



Facultad de Ciencias

Técnicas de aprendizaje automático profundo para la asignación de momento a muones altamente energéticos en el experimento CMS del LHC

(Deep learning techniques for muon momentum assignment with the CMS experiment at LHC)

TRABAJO DE FIN DE MÁSTER
PARA ACCEDER AL
**MÁSTER EN
CIENCIA DE DATOS**

Autor : Pedro José Fernández Manteca
Director : Pablo Martínez Ruíz del Árbol
Co-director : Alicia Calderón Tazón

Septiembre - 2020

*A Pablito por ser un director chuli
Para los agradecimientos
Para los agradecimientos
Para los agradecimientos
Para los agradecimientos
Para los agradecimientos*

Índice

1. Introducción	4
2. El experimento CMS	8
2.1. Tracker	9
2.2. Tubos de deriva: DTs	10
2.3. Cámaras de tiras catódicas: CSCs	10
2.4. Segmentos en las cámaras de muones	10
3. Aprendizaje automático y su uso en física de altas energías	13
4. Asignación de momento transverso en el experimento CMS	15
4.1. Reajuste TPFMS (<i>Tracker-plus-first-muon-station</i>)	15
4.2. Reajuste Picky	15
4.3. Reajuste DYT (<i>Dynamic truncation</i>)	16
4.4. Asignación final de momento transverso: el algoritmo TuneP	16
5. Método propuesto	17
5.1. Herramientas utilizadas para el análisis	17
5.2. Muestra de simulación utilizada	18
5.3. Preparación del conjunto de datos: selección de muones y segmentos .	19
5.4. Distribuciones de control	20
6. Implementación y entrenamiento de la DNN	26
6.1. Conjunto de datos y arquitectura de la red	26
6.2. Variables de entrenamiento	27
7. Resultados	29
8. Conclusiones	32

Resumen

El objetivo de este trabajo es aplicar técnicas de aprendizaje automático profundo supervisado para la asignación de momento transverso a muones altamente energéticos en el experimento CMS (*Compact Muon Solenoid*) del LHC (*Large Hadron Collider*) mediante regresión, teniendo como principal meta mejorar los resultados del procedimiento actual, y de ser así plantear incluir este tipo de metodologías de manera oficial en la futura toma de datos del experimento (Run 3).

Para el entrenamiento se ha utilizado una muestra de simulación Monte-Carlo, donde una partícula masiva se desintegra a muones de las características buscadas, de manera que el algoritmo predictivo se alimenta de la información que estos muones dejan a su paso por el detector CMS.

FILL ME: RESULTADOS

Palabras clave: aprendizaje automático profundo, muones altamente energéticos, CMS

Resumen

The aim of this work is to apply supervised deep machine learning techniques for the transverse momentum assignment to highly energetic muons in the CMS (*Compact Muon Solenoid*) experiment at LHC (*Large Hadron Collider*) collider through regression, with the main goal of improving the results of the current procedure, and if so, consider including this kind of methodologies for the future data taking of the experiment (Run 3).

For the training, a Monte-Carlo simulation sample has been used, where a massive particle decays into muons of the desired characteristics, so that the predictive algorithm feeds on the information that these muons leave as they pass through the CMS detector.

FILL ME: RESULTADOS

Key words: deep learning, highly energetic muons, CMS

1. Introducción

El objeto principal de estudio de este trabajo es el muón, una partícula elemental cargada con spin 1/2 y con masa aproximadamente 200 veces mayor que el electrón, siendo además una partícula inestable, con un tiempo de vida de $2.2\mu s$, que es elevado en comparación con otras partículas que poseen esta característica.

En este trabajo nos centraremos específicamente en la detección y reconstrucción de muones en el detector CMS (del inglés *Compact Muon Solenoid*), situado en el gran colisionador de hadrones LHC (del inglés *Large Hadron Collider*). La mayor parte de los muones medidos por CMS provienen típicamente de desintegraciones de quarks top, de hadrones o de desintegraciones leptónicas de bosones Z o W [1]. Estos muones se caracterizan normalmente por tener un momento transverso (p_T), la componente del momento del muon en el plano perpendicular a la dirección de choque de los protones, inferior a 200 GeV y se categorizan como muones de bajo momento o *low- p_T muons*.

Por otra parte, los muones de alto momento o *high- p_T muons* pueden tener como origen procesos físicos atípicos como la desintegración de partículas exóticas fuera del Modelo Estándard de la física de partículas [2] como bosones Z' o W' [3, 4] con masas en la escala del TeV. El descubrimiento de alguna de estas partículas sería un indicativo directo de nueva física, por lo que medir las propiedades de estos muones, de la manera más precisa posible en el detector, es de vital importancia. Concretamente, el objetivo principal del trabajo consiste en aplicar una técnica de regresión basada en una red neuronal profunda o DNN (del inglés *Deep Neural Network*) [5] para estimar el momento transverso de los *high- p_T muons* con precisión.

Experimentalmente, la medida del momento de los *high- p_T muons* plantea varias dificultades. Primero, hay que tener en cuenta que la resolución de la medida de p_T a partir de la traza empeora cuando el momento del muón aumenta. Esto es así debido a que el momento transverso de los muones se mide a través de la relación entre curvatura y momento que establece la fuerza de Lorentz. En presencia de un campo magnético uniforme B , y con un radio de curvatura de la traza R , el momento transverso p_T de un muón con carga q se puede expresar como:

$$p_T[\text{GeV}] = 0.3 \times B[\text{T}] \times R[\text{m}] \times q \quad (1)$$

El campo magnético dentro del solenoide de CMS es prácticamente uniforme y conocido con gran precisión ($B = 3.8$ T), mientras que el radio de curvatura se relaciona con la longitud del arco L y la distancia sagitta s definida en la Figura 1 a través de:

$$R[\text{m}] \approx L[\text{m}]^2 / 8s[\text{m}] \quad (2)$$

Esta aproximación es válida para $L/R \ll 1$. Combinando ahora las ecuaciones (1) y (2), se obtiene:

$$s[\text{m}] \approx (0.3B[\text{T}]L[\text{m}]^2/8)(q/p_T[\text{GeV}]) = (0.3BL^2/8) \times (q/p_T), \quad (3)$$

Se observa que s es inversamente proporcional al momento transverso, por lo que para mejorar la resolución en la medida del p_T en los casos con sagittas pequeñas, las trazas de los muones en CMS se reconstruyen en distintos subdetectores separados a varios metros del punto de colisión, como se detallará en la Sección 2, para así tener trazas de mayor longitud y por consiguiente mayores valores de s . Hay que tener en cuenta que medir s con precisión puede ser un gran desafío; por ejemplo un muón con $p_T = 200$ GeV que atraviesa CMS presenta un valor de sagitta del orden del milímetro, lo que requiere contar con detectores de muones con una resolución espacial excelente. Para muones en la escala del TeV, incluso los errores sistemáticos en el posicionamiento de los detectores puede introducir un sesgo en la medida final.

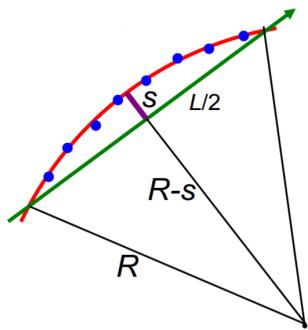


Figura 1: Definición de la distancia saggita, s , a partir de la longitud de la traza reconstruida L y de su radio R .

En segundo lugar, cuando los muones de alto momento atraviesan las componentes más densas del detector, formadas por hierro principalmente, no sólo sufren pérdida de energía por ionización sino que también sufren otros procesos como la producción de pares, Bremsstrahlung o interacciones fotonucleares, que alteran su energía haciendo más difícil el proceso de reconstrucción. En la Figura 2 se muestra la dependencia de la pérdida de energía del muón por unidad de distancia dE/dx al atravesar distintos medios como función de su energía.

Se observa que la energía crítica para el hierro, E_c^{iron} , donde la energía de ionización (en marrón) es igual a la suma de todas las pérdidas radiativas (en morado), ocurre aproximadamente a 300 GeV. Como consecuencia, la principal fuente de pérdida de energía para un muón con $E > E_c^{iron}$ que viaja por hierro a través de los distintos subdetectores de CMS es debida a radiación electromagnética fruto de la producción de electrones y fotones. Esta radiación electromagnética se manifiesta en el detector como una cascada de partículas (*muon shower*) que producen señales adicionales en los detectores, y puede incluso cambiar la dirección de la trayectoria del muón, afectando negativamente a la reconstrucción de su traza y degradando por consiguiente la medida de su momento.

Los algoritmos de reconstrucción de trazas que son utilizados para la asignación del momento transverso a los muones *high- p_T* en CMS evitan lidiar con casos de emisión de cascadas, como se explicará con más detalle en la Sección 4. De esta manera, por ejemplo, si se encuentran varios impactos en un subdetector que puedan ser indicativo de que una cascada ha tenido lugar, típicamente se ignoran estas señales a la

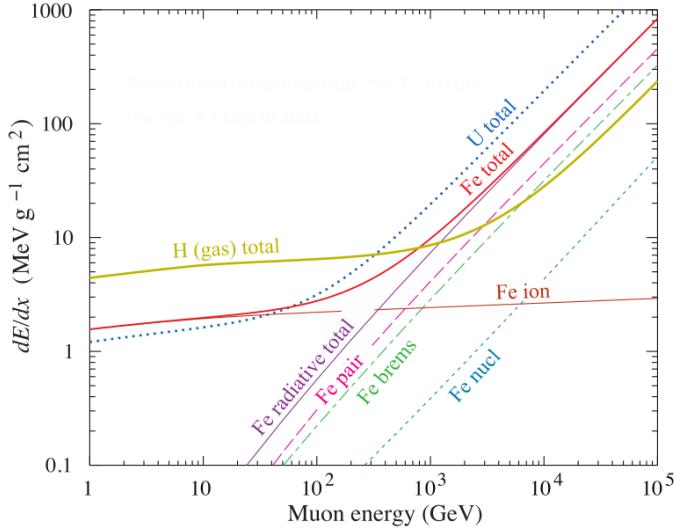


Figura 2: Media de la pérdida de energía por ionización y por radiación de un muón en hidrógeno, hierro y uranio como función de su energía. Average energy loss from ionization and radiative loss of a muon in hydrogen, iron and uranium as a function of its energy. En el caso del hierro, se separan las contribuciones a dE/dx para producción de pares, Bremsstrahlung e interacciones fotonucleares. Figura tomada de [6].

hora de reconstruir la traza del muón.

