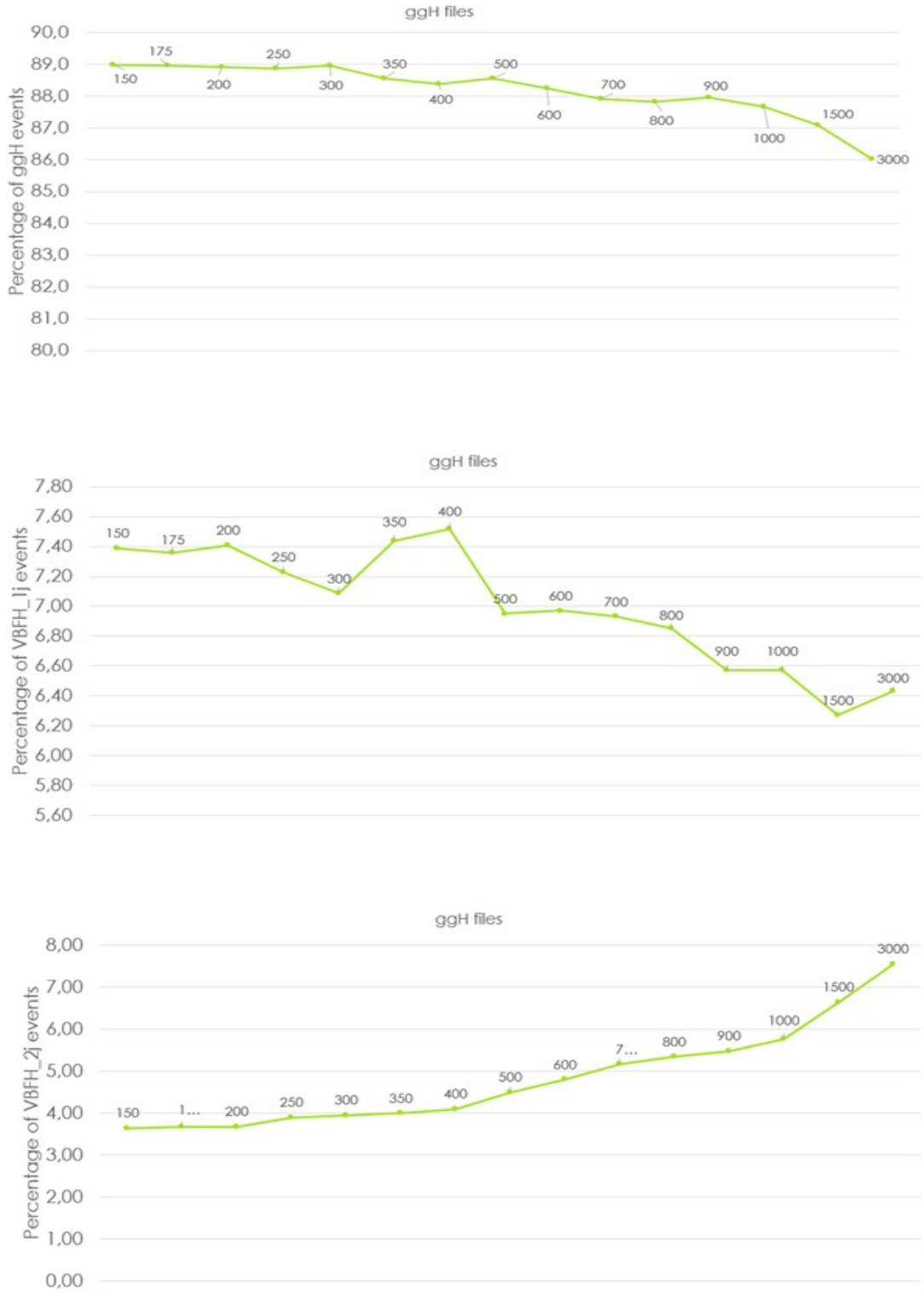


Figure 25: *Evolution of percentage of events in each category for ggH files.*

For ggH files, the difference in percentages is too large between over 85% for ggH and less than 8% for VBF, so I kept them in separate graphs, but for VBF it makes sense to put them together in one:

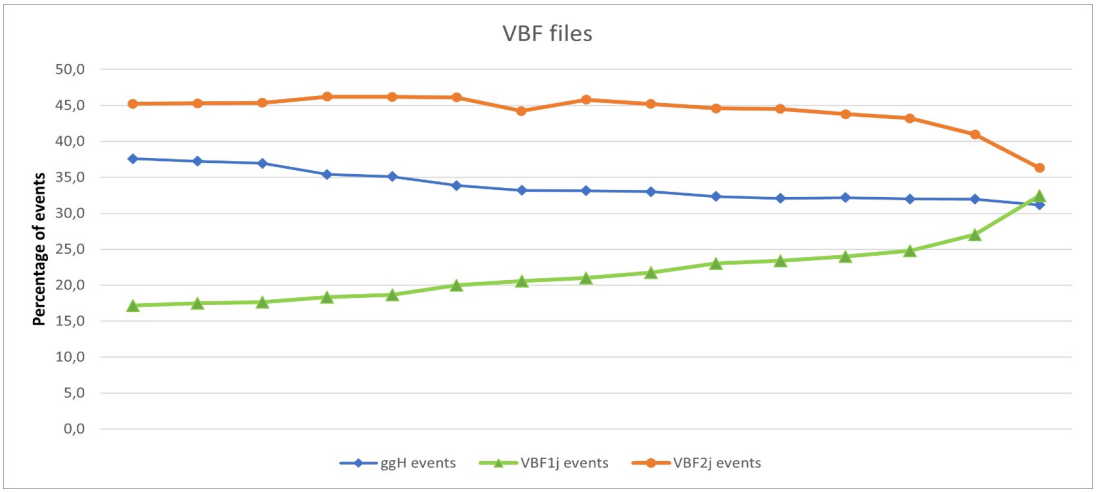


Figure 26: *Evolution of percentage of events in each category for VBF files.*

It is visible that for ggH files the percentage of ggH-like events is very high, although it diminishes a bit on higher masses. VBF1j-like events are fluctuating around 7%, while VBF2j-like events tend to rise in percentage with higher mass. For VBF files, on the other hand, there is a trend of ggH and VBF2j-like events decreasing, and VBF1j-like increasing, to the point where, for 3000 GeV, there is almost the same amount of each category.

Now, let's take a look at the distribution of the kinematic discriminants.

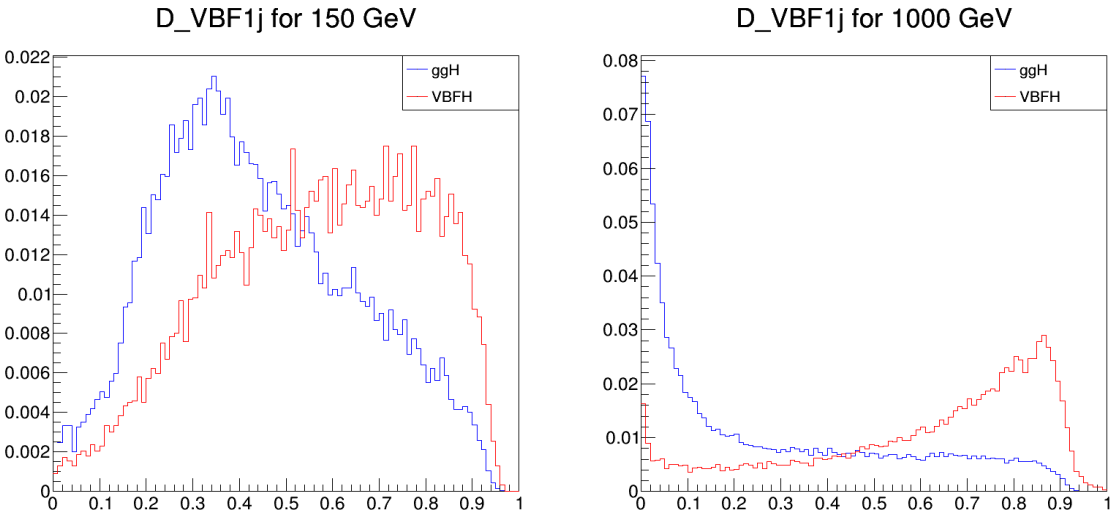


Figure 27: *Discriminating variable D_VBF1j shown for masses 150 (left) and 1000 GeV (right). All are normalised to 1. The rest of the plots can be found in Appendix C.*

As can be seen, the cut for D_VBF2j is good as it is. If you imagine a line dividing any of the plots at the point 0.46386 on the x-axis, it is clear that a good discrimination is achieved. As for D_VBF1j, a cut at 0.58442 is pretty good also, even though one might argue it could be a bit lowered. More detailed analysis is required for a final answer.

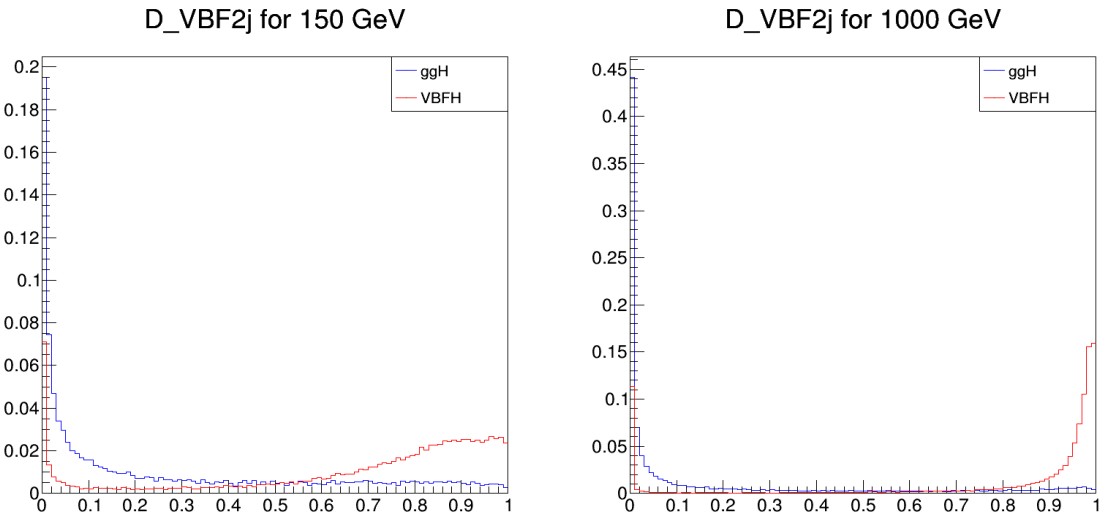


Figure 28: *Discriminating variable D_VBF2j shown for masses 150 (left) and 1000 GeV (right). The rest of the plots can be found in Appendix C.*

A useful thing to do is plot a ROC curve (ggH efficiency vs VBF efficiency) for 125 GeV and for each of the higher masses to get an idea of how similar they look and to try to find a way of getting the same efficiencies for high masses as for the optimised 125 GeV analysis.

