Facultad
de
Ciencias

ESTUDIO DE LA DESINTEGRACIÓN DEL BOSÓN DE HIGGS EN EL LHC

Study of the decay of the Higgs boson at LHC

Trabajo de Fin de Grado
para acceder al

GRADO EN FISICA

Autor: Guillermo Camacho de la Vega

Director: Jónatan Piedra Gómez

Septiembre - 2019

Resumen

Este trabajo tiene como fin mejorar la búsqueda del bosón de Higgs proveniente de la producción asociada (VH donde V puede ser el bosón $W^{\pm}$ o el bosón Z) con la desintegración $H \to WW \to 2\ell 2\nu$ y el bosón W o Z desintegrándose hadrónicamente, con colisiones protón-protón y energía de centro de masas $\sqrt{s}$ de 13 TeV. Este análisis se realiza usando datos tomados por el experimento CMS durante el Run del LHC de 2017, con una luminosidad integrada de $41{,}5\,\mathrm{fb}^{-1}$. Se ha realizado un análisis basado en dos variables clave, la masa invariante de los dos jets principales m_{jj} y el discriminante Quark-Gluon de los jets. Posteriormente se ha realizado otro estudio usando técnicas de análisis multivariante (MVAs), en concreto se han entrenado dos boosted decision trees (BDTs) para intentar separar nuestra señal de dos de los fondos principales, Higgs proveniente de fusión de gluones, y desintegración de quarks top.

Palabras clave: CMS, producción asociada, boosted decision trees, análisis por cortes, Higgs

Abstract

The objective of this project is to improve the search for the Higgs boson from the associated production mechanism (VH, where V can be a $W^{\pm}$ or a Z boson) in the $H \to WW \to 2\ell 2\nu$ decay channel where the W or Z decays hadronically, using proton-proton collisions with a center of mass energy $\sqrt{s}$ of 13 TeV. This analysis is made with data provided by CMS during the 2017 LHC Run, with an integrated luminosity of $41{,}5\,\mathrm{fb}^{-1}$. The first analysis is cut-based, focusing on two key variables: the mass of the two leading jets m_{jj} and the Quark-Gluon discriminant of said jets. Then, another study using multivariate analysis techniques (MVAs) has been made, specifically by training two boosted decision trees (BDTs) to separate our signal from two of the main backgrounds, Higgs from gluon fusion, and top quarks decay.

Key words: CMS, associated production, boosted decision trees, cut-based analysis, Higgs

Índice

1. Introducción

Este trabajo tiene como objetivo mejorar las medidas de la producción y desintegración del bosón de Higgs en CMS en un canal muy concreto, en el que el Higgs se produce junto a un bosón V (W o Z) y se desintegra en dos bosones W que a su vez dan lugar a dos leptones y dos neutrinos, y el V se desintegra en dos jets[1]. Para comprender los fenómenos con los que trabajamos, se ha de hacer un breve estudio del modelo estándar y el bosón de Higgs, así como conocer el colisionador LHC (Large Hadron Collider) y el detector CMS (Compact Muon Solenoid), sus partes y las funciones que cumplen.

Para este análisis se ha usado el paquete de ROOT [1], tanto para trabajar sobre las muestras y generar los histogramas como para aplicar el análisis multivariante.

1.1. El modelo estándar

El *Standard Model (SM)* o modelo estándar de la física de partículas nos da una visión unificada de las partículas elementales y las interacciones entre ellas (fuerzas), que se describen como intercambios de partículas [2].

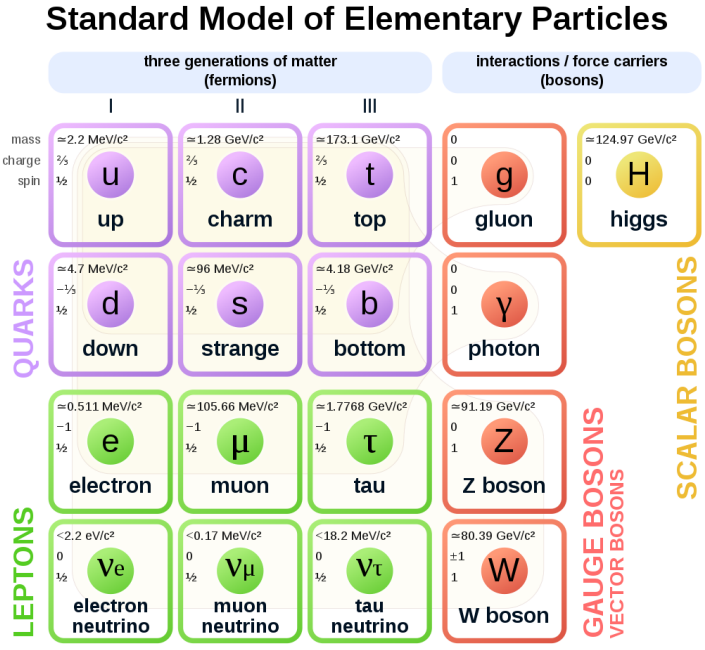

Figura 1: Componentes fundamentales del modelo estándar, con las tres generaciones de fermiones y los bosones [3].

[1] De ahora en adelante nos referiremos a este canal como "VH 2j" por brevedad.

Las partículas elementales se dividen en fermiones (spin semientero, constituyentes de la materia) y bosones (spin entero, mediadores de fuerzas). Los fermiones a su vez se pueden separar en leptones y quarks, o bien en tres generaciones (con dos leptones y dos quarks cada una). Para cada una de las cuatro partículas de la primera generación, existen dos copias que solo se diferencian por ser más masivas, en la segunda y tercera generación (ver las tres primeras columnas en la Figura 1). Debido a la naturaleza de su interacción (fuerza fuerte), los quarks solo pueden existir agrupados en estados ligados llamados hadrones, como el protón o el neutrón.

Como describe la ecuación de Dirac, que trata la dinámica de los fermiones, para cada uno de los 12 fermiones existe una antipartícula, de igual masa y carga opuesta, que se denota con una barra ($\bar{u}$) o por el signo de su carga (e^+).
El modelo estándar también desarrolla tres tipos de interacciones entre partículas, o fuerzas fundamentales. Cada una se describe mediante una teoría cuántica de campos (QFT) y el intercambio de una partícula virtual[2] que es bosón gauge (partícula de spin 1) . Estas tres fuerzas son:

- **La fuerza débil** afecta a todos los fermiones, su mediador puede ser el bosón Z (corriente neutra) o el bosón $W^{\pm}$ (corrientes cargadas).
- **La fuerza electromagnética** afecta a las partículas con carga eléctrica no neutra, su mediador es el fotón.
- **La fuerza fuerte** Afecta solo a quarks, su mediador es el gluón.

La fuerza de la gravedad no se incluye en el modelo estándar.

Partículas inestables y su desintegración

Algunas partículas se pueden producir en colisiones de alta energía como las del LHC (del que hablaremos en la Sección 2.1). Solo unas pocas de estas partículas son estables (el electrón, el protón, el fotón y los neutrinos), es decir, la mayoría son inestables y por lo tanto tienen un tiempo de vida limitado. Estas partículas se desintegran a estados de menor masa mediante procesos de interacción débil (ya que esta interacción permite cambios de sabor). Si la vida de una partícula inestable es larga (mayor que $\sim 10^{-10}$ s) viaja varios metros antes de desintegrarse y puede ser observada por los detectores (como muones, neutrones, piones...). Si su vida es muy corta, se desintegran antes de ser detectadas y solo se pueden identificar midiendo los productos de su desintegración.

[2]Partículas virtuales: partículas elementales cuyo tiempo de existencia está limitado por el principio de indeterminación de Heisemberg. No tienen la misma masa que una partícula real, pero conservan energía y momento.

1.2. El bosón de Higgs

El mecanismo de Higgs y el bosón de Higgs son una parte esencial del modelo estándar, necesarios para que sea una teoría consistente [2]. En el SM, las fuerzas electromagnética y débil se pueden unificar en la teoría electrodébil, pero ésta requiere que las partículas no tengan masa. Sin embargo, se observa experimentalmente que los bosones W y Z, asi como los leptones o los quarks, tienen una masa inercial que no se explica con la teoría electrodébil y el modelo estándar. Para solucionar esto se desarrolla el modelo del mecanismo Brout-Englert-Higgs [4], que da masa a las partículas pero no modifica las características ya conocidas de las interacciones entre partículas. De esta manera, se describe la interacción electrodébil a partir de tres bosones con masa que portan la fuerza débil ($\mathrm{W}^+, \mathrm{W}^-, \mathrm{Z}$), uno sin masa que porta la electromagnética (fotón γ) y un nuevo bosón escalar de spin 0, el Higgs. Las partículas masivas adquieren dicha masa mediante la interacción con el campo de Higgs.

El bosón de Higgs en el SM es una partícula escalar de carga neutra. Su masa según las medidas más recientes es de $125{,}18 \pm 0{,}16$ GeV [5]. Su vida media para esta masa es de solo $1{,}56 \cdot 10^{-22}$ segundos [6]. Se acopla con todos los fermiones y partículas masivas, y su constante de acoplamiento es proporcional a la masa de las mismas. Por lo tanto, el bosón Higgs se puede desintegrar a casi cualquier partícula del modelo estándar, aunque la desintegración más probable es a un par de quarks bottom ($b\bar{b}$) con una fracción de desintegración $\Gamma = 57{,}8\,\%$, y en menor medida a bosones W ($21{,}6\,\%$) o gluones ($8{,}6\,\%$). En la Figura 2 se pueden ver las diferentes fracciones de desintegración en función de la masa del bosón Higgs.

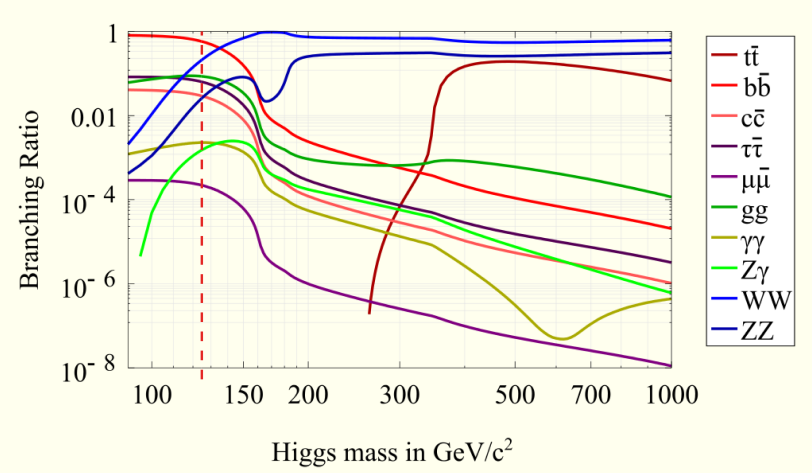

Figura 2: Fracciones de las diferentes desintegraciones del bosón de Higgs en función de su masa [7]. La linea punteada marca 125 GeV.

Mecanismos de producción

El bosón de Higgs del modelo estándar se puede producir en un colisionador de partículas como el LHC a través de diferentes procesos (ilustrados por los diagramas de Feynman en la Figura 3). Los cuatro más comunes son:

- **Producción por fusión de gluones** (*gluon fusion*, ggH): Se produce por la interacción entre dos gluones que forman un bucle de quarks virtuales (generalmente quarks pesados: t y b). Es el mecanismo dominante en el LHC.
- **Producción por fusión de bosones vectoriales** (*vector boson fusion*, VBF): Dos fermiones que colisionan pueden intercambiar un bosón W o Z virtual, que si lleva suficiente energía, puede emitir un Higgs. Es el segundo proceso más común en el LHC.
- **Proceso de Higgs-strahlung o producción asociada** (*associated production*, VH): Si un fermión colisiona con un anti-fermión (quark y anti-quark, electrón y positrón) ambos pueden unirse para formar un W o Z virtual, que puede emitir un Higgs. Al colisionar protón con protón, este proceso es menos común en el LHC.
- **Producción por fusión de un par top-antitop** (*top fusion*, ttH): Dos gluones que colisionan, cada uno se desintegra en un par quark y anti-quark pesados que pueden combinarse para formar un Higgs. Es el proceso menos común de los cuatro.

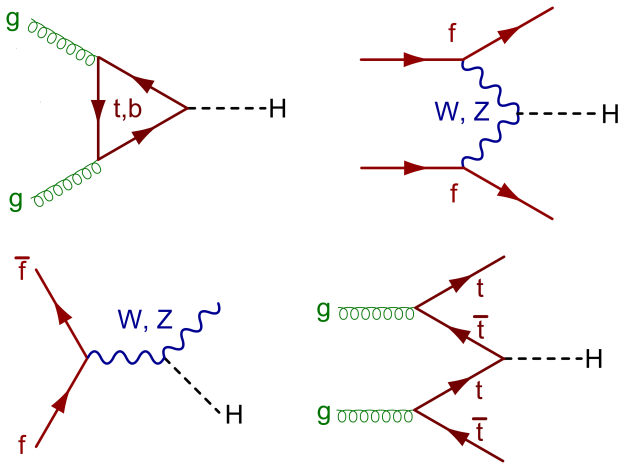

Figura 3: Diagramas de Feynman de los procesos de producción del Higgs: ggH (arriba izquierda), VBF (arriba derecha), VH (abajo izquierda) y ttH (abajo derecha) [8].

2. Estudio del bosón de Higgs en CMS

En julio de 2012, los experimentos ATLAS y CMS del LHC anunciaron el hallazgo de una partícula consistente con el bosón de Higgs en la zona de masa alrededor de 125 GeV [9–11]. Actualmente se estudian las propiedades del bosón de Higgs para comprobar la certeza de las predicciones del modelo estándar, y buscar evidencias de nueva física [4].

