Medición de la producción de pares de jets a 13 TeV y uso de redes neuronales adversarias para identificar jets masivos

Marceca, Gino

2019

Medición de la producción de pares de jets a 13 TeV y uso de redes neuronales adversarias para identificar jets masivos

Resumen

En esta tesis se presenta la medición de la sección eficaz de masa invariante entre pares de jets producidos en colisiones protón-protón a una energía de centro de masa de 13 TeV. Los datos fueron colectados por el detector ATLAS en el Gran Colisionador de Hadrones del Laboratorio CERN durante el año 2015. Las mediciones de la sección eficaz fueron comparados cuantitativamente con las predicciones teóricas del Modelo Estándar (ME) a orden siguiente al dominante (NLO) corregidas por efectos no pertubativos. Estos estudios ponen a prueba el ME por posible evidencia de nueva física. A su vez, la producción de jets contiene información acerca de la distribución partónica dentro del protón y de la constante de acoplamiento fuerte α_s, lo que permite mejorar el conocimiento de QCD a una escala de energía nunca antes alcanzada. Los resultados experimentales mostraron estar en acuerdo con las predicciones teóricas, validando por primera vez el ME a 13 TeV en lo que respecta a la producción de pares de jets.

Por otro lado, se describe la aplicación de técnicas de aprendizaje automático para la identificación de jets provenientes de quarks top o bosones W en medio de un fondo dominante de jets de QCD. Dicha tarea resulta fundamental para la búsqueda de partículas masivas producto de nueva física o para la mejora de precisión de las propiedades del Higgs. El modelo propuesto en esta tesis esta basado en redes neuronales adversarias, el cual permite lograr un clasificador no correlacionado con ciertos observables físicos de interés, como la masa del jet. Ciertas búsquedas de nueva física en ATLAS son sensibles a efectos no deseados introducidos por la correlación, inherente para un poder de clasificación optimo, entre la variable discriminante y la masa del jet, lo que resulta en la reducción de la significancia estadística del analisis. Esta técnica propone una solución a este tipo de problemas. Estudios basados en simulaciones Monte Carlo muestran mejoras significativas respecto a otras métodos analíticos y de multivariable utilizados tradicionalmente en ATLAS, resultando por lo tanto prometedor para futuras búsquedas de nueva física.

Palabras clave: ATLAS, LHC, Jets, Sección eficaz, Redes neuronales

Measurement of the di-jet production at 13 TeV and use of adversarial neural-networks to identify heavy jets

Abstract

This thesis presents the measurement of the dijet cross-section from proton-proton collisions at a center-of-mass energy of 13 TeV. The data were collected by the ATLAS detector from the Large Hadron Collider at CERN during 2015. The measurement of the cross-sections were compared quantitatively with the Standard Model (SM) theoretical predictions at next-to-leading order (NLO) corrected by non-perturbative effects. These studies test the SM for possible evidence of new physics. In addition, jets production contains information on the distributions of partons inside the proton and also on the strength of their interaction α_s, which allows to extend the knowledge of QCD at an unprecedented energy scale. The experimental results shown to be in agreement with the theoretical predictions, validating for the first time the SM at 13 TeV regarding dijet production.

On the other hand, machine learning techniques used for jets identification coming from quarks top or W bosons in a dominated QCD background are described. This task results essential for heavy particles searches beyond SM or to improve the precision of Higgs properties measurements. The technique proposed in this thesis is based on adversarial neural-networks, which allows to build a non-correlated classifier with some physical observables of interest, as the jet mass. Some new physics searches in ATLAS are sensitive to undesirable effects introduced by these correlations, needed to achieve a good classification power, between the discriminant observable and the jet mass, which results in a reduction of the statistical significance in the analysis. This method proposes a solution to this kind of issues. Studies based in Monte Carlo simulations showed significant improvements with respect to other analytical and multivariate techniques used in ATLAS, providing an encouraging first look for the application in future physics searches.

Keywords: ATLAS, LHC, Jets, Cross-section, Neural-Networks

Contents

Introducción (resumen)

El 20 de Mayo del 2015, el detector ATLAS colecto por primera vez datos de colisiones proton-proton a una energia de centro de masa de 13 TeV generadas por el Gran Colisionador de Hadrones (LHC). Mediciones de la sección eficaz de producción de jets resulta fundamental para extender el conocimiento de QCD a escalas de energía nunca antes exploradas. En particular, la producción de pares de jets permite brindar información acerca de la estructura interna partónica del protón y de la constante de acoplamiento fuerte α_s.

Esta tesis presenta, por primera vez, mediciones de la sección eficaz de pares de jets a 13 TeV. Dicha medición es comparada cuantitativamente con las predicciones teóricas del Modelo Estándar (ME) a orden siguiente dominante (NLO) corregidas por efectos no pertubativos.

Jets son también importantes para la búsqueda de resonancias de parículas masivas producto de nueva física. En particular, los jets que provienen de una partícula pesada como quark top o bsoón W. Resulta fundamental por ende la identificación de este tipo de jets en fondos de QCD dominantes. En esta tesis se presenta una técnica novedosa basada en redes neuronales adversarias para lograr un clasificador no correlacionado con ciertos observables físicos, como la masa del jet. Esto muestra una mejora en el rendimiento con respecto a otras técnicas utilizadas en ATLAS hasta le fecha, resultando por ende prometedor para el uso en futuros analisis de física.

Chapter 1

Introduction

On 20 May 2015, the ATLAS detector recorded for the first time proton-proton (pp) collisions at a centre of mass energy of 13 TeV delivered by the Large Hadron Collider (LHC). The proton collisions set a new energy record, overcoming the past Run 1 at 8 TeV in 2012, and so extending the exploration of quantum chromodynamics (QCD) at scales never reached before. Precision measurements of strong interactions are crucial in understanding physics at hadron colliders, in addition, QCD provides one of the main backgrounds to many New Physics searches; furthermore, it is also through tests of QCD that New Physics may be discovered.

The observable signatures of quarks and gluons produced at the LHC are hadronic particle showers called "jets". By measuring jets in particle detectors, one can make predictions on the nature of the original parton-parton interaction which produced the signature observed in the detector and test QCD theory.

Jets are used for a wide range of physics analyses. One way of classifying their uses is according to the different possible origins for the partons that give rise to the jets. At hadron colliders, one of the best studied jet observables is the dijet spectrum, related to the high-momentum-transfer $2 \rightarrow 2$ scattering of partons inside the colliding protons. In this kind of process the energy of the jet (in the partonic centre-of-mass frame) is closely related to that of the parton in the proton that underwent a hard scattering and the dijet spectrum therefore contains information on the distributions of partons inside the proton and also on the strength of their interaction α_s, e.g. refs. [19, 26, 6, 21].

This thesis presents, for the first time, measurements of the dijet cross-sections in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$ centre-of-mass energy by the ATLAS Collaboration at the LHC, using data collected in 2015 and corresponding to an integrated luminosity of $3.2\,\text{fb}^{-1}$. Next-to-leading-order (NLO) perturbative QCD (pQCD) predictions calculated using several parton distribution function (PDF) sets, corrected for electroweak and non-perturbative effects, are quantitatively compared to the measurement results.

Another origin for the partons that lead to jets come from the hadronic decay of a heavy particle, for example a top quark, a W boson, a Higgs boson, or some other yet-to-be discovered resonance. The identification of jets coming from a top quark or W boson against the large background QCD is fundamentally important to searches for new physics. We present a novel technique using neural-networks in order to improve the task of jet classification coming from a two-prong decay.

In order to compare the cross-section measurements with the theory, different jet calibration procedures need to be made. We describe briefly the different calibration steps done on the jet observables, focusing particularly on the jet energy scale calibration based on Monte Carlo simulation (MC-JES) which was one of the main task developed during the doctorate.

The thesis is organized as follows: Chapter 2 describes the theoretical framework, with emphasis in the theory of the strong interactions and the aspects that are important for the understanding of the hadronic final states in hadronic collisions. The LHC and the ATLAS detector are described in Chapter 3, as well as the experimental conditions during the 2015 data taking. Chapter 4 is about how jet reconstruction and calibration are performed in ATLAS, focusing particularly on the MC-JES calibration developed during the Ph.D. On Chapter 5 and 6 the two analyses involved as a main contributor are presented. The first one is about the measurement of the dijet cross-sections at 13 TeV. Here, qualitative and quantitative comparisons between data and theory are shown, also an explanation of each experimental systematic uncertainty propagated to the measurement, and the diverses techniques used for getting the final result. The second one, relates to performance studies using neural-networks for jet classification. Finally, Chapter 7 presents the conclusions.

Aspectos teóricos (resumen)

En este capítulo se presenta un breve resumen del Modelo Estándar (ME) y de las predicciones de QCD. A su vez, se describen las técnicas de Monte Carlo utilizadas para las simulaciones de las colisiones protón-protón y la interacción con el detector ATLAS.

El ME es una teoría cuántica de campos que describe el comportamiento de las paríoulas experimentalmente observadas bajo la interacción de la fuerza electromagnética (EM), débil y fuerte. El modelo esta descripto por el grupo de simetría de Poincare (homgeneidad, isotropía y covarianza) y simetrías de gauge local $SU(3) \times SU(2) \times U(1)$. Todo el modelo de simetrias e interacciones esta descripto por una densidad Lagrangiana, la cuál describe 3 familias de quarks, 3 familias de leptones, 4 intermediarios de las 3 fuerzas fundamentales (EM, debil y fuerte) y el boson de Higgs, descubierto en 2012 por ATLAS y CMS.

Debido a que las colisiones en el LHC son hadrónicas (formadas por muchos partones), se necesita de un método para extender las predicciones de nivel partónico a nivel hadrónico. El teorema de factorización descripto por la ecuación 2.1 permite la realización de tal propósito, indispensable para comparar las predicciones teóricas con las mediciones experimentales. Esta relación se utiliza para poder ajustar con un fit la dependencia de las distribuciones partónicas (PDFs) en función de la fracción de momento de los partones dentro del protón x y la escala de energía Q^2. De esta manera, se pueden mejorar las funciones de estructura del protón, las culaes no pueden ser predichas de primeros principios.

Simulaciones Monte Carlo basadas en lluvia partónicas son fundamentales para comparar con los observables fisicos medidos y reconstruidos en el detector. Dichos generadores parten de la simulación a nivel de elemento de matriz (LO o NLO) y radican en una lluvia partónica para facilitar la simulación del proceso de hadronización permitiendo obtener las partículas observables estables por un detector. El generador mas utilizado por ATLAS es `Pythia`, y dado el buen acuerdo que presenta con los datos reales, fue el elegido por defecto en esta tesis.

Una vez generados las partículas estables del evento, se simula la interacción con la materia resultando en el mismo tipo de señales registradas por el detector ATLAS. Para tal propósito, se utiliza el software `Geant4`, el cual es un programa flexible disenñado para simular geometrías complejas en entornos complejos como es el caso del detector ATLAS.

Chapter 2

Theoretical aspects

This chapter presents an overview of the Standard Model and theoretical QCD predictions, Sections 2.0.1 and 2.0.2. The Monte Carlo tools used to simulate the QCD processes and the interaction with the detector are discussed in Sections 2.0.3 and 2.0.4.

2.0.1 The Standard Model

The Standard Model (SM) is a quantum field theory that describes the behavior of all experimentally-observed particles under the influence of the electromagnetic, weak and strong forces. Mathematically speaking, its formulation is based on group theory that obeys the global Poincare symmetry postulated for all relativistic quantum field theories (translational, rotational symmetries and the inertial reference frame invariance), in addition to the local $SU(3) \times SU(2) \times U(1)$ gauge symmetry (internal symmetry). Roughly, the three factors of the gauge symmetry give rise to the three fundamental interactions, described as an exchange of particles (gauge bosons) between the objects affected, such as the gluon for the strong interaction, described by *Quantum Chromodynamics* (QCD), the W and Z bosons for the unified electroweak force, and the photon for the electromagnetic force, modeled by *Quantum Electrodynamics* (QED), respectively. These particles are known as the force carriers or messenger particles and all have spin of 1, therefore, all known gauge bosons are vector bosons. In addition to gauge bosons, there is a scalar boson called the Higgs field. The quantum excitation of this field (the Higgs particle) was predicted to exist in the 1960s and confirmed experimentally for the first time in 2012 by the ATLAS and CMS experiments at the LHC [61]. The confirmation of the existence of the Higgs field completed the last missing piece of the SM. Unlike other known fields such as the electromagnetic field, it has a non-zero expectation value in the vacuum and explains why some fundamental particles have mass, despite that the symmetries controlling their interactions impose that they should be massless. It also resolves several other long-standing puzzles, such as the reason for the extremely short range of the weak force [70]. All this model based on symmetries and interactions is totally contained in terms of a *Lagrangian density*.

The fundamental building blocks of matter described by the SM are fermions with spin 1/2:

- six leptons (and their antiparticles), organized in three families,

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix} \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix} \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}$$

- and six quarks (and their antiparticles), organized in three families,

$$\begin{pmatrix} u \\ d \end{pmatrix} \begin{pmatrix} c \\ s \end{pmatrix} \begin{pmatrix} t \\ b \end{pmatrix}$$

The three families are ordered from lower (1st generation) to higher (3er generation) mass values. These particles are considered point-like, as there is no evidence of any internal structure of leptons or quarks (q) to date. The six types of quarks are also known as the six quark flavors. Collectively, the u (up), d (down), and s (strange) quarks are frequently referred to as light quarks. The heaviest quark of the Standard Model, the quark t (top), was the last to be found [60, 63]. The electric charge Q of quarks take fractional values, i.e. $+2/3$ for quarks u, c and t and -1/3 for quarks d, s and b; yet they are only observed as the integer charge combinations of three quarks (baryons) or a quark and an antiquark (mesons) [1]. Quarks and gluons are collectively called partons and are the fundamental actors of the strong interactions.

Figure 2.1 shows the elemental table which summarizes the SM.

2.0.2 Theoretical QCD Predictions

Renormalization of fields appearing in the SM Lagrangian is a necessary technique which allows to treat infinities arising from mommentum space integral divergences in computing Feynman diagrams. Fortunately, the SM, in particular QCD, as in any Gauge theory, is *renormalizable*, in the sense that all divergences, particularly the *ultraviolet* ones, can be absorbed into redefinition of few physical parameters. Because of renormalization and also due to the *non-Abelian* property of QCD, the coupling constant of the strong interaction, α_s, decreases as the interaction energy increases, called the *running* of the coupling. As a result, in high-energy interactions, which means at short lenghts ($\lesssim$ size of the proton), the coupling constant in

[1]Although stable particles are regularly made of 3 quarks (baryons) or a quark and an antiquark (mesons), evidence for exotic baryons consisting of 5 quarks has been observed by the LHCb Collaboration. [16]

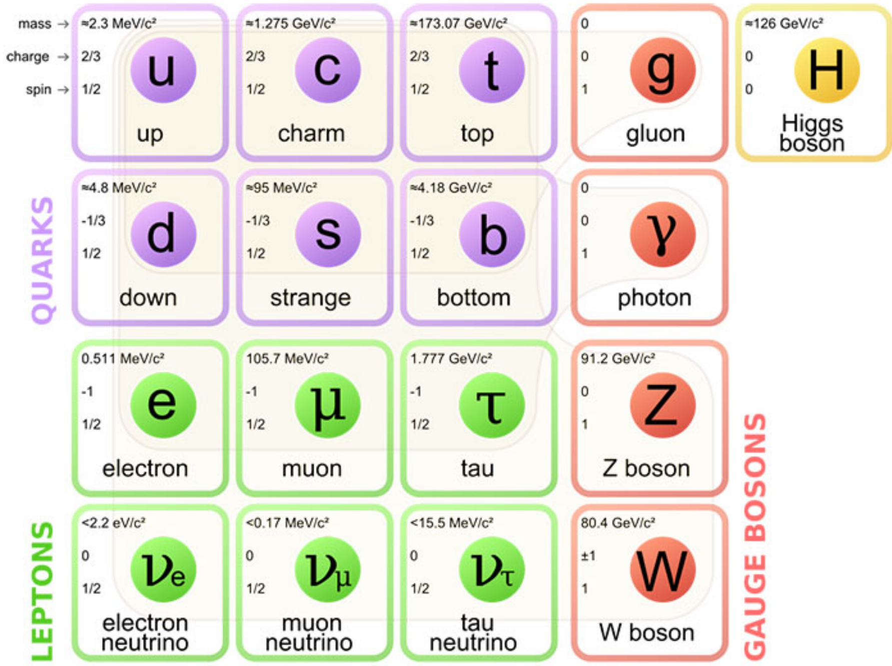


Figure 2.1: Elemental table of particles of the SM.

QCD becomes sufficiently small to warrant a perturbative series expansion. This is known as *asymptotic freedom* and makes quantitative predictions for the strong interaction possible in particular at the energy scale probed by hadron colliders. On the contrary, partons are coupled together more strongly as the distance between them increases. This effect, known as *confinement*, "explains" why quarks and gluons are not visible in their own right, but become trapped into color-neutral hadrons (baryons or mesons), known as *hadronization*, which is a non-perturbative effect and so cannot be computed from first principles.

In perturbative QCD (pQCD), each order of the expansion corresponds to an additional power in the coupling constant. By considering all possible Feynman diagrams for a given order of perturbation theory, all the terms in the calculation summed together give a definite *cross-section* at a fixed-order. In this context, leading-order diagrams (LO) are the first ones which appears in the perturbation series, without internal loops. The complexity of the process determines the precision of the calculation that has to be performed. For inclusive parton production, calculations are typically performed at next-to-leading order (NLO), although NNLO QCD predictions have recently been performed for the dijet cross-sections at 7 TeV [54].

