



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Uso de modelos de aprendizaje profundo para simular
colisiones de partículas en física de altas energías

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: García Ruiz, Francisco Tomás

Tutor/a: Gómez Adrian, Jon Ander

Cotutor/a externo: Salt Cairols, Jose Francisco

Director/a Experimental: Ruiz de Austri Bazan, Roberto

CURSO ACADÉMICO: 2023/2024

Resum

El Model Estàndard pretén explicar i classificar la matèria, les forces i les energies que la controlen. Tot i això, presenta certes mancances i no aconsegueix explicar alguns comportaments de la matèria, com la matèria fosca. Projectes com ATLAS al Gran Col·lisionador d'Hadrons busquen fissures al Model Estàndard, anomenats senyals, que podrien explicar les incògnites que presenta el Model Estàndard. A Física d'Altes Energies, s'utilitzen tècniques de simulació de Montecarlo per obtenir dades simulades, les quals, juntament amb dades experimentals, són vitals per descobrir nous senyals i validar les hipòtesis realitzades pels físics. Això no obstant, l'alt cost temporal i computacional d'aquesta tècnica limita el progrés científic a causa de la velocitat de generació d'esdeveniments i limitacions al pressupost. La generació d'esdeveniments es pot accelerar amb l'ajuda de models d'aprenentatge profund, cosa que permetria augmentar el volum d'esdeveniments generats i reduir el cost computacional associat, proporcionant una millora significativa a la velocitat d'estudi de senyals. Això permetria, per exemple, cerques més àmplies de senyals de la denominada nova física. Per tant, proposem l'ús de xarxes generatives adversarials com a alternativa econòmica i ràpida al mètode de Montecarlo per a la generació de nous esdeveniments.

Paraules clau: Aprenentatge profund, Simulació por Montecarlo, Models Generatius Profunds, Gran Col·lisionador de Hadrons, Xarxes Generatives Adversatives, Col·lisió de Partícules d'Alta Energia

Resumen

El Modelo Estándar pretende explicar y clasificar la materia, las fuerzas y las energías que la controlan. Sin embargo, presenta ciertas carencias y no logra explicar algunos comportamientos de la materia, como la materia oscura. Proyectos como ATLAS en el Gran Colisionador de Hadrones buscan fisuras en el Modelo Estándar, llamadas señales, que podrían explicar las incógnitas que presenta el Modelo Estándar. En Física de Altas Energías, se utilizan técnicas de simulación de Montecarlo para obtener datos simulados, los cuales, junto con datos experimentales, son vitales para descubrir nuevas señales y validar las hipótesis realizadas por los físicos. No obstante, el alto costo temporal y computacional de esta técnica limita el progreso científico debido a la velocidad de generación de eventos y limitaciones en el presupuesto. La generación de eventos se puede acelerar con la ayuda de modelos de aprendizaje profundo, lo que permitiría aumentar el volumen de eventos generados y reducir el coste computacional asociado, proporcionando una mejora significativa en la velocidad de estudio de señales. Esto permitiría, por ejemplo, búsquedas más amplias de señales de la denominada nueva física. Por lo tanto, proponemos el uso de redes generativas antagónicas como una alternativa económica y rápida al método de Montecarlo para la generación de nuevos eventos.

Palabras clave: Aprendizaje profundo, Simulación por Montecarlo, Modelos Generativos Profundos, Gran Colisionador de Hadrones, Redes Generativas Antagónica, Colisión de Partículas de Alta Energía

Abstract

The Standard Model aims to explain and classify matter, forces, and energies that control it. However, it has certain shortcomings and cannot explain some phenomena of matter, such as dark matter. Projects like ATLAS at the Large Hadron Collider are searching for cracks in the Standard Model, called signals, which could explain the uncertainties

presented by the Standard Model. In High-Energy Physics, Montecarlo simulation techniques are used to obtain simulated data, which along with experimental data are vital for discovering new signals and validating hypotheses made by physicists. However, the high temporal and computational cost of this technique limits scientific progress due to the speed of event generation and budget limitations. The events generation can be accelerated with the help of deep learning models, which would allow augmenting the volume of generated events and reducing the associated computational cost, providing a significant improvement in the speed of signal studies. This would allow, for example, broader searches for signals of the so-called Physics Beyond the Standard Model. Therefore, we propose the use of generative adversarial networks as an economic and fast alternative to the Montecarlo method for the generation of new events.

Key words: Deep Learning, Montecarlo Simulation, Deep Generative Models, Large Hadron Collider, Generative Adversarial Networks, Autoencoders, High-Energy Particle Collision

Índice general

Índice general	V
Índice de figuras	VII
Índice de tablas	VIII
<hr/>	
1 Introducción	1
1.1 Motivación	2
1.2 Objetivos	3
1.3 Estructura de la memoria	4
2 Estado del arte	5
2.1 Análisis del estado actual	6
3 Conjunto de datos	9
3.1 Obtención	9
3.2 Formato	10
3.3 Análisis Exploratorio	11
3.4 Resumen del capítulo	12
4 Solución propuesta	15
4.1 Redes Generativas Antagónicas	15
4.1.1 WGAN	16
4.1.2 WGAN-GP	17
4.2 Resumen del capítulo	18
5 Hardware y software usado	19
5.1 Hardware	19
5.2 Software	19
5.2.1 Análisis de datos y representación	20
5.2.2 Diseño de los modelos	21
5.3 Resumen del capítulo	22
6 Métricas de evaluación	23
6.1 Métricas de similitud	23
6.2 Métricas de generación	25
6.3 Resumen del capítulo	26
7 Experimentación	27
7.1 Experimento con <i>t\bar{t}</i> del SM	27
7.1.1 Preprocesado	27
7.1.2 Generador de máscaras	30
7.1.3 Arquitectura del modelo	30
7.1.4 Resultados	33
7.2 Aplicaciones en procesos BSM	41
7.2.1 Resultados	41
8 Conclusiones	49
8.1 Trabajos futuros	50
Bibliografía	51

Apéndices

A Componentes de los modelos	55
A.1 Funciones de activación	55
A.2 Optimizadores	56
B Distribuciones de los datos	59
B.1 Propiedades del evento	59
B.2 Propiedades de las partículas	59
C Objetivos de desarrollo sostenible	65

Índice de figuras

1.1	Reconstrucción de una colisión protón-protón en ATLAS[3].	2
1.2	Esquema del detector ATLAS en el CERN [4].	3
3.1	Histograma de las variables MET y MET ϕ para los conjuntos de datos antes y después del procesamiento.	11
3.2	Histograma de la Energía de los fotones antes del logaritmo.	12
3.3	Histograma de la variable ϕ de los fotones.	12
3.4	Histograma de la Energía de los fotones después del logaritmo.	12
3.5	Histograma de la variable η de los fotones.	12
4.1	Diagrama de entrenamiento de una GAN.	16
5.1	Diagrama de la arquitectura del chip GA102 [32].	20
5.2	Esquema de la tecnología usada [36].	21
7.1	Diagrama del preprocesado del experimento.	29
7.2	Comparación máscaras generadas y máscaras presentes en el conjunto de datos filtrado (SM).	30
7.3	Diagrama de la arquitectura GAN para el MET y MET ϕ en el experimento.	31
7.4	Diagrama de la arquitectura GAN para las primeras partículas en el experimento.	32
7.5	Diagrama del ejemplo del proceso de generación en el experimento.	34
7.6	Histogramas de las características de la energía faltante del evento (MET, MET ϕ).	35
7.7	Histogramas de las características de los primeros b-jets (b) en el experimento.	35
7.8	Histogramas de las características de los segundos b-jets (b) en el experimento.	36
7.9	Histogramas de las características de los primeros jets (j) en el experimento.	37
7.10	Histogramas de las características de los segundos jets (j) en el experimento.	37
7.11	Histogramas de las características de los electrones (e-) en el experimento.	38
7.12	Histogramas de las características de los positrones (e+) en el experimento.	39
7.13	Histogramas de las características de los muones (m-) en el experimento.	39
7.14	Histogramas de las características de los antimuones (m+) en el experimento.	40
7.15	Histogramas de las características de los fotones (g) en el experimento.	40
7.16	Comparación máscaras generadas y máscaras presentes en el conjunto de datos filtrado (BSM).	41
7.17	Histogramas de las características de la energía faltante del evento (MET, MET ϕ) en stop_02.	42
7.18	Histogramas de las características de los primeros b-jets (b) en stop_02.	42
7.19	Histogramas de las características de los segundos b-jets (b) en stop_02.	43
7.20	Histogramas de las características de los primeros jets (j) en stop_02.	44
7.21	Histogramas de las características de los segundos jets (j) en stop_02.	44
7.22	Histogramas de las características de los electrones (e-) en stop_02.	45

7.23	Histogramas de las características de los positrones (e^+) en stop_02.	46
7.24	Histogramas de las características de los muones (m^-) en stop_02.	46
7.25	Histogramas de las características de los antimuones (m^+) en stop_02.	47
7.26	Histogramas de las características de los fotones (g) en stop_02.	47
B.1	Histogramas de las características de la energía faltante del evento (MET, $MET\phi$) y la energía total.	59
B.2	Histogramas de las características de los b-jets (b).	60
B.3	Histogramas de las características de los jets (j).	60
B.4	Histogramas de las características de los electrones (e^-).	61
B.5	Histogramas de las características de los positrones (e^+).	61
B.6	Histogramas de las características de los muones (m^-).	62
B.7	Histogramas de las características de los antimuones (m^+).	62
B.8	Histogramas de las características de los fotones (g).	63

Índice de tablas

3.1	Símbolos que representan las partículas finales de cada evento.	10
3.2	Procesos de señales SM y BSM generados con su identificación.	14
7.1	Métricas de los resultados del experimento SM.	38
7.2	Métricas de los resultados del experimento BSM.	48

CAPÍTULO 1

Introducción

En el universo de la física, el Modelo Estándar, o *Standard Model* (SM), prevalece como el paradigma teórico más ampliamente aceptado en la actualidad, ofreciendo una explicación fundamental sobre la constitución de la materia y las fuerzas primordiales que orquestan los procesos físicos. No obstante, ciertos fenómenos observados escapan a sus predicciones teóricas y existen limitaciones intrínsecas en el modelo para abordar cuestiones esenciales como: el origen de la masa, el problema CP fuerte¹, la oscilación de neutrinos, la asimetría materia-antimateria, y la naturaleza de la materia oscura o la energía oscura.

La física de partículas y su exploración a través de experimentos de colisiones de alta energía han llevado a avances significativos en nuestra comprensión de las leyes fundamentales que rigen el universo. En particular, el Gran Colisionador de Hadrones, o *Large Hadron Collider* (LHC), emplazado en la Organización Europea para la Investigación Nuclear, o *Conseil Européen pour la Recherche Nucléaire* (CERN), ha sido un epicentro de descubrimientos, permitiendo la observación y el estudio de partículas subatómicas en condiciones extremas.

En el CERN, el mencionado colisionador de partículas se utiliza para adquirir **datos reales** de colisiones protón-protón (Figura 1.1), que luego se seleccionan y registran como eventos. La selección en línea de eventos se realiza a partir de un disparador, y luego los eventos se procesan y reconstruyen. El proceso de recolección de datos tiene lugar en cuatro puntos de colisión distintos a lo largo del anillo circular de 27 kilómetros del LHC, utilizando siete detectores empleados por diferentes experimentos [1], [2].

Durante el transcurso de un experimento, se generan datos simulados a través de varios métodos para facilitar la comparación entre datos de diferentes modelos teóricos y las características reales de los eventos. Posteriormente, se discuten los hallazgos derivados del análisis de los datos acumulados. Estas investigaciones tienen como objetivo identificar nuevos marcos teóricos, conocidos como Física Más Allá del Modelo Estándar, *Beyond Standard Model* (BSM), que puedan brindar explicaciones para las limitaciones mencionadas del modelo actual.

Sin embargo, la generación de datos reales en experimentos como el LHC es un proceso costoso y limitado. La premisa que planteamos radica en la implementación de Modelos Generativos Profundos para simular eventos físicos de forma más expedita en comparación con los enfoques tradicionales basados en Montecarlo. En este trabajo de grado, nos sumergimos en el emocionante mundo de las Redes Generativas Antagónicas, o *Generative Adversarial Networks* (GANs), para abordar precisamente este desafío.

¹El problema CP fuerte hace referencia de por qué la Cromodinámica Cuántica (QCD) no parece romper la simetría de la paridad de carga (CP del inglés Charge Parity).

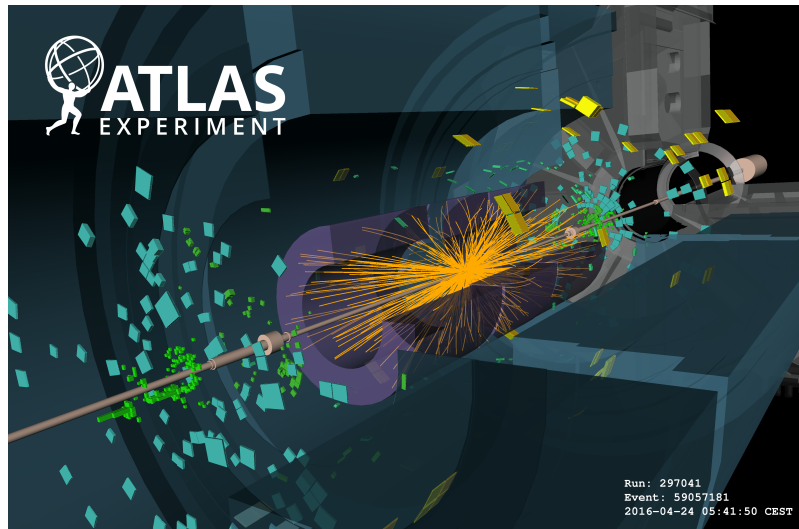


Figura 1.1: Reconstrucción de una colisión protón-protón en ATLAS[3].

Nuestra atención se dirige en particular hacia el experimento ATLAS (Figura 1.2), una iniciativa de propósito general con el objetivo de investigar el origen de las masas y la física BSM. Aunque hasta el momento no existen indicios categóricos de modelos de nueva física confirmados, futuras exploraciones podrían potencialmente revelar resultados prometedores.

1.1 Motivación

En el ámbito de la simulación de procesos físicos, la utilización de modelos de Aprendizaje Profundo se convierte en una alternativa fundamental. Por lo general, la simulación se realiza siguiendo el método de Montecarlo, que se desarrolla en dos etapas:

1. En primer lugar, se extraen números pseudoaleatorios de una distribución de probabilidad.
2. Posteriormente, los números generados son transformados por un algoritmo en eventos físicos simulados, que en este caso son colisiones de partículas de alta energía.

Esta metodología supone una alta demanda de recursos computacionales debido a la complejidad de las simulaciones numéricas. El esfuerzo conjunto para la investigación en este campo se ve mermado por las restricciones en la velocidad de simulación y los presupuestos necesarios. Para ilustrar, la generación de un evento de Montecarlo en experimentos de física de partículas, incluyendo la respuesta del detector, puede consumir hasta 10 minutos de tiempo de cómputo debido al uso de algoritmos no óptimos en herramientas como GEANT4 [5].

En experimentos futuros, la acumulación de datos reales generados con el colisionador requerirá la generación de miles de millones de eventos. Sin embargo, los métodos actuales para la generación de eventos requieren de unos tiempos de ejecución prohibitivos e incompatibles. Además, si se continuase empleando estos métodos, aumentaría sustancialmente la carga ambiental en términos de consumo de energía. Por lo tanto, la optimización de la generación de eventos se convierte en una tarea imperativa tanto para mejorar la eficiencia como para abordar estas preocupaciones.

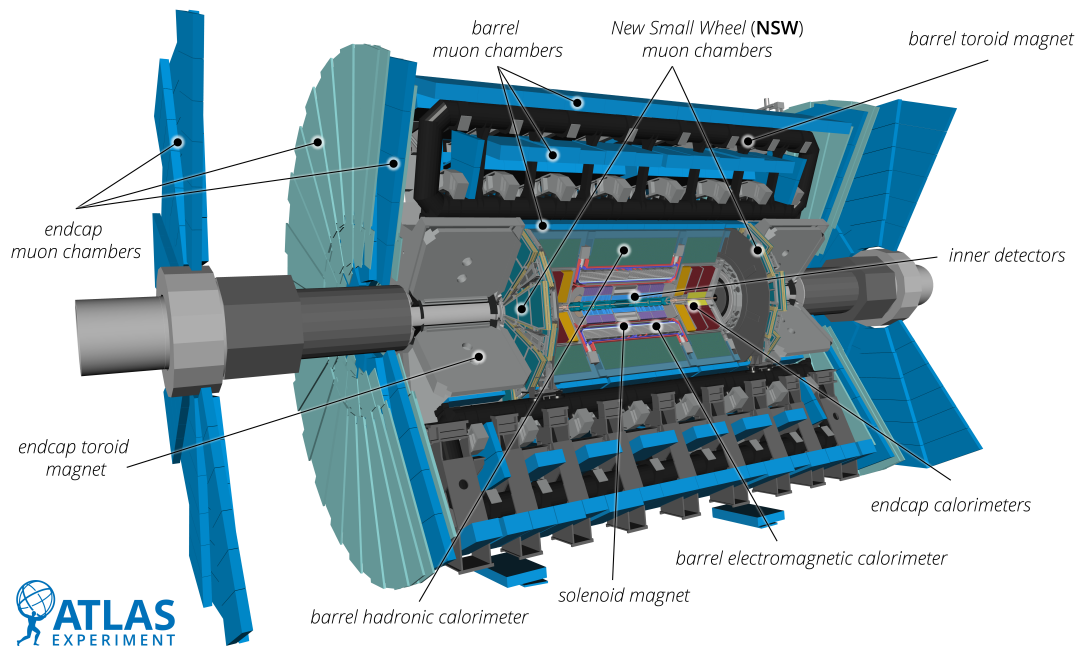


Figura 1.2: Esquema del detector ATLAS en el CERN [4].

El uso de modelos generativos ya ha demostrado ser de gran ayuda en la producción de datos simulados en grandes cantidades, lo que ahorra recursos temporales, energéticos y económicos durante los estudios y análisis de física. Además, a menudo es necesario generar eventos que exhiban características concretas observadas en los datos reales. Actualmente, los generadores de eventos producen una cantidad sustancial de eventos y posteriormente filtran los interesantes, descartando los eventos restantes. La exploración de métodos para minimizar la generación de eventos irrelevantes también es un tema de interés.

1.2 Objetivos

El objetivo principal de este trabajo de fin de grado es explorar y evaluar la efectividad de las GANs en la generación de datos de eventos de colisiones de partículas protón-protón. Para lograr este objetivo, nos planteamos una serie de metas específicas:

- 1. Entender los Fundamentos de las GANs y su Aplicación en Física de Partículas:** Iniciaremos este estudio adentrándonos en los conceptos fundamentales de las GANs y su funcionamiento. Investigaremos cómo esta técnica ha sido utilizada en otros campos y cómo podemos aplicarla de manera efectiva en la generación de datos de colisiones de partículas.
- 2. Adaptar Arquitecturas de Red e Hiperparámetros para Datos de Física de Partículas:** Dado que los datos de eventos de colisiones de partículas tienen características específicas, trabajaremos en ajustar las arquitecturas de red y los hiperparámetros de las GANs para maximizar su rendimiento en la generación de datos realistas y coherentes en el contexto de la física de partículas.
- 3. Comparar Resultados con Simulados por Montecarlo:** Una vez entrenados los modelos, compararemos los datos generados por las GANs con datos simulados por Montecarlo previamente existentes. Utilizaremos métricas específicas para evaluar

la similitud y calidad de los datos generados en comparación con los datos simulados por Montecarlo.

En resumen, nuestro trabajo se centra en crear un marco de trabajo eficiente y preciso para la generación de eventos a partir de modelos generativos profundos, específicamente GANs. A través de la comparación con enfoques convencionales y la exploración de diversas estrategias, buscamos mejorar tanto la eficiencia como la calidad de los eventos generados, contribuyendo, así, al avance en la simulación de procesos físicos en el ámbito de la física de partículas.

1.3 Estructura de la memoria

La presente memoria del Trabajo de Fin de Grado (TFG) se estructura en diversas secciones con el propósito de proporcionar una visión completa y detallada de la investigación realizada.

- Capítulo 1. La introducción contextualiza y resalta la necesidad de superar las limitaciones de los enfoques tradicionales de simulación en física de partículas.
- Capítulo 2. Sobre el estado del arte, examina investigaciones relevantes en el campo de modelos generativos profundos y su aplicación en la generación de eventos físicos, identificando avances y desafíos.
- Capítulo 3. Dedicado a presentar el conjunto de datos, detalla la obtención y formato de los datos utilizados, seguido de un análisis exploratorio para comprender sus características.
- Capítulo 4. Sobre la solución propuesta, se centra en las aplicaciones y soluciones ideadas para la generación de eventos, detallando su estructura y adaptación a la problemática.
- Capítulo 5. Sobre el hardware y software usado, describe los recursos utilizados en el desarrollo del trabajo.
- Capítulo 6. Explica con detalle las métricas de evaluación y los criterios para medir la calidad y eficiencia de los modelos generativos.
- Capítulo 7. Describe los experimentos realizados para validar la propuesta.
- Capítulo 8. Presenta las conclusiones, resume los resultados clave y esboza posibles direcciones futuras.

CAPÍTULO 2

Estado del arte

En la búsqueda de señales de nueva física en colisiones protón-protón, la aplicación de técnicas de aprendizaje profundo ha ganado relevancia en la comunidad científica. Se han realizado diversos intentos para detectar estas señales utilizando tanto algoritmos supervisados como no supervisados. A pesar de los esfuerzos, hasta el momento no se ha encontrado una señal clara de física más allá del Modelo Estándar (BSM) en los datos del Gran Colisionador de Hadrones (LHC). Este contexto motiva la necesidad de métodos innovadores.

Un experimento similar al que proponemos es una investigación previa enfocada en la generación de eventos a partir de un proceso físico que ya utilizó Modelos Generativos Profundos [6]. Explora la viabilidad de aprender la generación de eventos y la frecuencia de ocurrencia de dichos eventos utilizando varios modelos generativos de aprendizaje automático para producir eventos similares a los generadores de Montecarlo. Las arquitecturas propuestas se pueden clasificar en dos conjuntos distintos: por un lado, las *Generative Adversarial Networks* (GANs), que incluyen variantes como WGAN [7], WGAN-GP [8], LS-GAN [9] y MMDGAN [10]; y por otro lado, los *Variational Autoencoders* (VAEs), entre los que se destaca especialmente el β -VAE [11]. Este enfoque produjo distribuciones muy similares a los eventos generados por Montecarlo y generó eventos significativamente más rápido. Además cuenta con otras posibles aplicaciones como la estimación y muestreo genérico de densidad y la detección de anomalías.

Un acontecimiento relevante dentro de la comunidad científica en relación con este estudio fue *The Dark Machines Anomaly Score Challenge*, que se discute en detalle en [12]. El conjunto de datos utilizado en nuestro trabajo es el mismo que se empleó en este desafío, y sus características se detallan en el **Capítulo 3**. El objetivo principal del desafío era identificar indicios de nueva física en el LHC utilizando algoritmos de aprendizaje no supervisado. Inicialmente, se propuso la implementación de una puntuación de anomalía para definir regiones de señal independientes del modelo. Posteriormente, se examinó una amplia gama de algoritmos de detección de anomalías y estimación de densidad, desarrollados específicamente para el desafío de datos, y se evaluó su rendimiento en una variedad de entornos de análisis realistas. Este desafío sirve como base para nuevos puntos de referencia y el desarrollo de enfoques más efectivos en la detección de señales BSM.

El equipo del proyecto ATLAS en el CERN también investigó este tema, definiendo su propia estrategia para encontrar posibles indicaciones de física más allá del modelo estándar [13]. En su enfoque, los eventos fueron categorizados según su estado final, lo que resultó en múltiples clases de eventos. Para cada clase de evento, se utilizó un algoritmo de búsqueda automatizado para evaluar la compatibilidad de los datos con los resultados esperados simulados por las simulaciones de MC en diversas distribuciones

sensibles a los efectos de la nueva física. La significancia de cualquier desviación se cuantificó mediante pseudo-experimentos. Si una selección de datos exhibía una desviación significativa, se definiría una región de señal para un análisis posterior en profundidad, incorporando una mejor estimación del fondo. La interpretación estadística de las regiones de señal derivadas de los datos se realizó en un nuevo conjunto de datos. La sensibilidad de este enfoque se evaluó utilizando procesos del modelo estándar, así como señales de referencia que representaban posibles instancias de nueva física.