2.1. El acelerador LHC

El Large Hadron Collider (LHC) es el acelerador de partículas más grande y potente del mundo. Comenzó a operar en septiembre de 2008. Consiste en un anillo de 27 kilómetros de longitud con imanes superconductores y estructuras para acelerar las partículas durante su recorrido. En el interior del anillo, dos haces de protones viajan en sentidos contrarios hasta colisionar, con una energía de centro de masas $\sqrt{s} = 13$ TeV.

En el LHC hay varios detectores con diferentes funciones: ALICE, ATLAS, LHCb, y CMS (Figura 4). Este último se especializa en la detección de muones con el fin de explorar nueva física en la escala de energía del TeV y en especial la física del Higgs.

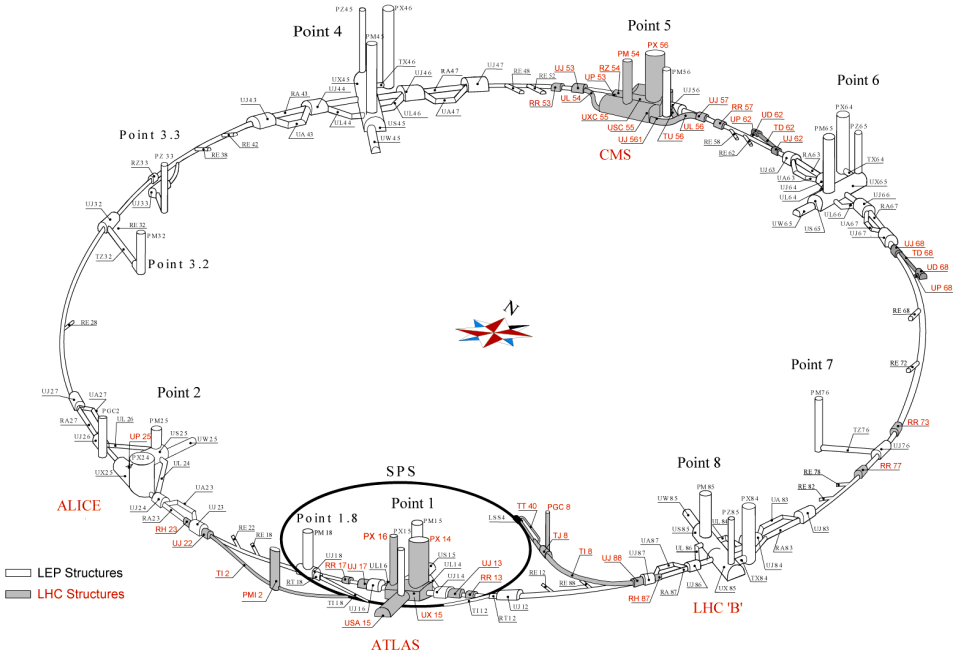

Figura 4: Esquema del LHC [12].

2.2. El detector CMS

El detector CMS (Compact Muon Solenoid) [13] se compone de las siguientes partes: detectores de trazas, calorímetros electromagnético y hadrónico y sistema de muones (Figura 5). Cada tipo de partícula interacciona de manera diferente con estas partes en función de sus propiedades. La mayoría de los detectores de partículas tienen una parte central cilíndrica (*barrel*) y dos tapas en los extremos (*endcaps*). Además de detectores, CMS tiene un gran iman superconductor solenoidal que permite determinar el momento lineal de las partículas cargadas en base a la curvatura de sus trayectorias, causada por el campo magnético de de 3,8 Teslas (Fuerza de Lorentz.).

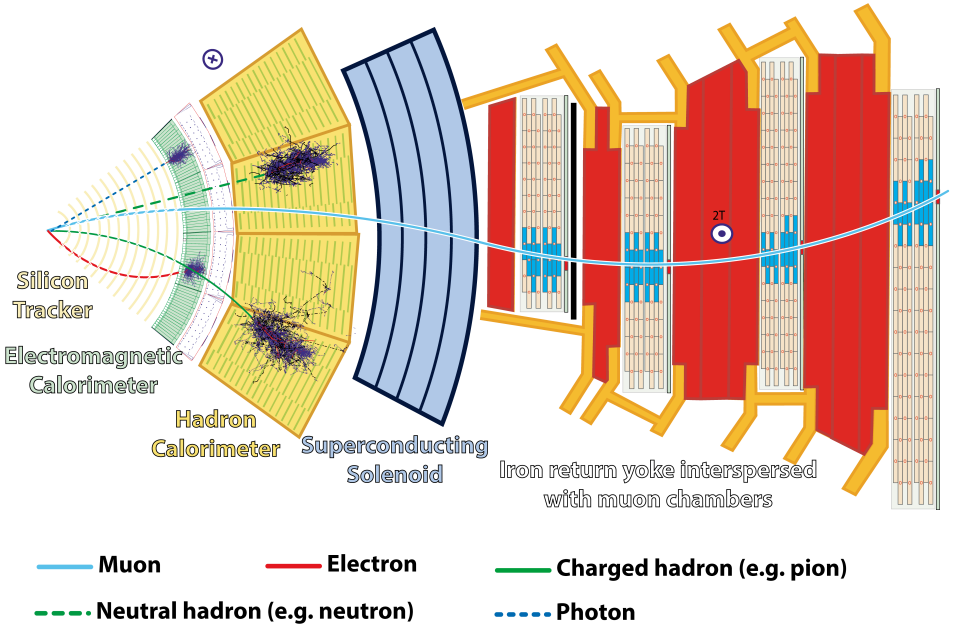

Figura 5: Sección transversal del detector CMS y sus partes [14].

Detector de trazas (Tracker)

Es la parte más interna y cercana al punto de colisión. Está compuesto de silicio, que produce señales eléctricas cuando una partícula lo atraviesa. Contiene sensores de píxel y de strip. El tracker puede medir la posición de muones, electrones, hadrones y partículas de vida corta como quarks b. Se divide en dos partes:

- Pixel Tracker: Está compuesto de millones de píxeles de silicio, tiene una resolución espacial muy alta y es esencial para medir partículas

de alto momento y los vértices secundarios producidos por desintegración de otras partículas.

- Strip Tracker: Está compuesto de cuatro capas de barras de silicio a baja temperatura.

Calorímetro electromagnético (ECAL)

Está compuesto de miles de cristales centelleadores de tungstato de plomo ($\mathrm{PbWO_4}$), un material muy denso pero transparente, capaz de frenar partículas muy energéticas. Mide con mucha precisión la energía y trayectoria de electrones y fotones, separando los de alta y baja energía.

Calorímetro hadrónico (HCAL)

Está formado por diferentes partes de materiales densos (bronce o acero) con centelleadores. Mide la energía y trayectoria de partículas hadrónicas (protones, neutrones, etc) que llegan en forma de chorros de partículas o jets. Ademas, permite hacer una medida indirecta de la presencia de partículas cuya interacción es tan débil que atraviesan todos los detectores sin producir señales, como los neutrinos, a través de la energía perdida transversa (ver Sección 2.2.3).

Sistema de muones

Los muones son partículas clave en la desintegración de otras partículas, entre ellas el Higgs. Son muy masivos y pueden penetrar varios metros de acero sin interactuar, por lo que atraviesan los detectores y calorímetros. Por ello se ponen cámaras de detección de muones en las capas más exteriores de CMS, intercaladas con el yugo de hierro ("iron yoke"). Para identificar muones y medir su momento se usan tres tipos de detectores [15]:

- Tubos de deriva: Consisten en tubos de 4 cm de ancho, que contienen un cable cargado eléctricamente y una atmósfera de gas. Cuando una partícula con carga los atraviesa, se desprenden electrones del gas que son atraídos por el cable, produciendo "hits" al contactar. Cubren la región central (barrel) con $0 < \eta < 1{,}3$ (η es la pseudorapidez, descrita en la Sección 2.2.2). Se divide en tres sub-unidades, dos que miden el ángulo ϕ y una que mide la posición en el eje z.

- Cámaras de tiras catódicas (CSC): Se trata de una serie de cables cargados positivamente (ánodos) y tiras cargadas negativamente (cátodos) rodeados de gas. Las partículas cargadas producen electrones e iones en el gas que generan "hits" en los ánodos y cátodos respectivamente. Se sitúan en las regiones externas (endcaps) con $0{,}9 < \eta < 2{,}4$.

Cada endcap tiene 4 estaciones de muones formadas por varias CSCs con forma trapezoidal.

- Cámaras de tiras resistivas (RPC): Son cámaras de gas con ánodos y cátodos. Cuando una partícula cruza el gas, se produce una avalancha de electrones que son recogidos por tiras metálicas tras un retardo conocido con mucha precisión. Tienen una resolución temporal muy buena (respuesta muy rápida) por lo que son muy útiles para el sistema de trigger (Sección 2.2.1). Cubren el barrel y el endcap ($0 < \eta < 2{,}1$) y sirven como detector adicional o complementario.

2.2.1. Recogida y procesado de datos

El numero de colisiones que se producen en el LHC es muy grande, del orden de 10^9 por segundo. Para poder almacenar los datos, se deben usar una serie de disparadores ("triggers") que seleccionen los eventos que nos interesan y descarten el resto en tiempo real. Con esto se puede reducir el numero de eventos a unos 400 por segundo, suficiente para poder almacenarlos. Los triggers se dividen en dos niveles: Level 1 (hardware), y High Level Trigger (HLT, software) [16].

Para el Level 1 se utilizan principalmente las cámaras de tiras resistivas del sistema de muones, y también las CSCs y los tubos de deriva. Se hace una selección rápida de candidatos según su momento en el plano transverso (p_T) y se eliminan fondos. Este trigger es muy rápido y hace una selección muy simple, quedando del orden de 10^5 sucesos por segundo.
EL HLT se compone de dos subniveles, Level 2 y Level 3. Se hacen reconstrucciones de los muones basadas en la información de los detectores, reduciendo el número de sucesos por segundo a unos 400 Hz [17].

Una vez superados los triggers, la información elemental recogida por los diferentes detectores de CMS se combina mediante algoritmos para reconstruir las trazas de las partículas y sus propiedades, identificándose "objetos" como electrones, muones, jets, etc con los que se puede trabajar. Los datos se almacenan y se copian en el LHC Computing Grid, para que puedan ser accedidos para análisis a través del sistema de software de CMS, CMSSW [18].

2.2.2. Observables y variables de los procesos

Para este trabajo se han usado datos de colisiones generadas con simulaciones de Montecarlo, y datos tomados por el detector CMS. La información de las propiedades de las partículas detectadas se recoge en diferentes variables como la masa, el momento, el ángulo respecto al eje z, etc. En la

Figura 6 se puede ver la definición del sistema de ejes y algunas variables angulares y cinemáticas. Sobre las variables se pueden hacer selecciones en forma de cortes para modificar el número de sucesos de un tipo determinado a la hora de hacer un análisis (Sección 3.3). Algunas de las variables que se han usado en este análisis son:

- **Momento transverso p_T:** es la componente (proyección) del momento de un partícula o conjunto de partículas que se mide en el plano transverso al eje del haz (plano xy). El superíndice denota la partícula correspondiente, por ejemplo, $p_T^{\ell_2}$ es el p_T del segundo leptón (están ordenados de mayor a menor momento).

- **Pseudorapidez η:** se trata de una variable que se relaciona con el ángulo que tiene la partícula respecto del eje del detector (eje z). Tiene la característica de que en el límite de partículas relativistas ($p_T \gg m$) es equivalente a la rapidez y las diferencias de rapidez son invariante Lorentz (no cambian bajo transformaciones de Lorentz). La pseudorapidez se define como:

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] = \operatorname{arctanh}\left(\frac{p_L}{|\vec{p}|}\right), \tag{1}$$

  donde θ es el ángulo entre el momento de la partícula y el eje z del detector, y p_L es la componente del momento paralela a este eje (momento longitudinal).

  Para ángulos θ entre 0º y 90º la pseudorapidez alcanza valores entre ∞ y 0, respectivamente. Generalmente en el detector se miden valores de $|\eta|$ menores que 5, y con buena resolución, menores que $2{,}5$ ($\theta \approx 10°$).

- **Ángulo acimutal ϕ:** es el ángulo entre el eje x y la dirección de la partícula proyectada sobre el plano transverso.

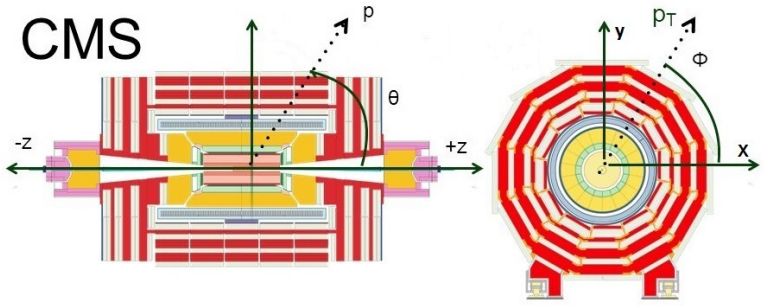

Figura 6: Representación en el detector de algunas variables básicas usadas en el análisis [19].

- **Separación angular ΔR:** se define a partir de la pseudorapidez y el ángulo acimutal como:

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \tag{2}$$

  Es invariante Lorentz para partículas sin masa.

- **Masa transversa del bosón de Higgs $m_{\mathrm{T}}^{\mathrm{H}}$:** se puede reconstruir a partir de las propiedades cinemáticas de la producción y desintegración del Higgs, como [20]:

$$m_{\mathrm{T}}^{\mathrm{H}} = \sqrt{2p_T^{\ell\ell}p_T^{miss}[1 - \cos\Delta\phi(\ell\ell, \vec{p}_T^{\,miss})]}, \tag{3}$$

  donde $\Delta\phi(\ell\ell, \vec{p}_T^{\,miss})$ es el ángulo acimutal entre el momento del par de leptones y el momento perdido transverso.