El objetivo de este trabajo, por contrapartida, es recopilar toda la información posible de la trayectoria de los muones a su paso por CMS (incluyendo los impactos provenientes de cascadas electromagnéticas), utilizando para ello muones simulados altamente energéticos con p_T generado conocido (el momento real asignado en la simulación de la partícula), y entrenar posteriormente una DNN que haga regresión al p_T reconstruido, de manera que se consiga una asignación de momento transverso mejor que el proporcionado por los algoritmos centrales de CMS.

Para cuantificar la calidad de la asignación del momento transverso, se hará uso de la resolución R , definida como:

$$R = \frac{|p_T^{GEN} - p_T^{RECO}|}{p_T^{GEN}} \quad (4)$$

Donde p_T^{GEN} hace referencia al momento generado del muón y p_T^{RECO} al momento reconstruido a través de su traza en el detector. El valor central de la distribución de R da cuenta del sesgo existente en la estimación del momento, mientras que su anchura da cuenta de la resolución del sistema.

La presente memoria está estructurada en las siguientes secciones. En la Sección 2 se describirá brevemente el dispositivo experimental utilizado: el detector CMS. Posteriormente, en la Sección 4 se presentarán los algoritmos actuales utilizados para la asignación de momento. En la Sección 5 se describirá el método propuesto, es decir,

cómo se realiza la selección de muones y hits sobre la muestra de simulación utilizada. En la Sección 6 se detallará cómo se construyen las variables de entrenamiento, la arquitectura de la red utilizada, y los resultados del entrenamiento. Por último, en las Secciones 7 y 8 se mostrarán los resultados finales y las conclusiones del trabajo.

2. El experimento CMS

El detector CMS, representado de manera esquemática en la Figura 3, se localiza en uno de los puntos del acelerador LHC donde se hacen colisionar los haces de protones. Está compuesto, de la zona más interna a la más externa, por un tracker de píxeles y tiras de silicio para la detección de partículas cargadas con gran resolución espacial, un calorímetro electromagnético de cristal de tungstato (ECAL) para la medida de electrones y fotones principalmente, un calorímetro hadrónico constituido de material denso y absorbente (HCAL) especializado en la medida de hadrones, y finalmente, en la parte más externa se encuentran las cámaras de muones. Entre el HCAL y las cámaras de muones se tiene un imán superconductor que alcanza un campo magnético de 3.8 T, suficiente para curvar partículas cargadas y permitir una buena resolución en la medida del momento de las mismas.

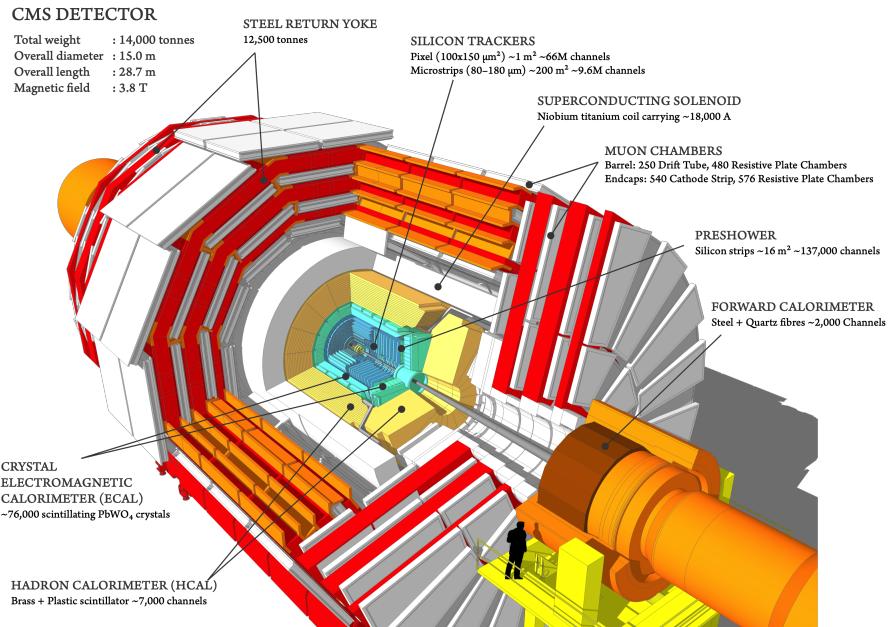


Figura 3: Representación gráfica de las distintas partes del detector CMS. Imagen tomada de [7].

En cuanto a su geometría, el sistema de coordenadas aceptado tiene como origen el punto de colisión, el eje y apunta verticalmente hacia arriba, el eje x radialmente desde el origen, y el eje z recorre la dirección del haz (ver Figura 3). El ángulo azimutal ϕ se mide a partir del eje x en el plano x-y transverso al haz, mientras que el ángulo polar θ se mide desde el eje z en el plano x-z. Otra variable angular importante que será utilizada en el análisis por ser invariante bajo transformaciones de Lorentz en el eje z, es la pseudorapidez η , que se define en función del ángulo polar como:

$$\eta = -\ln \left(\tan \frac{\theta}{2} \right) \quad (5)$$

De esta manera, se suelen usar variables definidas en el plano transversal a la dirección del haz de partículas como el momento transverso p_T o la energía transversa E_T .

En este trabajo nos centraremos en la medida de los muones, que al ser partículas cargadas dejan señal en el tracker interno, no interactúan apenas con el material denso de los calorímetros, y llegan a las cámaras de muones externas, situadas a unos cuatro metros del punto de colisión.

En las siguientes subsecciones se describirá brevemente el funcionamiento y características de los detectores de CMS que se utilizan en este trabajo: el tracker, los tubos de deriva o DTs (del inglés *Drift Tubes*), y las cámaras de tiras catódicas o CSCs (del inglés *Cathode Strip Chambers*).

2.1. Tracker

El tracker [8] se sitúa en la parte más interna de CMS y está formado por píxeles, que se sitúan en el núcleo del subdetector y reciben por tanto la mayor fluencia de partículas, y por tiras de silicio que rodean las capas de píxeles (ver Figura 4).

A medida que las partículas viajan a través del tracker, los píxeles y las tiras producen pequeñas señales eléctricas que se amplifican y detectan. De esta manera, el tracker cuenta con un total de 75 millones de canales electrónicos, dando lugar a unas 6000 conexiones por centímetro cuadrado, consiguiendo medir las trayectorias de las partículas cargadas con una precisión de unos $10 \mu\text{m}$. Cabe destacar que debido a que el tracker está compuesto por un material relativamente ligero como es el silicio, la probabilidad de que los muones produzcan un *muon shower* en esta zona del detector, es muy baja.

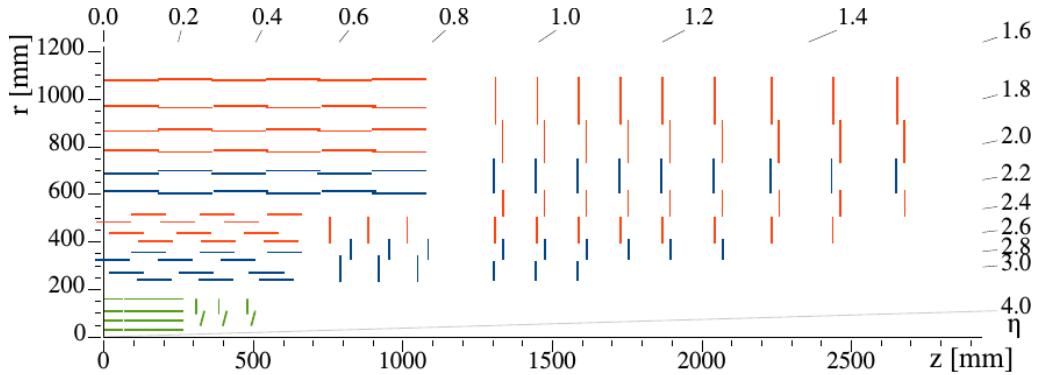


Figura 4: Cuadrante del tracker de CMS en el plano r - z . El detector de píxeles se muestra en verde, mientras que los módulos de tiras de una cara y de dos caras se muestran como segmentos rojos y azules respectivamente. Imagen tomada de [9].

2.2. Tubos de deriva: DTs

Los tubos de deriva [10] cubren el barril de CMS ($|\eta| < 0.9$) en la parte más externa del detector. Cada tubo se compone de un hilo colector de carga cargado positivamente y está lleno de gas, de forma que una partícula cargada a su paso por el tubo arranca electrones de los átomos del gas, que son atraídos eléctricamente y recolectados por el hilo conductor. De esta manera, se obtienen las coordenadas del paso del muón por el tubo a partir de la posición del hilo donde los electrones impactan y de la distancia del muón al hilo, que se calcula multiplicando la velocidad de deriva del electrón en el tubo por el tiempo de viaje hasta el hilo.

El subdetector completo consta de 250 cámaras, cada una de ellas con unas dimensiones en promedio de 2×2.5 m, estando formadas por de tres capas con unos 60 tubos repartidos en cuatro subcapas.

En la Figura 6 se representa gráficamente la posición espacial de las distintas cámaras que forman las DTs, que se denotan como MBZ/N/S, donde Z=-2...+2 se corresponde con el número de rueda a lo largo del eje z, N=1...4 hace referencia al número de estación concéntrica en el plano xy, y S=1...12 al número de sector circular.

2.3. Cámaras de tiras catódicas: CSCs

Las cámaras de tiras catódicas [11] se localizan en las tapas de CMS ($1.2 < |\eta| < 2.4$) y su funcionamiento es similar al de las DTs. En este caso se tienen cables de cargados positivamente (ánodos) cruzados con tiras de cobre cargadas negativamente (cátodos) dentro de un volumen de gas. Cuando los muones atraviesan la cámara arrancan electrones de los átomos de gas, produciéndose una avalancha de electrones que se dirigen a los cables del ánodo, mientras que los iones positivos se alejan del cable y se dirigen al cátodo de cobre, lo que también induce un pulso de carga en las tiras (en dirección perpendicular al ánodo). Debido a que las tiras y los cables son perpendiculares, se obtienen dos coordenadas de posición para cada partícula que atraviesa la cámara.

El subdetector completo de las CSCs contiene 540 cámaras en total, y está compuesto de anillos de cámaras trapezoidales de hasta 3.4 m de largo y 1.5 m de ancho, colocadas en ocho discos, cuatro en cada tapa. En este caso, las cámaras CSCs se denotan espacialmente como ME \pm S/R, donde el signo indica en qué tapa de CMS se encuentra, S=1...4 hace referencia al número de estación, paralelas en el eje z (ver Figura 6), y R se corresponde con el número de anillo, concéntricos en el plano xy. En la Figura 5 se muestra una imagen real de la rueda ME-1, con sus tres anillos concéntricos en el plano xy.