También cabe mencionar diversas aproximaciones que se han explorado en el entorno de física de partículas. Podemos agrupar estas aproximaciones en las siguientes tres:

Simulación de detectores: abarca enfoques basados en GAN [14]-[16], así como basados en VAE [17]-[19];

Generación de eventos: dentro de esta aproximación se han realizado numerosos trabajos utilizando modelos GAN [20]-[22] y VAE [23];

Integración de Montecarlo: se observa la prevalencia del uso de la regresión en muchos estudios [24], [25], además de algunos intentos de emplear Flujos Normalizadores [26]-[28].

2.1 Análisis del estado actual

Tras el análisis del estado del arte sobre algunas de las contribuciones más relevantes relacionadas con el estudio de la física de partículas, podemos obtener una perspectiva de la situación actual y algunas conclusiones:

- El desafío de datos *The Dark Machines Anomaly Score Challenge* [12] estableció un marco sólido para evaluar el rendimiento de diferentes enfoques en la generación de eventos, y proporcionó un conjunto de datos adecuado para utilizarlo como punto de partida para desarrollar nuevos modelos.
- En la actualidad, esta área está experimentando avances significativos debido a la intersección de técnicas de aprendizaje automático y la física de partículas. La adopción de modelos generativos, como las GANs y los VAEs, ha permitido abordar de manera más eficiente y precisa la simulación de eventos de colisión. Sin embargo, sigue siendo un área novel y persisten desafíos en términos de escalabilidad y representación precisa, lo que motiva la continua exploración y desarrollo de nuevos enfoques y técnicas.
- Algunos enfoques, como DijetGAN [15], han demostrado ser eficaces en la generación de eventos de un conjunto específico, como *dijets*¹, donde dos partículas altamente energéticas son producidas en la colisión. Sin embargo, estos modelos pueden tener un rendimiento limitado en comparación con otro tipo de modelos, en particular, cuando se aplican a clases de eventos más amplias y diversas, como se ha demostrado en [6].

A continuación, algunos trabajos destacados por sus enfoques para abordar el problema y los resultados obtenidos:

¹Un evento en el que dos partículas de alta energía, típicamente quarks o gluones, son producidas en direcciones opuestas como resultado de una colisión de partículas

- El trabajo presenta varios enfoques para la generación eficiente de eventos tanto del Modelo Estándar como Más Allá del Modelo Estándar, destacando el uso de modelos β -VAEs y la introducción de un nuevo modelo, el α -VAE. Si bien se obtuvieron resultados positivos, se reconoce la necesidad de ajustar la arquitectura y parámetros internos para mejorar la generación de algunas características de las partículas. Además, la combinación con un modelo adicional, el Modelo Gaussiano de Mezcla Bayesiana, no mejoró los resultados, aunque se lograron los objetivos inicialmente definidos de mejorar la eficiencia computacional y mantener la precisión [23].
- El estudio investiga el uso de flujos normalizadores en el espacio latente de VAEs para mejorar la precisión del proceso de aprendizaje previo. Se observa una mejora general en la identificación de anomalías al utilizar NF, especialmente con la variante Householder del modelo de flujo normalizador Sylvester. Esto representa una mejora sobre los modelos Conv-VAE previamente presentados, lo que sugiere un avance en la detección de anomalías en muestras de señales de referencia [28].
- Se presenta un modelo GAN basado en una unidad recurrente capaz de generar listas de partículas con longitudes variables. Aunque se enfrentaron limitaciones técnicas en cuanto a la longitud de la cadena de partículas devuelta, el modelo demuestra ser capaz de emular el efecto de la contaminación de fondo en colisiones del LHC, lo que sugiere una aplicación prometedora en la emulación de fenómenos complejos en la física de partículas [22].
- Se propone una estrategia de aumento de datos que utiliza un modelo generativo de aprendizaje profundo para convertir una representación de eventos de colisión en el nivel de generador a la representación correspondiente en el nivel de reconstrucción. Esta estrategia muestra una reducción sustancial en términos de recursos computacionales requeridos y mantiene la precisión en la generación de conjuntos de datos, lo que sugiere su utilidad potencial en la era de mediciones de alta precisión del LHC [21].
- El documento propone un método novedoso para simular rápidamente jets de partículas en el LHC mediante VAEs, evitando los pasos: simulación del detector y reconstrucción de colisiones. Al representar los chorros como listas de elementos y entrenar el VAE para generar la lista de elementos detectados, se logra una precisión destacada en el cuadrimomento del jet y los momentos de los elementos. La rapidez y precisión del método lo hacen útil para estudios en el LHC, aunque hay áreas potenciales de mejora, como la descripción de la subestructura del jet. En resumen, el documento ofrece una contribución significativa al proponer una alternativa rápida y precisa para simular eventos en el LHC, identificando áreas para investigaciones futuras [18].

La elección de centrarse en GANs en un trabajo relacionado con la generación de eventos de colisión se basa en la capacidad probada de estas redes para abordar desafíos específicos y complejos en la simulación de eventos. Así como la necesidad actual para ofrecer un marco de trabajo que presente versatilidad para adaptarse a una variedad de clases de eventos. Por estos motivos, hacemos uso del conjunto de datos ofrecido en *The Dark Machines Anomaly Score Challenge* [12], que incluye tanto procesos del Modelo Estándar como procesos Más Allá del Modelo Estándar, con el objetivo de construir un modelo viable capaz de ser entrenado para cualquier tipo de proceso.

CAPÍTULO 3

Conjunto de datos

El conjunto de datos utilizado en este trabajo es el mismo que se propuso en el “The Dark Machines Anomaly Score Challenge” [12], y el utilizado anteriormente por nuestro compañero Raúl Balanzá García [23], el cual está disponible para su descarga en [29]. Una explicación mucho más exhaustiva del procedimiento de generación de datos puede encontrarse en [30].

3.1 Obtención

Los eventos simulan el resultado de colisiones protón-protón, generando condiciones similares a las observadas en el LHC, con una energía en el centro de masa de 13 TeV¹. Los procesos de señal y fondo fueron desarrollados mediante métodos y herramientas convencionales. Tras el proceso de generación las partículas en su estado final, tal como se describe detalladamente en la Tabla 3.1, se guardan en un archivo de texto CSV, donde cada evento es representado en una única línea.

En el escenario de una colisión, se producen un máximo de 20 objetos, y para cada uno de ellos, se registran propiedades diversas como la energía total (E), el impulso transversal (p_T), la pseudorapidez (η) y el ángulo azimutal (ϕ). La inclusión de un evento en el conjunto de datos se efectúa solo si cumple al menos uno de los siguientes criterios:

- Uno o más **jets** o **b-jets** con $p_T > 60 \text{ GeV}$ y $|\eta| < 2,8$.
- Uno o más **electrones** con $p_T > 25 \text{ GeV}$ y $|\eta| < 2,47$, excepto para $1,37 < |\eta| < 1,52$.
- Uno o más **muones** con $p_T > 25 \text{ GeV}$ y $|\eta| < 2,7$.
- Uno o más **fotones** con $p_T > 25 \text{ GeV}$ y $|\eta| < 2,37$.

Los requisitos para las partículas presentes en el conjunto de datos fueron los siguientes:

- **jet** o **b-jet**: $p_T > 20 \text{ GeV}$ y $|\eta| < 2,8$.
- **electrón** o **muón**: $p_T > 15 \text{ GeV}$ y $|\eta| < 2,7$.
- **fotón**: $p_T > 20 \text{ GeV}$ y $|\eta| < 2,37$.

¹Un eV es la cantidad de energía cinética ganada o perdida por un solo electrón al ser acelerado desde el reposo a través de una diferencia de potencial eléctrico de 1 voltio en el vacío.

La intención era producir un conjunto de datos versátil, apto para diversos tipos de investigaciones, cada una posiblemente demandando criterios de selección diferentes. Los detalles completos de los procesos generados están presentados en la Tabla 3.2.

ID	Objeto
b	b-jet
j	jet
e-	electrón
e+	positrón
m-	muón
m+	antimuón
g	fotón

Tabla 3.1: Símbolos que representan las partículas finales de cada evento.

3.2 Formato

Como se mencionó en la sección anterior, los datos se presentan en un formato de archivo CSV, donde cada evento se representa en una sola línea. Cada línea varía en longitud y consta de tres especificadores de eventos seguidos de las características cinemáticas de cada objeto dentro del evento.

Cada archivo representa un proceso distinto. La cantidad de eventos generados para cada proceso está garantizada de ser igual o mayor que el requisito mínimo para obtener $10 fb^{-1}$ de datos² ($N_{10 fb^{-1}}$). Además, cuando $N_{10 fb^{-1}} < 20000$, se proporciona un segundo archivo con 20000 líneas. El formato de todos los archivos CSV es el siguiente:

ID de evento; ID de proceso; peso del evento; MET; METphi;
obj1, E1, pt1, eta1, phi1; ...

El ID de evento actúa como un identificador único para cada evento, representado por un valor entero. Su propósito es facilitar la depuración específica de eventos. Por otro lado, el ID del proceso es una cadena que corresponde al proceso responsable de generar el evento, como se describe en la Tabla 3.2. Si bien el peso del evento depende de la sección eficaz de un proceso específico y del número de eventos en un archivo, no se considera en nuestros experimentos. Estos componentes son los especificadores de eventos en cada línea del archivo CSV.

En cuanto a las características cinemáticas, las entradas MET y MET ϕ corresponden a la magnitud E_{miss}^T y al ángulo azimutal ϕ_{miss}^T del vector de energía transversal faltante dentro del evento. Estos valores representan la energía transversal y el ángulo azimutal de los objetos que efectivamente no se detectan. Dichos objetos se consideran que han evitado la detección.

Los identificadores de objetos (obj1, obj2, etc.) son cadenas que identifican de manera única a cada objeto dentro del evento, de acuerdo con las especificaciones detalladas en la Tabla 3.1. Después de cada identificador de objeto, hay cuatro valores separados por comas que describen de manera integral las características del objeto: E1 (E), pt1 (pT), eta1 (η) y phi1 (ϕ). Estas características se alinean con las mencionadas en la sección anterior. La disposición de las partículas dentro de cada evento sigue la secuencia de: b-jets, jets, leptones y fotones. Dentro de cada tipo de partícula, están ordenadas en orden descendente según su momento transversal (pT). En consecuencia, en la sección de resultados, cuando mencionamos la partícula líder, nos referimos a aquella con el pT más alto dentro de su categoría respectiva. Los eventos están categorizados por proceso

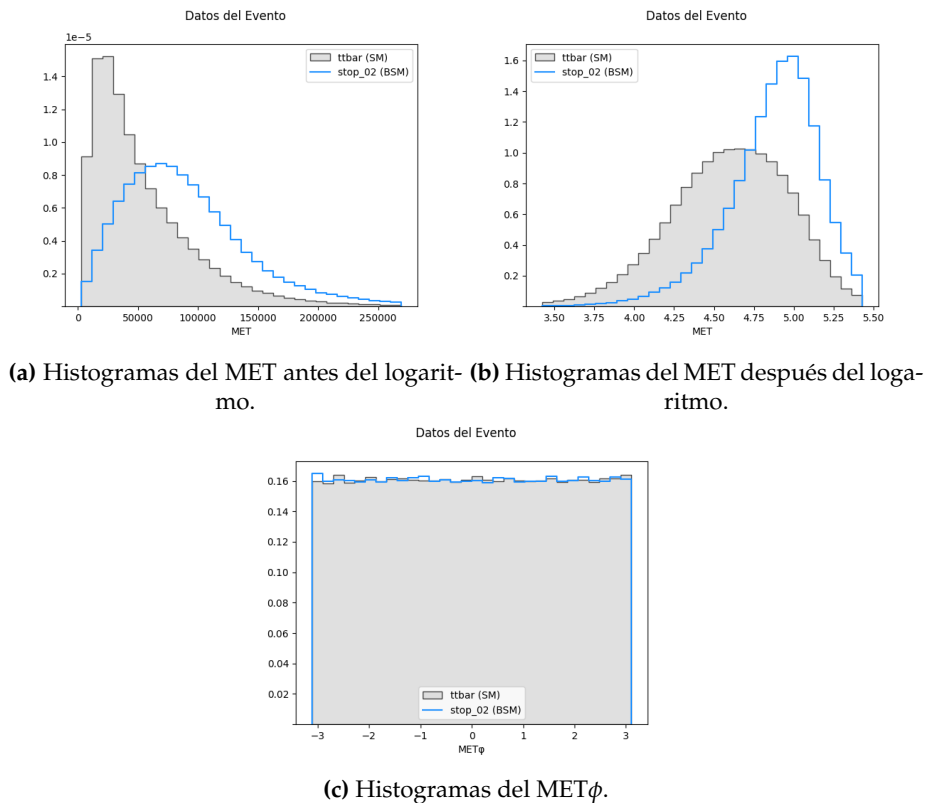


Figura 3.1: Histograma de las variables MET y $MET\phi$ para los conjuntos de datos antes y después del procesamiento.

y distribuidos en 65 archivos, con un tamaño total de 66.5 GB. Sin embargo, durante la fase de experimentación, solo se seleccionaron los archivos más grandes para cada tipo de proceso. Específicamente, para los procesos que alcanzaron $10 fb^{-1}$ con menos de 20K eventos, se eligió el archivo que contiene 20K eventos.

3.3 Análisis Exploratorio

Antes de llevar a cabo cualquier experimento, se realizó un análisis del conjunto de datos para determinar si las diferentes variables seguían alguna distribución conocida. Al realizar un estudio de las variables de la energía perdida del evento, encontramos que:

- MET tiene la forma de una distribución log-normal, como ilustra la Figura 3.1a, que puede convertirse en una distribución normal aplicando el logaritmo, como muestra la Figura 3.1b.
- $MET\phi$ seguía una distribución uniforme, como muestra la Figura 3.1c.

Después de considerar por separado todos los tipos de partículas que pueden encontrarse en los eventos (enumeradas en la Tabla 3.1), se observó que, entre las cuatro características que describen cada partícula:

- E y p_T , al igual que el MET, tienen la forma de una distribución log-normal, como puede verse en la Figura 3.2, y por tanto, pueden convertirse en una distribución normal aplicando el logaritmo, como se muestra en la Figura 3.4.
- ϕ seguía una distribución uniforme, como se ve en la Figura 3.3.

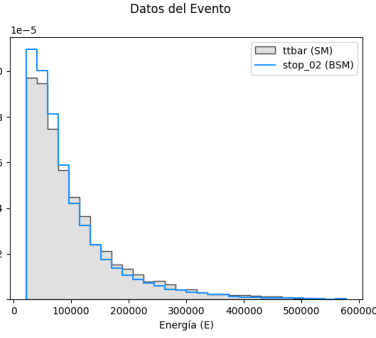


Figura 3.2: Histograma de la Energía de los fotones antes del logaritmo.

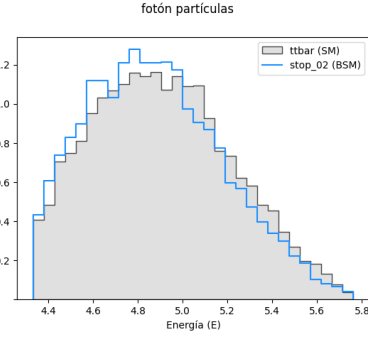


Figura 3.4: Histograma de la Energía de los fotones después del logaritmo.

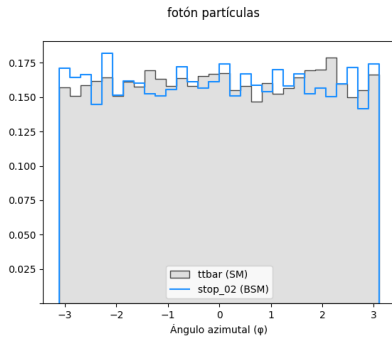


Figura 3.3: Histograma de la variable ϕ de los fotones.

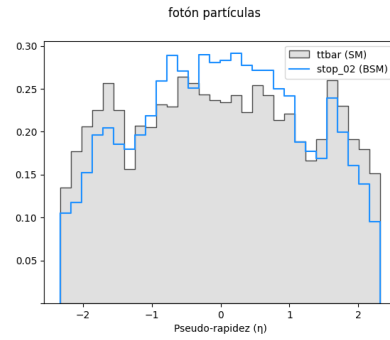


Figura 3.5: Histograma de la variable η de los fotones.

- η no parecía describir ninguna forma de distribución conocida, especialmente en los fotones, como podemos ver en la Figura 3.5.

Tras realizar el análisis descrito, esta información se tuvo en cuenta para convertir los datos en un formato adecuado que se proporcionaría como entrada a cada modelo, aplicando el logaritmo en las variables E y p_T de las partículas y MET del evento.

3.4 Resumen del capítulo

En resumen, en este capítulo, se ha detallado el conjunto de datos utilizado para simular colisiones de partículas, específicamente protones, en condiciones similares a las observadas en el LHC con una energía de 13 TeV. El conjunto de datos fue extraído de *The Dark Machines Anomaly Score Challenge* y presenta eventos en un formato CSV, donde cada evento está representado por una sola línea con especificadores de eventos y características cinemáticas de las partículas involucradas.

Se creó un conjunto de datos diverso, adecuado para múltiples investigaciones, con eventos categorizados en procesos del Modelo Estándar (SM) y más allá del Modelo Estándar (BSM).

En la fase de análisis exploratorio, se identificaron las distribuciones de las variables clave. Se encontró que la energía perdida del evento (MET) sigue una distribución log-normal, que se convierte en normal al aplicar el logaritmo, mientras que el ángulo azimutal de la energía transversal faltante ($MET\phi$) sigue una distribución uniforme. Características similares se observaron para las partículas individuales, con E y p_T siguiendo una distribución log-normal y ϕ una distribución uniforme, aunque η no mostró una distribución clara.

Esta comprensión de las distribuciones permitió transformar adecuadamente los datos, aplicando el logaritmo a las variables E y pT , así como a MET , para optimizar la entrada a los modelos de aprendizaje automático empleados en los experimentos posteriores.

SM	
Proceso	ID
$pp \rightarrow jj$	njets
$pp \rightarrow W^\pm(+2j)$	w_jets
$pp \rightarrow \gamma(+2j)$	gam_jets
$pp \rightarrow Z(+2j)$	z_jets
$pp \rightarrow t\bar{t}(+2j)$	ttbar
$pp \rightarrow W^\pm t(+2j)$	wtop
$pp \rightarrow W^\pm \bar{t}(+2j)$	wtopbar
$pp \rightarrow W^+W^-(+2j)$	ww
$pp \rightarrow t + jets(+2j)$	single_top
$pp \rightarrow \bar{t} + jets(+2j)$	single_topbar
$pp \rightarrow \gamma\gamma(+2j)$	2gam
$pp \rightarrow W^\pm\gamma(+2j)$	Wgam
$pp \rightarrow ZW^\pm(+2j)$	zw
$pp \rightarrow Z\gamma(+2j)$	Zgam
$pp \rightarrow ZZ(+2j)$	zz
$pp \rightarrow h(+2j)$	single_higgs
$pp \rightarrow t\bar{t}\gamma(+1j)$	ttbarGam
$pp \rightarrow t\bar{t}Z$	ttbarZ
$pp \rightarrow t\bar{t}h(+1j)$	ttbarHiggs
$pp \rightarrow \gamma t(+2j)$	atop
$pp \rightarrow t\bar{t}W^\pm$	ttbarW
$pp \rightarrow \gamma\bar{t}(+2j)$	atopbar
$pp \rightarrow Zt(+2j)$	ztop
$pp \rightarrow Z\bar{t}(+2j)$	ztopbar
$pp \rightarrow t\bar{t}\bar{t}$	4top

BSM	
Proceso	ID
$pp \rightarrow \tilde{g}\tilde{g}(1 \text{ TeV})$	gluino_01
$pp \rightarrow \tilde{g}\tilde{g}(1,2 \text{ TeV})$	gluino_02
$pp \rightarrow \tilde{g}\tilde{g}(1,4 \text{ TeV})$	gluino_03
$pp \rightarrow \tilde{g}\tilde{g}(1,6 \text{ TeV})$	gluino_04
$pp \rightarrow \tilde{g}\tilde{g}(1,8 \text{ TeV})$	gluino_05
$pp \rightarrow \tilde{g}\tilde{g}(2 \text{ TeV})$	gluino_06
$pp \rightarrow \tilde{g}\tilde{g}(2,2 \text{ TeV})$	gluino_07
$pp \rightarrow \tilde{t}_1\tilde{t}_1(220 \text{ GeV}), m_{\tilde{x}0} = 20 \text{ GeV}$	stop_01
$pp \rightarrow \tilde{t}_1\tilde{t}_1(300 \text{ GeV}), m_{\tilde{x}0} = 100 \text{ GeV}$	stop_02
$pp \rightarrow \tilde{t}_1\tilde{t}_1(400 \text{ GeV}), m_{\tilde{x}0} = 100 \text{ GeV}$	stop_03
$pp \rightarrow \tilde{t}_1\tilde{t}_1(800 \text{ GeV}), m_{\tilde{x}0} = 100 \text{ GeV}$	stop_04
$pp \rightarrow Z'(2 \text{ TeV})$	Zp_01
$pp \rightarrow Z'(2,5 \text{ TeV})$	Zp_02
$pp \rightarrow Z'(3 \text{ TeV})$	Zp_03
$pp \rightarrow Z'(3,5 \text{ TeV})$	Zp_04
$pp \rightarrow Z'(4 \text{ TeV})$	Zp_05

Tabla 3.2: Procesos de señales SM y BSM generados con su identificación.

CAPÍTULO 4

Solución propuesta

Tomando en consideración el panorama actual expuesto en el [Capítulo 2](#), y tras seleccionar y analizar detenidamente el conjunto de datos a utilizar, proponemos una solución basada en la implementación de GANs, o combinación de estas con otros modelos, con el propósito de generar datos de eventos de colisión de partículas protón-protón. La elección de GANs se fundamenta en su versatilidad y eficacia en la generación de datos realistas, así como en la flexibilidad que ofrecen para capturar las complejidades de las distribuciones de los datos. A pesar de ser un enfoque ampliamente estudiado, nuestra elección radica en la búsqueda de mejoras y refinamientos en su aplicación específica para nuestro dominio.

La relevancia de utilizar GANs en este contexto se acentúa debido a su habilidad para simular datos realistas, lo que puede acelerar y mejorar la simulación de eventos de colisión de partículas. Al generar datos que se asemejen a los reales, se facilita la investigación de nuevos fenómenos y se permite probar y verificar diversas hipótesis sin depender exclusivamente de los datos reales adquiridos por experimentos costosos y largos.

El objetivo último de esta investigación es proporcionar una herramienta valiosa para la comunidad científica en el campo de la física de partículas, permitiendo la generación eficiente de datos de eventos de colisión. Al lograr una generación más rápida y precisa de estos eventos, podremos acelerar la exploración de nuevas teorías y fenómenos, así como avanzar en nuestra comprensión de la naturaleza fundamental de la materia y las fuerzas que la gobiernan.

4.1 Redes Generativas Antagónicas

Las GANs [31] desempeñan un papel esencial como una técnica de modelado generativo profundo. Una GAN es un tipo de arquitectura de redes neuronales que se compone de dos partes principales: el generador y el discriminador.

- El **generador**, G , es responsable de crear datos sintéticos, como imágenes o secuencias, que se asemejen a los datos reales de un conjunto determinado. A través del entrenamiento, el generador aprende a mapear un espacio latente, generado arbitrariamente, en un espacio de datos, capturando las características y patrones subyacentes presentes en los datos reales.
- El **discriminador**, D , actúa como un “juez clasificador” que evalúa si un dato dado es real o generado por el generador. Está entrenado para distinguir entre datos auténticos y datos sintéticos.