- **Quark-Gluon likelihood discriminant QGL:** es una variable disponible solo para jets, se trata de un discriminante que busca distinguir entre jets creados por quarks ligeros, y jets creados por gluones. Esta variable asigna un valor entre 0 y 1 a cada jet; valores cercanos a 0 indican que es más probable que provenga de un gluón, valores cercanos a 1 indican que proviene de un quark [21].

2.2.3. Objetos (Physics Objects)

En este análisis se emplean "objetos" (physics objects) que son los elementos que se usan para reconstruir los procesos. Los objetos principales son electrones, muones, jets y la MET (Missing Transverse Energy). Descripciones con mayor detalle de cada uno de estos objetos se pueden encontrar en [22] y [23]. A continuación se resumen las características principales de estos objetos.

Electrones

Para identificar los electrones provenientes de un W o un Z ("prompt electrons") y separarlos de fondos como electrones provenientes de fotones o provenientes de quarks, se requiere una serie de características en sus variables, por ejemplo el que estén aislados (para evitar que sean parte de un jet). Estas características pueden ser observables del calorímetro, variables de aislamiento o variables para rechazar la creación de pares (conversión de fotones). Por ejemplo, se les pide $|\eta| < 2{,}5$.

Muones

De manera similar a los electrones, se requiere que los muones cumplan una serie de criterios para ser considerados señal ("prompt muons"). Entre

otras cosas se requiere $|\eta| < 2{,}4$ y un parámetro de impacto (distancia al vértice de la colisión) determinado dependiendo de su momento. Ademas, los muones pueden ser de tres tipos en función de dónde sean detectados:

- Tracker muon: Solo se recompone a partir de señales en el detector de trazas.
- Stand-alone muon: Solo se reconstruye a partir de información en el detector de muones.
- Global muon: Se reconstruye a partir de información de ambos detectores, el de trazas y el de muones.

Jets

Los jets son chorros colimados de partículas que se detectan en los calorímetro electromagnético y hadrónico del detector. Estos jets se generan a partir de los quarks y gluones que se forman en la colisión, y que inmediatamente después se hadronizan, transformándose en chorros de partículas hadrónicas que podemos observar [24]. Hay varias maneras de reconstruir y agrupar estos jets mediante diferentes algoritmos.

Transverse Missing Energy

La MET (E_T^{miss}, Tranverse Missing Energy, energía perdida transversa) es una medida de la energía faltante en el plano transverso del detector tras una colisión (por lo tanto, solo hay un valor de MET por cada evento). Se define como menos el valor absoluto de la suma del momento transverso de todas las partículas finales en un evento:

$$E_T^{miss} = -\left|\sum_i \overrightarrow{p}^i{}_T\right| \tag{4}$$

Esta energía faltante (en el plano transverso) es la diferencia entre la energía total final tras la colisión, y la energía inicial de los protones que colisionan, y por conservación de la energía debería ser cero. Sin embargo, aparece debido a la existencia de partículas que no podemos detectar: los neutrinos, y también debido a características del detector, problemas o falta de resolución, mala identificación de trazas, materia oscura, partículas supersimétricas que se escapan, etc.
Hay diversos algoritmos de reconstrucción de MET. Nosotros usamos PuppiMET (Pileup Per Particle Identification) para este análisis, ya que se ha encontrado que obtiene la mejor resolución y el mejor acuerdo de datos-MC [23].

3. Estudio del canal VH 2j

El canal VH 2j se da para el mecanismo de producción asociado a un bosón vectorial V (que puede ser W o Z). En concreto, este canal se centra en el caso en que el bosón V se desintegra a dos jets, y el Higgs se desintegra en un par de bosones W que a su vez se desintegran leptónicamente (a un electrón más un neutrino, o un muón más un neutrino):

$$\mathrm{pp} \rightarrow \mathrm{WH} \rightarrow \mathrm{WWW} \rightarrow \mathrm{q\bar{q}'}\ \mathrm{e}\nu_{\mathrm{e}}\ \mu\nu_{\mu}$$

$$\mathrm{pp} \rightarrow \mathrm{ZH} \rightarrow \mathrm{ZWW} \rightarrow \mathrm{q\bar{q}}\ \mathrm{e}\nu_{\mathrm{e}}\ \mu\nu_{\mu}$$

Este mecanismo de producción es poco probable, por lo que este análisis tiene una estadística pobre. En este trabajo se han usado los datos obtenidos por CMS en el Run del LHC de 2017.

3.1. Estudios previos

El análisis realizado con los datos del Run del LHC de 2016 [20] describe el proceso $\mathrm{H} \rightarrow \mathrm{WW} \rightarrow 2\ell 2\nu$, incluyendo los procesos de ggH (0, 1 y 2 jets), VBF y VH (subdividido en varios estados finales: WH3l, ZH4l y VH2j).
En el apartado de VH2j, se apunta al estado final con un V que se desintegra en dos jets y el Higgs a WW y estado final leptónico. Se tiene una región de señal ($e\mu$) y dos regiones de control (Top y $DY\tau\tau$) como se verá más adelante. Debido a la baja estadística, el análisis es un "shape analysis" en 1D, que consiste en realizar un ajuste a la forma de una distribución, en este caso de la masa invariante del sistema dileptónico ($m_{\ell\ell}$). La fracción de señal del VH frente a los demás mecanismos de producción es baja, y ggH es el dominante.

3.2. Sucesos de fondo más comunes

Para estudiar el bosón de Higgs es necesario tener en cuenta la existencia de otros procesos con secciones eficaces muchísimo mayores (mucho más probables, como se puede ver en la Figura 7) que contaminan la región de señal, y que se intentan reducir mediante cortes en variables (Sección 3.3). Algunos de estos procesos son:

- **Top:** Los quarks top principalmente se desintegran en quarks b más un bosón W, que puede desintegrarse leptónicamente a electrón más neutrino, muón más neutrino o tau mas neutrino, o hadrónicamente. Los procesos de top son el fondo dominante (tienen una sección eficaz muy alta) y además un estado final muy similar a la señal. Pueden tratarse de pares top y antitop o de quarks top solos. Se estima con MC.

- **Drell-Yan:** Se trata de un proceso causado por el scattering de hadrones a alta energía, y consiste en un par de quark y antiquark que se aniquilan, creando un fotón virtual o bosón Z que se desintegra en un par de leptón y antileptón. Se predice con MC.
- **Non-prompt:** Son partículas detectadas como leptones pero que no vienen de W o Z, o no son leptones. Pueden ser partículas de otro tipo (por ejemplo un jet que se identifica como un electrón) o leptones que provienen de la desintegración de un quark b o de la conversión de un fotón. Lo contrario son los leptones prompt, leptones reales provenientes de un W o Z. [26]. Para estimar este fondo se usa el método "fakeable object". Consiste en definir una región de control donde los requerimientos para los leptones son menos estrictos, por lo que hay más identificaciones incorrectas. Entonces se usa un factor de extrapolación para relacionar estos sucesos con la región de señal.
- **WW:** Este fondo es difícil de reducir ya que tiene el mismo estado final que la señal proveniente de un bosón de Higgs que se desintegra en un par WW. Se estima con MC.
- **Higgs proveniente de otros mecanismos de producción:** Como ya se vio en la Sección 1.2, hay otros mecanismos de producción del bosón de Higgs con mayor sección eficaz que el VH, en especial el Higgs proveniente de fusión de gluones (ggH). Se predicen mediante MC.

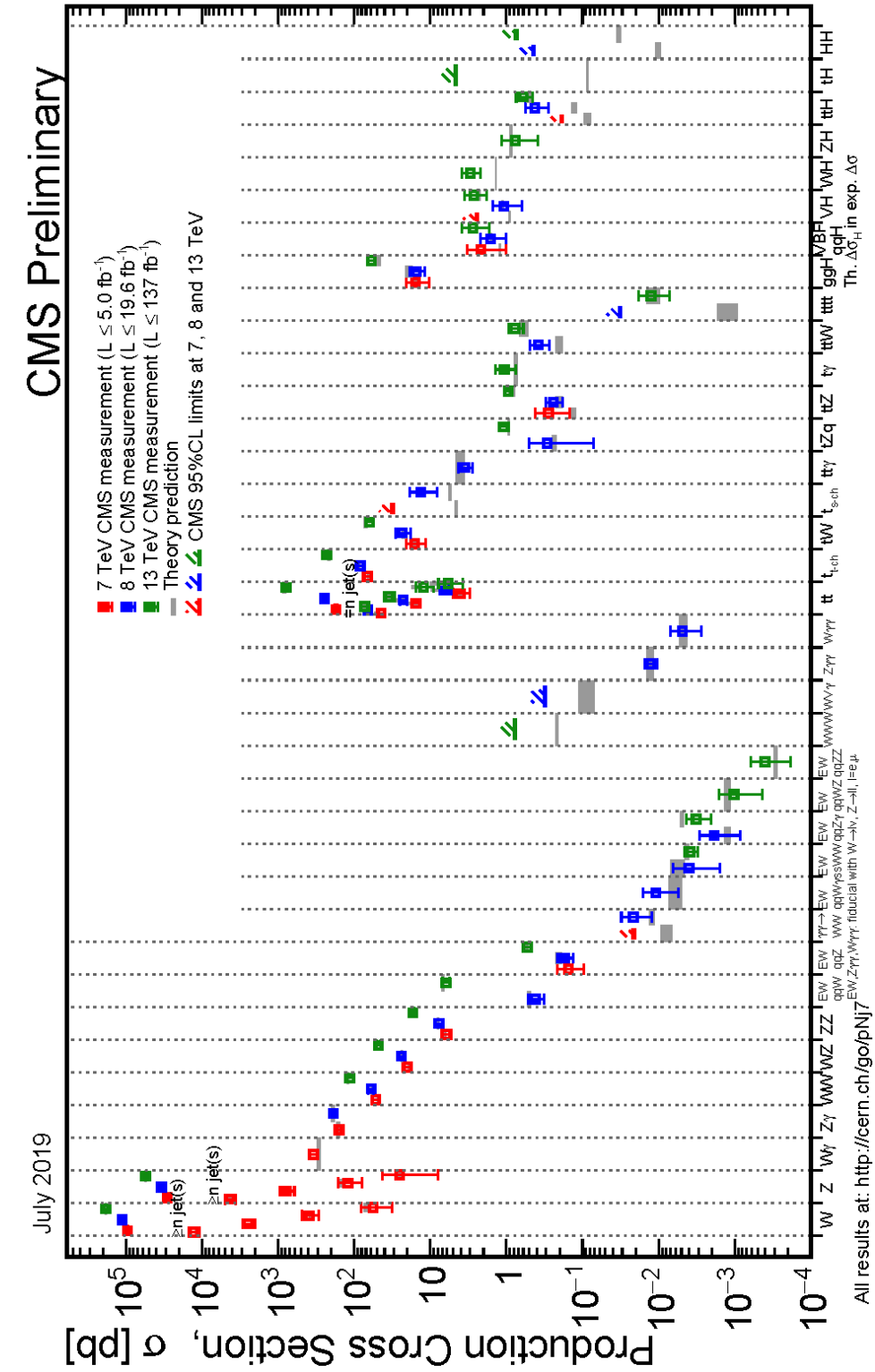

Figura 7: Sección eficaz de producción de diversos procesos en CMS [25].

3.3. Dataset y selección de sucesos

Se requiere que en el estado final haya dos leptones (para ello se usa una combinación de triggers de un leptón y dos leptones) [27]. El estado final tiene electrones, muones , E_T^{miss} y jets.
Los datos que se han usado corresponden al Run de 2017, con $\sqrt{s} = 13$ TeV y una luminosidad integrada de 41.5 fb^{-1}. Se han aplicado varios triggers: MuonEG, DoubleMuon, SingleMuon, DoubleEG,, SingleElectron. El trigger de MuonEG selecciona un electrón y un muon con unos requerimientos de p_T, identificación y aislamiento suaves, por lo que tienen una alta eficiencia de señal. SingleElectron y SingleMuon tienen requerimientos más duros y recogen sucesos en los que solo un leptón es identificado. DoubleElectron y DoubleMuon se usan para estimar la contaminación de WZ y Wγ^* en la region de señal. La eficiencia de los triggers se ha estimado en otros análisis en torno a 99 %.

Debido a la gran cantidad de fondos, para caracterizarlos se hace uso de simulaciones de Monte Carlo, se generan procesos de fondo y señal como los que se esperaría detectar en CMS y se hace análisis mediante cortes sobre ellos. Estos procesos simulados se han obtenido usando diferentes generadores como POWHEG [28], MADGRAPH [29] y PYTHIA [30]. En las Tablas 1 y 2 se enumeran los principales procesos de fondo y señal, respectivamente. También se muestra el orden de magnitud de sus secciones eficaces [25].

Proceso	Sección eficaz / fb^{-1}
DY $\to \ell\ell$	$\sim 10^5$
Top (t$\bar{\text{t}}$, single top)	$\sim 10^3$
WW $\to 2\ell 2\nu$	$\sim 10^2$
Vγ	$\sim 10^2$
WZ$\gamma^* \to 3\ell\nu$	$\sim 10^2$
VZ $\to 2\ell 2\nu / 4\ell / 2\ell 2q$	~ 10

Tabla 1: Principales procesos de fondo estudiados.