2.4. Segmentos en las cámaras de muones

Cada una de las cámaras de DTs y CSCs se compone de varias capas, y las señales o impactos que dejan los muones a su paso se reconstruyen en cada una de ellas.

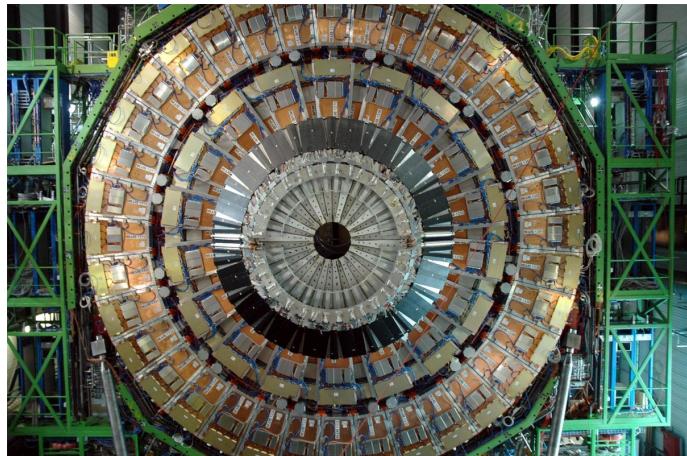


Figura 5: Foto completa de la estación ME-1. Imagen tomada de [12].

A partir de estos impactos, se construyen trazas rectas denominadas segmentos uniendo las señales encontradas en las distintas capas dentro de cada cámara DT o CSC. Estos segmentos están formados por la posición y dirección tanto en el plano transverso como en el coordenada longitudinal, salvo para las cámaras DT más exteriores, en las que la información longitudinal no está disponible. De esta manera, como se detallará en las sucesivas secciones, se utilizarán en el análisis los segmentos que los muones dejan a su paso por las distintas cámaras.

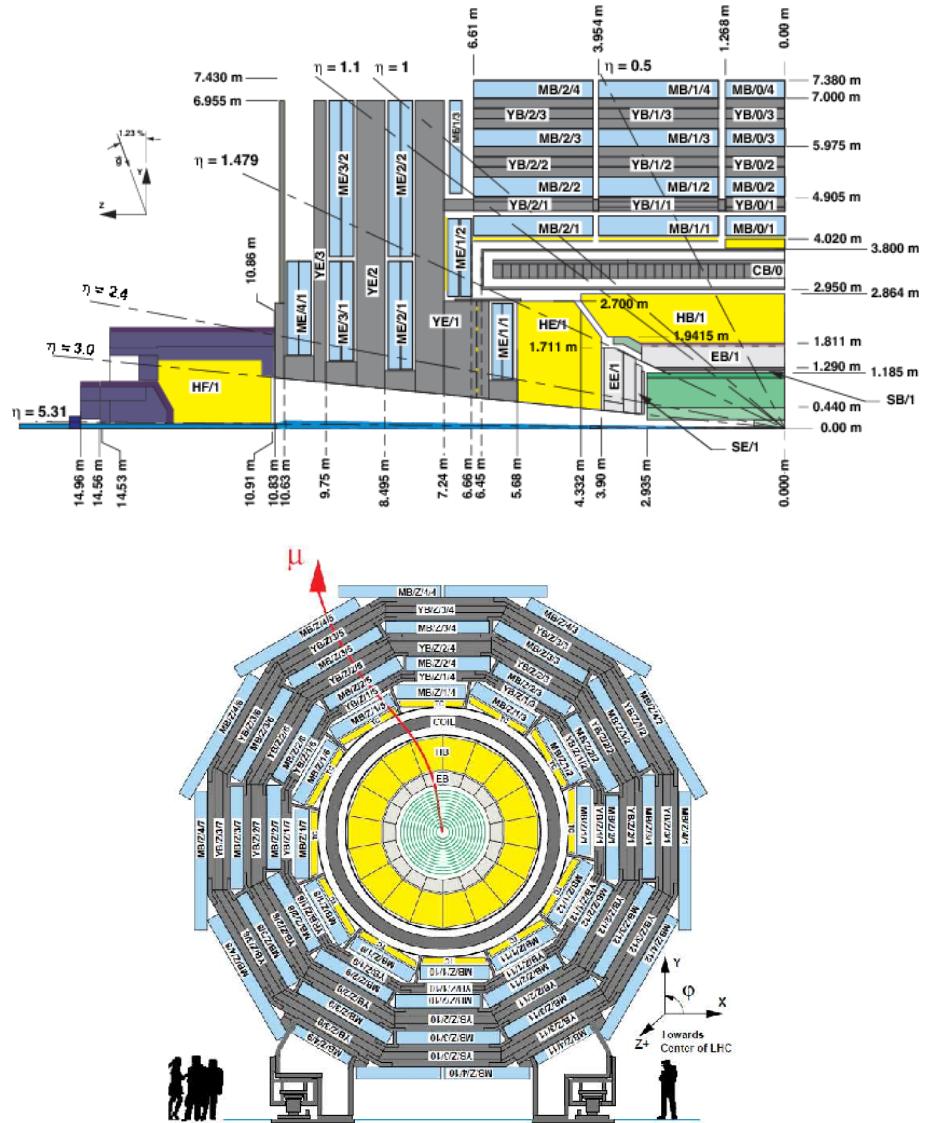


Figura 6: Vista esquemática del detector CMS. Arriba: vista longitudinal de un cuarto del detector. Abajo: vista transversal en $z = 0$. Ambas figuras han sido tomadas de [10].

3. Aprendizaje automático y su uso en física de altas energías

El aprendizaje automático es una rama de la inteligencia artificial que tiene como objetivo el desarrollo de algoritmos que permitan a la máquina aprender de la experiencia.

Este trabajo tiene como objetivo mejorar la medida del momento transverso de muones de *high-p_T*. Puesto que CMS dispone de simulaciones precisas sobre la producción, propagación y proceso de medida de dichos muones, en donde el momento real de la partícula es conocido, este trabajo se centrará en un método supervisado. De esta forma, al modelo matemático se le proporcionan, en la fase de aprendizaje o entrenamiento, tanto las características del muón (datos de entrada), como el momento real asignado en la simulación de la partícula, con el fin de que en esta etapa el modelo pueda encontrar correlaciones y patrones de comportamiento en los datos de entrada que mejoren la estimación del *p_T* del muón. Posteriormente, en el proceso de testeo el modelo ha de ser capaz de poder dar una predicción del *p_T* tomando como datos de entrada de muones con momento transverso desconocido.

Dentro de los distintos tipos de algoritmos de aprendizaje automático destacan las redes neuronales artificiales, cuyo funcionamiento se inspira en las redes neuronales biológicas que forman el cerebro, y se componen de unidades conectadas denominadas neuronas que reciben una cierta información, la procesan, y la transmiten a las siguientes neuronas con las que están conectadas. La estructura esquemática de una red neuronal se muestra en la Figura 7, y está formada esencialmente por capas de neuronas conectadas entre sí, donde cada conexión lleva asociado un peso que da cuenta de la importancia que tiene la información que va de una neurona a otra.

El proceso de transmisión de información va de izquierda a derecha: en primer lugar se inicializan los pesos de la red (habitualmente de manera aleatoria) y se tiene una primera capa de neuronas que se corresponden con las distintas variables de entrada. Si todas las neuronas de una capa están conectadas con las neuronas de la capa siguiente, a cada neurona de la segunda capa le llegará como entrada un sumatorio de la información de cada neurona de la capa anterior multiplicada por su peso asociado, de manera que a esta entrada se le aplica una función de activación no lineal y se sigue propagando la información a las siguientes capas. Una vez que la información llega a la capa final de la red, se evalúa el error en la predicción respecto al valor real conocido de la magnitud que se quiere predecir.

Por otra parte, el proceso de aprendizaje a partir de la experiencia se hace en la dirección contraria: una vez obtenido el error en la predicción, los pesos se van actualizando capa a capa hacia atrás acorde a la dirección del gradiente de la función de error, de forma que se busca que en la siguiente iteración el error en la predicción sea menor.

La red neuronal profunda o DNN se distingue de la red neuronal convencional por

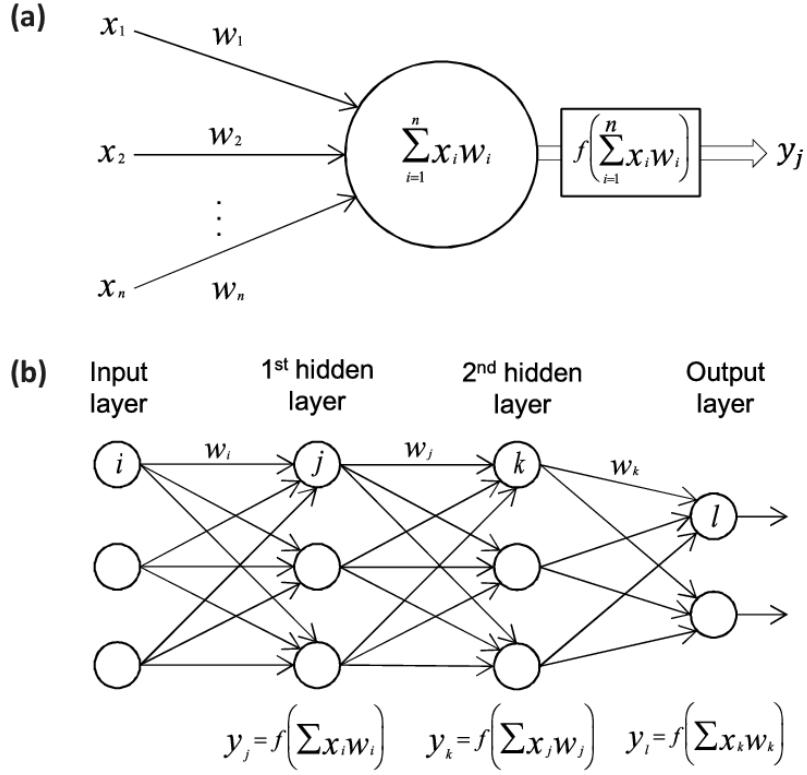


Figura 7: Representación gráfica del funcionamiento de una red neuronal artificial. a: procesamiento de una neurona, donde cada entrada x_i lleva asociado un peso w_i , y la suma de todas las entradas pesadas que llega a la neurona se pasa a la siguiente neurona tras aplicar una función de activación no lineal. b: ejemplo de una red neuronal multicapa donde las neuronas de cada capa están conectadas a todas las neuronas de la capa siguiente. La información es propagada desde la primera capa hasta la capa final, donde se evalúa el error en la predicción. Imagen tomada de [13].

tener más capas ocultas y un mayor número de neuronas en cada capa, aumentando así la cantidad de parámetros, nivel de abstracción y complejidad, y permitiendo obtener un mayor rendimiento en la predicción para conjuntos de datos de entrada de gran dimensionalidad.