Due to the fact that the LHC does not produce simple parton-parton interactions, but instead collisions of hadrons, one may wish a tool to extend perturbative calculations for parton interactions to proton-proton collisions. Fortunately, the factorization theorem [94] allows us to account for this. This theorem expresses the total cross section of two hadrons interaction as a weighted combination of parton cross-sections interaction. The weighting factors $f_i(x)$ are the parton distribution functions (PDFs), where $f_i(x)dx$ gives the differential probability of finding a parton of type i carrying a fraction of the total hadron momentum between x and $x + dx$. The hadronic cross-section, at some energy Q^2 that characterizes the interaction, can be written as:

$$\sigma(P_1, P_2) = \sum_{i,j} \int dx_1 dx_2 f_i(x_1, \mu_f^2) f_j(x_2, \mu_f^2) \hat{\sigma}_{ij}(p_1, p_2, \alpha_s(\mu_r^2), Q^2/\mu_r^2, Q^2/\mu_f^2). \tag{2.1}$$

Here, P_1 and P_2 are the momenta of the two incoming hadrons, x_1 and x_2 are the momentum fractions carried by the interacting partons, and $p_1 = x_1P_1$, $p_2 = x_2P_2$ are the interacting parton momenta. The partonic cross section $\hat{\sigma}_{ij}$, corresponding to the interaction of partons i and j, is calculated at a fixed order in α_s, which is evaluated at some renormalization scale, μ_R. The total cross section is obtained by summing over all possible parton flavors and integrating over all possible momentum fractions. The parton distribution functions, f_i and f_j, are evaluated at a factorization scale, μ_F, which can be thought of as the scale that separates short-distance, perturbative physics, from long-distance, non-perturbative physics. They are usually both taken to be equal, $\mu_F = \mu_R = \mu$, chosen at the typical scale Q^2 of the process, in order to minimize the contribution of (uncalculated) higher order terms. In the limit of a complete perturbation expansion, the hadronic cross-section would be exact and would not depend on μ_F, μ_R. Such dependence is assigned as a theoretical uncertainty at a given order expansion. The Figure 2.2 shows the dijet cross-section as a function of μ_R for different μ_F values. As it can be seen, both dependences decrease as the precision of the calculation increases, which results in a significant reduction of the theoretical uncertainty of NNLO with respect to NLO [54].

The NLO predictions used in this thesis were calculated using `NLOJET++` 4.1.3 [82] for various NLO PDF sets and various values of the renormalisation and factorization scales.

In order to be able to compare with the measurement, perturbative calculations need to be corrected by non-perturbative (NP) and electroweak (EW) effects. NP contributions account for hadronization or underlying event (UE), additional parton interactions which are not involved in the hard scattering process of two protons.

Precise measurements of hard scattering cross-sections are used in this way not only to test predictions but also to fit (in case of agreement between data and theory is found) the dependence of the PDFs on x and Q^2 [21], which can be obtained only through experimental results. Cross-section measurements can also be used to determine the strength of the interaction α_s [6]. Both exceed the scope of the studies developed in this thesis.

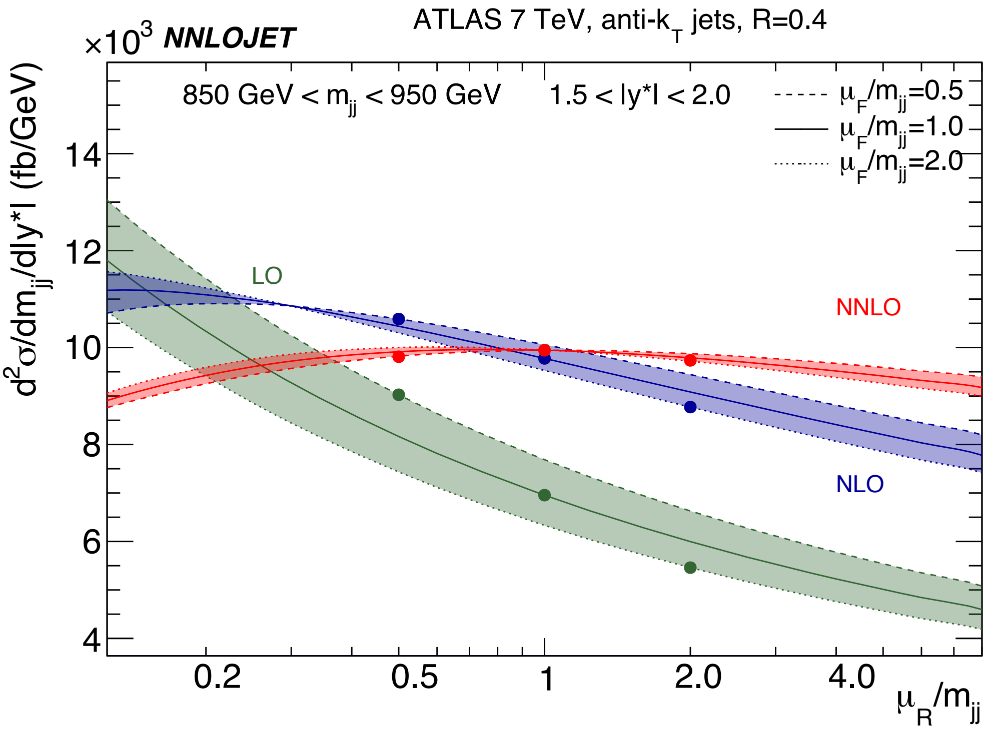


Figure 2.2: The dijet cross section evaluated at a specific kinematic range as a function of μ_R/m_{jj} LO (green), NLO (blue) and NNLO (red). The thickness of the bands shows the variation due to factorization scale.

2.0.3 Monte Carlo Simulation

When two partons collide, they can undergo a large angle ("hard") scattering generating two (or more) final-state partons at large momemtum. Partons at these states tend to radiate more partons nearly collinear with its parents. As the *parton shower* evolves they loose momentum and the α_s coupling becomes higher up to the non-perturbative QCD regime is reached. As a result, due to confinement partons hadronize, leading to a "spray" of colorless hadrons forming a cone structure that we label a *jet*.

Parton showering of jet formation exits the region of validity for pQCD, and must be therefore described by phenomenological models, to account for the length of time when showering occurs and the combination of colored partons into bound states of colorless hadrons. One example is the Lund string model [92], which is the one implemented in ATLAS *event generators*.

Event generators are based in Monte Carlos (MC) calculations and use LO (NLO) perturbative calculations of matrix elements for $2 \to 2$ ($2 \to 3$) processes, relying on the parton shower to produce the equivalent colorless-particle final state. The most common used parton shower generator is `Pythia`. This generator utilizes LO pQCD matrix elements, along with a leading-logarithmic parton shower, an UE simulation with multiple parton interactions, and the Lund string model for hadronisation. The nominal Pythia samples used in this thesis were produced using a set of tuned parameters called the A14 tune [49] and the NNPDF 2.3 LO PDF [37] set. Given the good agreement between data and Pythia, and because of the vast use through the Collaboration, this generator was used as the default for both the measurement and the neural-networks performance studies. Other general purpose MC generators were used as well in the measurement such as `Herwig++` [36], `Powheg` and `Sherpa` [90]. The last two are computed at NLO level in the matrix element. The specific use of each MC will be described throughout each chapter.

The MC samples are overlaid with additional pythia simulated *pp* collisions (pile-up) to roughly match the data pile-up conditions with a mean number of 24 collisions per bunch crossing for the 2015 data-taking. A formal definition of pile-up is shown in Sec. 3.0.7.

2.0.4 Detector simulation

To have simulated events at the same level that one might observed with the detector, the passage of particles through matter and digitalisation must be simulated. The particles output from event generators is passed through a detector simulation model, resulting in the same physical signals one might observe in real data. The detailed simulation of the ATLAS detector is based on the `Geant4` toolkit [22]. `Geant4` is a flexible program designed to handle complex geometries and enviroments. It simulates the magnetic fields and all the particle interaction processes with matter such as ionization, bremsstrahlung, photon conversions, multiple scattering, scintillation, absorption and transition radiation. The software is easily adaptable to different conditions at a given time and so it contains updated information regarding dead channels and misalignments.

Finally, the simulated events are processed using the same reconstruction algorithms and analysis filters as is used for real data.

El LHC y el detector ATLAS (resumen)

El Gran Colisionador de Hadrones (LHC) es el acelerador de partículas mas grande del mundo. Es un sincrotrón de protones ubicado en el Laboratorio CERN en la frontera Franco-Suiza (Ginebra). Esta ubicado 100 m bajo tierra y tiene un perímetro de 27 km. El mismo alberga varios detectores, entre ellos ATLAS, CMS, Alice y LHCb. Los dos primeros son detectores multipropósito, cuyo objetivo es buscar cualquier tipo de anomalia que haya en los datos. Por otro lado, Alice y LHCb son mas específicos y buscan comprender mejor el quark-gluon plasma y la naturaleza de los quarks b, respectivamente.

Los protones, previos a ser inyectados en el LHC, son generados por una botella de hidrógeno y separados mediante un campo eléctrico. Estos son acelerados por diversos aceleradores de menor energía hasta finalmente ser inyectado al anillo central del LHC.

El LHC colisiona protones a una frecuencia de 40 MHz. El numero de eventos total generados es proporcional a la sección eficaz del proceso (dependiente de la naturaleza) y a la luminosidad integrada en el tiempo (dependiente de las condiciones experimentales). Un efecto no deseado inherente del proceso de generación de datos es el denominado *pile-up*. Este consiste en múltiples colisiones protón-protón dentro de un mismo choque (*in-time* pile-up) y efectos introducidos por cruces vecinos (*out-of-time* pile-up). Estos efectos ocasionan pérdida de precisión en las mediciones y resulta un desafío experimental desarrollar técnicas que atenuen dicho fenómeno.

El detector ATLAS esta formado por varios sub-detectores. En orden de cercania a la zona de interacción se encuentran: el detector de trazas, el cual esta formado por un detector de pixeles de alta granularidad, seguido por discos de microfibra de silicio (SCT) y complementado por un detector de radiación (TRT). Estos se encargan de medir la curvatura y por ende el momento de las trazas producidas por partículas cargadas.

Luego del detector de trazas se encuentran los calorímetros electromagnético (em) y hadrónico (had). El calorímetro em se encarga de medir solo la energía depositada por fotones o electrones. Ambos pueden ser distinguidos al saber si dejaron o no una traza en el detector de trazas. Por otro lado, el calorímetro had mide la energía depositada por los hadrones, los cuales interactuan con la materia luego de pasar el calorímetro em.

Por último se encuentra el detector de muones, los cuales miden exclusivamente el momento de los muones.

Dada la alta frecuencia que opera el LHC (40 MHz), tecnológicamente no es posible almacenar todos los datos. Por tal motivo ATLAS cuenta con un sistema de trigger, el cual esta optimizado para de forma rápida decidir si un evento es o no interesante para ser almacenado. Este sistema funciona como un filtro de eventos, y reduce la frecuencia de almacenamiento de datos de 40 MHz a 1 KHz.

Chapter 3

The LHC and the ATLAS detector

3.0.5 The Large Hadron Collider

The Large Hadron Collider (LHC) [40] is the world's largest and most powerful particle accelerator. It is a proton-proton (pp) synchrotron located at CERN Laboratory, just outside the city of Geneva (Switzerland), approximately 100 m underground. It is designed to collide bunches of up to $\sim 10^{11}$ protons every 25 ns at a center-of-mass-energy of $\sqrt{s} = 14$ TeV. However, in order to not delay the delivery of particle collisions for physics research for the Run 2 period, scheduled for the beginning of 2015, the decision was to operate at 13 TeV [1]. The experiments analyzing the collisions produced by the LHC are distributed around the 27 km ring mostly opposite each other. The ATLAS experiment is closest to the main CERN site and is located opposite to the other general purpose detector, CMS. ALICE and LHCb experiments are specific purpose detectors. The former is designed to investigate heavy ion collisions; the latter, to investigate rare decays of b-mesons. The layout of these four experiments along the LHC ring is shown in Figure 3.1.

3.0.6 The Long Journey of a Proton

Proton beams are formed, before insertion into the main LHC ring, using a succession of machines that accelerate particles to increasingly higher energies. Each machine boosts the energy of a beam of particles, before injecting the beam into the next machine in the sequence. In the LHC, the last element in this chain, particle beams are accelerated up to the record energy of 6.5 TeV per beam. Most of the other accelerators in the chain have their own experimental halls where beams are used for experiments at lower energies.

The proton source is a simple bottle of hydrogen gas. An electric field is used to strip hydrogen atoms of their electrons to yield protons. Linac 2, the first accelerator in the chain, accelerates the protons to the energy of 50 MeV. The beam is then injected into the Proton Synchrotron Booster (PSB), which accelerates the protons to 1.4 GeV, followed by the Proton Synchrotron (PS), which pushes the beam to 25 GeV. Protons are then sent to the Super Proton Synchrotron (SPS) where they are accelerated to 450 GeV.

The protons are finally transferred to the two beam pipes of the LHC. The beam in one pipe circulates clockwise while the beam in the other pipe circulates anticlockwise. It takes 4 minutes and 20 seconds to fill each LHC ring, and 20 minutes for the protons to reach their maximum energy of 6.5 TeV. Beams circulate for many hours inside the LHC beam pipes under normal operating conditions. The two beams are brought into collision inside the four detectors: ALICE, ATLAS, CMS and LHCb, where the total energy at the collision point is equal to 13 TeV.

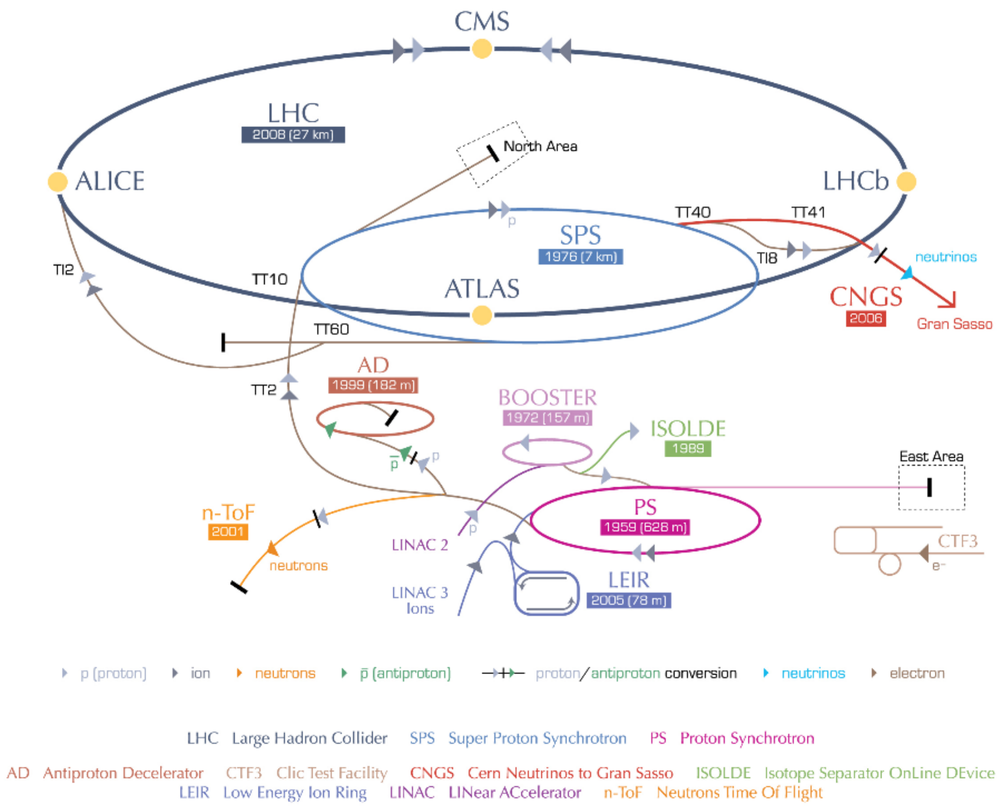


Figure 3.1: The CERN accelerator complex, showing the injection system, along with each component's date of construction, and the placement of the four main experiments.

3.0.7 Luminosity and pile-up

The rate of events produced by the colliding beams depends on the luminosity of the collisions, which is a measure of the number of events per second per unit cross section, typically measured in units of cm^{2}s^{-1}. The number of events (N) of a particular process, then, is given by the product of the integrated luminosity, $\int Ldt$, and the cross-section of the process, σ_{event}:

$$N = \sigma_{\text{event}} \int Ldt$$

The integrated luminosities are typically quoted in units of inverse picobarns, pb^{-1} = 10^{-36}cm^2. In order to measure processes with very little cross sections a very high luminosity is required.

The delivered luminosity can be written as [12]:

$$L = \frac{n_b f_r n_1 n_2}{2\pi \Sigma_x \Sigma_y}$$

where n_b is the number of colliding bunch pairs, n_1 and n_2 are the bunch populations (protons per bunch) in beam 1 and beam 2 respectively (together forming the bunch charged product), f_r is the machine revolution frequency, and Σ_x and Σ_y are the width and the height of the proton beams.

The number of protons per bunch, the number of bunches per beam, and the revolution frequency are all set by the beam operators. The widths of the proton beams are measured in a process known as a Van der Meer (vdM) scan [99]. In a vdM scan, the beams are separated by steps of a known distance. The collision rate is measured as a function of this separation, and the width of a gaussian fit to the distributions yields the width of the beams in the direction of the separation.

The total integrated luminosities provided by the LHC and recorded by ATLAS in 2015 are shown in Figure 3.2. These events form the dataset analyzed in this thesis. By means of the beam-separation or vdM scans, as well as other techniques to measure the bunch charged product, the ATLAS Collaboration has determined that the uncertainty on the 2015 luminosity measurement is $\delta L = 2.1\%$ [5].

Due to the cross-section for interaction and the large number of protons per bunch, the possibility to observe multiple pp interactions per bunch crossing increases proportionally. This phenomenon, referred to as "pile-up", can really occur in two distinct forms. The first form is the presence of multiple pp collisions (different from the interaction of interest) in the same bunch crossing, referred to as "in-time" pile-up. The second form of pile-up takes place due to electronic integration times within the detector. Certain detector components are actually sensitive to multiple bunch crossings due to the long electronic signals generated in the response to energy depositions or charge collection. One or more pp collisions in neighboring bunch-crossings can then affect the measurement. This form of pile-up is referred to as "out-of-time" pile-up. The total number of reconstructed primary vertices with 2 or more tracks, N_{PV}, provides a measure of in-time pile-up activity in each event. On the other hand, the mean number of interactions per crossing $\langle \mu \rangle$ is sensitive to both in-time and out-of-time pile-up, which corresponds to the mean of the poisson distribution of the number of interactions per crossing calculated for each bunch: $\mu = L_{\mathrm{bunch}} \times \sigma_{\mathrm{inel}}/f_r$, where L_{bunch} is the per bunch instantaneous luminosity, σ_{inel} is the inelastic cross section which we take to be 80 mb for 13 TeV collisions, and f_r is the LHC revolution frequency. The conditions for the 2015 data were $\langle \mu \rangle \sim 13.5$ and $N_{PV} \sim 20$.