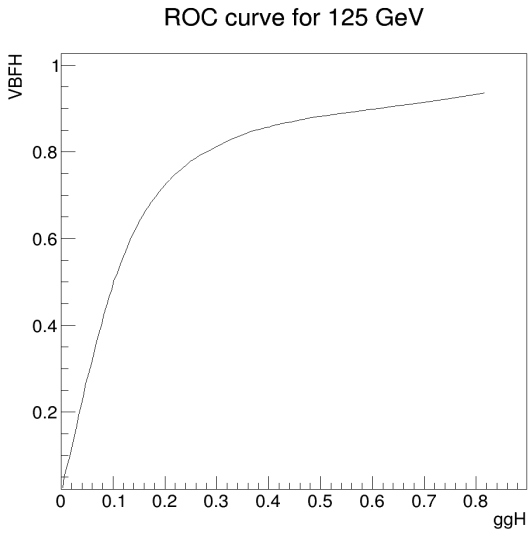


Figure 29: *ROC curve of the D_VBF2j, done for 125 GeV.*

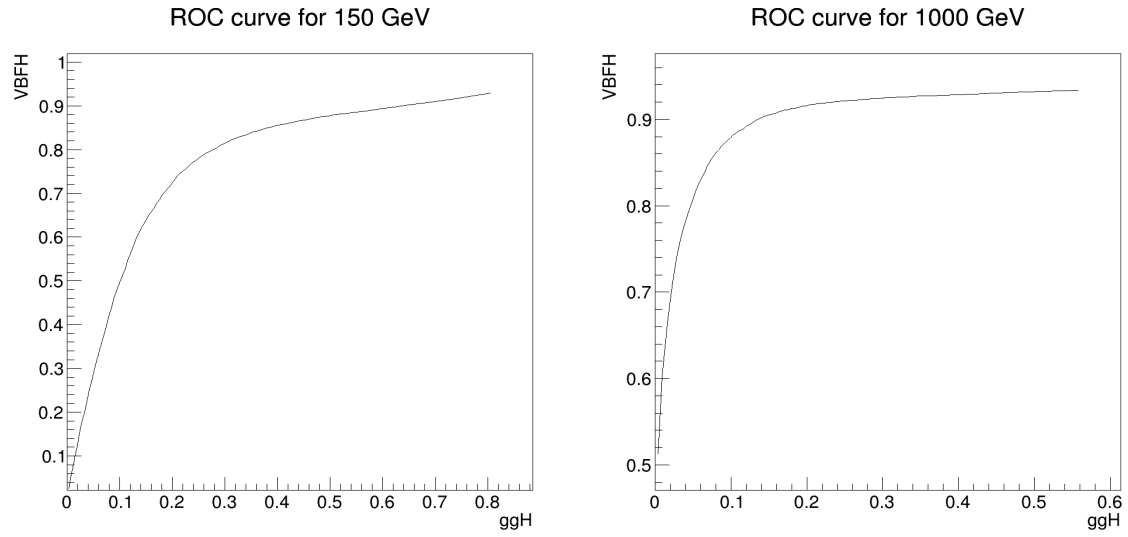


Figure 30: *ROC curves of D_VBF2j for higher masses, 150 (left) and 1000 GeV (right). The rest of the plots can be found in Appendix C.*

If you pay attention to the numbers on the x-axis of the ROC curves, you can see the differences between 125 GeV and higher masses. The idea is that we should have a cut for each mass that has the same efficiency in selecting VBF events with two jets as in the 125 GeV case (with only one jet information we get less separating power). Finding the optimal cut for each mass would improve the high mass analysis.

7 Conclusion

Some ways of possible improvements in the H->4l analysis regarding high mass Higgs boson search is presented. The main idea of the work done for this thesis is to optimise the already existing methods and codes created for Standard Model Higgs analysis in order to get better results in the off-shell analysis, so the recurring theme of the paper is - doing, for a higher mass, something previously done for a lower mass, with appropriate changes.

In the ReReco - UL comparison a good agreement has been found between the two samples. It is significant as this was the first time the process has been done for high mass with a range of both signal and background events showing satisfactory outcomes. This gives additional confidence that the Ultra Legacy for the full Run 2 can be used in research safely.

A SIP cut study was performed trying to determine if the limit value of it, best suited for a lower mass analysis can be altered or, more precisely, increased, in high mass analysis. The final decision, whether it should be done or not, is not stated here, but it is concluded that for a certain mass range it might show some degree of improvement. Possibly a deeper investigation by experts is required for a final resolution.

Finally, the discriminating power of some variables in categorising the events is shown, regarding the production modes of the Higgs boson. Also, the evolution of the percentages of events in each category with respect to the mass is presented. It is in no way a complete study, it is simplified but could still be somewhat useful. The main conclusion is that the behaviour is expected and no major changes are directly found to be needed. Work to be done next may include trying to find, for each mass, the exact cut that should be used to achieve the same efficiencies as the Standard Model Higgs analysis.

8 Bibliography

[1] *CERN page*, URL: https://home.web.cern.ch/about (18. 8. 2022.)

[2] *CERN page*, URL: https://home.cern/news/news/physics/cms-measures-higgs-bosons-mass-unprecedented-precision (18. 8. 2022.)

[3] Lorenzo, R. (2019.) *Search for new resonances in p-p collisions using fully leptonic W+ W- decays with the CMS detector*. Ph.D. thesis. Siena University

[4] Šćulac, T. (2018.) *Measurements of Higgs boson properties in the four-lepton channel in pp collisions at centre-of-mass energy of 13 TeV with the CMS detector*. PhD thesis. International dual doctorate; University of Zagreb, Faculty of Science and University Paris-Saclay, École Polytechnique

[5] Fontanesi, E. (2021.) *Precision measurements of the Higgs boson properties: from the H $\rightarrow$ ZZ* $\rightarrow$ 4l analysis with CMS at the LHC to the future large lepton colliders*. Ph.D. thesis. Università di Bologna

[6] Ehat K. and Veelken C. (2022.) *Stitching Monte Carlo samples*. National Institute for Chemical Physics and Biophysics, Tallinn, Estonia

[7] Dicaire N. (2015.) *Background Estimations in the Higgs to Four Leptons Decay Channel* CERN Summer Student 2015 Report https://cds.cern.ch/record/2053731/files/DicaireReport2015_CERN.pdf

[8] Boselli, S. (2016.) *Theoretical predictions for Higgs boson decay into four leptons*. Ph.D. thesis. University of Pavia

[9] Cavallari, F. and Rovelli, C. (on behalf of the CMS Collaboration) (2019.) *Calibration and Performance of the CMS Electromagnetic Calorimeter in LHC Run2*. Online article.

[10] *CERN page*, URL: https://cms.cern/news/jets-cms-and-determination-their-energy-scale (9.9. 2022.)

[11] Peck A. (2020.) *Photoproduction of η_c mesons in ultra-peripheral P_B + P_b collisions at $\sqrt{s_{NN}} = 5.02TeV$ at the LHC*. Master thesis. Creighton University

[12] Web page, "Taking a closer look at LHC", URL: https://lhc-closer.es/taking_a_closer_look_at_lhc/0.lhc_p_collisions/idioma/en_GB (9.9.2022.)

[13] Shlomi J. et al. (2021.) *Secondary vertex finding in jets with neural networks*. Regular article - "The European physical journal C", Springer

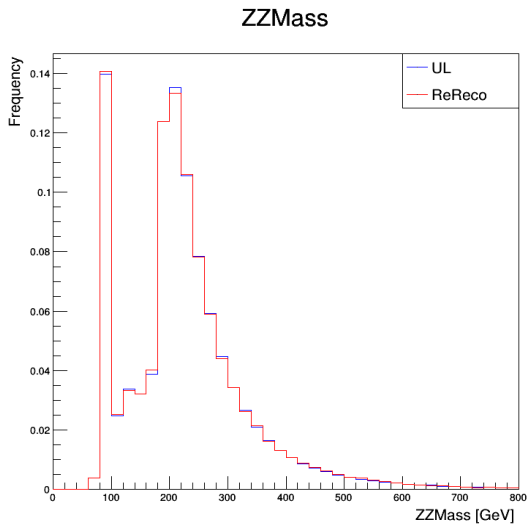
[14] Cowan, G. (2016.) *Some Statistical Tools for Particle Physics* MPI Seminar / Statistics for Particle Physics https://www.pp.rhul.ac.uk/~cowan/stat/cowan_munich16.pdf

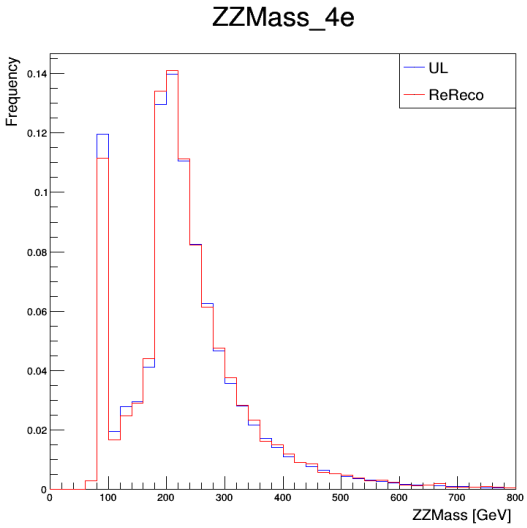
[15] *GitHub page*, URL: https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/part5/roofit/ (20.5.2022.)

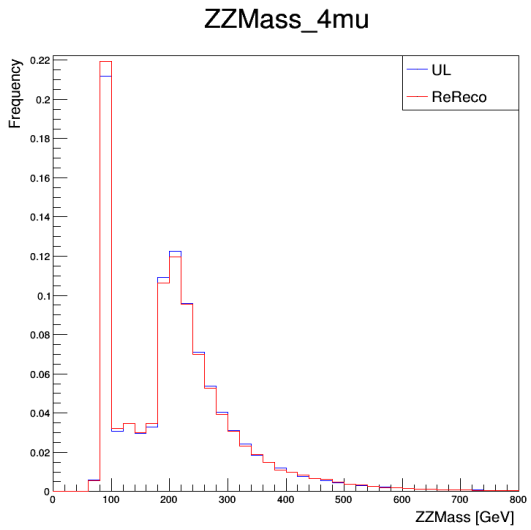
[16] *ATLAS page*, URL: https://atlas.cern/glossary (31.8.2022.)

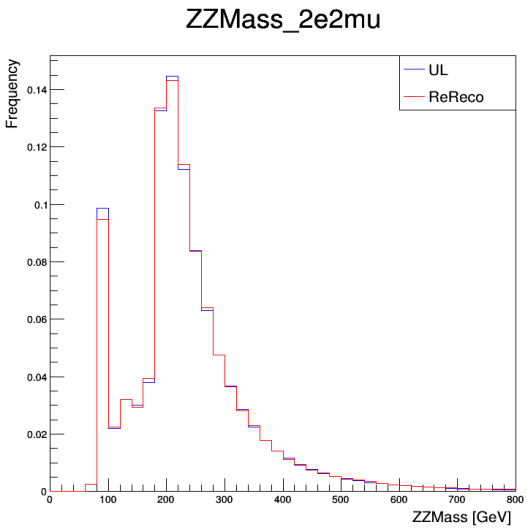
A UL vs ReReco

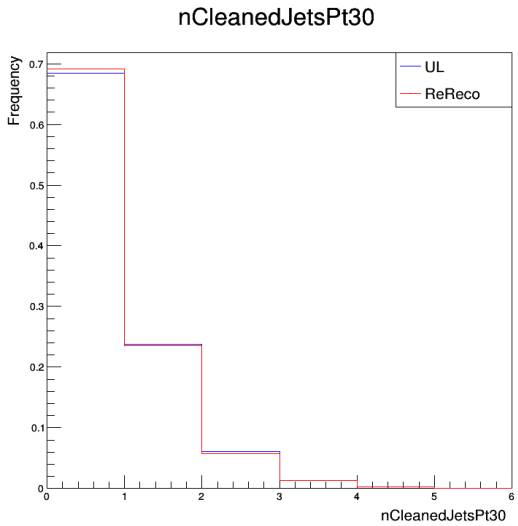
Background events:

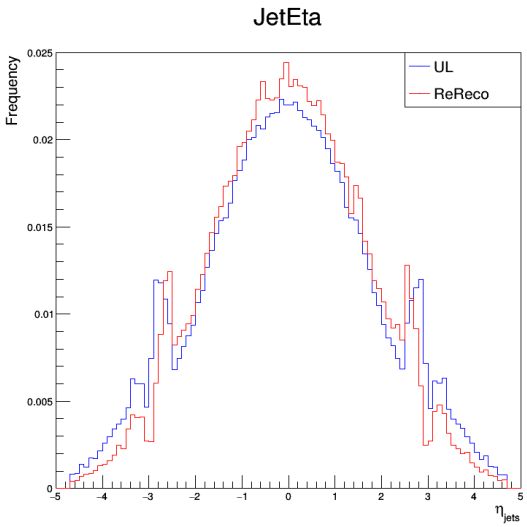


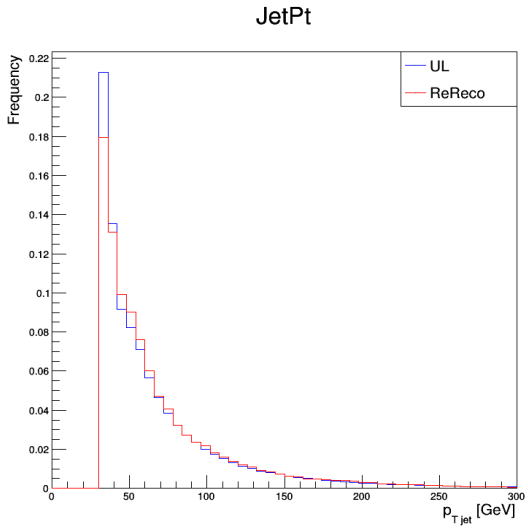


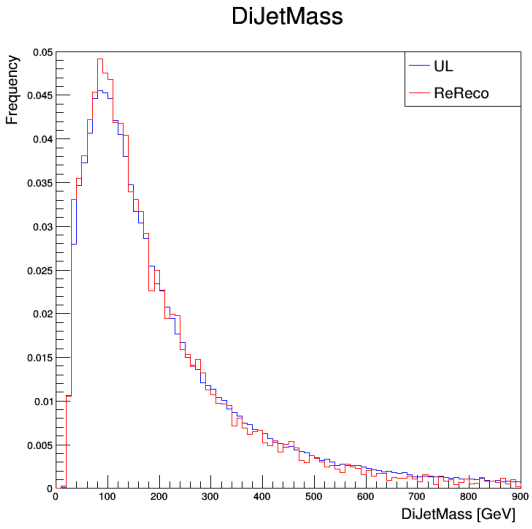


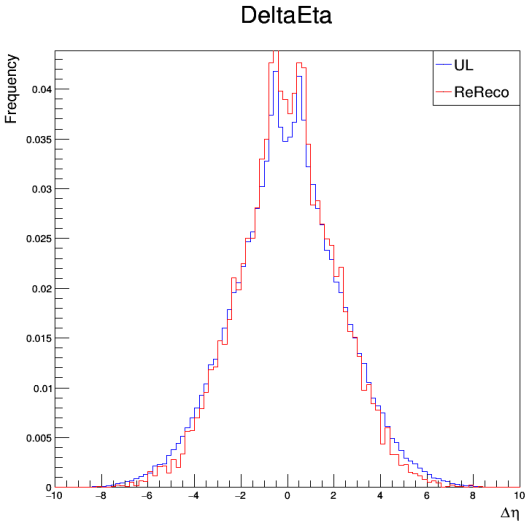


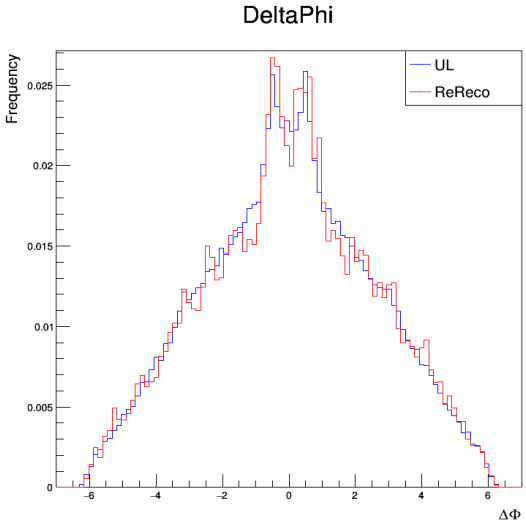

nCleanedJetsPt30
Frequency
UL
ReReco
nCleanedJetsPt30


JetEta
Frequency
UL
ReReco
η jets

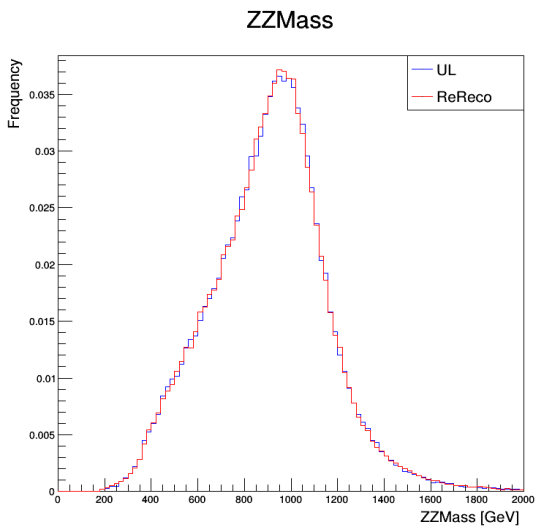

JetPt
Frequency
UL
ReReco
p T jet [GeV]

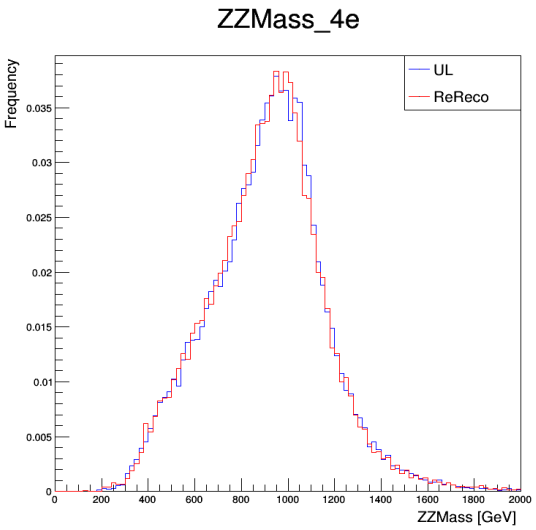

DiJetMass
Frequency
UL
ReReco
DiJetMass [GeV]

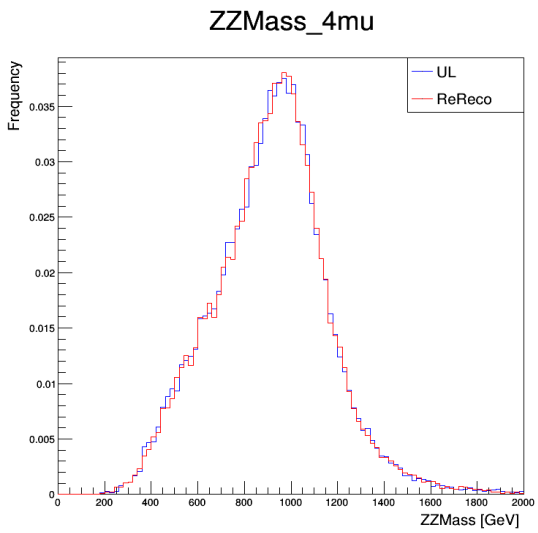

DeltaEta
Frequency
UL
ReReco
Δη

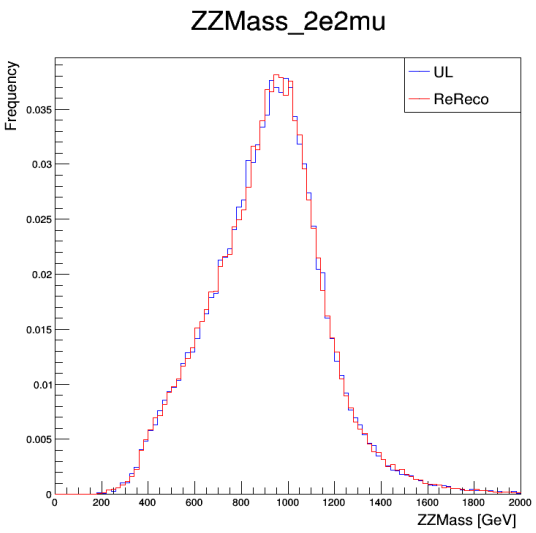

DeltaPhi
Frequency
UL
ReReco
ΔΦ

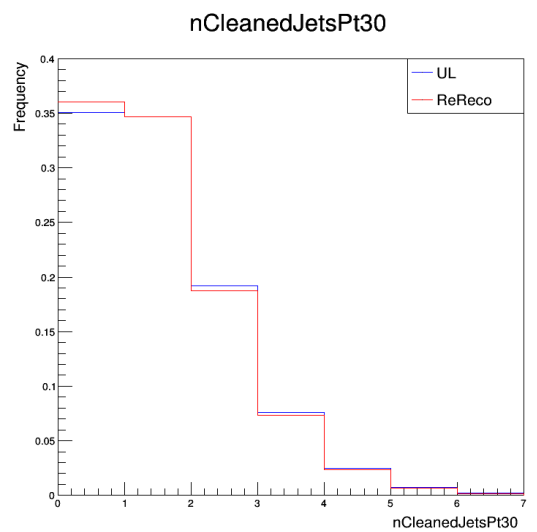
Signal events, ggH mode, mass 1000GeV:

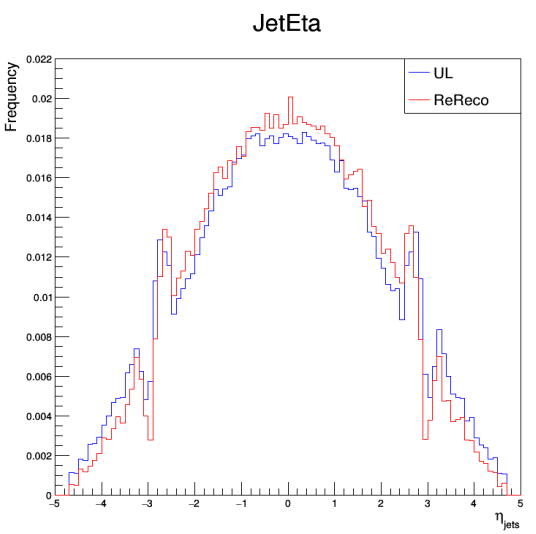


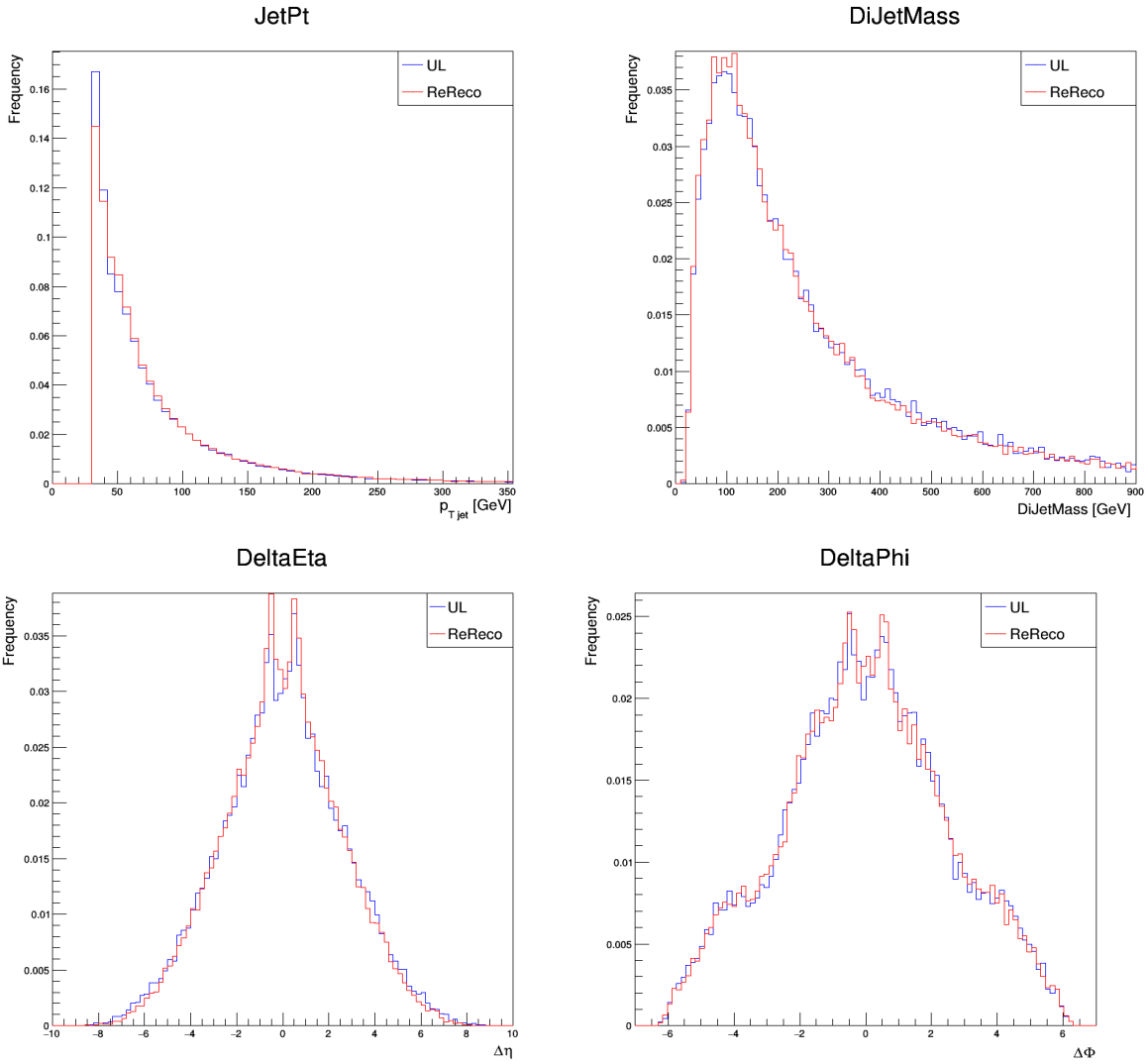




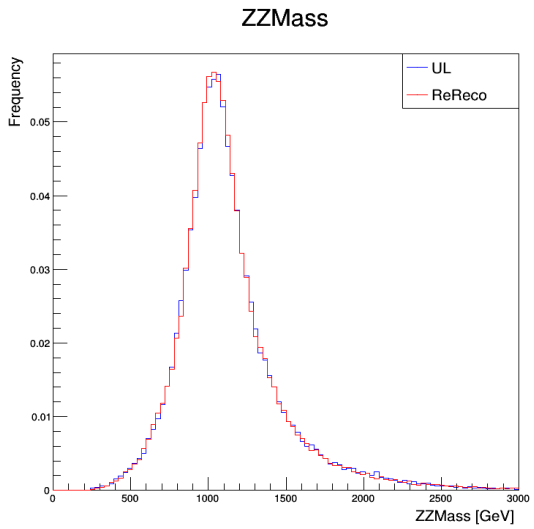


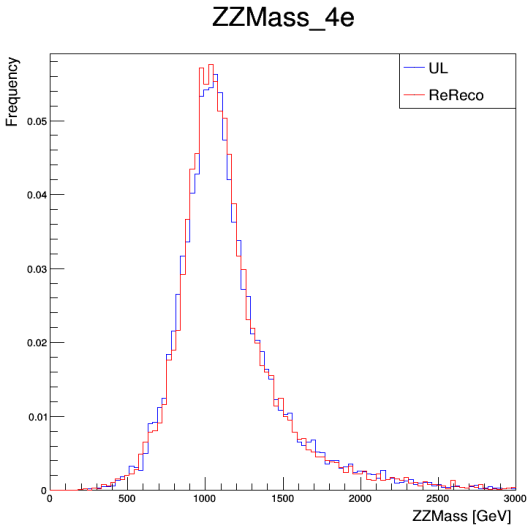


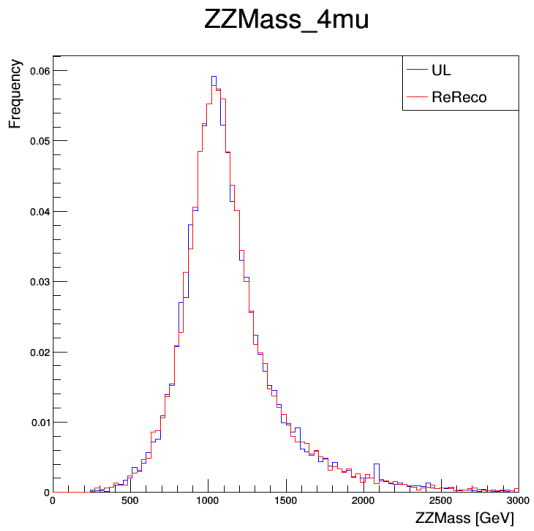


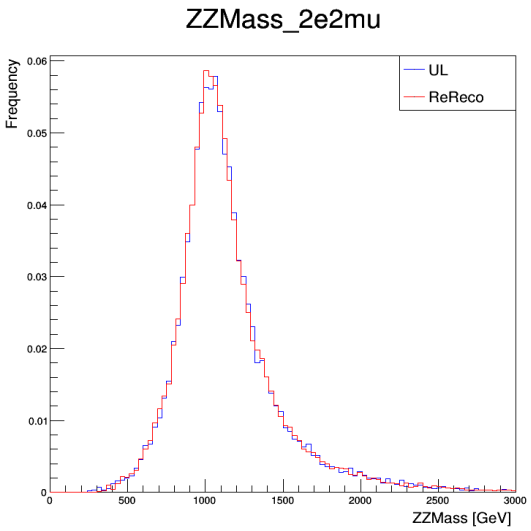


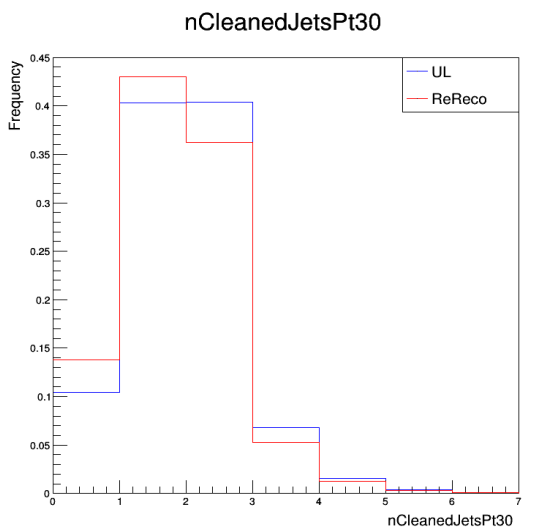
Signal events, VBFH mode, mass 1000 GeV:

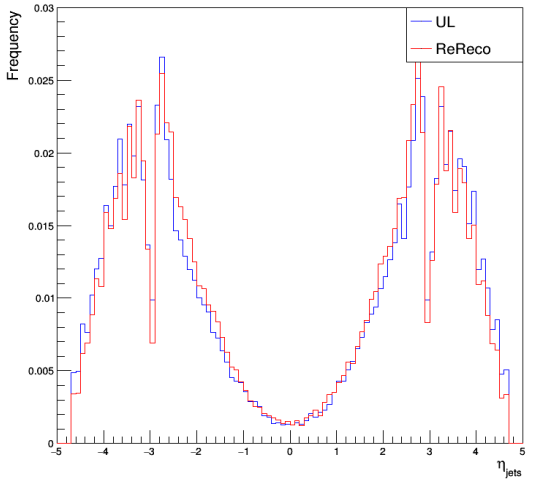

ZZMass
UL
ReReco
Frequency
ZZMass [GeV]

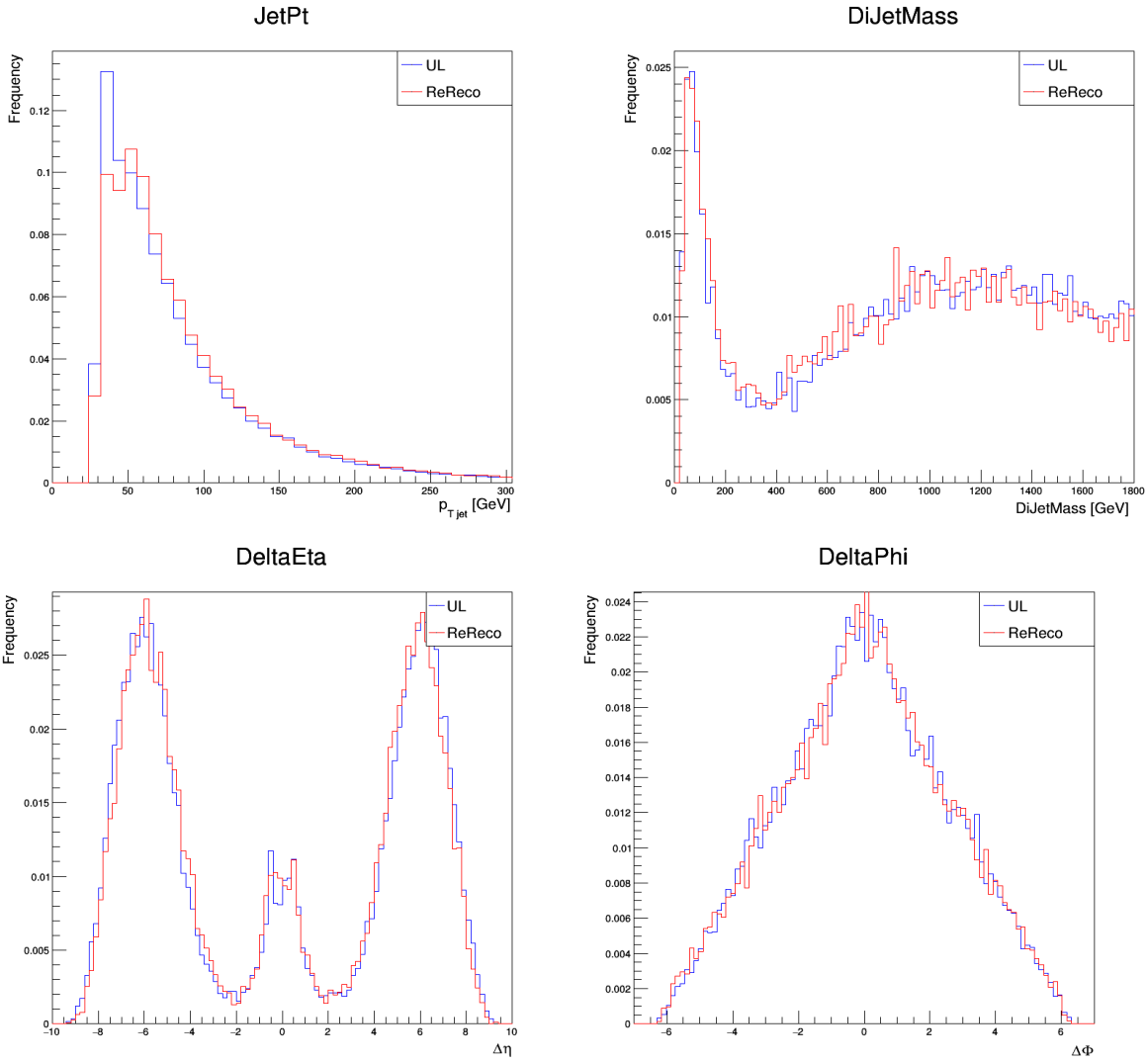

ZZMass_4e
UL
ReReco
Frequency
ZZMass [GeV]


ZZMass_4mu
UL
ReReco
Frequency
ZZMass [GeV]

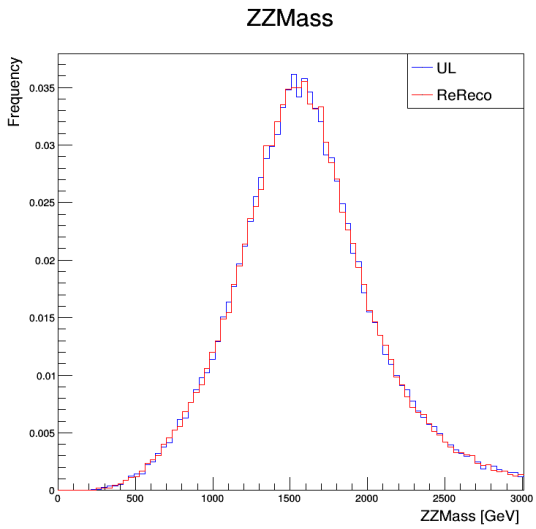

ZZMass_2e2mu
UL
ReReco
Frequency
ZZMass [GeV]