En física de altas energías, el uso de las redes neuronales profundas se ha extendido en los últimos años, siendo una de las herramientas más utilizadas en la reconstrucción de objetos físicos y en la clasificación de eventos en búsquedas de nuevas partículas (ver [14]).

4. Asignación de momento transverso en el experimento CMS

El momento de un muón *low- p_T* es extraído a partir de la reconstrucción de su trayectoria a través de las señales que deja en el tracker junto con los impactos en el sistema de muones mediante un ajuste por filtro de Kalman [15], que se trata de una técnica que permite obtener los parámetros asociados a una trayectoria a partir de una serie de puntos pertenecientes a dicha trayectoria y que son proporcionados de forma secuencial, produciéndose de forma dinámica una nueva estimación de los parámetros a medida que se añade más información.

En el caso de muones *high- p_T* , la probabilidad de emisión de partículas adicionales generadas en las cascadas electromagnéticas es mayor, y estas dejan típicamente señales adicionales (múltiples segmentos) en las distintas cámaras, provocando que estas señales adicionales sean probablemente utilizadas en el ajuste por filtro de Kalman en lugar de las señales que realmente provienen del muón que se quiere medir, o equivalentemente, un número elevado de segmentos en una cámara en particular puede dificultar seriamente la propia reconstrucción. Es por esto que los muones de alto momento requieren un tratamiento más cuidadoso de la información encontrada en el sistema de muones mediante algoritmos especiales denominados reajustes, que seleccionan qué impactos van a utilizarse en el proceso de reconstrucción de la traza y cuáles no.

En las siguientes subsecciones se da una descripción general de los distintos reajustes para muones *high- p_T* que actualmente se llevan a cabo en CMS, así como del algoritmo utilizado para la asignación final del momento transverso.

4.1. Reajuste TPFMS (*Tracker-plus-first-muon-station*)

El primer reajuste, que es el más sencillo en cuanto a su implementación, es el TPFMS, que selecciona únicamente las señales en el tracker y en la estación más interna del sistema de muones que contiene señales compatibles con las del tracker. De esta manera se pretende desechar aquellos impactos de estaciones más lejanas y eliminar así la posible contaminación debida a las cascadas electromagnéticas.

4.2. Reajuste Picky

El reajuste Picky tiene como objetivo encontrar posibles cascadas y eliminar aquellas señales adicionales que no pertenecen al muón original del ajuste de la traza. De esta manera, si más de n impactos se encuentran dentro de un cono en torno a un impacto concreto, la estación en la que se encuentran dichas señales se marca como contaminada. Así, a la hora de hacer el ajuste a la traza, si su χ^2 está por encima de un cierto umbral, se eliminan del mismo aquellos impactos que se encuentren en cámaras contaminadas y se repite el ajuste de nuevo.

Los parámetros del algoritmo n y χ están optimizados en base a estudios sobre simulaciones.

4.3. Reajuste DYT (*Dynamic truncation*)

Cuando un muón pierde una gran fracción de su energía durante su trayecto, su dirección puede cambiar y las señales encontradas en las siguientes estaciones pueden ser inconsistentes con la trayectoria inicial (ver Figura 8), provocando problemas a la hora de reconstruir la traza y consecuentemente al tratar de asignar un valor acertado a su momento transverso. En estos casos, el reajuste DYT considera más conveniente parar el filtro de Kalman una vez que un cambio en la trayectoria del muón es detectado.

En el reajuste DYT se define un operador E que da idea de la compatibilidad de un determinado segmento en una cámara con la extrolación de la traza interna a dicha cámara, de manera que si este operador supera un valor predeterminado, el filtro de Kalman se detiene y no tiene en cuenta aquellos impactos en las cámaras posteriores al cambio de trayectoria detectado.

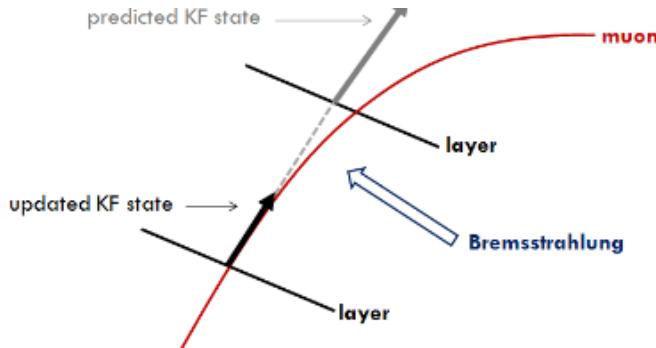


Figura 8: Representación gráfica del cambio en la trayectoria del muón tras una gran pérdida de energía por Bremsstrahlung.

4.4. Asignación final de momento transverso: el algoritmo TuneP

Finalmente, la recomendación central de la colaboración CMS para la asignación del momento transverso de los muones *high- p_T* se corresponde con el proporcionado por el algoritmo TuneP.

El objetivo de este algoritmo es simplemente elegir cuál es la mejor reconstrucción de la traza posible de entre la traza reconstruida únicamente en el tracker y las trazas obtenidas por los distintos reajustes (TPFMS, Picky y DYT). Dicha elección se hace teniendo en cuenta conjuntamente el χ^2/ndof y el σ_{p_T}/p_T de las trazas consideradas.

Con todas estas consideraciones, CMS consigue medir el momento transverso de los muones con $p_T \leq 1$ TeV con un error menor al 10 %.

5. Método propuesto

Este trabajo persigue implementar un algoritmo regresivo usando una DNN para predecir el momento transverso de muones de alto momento, intentando mejorar los algoritmos actuales de CMS.

Puesto que parte de los problemas de reconstrucción de este tipo de muones tiene que ver con la alta probabilidad de producción de cascadas electromagnéticas y por consiguiente la alta multiplicidad de segmentos en el sistema de muones, la propuesta tiene como elemento clave la distribución de segmentos en las cámaras de muones. En particular, cada elemento del conjunto de datos estará compuesto por la información de la traza del muón reconstruida únicamente en el tracker, Aquí entiendo que hay que anadir tambien el tuneP, así como sus extrapolaciones al sistema de muones, junto con información acerca del número y distribución de los segmentos en torno a dichas extrapolaciones.

En la presente sección se detallarán las herramientas y metodología utilizadas para el tratamiento de los datos.

5.1. Herramientas utilizadas para el análisis

Las herramientas utilizadas en este trabajo cumplen diversas funciones, pasando por la generación y simulación del paso de los muones por el detector, la reconstrucción y obtención de las variables de interés y su posterior análisis. En particular, pueden diferenciarse las siguientes herramientas:

- Pythia [16] es un generador de Monte-Carlo que tiene como objetivo proporcionar todos los elementos necesarios para la simulación de eventos fruto de la colisión de partículas en condiciones de alta energía como las del LHC. Concretamente, se utilizará Pythia para generar una muestra representativa de muones de alto momento de la que emergerá la muestra de entrenamiento y test. Las macroinstrucciones usadas para la producción de la muestra se pueden encontrar en el repositorio [17].
- CMSSW [18] es una colección de software abierto de la colaboración CMS basado en C++ y Pythia utilizado principalmente para la simulación, calibración, alineamiento, y reconstrucción de objetos físicos, como los muones, a partir de sus segmentos y trazas. CMSSW a su vez utiliza Geant4 [19] que es el programa encargado de simular el transporte de las partículas por el detector. Para este trabajo se ha creado un módulo en C++ que puede encontrarse en [20]. Este código utiliza diferentes herramientas de CMSSW para realizar la selección apropiada de muones y segmentos que se describirá en la subsección 5.3.
- ROOT [21] es el marco de trabajo más utilizado en física de altas energías para el procesado de datos, para análisis estadístico, y proporciona además herramientas de visualización y almacenamiento de los mismos. ROOT está

basado en programación orientada a objetos y escrito en C++, y se usará como formato de los datos de entrada y salida del código de selección [20], almacenando la información recogida como objetos de ROOT.

- Pandas [22] (del inglés *Python Data Analysis Library*), es una librería de Python que ofrece gran rendimiento en el manejo de estructuras tabulares de datos. En este caso, se usará Pandas, previa transformación de los datos en formato ROOT a formato de texto, para el procesado de los segmentos seleccionados (operaciones de limpiado de los datos, agregaciones, construcción de las variables de entrenamiento...etc). Los módulos creados para el procesado pueden encontrarse en [23].
- TensorFlow [24] es una librería de código abierto desarrollada por Google que contiene la estructura matemática de distintos algoritmos de aprendizaje automático como las redes neuronales artificiales. Se utilizará TensorFlow para entrenar el modelo de regresión al momento transverso de los muones.

La secuencia de trabajo en cuanto a las herramientas utilizadas se muestra esquemáticamente en la Figura 9.

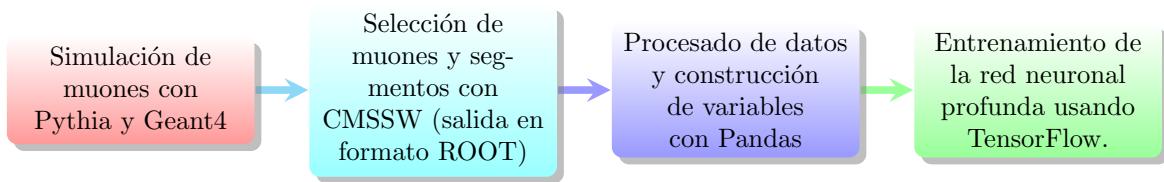


Figura 9: Uso secuencial de las distintas herramientas utilizadas en el trabajo.

5.2. Muestra de simulación utilizada

Se han generado un total de 965000 sucesos de colisiones protón-protón a una energía de centro de masas de 13 TeV¹ mediante simulación de Monte-Carlo utilizando el programa Pythia [17]. Cada suceso contiene un único muon. Se impone que los muones generados tengan un momento transverso aleatoriamente distribuido en el rango entre 20 y 2500 GeV y espacialmente repartidos en el rango $-2.5 < \eta < 2.5$, y se simula el paso de dichos muones por el detector CMS con el paquete Geant4.

La simulación de los procesos de colisión es computacionalmente costosa y en general requiere del uso de clusters de ordenadores. El tamaño de la muestra obtenida se ha considerado razonable teniendo en cuenta el tiempo de computación y el espacio de almacenamiento necesario. En la Figura 10 se muestra la distribución bidimensional del momento transverso dado por el algoritmo TuneP en función del momento transverso de generación para todos los muones de la muestra que pasan la selección detallada en la subsección 5.3.