3.0.8 The Atlas Detector

The Atlas (A Toroidal LHC Apparatus) detector of 7000 tones is one of the two general purpose particle detectors built for probing pp collisions at LHC. At 13 TeV center-of-mass energy, the LHC interactions result in high particle multiplicity, requiring fine detector granularity; and, particle production at small angles with respect to the interaction line, requiring large detector angular coverage. To achieve these performance goals, a design consisting of multiple

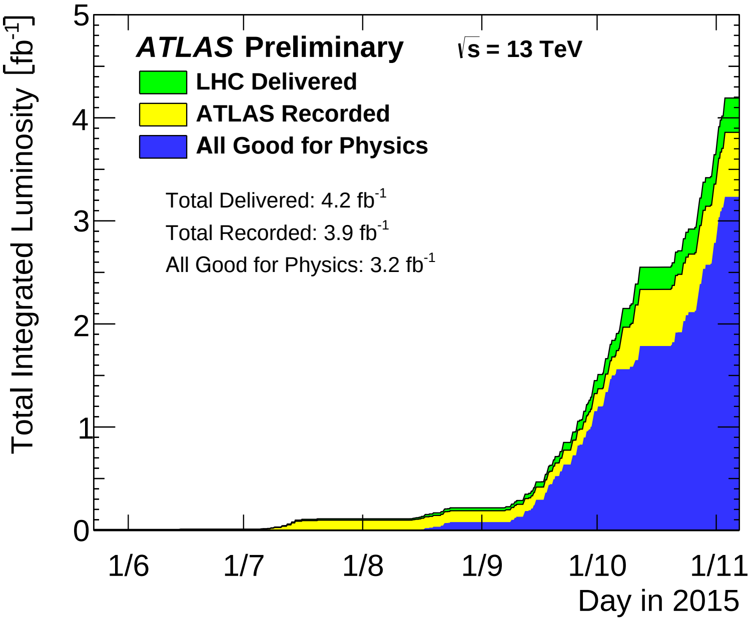


Figure 3.2: Cumulative luminosity versus time delivered to ATLAS (green), recorded by ATLAS (yellow), and certified to be good quality data (blue) during stable beams for pp collisions at 13 TeV centre-of-mass energy in 2015.

detector sub-systems with cylindrical symmetry around the incoming beams is used as shown in Figure 3.3. Closest to the interaction point the inner tracking detector is placed, providing charged particle reconstruction. The magnet configuration comprises a thin superconducting solenoid surrounding the inner detector cavity, and three large superconducting toroids (one barrel and two end-caps) arranged with an eight-fold azimuthal symmetry around the calorimeters. This fundamental choice has driven the design and size (44 m in length and 25 m in height) of the rest of the detector. Outside the solenoid, a calorimeter system performs electron, photon, tau, and jet (defined in Section 4) energy measurements. Finally, the calorimeter is surrounded by the muon spectrometer where an array of muon drift chambers perform muon identification and momentum measurements. A detailed description of its sub-systems can be found on its technical design report [50, 51].

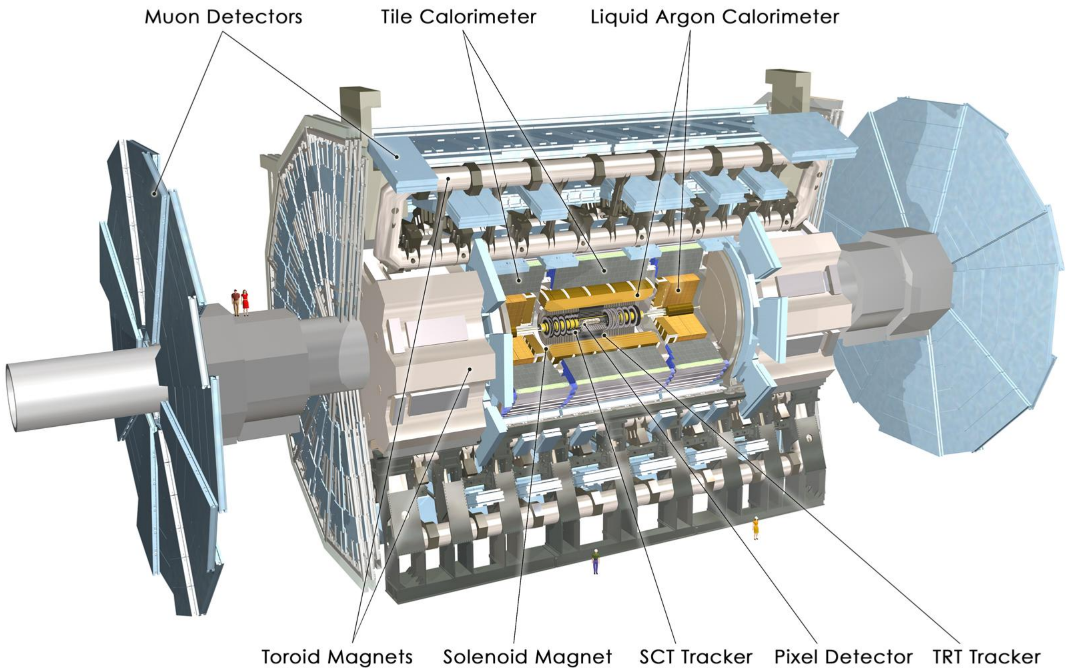


Figure 3.3: The ATLAS Detector.

The ATLAS detector coordinate system is used to describe the position of particles as they traverse these subdetectors. It is a right-handed coordinate system, with z pointing along the beam direction, positive x pointing toward the center of the LHC ring, and positive y pointing up. The $x - y$ plane is referred to as the transverse plane, and the z direction as the longitudinal direction. The azimuthal angle ϕ is measured as usual around the beam axis, and the polar angle θ is the angle from the beam axis. In experimental particle physics, the pseudorapidity η is commonly used as the polar angle instead of θ. They relates each other as following $\eta = -\ln(\tan\frac{\theta}{2})$. It can be shown that by doing a massless approximation ($m << |p|$), which is valid given our energy regime, η equals the rapidity $y \equiv \frac{1}{2}ln\big(\frac{E+p_L}{E-p_L}\big)$, where p_L is the momentum across the beam line. It can be seen that the rapidity y (or pseudorapidity in the massless approximation) are Lorentz invariant under boosts along the longitudinal axis. A measurement of a rapidity difference Δy between particles (or $\Delta\eta$) is hence not dependent on the longitudinal boost of the reference frame. This is an important feature, where the colliding partons carry different longitudinal momentum fractions, which means that the rest frames of the parton-parton collisions will have different longitudinal boosts. This justifies the use of η instead of θ as the polar angle representation.

Regions of low η are referred to as "central", and regions of high η are referred to as "forward". The transverse momentum p_{T} is defined as the momentum projected in the $x-y$ plane. The distance ΔR in the pseudorapidity-azimuthal angle space is defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

Inner Tracking System

The inner-detector system (ID), providing charged particle tracking, is immersed in a 2T axial magnetic field, generated by the central solenoid. It consists of high-granularity silicon pixel detector closest to the interaction point followed by the silicon microstrip tracker (SCT). These silicon detectors, covering the range $|\eta| < 2.5$, are complemented by the transition radiation tracker (TRT), which enables radially extended track reconstruction up to $|\eta| = 2.0$. The positions of the registered hits are combined to form tracks, with the radius of curvature of the tracks (caused by the presence of the magnetic field) providing a measurement of the particle's transverse momentum.

The pixel detector, SCT, and TRT sensors are arranged on concentric cylinders around the beam axis, known as barrel layers, and on disks perpendicular to the beam at either end of the barrel, known as end-caps. The overall layout of the inner detector is shown in Figure 3.4

One of the upgrades from Run 1 was the insertion of a new layer between the beam pipe (where particle collisions occur) and the pixel detector. This new sub-detector, known as the Insertable B-Layer (IBL), is located at 0.2 mm from the beam tube and 1.9 mm from the pixel. Thanks to this new fourth layer and dedicated algorithms developed, the performance for reconstructing charged particles has improved with respect to Run 1, resulting in particular in a better tagging efficiency for the decay of beauty quarks (b-tagging), crucial for physics analyses [8].

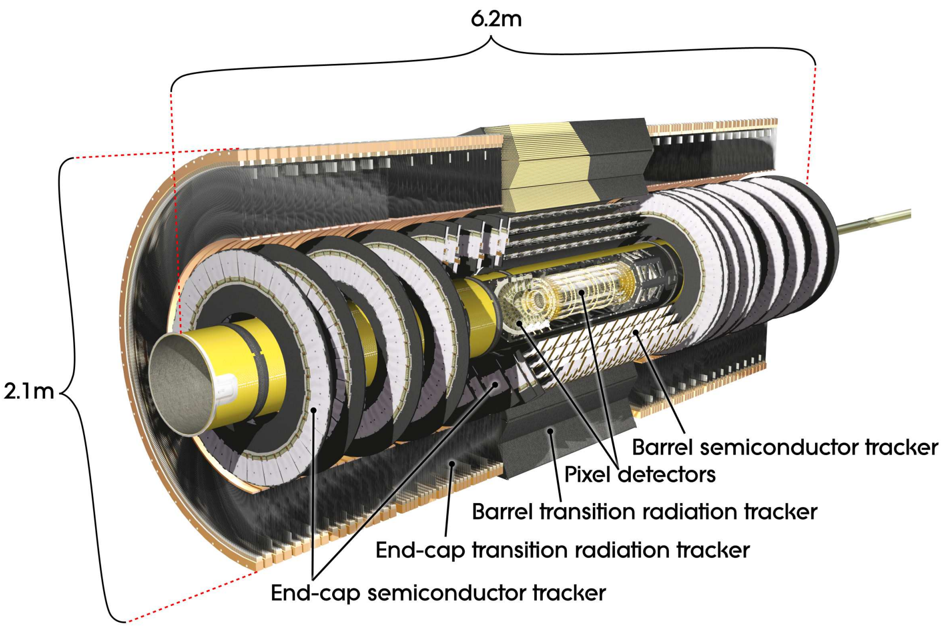


Figure 3.4: Layout of the ATLAS Inner Detector.

Calorimeter

For the studies presented in this thesis, the most important subsystem is the calorimetry, which covers the pseudorapidity range $|\eta| < 4.9$. It consists of an electromagnetic calorimetry (EM) which covers the region $|\eta| < 3.2$, provided by a barrel and endcap high-granularity lead/liquid-argon (LAr) detector, with an additional thin LAr presampler covering $|\eta| < 1.8$ to correct for energy loss in material upstream of the calorimeters. It only measures electrons/positrons and photons. Next to the EM calorimeter, it follows the Hadronic calorimeter, which measures the energy of hadronic particles and partially of muons. It is made of steel/scintillating-tile detectors, segmented into three barrel structures within $|\eta| < 1.7$, and two copper/LAr hadronic endcap calorimeters.

The EM and hadronic calorimeters are sampling calorimeters meaning that they utilize alternating layers of absorber material, composed of heavy atoms that interact with energetic particles and cause them to loose energy, and an active material, that produces a electrical signal in response to the deposited energy.

The calorimeters closest to the beam-line are housed in three cryostats, one barrel and two end-caps. The barrel cryostat contains the EM barrel calorimeter, and the two end-cap cryostats each contain an EM end-cap calorimeter (EMEC), a hadronic end-cap calorimeter (HEC), located behind the EMEC, and a forward calorimeter (FCal) to cover the region closest to the beam. These calorimeters use liquid argon as the active detector medium and need to be mantained at a constant temperature of $\sim$ 88K. Liquid argon has been chosen for its inertness property, which results in linear behaviour (production of ionization charge as a function of incident charge), stability of response over time and intrinsic radiation-hardness. An illustration of all these components can be found in Figures 3.5 and 3.6.

Each calorimeter is segmented both transverse to the particle direction, to give position information, and along the particle direction, to chart the development of the particle shower. This permits detailed mapping of EM and hadronic showers in the calorimeter, allowing for studies of the internal structure of hadronic jets and partially giving rise to the high resolution measurements of their energy.

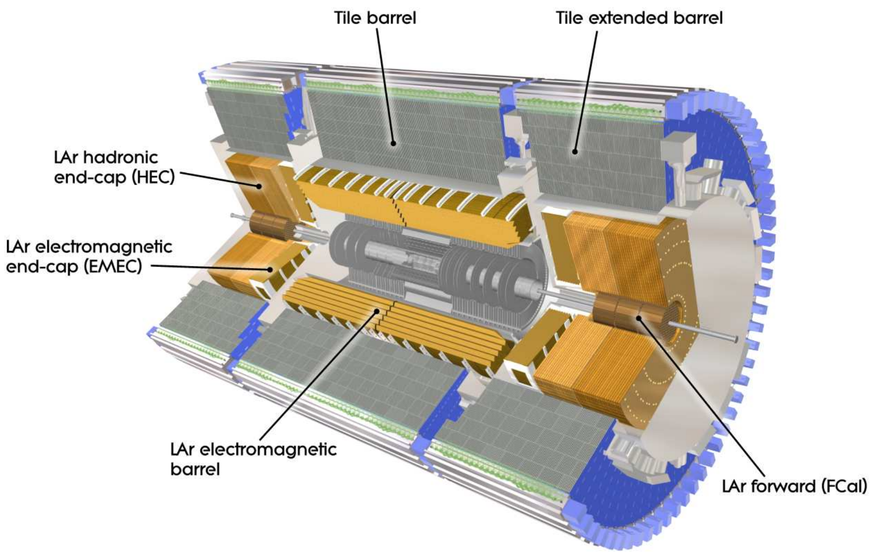


Figure 3.5: Layout of the ATLAS electromagnetic and hadronic calorimeter systems. The total length is ~ 12 m, extending to a maximum radius of 4.25 m.

The Muon Spectrometer

The muon system gives the ATLAS detector its overall shape and imposing nature, as depicted in Figure 3.7. Because of muons have much smaller cross section to interact in material than electrons and hadrons, they do not deposit all their energy in the calorimeters. The muon spectrometer was designed to detect muons within $|\eta| < 2.7$. Given that the spectrometer is located outside the calorimeter, it is expected that only muon particles will interact with the detector.

To provide a momentum measurement, the muons trajectories are bent in a toroidal magnetic field. This field is provided by one large barrel toroid and two large end-cap toroids, each toroid consisting of eight coils arranged symmetrically around the beam axis. The toroid system produces a magnetic field that is typically oriented in the ϕ direction and that is measured

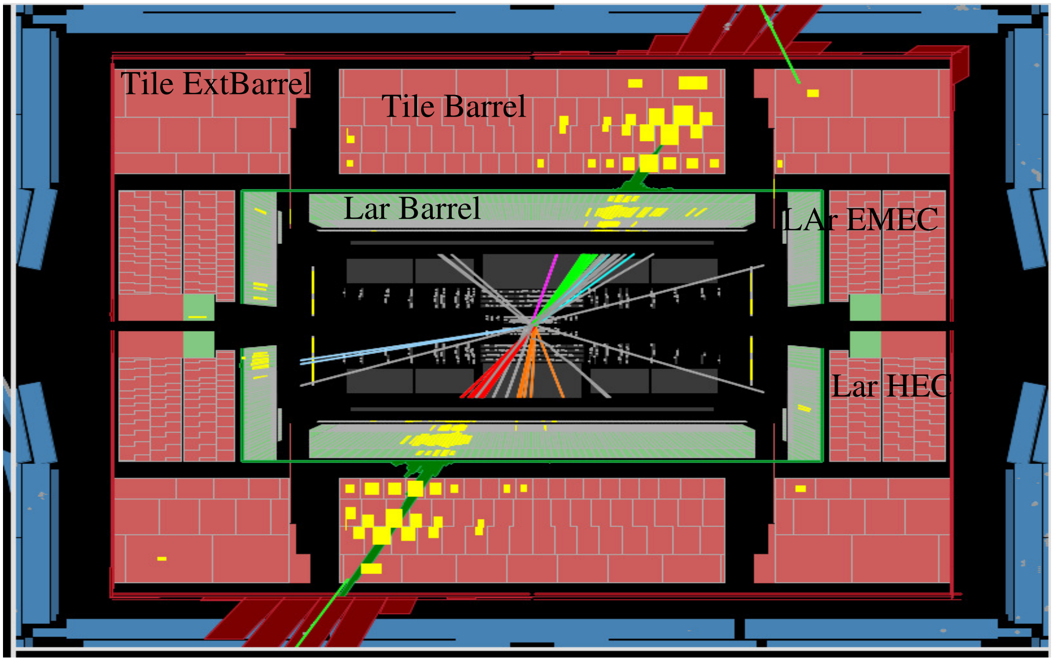


Figure 3.6: Diagram of the ATLAS detector together with a collision example in the plane $z - \eta$. Tracks are observed in different colours in the ID. The energy depositions (yellow) are observed in the em (green) and had (red) calorimeters.

with over 1800 Hall sensors placed throught the magnets. Under the influence of this field, muons are deflected in the $z - \eta$ plane and the transverse momentum of the muons is given then by the radius of curvature of the tracks. The trajectory is reconstructed by measuring the time of arrival of an electron cascade produced in drift tube chambers. These consist in cylindrical tubes (cathode) filled with a reactive gas and an anode wire in the middle along its direction. The time of arrival of electrons generated in the gas to the anode allow to reconstruct a spherical surface area where the muon might have passed. A combination of these tubes make it possible to reconstruct the exact trajectory of the muon. An example of these chambers are the *Monitored Drift Tubes* (MDT), which provides coverage for the more central region.

3.0.9 The Trigger System

At design luminoisty, the LHC will deliver approximately 40 million collision events every second. With an average ATLAS event size of ∼1.5 MB, this is far more information than can be saved into the finite data storage resources available. The goal of the trigger system is to move interesting physics events to permanent storage, while rejecting the vast majority of other events. Based on the technology and resources restrictions, the bandwidth kept by the trigger is approximately 1 kHz from up to 40 MHz of collisions. It is worth to mention that the

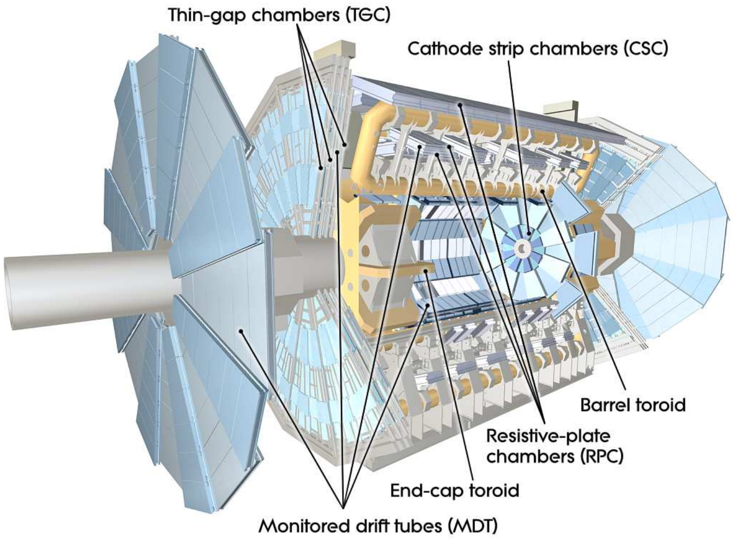


Figure 3.7: The Muon Chamber.

first long shutdown between LHC Run 1 and Run 2 operations was used to improve the trigger system with almost no component left untouched, in order to face the increased centre-of-mass energy of 13 TeV and higher luminosity operations.