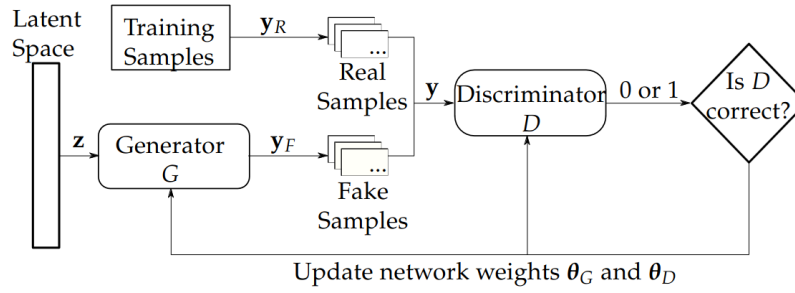


Figura 4.1: Diagrama de entrenamiento de una GAN.

El proceso de entrenamiento de una GAN es similar a un juego entre el generador y el discriminador. El generador busca mejorar su capacidad para engañar al discriminador al generar datos que se asemejen tanto como sea posible a los datos reales. Por otro lado, el discriminador busca mejorar su habilidad para discernir entre datos genuinos y generados. Esta competencia entre el generador y el discriminador impulsa la mejora continua de ambos componentes.

La función de pérdida juega un papel crucial en una GAN. Utiliza una función antagónica que guía al generador hacia la producción de datos cada vez más realistas, al mismo tiempo que el discriminador se vuelve más experto en distinguir entre lo real y lo sintético. Esta lucha constante entre el generador y el discriminador lleva a la convergencia de la GAN hacia un punto en el que el generador es capaz de producir datos que son indistinguibles de los datos reales.

La primera parte de la ecuación busca maximizar la probabilidad de que el discriminador clasifique correctamente los datos reales como reales. La segunda parte busca maximizar la probabilidad de que el discriminador clasifique correctamente los datos generados por el generador como falsos. En conjunto, la ecuación de pérdida representa la competencia entre el generador y el discriminador, donde el generador intenta minimizar esta función mientras que el discriminador intenta maximizarla. Las funciones de pérdida del sistema son las siguientes. Función de pérdida del **generador**:

$$\mathcal{L}_G = -\log(D(G(z))) \quad (4.1)$$

Función de pérdida del **discriminador**:

$$\mathcal{L}_D = -\log(D(x)) - \log(1 - D(G(z))) \quad (4.2)$$

Donde \mathcal{L}_G es la función de pérdida del generador, \mathcal{L}_D es la función de pérdida del discriminador, G es el generador, D es el discriminador, x representa una muestra de datos reales, y z es una muestra del espacio latente generada de manera aleatoria.

4.1.1. WGAN

Una *Wasserstein Generative Adversarial Network* (WGAN) [7] es una variante de las redes generativas antagónicas que introduce el concepto de distancia de Wasserstein o distancia del transportador terrestre, *Earth-Mover Distance* (EMD) en inglés, en lugar de utilizar la función de pérdida de entropía cruzada utilizada en las GANs convencionales. Esta modificación tiene como objetivo superar algunos de los problemas asociados con las GANs tradicionales, como la inestabilidad en el entrenamiento y la dificultad para medir la calidad de las muestras generadas. Además se elimina la figura del discriminador y se sustituye por la del crítico, ahora la red que realiza el papel de “juez o crítico” no devuelve una probabilidad sino una puntuación que indica como de real o falso es un dato. En una WGAN, en lugar de medir la probabilidad de que el discriminador clasifique

correctamente entre datos reales y generados, se calcula la distancia de Wasserstein entre las distribuciones de datos reales y generados, se encuentra más detallada en el [Capítulo 6](#). En el contexto de las redes neuronales, esta distancia se aproxima al valor promedio de las diferencias entre las salidas del crítico para las muestras reales y generadas.

La clave de la WGAN es la introducción de una restricción Lipschitz¹ en el crítico, es por ello que se introduce un método llamado *clipping* de pesos². Esta restricción asegura que el gradiente del crítico no crezca demasiado, lo que a su vez ayuda a estabilizar el proceso de entrenamiento y evita el colapso del gradiente como se describe en [7]. La función de pérdida de la WGAN se basa en el concepto de distancia de Wasserstein, y el generador busca minimizar esta distancia, mientras que el crítico busca maximizarla. Este proceso aborda algunos problemas de las GANs tradicionales al enfocarse en la distancia de Wasserstein y aplicar restricciones al crítico para estabilizar el entrenamiento. Esto permite obtener muestras generadas de mayor calidad y una mejor convergencia en el proceso de entrenamiento.

Función de pérdida del **generador**:

$$\mathcal{L}_G = -\mathbb{E}_{z \sim p_z}[C(G(z))] \quad (4.3)$$

Función de pérdida del **crítico**:

$$\mathcal{L}_C = \mathbb{E}_{x \sim p_{data}}[C(x)] - \mathbb{E}_{z \sim p_z}[C(G(z))] \quad (4.4)$$

Donde \mathcal{L}_G es la función de pérdida del generador, \mathcal{L}_C es la función de pérdida del crítico, G es el generador, C es el crítico, x representa una muestra de datos reales, z es una muestra del espacio latente generada de manera aleatoria, y p_z es la distribución de probabilidad del espacio latente y p_{data} es la distribución de probabilidad de las muestras reales. De tal forma que: $\mathbb{E}_{x \sim p_{data}}[C(x)]$ es la esperanza matemática (o valor esperado) de la salida del crítico C cuando se le alimenta con datos reales x provenientes de la distribución real $p_{data}(x)$; y $\mathbb{E}_{z \sim p_z}[C(G(z))]$ es la esperanza matemática de la salida del crítico C cuando se le alimenta con datos generados $G(z)$ provenientes del generador G , el cual se alimenta con ruido z .

4.1.2. WGAN-GP

Una *Wasserstein Generative Adversarial Network with Gradient Penalty* (WGAN-GP) [8] es una extensión de la WGAN que aborda aún más la estabilidad y el rendimiento del entrenamiento en redes generativas antagónicas.

Una WGAN-GP va un paso más allá al introducir una penalización de gradiente en el crítico para garantizar que el crítico tenga una limitación en sus gradientes. Esta penalización de gradiente resuelve un problema clave en el entrenamiento de GANs, conocido como el colapso modal. El colapso modal ocurre cuando el generador produce una variedad limitada de muestras de alta calidad, pero no logra capturar toda la diversidad de la distribución real de datos.

La penalización de gradiente en una WGAN-GP se logra tomando los gradientes del crítico con respecto a los datos reales y generados, y luego aplicando una restricción a esos gradientes para que no superen un valor máximo (generalmente 1). Esto se logra mediante el uso de la norma L2 (euclidiana) y asegura que el crítico sea 1-Lipschitz, lo

¹La restricción de Lipschitz se refiere a la propiedad de que la función del crítico debe ser Lipschitz continua con una constante de Lipschitz K .

²El clipping de pesos es una técnica en la que los parámetros del crítico se restringen a un pequeño rango fijo, por ejemplo, para asegurar que la función C sea Lipschitz continua

que significa que sus gradientes no pueden ser mayores que 1 en ningún punto. Esta penalización suave del gradiente garantiza una optimización más estable y evita problemas de divergencia o inestabilidad durante el entrenamiento.

Función de pérdida del **generador**:

$$\mathcal{L}_G = -\mathbb{E}_{z \sim p_z}[D(G(z))] \quad (4.5)$$

Función de pérdida del **crítico**:

$$\mathcal{L}_C = \mathbb{E}_{x \sim p_{data}}[C(x)] - \mathbb{E}_{z \sim p_z}[C(G(z))] + \lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}}[(\|\nabla_{\hat{x}} C(\hat{x})\|_2 - 1)^2] \quad (4.6)$$

Donde \mathcal{L}_G es la función de pérdida del generador, \mathcal{L}_C es la función de pérdida del crítico, G es el generador, C es el crítico, x representa una muestra de datos reales, z es una muestra del espacio latente generada de manera aleatoria, y p_z es la distribución de probabilidad del espacio latente generado de manera aleatoria, p_{data} es la distribución de probabilidad de las muestras reales, \hat{x} es una interpolación entre datos reales y generados y λ es un factor de ponderación para la penalización de gradiente. De tal forma que $\|\nabla_{\hat{x}} C(\hat{x})\|_2$ es la norma del gradiente de la salida del crítico con respecto a \hat{x} , y por tanto, $\mathbb{E}_{\hat{x} \sim p_{\hat{x}}}[(\|\nabla_{\hat{x}} C(\hat{x})\|_2 - 1)^2]$ es el término de penalización del gradiente. La norma del gradiente de C con respecto a \hat{x} se fuerza a ser cercana a 1. La interpolación de \hat{x} garantiza que se consideren puntos entre los datos reales y generados, proporcionando una evaluación más robusta de la continuidad de C .

4.2 Resumen del capítulo

Durante este capítulo hemos explicado como las Redes Generativas Antagónicas (GANs) han demostrado ser una herramienta poderosa en la generación de datos sintéticos realistas, especialmente en el contexto de la física de partículas para la simulación de eventos de colisión de partículas protón-protón. La elección de GANs se fundamenta en su capacidad para capturar las complejidades de las distribuciones de datos, lo que permite generar muestras que se asemejen a los datos reales adquiridos por experimentos costosos y largos.

También hemos explicado como durante el entrenamiento de una GAN, el generador aprende a mapear un espacio latente en datos que el discriminador intenta distinguir entre reales y generados. Esta competencia impulsa la mejora continua de ambos componentes hasta lograr la generación de datos prácticamente indistinguibles de los reales. La función de pérdida juega un papel crucial, donde el generador busca minimizar la probabilidad de que el discriminador clasifique incorrectamente los datos generados como falsos.

Además, las variantes como la Wasserstein GAN (WGAN) y la WGAN con penalización de gradiente (WGAN-GP) introducen mejoras significativas. La WGAN incorpora la distancia de Wasserstein para mejorar la estabilidad del entrenamiento, mientras que la WGAN-GP añade una penalización de gradiente para controlar la capacidad del discriminador, mejorando la calidad y la diversidad de las muestras generadas.

En resumen, las GANs y sus variantes representan una contribución significativa al campo de la física de partículas al permitir una generación eficiente de datos de eventos de colisión. Esta tecnología acelera la investigación al facilitar la exploración de nuevas teorías y fenómenos, fortaleciendo nuestra comprensión de la naturaleza fundamental de la materia y las fuerzas que la gobiernan.

CAPÍTULO 5

Hardware y software usado

Históricamente, la implementación de algoritmos de aprendizaje automático y estructuras de modelos exigía su creación desde cero, demandando la especificación meticulosa de cálculos intrincados esenciales durante tanto las fases de entrenamiento como de inferencia. No obstante, la investigación continua y los esfuerzos colaborativos dentro de la comunidad científica han fomentado avances notables con el paso del tiempo. En la actualidad, disponemos de una plétora de bibliotecas y herramientas que están a nuestra disposición, simplificando el desarrollo y la evaluación de modelos a través de interfaces de programación de alto nivel y fácil manejo.

Estos marcos instrumentales han propiciado un cambio de paradigma al abstraer las complejidades de la implementación de bajo nivel. Nos capacitan para centrar nuestra atención en conceptos fundamentales e innovaciones dentro de nuestros modelos, liberándonos de los pormenores de la definición de operaciones rudimentarias.

A través de la utilización de estas herramientas, el proceso de construcción y prueba de diversas arquitecturas de modelos se ha vuelto más sencillo. En el ámbito de esta investigación, se empleó un espectro de herramientas de hardware y software, facilitando la construcción y el entrenamiento de los modelos diseñados, todo ello sin requerir una inmersión en las profundas sutilezas de la implementación de bajo nivel.

5.1 Hardware

El proceso de entrenamiento y prueba de un modelo de IA requiere llevar a cabo una cantidad considerable de cálculos de tipo matricial, y usualmente demanda repetir experimentos mientras se exploran diferentes configuraciones. En lo que respecta al hardware, las GPUs se erigen como los dispositivos más óptimos para llevar a cabo los cálculos necesarios, logrando un rendimiento superior en comparación con las CPUs.

A raíz de esta consideración, la infraestructura de hardware asignada a este proyecto abarcó un servidor dotado de 32 GB de memoria RAM DDR4 y una potente tarjeta gráfica NVIDIA GeForce RTX 3090 TI. Esta tarjeta ostenta el chip GA102, repleto de componentes diseñados para tareas de aprendizaje automático, acompañados de 10,752 núcleos CUDA y 24GB GDDR6X [32].

5.2 Software

Dentro del ámbito de nuestra investigación, optamos por un entorno de software basado en Linux, el cual demostró ser una base confiable y eficiente para nuestra experi-

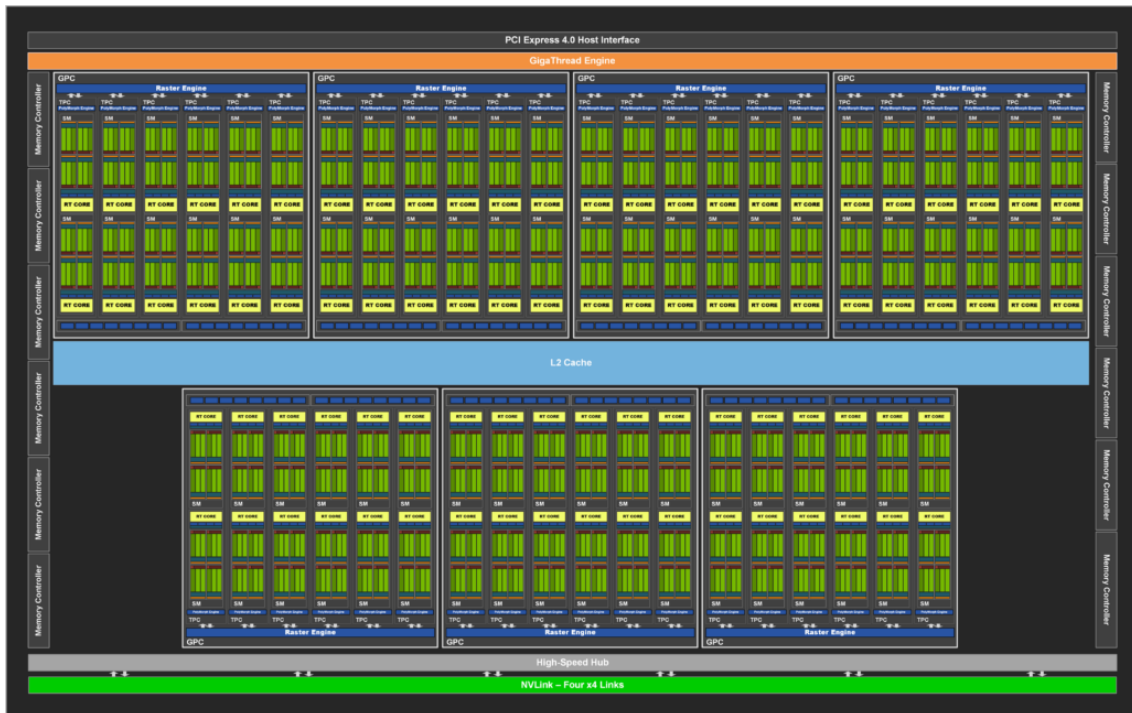


Figura 5.1: Diagrama de la arquitectura del chip GA102 [32].

mentación. Elegimos Python como nuestro lenguaje principal de programación de alto nivel. Esta elección se originó debido a la extensa gama de bibliotecas relacionadas con el aprendizaje automático disponibles, que facilitan la rápida formulación de arquitecturas complejas a través de una sintaxis comprensible.

5.2.1. Análisis de datos y representación

Con el propósito de evaluar la calidad de los resultados obtenidos y valorar su alineación con los objetivos deseados, se vuelve imperativo forjar una representación visualmente comprensible y de fácil interpretación de los datos. Para cumplir con este requisito, se empleó una gama de bibliotecas y herramientas de software. Concretamente, recurrimos a Matplotlib, Scikit-learn y Numpy para calcular y presentar estadísticas derivadas de nuestros datos generados.

Scikit-learn [33], también conocido como sklearn, es una biblioteca de aprendizaje automático ampliamente utilizada de código abierto para el lenguaje de programación Python. Está diseñada para ofrecer una interfaz completa y fácil de usar para diversas tareas de aprendizaje automático, incluyendo clasificación, regresión, agrupamiento, reducción de dimensionalidad, entre otras. Simplifica el proceso de aplicar técnicas y algoritmos de aprendizaje automático a datos del mundo real. Ofrece una API coherente y bien documentada que permite a investigadores y desarrolladores experimentar eficientemente con diferentes algoritmos y modelos sin tener que sumergirse en las implementaciones de bajo nivel.

Scikit-learn brinda un conjunto exhaustivo de métricas de distancia predefinidas que pueden integrarse sin inconvenientes con diversos algoritmos, además de incorporar herramientas para la división de los datos. Dichas características se han utilizado en el **Capítulo 6** y el **Capítulo 7**, respectivamente.

Numpy (Numerical Python) [34] es una biblioteca esencial en el ámbito de la programación y análisis numérico utilizando Python. Su enfoque principal es proporcionar

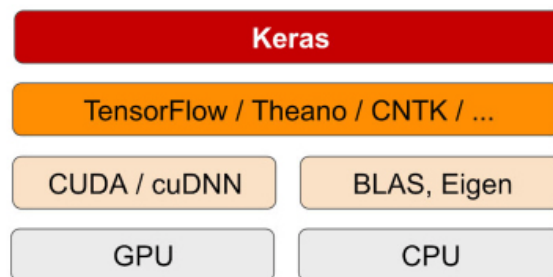


Figura 5.2: Esquema de la tecnología usada [36].

soporte para la creación y manipulación eficiente de matrices y arreglos multidimensionales, lo que la convierte en una herramienta fundamental para tareas relacionadas con matemáticas, estadísticas, ciencia de datos y cálculos científicos en general.

La característica distintiva de Numpy es su capacidad para trabajar con arreglos multidimensionales, conocidos como ndarrays (arrays n-dimensionales). Estos ndarrays permiten almacenar y operar sobre datos en forma de matrices o vectores con dimensiones variables. Esta estructura de datos altamente optimizada es esencial para la ejecución eficiente de cálculos numéricos y científicos en Python.

Una de las razones por las que Numpy es tan popular en el ámbito académico y científico es su capacidad para leer diversos formatos de archivos e integrarse sin problemas con otras bibliotecas en el ecosistema científico de Python, como Scipy, Pandas y Matplotlib. Esta integración facilita el flujo de trabajo en análisis de datos, modelado matemático y visualización.

Matplotlib [35] es una reconocida biblioteca de visualización de datos en Python. Ofrece una amplia gama de herramientas para crear diversos tipos de visualizaciones de datos. Su diseño está orientado a brindar flexibilidad y personalización.

Esta biblioteca permite crear una variedad de tipos de gráficos y utiliza un conjunto extenso de funciones y clases para controlar diferentes aspectos de los mismos. Esto incluye ejes, etiquetas, títulos, leyendas, colores, marcadores y líneas de cuadrícula, con el propósito de mejorar la representación visual de los datos. Matplotlib proporciona diferentes interfaces para crear gráficos, como pyplot, y la interfaz orientada a objetos, que ofrece mayor control y flexibilidad al manipular directamente los objetos de los gráficos.

Una de las características distintivas de Matplotlib es su capacidad para trabajar con varios formatos de salida, incluida la posibilidad de guardar los gráficos como archivos de imagen, como PNG, JPEG o SVG. Esta versatilidad permite a los usuarios seleccionar el formato más adecuado para sus necesidades y compartir sus visualizaciones de manera efectiva.

5.2.2. Diseño de los modelos

Hemos elegido TensorFlow y Keras como bibliotecas fundamentales para la formulación y el entrenamiento de nuestros modelos. Dotándonos de un marco robusto para desarrollar fácilmente todo el proceso de entrenamiento y generación.

Tensorflow [37], un marco de aprendizaje automático de código abierto desarrollado por Google, constituye el núcleo de esta selección. Presenta un amplio repertorio de herramientas, bibliotecas y recursos diseñados para la construcción y despliegue de modelos, atendiendo así a una amplia gama de aplicaciones. En esencia, los cálculos se representan como grafos computacionales: una disposición interconectada de nodos, cada

uno de los cuáles representa una función matemática distinta. Los datos se encapsulan como tensores, que se manifiestan como matrices multidimensionales.

Su arquitectura incluye una API de alto nivel que permite a los desarrolladores definir y entrenar sin problemas modelos de aprendizaje automático, dando cabida a un amplio espectro de arquitecturas de modelos. Una amplia colección de capas, optimizadores y funciones de pérdida simplifica la construcción de modelos con eficiencia. TensorFlow se ha convertido en uno de los marcos de trabajo más utilizados en el ámbito del aprendizaje automático y el aprendizaje profundo, impulsando innumerables aplicaciones en diversos campos.

Keras [38], un marco de aprendizaje profundo rico en interfaces situado sobre TensorFlow, es un complemento indispensable de nuestro conjunto de herramientas. De código abierto por naturaleza, Keras ha sido meticulosamente diseñado para ser intuitivo, modular y comprensible, lo que lo convierte en un favorito tanto de los neófitos como de los investigadores experimentados en aprendizaje profundo. Su intuitiva API sirve de escudo contra la complejidad de la construcción de redes neuronales, agilizando la creación de prototipos y la experimentación de diversas arquitecturas de red. Esto, a su vez, facilita la personalización y el refinamiento meticuloso de los modelos para adaptarlos a requisitos específicos.

Keras proporciona un amplio repertorio de capas fácilmente disponibles, lo que permite ensamblar sin esfuerzo intrincadas arquitecturas de redes neuronales. Además, permite la creación de redes con múltiples entradas o salidas de datos, lo que posibilita la orquestación de modelos sofisticados. El marco también incluye una amplia gama de optimizadores, funciones de activación, funciones de pérdida y métricas, junto con utilidades adaptadas al preprocesado de datos. Otra de sus virtudes es que admite paradigmas de programación tanto simbólicos como imperativos, lo que proporciona la adaptabilidad y la integración sin fisuras que caracterizan el panorama de las bibliotecas de Python.

5.3 Resumen del capítulo

En resumen, el uso efectivo de estas herramientas no solo simplifica el desarrollo y la evaluación de modelos de aprendizaje automático, sino que también permite a los investigadores enfocarse en la innovación y la mejora continua de las técnicas utilizadas para generar eventos de colisiones de partículas mediante Aprendizaje Automático.

CAPÍTULO 6

Métricas de evaluación

Al abordar cualquier tarea de aprendizaje automático, resulta esencial contar con herramientas que permitan evaluar el rendimiento de cada enfoque propuesto. Esta evaluación no solo brinda la capacidad de determinar la idoneidad de una propuesta para la tarea en cuestión, sino que también posibilita la comparación de múltiples enfoques para seleccionar el que mejor se ajuste al objetivo deseado.

Debido a ello proponemos la utilización de diversos indicadores numéricos, que nos permiten efectuar una comparación más meticulosa entre nuestros modelos. Además, incluimos métricas interesantes para nuestros objetivos, tales como la velocidad de generación de eventos o la cantidad de eventos descartados (inválidos) durante el proceso de generación. Cabe destacar que la propuesta inicial de estas métricas forma parte de la colaboración con nuestro compañero Raúl Balanzá aplicadas en su TFM [39], siendo este la continuación de [23], y han sido también adoptadas por nuestro compañero Joan Boix Avalos para su TFG [40]. Esto nos proporciona un marco de trabajo que habilita la capacidad de comparar resultados entre trabajos.

6.1 Métricas de similitud

Definimos algunas métricas para permitir una comparación precisa de los resultados en la fase de generación con la distribución real de los datos. Calculamos todas las métricas presentadas para cada gráfico, obteniendo diversas medidas que evalúan cómo se ajusta la distribución de datos generados en comparación con la distribución real. Para obtener el valor de las métricas para la totalidad del modelo calculamos un promedio ponderado cuyos pesos son la cantidad de partículas de cada tipo presentes en los datos reales. Definiéndose de la siguiente forma:

$$M = \frac{1}{N_{part}} \sum_i \frac{N(i)_{part}}{N_{part}} \cdot M(i)_{part} \quad (6.1)$$

En esta ecuación, M representa el valor general de la métrica, N_{part} es el número total de partículas de la verdad objetiva consideradas, $N(i)_{part}$ es la cantidad de partículas del tipo i en el conjunto de datos, y $M(i)_{part}$ es el valor de la métrica M para ese tipo de partícula.