¹condiciones actuales de funcionamiento del acelerador LHC

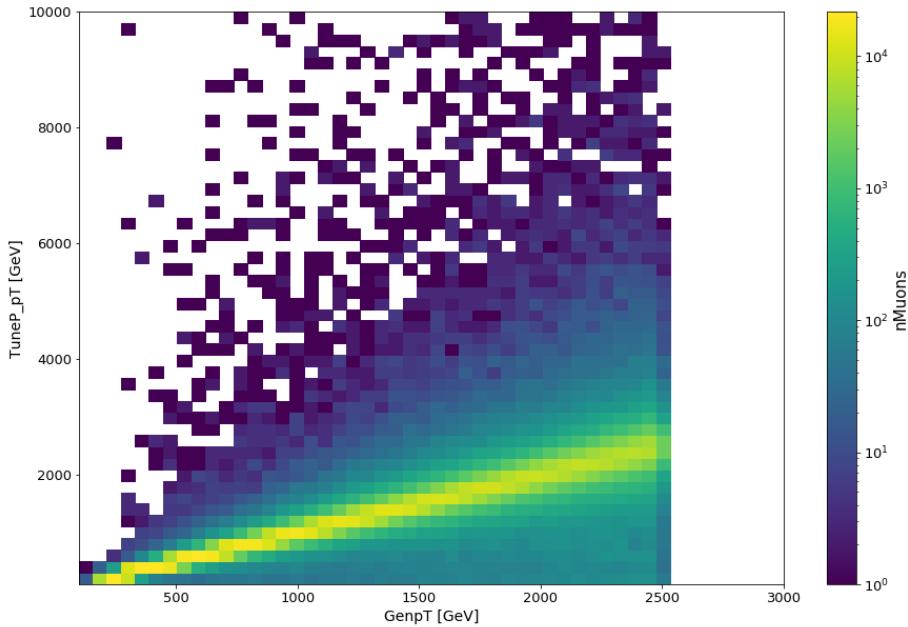


Figura 10: Distribución bidimensional del p_T proporcionado por el algoritmo TuneP en función del p_T de generación para los muones de la muestra.

5.3. Preparación del conjunto de datos: selección de muones y segmentos

El proceso de selección de los muones y segmentos así como de la producción de las extrapolaciones (ver código [20]) a partir de la muestra de simulación consta de las siguientes partes:

1. Lectura de datos: en los datos de entrada, los muones y segmentos reconstruidos por evento se almacenan en colecciones de ROOT, que funcionan a modo de contenedor de información. Por tanto, el primer paso es leer las colecciones para poder iterar sobre ellas.
2. Se recorre la colección de muones y se seleccionan aquellos que tengan una traza reconstruida en el tracker con $p_T > 200$ GeV, y que esta traza reconstruida esté próxima a la traza real generada del muón dentro de un cono de radio $\Delta R = \sqrt{(\Delta\eta^2 + \Delta\phi^2)} < 0.3$. De esta forma se pretende garantizar que el muon generado coincide exactamente con la traza considerada.
3. Posteriormente, se recorren las colecciones de segmentos encontradas en el suceso, y se almacenan sus coordenadas y los identificadores de los detectores en los que se encuentran, si los segmentos son válidos², y sus coordenadas se han medido correctamente.

²Que son compatibles con la trayectoria inicial de la traza, y con un χ^2 del ajuste de las señales que componen el segmento no muy elevado.

4. Se selecciona la posición más externa de la traza del tracker y se extrapola a la superficie de cada una de las cámaras de muones registradas donde se ha encontrado al menos un segmento. El proceso de extrapolación se lleva a cabo utilizando un software especializado que integra la trayectoria de los muones aplicando la ley de Lorentz y teniendo en cuenta sus interacciones con el material de los detectores, como la ionización o la dispersión múltiple Coulombiana.
5. Se seleccionan aquellas extrapolaciones que son válidas (compatibles con la dirección de la trayectoria inicial), vayan en la dirección del campo magnético impuesto por el solenoide de CMS, y cumplan que la distancia entre el centro de la cámara y la propia extrapolación no exceda el tamaño propio de la cámara. Este último requerimiento es de vital importancia, ya que las superficies a las que se extrae son planos de dimensión infinita (sin delimitar por las dimensiones reales de las cámaras), y una partícula cargada en movimiento sometida a un campo magnético siempre puede cortar un plano de dimensión infinita al curvarse.
6. Se guardan todas las variables de interés como objetos de ROOT, es decir, la información sobre la traza del muón en el tracker, las coordenadas y direcciones de los segmentos recogidos, y las coordenadas de las extrapolaciones.

En las Figuras 11 y 12 se muestran las posiciones espaciales de los segmentos y la propagaciones en los planos xy y xz una vez realizado el limpiado.

Para la construcción de las variables que serán utilizadas en el entrenamiento se usan agregaciones de Pandas (ver doStep1.py en [23]). Así, para cada muón se agregan los segmentos encontrados en cada estación de DTs y CSCs y se obtiene el número total de segmentos por estación, la media espacial de los segmentos, la desviación estándar de los mismos, la asimetría y la kurtosis.

5.4. Distribuciones de control

Tras la obtención de la muestra de entrenamiento y de test, se han realizado varias comprobaciones para asegurar su calidad y coherencia.

En primer lugar, se seleccionan muones con cuatro o más segmentos en las DTs o CSCs y exigiendo que tengan al menos un segmento en cada estación. Esta submuestra contiene trazas de muones que pueden considerarse completas. Posteriormente, se separan los muones que cumplen estas condiciones en dos categorías: aquellos que tienen exactamente cuatro segmentos (uno en cada estación), y aquellos que tienen más de cuatro segmentos. Los muones pertenecientes a la segunda categoría son susceptibles de haber producido una cascada electromagnética. De esta manera, si para cada muón se selecciona el máximo valor de distancia entre el segmento y la propagación, uno esperaría encontrar valores relativamente bajos en la primera categoría, y valores más altos en la segunda, debido a la presencia de segmentos adicionales. Las distribuciones para ambos grupos se muestran en la Figura 13.

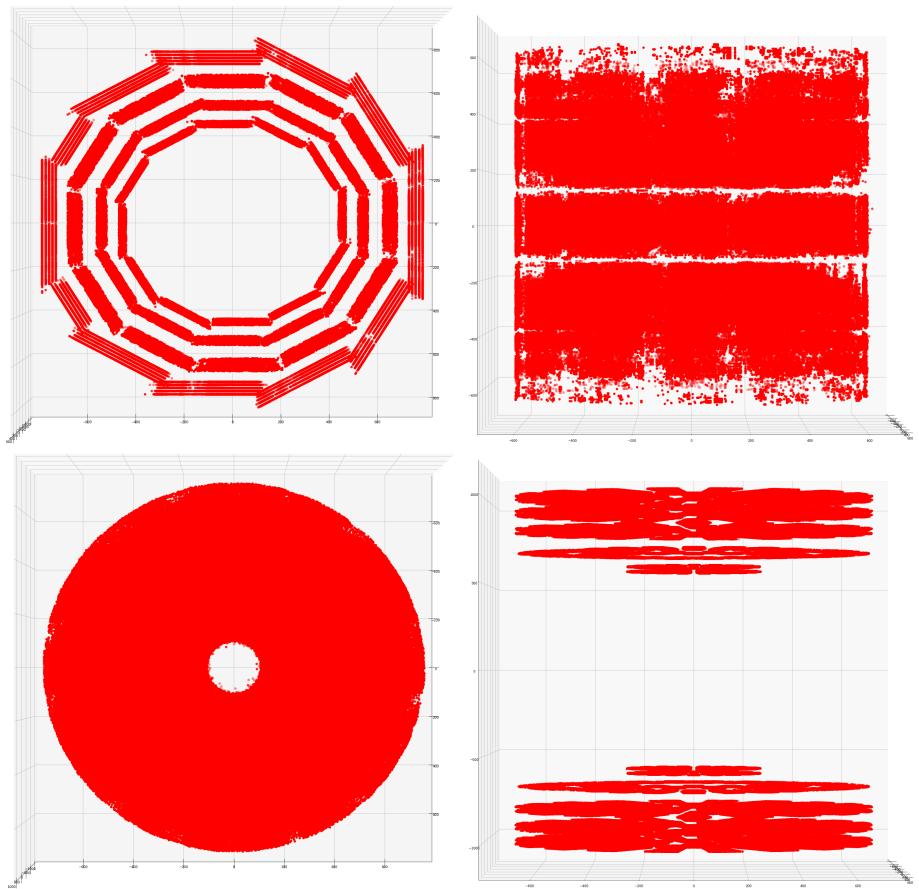


Figura 11: Posiciones geométricas de todos los segmentos seleccionados. Arriba izquierdo: Segmentos en DTs en el plano xy. Arriba derecha: Segmentos en DTs en el plano xz. Abajo izquierdo: Segmentos en CSCs en el plano xy. Abajo derecho: Segmentos en CSCs en el plano xz.

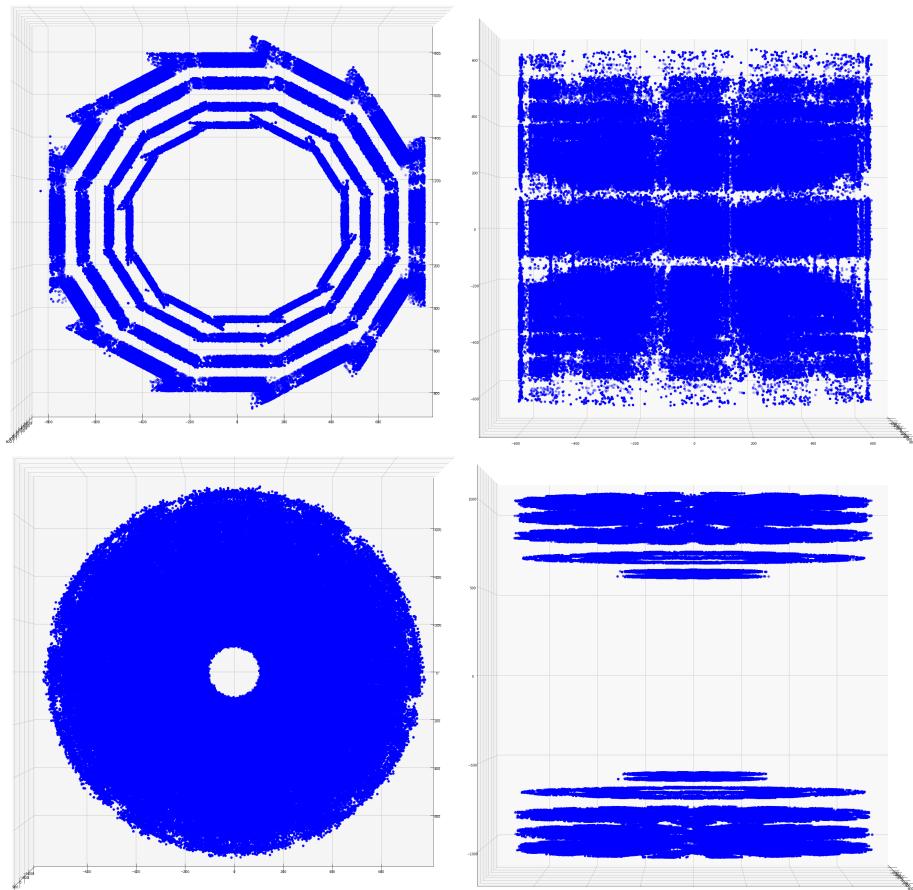


Figura 12: Posiciones geométricas de todos los propagaciones seleccionados. Arriba izquierda: Propagaciones en DTs en el plano xy. Arriba derecha: Propagaciones en DTs en el plano xz. Abajo izquierda: Propagaciones en CSCs en el plano xy. Abajo derecha: Propagaciones en CSCs en el plano xz.