The current system has two sequential levels; the first level (L1) is a hardware-based system using information from the calorimeter and muon subdetectors, reducing the data rate to ~ 100 kHz. The second *High Level Trigger* (HLT) is a software-based system which operates on the events already selected from the L1 using information from all subdetectors. HLT physics output rate is 1 kHz (400 Hz in Run 1).

For each bunch crossing, the trigger system verifies if at least one of hundreds of conditions (triggers) based on different *signatures* such as electrons, photons, muons, jets, jets with b-flavour tagging (b-jets) or specific B-physics decay modes is satisfied. In this thesis we were particularly interested on jets triggers, which were built at online level using the same reconstruction algorithm and calibration as in offline described in Section 4.0.11.

One of the important upgrades for Run 2 was the inclusion of a new topological trigger module (L1Topo), consisting of two FPGA-based (Field-Programmable Gate Arrays) processors. The modules are identical hardware-wise and each is programmed to perform selections based on geometric or kinematic association between trigger objects received from the L1Calo or L1Muon systems. A complete description of the trigger system used in Run 1 and Run 2 can be found here [13, 8].

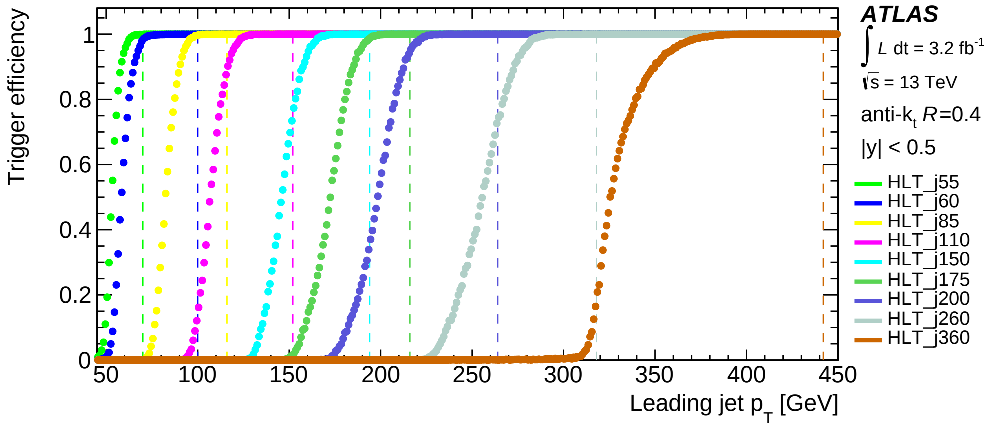


Figure 3.8: Turn-on curves of the trigger efficiency as a function of the jet p_{T} for the central rapidity bin, obtained by emulating the single jet triggers in data. The dashed lines represent the trigger thresholds used in measurement.

Single Jet Triggers

The triggers used for the measurement presented in this thesis were based on single jet transverse energy $E_T \equiv \frac{E}{\cosh(\eta)}$ thresholds. Values varying from 55 GeV to 360 GeV are used to record events with at least one jet with E_T above the threshold in the region $|\eta| < 3.2$.

Given the orders of magnitude difference on the cross-section between low and high energy jet events produced, triggers must be prescaled, meaning that only a fraction of the events fired by a trigger should be kept. This fraction depends on the prescale factor of each trigger, e.g a prescale of 1 means all events are kept, while a prescale of 10 means that just 1 over 10 events which passed the trigger is stored. The prescale factor decreases with the energy threshold to populate more the high p_{T} region, which are the more interesting events one wants to analyse. The highest threshold trigger (`HLT_j360`) was activated without prescales during the 2015 data-taking.

To ensure a fully efficient phase-space of the measurement in terms of the trigger, the efficiency for each trigger was measured as a function of jet p_{T}, as seen in Figure 3.8. It was computed by emulating the trigger in data using an inclusive jet kinematic selection. Each trigger corresponds in this way to a jet p_{T} region determined by the dash vertical line. These contiguous sets ensures an efficiency of 99.9% for each trigger.

The resulted p_{T} ranges for each of the trigger items used in the measurement are shown in Table 3.9. These corresponds to the regions determined by the dash line thresholds in Figure 3.8. The right column represents the equivalent luminosity for each trigger defined as

p_{T} range, [GeV]	Trigger	Luminosity [pb^{-1}]
75–100	HLT_j55	0.0807252
100–134	HLT_j60	0.118673
134–152	HLT_j85	0.483386
152–194	HLT_j110	1.37096
194–240	HLT_j150	5.09517
240–264	HLT_j175	10.1739
264–318	HLT_j200	18.7346
318–442	HLT_j260	65.0237
442– END	HLT_j360	3158.32

Figure 3.9: p_{T} ranges for each trigger item determined by its region of fully efficiency. The right column shows the equivalent luminosity for each trigger item used in the measurement.

$\mathcal{L}_i = \frac{\mathcal{L}}{\pi_i}$, where π_i is the prescale factor of item i. As it can be seen, for the item `HLT_j360` the prescale factor $\pi_i = 1$ and so $\mathcal{L}_i = \mathcal{L} \sim 3.2\ \mathrm{fb}^{-1}$.

Jets (resumen)

Debido a la propiedad de confinamiento de QCD, los quarks y gluones no pueden ser observados de forma directa, sino a través de lluvia de partículas incoloras, que reciben el nombre de *jets*. Para relacionar las señales observadas con los partones a nivel de elemento de matriz, una definicón precisa y consistente de un algoritmo de recombinación de partículas para formar un jet (algoritmo de jet) es necesaria. Este toma como input una lista de cuadrivectores (partones, partículas, trazas, clusters calorimétricos) y los recombina para obtener un jet el cual contiene información de la cinemática del elemento originario.

El algoritmo de jet mas utilizado en ATLAS, y en particula el utilizado en esta tesis, es el denominado *anti-*k_t, dado a su no sensibilidad con el pile-up. Tal algoritmo se basa en una definición de distancia dado por la ecuación 4.1.

Los inputs al algoritmo de jets puede ser cualquier lista de objetos que puedan ser determinados por un cuadrivector. En particular clusters calorimétricos. La recombinación de estos clusters da lugar a jets calorimétricos. Aquellos clusters en donde se asume que la energía es depositada por partículas electromagnéticas se los denomina EM clusters, dando lugar a EM jets. Otro tipo de jets son los llamados LCW, los cuales fueron formados por clusters corregidos por la respuesta del calorímetro a hadrones. Por último, se encuentran los denominados p-flow jets, en donde se utilizo no información del calorímetro (EM clusters), sino que también las trazas asociadas a los mismos. Esta técnica, novedosa en ATLAS, permite mejorar la resolución a bajo momento ($\leq$ 100 GeV) de los jets. Por último se encuentran los jets de partículas, los cuales fueron reconstruidos por las partículas estables simuladas por los Monte Carlos.

La calibración de los jets resulta fundamental para su posterior comparación con los modelos teóricos. La corrección consiste en llevar la energia de los jets reconstruidos en el detector al su correspondiente jet de partícula. Unas de las etapas de la calibración es la basada en simulaciones Monte Carlo denominada *MC Jet Energy Scale (MC-JES).* Dicha calibración fue unas de las tareas principales ejercidas durante el doctorado y se describe en detalle en este capítulo.

La calibración de los jets introduce incertezas sistemáticas en a medición. La mas relevante es la que proviene de la determinación de la escala de energía, la denomidada *Jet Energy Scale* que proviene de la propagación a la sección eficaz de las correcciones realizadas en el cuadrimomento de los jets.

Chapter 4

Jets

4.0.10 From Particles to Jets

Because of QCD confinement, quarks and gluons cannot be directly observed. They fragment and hadronise, leading to a collimated spray of energetic hadrons, a jet. By measuring the jet energy and direction in a particle detector one can get close to the idea of the originating parton kinematics. However, in order to relate the observable signals to the theoretical partons from the matrix element, one needs a precise and robust *jet definition* algorithm to perform the comparison on an equal footing.

A jet algorithm is made of a set of parameters and rules on how to combine two four-momentum objects. It takes a list of partons, hadrons, tracks or calorimeter depositions, and returns parton, particle, track or calorimeter jets, respectively. To avoid theoretical issues and allow an easier connection between experimental results to the expectations at hadron-level, a jet definition should be *infrared* and *collinear* safety (IRC) [89]. This property means that if one modifies an event by a collinear splitting or the addition of a soft emission, the set of hard jets that are found in the event with the algorithm should remain unchanged, as shown in Figure 4.1.

Sequential recombination algorithms

Recombination algorithms are both collinear and infrared safe. For this reason, they can be used in calculations to any order in perturbation theory. The term recombination is used since they attempt to follow the parton shower branchings (forward or backward) which become progressively softer as the shower evolves. The resulting jet can be thought of as the final stage of this process and the algorithm is the device used to retrace the tree of sequential branchings. In general, recombination algorithms operate by successively combining pairs of particles using

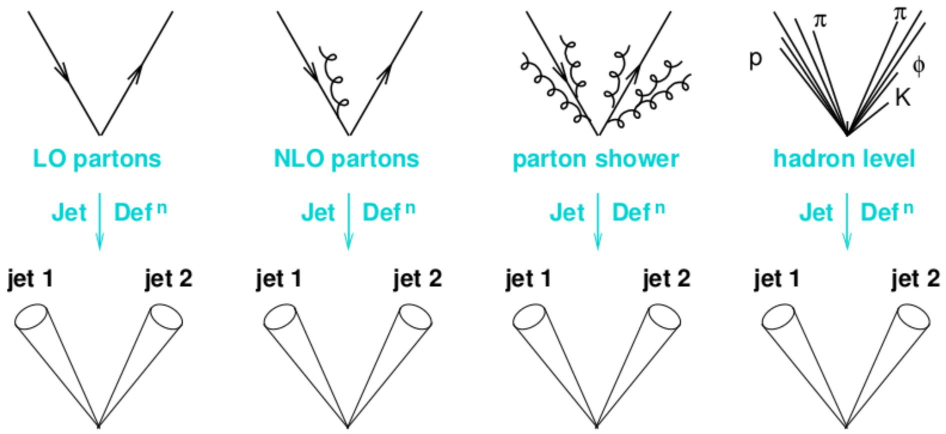


Figure 4.1: The application of a jet definition to a variety of events that differ just through soft/collinear branching and hadronization should give identical jets in all cases [88]

a distance metric d_{ij} between entities (four-momentum objects) and between entity i and the "beam" d_{iB}:

$$\begin{aligned} d_{ij} &= \min(p_{\mathrm{T},i}^{2p}, p_{\mathrm{T},j}^{2p}) \frac{\Delta R_{ij}^2}{R^2} \\ d_{iB} &= p_{T,i}^{2p} \end{aligned} \tag{4.1}$$

where ΔR is the angular distance $\Delta R^2 = \Delta y^2 + \Delta\phi^2$, and R is a parameter that represents the radius of the jet. The choice of the parameter p results in three different algorithms which can be obtained: The k_t algorithm with $p = 1$ [95, 87], The Cambridge-Aachen (C/A) algorithm [100], with $p = 0$, and the anti-k_t algorithm, with $p = -1$ [78].

The general algorithm procedure consists in:

1. Compute the minimum d_{ij} and d_{iB} among all particles.

2. If it is a d_{ij}, then recombine i and j into a single particle and repeat from step 1.

3. Otherwise, if it is a d_{iB}, then identify i as a jet and remove it from the list of particles. Return to step 1.

4. Continue until no particles remain.

The jet algorithm used in this thesis was the anti-k_t algorithm, which is generally the one adopted for the reconstruction of physics objects for an analysis in ATLAS. The anti-k_t first clusters hard objects together which results in more regular jets, in contrast with the k_t algorithm, which starts with soft particles, and so it results in irregular shapes (Figure 4.2). This quality makes the anti-k_t algorithm less sensitivy to pile-up effects [43]. The choice of the

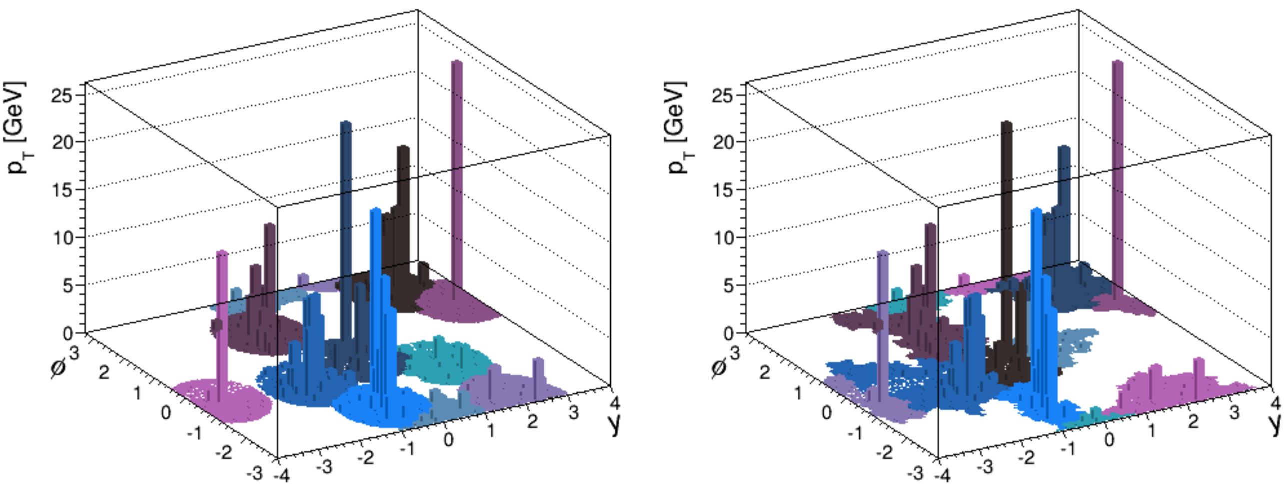


Figure 4.2: Active area of jets reconstructed with the anti-k$_t$ (left) and k$_t$ algorithms (right) with p_T^{jet} >10 GeV.

radius parameter was limited to the one used to derive the jet calibration results, as they take a long time and efforts to produce them for a given radius. The value used was $R = 0.4$.

These algorithms were implemented with the *FastJet 2.4.3* [42, 79] software package.

Jet Algorithm Inputs

Hadronic jets used for ATLAS analyses are reconstructed starting from the energy depositions of electromagnetic and hadronic showers in the calorimeters. The input to jet reconstruction can be topological cell clusters, charged particle tracks or any object which can be described by a four-momentum. A collection of three-dimensional, massless, positive-energy topological clusters (topo-clusters) made of calorimeter cell energies are used as input to the anti-k_t algorithm to form the calorimeter jets. Topo-clusters are built from neighboring calorimeter cells containing a significant energy above a noise threshold that is estimated from measurements of calorimeter electronic noise and simulated pile-up noise [3]. The calorimeter cell energies are measured at the EM scale, which assumes the energy is deposited by electromagnetically interacting particles. The resulting jets formed by these clusters are named EM-scale calorimeter jets (EM jets).

A second topo-cluster collection is built by taking the previous EM scale cluster collection, classifying them as being electromagnetic or hadronic and calibrating the calorimeter cells in these clusters accordingly such that the response of the calorimeter to hadrons is correctly reconstructed. This calibration uses the so called local cell signal weighting (LCW) method that aims at an improved resolution compared to the EM scale by correcting for a variety of effects in the calorimeter [14]. LCW corrections are determined from Monte Carlo simulations of charged and neutral pions. The jets built by these clusters are known as LCW-scale calorimeter jets (LCW jets).

Another topo-cluster collection built from EM scale cells, i.e without the LCW correction, are the particle flow clusters (p-flow). This is a novel technique implemented for the first time in Run 2 which is still evolving. In this case, tracking and calorimetric information are combined in order to remove overlaps between the momentum and energy measurements made in the inner detector and calorimeters, respectively. As a result, this improves the accuracy of the charged-hadron measurement, while retaining the calorimeter measurements of neutral-particle energies, and so improving the resolution of p-flow jets at low momenta [7].

Jets referred to as truth jets are reconstructed using stable, final-state particles from MC generators as input. Candidate particles are required to have a lifetime of $c\tau > 10$ mm and muons, neutrinos, and particles from pile-up activity are excluded. Truth jets are therefore measured at the particle-level energy scale.

Jet Substructure

A major field of jet physics is the identification of hadronically decaying boosted top quarks or W bosons, reconstructed as a single jet in the calorimeter. These boosted particles are the product of beyond SM particle decays and so they are crucial to searches for new physics and measurements at the energy frontier. Such high-mass resonances result in high p_{T} particles which decay in collimated products. The angular separation of the decay products is $R \sim 2M/p_{\mathrm{T}}$, where M and p_{T} are the mass and transverse momentum of the boosted, top or W particle. This boost makes it possible to reconstruct the entire hadronic decay in a single, large-radius ($R = 1.0$) jet.

The identification of resonance decays to jets relies on jet substructure, which is generally derived from the correlations between jet constituents. These correlations reflect differences in the radiation patterns for jets produced through resonant (signal) and non-resonant (background) events, and can therefore be used to distinguish the two. A substantial number of jet substructure variables have been proposed based on theoretical considerations, many of which have been used for jet classification in ATLAS [31, 33, 32, 29, 30]. Furthermore, it has been shown that improvements in the identification of hadronically decaying W bosons versus light quark- and gluon-initiated multijets can be achieved through a combination of several jet substructure variables using multivariate analysis (MVA) approaches, in particular using boosted decision trees (BDT) and neural networks (NN) [34, 10].

Table 4.1 shows the substructure variables used in this thesis for the neural network performance analysis developed in Chapter 6.

Variable	**Type**	**Reference**
C_2, D_2	Energy correlation ratios	[73]
τ_{21}	N-subjettiness	[97]
R_2^{FW}	Fox–Wolfram moment	[64]
$\mathcal{P}$	Planar flow	[25]
a_3	Angularity	[27]
A	Aplanarity	[45]
Z_{cut}, $\sqrt{d_{12}}$	Splitting scales	[98, 28]
$KtDR$	k_t-subjet ΔR	[44]

Table 4.1: Substructure variables used for the neural network- and BDT-based MVA jet classifiers. Feature extraction based on Ref. [10].