Proceso	Sección eficaz / fb^{-1}
ggH $\to$ WW $\to 2\ell 2\nu$	$\sim 10^2$
VBF $\to$ WW $\to 2\ell 2\nu$	~ 10
ZH $\to$ WW	~ 1
W$^{\pm}$H $\to$ WW	~ 1
ttH $\to$ non $-$ bb	$\sim$ 0,1

Tabla 2: Principales procesos de señal estudiados.

A partir de la sección eficaz σ, la Luminosidad $\mathcal{L}$ y la eficiencia de los cortes ϵ se puede estimar el numero de sucesos que esperamos de un determinado proceso mediante la Ecuación 5:

$$N = \sigma \cdot \mathcal{L} \cdot \epsilon \tag{5}$$

Debido a la gran sección eficaz de muchos de los fondos en comparación con los procesos de Higgs y nuestra señal, VH, se obtienen muchos más sucesos de fondo que de señal. Es necesario hacer cortes (ϵ) y selecciones que reduzcan el numero de sucesos de fondo para poder estudiar la señal. Estos cortes tienen el efecto secundario de reducir también el número de sucesos de señal.

En este análisis se busca optimizar la selección del mecanismo de producción VH2j. En la Tabla 3 se detallan los cortes que se aplican en la preselección, la región de señal y las dos regiones de control. Estos cortes iniciales se toman de [31] con algunas ligeras modificaciones.

Selección	Requerimientos
Preselección	$m_{\ell\ell} > 12$ GeV, $p_T^{\ell_1} > 25$ GeV,
	$p_T^{\ell_2} > 10$ (13) GeV para μ (e),
	PUPPI $E_T^{miss} > 20$ GeV, $p_T^{\ell\ell} > 30$ GeV, $p_T^{\ell_3} < 10$ GeV
	un electrón y un muón con cargas eléctricas opuestas
Región de señal	al menos dos jets con $p_T > 30$ GeV,
	los dos jets principales [3] con $\|\eta\| < 2{,}5$ y $\Delta\eta_{jj} < 3{,}5$,
	$60\text{GeV} < m_{\mathrm{T}}^{\mathrm{H}} < 125$ GeV y $\Delta R_{\ell\ell} < 2$,
	ningún jet b-tagged con $p_T > 20$ GeV,
	$m_{jj} < 200$ GeV
Reg. control $\mathrm{t\bar{t}}$	como la señal pero sin el corte de $m_{\mathrm{T}}^{\mathrm{H}}$,
	sin el corte de $\Delta R_{\ell\ell}$ y con $m_{\ell\ell} > 50$ GeV,
	al menos uno de los dos jets pricipales es b-tagged
Reg. control DY$\tau\tau$	como la señal pero con $m_{\mathrm{T}}^{\mathrm{H}} < 60$ GeV,
	$40 < m_{\ell\ell} < 80$ GeV,
	ningún jet b-tagged con $p_T > 30$ GeV

Tabla 3: Criterios de selección de la region de señal y las de control. La preselección se aplica a todo el análisis.

[3] Jets principales o "Leading" jets en este caso son los dos jets más energéticos.

3.4. Región de señal

La región de señal busca únicamente el estado final en el que un W se desintegra en un electrón y el otro a un muón (diferente sabor) ya que este estado final $e\mu$ tiene mucha menos contaminación de Drell-Yan que los estados e^+e^- o $\mu^+\mu^-$ (mismo sabor). Es posible hacer una análisis en la región de mismo sabor, pero requiere unos cortes y regiones de control diferentes [31].

Se aplican una serie de cortes para reducir los fondos dominantes (los que tienen sección eficaz mucho mayor que la señal). La selección que se lleva a cabo consiste en:

- Al menos dos jets con momento transverso mayor que 30 GeV.
- Los dos jets principales [3] deben ser centrales: $|\eta| < 2{,}5$.
- Masa invariante de estos jets (m_{jj}) menor que 200 GeV.
- Diferencia de pseudorapidez entre estos jets ($\Delta\eta_{jj}$) menor que $3{,}5$.
- Masa transversa $m_{\mathrm{T}}^{\mathrm{H}}$ entre 60 y 125 GeV.
- Distancia angular entre los dos leptones $\Delta R_{\ell\ell}$ menor que 2.
- No puede haber ningún jet b-tagged [4] (con p_T mayor que 20 GeV).

Los cortes en $\Delta\eta$ y m_{jj} hacen que nuestro espacio de fases sea ortogonal respecto a otros análisis del experimento CMS, como el de VBF. Esta segunda variable (m_{jj}) además es un buen discriminante entre los procesos de VH y ggH, como se verá en la Sección 4.1.

En la Figura 8 se muestran distribuciones de algunas variables en la región de señal. No muestran datos, solo las predicciones de MC, para evitar sesgar el análisis.

[4] Un "b-jet" o "b-tagged jet" es aquel que se ha identificado que proviene de un quark bottom (b).

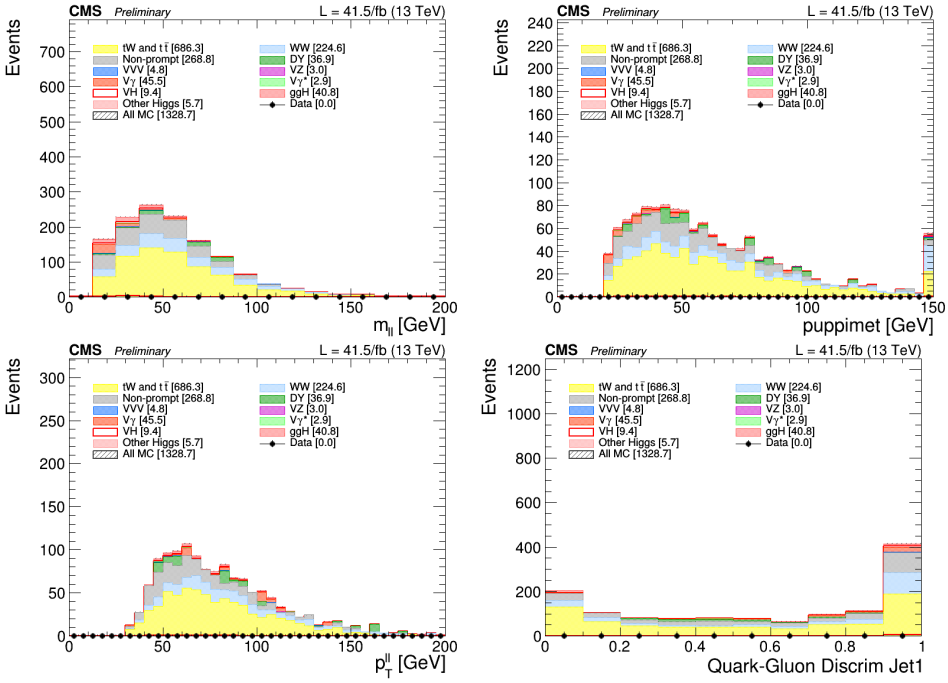

Figura 8: Distribuciones de la masa invariante del par de leptones (arriba izquierda), p_T de la MET (arriba derecha), momento transverso del par de leptones $p_T^{\ell\ell}$ (abajo izquierda) y discriminante QGL del jet principal (abajo derecha) para la región de señal. Los histogramas sólidos son las predicciones de MC.

3.5. Regiones de control

Las regiones de control son regiones cuyo espacio de fases es lo más parecido posible al espacio de fases de la región de señal, pero con cambios en unos pocos cortes concretos, con el fin de enriquecer estas regiones con un fondo determinado. Estas regiones de control se utilizan para estimar la normalización de los fondos de top ($\mathrm{t\bar{t}}$) y $\mathrm{DY}\tau\tau$ a partir de datos. Se han definido dos regiones de control:

- Región Top: Se requiere que al menos uno de los dos jets principales sea un jet b-tagged. Además se realiza un corte en la masa del sistema dileptónico: $m_{\ell\ell} > 50\ \mathrm{GeV}$ y se eliminan los cortes en $m_{\mathrm{T}}^{\mathrm{H}}$ y $\Delta R_{\ell\ell}$.
- Región $\mathrm{DY} \to \tau\tau$: Se requiere otro corte en la masa transversa: $m_{\mathrm{T}}^{\mathrm{H}} < 60\,\mathrm{GeV}$ y en la masa invariante del sistema dileptónico $40\,\mathrm{GeV} < m_{\ell\ell} < 80\,\mathrm{GeV}$. Se rechazan b-tagged jets con $p_T > 30\,\mathrm{GeV}$.

El resto de selecciones son idénticas entre la región de señal y las de control, incluidos los cortes realizados sobre los jets, para que los espacios de fase de señal y control sean lo más parecidos posible.
En las Figuras 9 y 10 se muestran algunas distribuciones de las regiones de control (Top y DY) con datos y MC.

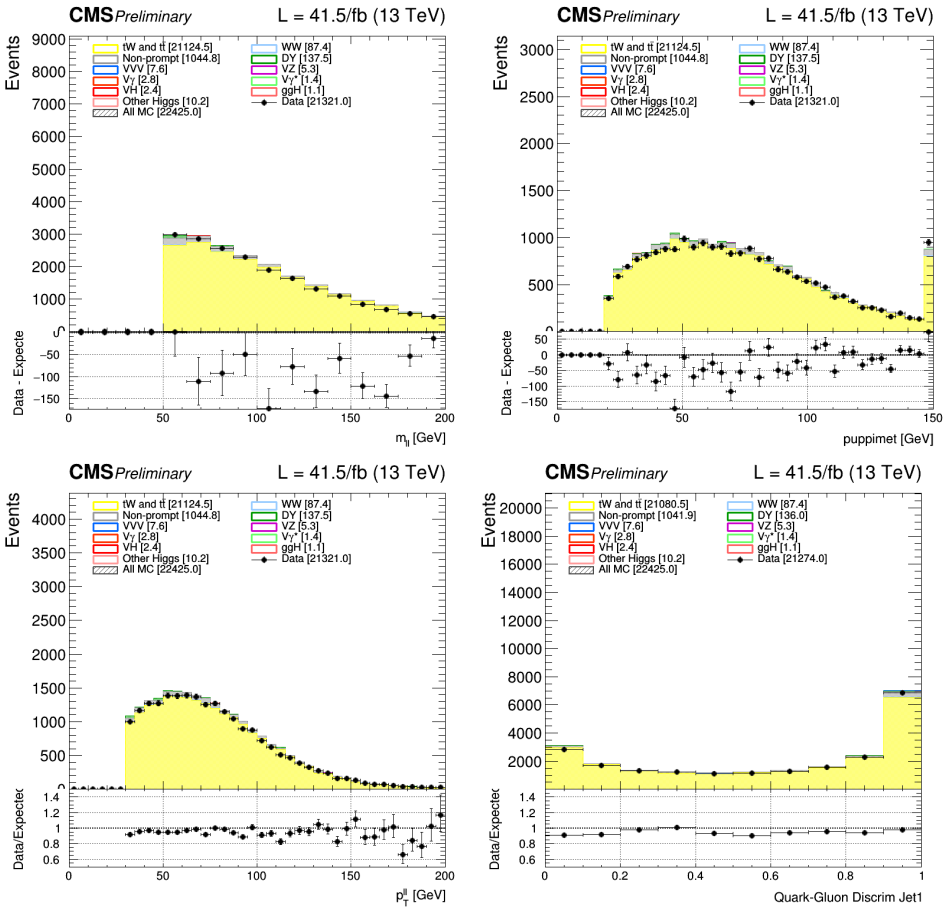

Figura 9: Distribuciones de la masa invariante del par de leptones (arriba izquierda), p_T de la MET (arriba derecha), momento transverso del par de leptones $p_T^{\ell\ell}$ (abajo izquierda) y discriminante QGL del jet principal (abajo derecha) para la región de control de top. Los puntos son los datos, los histogramas sólidos son las predicciones de MC. Aproximadamente el 95 % de esta región son sucesos de top.

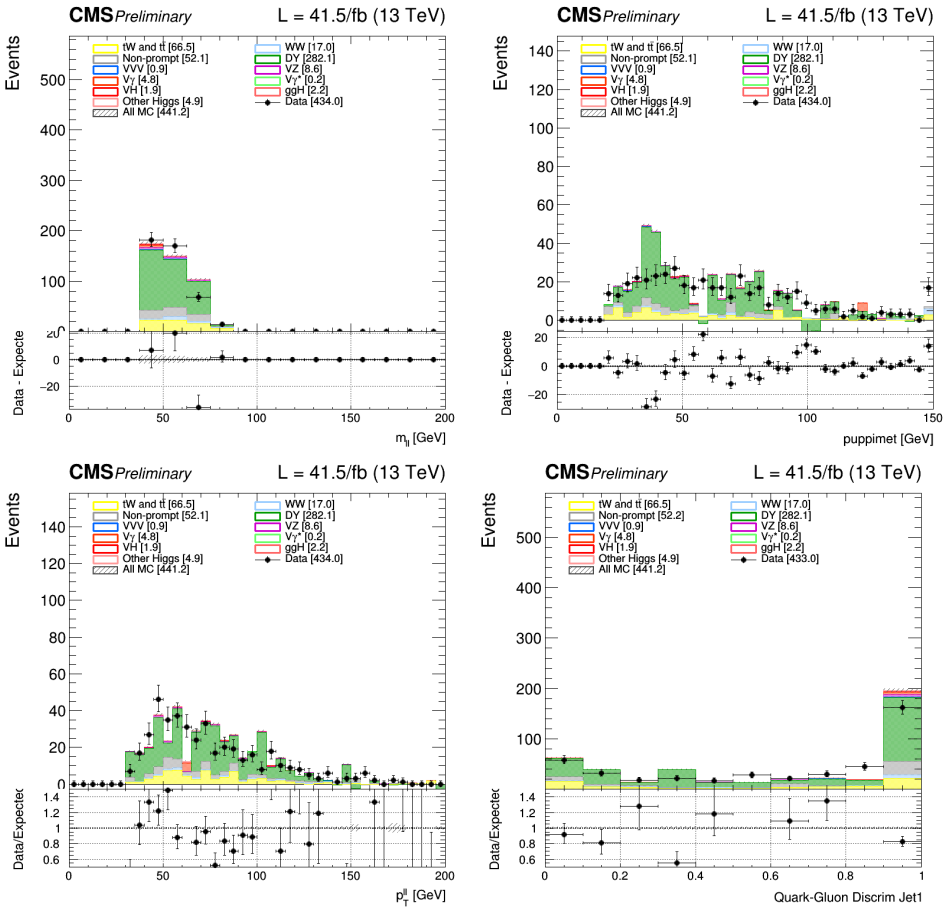

Figura 10: Distribuciones de la masa invariante del par de leptones (arriba izquierda), p_T de la MET (arriba derecha), momento transverso del par de leptones $p_T^{\ell\ell}$ (abajo izquierda) y discriminante QGL del jet principal (abajo derecha) para la región de control de Drell-Yan. Los puntos son los datos, los histogramas sólidos son las predicciones de MC. Aproximadamente el 65 % de esta región son sucesos de Drell Yan.