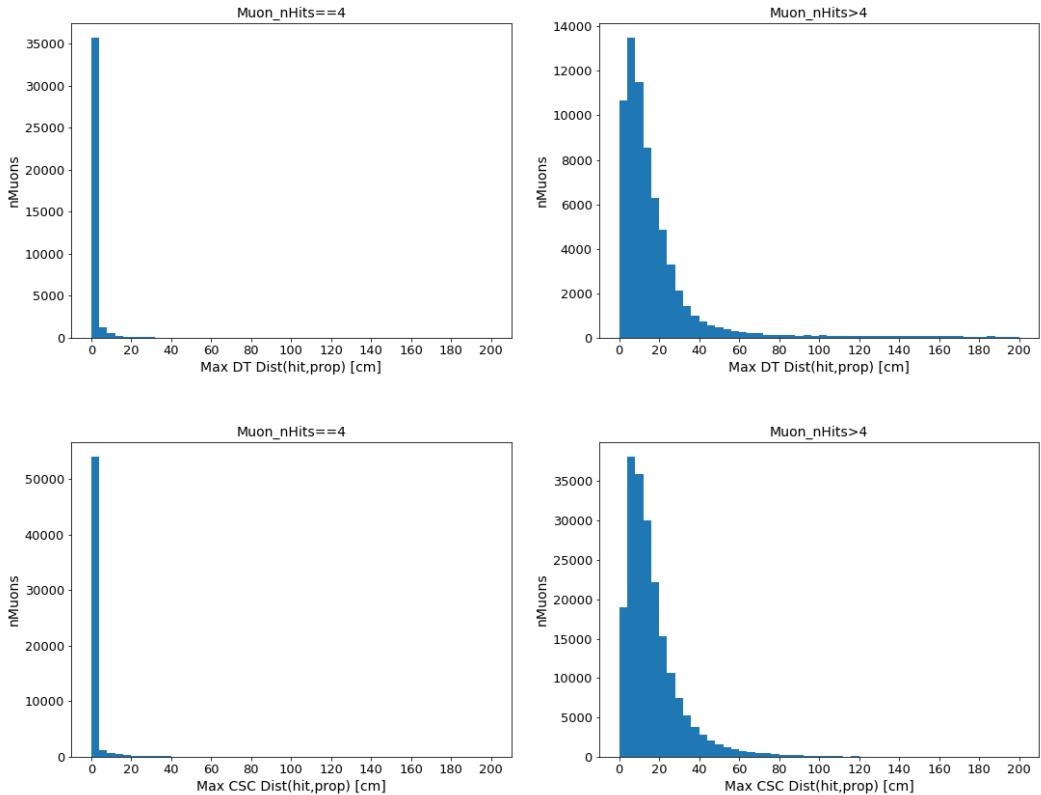


Figura 13: Distribuciones del máximo valor encontrado de distancia entre el segmento y la extrapolación (por muón). Arriba izquierda: Muones con cuatro segmentos en las DTs, uno en cada estación. Arriba derecha: Muones con más de cuatro segmentos en las DTs, y con al menos un segmento por estación. Abajo izquierda: Muones con cuatro segmentos en las CSCs, uno en cada estación. Abajo derecha: Muones con más de cuatro segmentos en las CSCs, y con al menos un segmento por estación.

Por otra parte, en la Figura 14 se muestra la dependencia de la media del número de segmentos producidos por cada muon con el p_T de generación para todos los muones de la muestra. Se observa en este caso que hay una clara tendencia ascendente del promedio del número de segmentos con el p_T , ya que como se indicó en la sección 1, a mayor p_T mayor es la probabilidad de que el muón emita una cascada y se encuentren por tanto más señales en las cámaras de muones. Comentamos aquí que las CSCS son mas problemáticas para así justificar un poco que para el trabajo nos centramos en las DT?.

Equivalentemente, en la Figura 15 se muestra la dependencia del promedio de la resolución R con el p_T de generación, donde puede observarse que la resolución en la asignación de momento transverso se va degradando con el p_T según lo esperado.

Por último, en la Figura 16 se muestra la dependencia del promedio del número de cascadas encontradas por muón, definidas por muón como el número de veces que

se tienen más de 20 señales³ en alguna de las estaciones de DTs o CSCs, con el p_T de generación. De nuevo se observa la tendencia ascendente acorde con lo esperado.

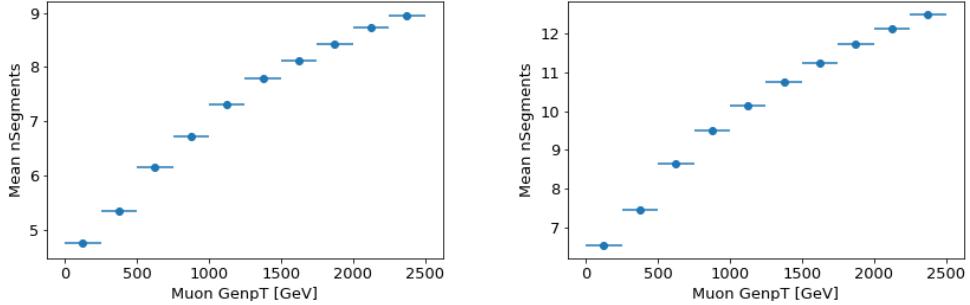


Figura 14: Valor medio de segmentos encontrados por muón en función del momento transverso generado, donde la incertidumbre en el eje de abscisas se corresponde con la desviación estándar de la distribución del número de segmentos por bin dividida entre la raíz cuadrada de número de muones en cada bin. Izquierda: Segmentos en las DTs. Derecha: Segmentos en las CSCs.

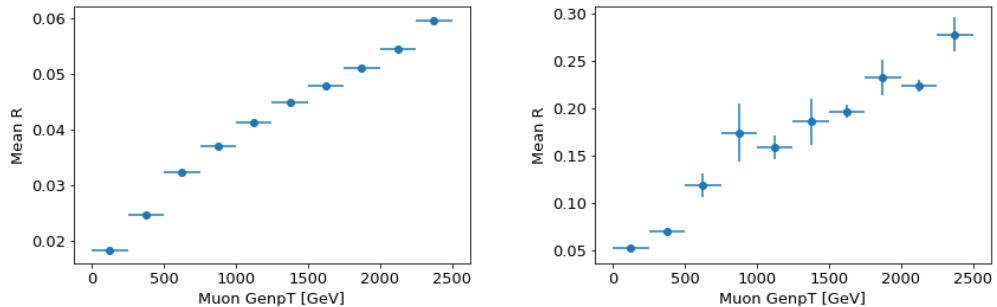


Figura 15: Valor medio de R en función del momento transverso generado, donde la incertidumbre en el eje de abscisas se corresponde con la desviación estándar de la distribución de R por bin dividida entre la raíz cuadrada de número de muones en cada bin. Izquierda: Segmentos en las DTs. Derecha: Segmentos en las CSCs.

³Número arbitrario de señales que puede dar un indicio de que una cascada ha tenido lugar.

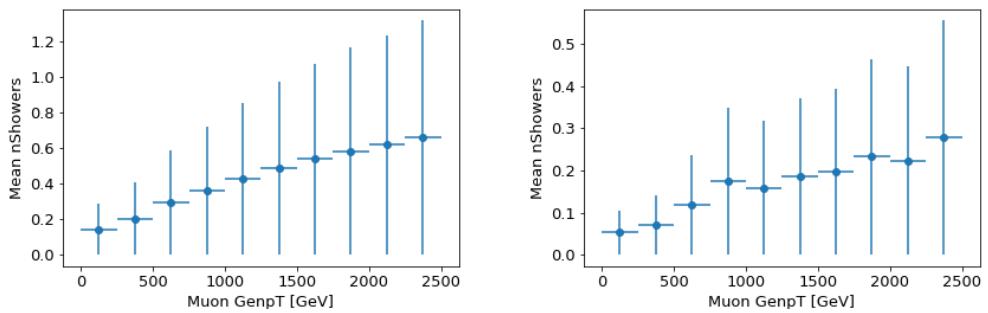


Figura 16: Valor medio del número de cascadas en función del momento transverso generado, donde la incertidumbre en el eje de abscisas se corresponde con la desviación estándar de la distribución del número de cascadas por bin dividida entre la raíz cuadrada de número de muones en cada bin. Izquierda: Segmentos en las DTs. Derecha: Segmentos en las CSCs.

6. Implementación y entrenamiento de la DNN

Una vez obtenida toda la información posible del muón a su paso por CMS, el objetivo es entrenar una red neuronal profunda tomando como entrada las propiedades de la traza en el tracker y la información espacial de las señales recogidas en el sistema de muones.

En este caso, la función de pérdida a minimizar será una función que dependa de la diferencia entre el p_T de generación conocido que se quiere predecir y el valor de p_T que devuelve la red, para así aprender las características de los muones (especialmente de aquellos que emitan cascadas) y hacer regresión a su momento transverso.

En cuanto al software, para el entrenamiento se utiliza la librería de Python de código abierto Keras [25], que se caracteriza principalmente por ofrecer sencillez de uso para el usuario, y la red se ejecuta sobre TensorFlow.