En este contexto, a MET y MET ϕ se les asigna un peso igual al número de eventos en el conjunto de datos, ya que están presentes de manera constante. De esta manera, los modelos pueden ser comparados entre sí de manera global.

Divergencia Kullback-Leibler

La divergencia de Kullback-Leibler [41], también conocida como KL-divergencia o entropía relativa, es una medida utilizada para cuantificar la diferencia entre dos distribuciones de probabilidad. Esta medida se enfoca en medir cuánto difiere una distribución de probabilidad P de otra distribución Q , teniendo en cuenta la información adicional requerida para codificar los datos generados por P utilizando una codificación óptima para Q .

La divergencia de Kullback-Leibler se define como:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (6.2)$$

Donde $P(i)$ y $Q(i)$ son las probabilidades asociadas al evento i en las distribuciones P y Q , respectivamente.

La divergencia de Kullback-Leibler es útil en una variedad de aplicaciones, como la teoría de la información, la estimación de parámetros en modelos estadísticos, el aprendizaje automático y la evaluación de modelos generativos, como las GANs. En el contexto de GANs, la divergencia KL puede utilizarse para medir la discrepancia entre la distribución de las muestras generadas y la distribución de los datos reales.

Cuando esta divergencia es cero, indica que las dos distribuciones son idénticas. Sin embargo, la divergencia de Kullback-Leibler no es simétrica; es decir, $D_{KL}(P \parallel Q)$ puede ser diferente de $D_{KL}(Q \parallel P)$.

Jensen-Shannon distance

La distancia Jensen-Shannon [42] es una medida utilizada para cuantificar la similitud entre dos distribuciones de probabilidad. Es una medida simétrica que se deriva de la divergencia de Kullback-Leibler (KL-divergencia) y se utiliza para medir la distancia entre dos distribuciones de probabilidad, teniendo en cuenta tanto sus similitudes como sus diferencias.

La ecuación de la distancia Jensen-Shannon se define en función de la divergencia de Kullback-Leibler y la entropía de las distribuciones P y Q de la siguiente manera:

$$JSD(P \parallel Q) = \frac{1}{2} (D_{KL}(P \parallel M) + D_{KL}(Q \parallel M)) \quad (6.3)$$

Donde $M = \frac{1}{2}(P + Q)$ es la distribución promedio entre P y Q , y D_{KL} es la divergencia de Kullback-Leibler.

Distancia de Wasserstein (Earth-Mover)

La distancia de Wasserstein [43], [44], también conocida como distancia del transportista de tierra, o *Earth-Mover Distance* (EMD) en inglés, es una métrica utilizada en teoría de probabilidades y estadística para medir la diferencia entre dos distribuciones de probabilidad. A diferencia de algunas métricas de distancia que se centran en propiedades puntuales, la distancia de Wasserstein se enfoca en cuánto trabajo es necesario para transformar una distribución en otra.

La distancia de Wasserstein se considera útil en situaciones donde se desea comparar dos distribuciones en función de la cantidad de cambio requerida para convertir una en la otra. En el contexto de GANs, donde se generan muestras sintéticas que se asemejan a datos reales, la distancia de Wasserstein puede proporcionar una medida cuantitativa de la similitud entre las distribuciones de las muestras generadas y las muestras reales. Esto es especialmente relevante cuando se trata de evaluar la calidad de generación de GANs, como hemos comentado anteriormente.

La distancia de Wasserstein se define como:

$$W(p, q) = \inf_{\gamma \in \Gamma(p, q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\| d\gamma(x, y) \quad (6.4)$$

Donde p y q son las distribuciones de probabilidad que se comparan, y $\Gamma(p, q)$ es el conjunto de todas las distribuciones conjuntas γ con marginales p y q . Cuando esta distancia es cero, indica que las dos distribuciones son idénticas.

6.2 Métricas de generación

Definimos dos métricas adicionales que ponen de manifiesto la eficiencia y calidad de nuestros eventos generados al utilizar los distintos modelos.

Calidad de los eventos

Para definir una métrica de calidad, observamos las condiciones del conjunto de datos para que un evento sea considerado válido, tal como se mencionó en la Sección 3.1.

En este caso, la métrica es bastante directa y describe el porcentaje de eventos que no cumplen con las condiciones mencionadas. Esta métrica, que denotamos como Tasa de Eventos Inválidos (TEI), o *Invalid Event Rate (IER)*, se puede definir de la siguiente manera:

$$TEI = \frac{N_{invalidos}}{N_{total}} \times 100 \quad (6.5)$$

Donde $N_{invalidos}$ es la cantidad de eventos que no cumplen con las condiciones para considerarse válidos, y N_{total} es la cantidad total de eventos generados.

Es importante destacar que cumplir con las condiciones de este conjunto de datos no garantiza que el evento sea físicamente correcto y viceversa.

Eficiencia de generación

Para evaluar la eficiencia del proceso de generación de eventos, hemos decidido obtener el tiempo necesario para generar un evento. Debido a que los eventos tienen formas variables en términos del número de objetos, diferentes eventos pueden requerir diferentes cantidades de tiempo para generarse.

Para obtener una métrica confiable, decidimos generar 1,000 eventos con cada modelo y luego calcular el tiempo promedio de generación entre esa cantidad. Con esta estrategia, nos aseguramos de que se generen eventos de diferentes longitudes y, por lo tanto, este valor puede utilizarse para comparar diferentes modelos. La métrica se define formalmente como:

$$T_{Generacion} = \frac{T_{1K}}{1000} \quad (6.6)$$

Donde T_{1K} es la cantidad de tiempo necesario para generar 1,000 eventos. Es importante tener en cuenta que la razón de incluir esta métrica en nuestra comparación radica en que los diferentes enfoques para generar eventos pueden presentar diferencias significativas en el tiempo de generación, dependiendo de la posibilidad de beneficiarse de las capacidades de generación en lotes de la GPU.

6.3 Resumen del capítulo

Concluyendo, hemos definido y calculado diversas métricas para evaluar la calidad y similitud de los eventos generados por modelos de aprendizaje automático con respecto a los datos reales. Estas medidas permiten una evaluación precisa de cómo se ajustan los datos generados a la distribución real, proporcionando así una comparación global entre modelos.

Estas métricas no solo son aplicables en la evaluación de modelos generativos como las GANs, sino que también son cruciales para la validación de eventos generados en nuestro contexto de estudio.

Para complementar estas métricas de similitud, definimos métricas adicionales para evaluar la calidad de los eventos generados. La Tasa de Eventos Inválidos (TEI) proporciona un porcentaje de eventos que no cumplen con las condiciones predefinidas; y la Eficiencia de Generación calcula el tiempo promedio necesario para generar eventos, lo que permite comparar la eficiencia de diferentes modelos en términos de generación de eventos complejos.

En resumen, estas métricas proporcionan un marco robusto para la evaluación cuantitativa y cualitativa de modelos de generación de eventos, facilitando así la mejora continua y la optimización de técnicas de aprendizaje automático aplicadas a nuestro problema específico.

CAPÍTULO 7

Experimentación

En esta sección, discutiremos el desarrollo y experimentación llevados a cabo basados en las soluciones propuestas en el [Capítulo 4](#) entorno a los datos discutidos en el [Capítulo 3](#). Utilizamos las tecnologías descritas en el [Capítulo 5](#) y analizamos los resultados de cada experimento mediante las métricas explicadas en el [Capítulo 6](#). Durante nuestra investigación para determinar el modelo óptimo para la generación de eventos, optamos por concentrarnos en *t \bar{t}* , un proceso específico del Modelo estándar, debido a la gran cantidad de datos disponibles. Posteriormente, aplicaremos este modelo a otro tipo de eventos dentro del BSM. En los procesos del tipo *t \bar{t}* , definido en la [Tabla 3.2](#), la colisión de protones genera dos quarks top que se desintegran en otras partículas.

7.1 Experimento con *t \bar{t}* del SM

Este primer experimento se focaliza en la utilización de múltiples modelos centrados en diferentes partes de cada evento, con el propósito de que cada modelo pueda capturar fielmente la distribución de una sección específica. De esta manera, se busca generar eventos de mayor calidad al combinar las representaciones aprendidas por cada modelo. Cabe mencionar que las partes del evento se generan de forma secuencial, de forma que son necesarias las partes anteriores, esto será explicado en más detalle en la parte de preprocesado y generación de eventos.

7.1.1. Preprocesado

En la etapa inicial de la metodología, se realiza un proceso de filtrado para seleccionar las combinaciones de partículas más comunes, hasta alcanzar una representación del 95 % de los datos. A partir de esta selección, se procede a dividir el conjunto de datos resultante en dos componentes esenciales: el conjunto de entrenamiento y el conjunto de prueba. El primero se destina a la fase de entrenamiento de los modelos, lo que les permite aprender los patrones y las relaciones presentes en los datos. Por otro lado, el conjunto de prueba se reserva para la evaluación de los modelos, utilizándolo para generar gráficos y calcular métricas de calidad de los modelos.

Cabe mencionar que tanto el conjunto de entrenamiento como el de prueba han sido diseñados de manera que reflejen la distribución de las combinaciones de partículas que se encuentran en el conjunto de datos filtrado inicialmente.

En el interior del conjunto de entrenamiento, se lleva a cabo una subdivisión adicional en varios archivos, según el número de modelos y las partes específicas de los eventos en los que se centran. Esta subdivisión desempeña un papel crucial para asegurar un proce-

so de entrenamiento organizado y eficiente. Cada archivo se enfoca en partes particulares de los eventos, lo que permite a los modelos capturar de manera óptima las características únicas de cada sección. En resumen, el proceso resultante se puede resumir de la siguiente manera, tal y como aparece en la Figura 7.1:

1. Se crea un archivo que alberga detalles cruciales como MET (Momento Transversal) y $MET\phi$ (Ángulo Azimutal Transversal) de un evento, junto con una máscara binaria ordinal de partículas. Esta máscara binaria ordinal, compuesta por 17 dígitos, permite representar todas las combinaciones de partículas presentes en el conjunto de datos filtrado. La máscara presenta la siguiente cantidad de cifras dedicadas a cada partícula en el mismo orden que en la Tabla 3.1¹: 3, b; 9, j; 1, e-; 1, e+; 1, m-; 1, m+; 1, g. Para esclarecer, la máscara {000 111000000 0 1 0 1 0} sería equivalente a 3 jets (j), 1 positrón (e+) y un antimuón (m+).
2. Se crea un archivo separado para la primera partícula de cada tipo, el cual proporciona información detallada sobre dicha partícula en el contexto del evento. Cada línea en estos archivos contiene el MET y $MET\phi$ del evento en el que se encontró la partícula, seguido por la suma de las cuatro características de las partículas previas² y, posteriormente, las cuatro características específicas de la partícula (Energía, Momento Transversal, η y ϕ). Es importante destacar que el orden en el que se calcula la suma de las características de las partículas anteriores sigue la disposición indicada en la Tabla 3.1. Un aspecto relevante a considerar es que, debido al orden de generación, al generar el primer b-jet, no existen partículas previas para ser sumadas. Para brindar claridad, consideremos el ejemplo del archivo correspondiente a los electrones. Este archivo incluirá el MET y $MET\phi$ del evento, la suma de las características de todas las partículas previas, en este caso los b-jets y jets presentes en el evento, y finalmente, las propias características distintivas de la partícula electrón. Esta estructuración organizada asegura la representación integral y organizada de los datos, lo cual es esencial para el éxito de los experimentos y análisis subsiguientes.
3. En la situación particular de los jets y los b-jets, es posible que en un evento se generen más de una partícula de estos tipos. Por este motivo, hemos creado dos archivos adicionales para almacenar la información de las partículas que vienen después de las primeras. Cada línea en estos archivos contiene el MET y $MET\phi$ del evento en el que se encontró la partícula, seguido de la suma de las características de las partículas previas, las cuatro características específicas de la partícula previa y, finalmente, las características específicas de la partícula actual.

Esta separación de datos da lugar a un total de 10 archivos en conjunto. Esto incluye 1 archivo dedicado a los datos del evento, 2 archivos para los datos de los b-jets, otros 2 archivos para los datos de los jets, y un archivo individual para cada una de las partículas restantes.

Además, es relevante destacar que las características de las partículas fueron estandarizadas mediante un proceso de escalado Min-Max. Este procedimiento implicó transformar las características a un rango $[-1, 1]$, preservando su distribución original. Este escalado se realizó antes del entrenamiento de los modelos y se revierte después de la generación de eventos.

¹El orden establecido en las partículas es escogido arbitrariamente en la tabla y carece de explicación física, las partículas no se manifiestan físicamente en los eventos en un orden específico

²Debemos recordar que el orden establecido en las partículas es escogido arbitrariamente en la tabla y carece de explicación física

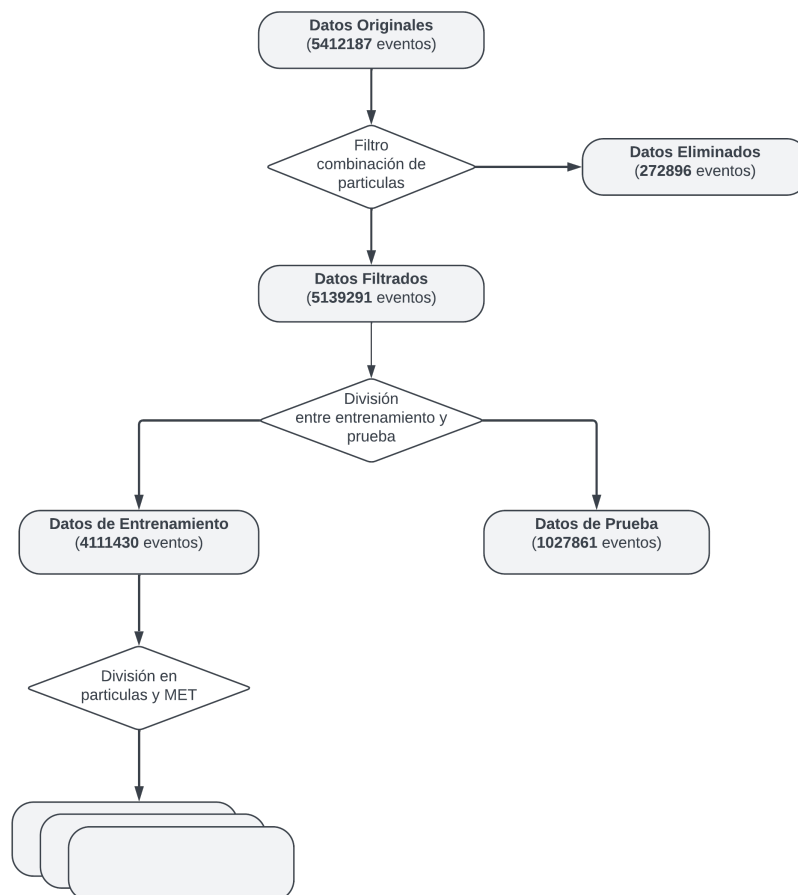


Figura 7.1: Diagrama del preprocesado del experimento.

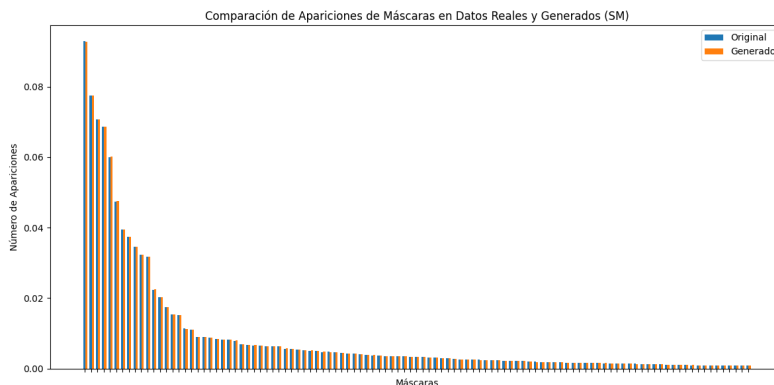


Figura 7.2: Comparación máscaras generadas y máscaras presentes en el conjunto de datos filtrado (SM).

7.1.2. Generador de máscaras

Durante la generación de datos, la definición de las máscaras que determinan la cantidad de partículas de cada tipo se llevará a cabo mediante un generador estadístico. Este generador se basa en la proporción de aparición de estas máscaras en el conjunto de datos filtrados. Es decir, se observa cuántas veces aparece cada combinación de máscaras en el conjunto de datos original y se utiliza esta información para generar las máscaras correspondientes en la fase de generación. En estadística este método se conoce como **alias** [45]. Podemos ver la comparación de las máscaras generadas y las presentes en el conjunto de datos filtrados en la Figura 7.2

7.1.3. Arquitectura del modelo

Los modelos utilizados en este experimento se basan en la arquitectura WGAN, que se describe en el Capítulo 4. En total, se entrenan 10 modelos, correspondientes a cada archivo, y comparten una estructura muy similar entre sí.

En la Figura 7.3, podemos observar que el modelo referente al MET y $MET\phi$ se compone de los siguientes elementos:

- El **Generador** recibe un vector del espacio latente generado arbitrario y la máscara binaria ordinal mencionada en la sección 7.1.1, y su función es producir los valores de MET y $MET\phi$. Para lograrlo, se emplean 7 capas densas en su estructura. Cada capa contiene 1024 neuronas, excepto la última capa que consta de solo 2 neuronas. Luego de cada capa densa, se aplica una capa de activación LeakyReLU con un valor α de 0.2, excepto en la última capa donde se utiliza una función tangente hiperbólica.
- El **Crítico** recibe un vector que contiene los valores de MET y $MET\phi$ del evento, ya sea real o generado, junto con la máscara ordinal binaria mencionada anteriormente. Su función es proporcionar una puntuación que se usa para calcular tanto la función de pérdida del crítico como la del generador. La estructura del crítico consiste en 7 capas densas, cada una con 1024 neuronas, a excepción de la última que tiene solo una neurona. Después de cada capa densa, se aplica una capa de activación LeakyReLU con un valor α de 0.2, excepto en la última capa donde no se utiliza ninguna función de activación.

En la Figura 7.4, podemos observar que la arquitectura usada en los modelos de las partículas. Algunos cambios respecto al modelo anterior y hechos importantes son:

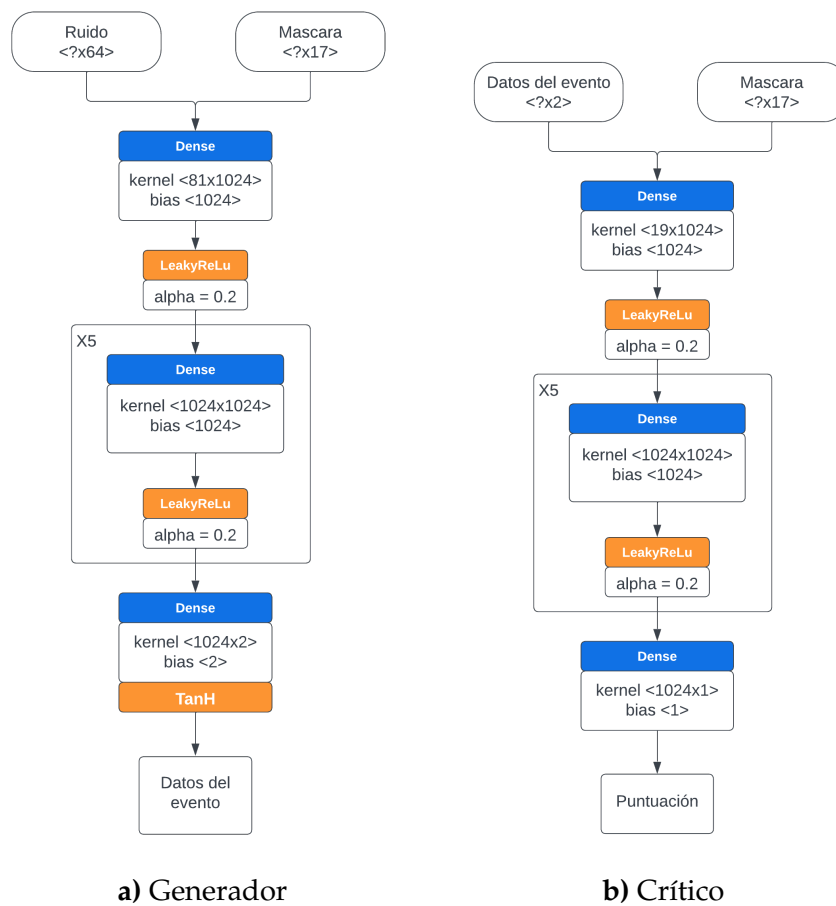


Figura 7.3: Diagrama de la arquitectura GAN para el MET y MET ϕ en el experimento.

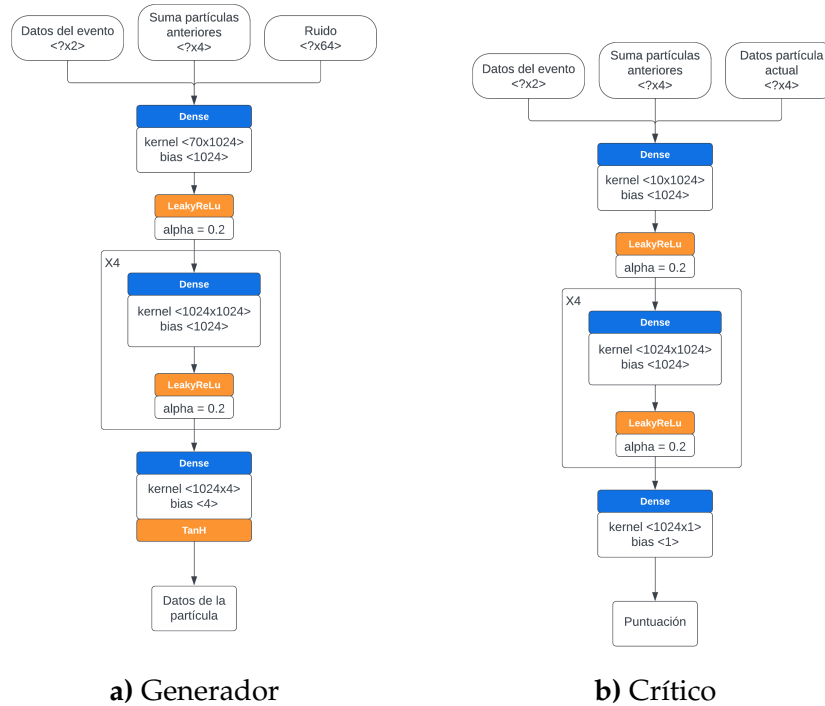


Figura 7.4: Diagrama de la arquitectura GAN para las primeras partículas en el experimento.

- Para los modelos de partículas, se han introducido ajustes en la arquitectura tanto del crítico como del generador. En el crítico, se ha considerado que reciba MET y $MET\phi$ del evento, la suma de los valores de las partículas anteriores y el valor de la partícula actual. Además, después de cada capa densa del crítico, se ha implementado una capa de activación LeakyReLU con un α de 0.2, excepto en la última capa que es lineal. Por su parte, el generador de estos modelos también recibe MET y $MET\phi$ del evento, la suma de los valores de las partículas anteriores y un vector generado aleatoriamente. Tras cada capa de activación en el generador, se ha incorporado una LeakyReLU con un α de 0.2, excepto en la última capa donde se utiliza una función tangente hiperbólica.
- Es importante destacar que en el caso del primer b-jet, el crítico no recibe la suma de los valores de las partículas anteriores debido a la ausencia de partículas previas. Del mismo modo, el generador del primer b-jet tampoco incorpora la suma de valores de partículas anteriores. Sin embargo, para los b-jets y jets que siguen al primero, las características del b-jet o jet inmediatamente anterior se consideran en tanto el crítico como el generador para una representación más precisa de la información secuencial.