4.0.11 Jet calibration

Calorimeter jets need to be calibrated in order to account for various known limitations of the detector as well as the inherent experimental setup, such as:

- **Calorimeter non-compensation:** correction for the different scales of the energy measured from hadronic and electromagnetic showers. Hadronic interactions deposit less ionization energy in the detector than EM ones due to the energy used to breake a nucleus.

- **Dead material:** energy deposited in non-instrumented regions, such as wires or services (support structures and cryostats).

- **Leakage:** showers reaching and escaping the outer edge of the calorimeters.

- **Energy deposits below noise thresholds:** clusters are only formed by energy deposits which are well above the background noise. Therefore it is required to correct for particles that do not form clusters.

- **Out of cone radiation:** particles which are included in the truth jet but not in the reconstructed jet. Part of the shower in the calorimeter develop far from the jet cone yielding topo-clusters that fail to be added to the jet in the clustering process.

- **Pile-up:** energy deposition in jets is affected by the presence of multiple proton-proton collisions in the same bunch crossing (in-time pile-up) and in previous bunch crossings (out-of-time pile-up).

In order to account for these effects, a combination of methods based on Monte-Carlo (MC) and data-driven techniques are employed to calibrate detector signals (reco jets) to physics-level objects (truth jets). Figure 4.3 presents an overview of the 2015 ATLAS calibration scheme for EM-scale calorimeter jets. Each stage of the calibration corrects the full four-momentum unless otherwise stated, scaling the jet p_{T}, energy, and mass.

The jet calibration techniques are discussed in detail in reference [4]. The following presents a summary about the jet calibration steps.

First, the origin correction recalculates the four-momentum of jets to point to the hard-scatter primary vertex[1] rather than the center of the detector, while keeping the jet energy

[1]The "primary vertex" is defined by that which has the highest $\sum p_T^2$ of tracks associated with it.

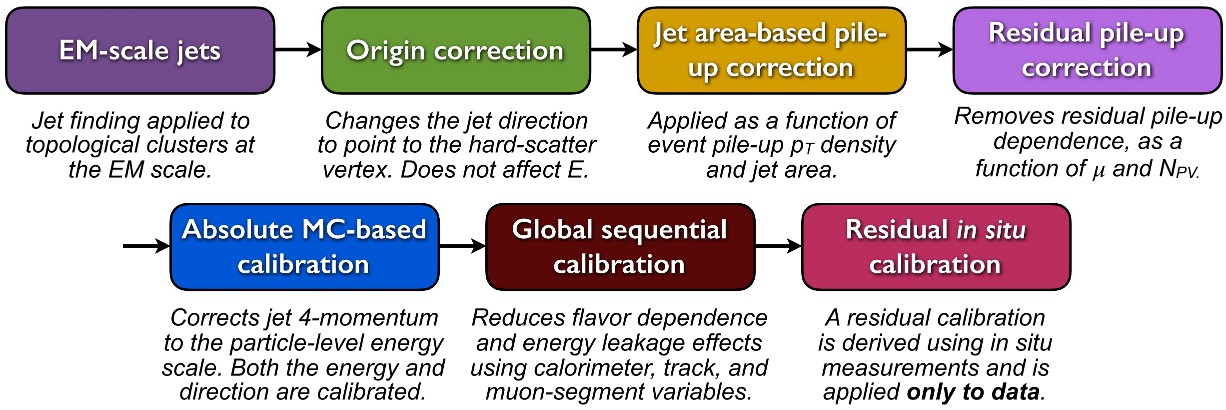


Figure 4.3: Calibration stages for EM-scale jets. Other than the origin correction, each stage of the calibration is applied to the four-momentum of the jet.

constant. This correction improves the η resolution of jets, as measured from the difference between reco jets and truth jets in MC simulation. The η resolution improves from roughly 0.06 to 0.045 at a jet p_{T} of 20 GeV and from 0.03 to below 0.006 above 200 GeV.

Next, the pile-up correction removes the excess energy due to in-time and out-of-time pile-up. It consists of two components: an area-based method subtraction based on the median p_{T} density (ρ) [43] and a residual correction derived from MC simulation, both derived at the per-event level.

The MC-JES calibration corrects, on average, the jet four-momentum to the particle-level energy scale, as derived using truth jets in dijet MC events, and is discussed in detail in the next section. Further improvements to the reconstructed energy and related uncertainties are achieved through the use of calorimeter information of longitudinal development of the jet, muon segments, and track-based variables in the global sequential calibration (GSC).

Finally, a residual in situ calibration is applied to correct jets in data using well-measured reference objects, including photons, Z bosons, and calibrated jets.

Monte Carlo based jet energy calibration

This section describes the technique used to implement the jet energy scale (JES) based on MC in order to recover the jet particle-level (true) scale from the reconstructed (reco) one. The calibration is reduced to an energy- (E) and η-dependent correction. It consists in measuring and inverting the jet energy response function defined as:

$$R(E) = \left\langle \frac{E^{reco}}{E^{true}} \right\rangle$$

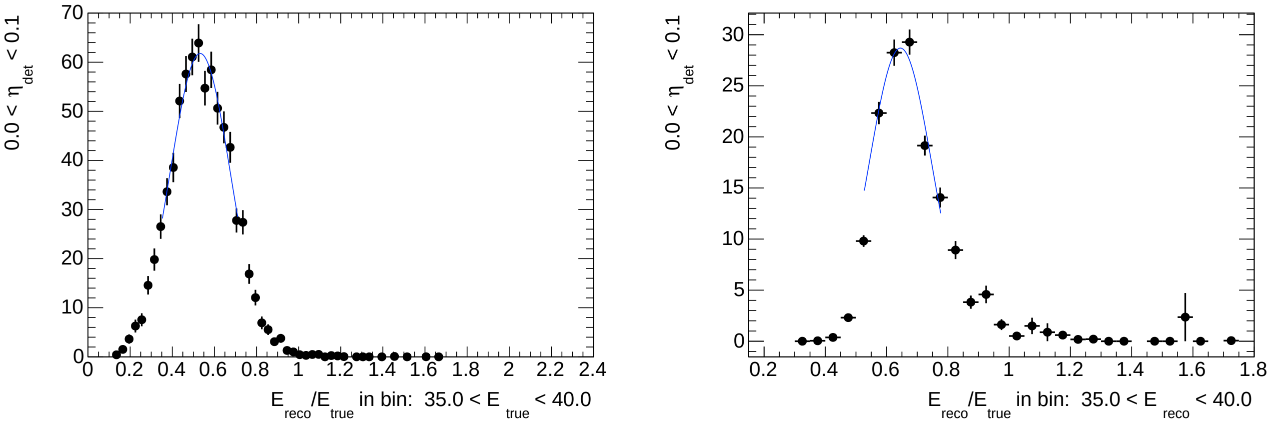


Figure 4.4: Response distribution for a given E^{true} (left) and E^{reco} (right) bin.

The E correction is derived in η bins of size 0.1 from -4.5 to 4.5. Calibrations using different reconstruction jet scale choices such as EM (nominal), LCW and pflow and various jet radii $R = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1.0)$ were derived and released for the use of the whole collaboration. We will focus on the results obtained for jets reconstructed at EM scale with $R = 0.4$, which were the default recommendation proposed by the calibration team to use in jet analyses. The studies were performed using a high statistics di-jet MC sample simulated with `Pythia`.

Numerical Inversion Technique

To estimate the true energy of a jet based on reco information only, which is what we want, we need to estimate first the response function $R(E^{reco})$. Once we get this, the true energy of a jet can be estimated as $\frac{E^{reco}}{R(E^{reco})}$. However, it is not straightforward how to evaluate the response as a function of E^{reco}, because for fixed E^{reco} the response distribution is not Gaussian. On the other hand, the calorimeter response for jets of fixed E^{true} is Gaussian, as illustrated in Figure 4.4. This is related to the fact that the cross-section is a steeply falling function with E^{true}, then for a fixed E^{reco} bin, there will be more jets with low E^{true} (i.e: higher response), causing a bias towards higher values of the response distribution.

The *numerical inversion* technique attempts to solve this issue by estimating $R(E^{reco})$ from $R(E^{true})$. This is a general technique and allows applying E-dependent calibrations to reconstructed jets when the average response of these jets has a dependence on E^{true}. It can be summarized as follows:

1. Calculate $R(E^{true})$: First, the distributions of $\frac{E^{reco}}{E^{true}}$ are evaluated in different E^{true} bins. Next, each distribution is fit by a Gaussian and the mean value is computed as a function

of each E^{true} bin center. Finally, in order to get a continuous response function to be evaluated on a jet-by-jet basis, a polynomial fit of the form of Eq. 4.2 is performed.[2]

2. Estimate the energy reco of a jet from its true energy as: $E^{reco}_{est} = R(E^{true}) \cdot E^{true}$, which is an unbiased estimator for E^{reco} assuming $R(E^{true})$ to be Gaussian.

3. Compute the response obtained in step 1 in bins of E^{reco}_{est} and fit it with Eq. 4.2.

Notice that the estimated $R(E^{reco})$ was in fact obtained from E^{reco}_{est}. However, when it is used as part of the calibration it is evaluated on a jet E^{reco}.

$$R(E) = a_0 + \sum_{i=1}^{n} a_i (\log E)^i \tag{4.2}$$

The effect of the numerical inversion technique is ilustrated in Figure 4.5. The response function (left) is used to get the final calibration curve (right). The transformation corresponds to displacing the points parallel to the x-axis to the left ($E^{reco} < E^{true}$), leaving the y-axis unchanged.

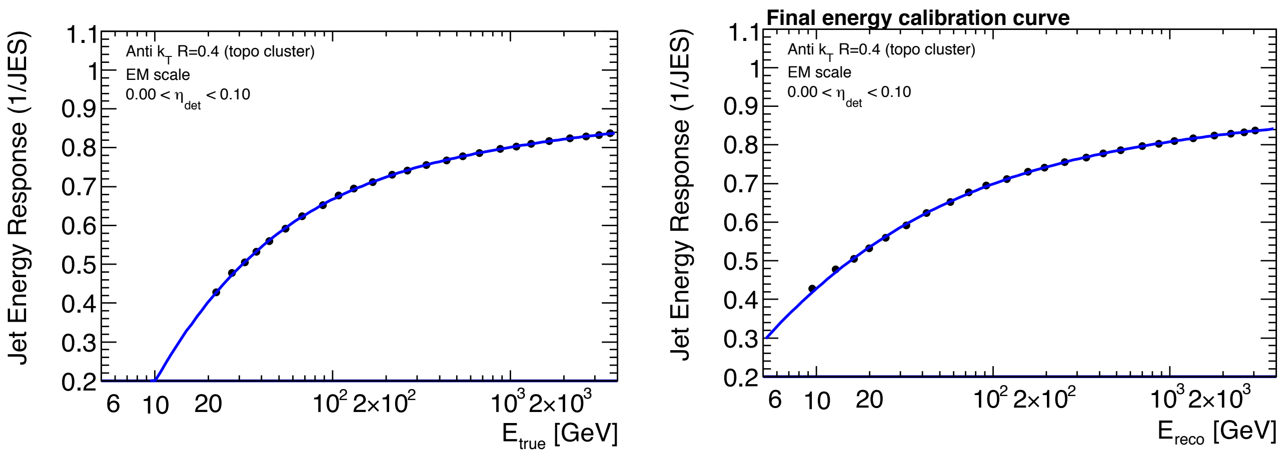


Figure 4.5: Average response as a function of E^{true} (left) and E^{reco} (right). The right plot is the final calibration curve.

The shape of the response function depends on the clusters used as inputs to the jet algorithm and on the parameters as the jet radius R. Responses for different jet collections can be seen in Figure 4.6. As it can be seen, LCW jets have a higher response (closer to the particle-level) than EM jets. This is expected as LCW jets are built from calibrated calorimeter cells [14]. Both curves increase with energy, as expected from Appendix A. On the other hand, p-flow jets present a better response than EM and LCW jets for $E^{true} < 100$ GeV and

[2]A physical motivation of this functional dependence can be found in Appendix A.

then decreases smoothly up to the EM response. This is contemplated given that in p-flow jets, the energy measured in the calorimeter of EM topo clusters that have associated tracks is replaced by the measured from tracks in the inner detector, yielding to an improvement of the jet resolution and energy response particularly for jets with $p_{\mathrm{T}} < 100$. The figure in the right shows the jet response for different jet radii. It is clearly seen how the response decreases as the jet radius increases. This could be related to the fact that the higher the radius the more clusters the jet algorithm takes, strengthen the calorimeter non-compensation effects as one would have "more constituents to correct for".

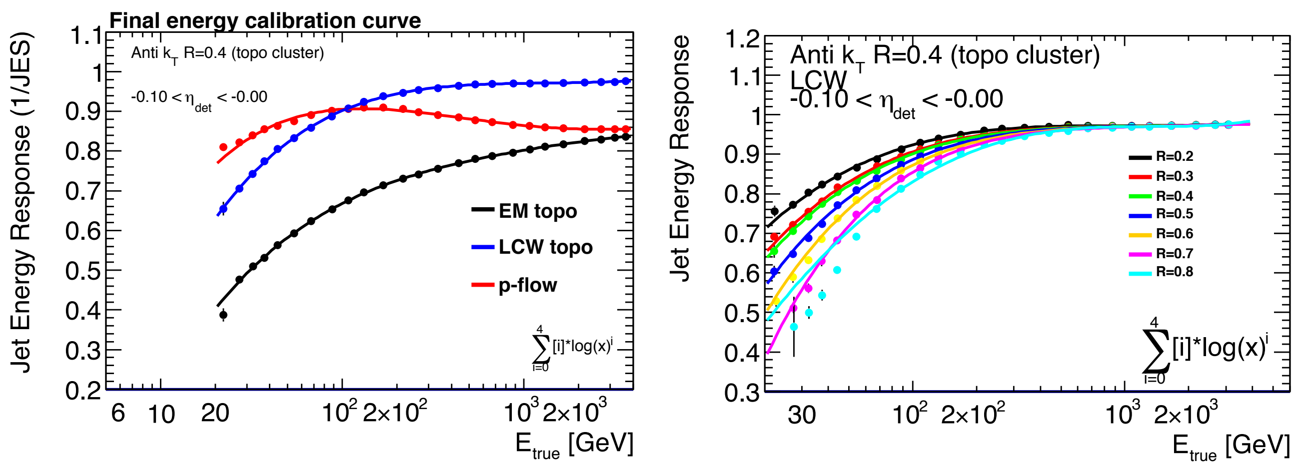


Figure 4.6: Response function for EM, LCW and p-flow jets (left), and different jet radius parameters (right)

E and η calibrations

With an expression for $R(E^{reco})$, the jet can be calibrated by applying a correction factor as

$$E^{calib} = E^{reco}/R(E^{reco})$$

The jet four-momentum is calibrated in an equivalent way: using $1/R(E^{reco})$ as a scale factor for each of its components.

Once the energy calibration was derived, the following step is to correct η. The η response is defined as $\langle \eta^{reco} - \eta^{true} \rangle$ and is parametrized also with Equation 4.2 in terms of E^{true}. There is no need of numerical inversion in this case since the correction is evaluated in E^{calib} which was obtained in order to match E^{true}. The correction is applied on a jet-by-jet basis as:

$$\eta^{calib} = \eta^{reco} - \langle \eta^{reco} - \eta^{true} \rangle$$

Closure

The E and η calibrations derived were validated by comparing E^{calib} and η^{calib} with their respective true values. In concrete, the distribution of E^{calib}/E^{true} was computed in different E^{true} and η bins. The averages of these distributions are then plotted for each E^{true} bin as a function of η, as shown in Figure 4.7. The average response agrees with unity within 1% for jets with $p_{\mathrm{T}} > 30$ GeV, presenting a non-closure below that range, similar to what obtained in the 2012 calibration [52]. This concerns to a non-gaussianity effect of the response distribution at low p_{T} coming from a $p_{\mathrm{T}} < 5$ GeV cut imposed at the jet constituent level. This is done in order to reduce the sample size, although nowadays lower thresholds are being used to achieve a better calibration at low p_{T}.

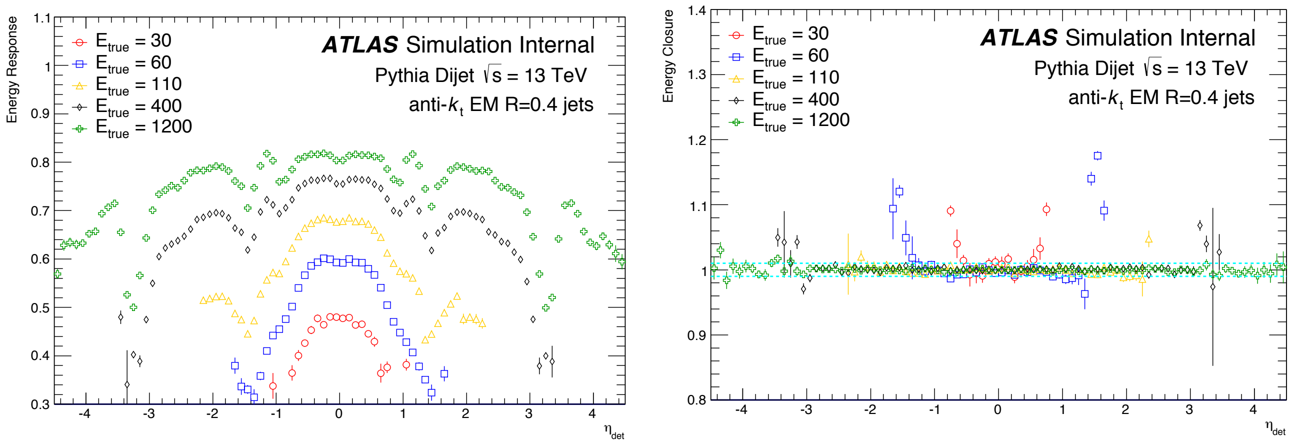
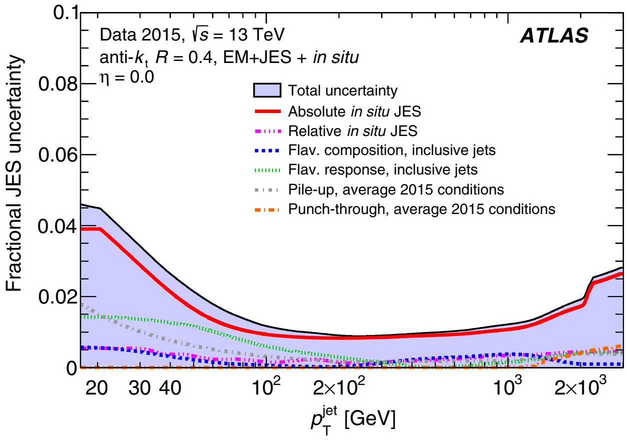


Figure 4.7: Average response in different E^{true} bins as a function of η before (left) and after (right) the calibration.