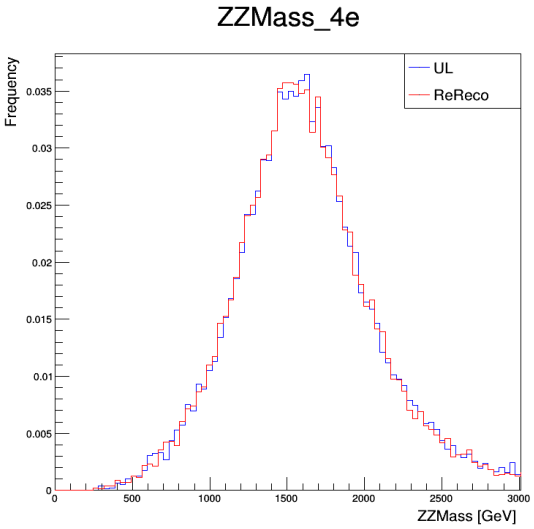

nCleanedJetsPt30
UL
ReReco
Frequency
nCleanedJetsPt30

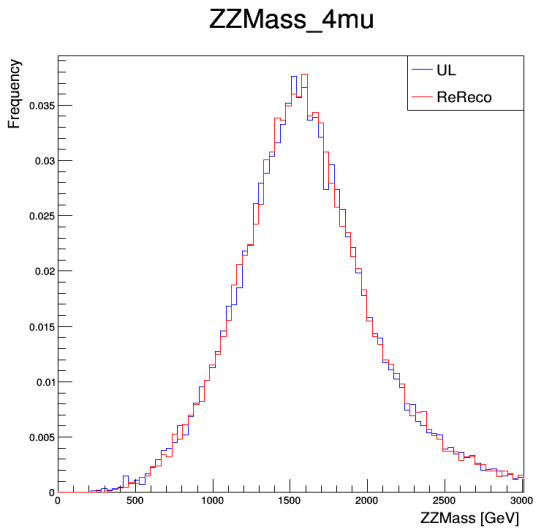

JetEta
UL
ReReco
Frequency
η_jets

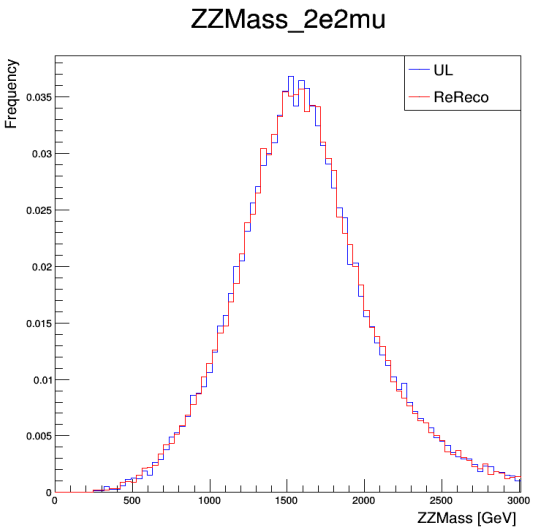


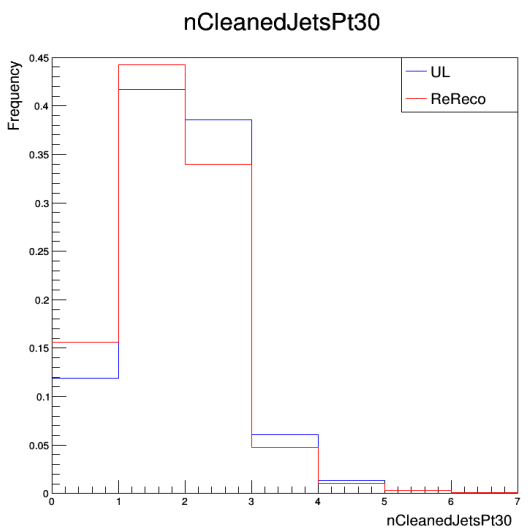
Signal events, VBFH mode, mass 1500 GeV:

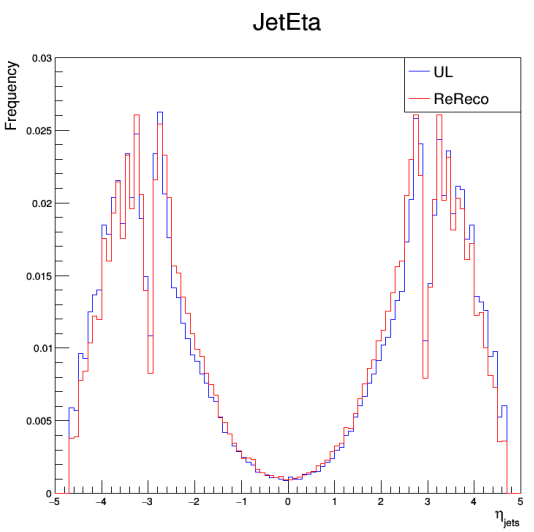

ZZMass
UL
ReReco
Frequency
ZZMass [GeV]

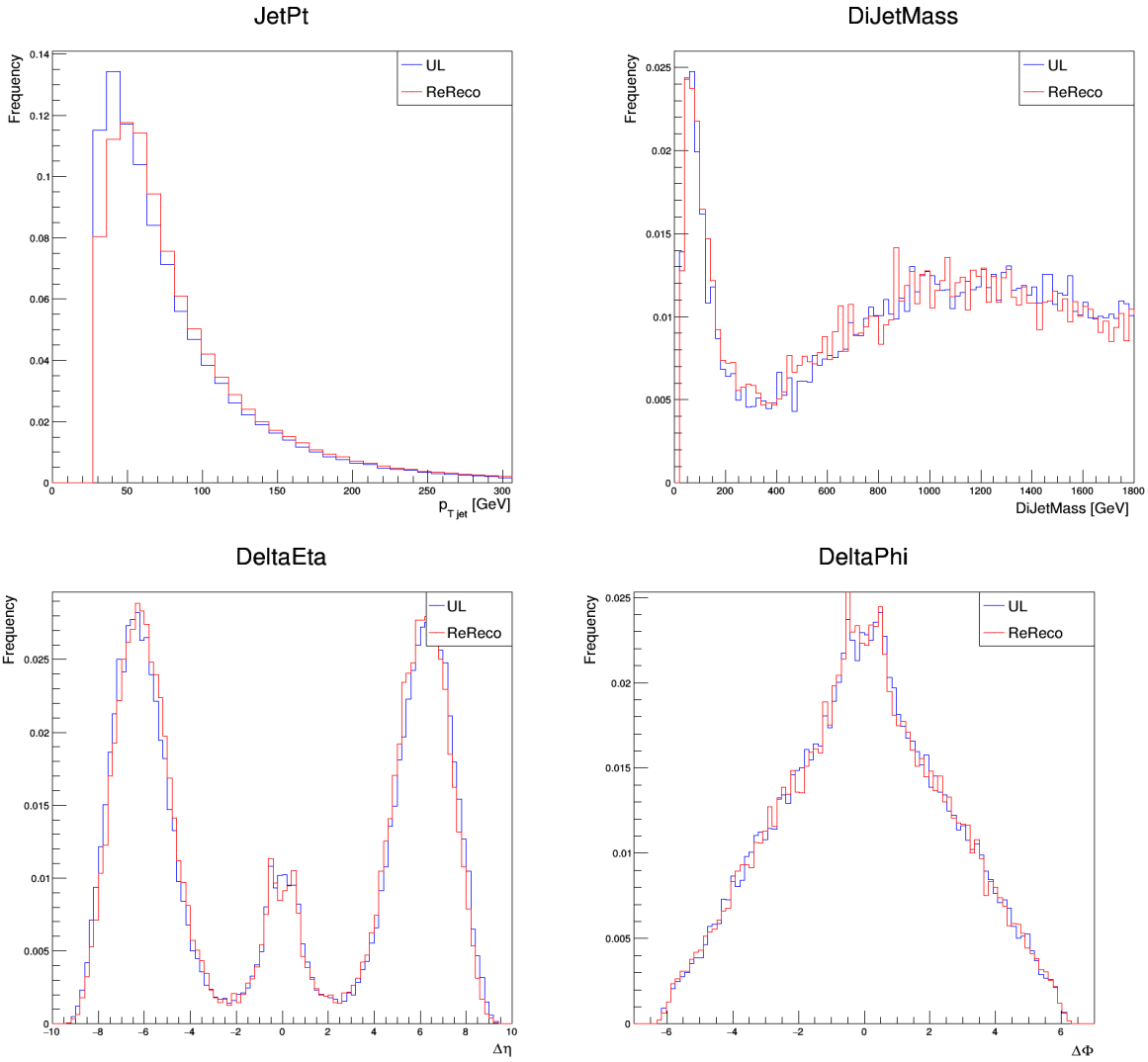

ZZMass_4e
UL
ReReco
Frequency
ZZMass [GeV]


ZZMass_4mu
UL
ReReco
Frequency
ZZMass [GeV]


ZZMass_2e2mu
UL
ReReco
Frequency
ZZMass [GeV]


nCleanedJetsPt30
UL
ReReco
Frequency
nCleanedJetsPt30


JetEta
UL
ReReco
Frequency
η_jets


JetPt
Frequency
UL
ReReco
$p_{T\,jet}$ [GeV]
DiJetMass
Frequency
UL
ReReco
DiJetMass [GeV]
DeltaEta
Frequency
UL
ReReco
Δη
DeltaPhi
Frequency
UL
ReReco
ΔΦ

B SIP

Tables of efficiencies:

Table 10: *Variation of signal and background for 135 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 135 GeV	0.0356	0.0527	0.0623	0.0682	0.0721	0.0747
Z+X bkg	0.0469	0.0977	0.1407	0.1790	0.2149	0.2436

Table 11: *Variation of signal and background for 140 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 140 GeV	0.0357	0.0536	0.0636	0.0694	0.0732	0.0755
Z+X bkg	0.0469	0.0977	0.1407	0.1790	0.2149	0.2436

Table 12: *Variation of signal and background for 145 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 145 GeV	0.0359	0.0534	0.0637	0.0699	0.0741	0.0769
Z+X bkg	0.0469	0.0977	0.1407	0.1790	0.2149	0.2436

Table 13: *Variation of signal and background for 150 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 150 GeV	0.0376	0.0553	0.0653	0.0716	0.0753	0.0780
Z+X bkg	0.0469	0.0977	0.1407	0.1790	0.2149	0.2436

Table 14: *Variation of signal and background for 175 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 175 GeV	0.0377	0.0566	0.0683	0.0748	0.0794	0.0826
Z+X bkg	0.0457	0.0943	0.1354	0.1726	0.2072	0.2350

Table 15: *Variation of signal and background for 230 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 230 GeV	0.0386	0.0582	0.0699	0.0767	0.0814	0.0846
Z+X bkg	0.0418	0.0851	0.1203	0.1541	0.1850	0.2097

Table 16: *Variation of signal and background for 270 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 270 GeV	0.0394	0.0591	0.0709	0.0782	0.0834	0.0872
Z+X bkg	0.0399	0.0805	0.1135	0.1457	0.1755	0.1987

Table 17: *Variation of signal and background for 300 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 300 GeV	0.0399	0.0599	0.0721	0.0797	0.0845	0.0884
Z+X bkg	0.0419	0.0826	0.1150	0.1474	0.1770	0.1999

Table 18: *Variation of signal and background for 350 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 350 GeV	0.0407	0.0604	0.0722	0.0799	0.0854	0.0892
Z+X bkg	0.0351	0.0684	0.933	0.1196	0.1447	0.1634

Table 19: *Variation of signal and background for 400 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 400 GeV	0.0401	0.0595	0.0718	0.0796	0.0849	0.0886
Z+X bkg	0.0400	0.0772	0.1072	0.1373	0.1652	0.1863

Table 20: *Variation of signal and background for 450 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 450 GeV	0.0400	0.0600	0.0724	0.0799	0.0852	0.0890
Z+X bkg	0.0442	0.0831	0.1134	0.1454	0.1730	0.1946

Table 21: *Variation of signal and background for 500 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 500 GeV	0.0407	0.0598	0.0719	0.0798	0.0855	0.0895
Z+X bkg	0.0388	0.0729	0.978	0.1240	0.1488	0.1661

Table 22: *Variation of signal and background for 600 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 600 GeV	0.0394	0.0599	0.0722	0.0801	0.0857	0.0898
Z+X bkg	0.0352	0.0719	0.0980	0.1262	0.1513	0.1707

.

Table 23: *Variation of signal and background for 700 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 700 GeV	0.0401	0.0605	0.0725	0.0808	0.0865	0.0905
Z+X bkg	0.0474	0.0818	0.1099	0.1348	0.1587	0.1757

Table 24: *Variation of signal and background for 800 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 800 GeV	0.0405	0.0606	0.0728	0.0808	0.0864	0.0904
Z+X bkg	0.0385	0.0753	0.1047	0.1274	0.1530	0.1703

Table 25: *Variation of signal and background for 900 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 900 GeV	0.0409	0.0613	0.0733	0.0808	0.0863	0.0907
Z+X bkg	0.0294	0.0573	0.0760	0.1039	0.1289	0.1439

Table 26: *Variation of signal and background for 1000 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 1000 GeV	0.0395	0.0597	0.0721	0.0800	0.0860	0.0899
Z+X bkg	0.0461	0.0833	0.1150	0.1427	0.1715	0.1914

Table 27: *Variation of signal and background for 1500 GeV.*

Variance	SIP<5	SIP<6	SIP<7	SIP<8	SIP<9	SIP<10
ggH 1500 GeV	0.0396	0.0592	0.0713	0.0792	0.0853	0.0893
Z+X bkg	0.0453	0.0694	0.0919	0.1222	0.1719	0.2085

C Categorisation

Percentages of events :

Table 28: *Percentage of events in each category for 150 GeV. Files are specified on top of columns, categories on the beginning of rows.*

150 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 29: *Percentage of events in each category for 175 GeV. Files are specified on top of columns, categories on the beginning of rows.*

175 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 30: *Percentage of events in each category for 200 GeV. Files are specified on top of columns, categories on the beginning of rows.*

200 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 31: *Percentage of events in each category for 250 GeV. Files are specified on top of columns, categories on the beginning of rows.*

250 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 32: *Percentage of events in each category for 300 GeV. Files are specified on top of columns, categories on the beginning of rows.*

300 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 33: *Percentage of events in each category for 350 GeV. Files are specified on top of columns, categories on the beginning of rows.*

350 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 34: *Percentage of events in each category for 400 GeV. Files are specified on top of columns, categories on the beginning of rows.*