Después de definir los distintos componentes, se procede a entrenar cada GAN utilizando el conjunto de datos *t \bar{t}* , con un tamaño de batch de 2048 eventos durante 160 epochs para el modelo del MET y $MET\phi$, y un tamaño de batch de 1024 eventos 320 epochs para cada modelo de partículas. Se emplea el optimizador Adam [46] para todos los modelos, con una tasa de aprendizaje de 0.000005, β_1 de 0.5 y β_2 de 0.9. Adicionalmente, el crítico o discriminador se actualiza 6 veces por cada vez que se actualiza el generador. En relación a los hiper-parámetros de las WGAN-GP, la penalización del gradiente es 10, en cuanto a los hiper-parámetros propios de las WGAN, la restricción aplicada a los pesos es de 0.01.

Una vez finalizado el proceso de entrenamiento, los generadores de máscaras y modelos se emplean para generar nuevos eventos mediante la decodificación de números generados aleatoriamente. La mitad de estos números se extraen de una distribución uniforme entre -1 y 1, mientras que la otra mitad se obtiene de una distribución normal con media cero y varianza unitaria. Esta combinación de distribuciones aleatorias ha demostrado mejorar el rendimiento en variables uniformes. Este proceso se ejecuta por evento, siguiendo el procedimiento secuencial aquí descrito:

1. Generar una máscara aleatoria utilizando el generador estadístico descrito en la Sección 7.1.2.
2. Generar un tensor aleatorio y proporcionarlo como entrada, junto con la máscara, al generador del modelo MET para generar un nuevo par de MET y $MET\phi$.
3. Si la máscara indica la presencia de b-jets, generar un tensor aleatorio y proporcionarlo como entrada, junto con el MET y $MET\phi$ del evento, al modelo del primer b-jet para generar las cuatro características de la partícula (Energía, Momento Transversal, η y ϕ).
4. Si la máscara indica la presencia de más b-jets, para cada b-jet adicional al primero genera de forma secuencial un tensor aleatorio y proporcionarlo como entrada, junto con el MET y $MET\phi$ del evento, los valores del b-jet anterior y la suma de las características de todas las partículas generadas hasta ese momento, al modelo de los b-jets posteriores al primero para generar las cuatro características de la partícula (Energía, Momento Transversal, η y ϕ).
5. Si la máscara indica la presencia de jets, generar un tensor aleatorio y proporcionarlo como entrada, junto con el MET y $MET\phi$ del evento y la suma de las características de todas las partículas generadas hasta ese momento, al modelo del primer jet para generar las cuatro características de la partícula (Energía, Momento Transversal, η y ϕ).
6. Si la máscara indica la presencia de más de un jet, para cada jet generar de forma secuencial un tensor aleatorio y proporcionarlo como entrada, junto con el MET y $MET\phi$ del evento, los valores del jet anterior y la suma de las características de todas las partículas generadas hasta ese momento, al modelo de los jets posteriores al primero para generar las cuatro características de la partícula (Energía, Momento Transversal, η y ϕ).
7. Para las partículas restantes presentes en la máscara (electrón, positrón, muon, antimuón, fotón), generar un tensor aleatorio y proporcionarlo como entrada, junto con el MET y $MET\phi$ del evento y la suma de las características de todas las partículas generadas hasta ese momento, al modelo correspondiente para generar las cuatro características de la partícula (Energía, Momento Transversal, η y ϕ).

A modo de ejemplo, podemos ver el proceso descrito en la Figura 7.5 para la máscara $\{110\ 111000000\ 0\ 1\ 0\ 0\ 0\}$, o lo que es lo mismo 2 b-jets, 3 jets y 1 positrón.

7.1.4. Resultados

En relación a las características de la energía faltante del evento (MET y $MET\phi$) presentes en la Figura 7.6, se observa una buena correspondencia entre los datos generados y las distribuciones originales en ambos modelos. Aunque en términos generales se logra una adaptación satisfactoria, es perceptible un decremento e incremento en el borde

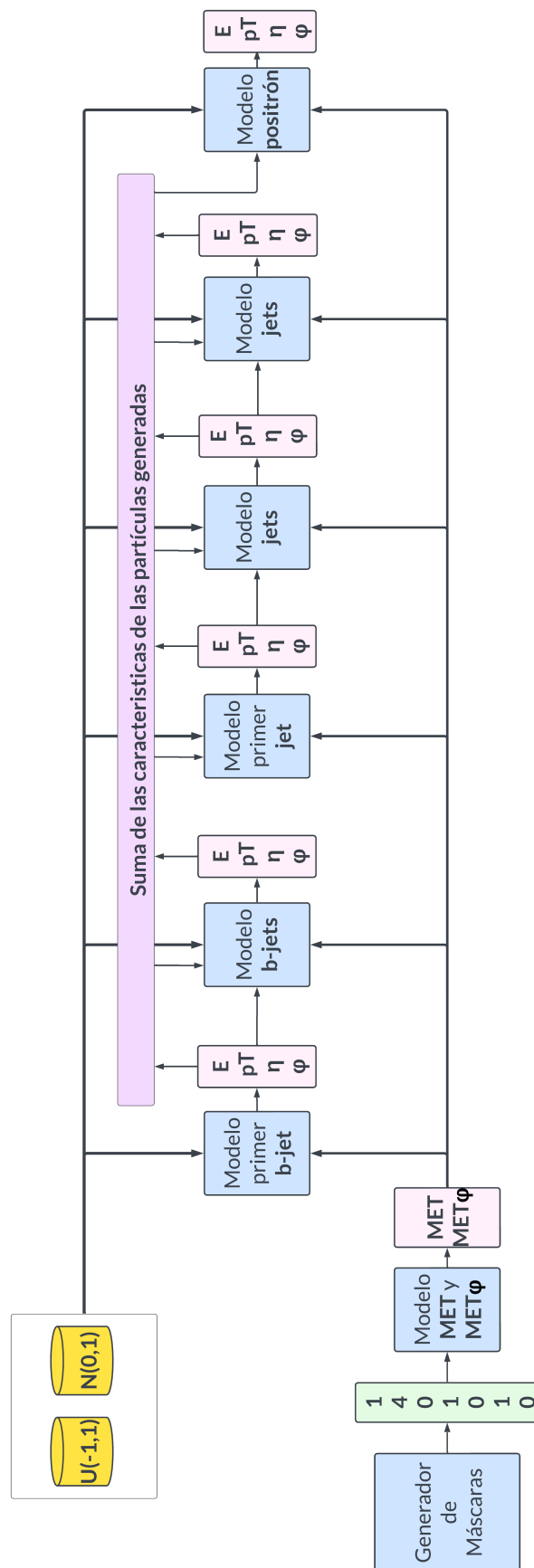


Figura 7.5: Diagrama del ejemplo del proceso de generación en el experimento.

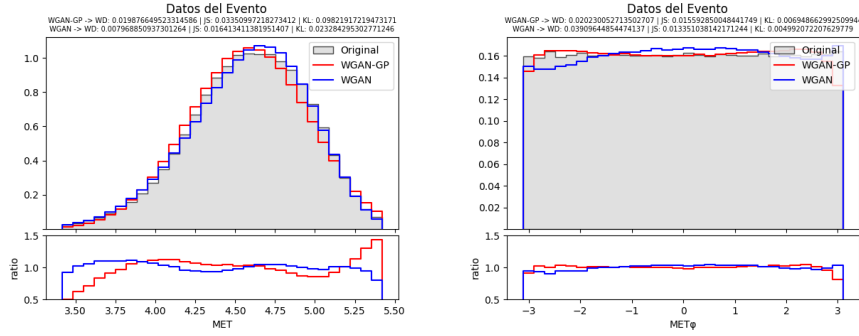


Figura 7.6: Histogramas de las características de la energía faltante del evento (MET, $MET\phi$).

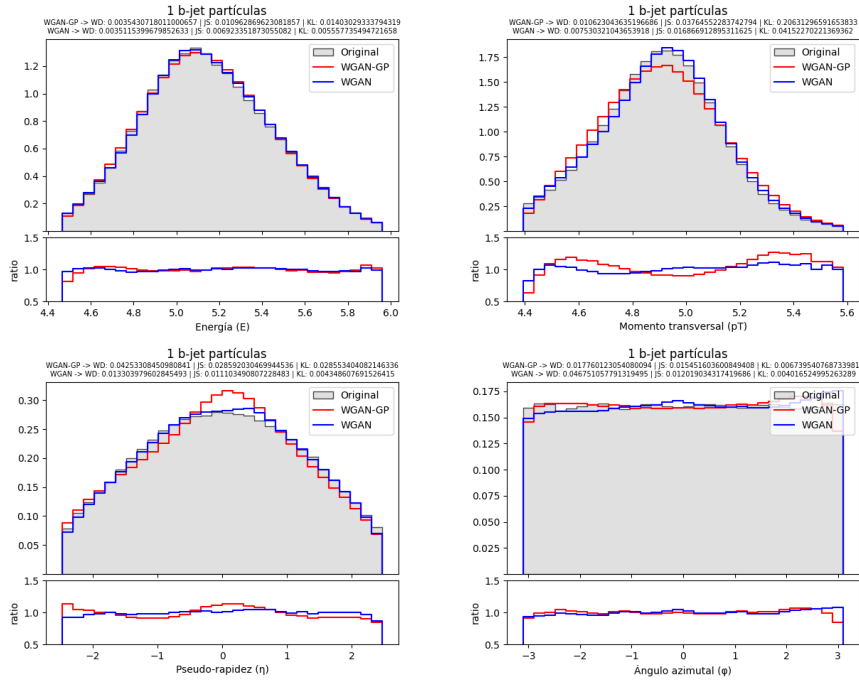


Figura 7.7: Histogramas de las características de los primeros b-jets (b) en el experimento.

izquierdo y derecho de la distribución, respectivamente, en el modelo WGAN-GP. Este ligero desvío podría ser considerado como un punto de mejora en futuros ajustes del modelo.

En lo que respecta a los b-jets, presentados en la Figura 7.7, se observa una buena generación en las cuatro características de la partícula: Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ). Aunque en términos generales se logra una adaptación satisfactoria, es importante destacar que se evidencia una tendencia en el modelo WGAN-GP a generar más partículas con una pseudo-rapidez cercana al centro de la distribución (valor 0). Además, la distribución del momento transversal (pT) parece estar ligeramente más achatada en comparación con los datos originales y los generados por el modelo WGAN, lo que se traduce en la generación de más valores fuera del centro de la distribución en relación a lo observado en los datos reales. Los resultados generados por el modelo WGAN son, aparentemente, más estables.

Para los segundos b-jets, analizados en la Figura 7.8, se observa un comportamiento en línea con lo mencionado anteriormente para los b-jets iniciales. Las cuatro características, es decir, Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ), son generadas de manera razonablemente precisa. No obstante, al igual que en

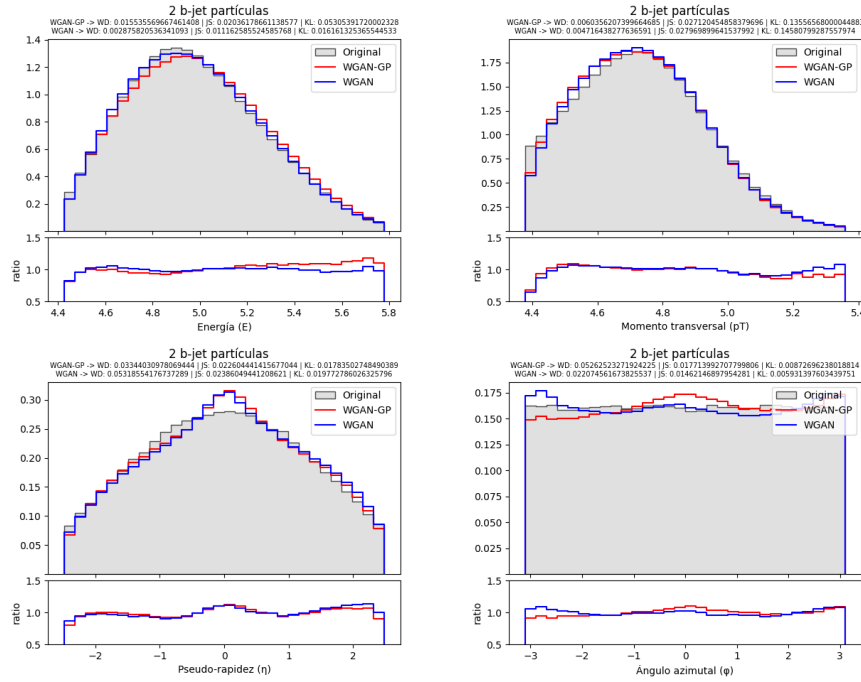


Figura 7.8: Histogramas de las características de los segundos b-jets (b) en el experimento.

el caso anterior, se nota una tendencia a generar más partículas con una pseudo-rapidez cercana al valor cero, que corresponde al centro de la distribución, y como novedad este fenómeno también se manifiesta en los resultados del modelo WGAN.

Para los primeros jets, que pueden examinarse en la Figura 7.9, se aprecia un comportamiento generalmente aceptable en la generación de sus cuatro características principales: Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ). No obstante, la distribución del momento transversal (pT) muestra un desplazamiento a la izquierda del centro de la distribución en comparación con los datos originales, esto es especialmente notable para el modelo WGAN-GP. El ratio respecto a los datos reales parece ser más estable en el modelo WGAN para todas las características.

Para los segundos jets, examinados en la Figura 7.10, se observa una mejora significativa con lo mencionado anteriormente para los jets iniciales. Las cuatro características, es decir, Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ), son generadas de manera razonablemente precisa. No obstante, se observa una tendencia a generar más partículas con una pseudo-rapidez cercana al valor cero, que corresponde al centro de la distribución.

En el caso de los electrones de la Figura 7.11, el modelo logra generar eventos que concuerdan de manera satisfactoria en términos de Energía (E), Pseudo-rapidez (η) y Ángulo azimutal (ϕ). Sin embargo, se observa en el ratio una pequeña tendencia a generar partículas con un Momento transversal (pT) perteneciente a los extremos de la distribución. Además, aunque la generación en términos de Energía (E), Momento transversal (pT) y Pseudo-rapidez (η) es satisfactoria, la generación del Ángulo azimutal (ϕ) no llega a adoptar una distribución uniforme en ambos modelos.

En el caso de los positrones en la Figura 7.12, se observa una buena adaptación en términos de Energía (E), Pseudo-rapidez (η) y Ángulo azimutal (ϕ) para ambos modelos. A diferencia de los electrones, no se observa una discrepancia significativa en la distribución del Momento transversal (pT). Además, los positrones presentan un buen ajuste en términos de Energía (E), Pseudo-rapidez (η) y Momento transversal (pT), y parece ha-

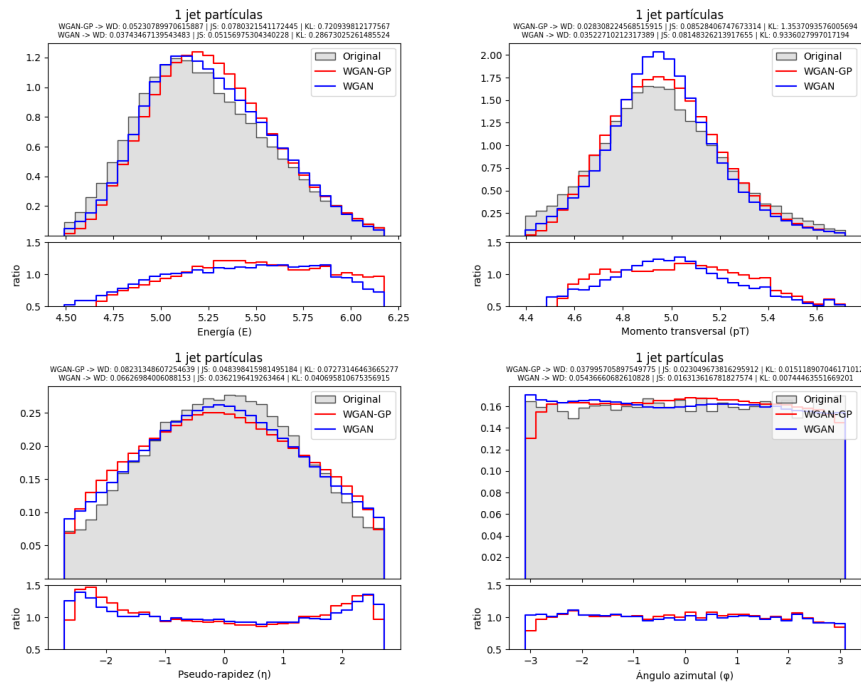


Figura 7.9: Histogramas de las características de los primeros jets (j) en el experimento.

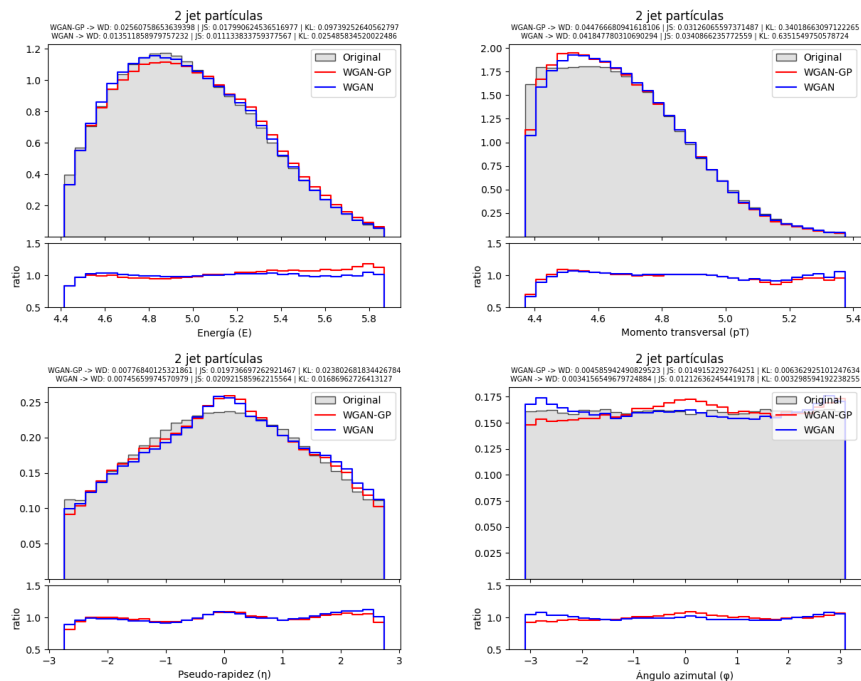


Figura 7.10: Histogramas de las características de los segundos jets (j) en el experimento.

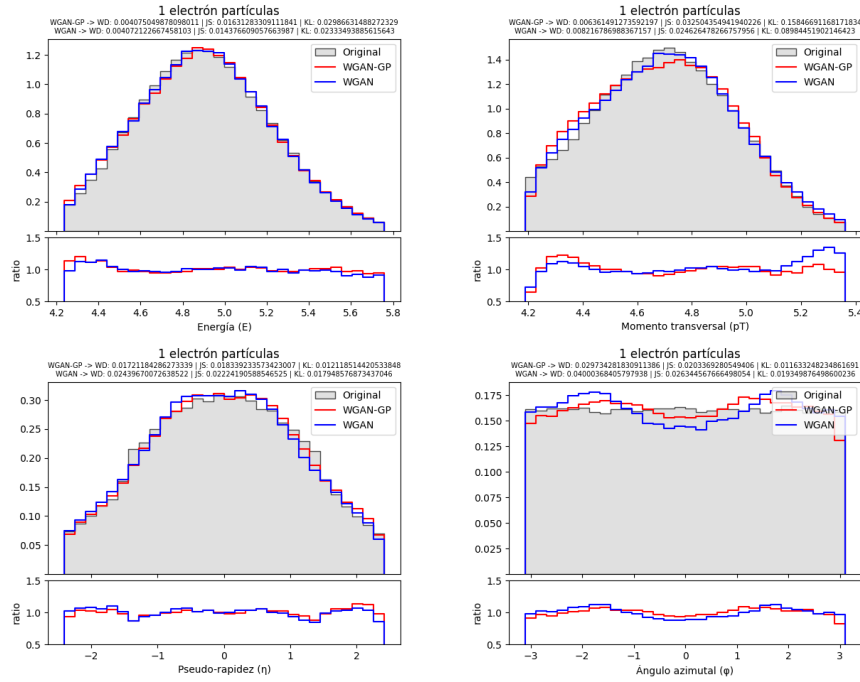


Figura 7.11: Histogramas de las características de los electrones (e^-) en el experimento.

berse capturado con mayor claridad la distribución uniforme en el Ángulo azimutal (ϕ), a diferencia de lo observado en los electrones.

En cuanto a los muones, se observa, en la Figura 7.13, una notable correspondencia entre los datos generados y las distribuciones originales en todas las características: Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ). Sin embargo, respecto a la Pseudo-rapidez (η), se observa una tendencia mayor a generar eventos en el centro y extremo derecho de su distribución. Para el modelo WGAN se aprecia una tendencia a generar valores en el extremo derecho de la distribución del Momento transversal (pT).

Al igual que en los muones, los antimuones, Figura 7.14, presentan una similitud excepcional entre los datos generados y las distribuciones originales en todas las características: Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ). La Pseudo-rapidez (η) no presenta el mismo comportamiento que en los muones, no mostrando una tendencia a generar eventos en el centro de su distribución.

El caso de los fotones, analizado en la Figura 7.15, presenta un reto particular. Dada la distribución peculiar de su Pseudo-rapidez (η), el modelo captura en su mayoría la tendencia de esta característica. Aunque en general se mantiene una coherencia, no se ajusta completamente debido posiblemente a la singularidad de su distribución y a la menor ocurrencia de fotones en el conjunto de datos. El resto de sus características presentan un ajuste aceptable. Aunque cabe mencionar que ninguno de los dos modelos parecen capturar el escalón presente en la cola derecha del Momento transversal (pT).

Modelo	Métricas de similitud			Métricas de generación	
	WS	JS	KL	TEI	TG
WGAN-GP	0.0214	0.0235	0.0929	1.14 %	48.416ms
WGAN	0.0197	0.0181	0.0941	1.06 %	47.334ms

Tabla 7.1: Métricas de los resultados del experimento SM.

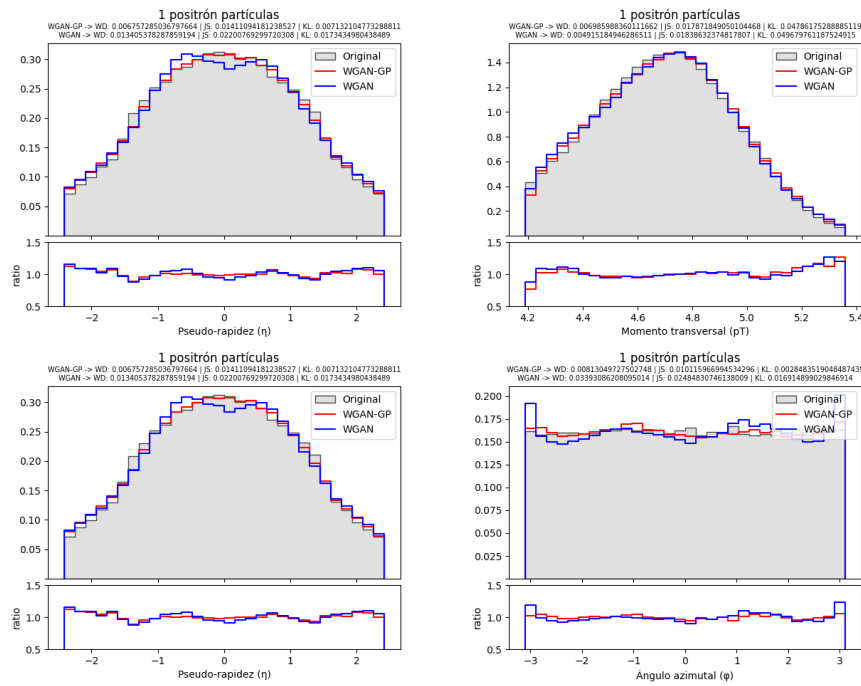


Figura 7.12: Histogramas de las características de los positrones (e^+) en el experimento.