4.0.12 Jet systematic uncertainties

The most relevant jet uncertainty comes from the Jet Energy Scale (JES) determination based on the calibration chain described in Fig. 4.3. The final calibration ends up with a set of 80 JES uncertainty terms propagated from each calibration stage. Most of them (70) come from the *in situ* corrections and account for assumptions made in the event topology, MC simulation, sample statistics, and electron, muon, and photon energy scale propagated uncertainties. The remainings come from pile-up corrections, to account for potential MC mismodelling of μ, N_{pv} and ρ, *punch-through* GSC correction, differences in the flavour response between gluon and light-quarks initiated jets and high p_{T} single-particle response studies. A detailed list of each uncertainty sources can be found here [4].

The combination (sum in quadrature) of all relative jet p_{T} uncertainty terms as a function of jet p_{T} and η is shown in Figure 4.8.



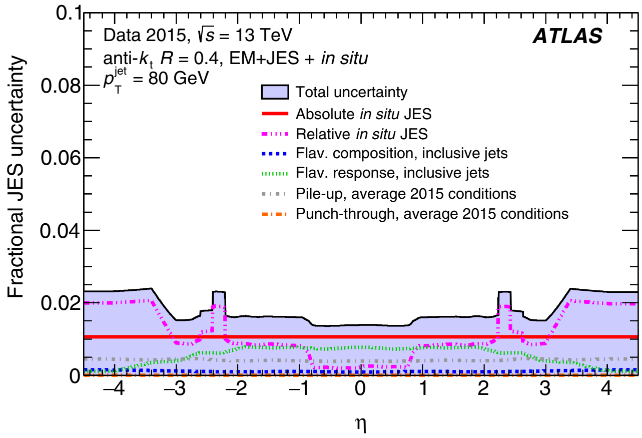


Figure 4.8: JES p_{T} uncertainty of fully calibrated jets as a function of jet p_{T} at $\eta = 0$ (left) and η at $p_{\mathrm{T}} = 80$ GeV (right).

The uncertainty coming from the statistical variations of the relative *in situ* calibration (η-Intercalibration) was split by 247 single components corresponding to the statistical uncertainty of each p_{T}-η bins. The use of this decomposition is very important to do particularly for precise measurements, which results in a better estimation of the degree of agreement between data and theory coming from a better modelling of the uncertainties.

Another less but significant contribution comes from the uncertainty on the precision of the measurement of the energy of a jet. In other words, the Jet Energy Resolution (JER) uncertainty. This will be covered in detail in Sec. 5.0.18.

Medición de la Producción de Pares de Jets a 13 TeV (resumen)

En este capítulo se describe en detalle la medición de la sección eficaz de producción de pares de jets a 13 TeV. Esta fue medida para distintas regiones de separación angular entre jets, y^*, en función de la masa invariante del sistema, m_{jj}.

Para poder comparar la sección eficaz al nivel de partícula, es necesario implementar correcciones por efectos de resolución y eficiencia del detector. Tal técnica se denomina *unfolding*. La que se utilizo en esta tesis es un método iterativo basada en una matriz de transferencia que cuantifica la probabilidad de un jet reconstruido en el detector provenir de un dado jet de partícula y vice versa. La misma es una simplificación de una técnica mas general utilizada para distintos tipos de mediciones.

La estimación de las incertezas estadísticas se hizo mediante la técnica de *bootstrapping*, la cuál consiste en pesar los eventos por una variable aleatoria proveniente de una distribución de *Poison* con $\mu = 1$ para generar *toys* y propagar los mismos al observable final. Se construye con los toys una matriz de covariancia, obteniendo de esta forma la incerteza estadística empírica sin necesidad de generar mas datasets.

La obtención de las incertezas sistemáticas se realiza variando el cuadrimomento del jet por su respectiva incerteza y propagando el jet con momento modificado a su correspondiente sección eficaz. Las incertezas del momento del jet son naturalmente amplificadas al ser propagadas a la sección eficaz, debido a su dependencia exponencial decreciente con la masa invariante.

Habiendo cuantificado todas las incertezas, se procede a la comparación con los cálculos teóricos a orden siguiente al dominante (NLO) corregidos por efectos no perturbativos. El nivel de acuerdo entre datos y teoría se cuantifica mediante un test de χ^2, el cual tiene en cuenta las varianzas (estadísticas y sistemáticas) y sus correlaciones (covarianzas). El estadístico utilizado es sensible también a la asimetría presentes en algunas incertezas.

Las mediciones dieron consistentes con el Modelo Estándar (p-values del orden porcentual), validando el ME de producción de pares de jets a 13 TeV.

Chapter 5

Measurement of dijet production at 13 TeV

5.0.13 Introduction

Precise measurements of jet cross-sections are crucial in understanding QCD physics at hadron colliders. In particular they provide valuable information about the strong coupling constant, α_s, and the structure of the proton (PDFs). What is more, dijet events represent a background to many other processes at hadron colliders. The predictive power of fixed-order QCD calculations is therefore also relevant in many searches for new physics and so it must be tested with experimental data.

In this chapter the measurement of double-differential dijet production cross-sections are presented as a function of the invariant mass of the dijet system, m_{jj}, and as a function of half the absolute rapidity separation, denoted y^*, defined as $|y_1 - y_2|/2$, where the subscripts 1,2 label the highest and second highest-p_{T} jets in the event satisfying $|y| < 3.0$, respectively. This quantity is invariant under a Lorentz boost along the z-direction and is equal to the absolute rapidity of each jet in the dijet rest frame.

Events with at least two jets with $p_{\mathrm{T}} > 75$ GeV and $|y| < 3$ are selected. In addition, the scalar sum of the p_{T} of the first and second leading jets, $H_{T,2} = p_{T1} + p_{T2}$, is required to be above 200 GeV. This requirement avoids instabilities in the NLO cross-section calculations due to the symmetric p_{T} requirement applied to the leading and sub-leading jets [66, 24].

The double-differential dijet cross-section is calculated as the ratio of the number of dijet events after correcting for detector effects (Section 5.0.16), N_{dijet}, to the integrated luminosity of the data, in a given interval of the invariant mass and y^*, Δm_{jj} and Δy^* respectively:

$$\frac{d^2\sigma}{dm_{jj}dy^*} = \frac{N_{\text{dijet}}}{L\Delta m_{jj}\Delta y^*}$$

A detailed description of the experimental techniques used through the measurement and the results on the comparison with NLO perturbative QCD predictions, corrected for electroweak and non-perturbative effects, are featured below.

5.0.14 Trigger Analysis Strategy

As mentioned in Sec. 3.0.9, single-jet triggers with E_T thresholds varying from 55 GeV to 360 GeV were used to record events. The trigger strategy for the dijet production accounts for different prescale combinations for dijet events in a given (m_{jj}, y^*) bin, which can be accepted by up to two jet triggers depending on the transverse momenta and pseudorapidities of the leading and sub-leading jets. Each pairing of triggers has a unique corresponding luminosity, which is used to calculate the differential cross section for that pairing. The separate cross sections from all pairings are then summed to obtain the physical distribution. This strategy is described as the *Inclusive method for fully efficient combinations* in [74]. As an example, consider an event which has a leading and sub-leading jet with 318 GeV $< p_T^{(1)} <$ 442 GeV and 264 GeV $< p_T^{(2)} <$ 318 GeV. Each jet belongs to a different fully efficient trigger region described in Table 3.9. In this case the corresponded triggers are: t1=`HLT_j260` and t2=`HLT_j200`. When the two jets fall in different trigger regions, an equivalent luminosity of the "*OR*" from the two triggers is calculated by weighting the event by $w = 1/P$, where P is the probability of the event to pass either t1 or t2 prescales, determined by its prescale factors (π_{t1}, π_{t2}) respectively:

$$P(t1 \cup t2) = P(t1) + P(t2) - P(t1 \cap t2) = \frac{1}{\pi_{t1}} + \frac{1}{\pi_{t2}} - \frac{1}{\pi_{t1}} \times \frac{1}{\pi_{t2}} \tag{5.1}$$

When the leading and sub-leading jets belong to the same trigger region, i.e: 318 GeV $< p_T^{(1)}, p_T^{(2)} <$ 442, the equation 5.1 is reduced to $P = \frac{1}{\pi_{t1}}$. The latter is achieved by considering that $P(t1 \cap t1) = P(t1)$ when $t1 = t2$.

5.0.15 Data vs. Monte Carlo cross-section comparisons

Before entering in the description of the unfolding procedure, lets show some results for the m_{jj} cross-sections in data compared with the obtained in `Pythia` MC simulation at the detector level. The distributions for the six y^* bins are depicted in Figure 5.1. Although the agreement looks reasonable over the six y^* bins, there are some shape differences particularly at high p_{T}. Those differences will enter as a systematic uncertainty from the unfolding procedure, described later in Section 5.0.16. Notice how the first m_{jj} bin moves to higher values as y^* increases. This arises from the fact that at high y^* the jets are more separated in rapidty and for a fixed energy (imposed by the $p_{\mathrm{T}} > 75$ GeV cut from the analysis selection) the m_{jj} becomes higher. Mathematically it can be easily derived from the following fundamental kinematic relation:

$$m_{jj} = (P_1 + P_2)_\mu (P_1 + P_2)^\mu = m_1^2 + m_2^2 + 2(E_1 E_2 - |\vec{P}_1||\vec{P}_2| \cos\theta)$$

where the indexes (1) and (2) labels the leading and sub-leading jets respectively. The angle θ is the separation between the two jet momenta, which is proportional to y^*. Here it can be seen that when the separation increases the third term of the equation becomes higher.

The comparison with Powheg+Pythia 8 Monte Carlo simulation has also been made (Figure 5.2). The agreement in this case is much worse than Pythia. This was one of the main reasons of the choice of Pythia as the nominal MC used in the unfolding.

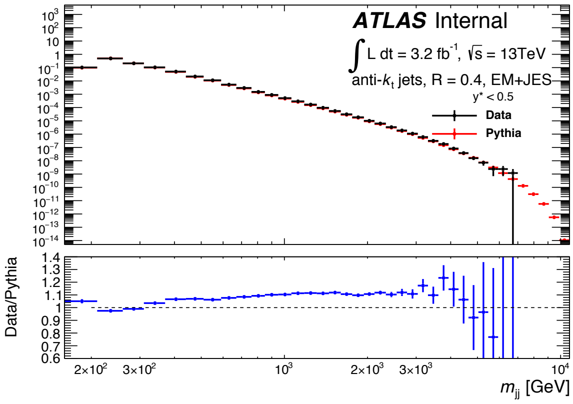
5.0.16 Data Unfolding

The comparison of data and theoretical predictions at the particle level rather than at the reconstruction level has the advantage that data can be used to test any theoretical model without the need for further detector simulation. In other words, unfolded distributions provide a result which can directly be compared with other experiments as well as with theoretical predictions, without requiring any detector information in order to retain the result.

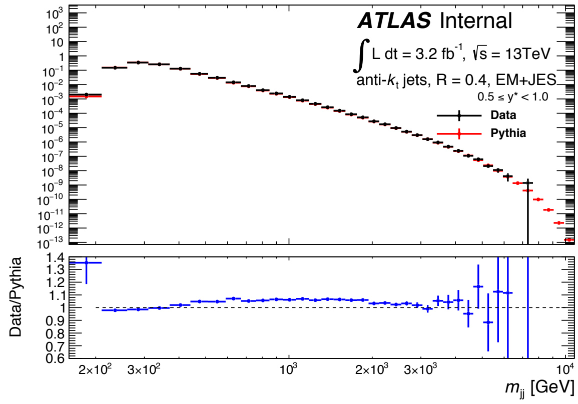
It is worth to mention that the **unfolding** also called **de-convolution** or **unsmearing**, can be found in mathematics under the general heading of **inverse problems**, which is the process of calculating from a set of observations the causal factors that produced them. As one can imagine, this has many applications not only on particle physics, but in fields such as optical image reconstruction, radio astronomy, crystallography and medical imaging.

In this chapter, the description of the unfolding technique used to de-convolute the data by detector effects is presented.

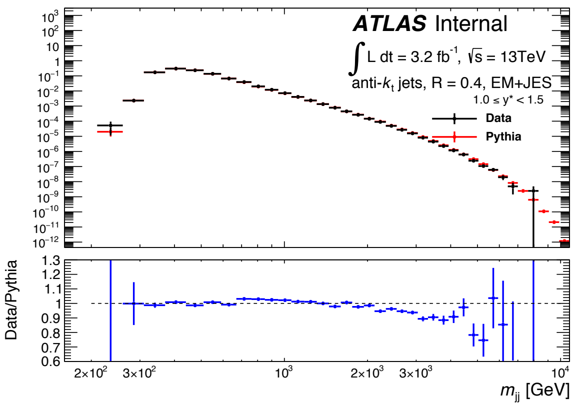
It is based on an iterative unfolding method which is detailed here [77]. This technique presents many features and is capable to be applied on any general analysis. In order to make the explanation simpler, the description is limited to this particular measurement, and so topics such as background substraction, different binning treatments, regularization function will be omitted.



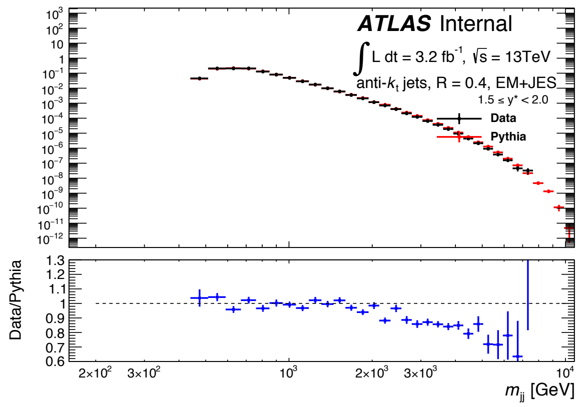
(a) $y^* < 0.5$



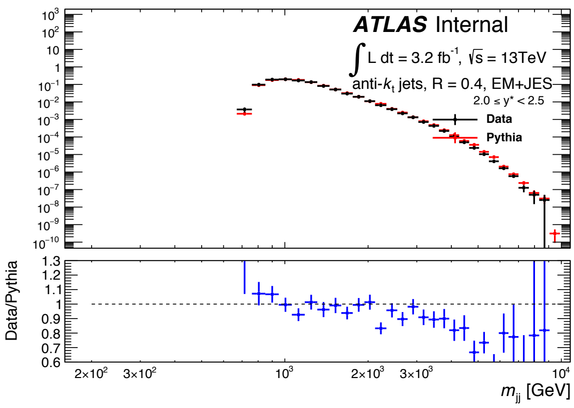
(b) $0.5 < y^* < 1.0$



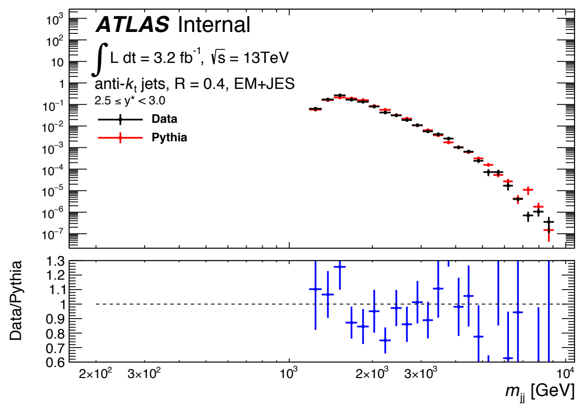
(c) $1.0 < y^* < 1.5$



(d) $1.5 < y^* < 2.0$



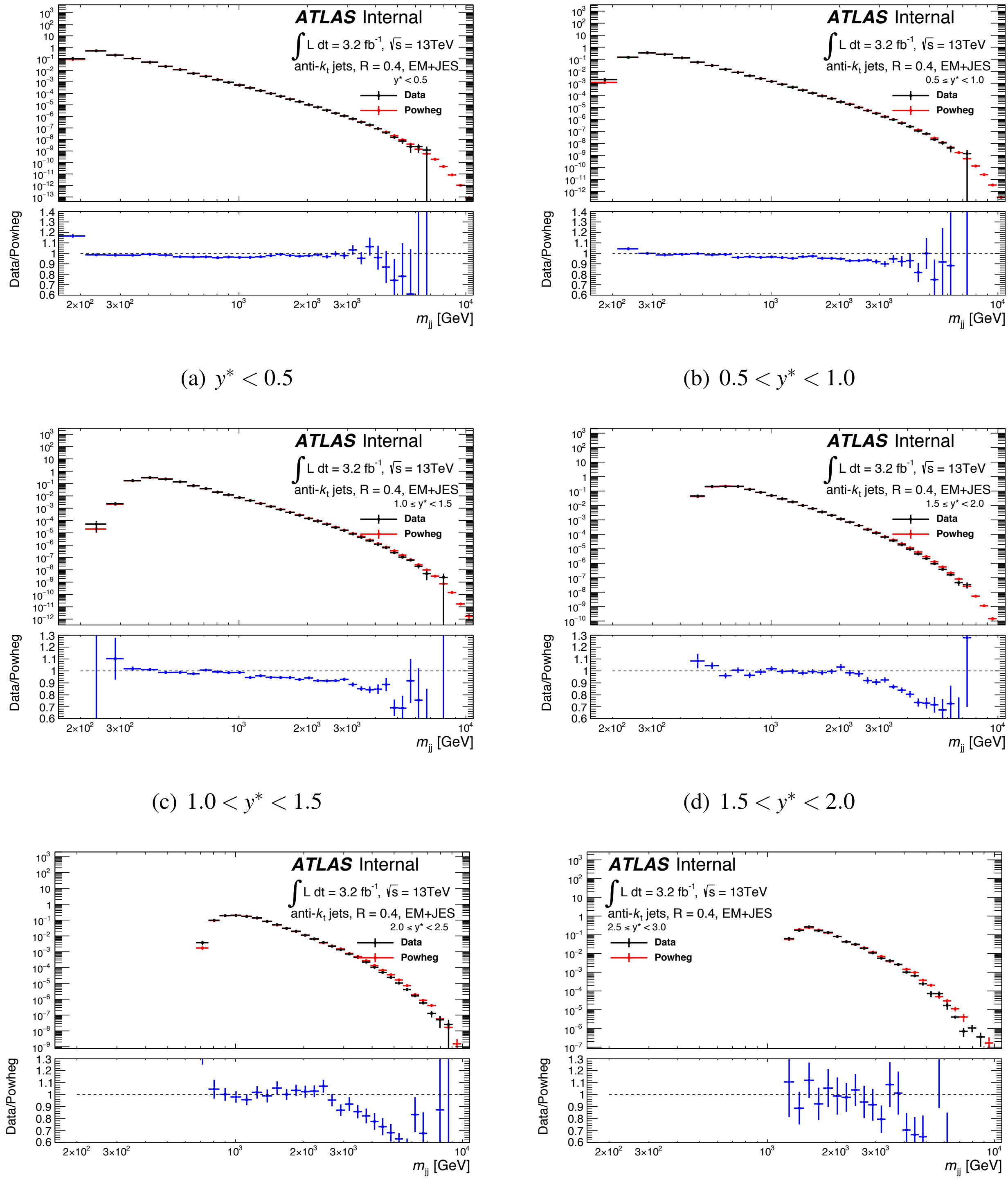
(e) $2.0 < y^* < 2.5$



(f) $2.5 < y^* < 3.0$

Figure 5.1: The m_{jj} cross-section at the detector level for anti-k_t $R = 0.4$ jets in different bins of y^* in data (black points) and `Pythia` (red points) normalised to the total area. The ratio of the two (blue points) is shown in the bottom panel.