6.1. Conjunto de datos y arquitectura de la red

Para el entrenamiento se han seleccionado los muones de la muestra de simulación con $p_T > 200$ GeV y con al menos 20 señales encontradas en alguna de las estaciones, de manera que se tenga un indicio de que una cascada electromagnética ha tenido lugar. Además, por simplicidad, el entrenamiento presentado en este trabajo se llevará a cabo únicamente con muones detectados en las DTs ($|\eta| < 0.9$).

El conjunto de datos resultante, compuesto por un total de 120225 muones, se ha dividido en un 80 % para el entrenamiento del modelo (del que un 10 % será usado como conjunto de datos de validación), mientras que un 20 % es utilizado para el testeo del mismo.

En cuanto al tipo de red utilizada, se ha entrenado una red de tipo *fully-connected*, donde todas las neuronas de cada capa están conectadas con las neuronas de las capas contiguas, con la siguiente arquitectura:

- Capa de entrada con 53 neuronas (variables de entrenamiento), un total de 9 capas ocultas con 512, 512, 256, 256, 128, 128, 64, 64 y 16 neuronas respectivamente, y una neurona en la capa de salida.
- Activación: Para las capas 1-10 se utiliza la función de activación ReLu [26], ya que es una de las funciones de activación más apropiadas para problemas de aprendizaje profundo por tener siempre derivada igual a la unidad y así evitar el problema del *vanishing gradient* [27]. En la última capa 11 se utiliza una función de activación lineal para hacer regresión al momento transverso.
- Función de pérdida: MSE (del inglés *Mean Squared Error*), que se define como el promedio de los errores al cuadrado de la siguiente manera:

$$MSE = \frac{1}{n} \sum_{i=1}^n \left(p_T^{GEN}_i - p_T^{PRED}_i \right)^2 \quad (6)$$

Donde n se corresponde con el número de muones de la muestra, $p_T^{GEN}_i$ hace referencia al momento transverso generado y conocido del muón i , y $p_T^{PRED}_i$ al momento transverso que predice la red para el mismo.

- Descenso del gradiente: para agilizar el aprendizaje se hace uso de la técnica del *mini-batch gradient descent* [28], que consiste en dividir el conjunto de datos de entrenamiento en fragmentos pequeños denominados *mini-batches*, de tal forma que los parámetros de la red se van actualizando para cada fragmento sin tener que recorrer toda la muestra, mejorando considerablemente la velocidad en el entrenamiento y consiguiendo una convergencia al mínimo de la función de pérdida razonablemente buena. El tamaño del mini-batch elegido ha sido de 2000 muones.
- Como optimizador en la búsqueda del mínimo de la función de pérdida se usa Adam (del inglés *Adaptive Moment estimation*), que combina el descenso del gradiente con momento [29] y el descenso del gradiente con RMSprop [30], con tasa de aprendizaje = 0.0005 y el resto de hiperparámetros los recomendados por el artículo original de Adam [31].
- Épocas de entrenamiento: 1000.
- Para regularizar se utiliza la técnica de *Early stopping* en la pérdida del conjunto de datos de validación, con una paciencia de 50 épocas.
De esta manera, para cada época se obtiene el valor del MSE en el conjunto de datos de validación (que no se utilizan para entrenar), y si el MSE no mejora después de 50 épocas se guarda el modelo encontrado con menor MSE en validación y se detiene abruptamente el entrenamiento para evitar así que el modelo sobreentrene.

El valor de los hiperparámetros se ha tomado tras probar distintas configuraciones aleatorias de los mismos (búsqueda de tipo cuadrícula), y se ha elegido aquella que proporciona un MSE menor en el conjunto de datos de testeo.

6.2. Variables de entrenamiento

Las variables utilizadas para el entrenamiento se dividen en tres categorías según su naturaleza, dando lugar a un total de 53 variables:

- Características de la traza interna reconstruida en el tracker: p_T , η , ϕ , carga.
- Momento transverso dado por la traza seleccionada por el algoritmo TuneP (ver Sección 4.4).
- Información del conjunto de segmentos recogidos en cada estación de las DTs: número total de segmentos en la estación, media espacial en x, y, z, desviación estándar en x, y, z, asimetría en x, y, z, kurtosis en x, y, z.

En la Figura 17 se muestran las distribuciones de las variables utilizadas en el entrenamiento provenientes de la traza del tracker y de las dos primeras estaciones de las DTs.

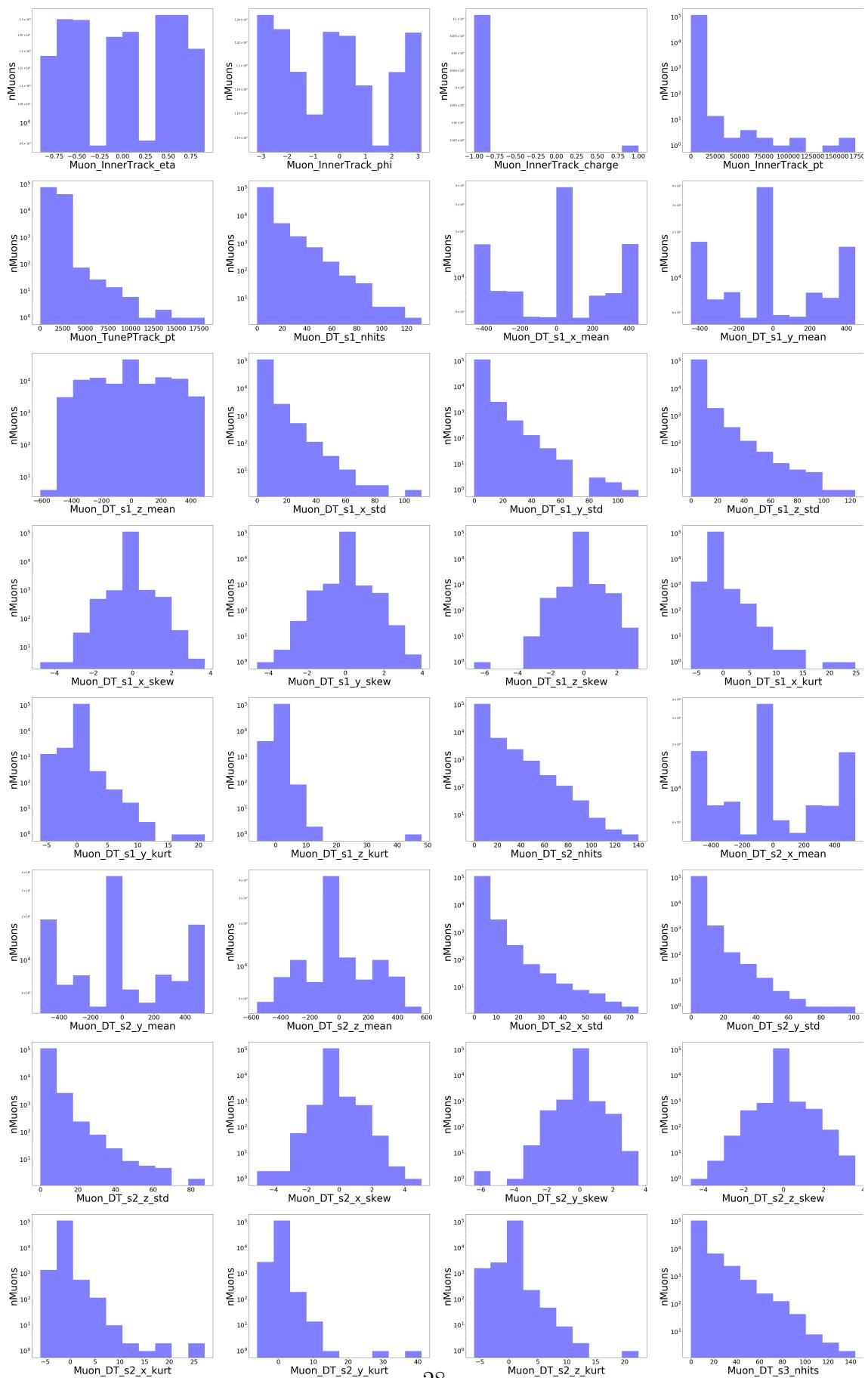


Figura 17: Distribuciones de las variables de entrenamiento provenientes de la traza del tracker y de las dos primeras estaciones de las DTs.

7. Resultados

En la Figura 18 se muestra la el valor de la pérdida MSE en función de la época de entrenamiento.

El mejor modelo de obtiene después de 175 épocas, con un valor de pérdida para el conjunto de datos de entrenamiento de 11276.5, para el conjunto de datos de validación de 11446.9, y para el conjunto de datos de testeo de 12357.7.

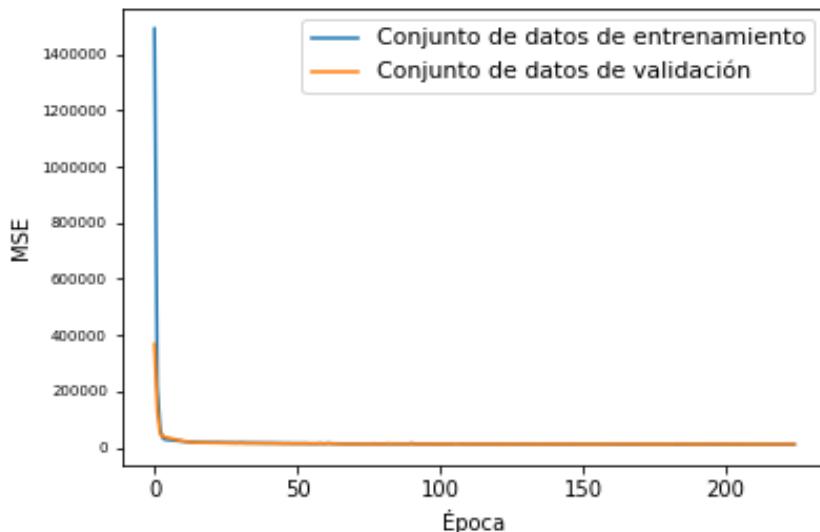


Figura 18: Valor del MSE en función de la época de entrenamiento para el conjunto de datos de entrenamiento (en azul) y para el conjunto de datos de validación (en naranja).

En la Figura 19 se muestra la distribución bidimensional del momento transverso de predicho por la red en función del momento transverso de generación para los muones del conjunto de datos de testeo, directamente comparable con la Figura 20, donde se muestra para el mismo conjunto de datos el momento transverso dado por el algoritmo TuneP en función del momento transverso de generación.

A la hora de cuantificar los resultados obtenidos ha de tenerse en cuenta que la red ha sido alimentada con muones generados con un valor máximo para el p_T de 2500 GeV, por tanto es natural pensar que el modelo va a tender a asignar un p_T igual o menor a este valor máximo a los muones cuyo p_T sea cercano a 2500 GeV. Por este motivo, y para tener resultados más coherentes, la evaluación del método se hará con muones con momento transverso generado en el rango $1800 \leq p_T \leq 2300$ GeV.