4. Análisis del canal VH 2j

Se ha trabajado sobre el canal VH 2j de dos maneras distintas: Con un análisis secuencial, basado en hacer cortes sobre variables para obtener los mejores resultados (mayor proporción de sucesos de señal frente a sucesos de fondo o significancia[5]), y con análisis multivariante para tratar de obtener las mejores selecciones posibles.

4.1. Análisis secuencial

Se han realizado dos análisis basados en cortes: Uno de ellos con los cortes de la región de señal (Tabla 3) más un corte en m_{jj}, y otro además con cortes en el *Quark-Gluon Likelihood Discriminant* (QGL) de los dos jets principales. El objetivo de los cortes aplicados es aumentar la pureza de la señal. Sin embargo, tienen el efecto negativo de reducir la eficiencia. Para encontrar los mejores valores posibles para cada corte se intenta maximizar la significancia, que es nuestra figura de mérito.

La variable m_{jj} es el mejor discriminador entre los mecanismos de producción VH (nuestra señal) y ggH (el fondo de Higgs más abundante), y por lo tanto cortar en esta variable nos permite aumentar la pureza de señal VH. Esto se debe a que en el proceso VH2j, los dos jets provienen de una resonancia del bosón W o Z, cuyas masas son: 80.4 GeV para el W y 91.2 GeV para el Z. Los jets en los procesos de ggH vienen de fondo no resonante de quarks y por ello su masa no tiene un pico. Ademas, se ha comprobado que este corte también elimina gran parte de otros fondos, como top. El corte que se realiza es: $65\,\mathrm{GeV} < m_{jj} < 105\,\mathrm{GeV}$. En la Figura 11 se puede observar la distribución (normalizada) de m_{jj} para VH y ggH.

En la Figura 12 se muestran algunas distribuciones de variables para la región de señal, aplicando los cortes de la Tabla 3 y el corte en m_{jj}.

Al análisis anterior se pueden añadir cortes en el discriminador QGL (Sección 2.2.2) para reducir aun más el fondo de ggH, aunque a cambio también perdemos parte de señal. El resultado es una pureza de señal aún más alta, pero con menos sucesos.

La variable $JetQGL$ nos da una idea del origen de los jets. En los procesos de VH, los jets deben provenir de quarks (QGL cercano a 1), mientras que en procesos ggH suelen provenir de gluones (QGL cercano a 0) como se ve en la Figura 13. Por lo tanto, los cortes que se aplican son: para el jet principal, $QGL > 0{,}4$, y para el segundo jet más energético, $QGL > 0{,}3$. Estos valores se optimizan para obtener la mayor significancia posible de señal.

[5] La significancia se define como $S/\sqrt{S+B}$, donde S es el número de sucesos de señal y B de fondo.

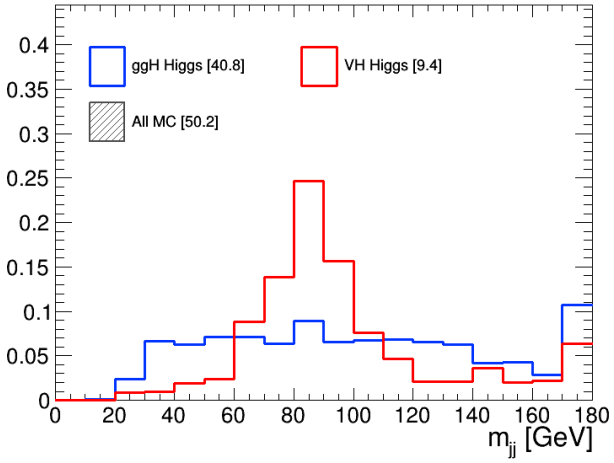

Figura 11: Distribución normalizada de m_{jj} para ggH (azul) y VH (rojo).

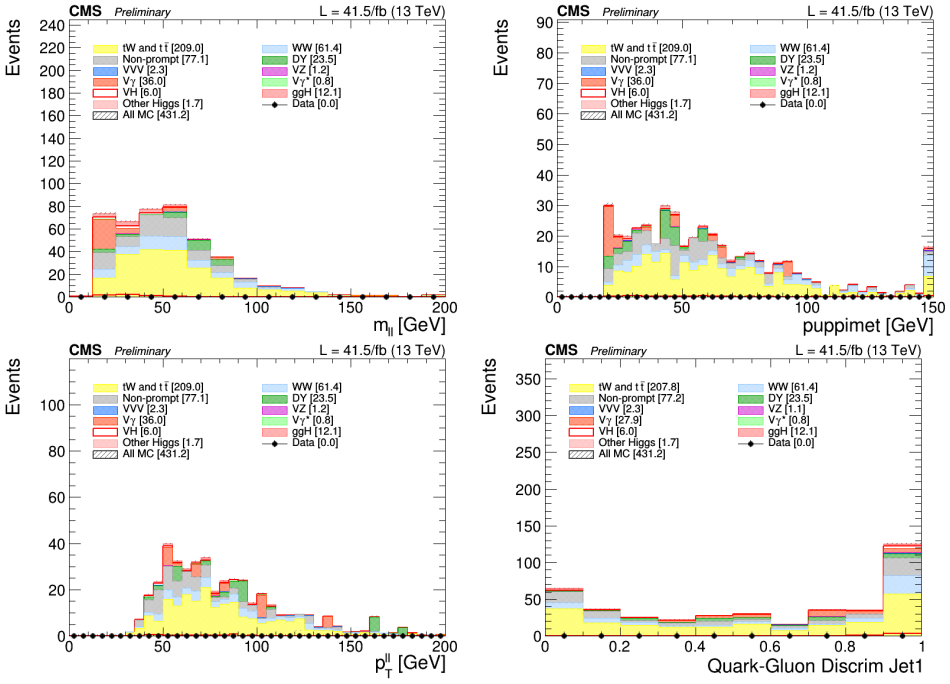

Figura 12: Distribuciones de la masa invariante del par de leptones (arriba izquierda), p_T de la MET (arriba derecha), momento transverso del par de leptones $p_T^{\ell\ell}$ (abajo izquierda) y discriminante QGL del jet principal (abajo derecha) para la región de señal $e\mu$, aplicando los cortes en m_{jj}.

En la Figura 14 se muestran algunas distribuciones de variables para la región de señal, aplicando los cortes de la Tabla 3 y los corte en m_{jj} y QGL. Las distribuciones no muestran datos (son *blinded*).

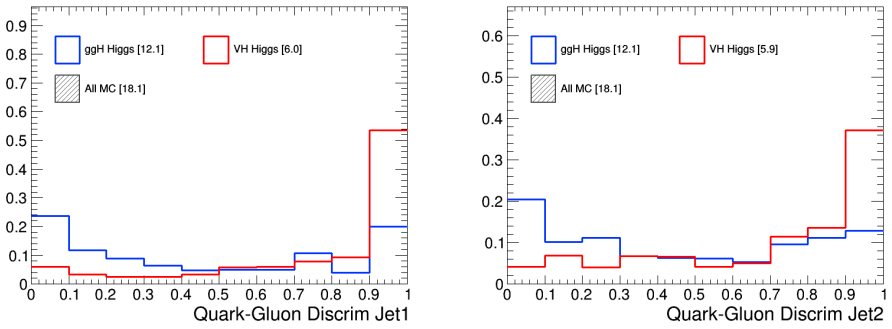

Figura 13: Distribución normalizada del discriminante QGL para el jet principal (izquierda) y el segundo jet más energético (derecha).

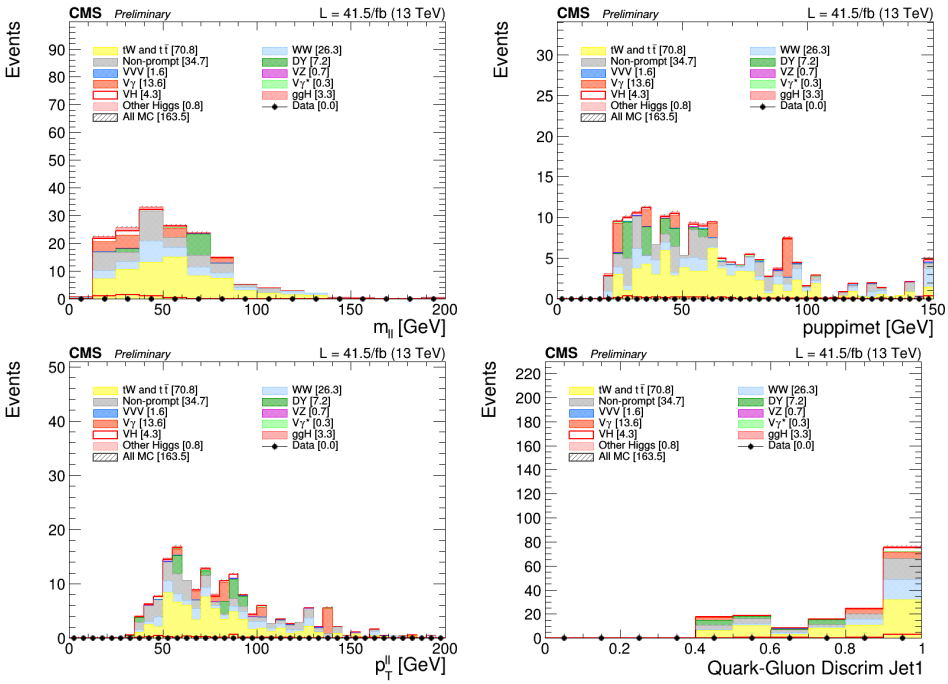

Figura 14: Distribuciones de la masa invariante del par de leptones (arriba izquierda), p_T de la MET (arriba derecha), momento transverso del par de leptones $p_T^{\ell\ell}$ (abajo izquierda) y discriminante QGL del jet principal (abajo derecha) para la región de señal $e\mu$, aplicando los cortes en el QGL y m_{jj}.

Para tener una idea de la efectividad de cada selección, se han separado los cortes de preselección y señal, y se ha calculado la eficiencia que se obtiene al aplicarlos, tanto individualmente, como al acumular cada corte con los anteriores. La base de la que se parte son todos los sucesos que contienen en su estado final un electrón y un muón de carga opuesta. Los resultados se muestran en la Tabla 4 para los sucesos de señal VH y Top, un fondo muy abundante. Se puede comprobar que los cortes eliminan una gran parte de señal, lo cual es un efecto secundario indeseado, pero eliminan una cantidad aún mayor de fondo. Tambien se ve como al eliminar los b-jets se reduce enormemente (92 %) el fondo top.

Selección VH	Ef. Individual	Ef. Acumulativa
μ y e de carga opuesta	100 %	100 %
$p_T^{\ell_1} > 25$, $p_T^{\ell_2} > 10\,(13)$ $\mu\,(e)$, $N_{lep} \geq 2$, $p_T^{\ell_3} < 10$	86,6 %	86,6 %
$m_{\ell\ell} > 12$, $p_T^{\ell\ell} > 30$	82,9 %	72,3 %
$p_T^{miss} > 20$	91,4 %	66,7 %
$N_{jets} \geq 2$	41,8 %	30,3 %
Jet $\eta < 2{,}5$	54,5 %	25,4 %
$\Delta\eta_{jj} < 3{,}5$	93,0 %	25,1 %
$60 < m_{\mathrm{T}}^{\mathrm{H}} < 125$	47,7 %	12,7 %
$\Delta R_{\ell\ell} < 2$	60,9 %	10,2 %
No bjets con $p_T > 20$	72,6 %	5,9 %
$65 < m_{jj} < 105$	27,7 %	3,2 %
Jet 1 QGL $> 0{,}4$, Jet 2 QGL $> 0{,}3$	32,4 %	2,3 %
Selección Top	**Ef. Individual**	**Ef. Acumulativa**
μ y e de carga opuesta	100 %	100 %
$p_T^{\ell_1} > 25$, $p_T^{\ell_2} > 10\,(13)$ $\mu\,(e)$, $N_{lep} \geq 2$, $p_T^{\ell_3} < 10$	96,1 %	96,1 %
$m_{\ell\ell} > 12$, $p_T^{\ell\ell} > 30$	87,3 %	84,0 %
$p_T^{miss} > 20$	94,0 %	79,0 %
$N_{jets} \geq 2$	73,6 %	58,7 %
Jet $\eta < 2{,}5$	75,7 %	49,6 %
$\Delta\eta_{jj} < 3{,}5$	93,0 %	48,8 %
$60 < m_{\mathrm{T}}^{\mathrm{H}} < 125$	39,2 %	20,0 %
$\Delta R_{\ell\ell} < 2$	38,3 %	8,4 %
No bjets con $p_T > 20$	8,0 %	0,25 %
$65 < m_{jj} < 105$	16,0 %	0,055 %
Jet 1 QGL $> 0{,}4$, Jet 2 QGL $> 0{,}3$	40,2 %	0,019 %

Tabla 4: Eficiencia de los cortes sobre la señal: individual (aplicando únicamente ese corte sobre todos los eventos) y acumulativa (aplicando el corte sobre todos los anteriores) para la señal VH y el fondo Top.