The unfolding attempts to correct physics observables by detector effects due to limited acceptance, finite resolution, or other systematic effects producing a transfer of events between



(a) $y^* < 0.5$

(b) $0.5 < y^* < 1.0$

(c) $1.0 < y^* < 1.5$

(d) $1.5 < y^* < 2.0$

(e) $2.0 < y^* < 2.5$

(f) $2.5 < y^* < 3.0$

Figure 5.2: The m_{jj} cross-section at the detector level for anti-k_t $R = 0.4$ jets in different bins of y^* in data (black points) and `Powheg+Pythia` (red points) normalised to the total area. The ratio of the two (blue points) is shown in the bottom panel.

different regions of the spectra. All these effects were simulated in MC which were used to correct later the data. The assumption is that MC simulates well the data, which is justified by the calibrations previously applied, covered in Section 4.0.11, such as the *in-situ* calibration.

Nevertheless, as it will be shown, the technique can be iterated until the MC shape gets closer to the data. In addition, a systematic uncertainty due to possible residual MC mismodelling is derived which will be described in detail on Section 5.0.16.

In order to describe the migrations of events across m_{jj} bins between the particle- (part) and reconstructed- (reco) level, a transfer matrix (TM) was built. It was filled with events that pass the selection cuts both at reco and part level, and matches each other (i.e they belong to the same y^* bin). Figure 5.3 shows the TM for the central y^* bin.

The unfolding technique can be summarized in three steps:

First, the data d_i is corrected by the matching inefficiency at reco level, also known as purity, which is the conditional probability that a given reco event matches a true event $P(\text{true} \mid \text{reco})_i$[1]. The i labels the mjj bin index. The corrected data $\tilde{d}_i$ spectra results in:

$$\tilde{d}_i = d_i \times P(\text{true} \mid \text{reco})_i$$

Next, the data is corrected by migrations of events between different reco m_{jj} bins due to resolution effects. For this, a folding and unfolding matrices are computed: $P_{ji} = \frac{A_{ji}}{\sum_{k=1}^{n_d} A_{ki}}$, $\tilde{P}_{ji} = \frac{A_{ji}}{\sum_{k=1}^{n_d} A_{jk}}$, where A_{ji} is the TM for an event generated in bin j and reconstructed in bin i, n_d is the number of bins in data. The folding (unfolding) probability matrix P_{ji} ($\tilde{P}_{ji}$) correspond to the probability of an event generated (reconstructed) in the bin j (i) to be reconstructed (generated) in the bin i (j). Since this is a MC based correction to the data, one must determine a MC normalization coefficient $N = \frac{N_d}{N_{MC}}$ where N_d and N_{MC} are the total events in data and MC respectively. The unfolded result corrected by purity and migrations effects is then given by:

$$\mu_i = t_i N + \sum_j (\tilde{d}_j - r_j N)\tilde{P}_{ji}$$

where, for a given bin i, t_i and r_i are the number of true and reco MC events respectively.

Finally, a correction due to matching inefficiency at part level (ϵ_i) is applied:

$$\tilde{\mu}_i = \mu_i/\epsilon_i = \mu_i/P(\text{reco} \mid \text{true})_i$$

[1]The conditional probability was obtained by using the definition $P(A|B) = \frac{P(A \cap B)}{P(B)}$. In this way, $P(\text{reco} \mid \text{true})_i = \frac{\sum_{k=1}^{n_d} A_{ik}}{T_i}$, where T_i are all true events which pass the selection but not necesarily matches a reco event

$\tilde{\mu}_i$ is in this way the unfolded spectra which will be compared with the theoretical predictions.

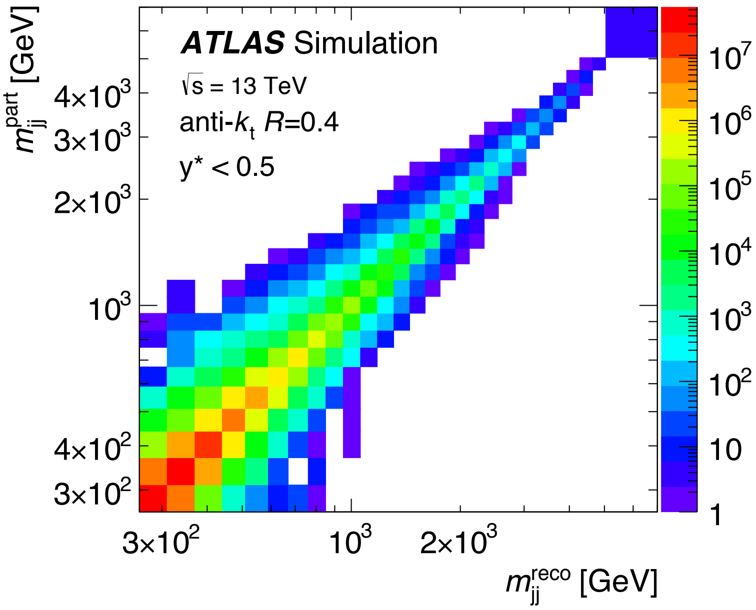


Figure 5.3: Transfer matrix used for unfolding for the central y^* bin.

The improvement of the unfolding probability matrix

As said at the beginning of this section, if the initial true MC distribution does not contain or badly describes some structures which are present in the data, one can iteratively improve it, and hence the transfer matrix. This can be done by using a better (weighted) true MC distribution, with the same folding matrix describing the physics of the detector, which will yield an improved unfolding matrix.

The improvement is performed for one bin j at the time, exploiting the difference between an intermediate unfolding result and the true MC

$$A'_{ij} = A_{ij} + \Delta\mu_j P_{ij} \frac{N_{MC}}{N_{\mu d}}$$

with $\Delta\mu_j = \mu_j - t_j \frac{N_{\mu d}}{N_{MC}}$ and $N_{\mu d}$ the total number of unfolded events.

This method allows an efficient improvement of the folding matrix, without introducing spurious fluctuations. Actually, the amplification of small fluctuations can be prevented at this step of the procedure as well.

The number of iterations in the unfolding is chosen such that the residual bias, evaluated through a data-driven closure test (described below), is within a tolerance of 1% in the bins with less than 10% statistical uncertainty. In these measurements this is achieved with only one iteration.

Closure Test

The transfer matrix was improved by one iteration to account for shape differences between data and MC on the m_{jj} spectra. In addition, a data-driven closure test was performed to account for any residual MC mismodelling.

In this study the part level m_{jj} spectrum in the MC simulation is reweighted and convolved through the folding matrix. The weights are chosen such that significantly improved agreement between the resulting reconstructed spectrum and data is attained. The reconstructed spectrum in this reweighted MC is then unfolded with the nominal transfer matrix as used for the data. Comparison of this result with the reweighted part level spectrum provides an estimate of the bias, which is interpreted as the systematic uncertainty. Figure 5.4 shows the uncertainty for the central y^* bin.

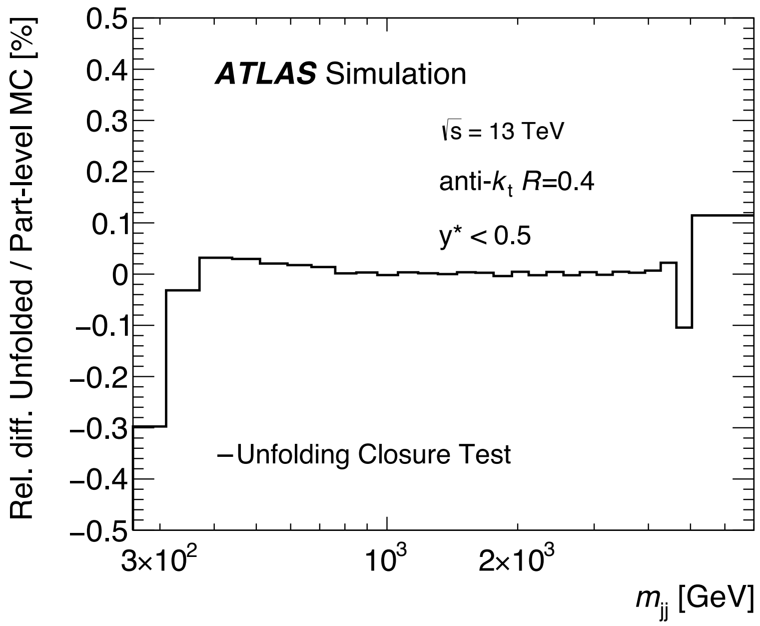


Figure 5.4: Unfolding systematic uncertainty obtained from the data-driven closure test for the central y^* bin.

The bias uncertainty were under 1% for all y^* bins. Due to lack of data and MC statistics, particularly at high y^*, a smoothing using a Gaussian kernel technique was applied, which is detailed in Appendix B.

5.0.17 Estimation of Statistical Uncertainties: Bootstrapping

Due to the complexity of data analysis methods (e.g: the unfolding) it may result practically imposible to compute the statistical uncertainty of estimates analytically. One option would be to collect more data and repeat the experiment multiple times, or divide the sample space in N subsamples and compute the variability σ in order to get the error as $\sigma/\sqrt{N}$. The first approach is not possible since we have limited data. The second one will be a poor estimate of the variance, in the sense that the result will present high statistical fluctuations. A more intelligent solution, which does not need to divide in subsamples (or collecting more data), is the *bootstrapping* technique, a computational method that simulates new samples, to help to determine how estimates from replicate experiments might be distributed and so allow to answer questions about precision and bias. The bootstrap technique used here does not make any assumption on the underlying distributions of the observables, and so it can be used generally for any kind of observables. The method consists basically in generating N replicas (toys) from the original sample by weighting each event by a random variable k which comes from a *Poison* distribution with $\mu = 1$.

$$\text{event}_j^{toy} = \text{event}_j \times k$$

We obtain in this way, the nominal distribution ($k = 1,\ \forall j$) and N replicas of the same observable. The unfolding is performed for each toy and a covariance matrix is constructed for the cross-section in each y^* bin. The total statistical uncertainty for each m_{jj} bin is obtained from the diagonal elements of the covariance matrix (i.e the variance). The separate contributions from the data and from the MC statistics are obtained from the same procedure by varying either the data or the simulated events. Furthermore an overall covariance matrix is constructed to describe the full statistical covariance among all m_{jj}, y^* bins. For this analysis, $N = 1000$ was used.

In order to validate this approach, a MC sample was divided in 32 MC subsamples and the variability of a given observable was computed analitically (by computing the covariance matrix from the 32 subsamples). The result was then compared with the one obtained from bootstrapping one of the MC subsamples. Figure 5.5 shows the result by the first approach (red curve) and the second one (blue), the black curve corresponds to bootstrapping the whole sample, which was multiplied by $\sqrt{32}$ in order to compare with the other curves. As expected, the three curves overlap each other.

The bootstrapping method was used to estimate the statistical uncertainties of data, MC,

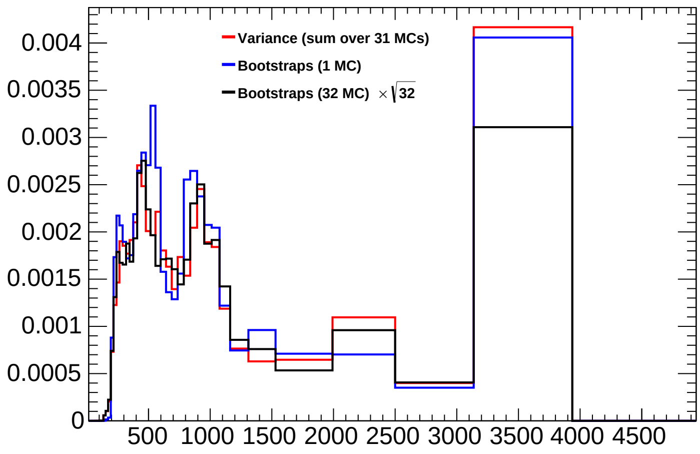


Figure 5.5: Statistical uncertainty of a given observable computed with three different approaches. Analytical calculation of 32 MC subsamples (red curve), bootstrapping of one subsample (blue curve), and bootstrapping of the whole sample (black curve).

and systematic uncertainties, and also to compute the level of agreement between data and theory (see Appendix C).

5.0.18 Propagation of Systematics Uncertainties

Jet Energy Scale

All the components of the JES uncertainty (see Section 4.0.12) are propagated through the unfolding technique. Each jet p_{T} is scaled up and down by one standard deviation of each component.

$$p_T'^{\,\mathrm{jet}} = p_T^{\mathrm{jet}} \pm \Delta$$

where Δ is the jet uncertainty from a given source of systematic.

The p_{T} shift is propagated to the cross-section and the resulted spectra is unfolded using the nominal transfer matrix. The difference between the nominal unfolded cross-section and the systematically shifted unfolded cross-section is taken as a systematic uncertainty.

A breakdown of the propagated JES uncertainties to the dijet cross-section can be seen in Figure 5.6. The dominant contributions come from the LAr, jet flavour, E/P (single-particle response) at high mass, η-Intercalibration and the absolute *in situ* (γ/Z+jets). One point to note is that the total cross-section uncertainty for the highest m_{jj} bin is $\sim 30\%$, which compared to the total p_{T} uncertainty for the last p_{T} bin in Figure 4.8 is $\times 10$ higher. This is related to the steeply falling dijet cross-section as a function of m_{jj}: $\sigma(m_{jj}) = km_{jj}^{-a}$. Which results in:

$$\frac{\Delta\sigma}{\sigma} = a\frac{\Delta m_{jj}}{m_{jj}}$$

where a is the slope of the falling spectra and $\Delta\sigma$ (Δm_{jj}) are the absolute uncertainties of the cross-section and m_{jj} respectively.

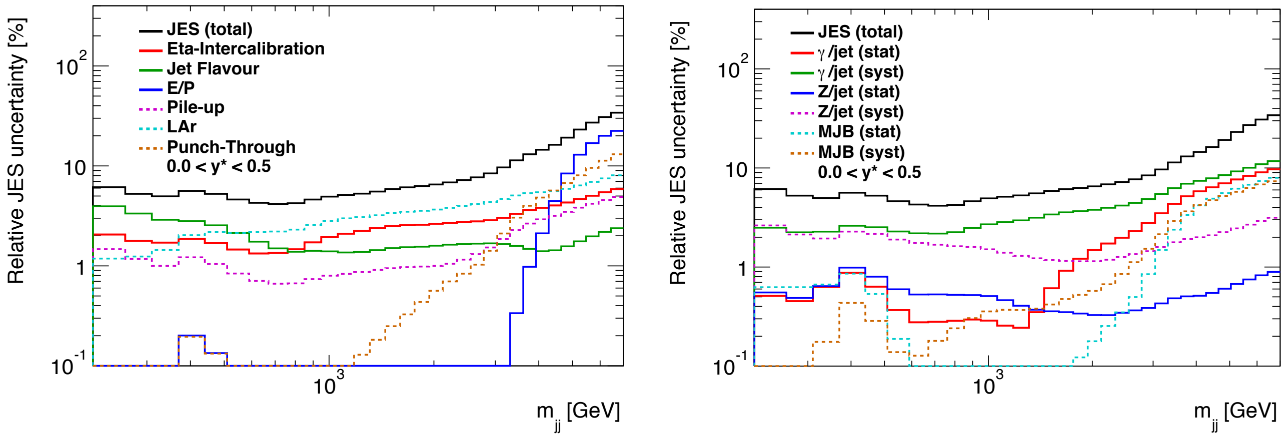


Figure 5.6: Breakdown of relative JES systematic uncertainties in the dijet cross-section for the first y^* bin. As a reference, the black line shows the total uncertainty (sum in quadrature of each contribution).

Jet Energy Resolution

Because the final result relies on the unfolding, an uncertainty on the JER must be taken into account. It was computed based on the JER derived in 2012 and so an additional nuisance parameter (NP) was considered, which captures the changes in detector conditions from 2012 to 2015 and differences in the underlying event tune due to the increase from 8 TeV to 13 TeV in center-of-mass energy.

There are 11 JER systematic uncertainty components, which is the result of 9 NPs from Run 2, 1 reduced NP from the Run1-Run2 cross calibrations and 1 NP for the forward region derived in Run 1. Some of the components are positive in part of the phase-space and negative in the complementary part, which allows to account for (anti-)correlations. This requires an enhancement of the JER in some phase-space region and a reduction of the JER in the complementary part. The effect of each NP is evaluated by smearing the p_{T} of the reconstructed jets in the MC simulation such that the resolution is enhanced (where relevant) by the size of each uncertainty component. To do so, the jet p_{T} is smeared by the factor σ_{smear}, calculated as:

$$\sigma^2_{\mathrm{smear}} + \sigma^2_{\mathrm{nominal}} = (\sigma_{\mathrm{nominal}} + \Delta\sigma)^2$$

where $\sigma_{\mathrm{nominal}}$ is the 2012 fractional resolution and $\Delta\sigma$ the resolution uncertainty for a given NP component.

Since a consistent reduction (on average) of the MC resolution is currently not trivial to implement, an enhancement of the resolution in a MC simulation is performed. The enhancement of the MC JER is accounted by constructing a new transfer matrix using smeared jets and is used to unfold a pseudo-data spectra. The ratio of the unfolded pseudo-data done with the enhanced JER transfer matrix and the nominal MC unfolded spectra is taken as the systematic uncertainty.