En la Figura 21 se comparan la distribuciones de la resolución R, definida en (4), calculadas sobre el conjunto de datos de testeo con $1800 \leq p_T^{GEN} \leq 2300$ GeV, a partir del p_T que da el algoritmo TuneP y a partir del p_T que predice la red neuronal

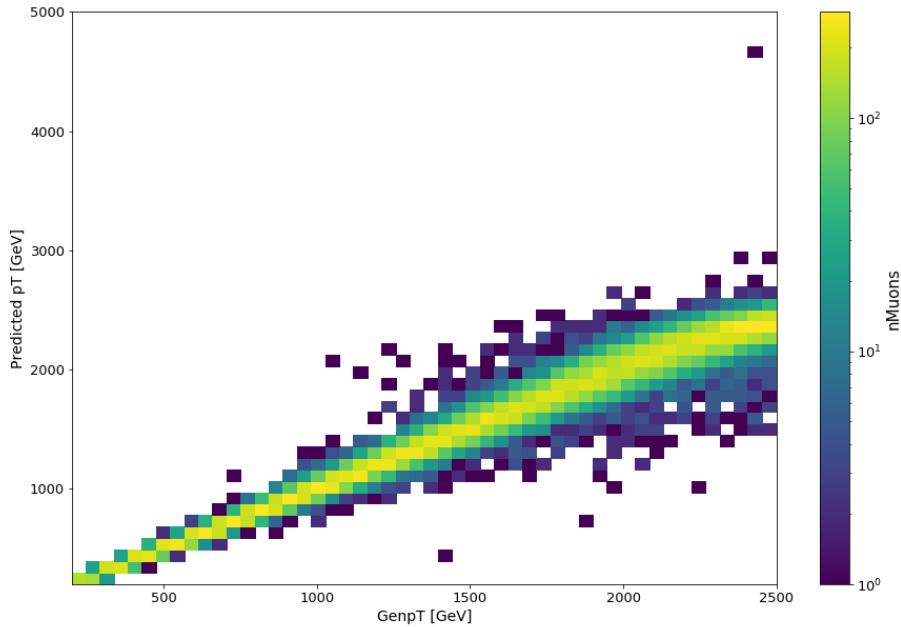


Figura 19: Distribución bidimensional del p_T predicho por la red neuronal en función del p_T de generación para los muones del conjunto de datos de testeo.

respectivamente.

Se obtiene un valor promedio para la resolución promedio en el conjunto de muones de testeo de $R = 0.0575 \pm 0.0012$ usando el p_T del algoritmo TuneP, y de $R = 0.04845 \pm 0.0006$ usando el p_T predicho por el modelo de regresión. Por lo tanto, se consigue una mejora en la asignación del momento transverso de aproximadamente un 20 % para estos muones altamente energéticos.

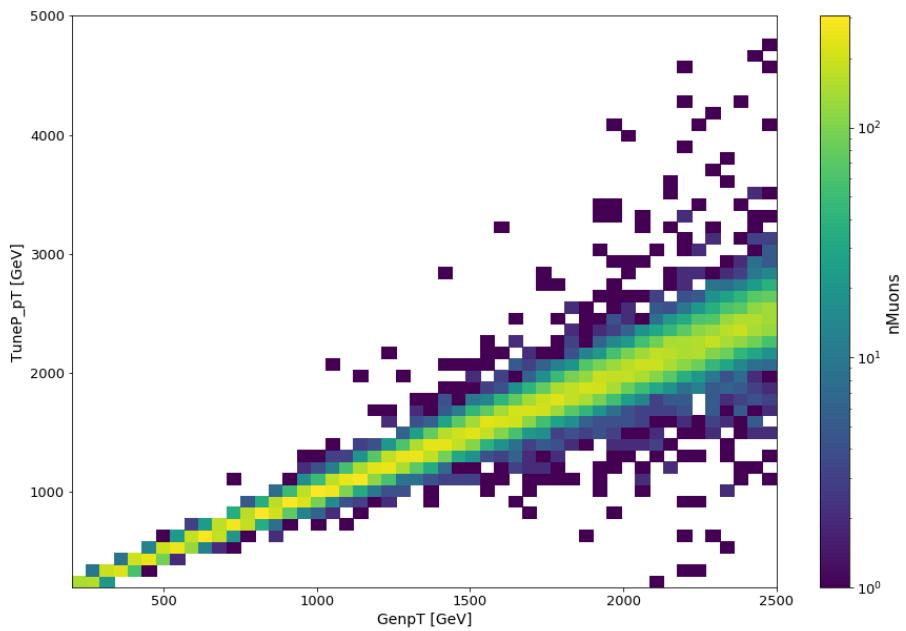


Figura 20: Distribución bidimensional del p_T dado por el algoritmo TuneP en función del p_T de generación para los muones del conjunto de datos de testeo.

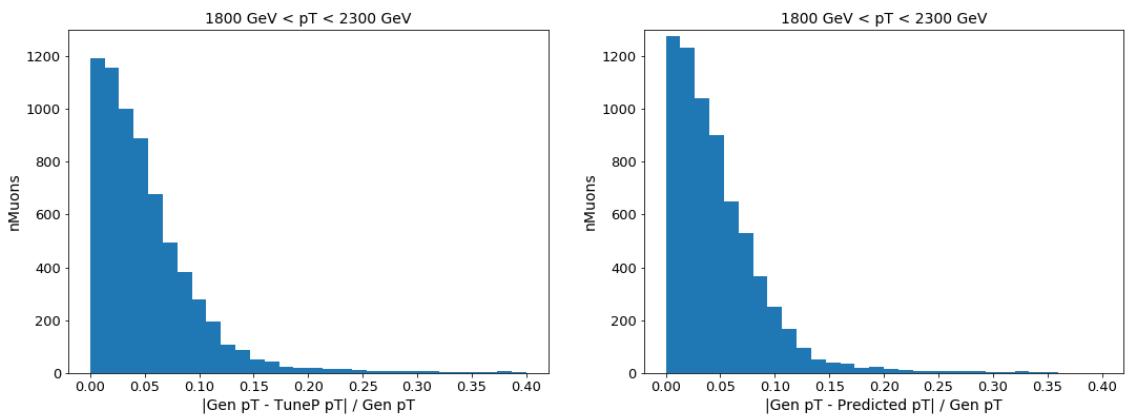


Figura 21: Distribución de la resolución R dada por (4) sobre el conjunto de muones de testeo con $1800 \leq p_T^{GEN} \leq 2300$ GeV. Izquierda: tomando el p_T que da el algoritmo TuneP. Derecha: tomando el p_T predicho por el modelo de regresión al p_T .

8. Conclusiones

Resumen de los resultados

Mejoras futuras: incluir CSCs en el entrenamiento, optimización de la red.

Referencias

- [1] M. et al. Tanabashi. Review of particle physics. *Phys. Rev. D*, 98:030001, Aug 2018.
- [2] Mary K Gaillard, Paul D Grannis, and Frank J Sciulli. The standard model of particle physics. *Reviews of Modern Physics*, 71(2):S96, 1999.
- [3] The CMS Collaboration. Search for a narrow resonance in high-mass dilepton final states in proton-proton collisions using 140 fb^{-1} of data at $\sqrt{s} = 13 \text{ TeV}$. Technical Report CMS-PAS-EXO-19-019, CERN, Geneva, 2019.
- [4] The CMS Collaboration. Search for heavy gauge w bosons in events with an energetic lepton and large missing transverse momentum at $s=13 \text{ tev}$. *Physics Letters B*, 770:278 – 301, 2017.
- [5] J. Schmidhuber. Deep Learning. *Scholarpedia*, 10(11):32832, 2015. revision #184887.
- [6] M. Tanabashi et al. Review of Particle Physics. *Phys. Rev.*, D98(3):030001, 2018.
- [7] Tai Sakuma. Cutaway diagrams of CMS detector. May 2019.
- [8] The CMS Collaboration. CMS Tracking Performance Results from Early LHC Operation. *Eur. Phys. J. C*, 70(arXiv:1007.1988. CERN-PH-EP-2010-019. CMS-TRK-10-001):1165. 29 p, Jul 2010.
- [9] Sketches of the cms tracker detector. <https://twiki.cern.ch/twiki/bin/view/CMSPublic/DPGResultsTRK>. Accessed: 2020-07-27.
- [10] The CMS Collaboration. Performance of the CMS Drift Tube Chambers with Cosmic Rays. *JINST*, 5:T03015, 2010.
- [11] The CMS Collaboration. Performance of the CMS Cathode Strip Chambers with Cosmic Rays. *JINST*, 5:T03018, 2010.
- [12] R Breedon and D Wenman. Images of the CMS Muons - Cathode Strip Chambers (CSC). CMS Collection., Nov 2008.
- [13] Sandra Vieira, Walter Pinaya, and Andrea Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience and Biobehavioral Reviews*, 74, 01 2017.
- [14] The CMS collaboration. Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques. 2020.
- [15] R. E. Kalman. A new approach to linear filtering and prediction problems”transaction of the asme journal of basic. 1960.

- [16] T. Sjöstrand, S. Mrenna, and P. Skands. A brief introduction to pythia 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.
- [17] Repositorio: scripts para la generación de la muestra. https://github.com/fmanteca/Zprime_production. Accessed: 2020-07-27.
- [18] Cmssw. <https://github.com/cms-sw/cmssw>. Accessed: 2020-07-27.
- [19] S. Agostinelli et al. GEANT4—a simulation toolkit. *Nucl. Instrum. Meth. A*, 506:250–303, 2003.
- [20] Código: selección de muones y segmentos. <https://github.com/fmanteca/RECOAnalysis/blob/master/plugins/RECOAnalysis.cc>. Accessed: 2020-07-27.
- [21] Fons Rademakers and René Brun. Root: an object-oriented data analysis framework. *Linux Journal*, 1998:6, 1998.
- [22] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.
- [23] Código: procesado de datos. https://github.com/fmanteca/HighPt_DNN/tree/master/processing. Accessed: 2020-07-27.
- [24] Martín Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [25] François Chollet et al. keras, 2015.
- [26] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.
- [27] S. Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München, 1991.
- [28] Michael P Perrone, Haidar Khan, Changhoan Kim, Anastasios Kyrillidis, Jerry Quinn, and Valentina Salapura. Optimal mini-batch size selection for fast gradient descent. *arXiv preprint arXiv:1911.06459*, 2019.
- [29] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks : the official journal of the International Neural Network Society*, 12(1):145—151, January 1999.
- [30] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.
- [31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.