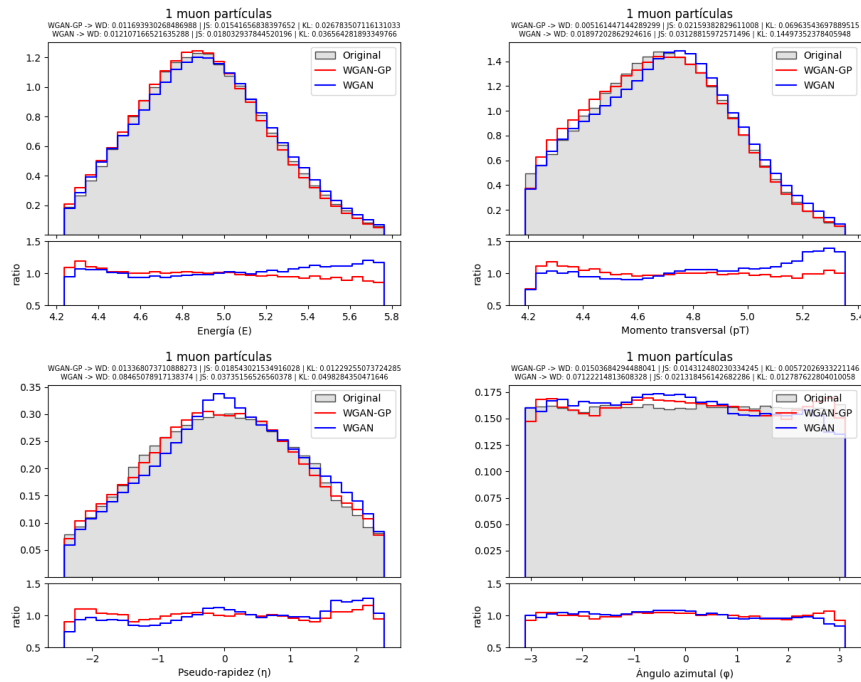


Figura 7.13: Histogramas de las características de los muones (m^-) en el experimento.

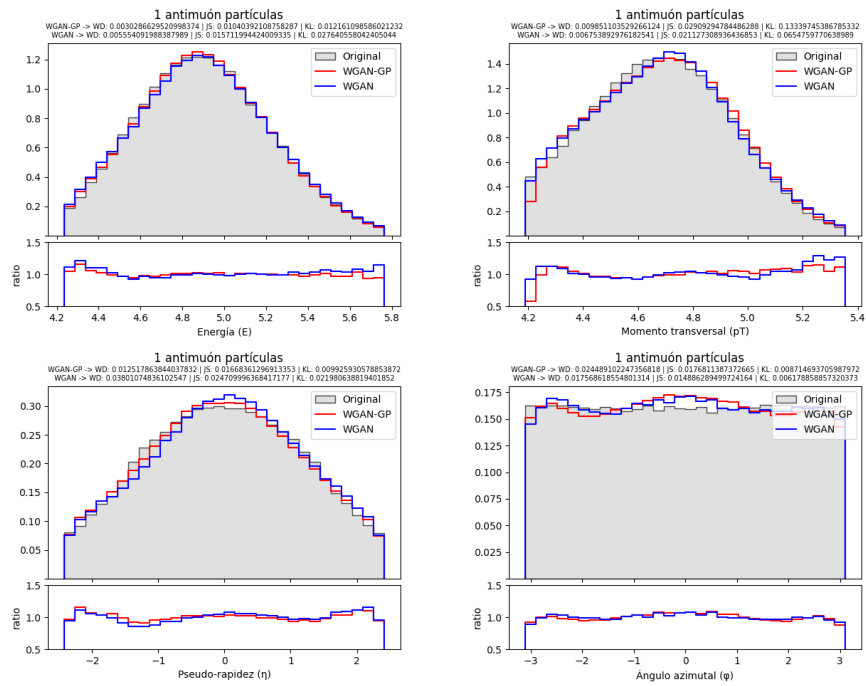


Figura 7.14: Histogramas de las características de los antimuones (m^+) en el experimento.

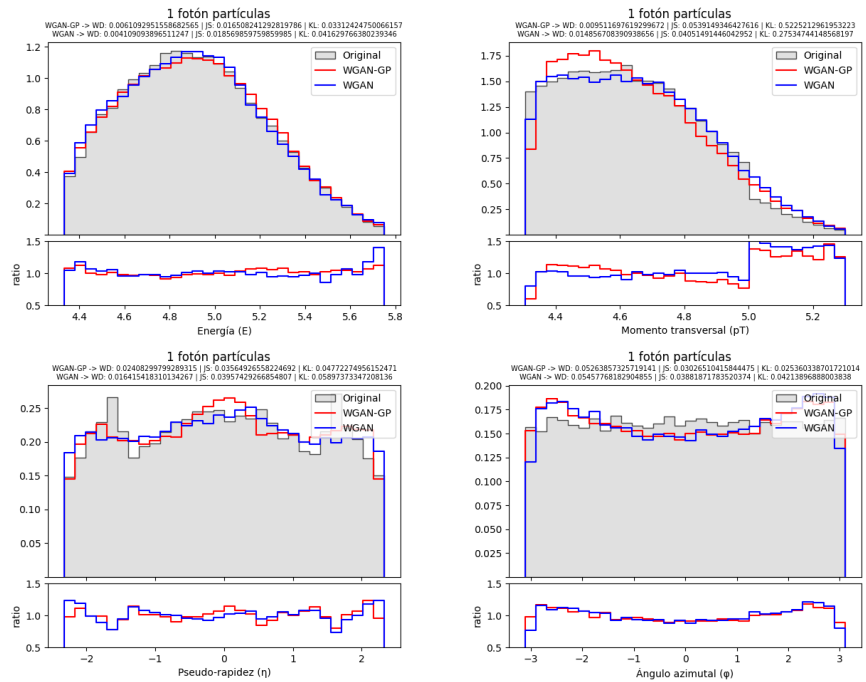


Figura 7.15: Histogramas de las características de los fotones (g) en el experimento.

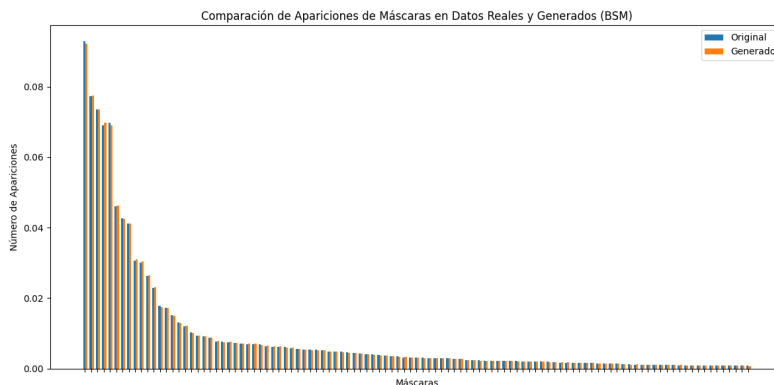


Figura 7.16: Comparación máscaras generadas y máscaras presentes en el conjunto de datos filtrado (BSM).

7.2 Aplicaciones en procesos BSM

En la sección anterior, llevamos a cabo el entrenamiento de los modelo utilizando un conjunto de datos específico de un tipo particular de evento, logrando resultados que se ajustaron de manera adecuada a las distribuciones de los datos originales. Con el objetivo de completar nuestra investigación, hemos tomado la decisión de reentrenar los modelos utilizando un tipo de evento perteneciente al Más Allá del Modelo Estándar (BSM), con el fin de evaluar la versatilidad de la arquitectura propuesta. El conjunto de datos seleccionado es stop_02, proveniente del conjunto de datos de eventos Más Allá del Modelo Estándar (descrito en la Tabla 3.2).

En los resultados que se presentarán a continuación, llevamos a cabo un proceso de ajuste (finetuning) sobre los modelos previamente expuestos, utilizando el nuevo conjunto de datos durante 40 epochs.

En el proceso de generación de datos, la generación de las máscaras, se realiza con la metodología explicada en la sección 7.1.2. La distribución del generador puede apreciarse en la imagen 7.16.

7.2.1. Resultados

En las características de la energía faltante del evento (MET y $MET\phi$), se observa en la Figura 7.17 una correspondencia aceptable entre los datos generados y las distribuciones originales. Sin embargo, se nota una marcada dispersión del modelo al generar eventos, como se observa en la forma achatada del MET. Esta dispersión es un punto de mejora en futuros ajustes del modelo y podría deberse a las diferencias significativas entre las distribuciones de los conjuntos de datos originales, ttbar_10fb y stop_02.

Como podemos ver, en la Figura 7.18 se observa un buen ajuste de las cuatro características: Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ). Sin embargo, es importante destacar un desplazamiento a la izquierda de la distribución original, con valores menores a los originales, en los valores de Energía (E) y Pseudo-rapidez (η). Además existe una tendencia, en ambos modelos, a generar valores centrales de la distribución para los valores de Pseudo-rapidez (η).

Para los segundos b-jets, en la Figura 7.19 se observa un comportamiento similar al de los b-jets iniciales. Las cuatro características, es decir, Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ), son generadas de manera razonablemente precisa. Sin embargo, se observa una diferencia notable en el momento transversal

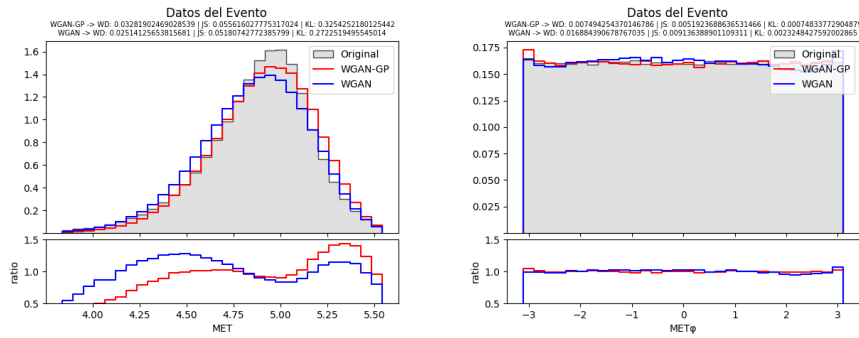


Figura 7.17: Histogramas de las características de la energía faltante del evento (MET, $MET\phi$) en stop_02.

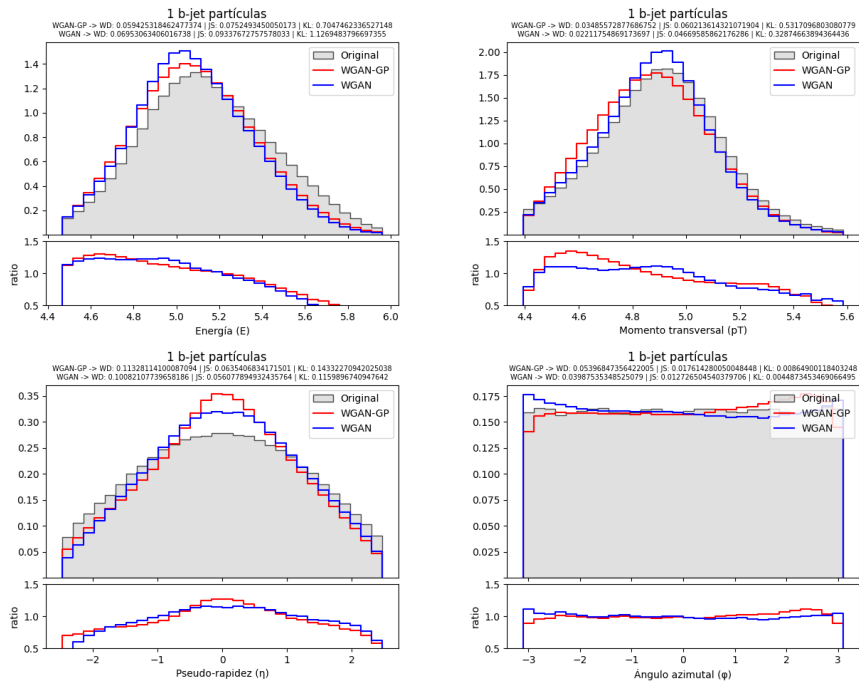


Figura 7.18: Histogramas de las características de los primeros b-jets (b) en stop_02.

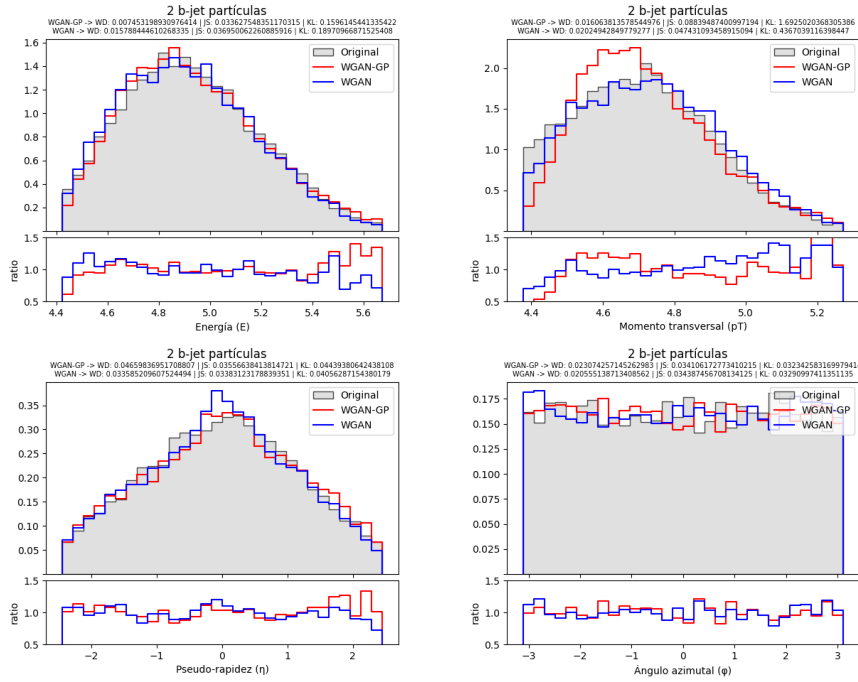


Figura 7.19: Histogramas de las características de los segundos b-jets (b) en stop_02.

(pT), que presenta una mayor cantidad de partículas generadas en el centro de la distribución para el modelo WGAN-GP, curiosamente el modelo WGAN no refleja el mismo fenómeno.

Además, se observa una diferencia en la calidad respecto a los primeros b-jets, probablemente debido a una cantidad menor de eventos observados que incluyan esta partícula. Esto podría estar influyendo en la precisión del modelo para estos casos específicos.

Para los primeros jets analizados en la Figura 7.20, se aprecia un buen comportamiento, en ambos modelos, en la generación de sus cuatro características principales: Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ). Sin embargo, se observa en ambos modelos una tendencia a generar más partículas con valores inferiores al centro de la distribución para el Momento transversal (pT). El modelo WGAN parece favorecer la generación en las partes más extremas de la uniforme de la característica Ángulo azimutal (ϕ).

En los segundos jets, observados en la Figura 7.21, se aprecia un mejor ajuste que en los primeros jets. Se observa un desplazamiento a la izquierda del centro de la distribución en los valores generados para las características Momento transversal (pT) y Pseudo-rapidez (η), además de una tendencia a generar valores más extremos. Las otras dos características, Energía (E) y Ángulo azimutal (ϕ), se han ajustado satisfactoriamente.

En el caso de los electrones, encontrado en la Figura 7.22, el modelo logra generar eventos que concuerdan de manera satisfactoria en términos de Energía (E), Pseudo-rapidez (η) y Ángulo azimutal (ϕ). Sin embargo, se observa que el modelo tiende a generar datos más extremos, aumentando el tamaño de la distribución en los bordes de la misma. Además, aunque la generación en términos de Energía (E) y Pseudo-rapidez (η) es satisfactoria, parece que no termina de ajustarse la distribución uniforme presente en el Ángulo azimutal (ϕ).

En el caso de los positrones, en la Figura 7.23 se observa un buen ajuste en términos de Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ). A diferencia de los electrones, no se manifiesta el aumento en los bordes de la distribución

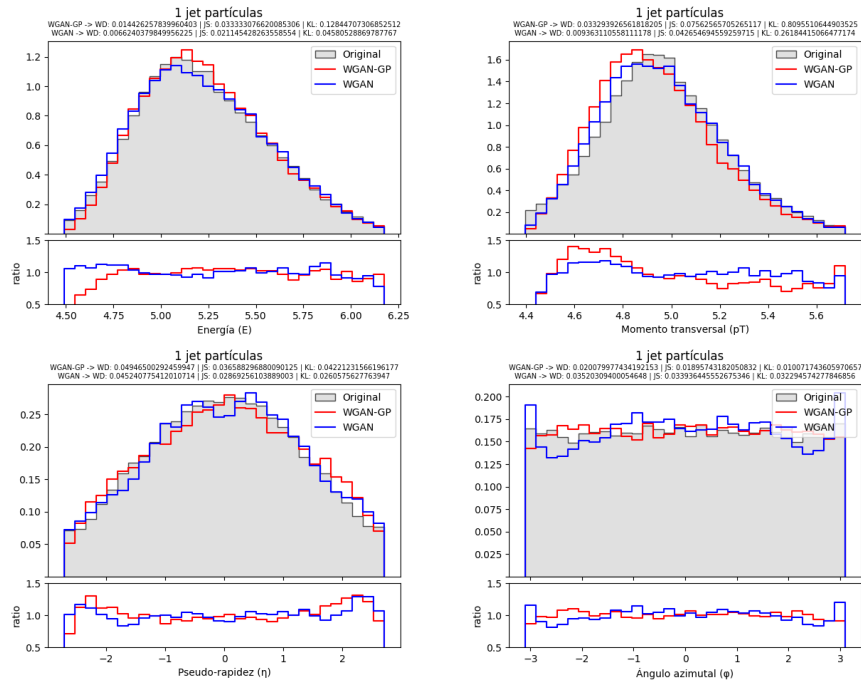


Figura 7.20: Histogramas de las características de los primeros jets (j) en stop_02.

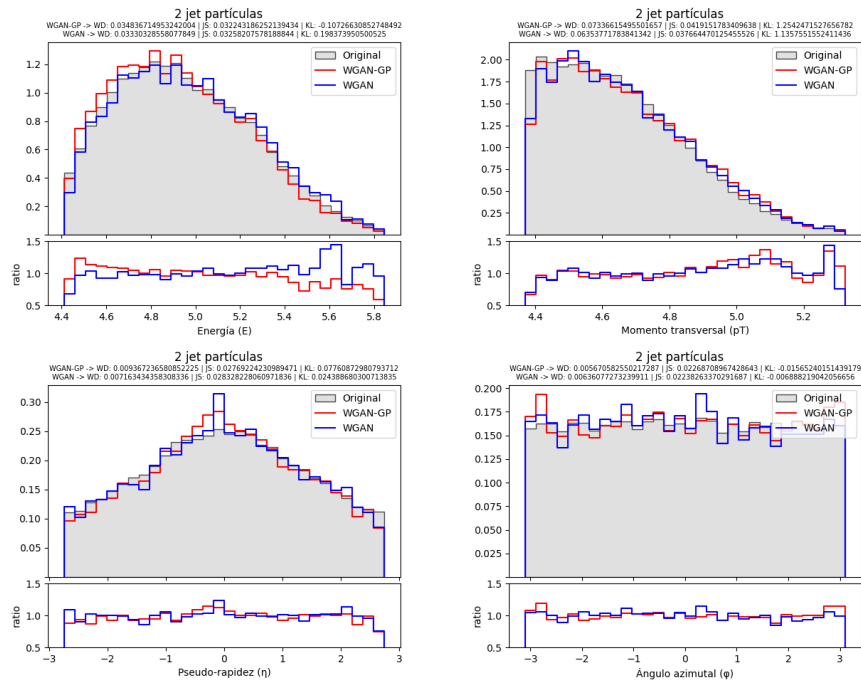


Figura 7.21: Histogramas de las características de los segundos jets (j) en stop_02.

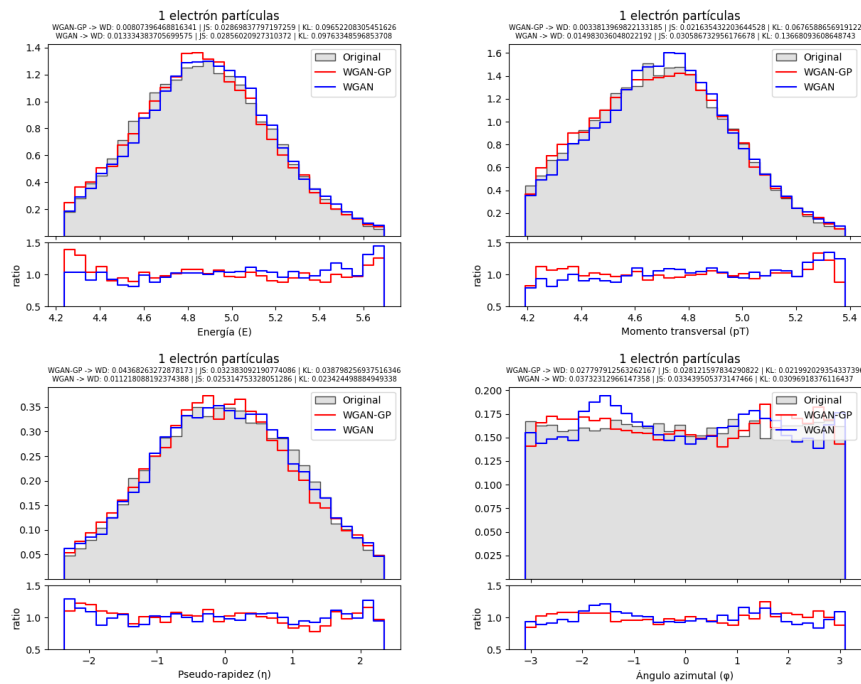


Figura 7.22: Histogramas de las características de los electrones (e^-) en stop_02.

de los valores de Energía (E), Pseudo-rapidez (η). Además, parece capturarse mejor la distribución uniforme presente en el Ángulo azimutal (ϕ).

En el caso de los muones, al igual que en los positrones, es notable la correspondencia excepcional entre los datos generados y las distribuciones originales en todas las características reflejadas en la Figura 7.24: Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ). Cabe destacar la forma de punta de flecha presente en la Pseudo-rapidez (η), que indica una mayor tasa de individuos en el centro de la distribución y en sus extremos.

En el caso de los antimuones de la Figura 7.25, se mantiene la excepcional similitud presentada en los muones y positrones, entre los datos generados y las distribuciones originales en todas las características: Energía (E), Pseudo-rapidez (η), Momento transversal (pT) y Ángulo azimutal (ϕ). Poco hay que destacar excepto el aumento de valores extremos en los bordes de la distribución, fenómeno ya observado anteriormente.

El caso de los fotones presenta un reto particular, tal como podemos observar en la Figura 7.26. Dada la distribución peculiar de su Pseudo-rapidez (η), el modelo captura en su mayoría la tendencia general de esta característica. Aunque en general se mantiene una coherencia, no se ajusta completamente debido posiblemente a la singularidad de su distribución y a la menor ocurrencia de fotones en el conjunto de datos. El resto de sus características presentan un buen ajuste.

Los resultados obtenidos respecto a la generación de fotones son similares a los observados en el modelo WGAN-GP. En este caso, parece existir un mayor ajuste en la Energía (E) y el Momento transversal (pT). No obstante, la forma de la Pseudo-rapidez (η) está más desdibujada.