400 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 35: *Percentage of events in each category for 500 GeV. Files are specified on top of columns, categories on the beginning of rows.*

500 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 36: *Percentage of events in each category for 600 GeV. Files are specified on top of columns, categories on the beginning of rows.*

600 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 37: *Percentage of events in each category for 700 GeV. Files are specified on top of columns, categories on the beginning of rows.*

700 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 38: *Percentage of events in each category for 800 GeV. Files are specified on top of columns, categories on the beginning of rows.*

800 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 39: *Percentage of events in each category for 900 GeV. Files are specified on top of columns, categories on the beginning of rows.*

900 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

Table 40: *Percentage of events in each category for 1000 GeV. Files are specified on top of columns, categories on the beginning of rows.*

1000 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

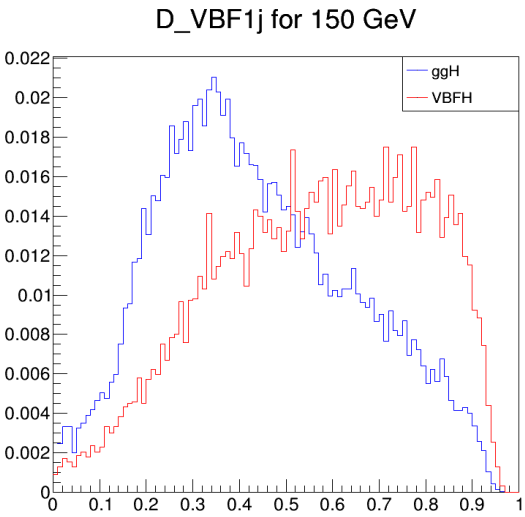
Table 41: *Percentage of events in each category for 1500 GeV. Files are specified on top of columns, categories on the beginning of rows.*

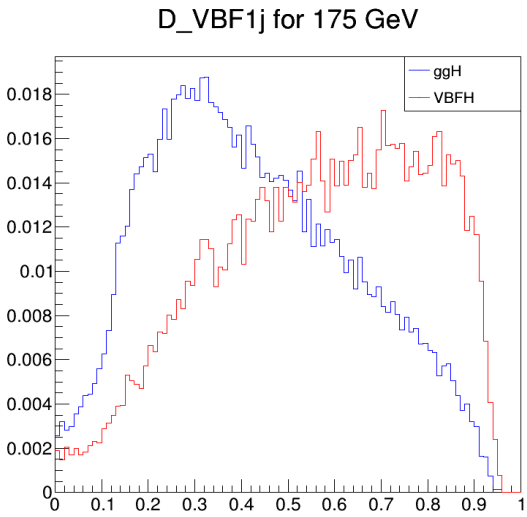
1500 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

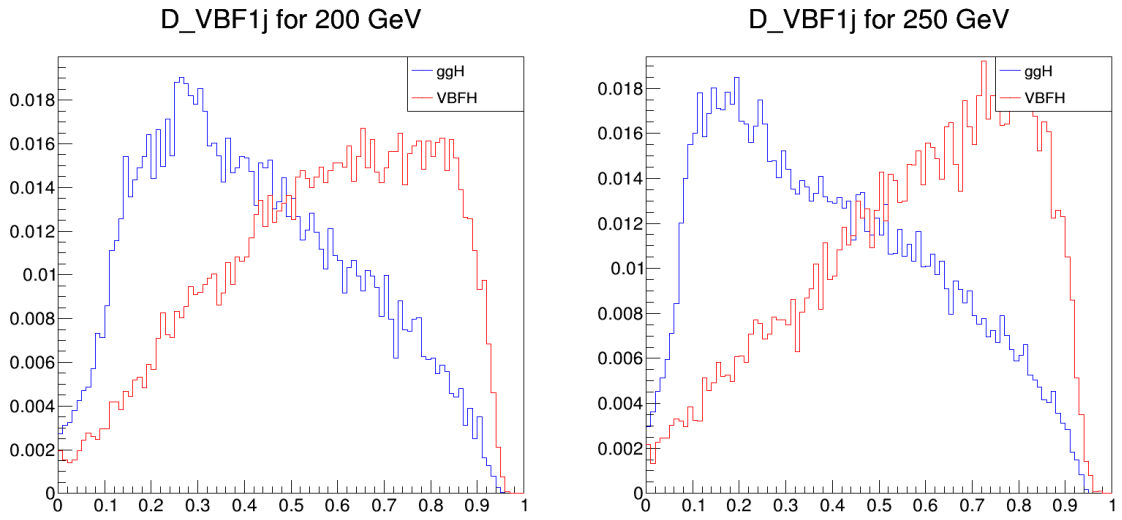
Table 42: *Percentage of events in each category for 3000 GeV. Files are specified on top of columns, categories on the beginning of rows.*

3000 GeV	ggH	VBF
ggH	88.97%	37.59%
VBF 1 jet	7.39%	17.17%
VBF 2 jets	3.64%	45.24%

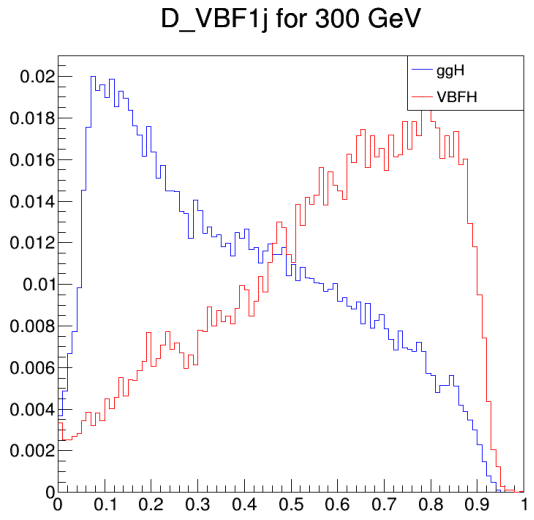
Kinematic discriminants:

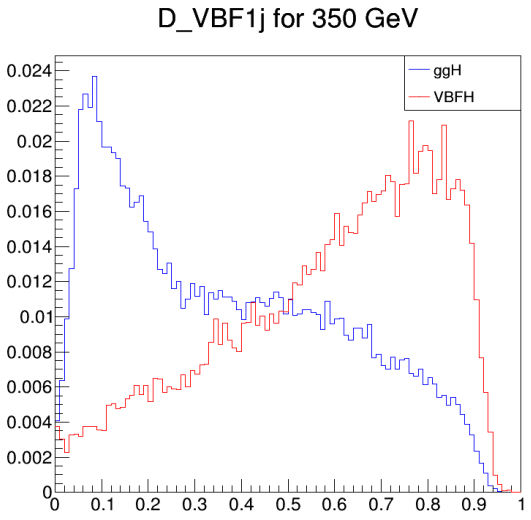


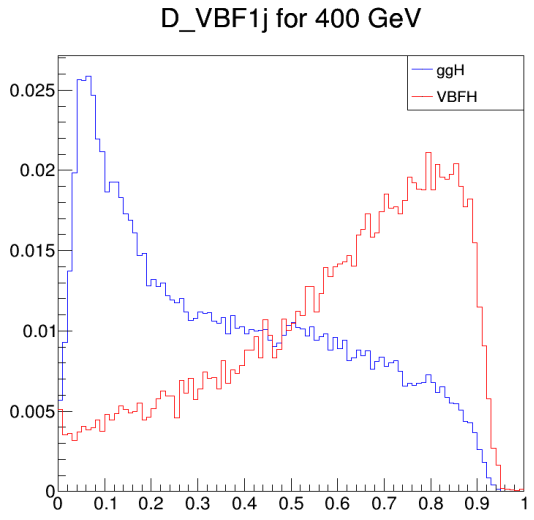


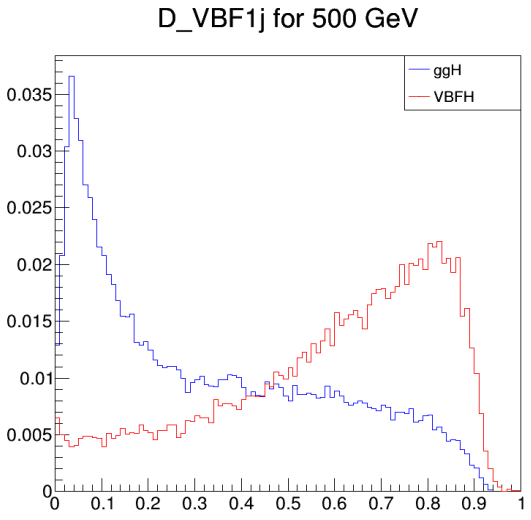

D_VBF1j for 200 GeV
ggH
VBFH

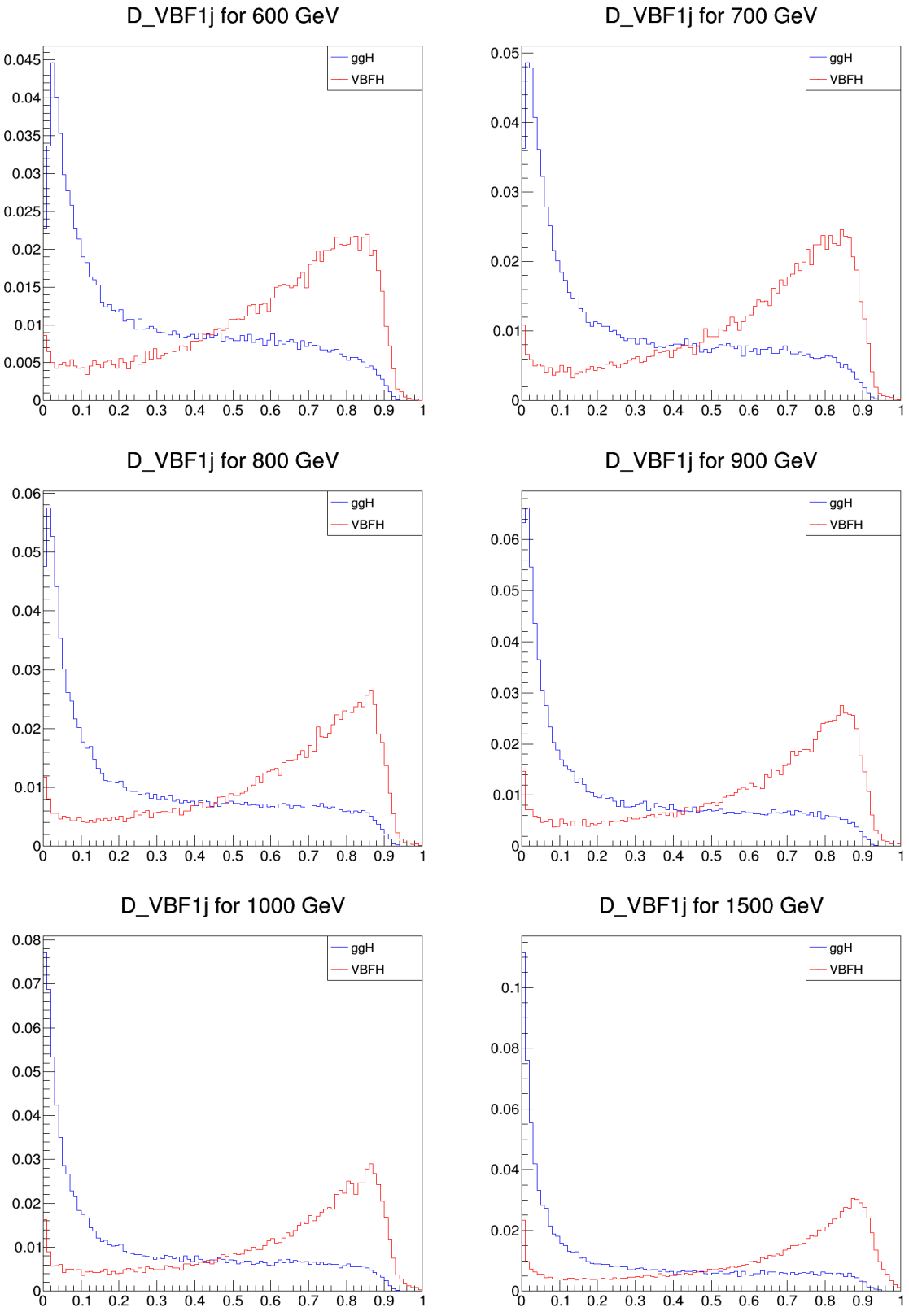
D_VBF1j for 250 GeV
ggH
VBFH


D_VBF1j for 300 GeV
ggH
VBFH


D_VBF1j for 350 GeV
ggH
VBFH


D_VBF1j for 400 GeV
ggH
VBFH


D_VBF1j for 500 GeV
ggH
VBFH


D_VBF1j for 600 GeV
D_VBF1j for 700 GeV
D_VBF1j for 800 GeV
D_VBF1j for 900 GeV
D_VBF1j for 1000 GeV
D_VBF1j for 1500 GeV
ggH
VBFH