4.2. Análisis multivariante

Para el análisis multivariante también se ha usado el framework de ROOT [1], y en concreto, el paquete TMVA (Toolkit for Multivariate Data Analysis) [32].
Se ha usado el metodo BDT (boosted decision trees) [33], un algoritmo de clasificación basado en los llamados *decision trees* (árboles de decisión). Un árbol de decisión es una representación que se usa para clasificar datos. Se basa en dividir un set en dos regiones (nodos) en base a un corte en una cierta variable. Cada nodo se divide a su vez en dos con otro corte, y este proceso se repite hasta que todos los nodos satisfacen un criterio de clasificación. Un ejemplo de árbol real obtenido de una BDT en este trabajo se muestra en la Figura 15.
Este tipo de algoritmo es susceptible a fluctuaciones estadísticas en los datos. Para evitar esto se usa el *boosting*. Consiste en usar varios *decision trees* (clasificadores débiles) para obtener un algoritmo de clasificación mucho más potente.

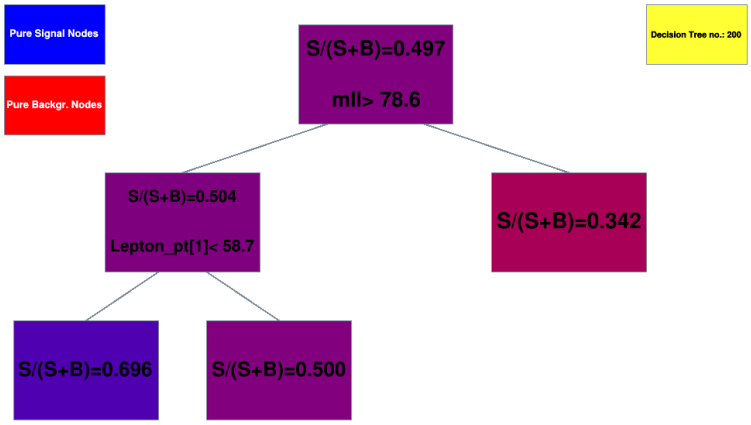

Figura 15: Ejemplo de un decision tree de la BDT que separa Higgs de Top.

Se pueden variar las características de los trees en el clasificador. Por simplicidad hemos usado los valores por defecto.

A la hora de aplicar una BDT a nuestros datos se hace una división en sucesos de entrenamiento y sucesos de test. El conjunto de entrenamiento se usa para ajustar los parámetros o pesos del clasificador. El conjunto de test es independiente del de entrenamiento, y se usa para comprobar la efectividad del modelo. Si el set de entrenamiento se ajusta mucho mejor que el de test, puede tratarse de un caso de sobre-entrenamiento (Figura 16).

En primer lugar se ha entrenado una BDT que separa los sucesos de

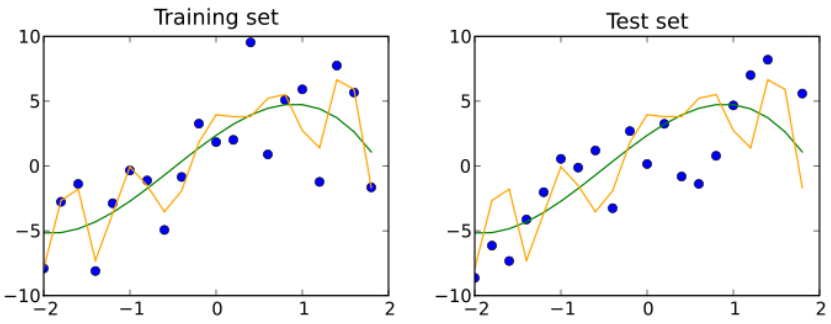

Figura 16: Ejemplo de conjuntos de entrenamiento y test (puntos azules) ajustados por dos modelos. El ajuste verde es correcto mientras que el naranja hace sobre-entrenamiento. [34].

Higgs del fondo principal, top. También se ha entrenado una BDT para separar nuestra señal (VH) del fondo ggH, que es la muestra de Higgs más abundante en nuestro análisis (al igual que ocurría con el análisis de 2016). Estas dos BDTs se aplican consecutivamente con el objetivo de tener unos resultados mejores que los obtenidos en el análisis por cortes.

Se ha probado con otros tipos de técnicas de análisis multivariante como una red neuronal profunda (DNNs) y un multi-layer perceptron (MLP) pero se escogió lal BDT porque daba resultados ligeramente mejores como se ve en la Figura 17.

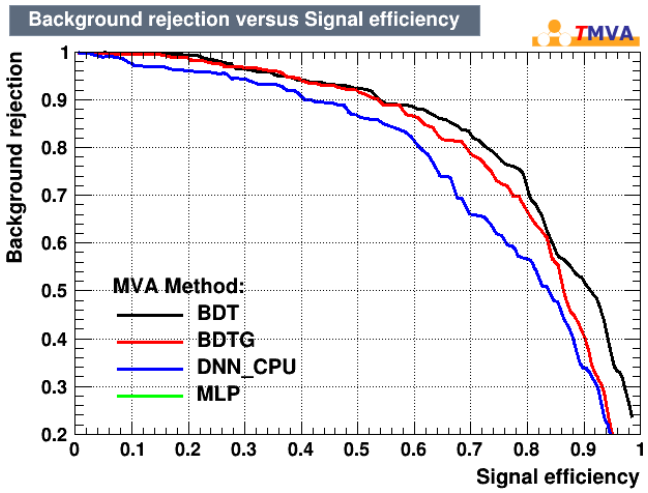

Figura 17: Representación de la eficiencia de señal frente a reducción de fondo para varios tipos de analisis MV.

4.2.1. BDT contra el fondo top

La BDT que separa el fondo de top se aplica en primer lugar, sobre la región de señal (Tabla 3). En esta selección se da prioridad a conservar el mayor numero posible de sucesos de señal (que en este caso son todos los Higgs: VH, ggH, qqH y ttH), aunque esto suponga reducir menos el fondo de top. Se aplica un corte muy suave, que reduce el número de sucesos de top en un $32\,\%$ y consigue una eficiencia de señal del $93\,\%$. En la Figura 18 se muestran las eficiencias en función del corte aplicado en la salida de la BDT y en la Figura 19 se muestra la distribución de señal y fondo. Los resultados del corte se recogen en la Tabla 5.

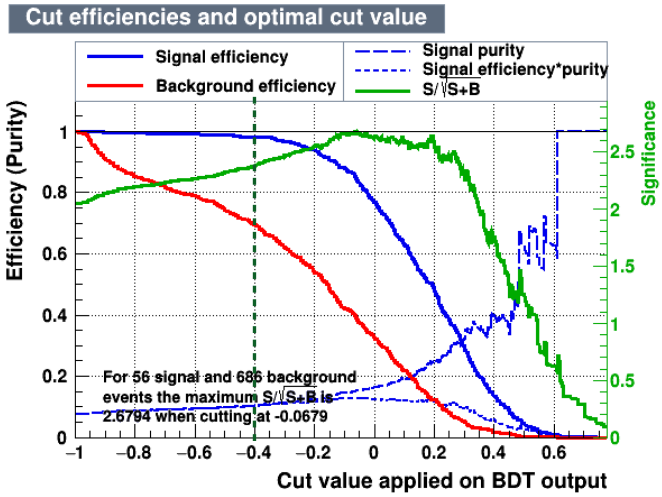

Figura 18: Valores de la eficiencia de señal (azul) y fondo (rojo) y significancia (verde) en función del corte que se aplique en el discriminante de la BDT. El corte en -0.4 se ha señalado con una linea punteada.

Corte	N Higgs	N top	Significancia	Ef. señal	Ef. fondo
$--$	55,9	686,3	2,05	$100\,\%$	$100\,\%$
$BDT > -0{,}4$	51,9	464,7	2,28	$93\,\%$	$68\,\%$

Tabla 5: Resultados de la BDT de top. Los valores iniciales (sin corte) se muestran en la primera fila (sin corte).

Para entrenar esta BDT se han usado las variables de la Tabla 6. Estas variables se han seleccionado tras varios entrenamientos con diferentes grupos de variables porque se ha encontrado que dan los mejores resultados. También se muestran los valores de importancia y separación que les asocia TMVA para dar una idea de su efectividad. La separación es una medida de la diferencia entre señal y fondo para una variable (se calcula

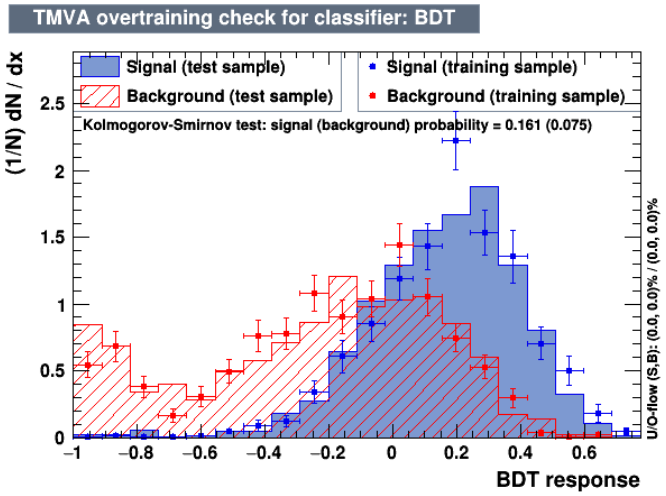

Figura 19: Distribución de muestras de test (histograma) y entrenamiento (puntos) de los sucesos de señal (azul) y fondo (rojo) para la BDT contra top.

cada vez que la BDT aplica un corte). La importancia es una medida de la frecuencia con que una variable se utiliza para separar dos nodos en un tree [35].

Variable	Separación	Importancia
$m_{\ell\ell}$	0,195	0,366
$\Delta\eta_{\ell\ell}$	0,083	0,053
$\Delta\phi_{\ell\ell}$	0,082	0,058
$m_{\mathrm{T}}^{\mathrm{H}}$	0,061	0,112
$min\Delta\eta_{j\ell}$	0,058	0,081
mjj	0,046	0,053
η_{ℓ_1}	0,044	0,036
$\Delta\phi_{jj}$	0,040	0,063
p_T^{j1}	0,033	0,054
η_{j_2}	0,029	0,049
$p_T^{\ell_2}$	0,027	0,074

Tabla 6: Lista de variables ordenadas por su capacidad de separación en la BDT.

En las Figuras 20 y 21 se muestran las distribuciones de las variables que se han utilizado para entrenar la BDT y las matrices de correlación de estas. La variable $min\Delta\eta_{j\ell}$ calcula la diferencia de η entre uno de los dos jets principales y uno de los dos leptones de mayor momento ($\Delta\eta_{j_1\ell_1}$, $\Delta\eta_{j_1\ell_2}$,

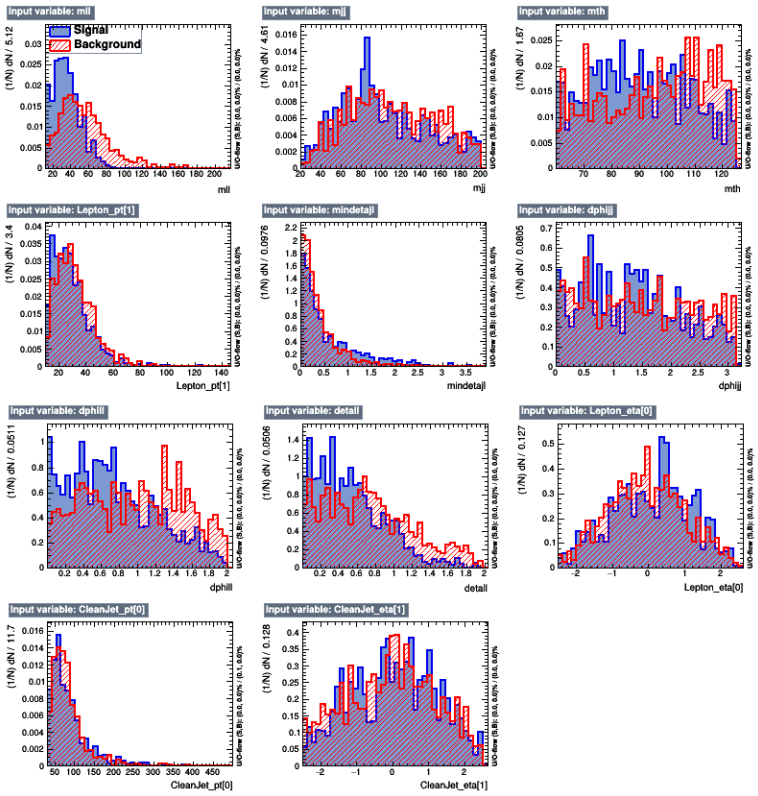

Figura 20: Distribución de las variables normalizadas de la BDT para señal (azul) y fondo top (rojo). De izquierda a derecha y de arriba a abajo: $m_{\ell\ell}$, m_{jj}, $m_{\mathrm{T}}^{\mathrm{H}}$, $p_T^{\ell_1}$, $min\Delta\eta_{j\ell}$, $\Delta\phi_{jj}$, $\Delta\phi_{\ell\ell}$, $\Delta\eta_{\ell\ell}$, η_{ℓ_1}, $p_T^{j_1}$, η_{j_2}.