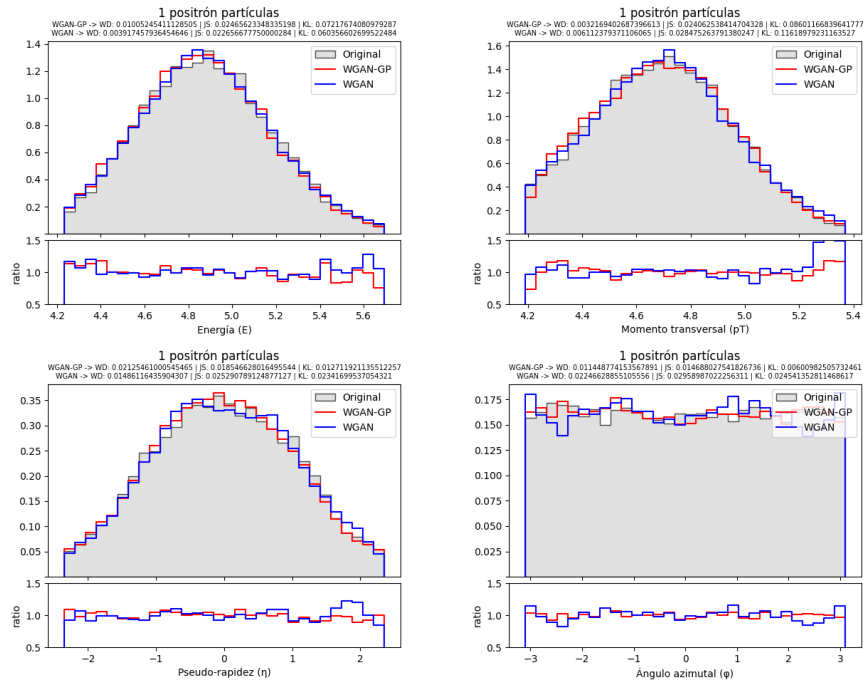


Figura 7.23: Histogramas de las características de los positrones (e+) en stop_02.

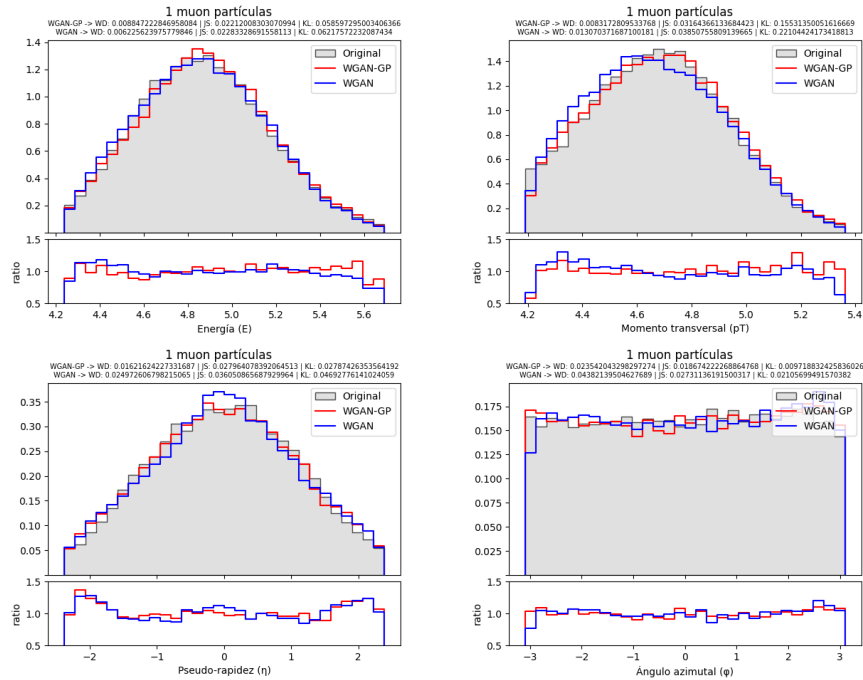


Figura 7.24: Histogramas de las características de los muones (m-) en stop_02.

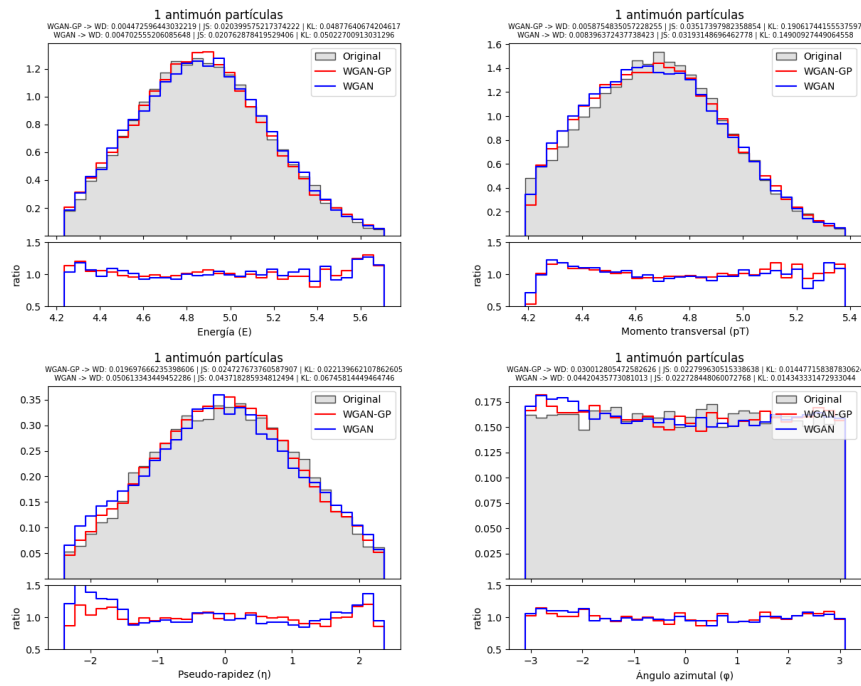


Figura 7.25: Histogramas de las características de los antimuones (m+) en stop_02.

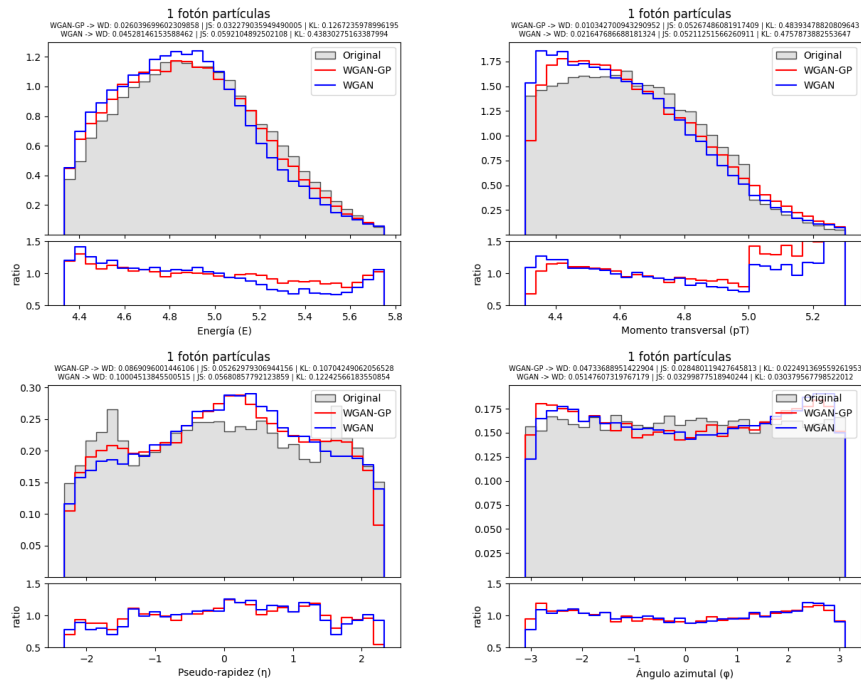


Figura 7.26: Histogramas de las características de los fotones (g) en stop_02.

Modelo	Métricas de similitud			Métricas de generación	
	WS	JS	KL	TEI	TG
WGAN-GP	0.02656	0.0325	0.2601	3.904 %	47.637ms
WGAN	0.02665	0.0300	0.2084	3.826 %	48.103ms

Tabla 7.2: Métricas de los resultados del experimento BSM.

CAPÍTULO 8

Conclusiones

En conclusión, hemos logrado cumplir con éxito los objetivos planteados en este trabajo de fin de grado, cuyo enfoque principal era explorar y evaluar la efectividad de las GANs en la generación de datos de eventos de colisiones de partículas protón-protón. A lo largo de este estudio, hemos comprendido los fundamentos de las GANs y su aplicación en el campo de la física de partículas, adaptando cuidadosamente las arquitecturas de red y ajustando los hiper-parámetros para acomodar las características específicas de los datos de colisiones de partículas. Además, llevamos a cabo una comparación exhaustiva de los resultados obtenidos con los datos generados por las GANs y los datos simulados por Montecarlo, empleando métricas apropiadas para evaluar la calidad y similitud entre ambas fuentes de datos. Aunque hemos obtenido resultados prometedores, también reconocemos que existen áreas para mejorar y afinar nuestros modelos.

Los resultados obtenidos con los modelos generativos han sido muy positivos y satisfactorios, aunque también reconocemos que existen áreas para mejorar y afinar nuestros modelos. A lo largo de este proyecto, hemos aprendido no solo el desarrollo, entrenamiento y evaluación de modelos GANs, sino también información característica respecto al conjunto de datos de colisiones de partículas protón-protón. Este proceso nos ha permitido comprender en profundidad cómo se generan y distribuyen estos eventos, así como las complejidades involucradas en replicar estas distribuciones de manera realista. La experiencia adquirida nos ha preparado mejor para abordar problemas similares en el futuro y ha sido un componente esencial de nuestra formación en ciencia de datos y aprendizaje automático.

Diversas asignaturas previas han sido particularmente útiles para este proyecto, proporcionando las bases necesarias en programación, desarrollo de modelos de *machine learning*. Asignaturas como Proyecto, Técnicas escalables en aprendizaje automático, Modelos descriptivos y predictivos nos han dado las herramientas para enfrentar los desafíos técnicos y conceptuales con mayor confianza y eficacia. El conocimiento adquirido en estas asignaturas fue crucial para el ajuste de hiper-parámetros, la implementación de diferentes arquitecturas de red y la evaluación de resultados, permitiéndonos realizar experimentaciones más informadas y sistemáticas.

Durante el desarrollo del proyecto, enfrentamos varias dificultades iniciales, especialmente en la etapa de experimentación. Inicialmente, intentamos implementar un modelo global que generara todos los eventos, pero esta estrategia no tuvo éxito debido a la complejidad y diversidad de los datos. Este fracaso inicial nos enseñó la importancia de segmentar el problema y abordar la generación de eventos en modelos especializados, lo que resultó ser una estrategia más efectiva, como se ha demostrado en [23], [39], [40]. También probamos arquitecturas más simples, como las GANs básicas y las LSGANs, antes de llegar a la configuración óptima. Estos ensayos y errores fueron fundamentales

para refinar nuestras metodologías y alcanzar los resultados prometedores que finalmente obtuvimos.

En resumen, este trabajo representa un primer paso sólido en la exploración de las GANs para la generación de eventos de colisiones de partículas en el ámbito de la física de partículas. Aunque hemos logrado avances importantes, reconocemos que hay espacio para mejoras y oportunidades futuras. En última instancia, nuestra contribución busca avanzar en la simulación de procesos físicos y en la generación de datos realistas para la comunidad científica en este emocionante campo.

8.1 Trabajos futuros

A pesar de los avances logrados en este trabajo, existen diversas oportunidades de investigación y desarrollo que pueden expandir aún más el enfoque propuesto. A continuación, se presentan algunas direcciones potenciales para futuros trabajos en el campo de generación de eventos de colisiones protón-protón utilizando GAN y otras técnicas de aprendizaje profundo.

Una futura continuación de este trabajo es evaluar la solución propuesta en un abanico más grande de procesos dentro de los eventos SM y BSM, enfatizando que se pueden adaptar los resultados a cualquier tipo de evento.

Una posible extensión es la exploración de enfoques de combinación de modelos, como la utilización de GANs en conjunto con otros tipos de modelos generativos, como los Autoencoders, (AEGANs) [47], o los Autoencoders Variacionales, (VAE-GANs) [48]. La combinación de estos modelos podría aprovechar las fortalezas individuales de cada enfoque y posiblemente mitigar las limitaciones de un solo modelo. Esto podría conducir a una mejora en la calidad de las muestras generadas y una mayor eficiencia en la simulación.

Otro camino a explorar es la aplicación de técnicas de reducción de dimensionalidad. Al reducir la dimensionalidad de los datos, podrían obtenerse representaciones más compactas y eficientes que aún conserven la información relevante. Métodos como Análisis de Componentes Principales (PCA) [49], [50] y t-SNE [51] podrían aplicarse obtener un espacio de menor dimensión.

Además de GANs, existen otros enfoques menos explorados en la generación de datos, como los Modelos de Flujos Normalizados, *Normalizing Flows* (NF) [52], [53]. Estos modelos ofrecen ventajas en términos de capacidad de generación y posibilidad de muestreo inverso. Investigar la aplicabilidad de los NF en la simulación de eventos de colisiones podría abrir nuevas oportunidades para mejorar la calidad y diversidad de las muestras generadas.

Si bien este trabajo se centra en la generación de eventos de colisiones protón-protón, el enfoque propuesto podría aplicarse a otros tipos de colisiones y experimentos. Explorar la adaptabilidad de los modelos generativos a diferentes contextos podría ampliar su utilidad y relevancia en la comunidad científica.

En resumen, el camino hacia la detección de señales de nueva física mediante técnicas de aprendizaje profundo presenta varias oportunidades de desarrollo. La combinación de modelos, la reducción de dimensionalidad y la exploración de modelos menos tradicionales son solo algunas de las direcciones prometedoras que podrían enriquecer y expandir el alcance de este campo de investigación.

Bibliografía

- [1] S. Rappoccio, «The experimental status of direct searches for exotic physics beyond the standard model at the Large Hadron Collider,» *Reviews in Physics*, vol. 4, pág. 100 027, 2019.
- [2] J. Sonneveld, «Searches for physics beyond the standard model at the LHC,» *The African Review of Physics*, vol. 13, 2018.
- [3] A. Collaboration et al., «ATLAS Events at 13 TeV-First 2016 Stable Beams,» inf. téc., 2016. dirección: <http://cds.cern.ch/record/2148236>.
- [4] R. M. Bianchi, A. Collaboration et al., «ATLAS experiment schematic illustration,» inf. téc., 2022. dirección: <https://cds.cern.ch/record/2837191>.
- [5] S. Agostinelli, J. Allison, K. a. Amako et al., «GEANT4—a simulation toolkit,» *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 506, n.º 3, págs. 250-303, 2003.
- [6] S. Otten, S. Caron, W. de Swart et al., «Event generation and statistical sampling for physics with deep generative models and a density information buffer,» *Nature communications*, vol. 12, n.º 1, pág. 2985, 2021.
- [7] M. Arjovsky, S. Chintala y L. Bottou, «Wasserstein Generative Adversarial Networks,» en *Proceedings of the 34th International Conference on Machine Learning*, D. Precup e Y. W. Teh, eds., ép. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, págs. 214-223. dirección: <https://proceedings.mlr.press/v70/arjovsky17a.html>.
- [8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin y A. C. Courville, «Improved Training of Wasserstein GANs,» en *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio et al., eds., vol. 30, Curran Associates, Inc., 2017. dirección: https://proceedings.neurips.cc/paper_files/paper/2017/file/892c3b1c6dcd52936e27cbd0ff683d6-Paper.pdf.
- [9] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang y S. Paul Smolley, «Least squares generative adversarial networks,» en *Proceedings of the IEEE international conference on computer vision*, 2017, págs. 2794-2802.
- [10] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang y B. Póczos, «Mmd gan: Towards deeper understanding of moment matching network,» *Advances in neural information processing systems*, vol. 30, 2017.
- [11] D. P. Kingma y M. Welling, «Auto-Encoding Variational Bayes,» *stat*, vol. 1050, pág. 1, 2014.
- [12] T. Aarrestad, M. van Beekveld, M. Bona et al., «The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider,» *SciPost Phys.*, vol. 12, pág. 043, 2022. DOI: [10.21468/SciPostPhys.12.1.043](https://doi.org/10.21468/SciPostPhys.12.1.043). dirección: <https://scipost.org/10.21468/SciPostPhys.12.1.043>.

- [13] M. Aaboud, G. Aad, B. Abbott et al., «A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment,» *The European Physical Journal C*, vol. 79, n.º 2, págs. 1-45, 2019.
- [14] H. Hashemi, N. Hartmann, S. Sharifzadeh, J. Kahn y T. Kuhr, «Ultra-high-resolution detector simulation with intra-event aware GAN and self-supervised relational reasoning,» *arXiv preprint arXiv:2303.08046*, 2023.
- [15] R. Di Sipio, M. F. Giannelli, S. K. Haghghat y S. Palazzo, «DijetGAN: a generative-adversarial network approach for the simulation of QCD dijet events at the LHC,» *Journal of high energy physics*, vol. 2019, n.º 8, 2019.
- [16] L. Vaslin, V. Barra y J. Donini, «GAN-AE: an anomaly detection algorithm for New Physics search in LHC data,» *The European Physical Journal C*, vol. 83, n.º 11, pág. 1008, 2023.
- [17] G. Kasieczka, B. Nachman, D. Shih et al., «The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics,» *Reports on progress in physics*, vol. 84, n.º 12, pág. 124 201, 2021.
- [18] M. Touranakou, N. Chernyavskaya, J. Duarte et al., «Particle-based fast jet simulation at the LHC with variational autoencoders,» *Machine Learning: Science and Technology*, vol. 3, n.º 3, pág. 035 003, 2022.
- [19] M. van Beekveld, S. Caron, L. Hendriks et al., «Combining outlier analysis algorithms to identify new physics at the LHC,» *Journal of High Energy Physics*, vol. 2021, n.º 9, págs. 1-33, 2021.
- [20] A. Butter, T. Plehn, S. Schumann et al., «Machine learning and LHC event generation,» *SciPost Physics*, vol. 14, n.º 4, pág. 079, 2023.
- [21] C. Chen, O. Cerri, T. Nguyen, J. Vlimant y M. Pierini, «Data augmentation at the LHC through analysis-specific fast simulation with deep learning,» *GEN*, vol. 16, págs. 24-4, 2020.
- [22] J. A. Martínez, T. Q. Nguyen, M. Pierini, M. Spiropulu y J.-R. Vlimant, «Particle Generative Adversarial Networks for full-event simulation at the LHC and their application to pileup description,» en *Journal of Physics: Conference Series*, IOP Publishing, vol. 1525, 2020, pág. 012 081.
- [23] R. Balanzá García, «Use of Deep Learning generative models for Monte Carlo event simulation in the context of LHC experiments,» Universitat Politècnica de València, 2022.
- [24] C. Salaun, A. Gruson, B.-S. Hua, T. Hachisuka y G. Singh, «Regression-based Monte Carlo integration,» *ACM Transactions on Graphics (TOG)*, vol. 41, n.º 4, págs. 1-14, 2022.
- [25] R. Cranley y T. Patterson, «A regression method for the Monte Carlo evaluation of multidimensional integrals,» *Numerische Mathematik*, vol. 16, págs. 58-72, 1970.
- [26] C. Gao, J. Isaacson y C. Krause, «i-flow: High-dimensional Integration and Sampling with Normalizing Flows,» *Machine Learning: Science and Technology*, vol. 1, n.º 4, pág. 045 023, 2020.
- [27] R. Verheyen, «Event Generation and Density Estimation with Surjective Normalizing Flows,» *SciPost Phys.*, vol. 13, pág. 047, 2022. DOI: [10.21468/SciPostPhys.13.3.047](https://doi.org/10.21468/SciPostPhys.13.3.047). dirección: <https://scipost.org/10.21468/SciPostPhys.13.3.047>.

- [28] P. Jawahar, T. Aarrestad, N. Chernyavskaya et al., «Improving Variational Autoencoders for New Physics Detection at the LHC With Normalizing Flows,» *Frontiers in Big Data*, vol. 5, 2022, ISSN: 2624-909X. DOI: [10.3389/fdata.2022.803685](https://doi.org/10.3389/fdata.2022.803685). dirección: <https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2022.803685>.
- [29] D. community, *LHC simulation Project*, 2020. DOI: [10.5281/zenodo.3685861](https://doi.org/10.5281/zenodo.3685861). dirección: <https://zenodo.org/record/3685861>.
- [30] S. Amoroso, P. Azzurri, J. Bendavid et al., «Les Houches 2019: physics at TeV colliders: standard model working group report,» *arXiv preprint arXiv:2003.01700*, 2020.
- [31] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., «Generative Adversarial Nets,» en *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence y K. Weinberger, eds., vol. 27, Curran Associates, Inc., 2014. dirección: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.
- [32] Nvidia, *Nvidia Ampere GA102 GPU architecture*. dirección: <https://images.nvidia.com/aem-dam/en-zz/Solutions/geforce/ampere/pdf/NVIDIA-ampere-GA102-GPU-Architecture-Whitepaper-V1.pdf>.
- [33] E. Bisong, «Introduction to Scikit-learn,» en *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Berkeley, CA: Apress, 2019, págs. 215-229, ISBN: 978-1-4842-4470-8. DOI: [10.1007/978-1-4842-4470-8_18](https://doi.org/10.1007/978-1-4842-4470-8_18). dirección: https://doi.org/10.1007/978-1-4842-4470-8_18.
- [34] T. E. Oliphant, *A Guide to NumPy*. Trelgol Publishing USA, 2006, vol. 1.
- [35] A. Yim, C. Chung y A. Yu, *Matplotlib for Python Developers: Effective techniques for data visualization with Python*. Packt Publishing Ltd, 2018.
- [36] F. Chollet, *Deep learning with Python*. Simon y Schuster, 2021.
- [37] M. Abadi, P. Barham, J. Chen et al., «TensorFlow: A System for Large-Scale Machine Learning,» en *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, Savannah, GA: USENIX Association, 2016, págs. 265-283, ISBN: 978-1-931971-33-1. dirección: <https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi>.
- [38] A. Gulli y S. Pal, *Deep learning with Keras*. Packt Publishing Ltd, 2017.
- [39] R. Balanza García, «Improving particle physics event simulation by using variational autoencoders in the context of LHC experiments,» Tesis de mtría., Universitat Politècnica de València, 2023.
- [40] J. Boix Avalos, «Application of generative models based on normalizing flows for the simulation of events in LHC experiments as an alternative to Monte Carlo methods,» Universitat Politècnica de València, 2023.
- [41] S. Kullback y R. A. Leibler, «On Information and Sufficiency,» *The Annals of Mathematical Statistics*, vol. 22, n.º 1, págs. 79-86, 1951, ISSN: 00034851. dirección: <http://www.jstor.org/stable/2236703> (visitado 25-06-2024).
- [42] J. Lin, «Divergence measures based on the Shannon entropy,» *IEEE Transactions on Information Theory*, vol. 37, n.º 1, págs. 145-151, 1991. DOI: [10.1109/18.61115](https://doi.org/10.1109/18.61115).
- [43] C. Villani et al., *Optimal transport: old and new*. Springer, 2009, vol. 338.
- [44] E. Levina y P. Bickel, «The earth mover's distance is the mallows distance: Some insights from statistics,» en *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, IEEE, vol. 2, 2001, págs. 251-256.

- [45] A. J. Walker, «An efficient method for generating discrete random variables with general distributions,» *ACM Transactions on Mathematical Software (TOMS)*, vol. 3, n.º 3, págs. 253-256, 1977.
- [46] D. P. Kingma y J. Ba, «Adam: A method for stochastic optimization,» *arXiv preprint arXiv:1412.6980*, 2014.
- [47] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow y B. Frey, «Adversarial autoencoders,» *arXiv preprint arXiv:1511.05644*, 2015.
- [48] X. Yu, X. Zhang, Y. Cao y M. Xia, «VAEGAN: A collaborative filtering framework based on adversarial variational autoencoders,» en *IJCAI*, 2019, págs. 4206-4212.
- [49] K. Pearson, «LIII. On lines and planes of closest fit to systems of points in space,» *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, n.º 11, págs. 559-572, 1901.
- [50] I. T. Jolliffe, «Principal component analysis for special types of data,» en *Principal Component Analysis*. New York, NY: Springer New York, 2002, págs. 338-372, ISBN: 978-0-387-22440-4. DOI: [10.1007/0-387-22440-8_13](https://doi.org/10.1007/0-387-22440-8_13). dirección: https://doi.org/10.1007/0-387-22440-8_13.
- [51] L. Van der Maaten y G. Hinton, «Visualizing data using t-SNE,» *Journal of machine learning research*, vol. 9, n.º 11, 2008.
- [52] D. Rezende y S. Mohamed, «Variational inference with normalizing flows,» en *International conference on machine learning*, PMLR, 2015, págs. 1530-1538.
- [53] R. T. Chen, Y. Rubanova, J. Bettencourt y D. K. Duvenaud, «Neural ordinary differential equations,» *Advances in neural information processing systems*, vol. 31, 2018.