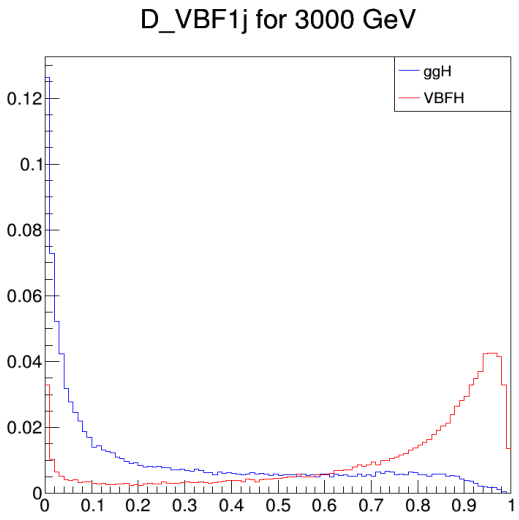
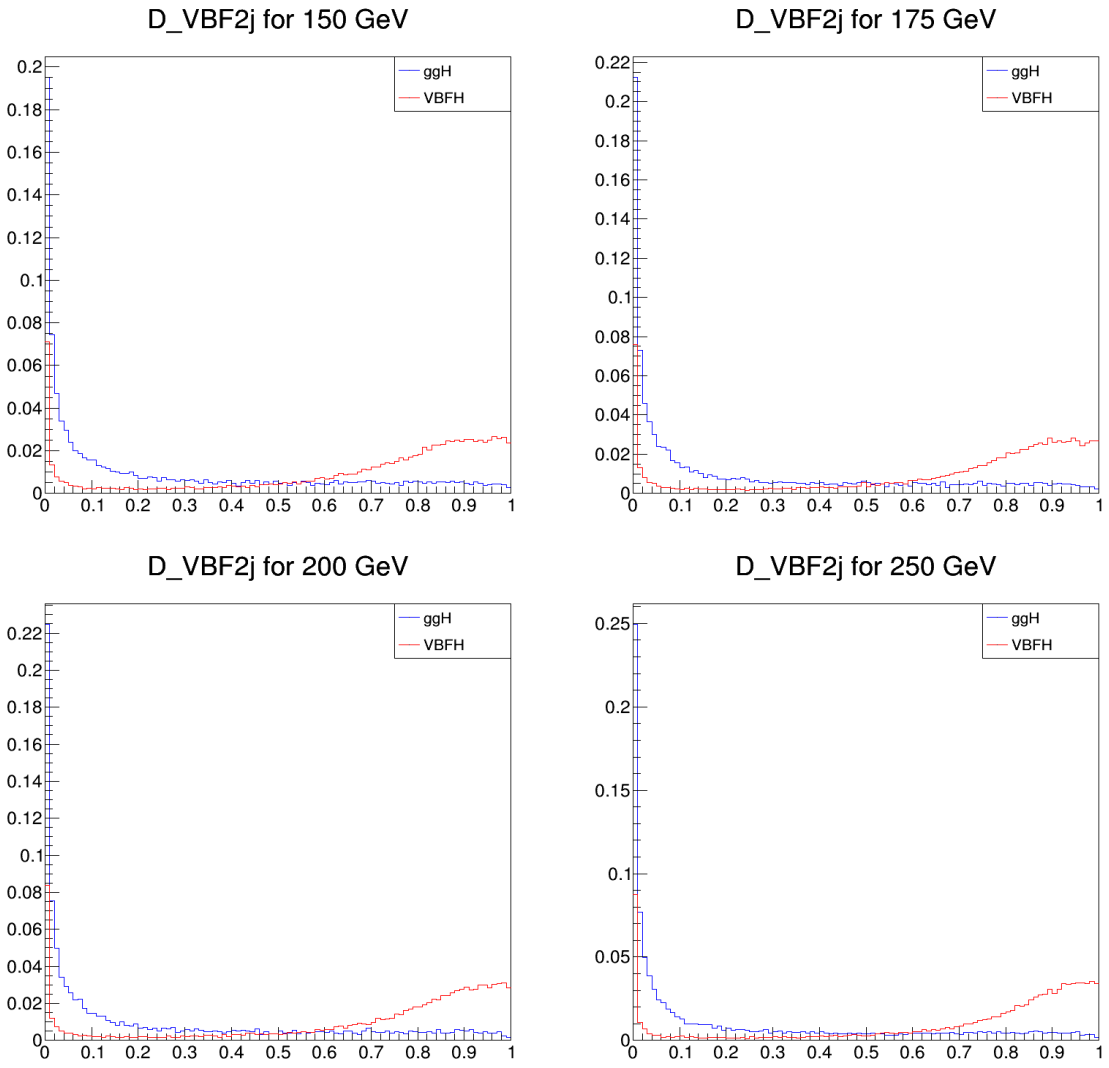
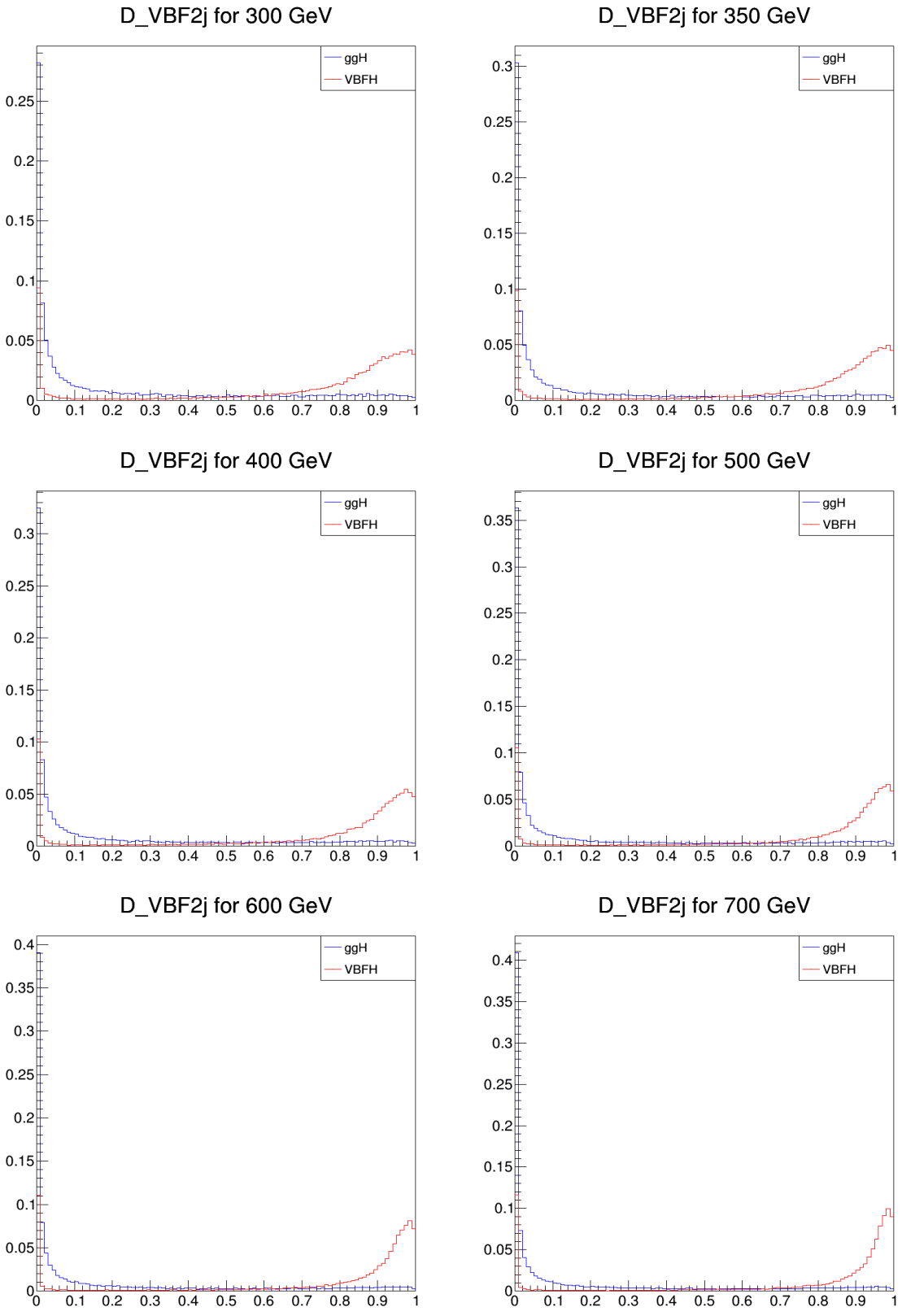


Figure 31: *Discriminating variable D_VBF1j shown for a range of masses.*

.




D_VBF2j for 300 GeV
D_VBF2j for 350 GeV
D_VBF2j for 400 GeV
D_VBF2j for 500 GeV
D_VBF2j for 600 GeV
D_VBF2j for 700 GeV
ggH
VBFH