$\Delta\eta_{j_2\ell_1}$, $\Delta\eta_{j_2\ell_2}$), y devuelve el menor valor.

En las matrices de correlación no se observa ningún comportamiento inusual, la correlación es baja excepto en algunos pares de variables que están relacionados, como $\Delta\phi_{\ell\ell}$, $\Delta\phi_{jj}$ y la masa del par de leptones o jets.

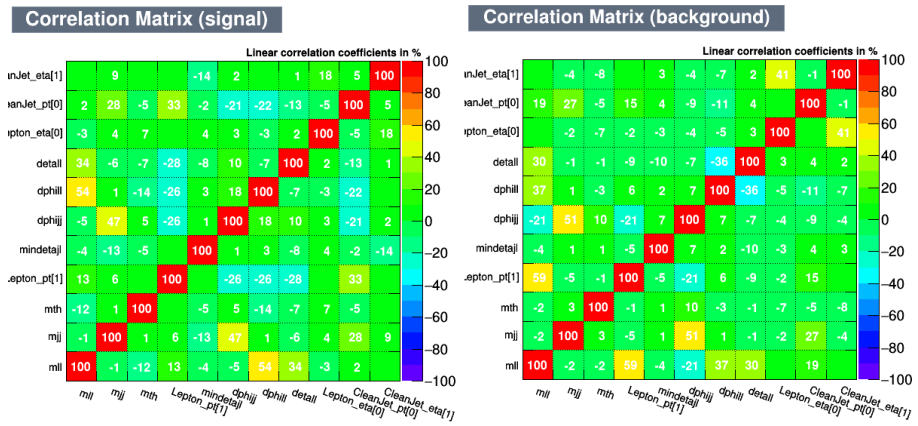

Figura 21: Matrices de correlación de las variables para señal (Higgs) y fondo (top).

4.2.2. BDT contra el fondo ggH

Esta BDT se ha entrenado de dos maneras: Primero se entrena con aplicando unicamente la selección de la región de señal sobre los sucesos. Con esto se obtienen unos resultados en los que hay gran presencia de otros fondos, como top. Entonces se hace un nuevo entrenamiento de la BDT aplicando los cortes de la región de señal y también el corte en la BDT de top. Para hacer esto, una vez calculado el corte en la BDT de top, se aplica directamente sobre los sucesos, obteniéndose un subconjunto de estos con el corte aplicado. Con estos nuevos sucesos se entrena de nuevo la BDT que separa ggH y VH. El objetivo es obtener la mejor significancia posible con VH como señal y ggH como fondo. El corte que logra esto nos lo da automáticamente TMVA, y los resultados se anotan en la Tabla 7.

Corte	N VH	N ggH	Significancia	Ef. señal	Ef. fondo
$--$	8,8	38,1	1,28	100 %	100 %
$BDT > 0{,}12$	5,14	3,93	1,71	58 %	10 %

Tabla 7: Resultados de la BDT de VH. Los valores iniciales, sin cortes, sobre los que se aplica son: $8{,}8$ sucesos de señal y $38{,}1$ de fondo, y se muestran en la primera fila.

En la Figura 22 se muestran las eficiencias y significancia en función del corte que se aplique. En la Figura 23 se muestra la distribución de los sucesos de VH y ggH.

Para entrenar esta BDT se han usado otra serie de variables, anotadas en la Tabla 8. También se muestran los valores de importancia y separación que obtienen.

Variable	Separación	Importancia
m_{jj}	0,250	0,143
QGL Jet 2	0,206	0,128
QGL Jet 1	0,198	0,144
$m_{\mathrm{T}}^{\mathrm{H}}$	0,130	0,061
$\Delta\eta_{jj}$	0,123	0,081
$min\Delta\eta_{j\ell}$	0,119	0,227
$m_{\ell\ell}$	0,091	0,075
p_T^{WW}	0,074	0,034
$p_T^{\ell_1}$	0,070	0,069
p_T^{jj}	0,064	0,000
$p_T^{\ell_2}$	0,059	0,039

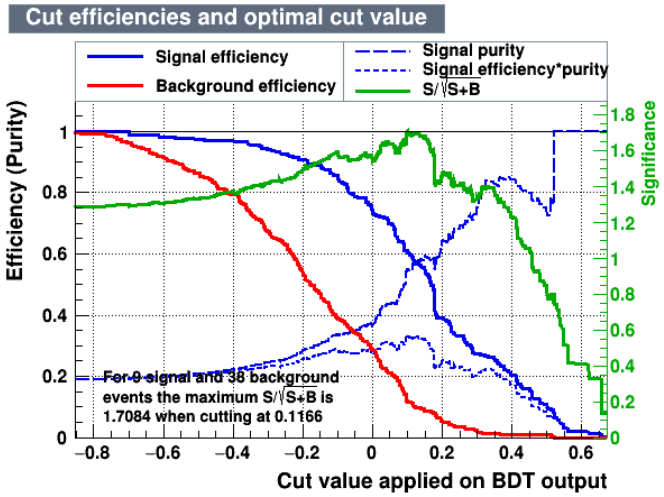

Figura 22: Valores de la eficiencia de señal (azul) y fondo (rojo) y significancia (verde) en función del corte que se aplique en el discriminante de la BDT.

Tabla 8: Lista de variables ordenadas por su capacidad de separación en la BDT de VH.

En las Figuras 24 y 25 se muestran las distribuciones de las variables que se han utilizado para entrenar la BDT de VH contra ggH y las matrices de correlación de estas.

En la Figura 26 se muestran algunas distribuciones de variables en la región de señal despues de aplicar los cortes en las BDTs.

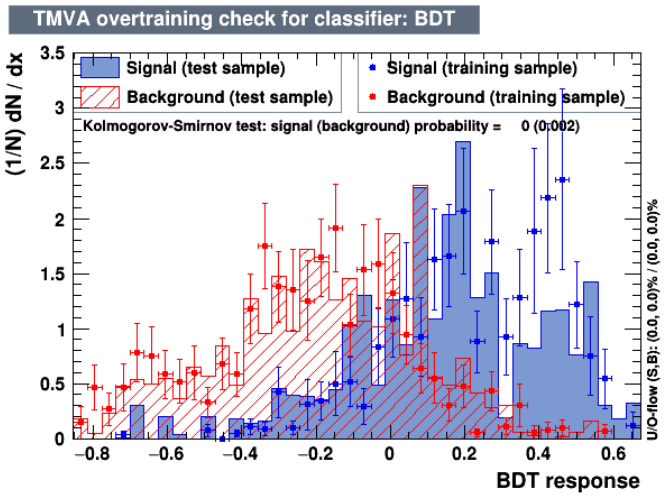

Figura 23: Distribución de muestras de test (histograma) y entrenamiento (puntos) de los sucesos de señal (azul) y fondo (rojo) para la BDT contra ggH.

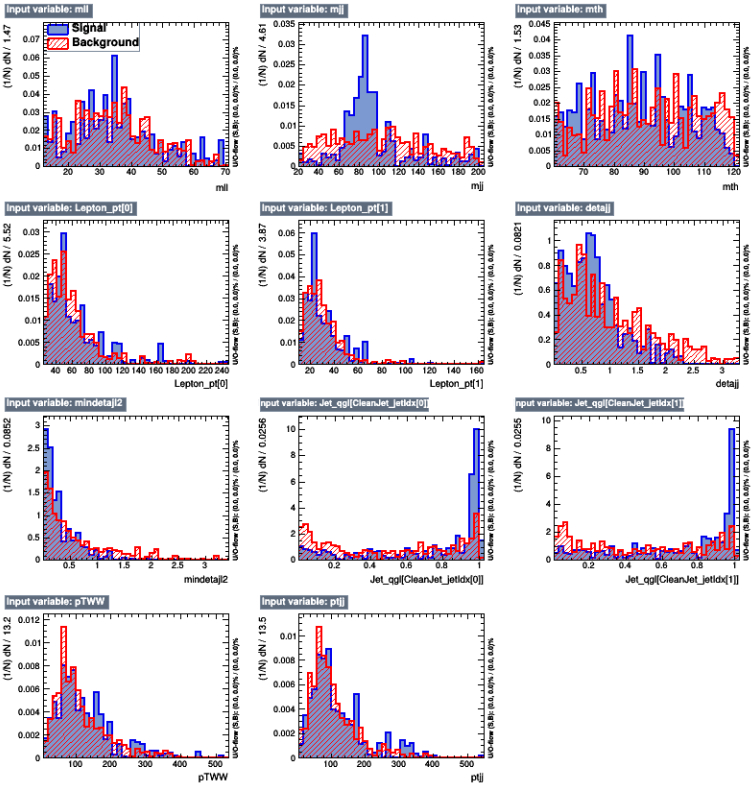

Figura 24: Distribución de las variables normalizadas de la BDT para señal (azul) y fondo ggH (rojo). . De derecha a izquierda y de ariba a abajo: $m_{\ell\ell}$, m_{jj}, $m_{\mathrm{T}}^{\mathrm{H}}$, $p_T^{\ell_1}$, $p_T^{\ell_2}$,$\Delta\eta_{jj}$, $min\Delta\eta_{j\ell}$, QGL Jet 1, QGL Jet 2, p_T^{WW} , p_T^{jj}.

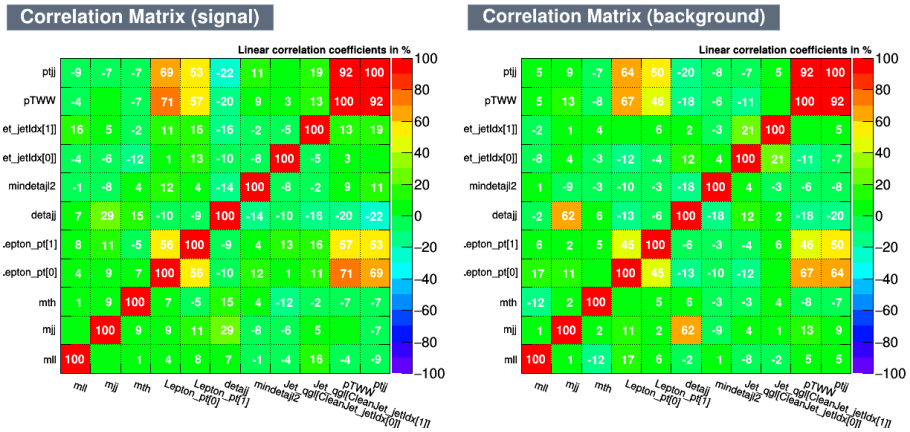

Figura 25: Matrices de correlación de las variables para señal (VH) y fondo (ggH).

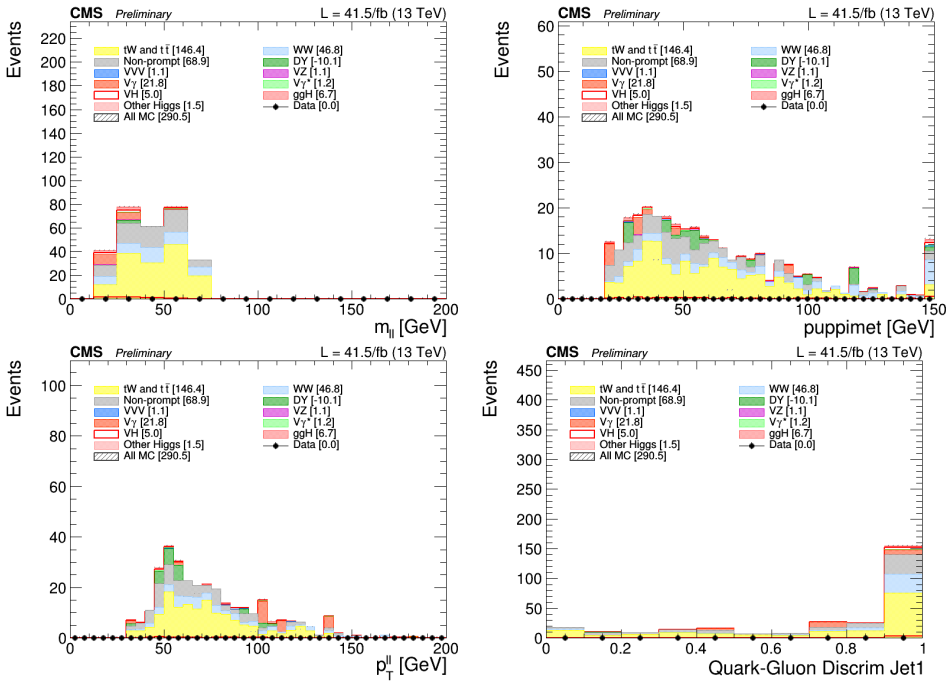

Figura 26: Distribuciones de la masa invariante del par de leptones (arriba izquierda), p_T de la MET (arriba derecha),c momento transverso del par de leptones $p_T^{\ell\ell}$ (abajo izquierda) y discriminante QGL del jet principal (abajo derecha) para la región de señal con cortes en ambas BDTs.