APÉNDICE A

Componentes de los modelos

En este apéndice, describimos los principales componentes que hemos utilizado para construir nuestros modelos durante la fase de experimentación de este trabajo, con el fin de aclarar su definición y propósito.

A.1 Funciones de activación

Una función de activación en una red neuronal es un elemento fundamental que introduce no linealidad en el modelo, permitiendo a la red capturar relaciones y patrones complejos en los datos. Cuando se reciben las entradas de una neurona, la función de activación determina si la neurona debe activarse y cuánta información se propaga a través de ella hacia las capas posteriores. Al aplicar una función no lineal, como la ReLU (Rectified Linear Unit) o la tangente hiperbólica, la red puede aprender a modelar relaciones no triviales entre las características de entrada y las salidas, lo que la capacita para abordar problemas más complejos y realizar transformaciones más ricas en los datos a medida que fluyen a través de las capas neuronales.

Lineal

La función de activación lineal es una función simple que calcula una salida proporcional a la entrada sin aplicar ninguna transformación no lineal. Matemáticamente, la función de activación lineal se define como:

$$f(x) = x \tag{A.1}$$

A diferencia de otras funciones de activación no lineales como ReLU o tangente hiperbólica, la función lineal no introduce ninguna no linealidad en los datos. Esto significa que las capas que utilizan esta función de activación simplemente realizarán una transformación lineal de los datos de entrada.

Rectified Linear Unit (ReLU)

La función de activación ReLU (Rectified Linear Unit) es una función ampliamente utilizada en redes neuronales debido a su simplicidad y eficacia en la solución del problema de desvanecimiento de gradientes. Se define matemáticamente como:

$$f(x) = \text{máx}(0, x) \tag{A.2}$$

Donde x es la entrada a la función. La función ReLU toma el valor de x si x es mayor o igual a cero, y si x es negativo, la función asigna un valor de cero. Esto significa que la función ReLU activa o enciende las neuronas si la entrada es positiva y las desactiva si la

entrada es negativa. La simplicidad de la función ReLU y su capacidad para acelerar el proceso de entrenamiento al abordar los gradientes cercanos a cero la convierten en una opción popular en muchas arquitecturas de redes neuronales.

Leaky Rectified Linear Unit (LeakyReLU)

La función de activación LeakyReLU es una variante de la función ReLU (Rectified Linear Unit). Al igual que en ReLU, la función LeakyReLU también se utiliza en redes neuronales para introducir no linealidades en las salidas de las neuronas. Sin embargo, a diferencia de ReLU, en LeakyReLU, la función no es completamente cero cuando la entrada es negativa, sino que permite un pequeño valor negativo multiplicado por un factor llamado "alpha". Esto ayuda a mitigar el problema de las neuronas muertas, donde las neuronas ReLU pueden quedarse inactivas durante el entrenamiento al tener una salida cero para todas las entradas negativas. La función LeakyReLU se define como:

$$f(x) = \begin{cases} x, & \text{si } x \geq 0 \\ \alpha x, & \text{si } x < 0 \end{cases} \quad (\text{A.3})$$

Donde α es un valor pequeño positivo en nuestro caso 0.2. La ventaja de LeakyReLU es que evita la inactivación completa de las neuronas y permite un gradiente no nulo incluso para entradas negativas, lo que puede facilitar el entrenamiento en comparación con ReLU en algunas situaciones.

Tangente Hiperbólica (TanH)

La función de activación tangente hiperbólica (tanh) es una opción común utilizada en las redes neuronales. La función tangente hiperbólica tiene un rango entre -1 y 1, y es antisimétrica con respecto al origen. Su ecuación matemática es:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (\text{A.4})$$

Esta función de activación se utiliza en las capas ocultas de las redes neuronales para introducir no linealidades y permitir que la red modele datos de manera más compleja. La función tangente hiperbólica toma valores negativos en el rango (-1, 0) y valores positivos en el rango (0, 1), con un punto de inflexión en el origen donde su valor es cero. Esto permite que la función capture relaciones no lineales y puede ser útil en situaciones donde se necesita una respuesta no lineal y simétrica en torno al origen.

A.2 Optimizadores

Adam

El optimizador Adam (*Adaptive Moment Estimation*) [46] es un algoritmo de optimización utilizado comúnmente para entrenar redes neuronales en el campo del aprendizaje profundo. Adam combina las ventajas de dos técnicas de optimización, el algoritmo de descenso de gradiente estocástico (SGD) y el método de momento adaptativo (Adagrad). El algoritmo adapta automáticamente las tasas de aprendizaje para cada parámetro individual, calcula momentos de primer y segundo orden de los gradientes y los utiliza para ajustar los pesos de la red de manera eficiente.

Las ecuaciones que describen el algoritmo Adam son las siguientes:

1. Cálculo del momento del primer orden (media móvil del gradiente):

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \quad (\text{A.5})$$

2. Cálculo del momento del segundo orden (media móvil de los gradientes al cuadrado):

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \quad (\text{A.6})$$

3. Corrección de sesgo de momento del primer orden:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \quad (\text{A.7})$$

4. Corrección de sesgo de momento del segundo orden:

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (\text{A.8})$$

5. Actualización de los pesos del modelo:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t \quad (\text{A.9})$$

Donde t es el paso de tiempo actual, g_t es el gradiente del parámetro en el paso t , α es la tasa de aprendizaje, β_1 y β_2 son los coeficientes de decaimiento para los momentos del primer y segundo orden respectivamente y ϵ es una pequeña constante para evitar la división por cero.

APÉNDICE B

Distribuciones de los datos

Antes de comenzar con la experimentación, se llevó a cabo un proceso de análisis de datos, descrito en la Sección 3.3. Como parte de ese proceso, se calcularon las distribuciones estadísticas de los datos fuente de los conjuntos de datos utilizados, tanto del SM como del BSM. En este apéndice, presentamos la distribución de los datos reales de ambos conjuntos de datos.

Es importante mencionar que, tal como se expuso en la Sección 3.3, notamos que las características de Energía (E) y Momento Transversal (p_T) seguían una distribución log-normal. Por lo tanto, se aplicó el logaritmo en base 10 a los histogramas de estas características de eventos y partículas para lograr una mejor representación gráfica y comprensión de sus distribuciones.

B.1 Propiedades del evento

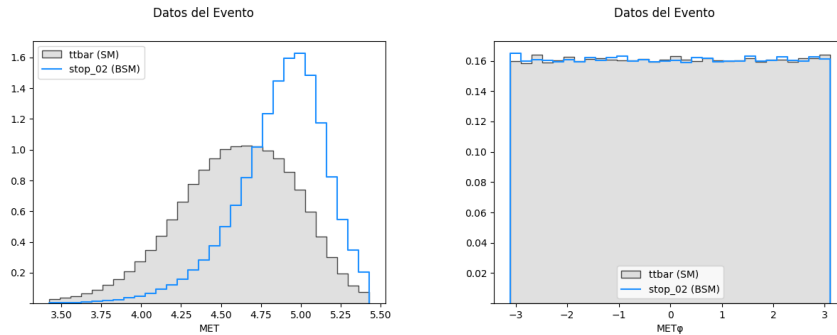


Figura B.1: Histogramas de las características de la energía faltante del evento (MET , $MET\phi$) y la energía total.

B.2 Propiedades de las partículas

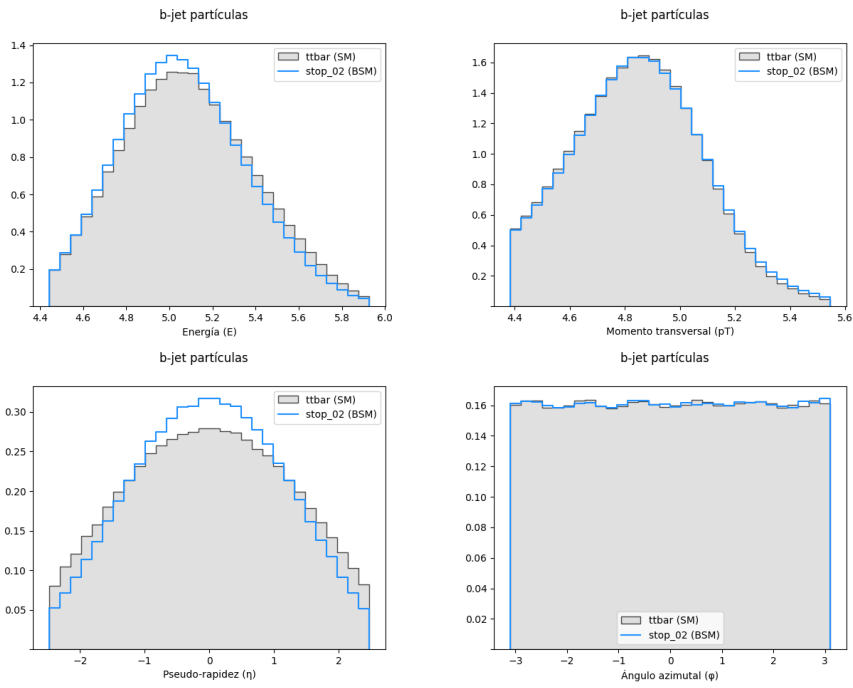


Figura B.2: Histogramas de las características de los b-jets (b).

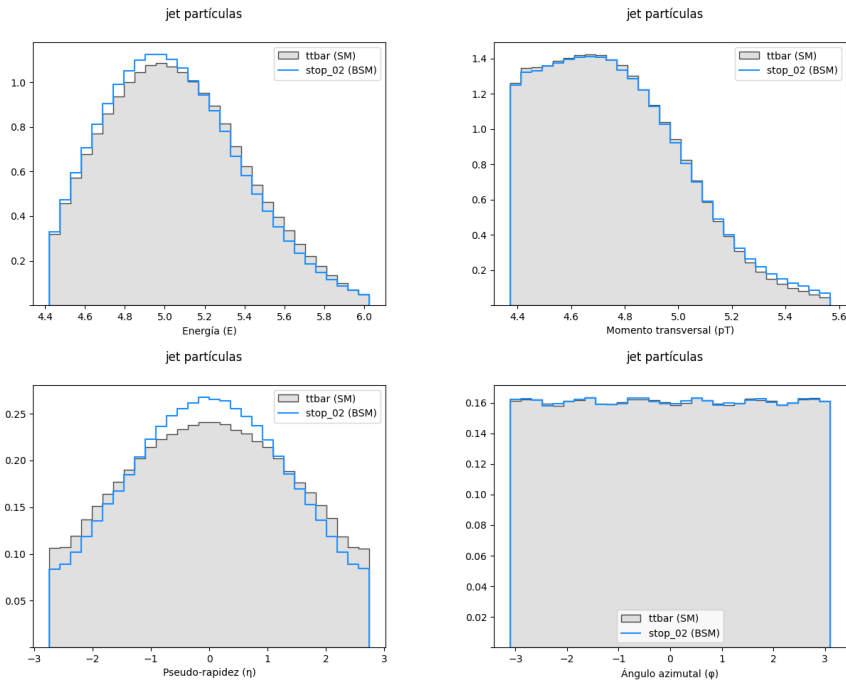


Figura B.3: Histogramas de las características de los jets (j).

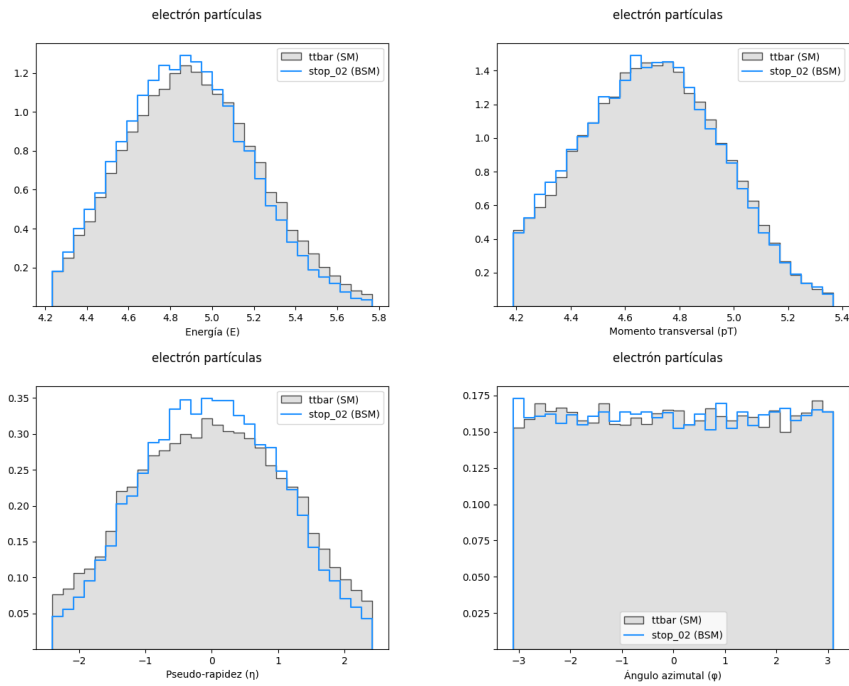


Figura B.4: Histogramas de las características de los electrones (e-).

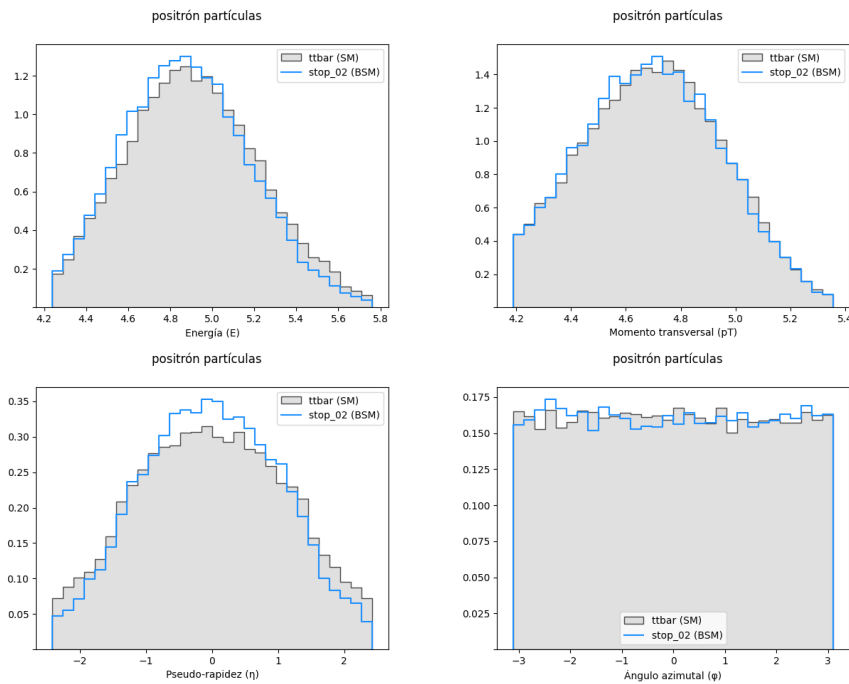


Figura B.5: Histogramas de las características de los positrones (e+).

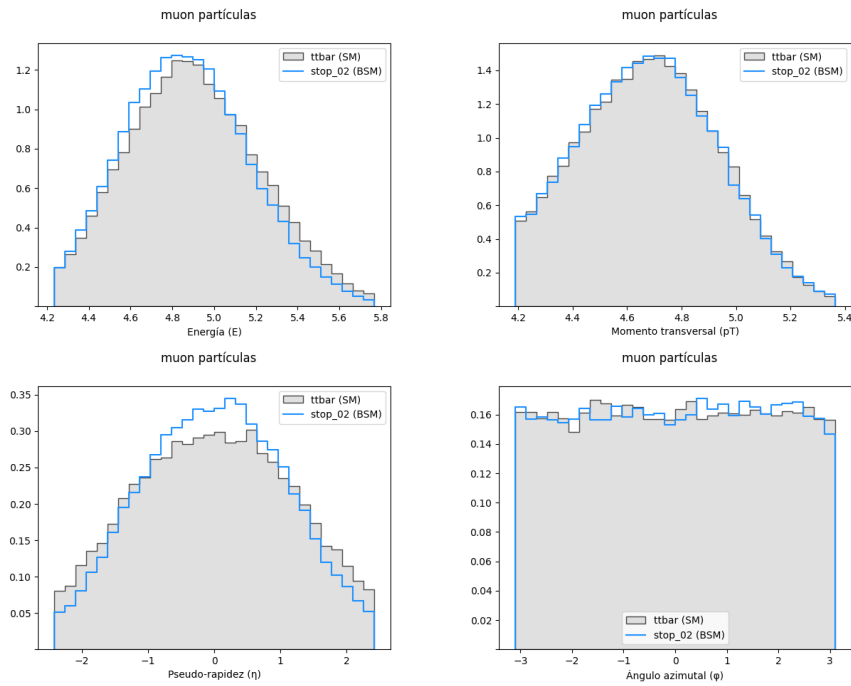


Figura B.6: Histogramas de las características de los muones (m^-).

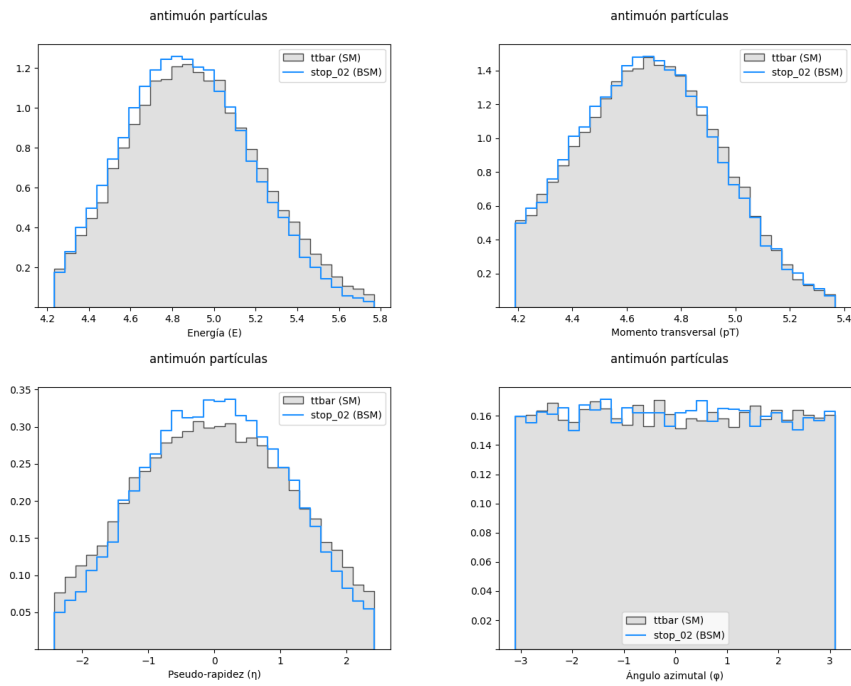


Figura B.7: Histogramas de las características de los antimuones (m^+).

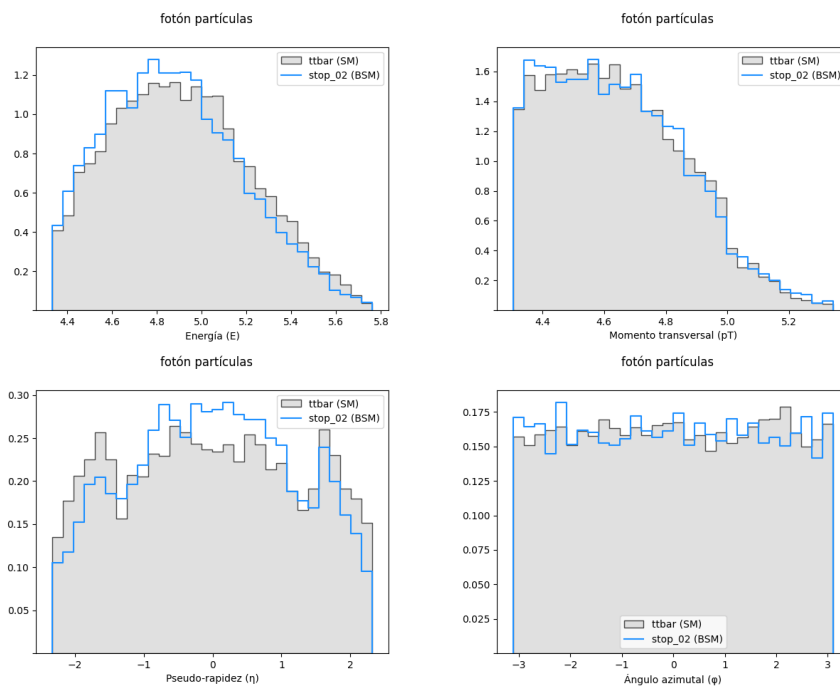


Figura B.8: Histogramas de las características de los fotones (g).

APÉNDICE C

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenible	Alto	Medio	Bajo	No procede
ODS 1. Fin de la pobreza.				X
ODS 2. Hambre cero.				X
ODS 3. Salud y bienestar.				X
ODS 4. Educación de calidad.				X
ODS 5. Igualdad de género.				X
ODS 6. Agua limpia y saneamiento.				X
ODS 7. Energía asequible y no contaminante.	X			
ODS 8. Trabajo decente y crecimiento económico.				X
ODS 9. Industria, innovación e infraestructuras.	X			
ODS 10. Reducción de las desigualdades.				X
ODS 11. Ciudades y comunidades sostenibles.				X
ODS 12. Producción y consumo responsables.				X
ODS 13. Acción por el clima.		X		
ODS 14. Vida submarina.				X
ODS 15. Vida de ecosistemas terrestres.				X
ODS 16. Paz, justicia e instituciones sólidas.				X
ODS 17. Alianzas para lograr objetivos.	X			



Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

El trabajo tiene una relación directa y significativa con varios Objetivos de Desarrollo Sostenible (ODS) establecidos por las Naciones Unidas en su Agenda 2030. Estos ODS abarcan una amplia gama de áreas, desde la protección del medio ambiente hasta el fomento de la innovación y la igualdad.

En primer lugar, el proyecto está fuertemente alineado con el ODS 7: Energía Asequible y No Contaminante. Aunque el trabajo de generación de eventos de colisión de partículas requiere una cantidad significativa de energía, es esencial destacar que el lugar donde se desarrolla gran parte de dicho trabajo, el CERN, obtiene la mayoría de su electricidad de fuentes con bajas emisiones de carbono. Además, el proyecto contribuye a la eficiencia energética al proponer un enfoque de generación de eventos más rápido y con menor consumo de recursos computacionales, lo que en última instancia conduce a una reducción de la energía requerida en comparación con los métodos tradicionales.

El ODS 9: Industria, Innovación e Infraestructura también está estrechamente relacionado con el proyecto. El trabajo no solo implica la implementación de tecnologías innovadoras, como las GANs, sino que también tiene la intención de mejorar los métodos utilizados en los experimentos realizados en el detector ATLAS en el CERN. Al optimizar la generación de eventos de colisión, el proyecto contribuye al progreso de la investigación en física de partículas y puede tener un impacto positivo en la industria y la innovación al validar nuevos modelos de física que podrían revolucionar diversas áreas.

El ODS 13: Acción por el Clima también es de gran relevancia para este proyecto. Aunque la naturaleza computacionalmente intensiva del proceso de generación de eventos puede consumir energía, la implementación de modelos más eficientes y rápidos podría reducir la demanda energética a largo plazo. Al reducir el tiempo requerido para la generación de eventos, se disminuye la duración de las operaciones de las computadoras, lo que conlleva un menor consumo de energía. Esta eficiencia contribuye indirectamente a la acción climática al reducir las emisiones de gases de efecto invernadero asociadas a la energía.

Por último, el ODS 17: Alianzas para lograr los Objetivos también se ve reflejado en el proyecto. El trabajo colaborativo con el Instituto de Física Corpuscular (Centro Mixto CSIC-Universitat de València) ilustra la cooperación entre diferentes entidades para lograr un objetivo común: mejorar la simulación de datos de física de partículas utilizando técnicas de Aprendizaje Automático. Esta colaboración cruza las fronteras disciplinarias y promueve la sinergia entre la física de partículas y la ciencia de datos, con la intención de alcanzar resultados más precisos y eficientes.