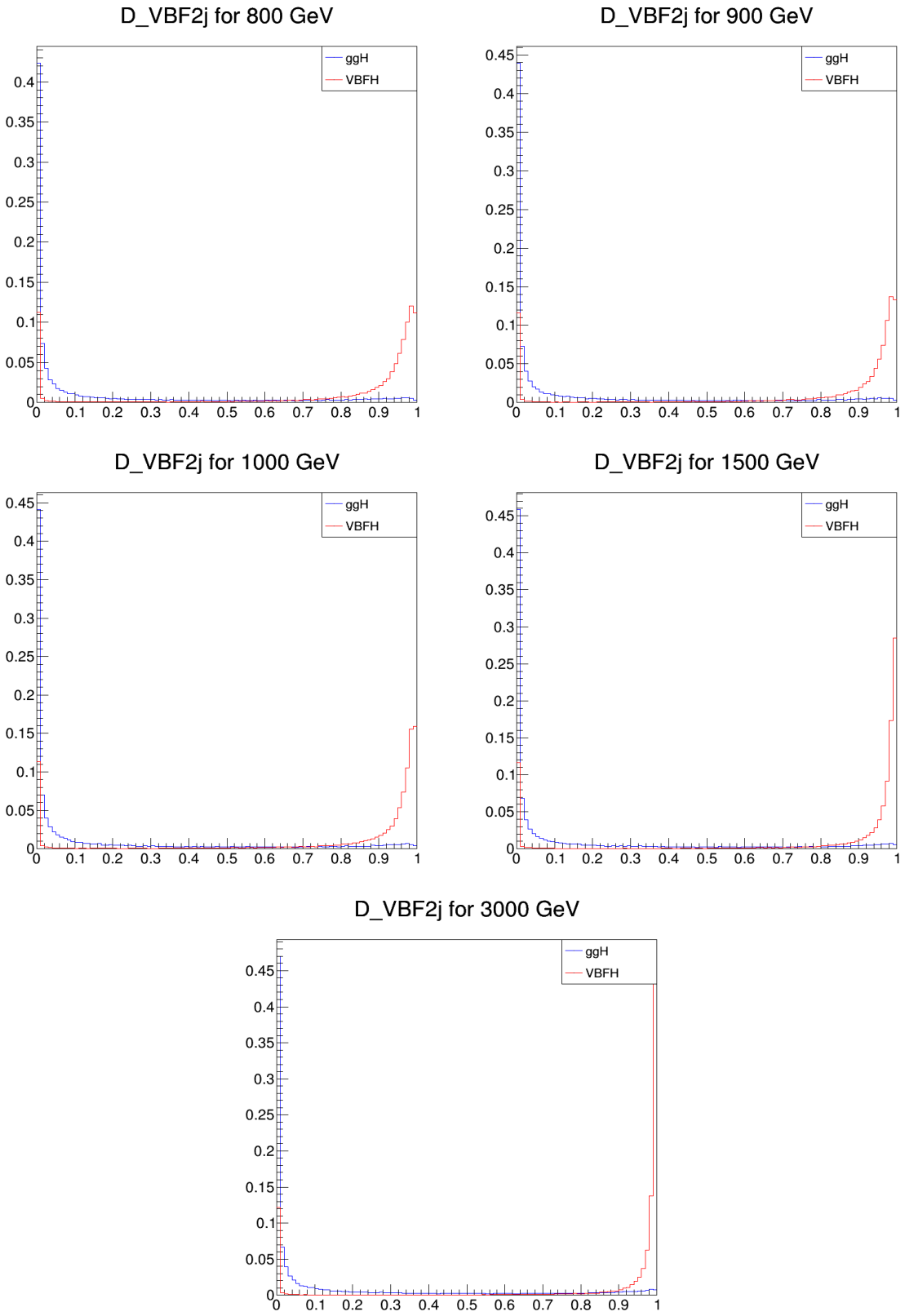
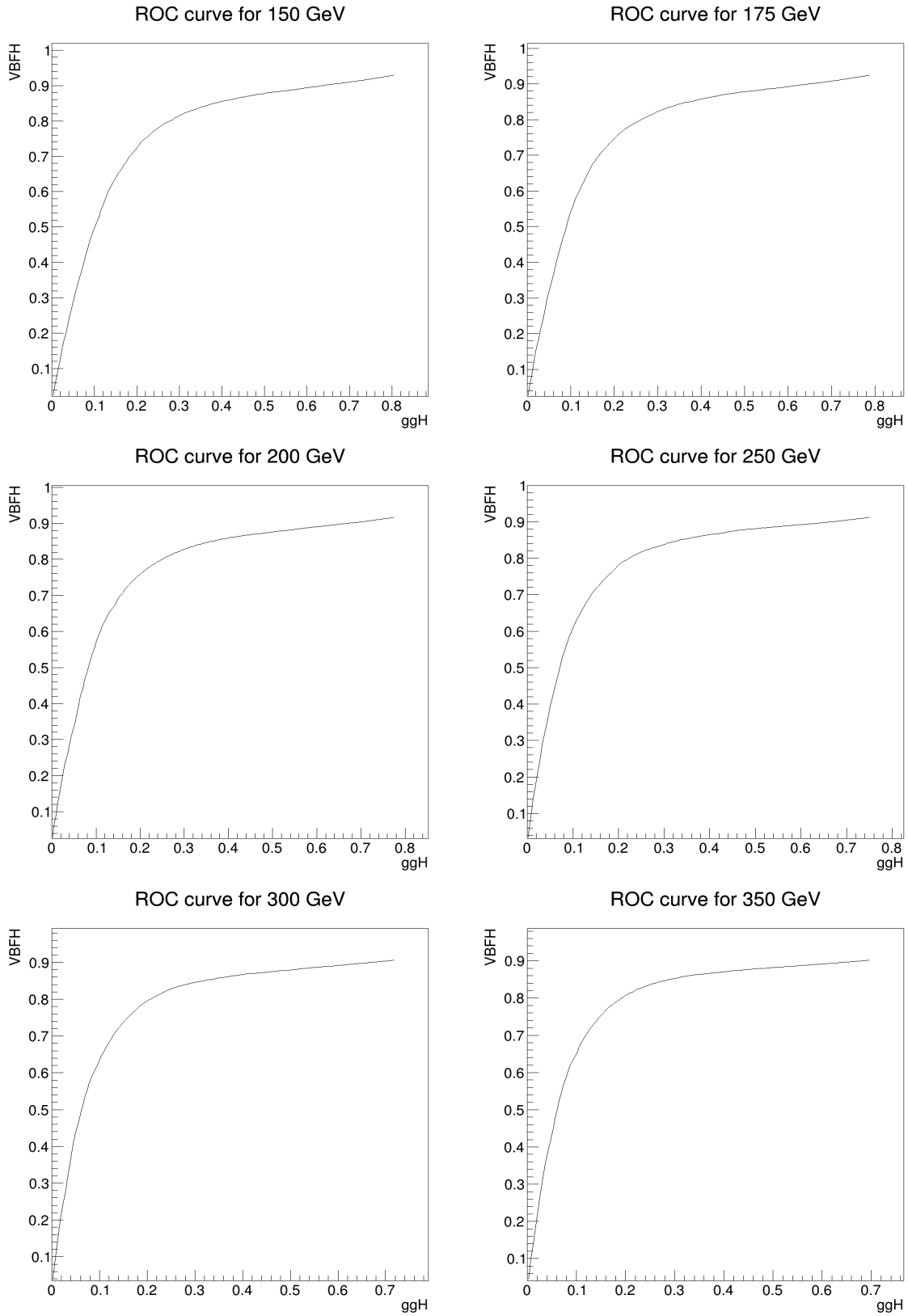
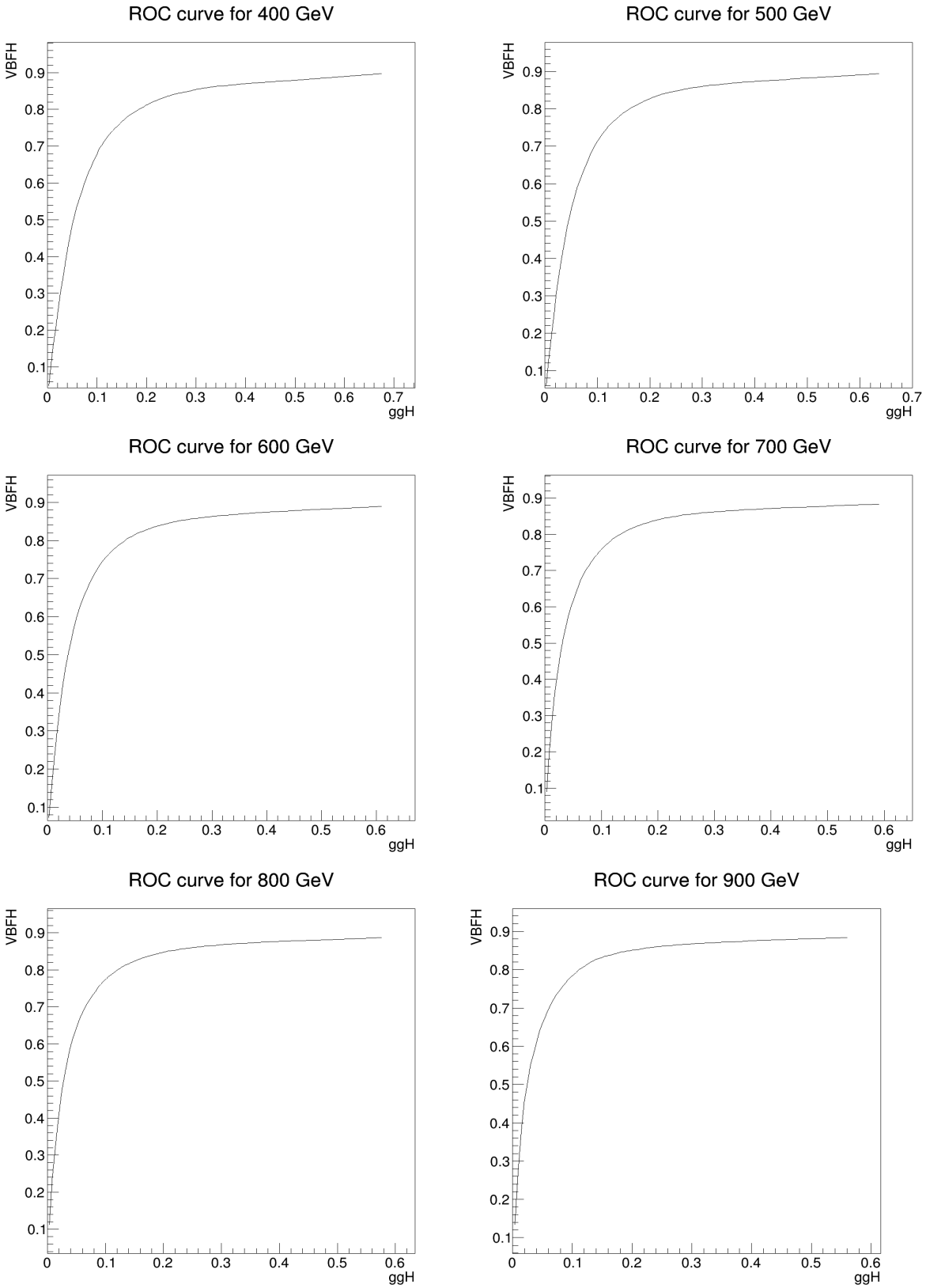


Figure 32: *Discriminating variable D_VBF2j shown for a range of masses.*

ROC curves:


ROC curve for 150 GeV
ROC curve for 175 GeV
ROC curve for 200 GeV
ROC curve for 250 GeV
ROC curve for 300 GeV
ROC curve for 350 GeV
VBFH
ggH


ROC curve for 400 GeV
ROC curve for 500 GeV
ROC curve for 600 GeV
ROC curve for 700 GeV
ROC curve for 800 GeV
ROC curve for 900 GeV
VBFH
ggH

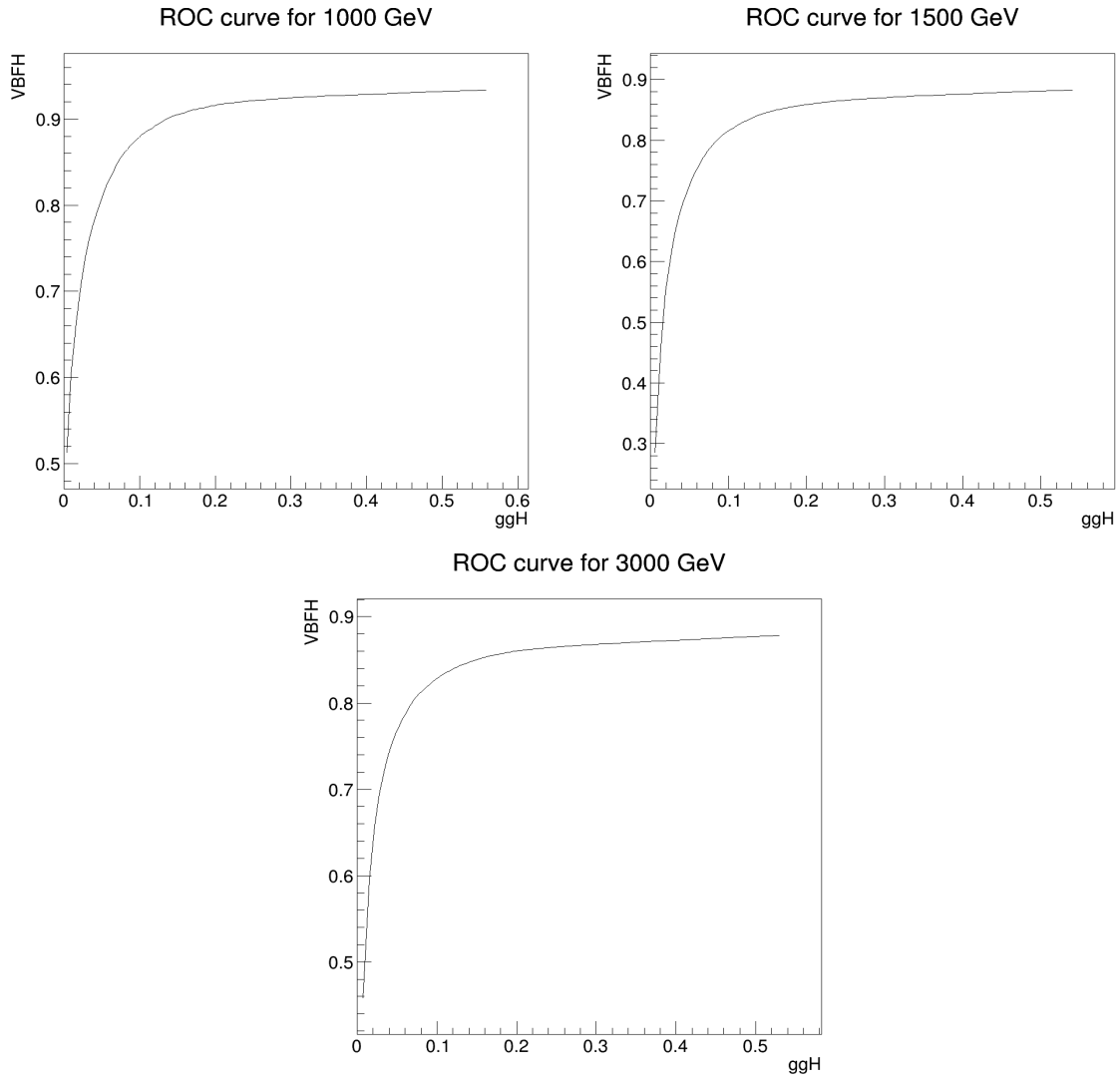


Figure 33: *ROC curves for higher masses.*