5. Resultados

En la Tabla 9 se muestra el numero de sucesos de señal (VH2j) y fondo que se obtienen tras aplicar las distintas selecciones que hemos usado. Siempre aplicamos los cortes de la región de señal, S, y luego hacemos una selección con el corte en m_{jj}, otra añadiendo el corte en QGL y otras dos mediante el análisis con BDTs: primero cuando se aplica la BDT que separa ggH sola, y después ambas BDTs (la de ggH y la de top) aplicadas simultáneamente. También se muestran la fracción de señal y la significancia de cada una. Con estos datos podemos valorar la eficacia de cada selección. En la Tabla 10 se muestran los mismos parámetros pero teniendo solo en cuenta el fondo de ggH.

Selección	N Señal	N Fondo	Frac. Señal	Significancia
S + Cortes m_{jj}	6,0	425,2	1,39 %	0,29
S + Cortes $QGL + m_{jj}$	4,3	159,2	2,70 %	0,34
S + BDT ggH	6,4	766,3	0,82 %	0,23
S + Ambas BDTs	5,0	285,5	1,75 %	0,30

Tabla 9: Eventos de señal y fondo, fracción de señal y significancia para los análisis por cortes y las BDTs. Se incluyen todos los fondos.

Selección	N Señal	N Fondo	Frac. Señal	Significancia
S + Cortes m_{jj}	6,0	12,1	33,1 %	1,41
S + Cortes $QGL + m_{jj}$	4,3	3,3	56,6 %	1,56
S + BDT ggH	6,4	7,4	46,4 %	1,72
S + Ambas BDTs	5,0	6,7	42,7 %	1,46

Tabla 10: Eventos de señal y fondo, fracción de señal y significancia para los análisis por cortes y la BDT. Solo se tiene en cuenta el fondo ggH.

Con estos resultados no podemos definir un claro ganador entre las selecciones. La significancia mas alta frente a todos los fondos (0,34) se consigue mediante cortes en m_{jj} y el discriminador QGL, también se tiene la mayor fracción de señal (2,7 %), sin embargo, este método también deja el menor numero absoluto de sucesos de señal (solo 4,3). Al aplicar ambas BDTs se obtiene una significancia también alta (0,3) con un numero de sucesos de señal ligeramente mayor (5,0) pero menor fracción de señal. La selección mas básica, con un único corte en m_{jj}, obtiene una significancia casi tan alta como las BDTs (0,29) con mas sucesos de señal y menor fracción de señal (bastante más sucesos de fondo).
Finalmente, al aplicar solo la BDT de VH frente a ggH se obtienen los resultados esperados: los mejores resultados de significancia cuando se tiene

en cuenta el fondo de ggH únicamente, pero obtenemos resultados no tan buenos con todos los fondos. Por último, también se debe tener en cuenta que aplicar BDTs añade una mayor complejidad al análisis: Se debe realizar un proceso de optimización, se añaden fuentes de error sistemático, nuevos pesos y variables usados por la BDT, etc. Por lo tanto, se suele considerar necesario obtener una mejora considerable en los resultados para que merezca la pena usar una BDT en el análisis.

En la Figura 27 se muestran la fracción de los diferentes modos de producción del Higgs que se obtienen en los dos análisis por cortes. Nuestra señal VH está separada en WH y ZH.

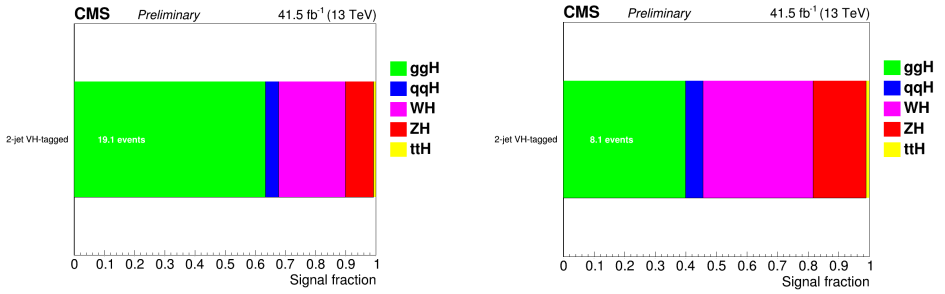

Figura 27: Plots de las fracciones de Higgs para el análisis con cortes en m_{jj} (izquierda) y el análisis con cortes en m_{jj} y QGL (derecha). Se muestran ggH (verde), VBF (azul), VH (rosa y rojo) y ttH (amarillo).

En la Figura 28 se muestran la fracción de los diferentes modos de producción del Higgs que se obtienen mediante el análisis multivariante, aplicando la BDT de ggH por separado, y ambas BDTs simultáneamente.

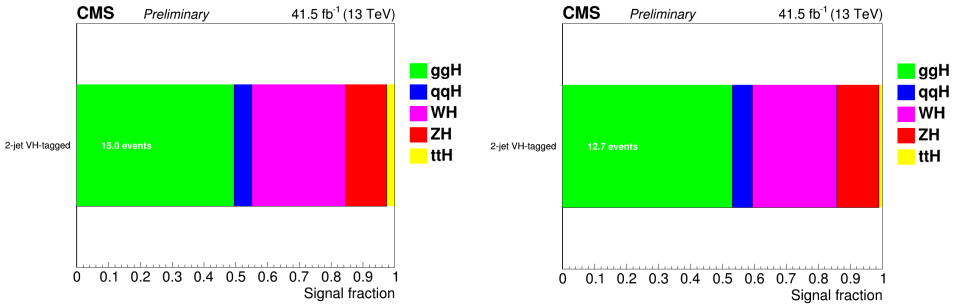

Figura 28: Plots de las fracciones de Higgs para el análisis con BDT de ggH (izquierda) y el análisis con ambas BDTs (derecha).

6. Conclusiones

Este trabajo se ha enmarcado en el estudio de la producción y desintegración del bosón de Higgs en el detector CMS del CERN. Se ha hecho un estudio del canal VH2j de desintegración del bosón de Higgs a partir de los datos obtenidos por CMS en 2017. Se han realizado dos análisis basados en cortes y uno mediante análisis multivariante.

En primer lugar se han mostrado los objetos físicos y las variables que son fundamentales a la hora de realizar este análisis, y que son comunes a varios de los grupos de trabajo de CMS. También se han mostrado las regiones de señal y control con las que hemos trabajado, con sus correspondientes selecciones.

El primer análisis que se ha realizado consiste en la aplicación de cortes directamente sobre dos variables, en primer lugar sobre la masa invariante de los dos jets principales, m_{jj}, y después un corte sobre el discriminante quark-gluon (QGL) de estos dos jets. El corte sobre m_{jj} aumenta la significancia de nuestra señal, VH, reduciendo considerablemente el fondo de ggH y otros en menor medida. Aplicando ambos cortes (m_{jj} y QGL) se consigue aumentar más aun la significancia de la señal, reduciendo mucho todos los fondos, aunque a cambio disminuye el número de sucesos total de VH.

El segundo análisis principal que se ha realizado consiste en el entrenamiento de dos BDTs con el fin de reducir dos fondos principales: el fondo de quarks top se reduce con una primera BDT de manera suave, y posteriormente el fondo de bosones de Higgs provenientes de la fusión de gluones se reduce considerablemente con una segunda BDT. El resultado es un aumento de la significancia de nuestra señal, respecto al análisis con corte en m_{jj}. No llega a tener una significancia tan alta como el análisis con cortes en m_{jj} y QGL, sin embargo, tiene un mayor número de sucesos de señal. Es posible que se puedan mejorar estos resultados optimizando aún mas las BDTs o utilizando un método TMVA diferente.

Los próximos pasos a realizar serían los siguientes. Para cada selección estudiada (dos secuenciales y dos usando BDT) habría que considerar los errores estadísticos y definir los errores sistemáticos. Con las diferentes fuentes de error (teóricas y experimentales) bien caracterizadas habría que realizar un ajuste de los sucesos esperados a los datos tomados por CMS. En el ajuste los parámetros libres serían las normalizaciones de las regiones de control y el número total de bosones de Higgs esperados (incluyendo, entre otros, VH y ggH). Esto no se ha podido realizar por falta de tiempo. Una vez realizado el procedimiento anterior para las cuatro selecciones do-

cumentadas en este trabajo se podrá ver, con precisión, cuál presenta una mayor pureza de producción asociada del bosón de Higgs con la menor incertidumbre (estadística y sistemática) posible. Finalmente, todo el trabajo realizado con los datos de 2017 habrá de ser completado con datos de 2016 y 2018. Hacerlo disminuirá el error estadístico del resultado, pero aumentará la complejidad del análisis, pues el comportamiento del detector CMS no es exacto entre los tres diferentes años. Por ejemplo, las condiciones de las colisiones han empeorado de 2016 a 2017 y 2018, concretamente el LHC ha proporcionado mayores luminosidades al precio de que las colisiones están más contaminadas de colisiones secundarias simultáneas.

Una vez completado, este trabajo formará parte de la publicación de CMS para el análisis $H \rightarrow WW$ con datos de todo el Run 2 (2016, 2017 y 2018).

Referencias

[1] About ROOT. https://root.cern.ch/about-root.

[2] Mark Thomson. *Modern particle physics*. 2013, Cambridge University Press.

[3] United States Department of Energy Particle Data Group Fermilab, Office of Science.
. https://commons.wikimedia.org/w/index.php?curid=4286964.

[4] Heather Gray; Bruno Mansoulié.
The Higgs boson: the hunt, the discovery, the study and some future perspectives . https://atlas.cern/updates/atlas-feature/higgs-boson.

[5] M. Tanabashi et al. (Particle Data Group). 2019 Review of Particle Physics., 2018. Phys. Rev. D 98, 030001.

[6] LHC Higgs Cross Section Working Group; Dittmaier; Mariotti; Passarino; Tanaka; Alekhin; Alwall; Bagnaschi; Banfi . Handbook of LHC Higgs Cross Sections: 2. Differential Distributions, 2012. CERN Report 2.

[7] Timothy Rias.
Higgs branching ratios. https://commons.wikimedia.org/w/index.php?curid=20667408.

[8] Timothy Rias.
Feynmann Diagram of Higgs production. https://commons.wikimedia.org/w/index.php?curid=20417763.

[9] The Higgs Boson . https://home.cern/science/physics/higgs-boson.

[10] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, 2012. Physics Letters B, Volume 716, Issue 1, Pages 1-29.

[11] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, 2012. Physics Letters B, Volume 716, Issue 1, Pages 30-61.

[12] S. Myers R. Ostojic J. Poole P. Proudlock O. Bruning P. Collier, P. Lebrun. LHC Design Report Volume II: The LHC Infrastructure and General Services, Noviembre 2004. CERN-2004-003.

[13] The CMS Collaboration. CMS Physics Technical Design Report Volume I: Detector Performance and Software, 2006. CMS TDR 8.1.

[14] CMS Collection David Barney. CMS Detector Slice. CMS-PHO-GEN-2016-001-1.

[15] The CMS Collaboration. CMS The Muon Project Technical Design Report, 1997. CMS-TDR-3.

[16] The CMS Collaboration. CMS Technical Design Report for the Level-1 Trigger Upgrade, 2013. CMS-TDR-12.

[17] V. Khachatryan et al. The CMS trigger system, 2017. INST12 P01020.

[18] CMSSW Twiki, The CMS Offline WorkBook.
https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBook.

[19] Francisco R. Villatoro.
CMS no observa el efecto magnético quiral sugerido por STAR y ALICE . https://francis.naukas.com/2017/06/21/cms-no-observa-efecto-magnetico-quiral-sugerido-star-alice/.

[20] The CMS Collaboration. Measurements of propierties of the Higgs boson decaying to a W boson pair in pp collisions at $\sqrt{s} = 13$ TeV. 2018.

[21] Twiki: Quark-gluon likelihood at 13 TeV .
https://twiki.cern.ch/twiki/bin/viewauth/CMS/QuarkGluonLikelihood.

[22] The CMS Collaboration. Common analysis object definitions and trigger efficiencies for the $H \rightarrow WW$ analysis with 2016 full data, 2017. AN-2017/082.

[23] The CMS Collaboration. Common analysis object definitions and trigger efficiencies for the $H \rightarrow WW$ analysis with full Run-II data, 2019. AN-2019/105.

[24] Clara Jordá Lope. Medida de la sección eficaz de producción del proceso WW en colisiones pp a $\sqrt{s} = 7$ TeV en el experimento CMS del LHC. Master's thesis, Universidad de Cantabria, 2012.

[25] CMS.
https://twiki.cern.ch/twiki/pub/CMSPublic/PhysicsResultsCombined/SigmaNew_v0.pdf, 2019.

[26] H. Bakhshian et al. Computing the contamination from fakes in leptonic final states, 2010. AN-2010/261.

[27] The CMS Collaboration. Higgs to WW leptonic differential measurements using 2016 and 2017 data sets, 2019. AN-2019/006.

[28] S. ALioli, P. Nason, C. Oleari, and E. re. NLO Higgs boson production via gluon fusion matched with shower in POWHEG, 2009. JHEP 04 (2009) 002.

[29] MG/ME Development team. The MadGraph5_aMC@NLO homepage, 2019. http://madgraph.phys.ucl.ac.be/.

[30] T. Sjostrand, S. Mrenna, and P. Z. Skands. A brief introduction to PYTHIA 8.1, 2008. Comput. Phys. Commun. 178(2008) 852-867.

[31] HWW team. Higgs to WW Measurements at 13 TeV Using Full 2016 Dataset, 2018. AN-2017/260.

[32] Kim Albertsson.
TMVA summary. https://root.cern.ch/tmva/summary.

[33] Decision trees . https://en.wikipedia.org/wiki/Decision_tree.

[34] Skbkekas.
Fig. https://commons.wikimedia.org/w/index.php?curid=6758064.

[35] Kim Albertsson et al. TMVA 4 Toolkit for Multivariate Data Analysis with ROOT, Users guide, 2018. CERN-OPEN-2007-007.