Total Experimental Uncertainty

In order to assess the statistical precision of the systematic uncertainties, each component is evaluated using a set of pseudo-experiments from bootstrapping. The statistical fluctuations of the systematic uncertainties are minimised using a smoothing procedure (see Appendix B).

The individual components of the systematic uncertainties are taken as uncorrelated, and are added in quadrature. The systematic uncertainties for the first and last y^* bins are shown in Figure 5.7. The jet energy scale is the dominant uncertainty for $m_{jj} < 4000$ GeV in the first y^* bin. In the complementary regions, including the whole m_{jj} range for the last y^* bin, the dominant source of uncertainty is statistics. The total uncertainty on the dijet measurement is about 5% (10%) at medium m_{jj} of 500-1000 GeV (2000-3000 GeV) in the first (last) y^* bin. The uncertainty increases both towards lower and higher m_{jj} reaching to 6% at low-mjj and 30% at high mjj.

5.0.19 Results

In the following section, the comparisons of the measured dijet cross-sections with the NLO pQCD predictions as a function of m_{jj} for the six y^* bins are shown.

The NLO pQCD predictions are calculated using several PDFs provided by the LHAPDF6 [41] library: the NLO CT14 [58], MMHT 2014 [69], NNPDF 3.0 [38], HERAPDF 2.0 [20] and the NNLO ABMP16 [23] set.

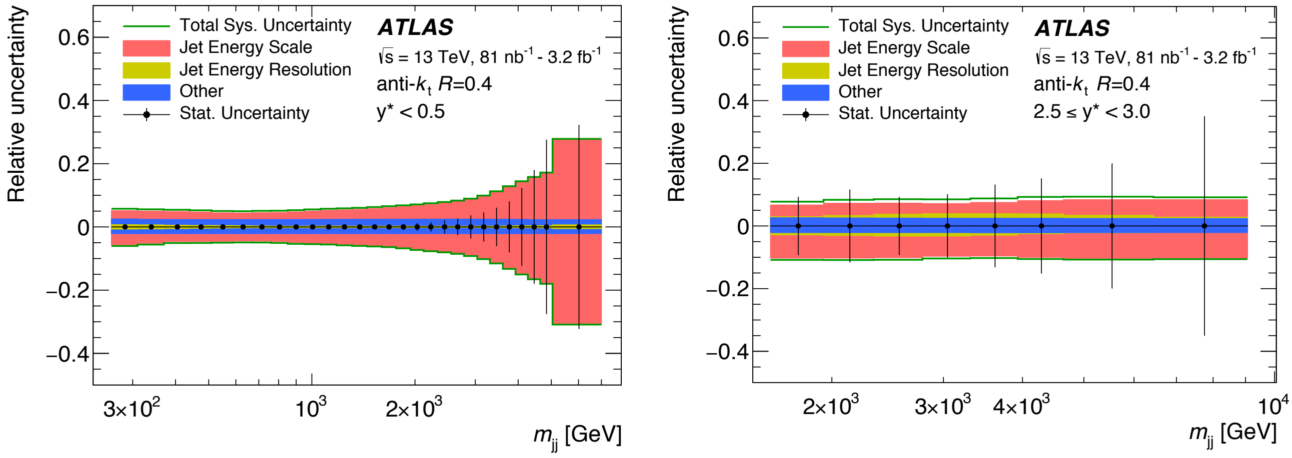


Figure 5.7: Relative systematic uncertainty for the first and last y^* bins (left and right respectively). The total systematic uncertainty is shown as black line. The individual uncertainties are shown in colors: the jet energy scale (red), jet energy resolution (green) and the other uncertainties (jet cleaning, luminosity and unfolding bias) added up in quadrature (blue). The statistical uncertainty is shown as vertical black line.

Qualitative Comparisons

The measured double-differential dijet cross-sections are shown in Figure 5.8. The measurement cover the m_{jj} range from 300 GeV to 9 TeV for $y^* < 3.0$, thus attaining a significantly higher reach than the previous ATLAS measurement [15]. Superimposed are the NLO predictions computed with the CT14 PDF set.

The ratios of the NLO pQCD predictions to the measured dijet cross-sections are shown in Figures 5.9 and 5.10. No significant deviation of the data points from the predictions is seen, the NLO pQCD predictions and data agree within uncertainties.

Quantitative Comparisons

The NLO pQCD predictions, corrected for non-perturbative and electroweak effects, are quantitatively compared to the measurement using the method described in C. The χ^2 value and the corresponding observed p-value, P_{obs}, are computed taking into account the asymmetries and the (anti-)correlations of the experimental and theoretical uncertainties. The individual experimental and theoretical uncertainty components are assumed to be uncorrelated among each other and fully correlated across the m_{jj} and y^* bins. The correlations of the statistical uncertainties across different phase-space regions are taken into account using covariance matrices derived from 1000 pseudo-experiments obtained by fluctuating the data and the MC

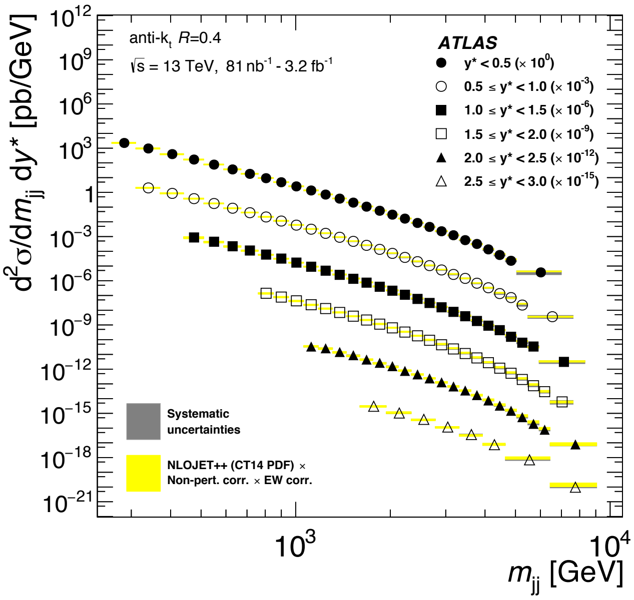


Figure 5.8: Dijet cross-sections as a function of m_{jj} and y^*, for anti-k_t jets with $R = 0.4$. The statistical uncertainties are smaller than the size of the symbols used to plot the cross-section values. The dark gray shaded areas indicate the experimental systematic uncertainties. The data are compared to NLO pQCD predictions calculated using `NLOJET++` with $p_T^{\max}\exp(0.3y^*)$ as the QCD scale and the CT14 NLO PDF set, to which non-perturbative and electroweak corrections are applied. The yellow shaded areas indicate the predictions with their uncertainties. In most m_{jj} bins the experimental systematic uncertainty is smaller than the theory uncertainties and is therefore not visible.

simulation, as described in Section 5.0.17.

Table 5.11 shows the summary of the observed P_{obs} values for each y^* bin of the dijet measurement, as well as those from a global fit using all the m_{jj} and y^* bins. Fair agreement is seen (with p-values in the percent range) when considering events from individual, as well as in aggregate, y^* regions. This is consistent with previous ATLAS measurement at $\sqrt{s} = 7$ TeV [15].

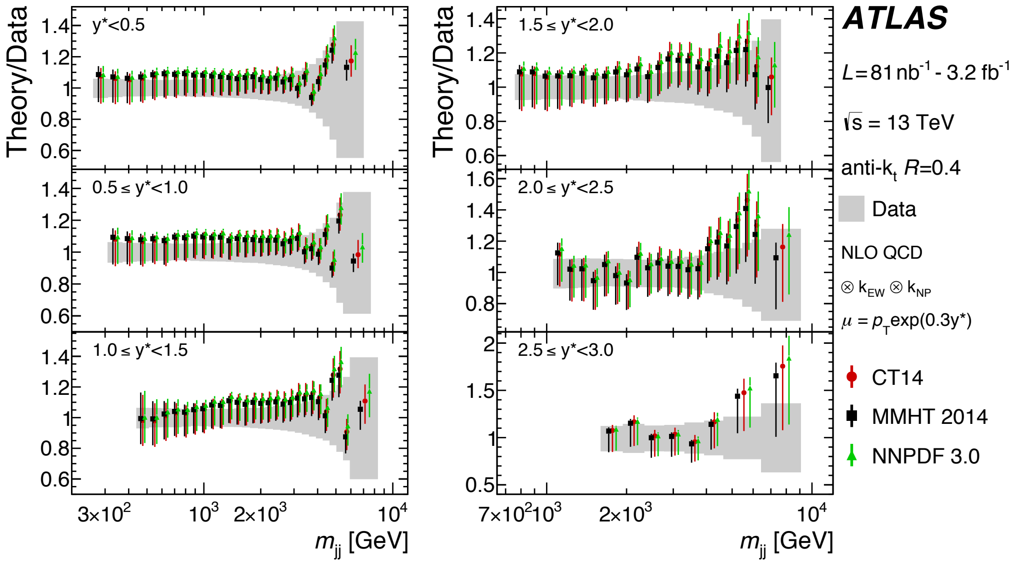


Figure 5.9: Comparison of the measured dijet cross-sections and the NLO pQCD predictions shown as the ratios of predictions to the measured cross-sections. The ratios are shown as a function of the jet m_{jj} in six y^* bins for anti-k_t jets with $R = 0.4$. The predictions are calculated using `NLOJET++` with three different PDF sets (CT14, MMHT 2014, NNPDF 3.0) and non-perturbative and electroweak corrections are applied. The uncertainties of the predictions are shown in coloured lines. The grey bands show the total data uncertainty including both the systematic and statistical uncertainties.

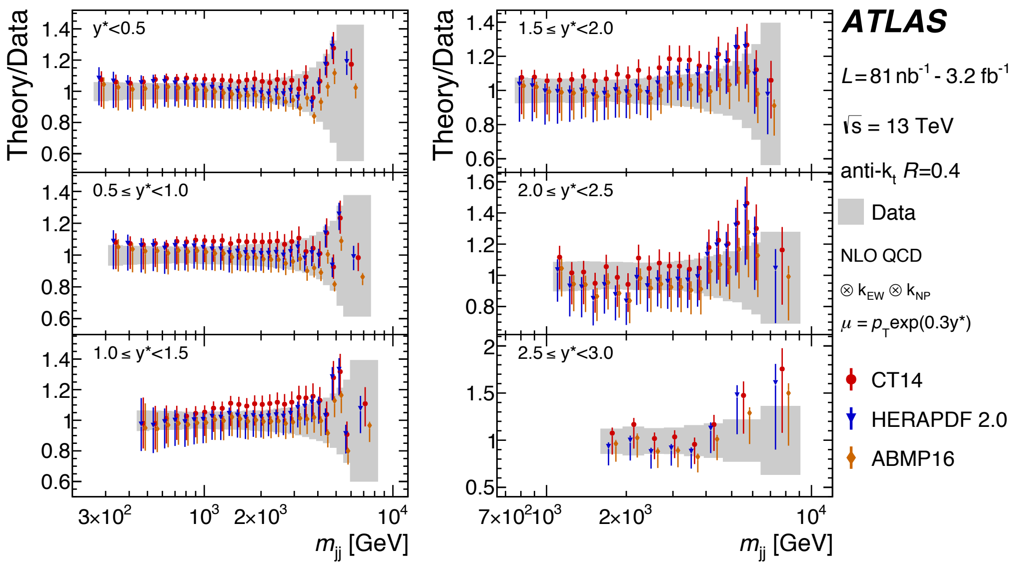


Figure 5.10: Comparison of the measured dijet cross-sections and the NLO pQCD predictions shown as the ratios of predictions to the measured cross-sections. The ratios are shown as a function of the jet m_{jj} in six y^* bins for anti-k_t jets with $R = 0.4$. The predictions are calculated using `NLOJET++` with three different PDF sets (CT14, HERAPDF 2.0, ABMP16) and non-perturbative and electroweak corrections are applied. The uncertainties of the predictions are shown in coloured lines. The grey bands show the total data uncertainty including both the systematic and statistical uncertainties.

y^* ranges	P_{obs} CT14	MMHT 2014	NNPDF 3.0	HERAPDF 2.0	ABMP16
$y^* < 0.5$	79%	59%	50%	71%	71%
$0.5 \leq y^* < 1.0$	27%	23%	19%	32%	31%
$1.0 \leq y^* < 1.5$	66%	55%	48%	66%	69%
$1.5 \leq y^* < 2.0$	26%	26%	28%	9.9%	25%
$2.0 \leq y^* < 2.5$	41%	34%	29%	3.6%	20%
$2.5 \leq y^* < 3.0$	45%	46%	40%	25%	38%
all y^* bins	9.4%	6.5%	11%	0.1%	5.1%

Figure 5.11: Summary of the observed P_{obs} values obtained from the comparison of the dijet cross-section and the NLO pQCD prediction corrected for non-perturbative and electroweak effects for various PDF sets and for each individual y^* range. The last row of the table corresponds to a global fit using all m_{jj} and y^* bins of the measurement.

Redes Neuronales Adversarias para Identificar Jets Masivos (resumen)

Este capítulo comienza con una descripción matemática general del aprendizaje automático. Fundamentando el uso de las redes neuronales y el porque las ventajas de una arquitectura profunda.

En esta tesis se propone el uso de redes neuronales adversarias (ANN) para construir un clasificador de jets masivos (provenientes de quarks top o bosones W) no correlacionados con la masa del jet. Los clasificadores tradicionales utilizados en ATLAS, como redes neuronales y arboles de decisión (*boosted decision trees*), presentan correlaciones no lineales con la masa del jet, lo que provoca la aparición de picos espurios en la distribución del ruido (QCD), lo cual conlleva a una reducción de la significancia estadística de un analisis de búsqueda de nueva física. El método propuesto basado en ANN, propone una solución a este tipo de problemas. El mismo es comparado con otras técnicas de descorrelación, como *uBoost*, regresión k-NN, DDT y CSS. Todos estos métodos se basan en variables de sub-estructura de los jets, los cuales son características poderosas que permite distinguir entre jets de señal y de ruido.

El analisis se restringe a la región de masa 50 GeV $< m <$ 300 GeV. El momento transverso del jet se limito al rango $200 < p_{\mathrm{T}} < 2000$ GeV. Con estos criterios cinemáticos se garantiza que el jet resultante sea colimado (de forma de contener las partículas b y W del decaimiento de un quark top). Para el entrenamiento de la red se utilizarion 10^6 eventos de seãl y de fondo. para la evaluación del rendimiento se utilizo un set separado consistente de 10^6 y 10^7 eventos de seãl y de fondo respectivamente.

Para medir el poder de clasificación se utilizaron curvas ROC, en donde se grafica el rechazo del fondo a una dada eficiencia de señal. El punto de trabajo elegido para la optimización fue de 50%. Por otro lado, se midio el poder de descorrelación mediante la divergencia de Kullback-Leibler. La misma mide la entropía relativa entre distribuciones de fondo antes y luego del corte por el clasificador.

La función de costo utilizada para la optimización de los parámetros consiste en dos funciones competitivas entre si. Por un lado una se encarga de optimizar el poder de clasificación (*binary cross-entropy loss*) y la adversarial esta condicionada a la primera para hacer inferencia de la masa del jet. Esto permite introducir una regularicación en los parámetros y restringir los pesos de clasificación a un sub-espacio el cual resulta dificultoso (sino imposible) determinar la masa del jet. De esta forma es que se logra un clasificador independiente del la masa, lo que resulta en una eficiencia de selección de fondo uniforme con la masa del jet, como se muestra en la Figura 6.10.

La performance final se muestra en un gráfico bidimensional, donde en el eje vertical se indica el poder de descorrelación y el eje horizontal el poder de clasificación alcanzado, evaluado en el set de testeo. Se puede observar que ANN logra el mayor poder de clasificación a un poder de descorrelación dado. Por otro lado, el máximo poder de descorrelación se logra con la técnica analítica de regresión k-NN.

Chapter 6

Adversarial Neural Networks for Decorrelating Jet Observables

6.0.20 Theoretical Aspects of Machine Learning

In this section, I will mention some words about Machine Learning, with focus on neural nets. This short intoduction will help to understand Section 6.0.24 in case the reader is not familiar with the field.

Machine Learning (ML) has become a very popular paradigm in computer science during the last decades. The reason of its success is mainly because, in contrast with traditional computer programs, ML does not need explicit rules for a given specific task, but it has the flexibility to *learn* its own set of rules based on experience, some of them might be too complex to program in a computer. What is more, ML has the properties of adaptation and generalization. The first means that the rules it learns can adapt easily for a given change in the inputs, by retraining on them, and the second that it can learn no by memorization, and so behaving well performant also for unseen data of the same nature. ML is closed related with computational statistics and mathematical optimization, which focus in making data-driven predictions, decisions, statistical and causal inferences.

ML can be divided into three broad groups of algorithms. *Supervised learning* aims to predict an output when given an input vector based on example input-output pairs. In other words, it learns a function that maps an input to an output based on labeled training data consisting of a set of training examples. *Unsupervised learning* discovers good internal representation of inputs that has not been labeled, classified or categorized. *Reinforcement learning* select actions to maximize rewards which may only occur occasionally. We will focus just on supervised learning, which was the technique developed in this thesis.

Supervised learning deals with two kind of problems. In *regression* the target output is a real number. Given a training set of points x with its labels y, $(x_i, y_i)_{i=1}^n$ with $(x_i, y_i) \in \mathbb{R}^N \times \mathbb{R}$, the task of the learning algorith is to determine a function $f : \mathbb{R}^N \to \mathbb{R}$ which would be "close" to the corresponding y value for a new x example. In *classification* the target output is a class label, and so the $f(x)$ will predict an integer number that represents some category, $f : \mathbb{R}^N \to \mathbb{Z}$.

Mathematical Formulation

For simplicitly, lets restrict to a classification problem. Let us assume that the pairs of examples-categories are modeled as random variables:

$$(X, Y) \sim p_{XY}(x, y), (x, y) \in \mathcal{X}, \mathcal{Y} \tag{6.1}$$

with $\mathcal{Y} = 1, 2, ..., M$ a finite set of labels and $p_{X,Y}(x, y)$ a probability model, which in general is unknown. The goal is to find the best estimator $\hat{y} : \mathcal{X} \to \mathcal{Y}$, known as the *decision rule*, such that

$$\mathcal{P}_{XY}\{\hat{y}(X) \neq Y\} := \int_{\mathcal{X}} p_X(x) P_{Y|X}(Y \neq \hat{y}(X)|X = x)dx \tag{6.2}$$

is small as possible. The *Bayes classifier theorem* guarantees the existence of the best solution for the stated problem. The optimal decision rule $\hat{y}^*(X)$ is given by:

$$\hat{y}^*(X) = \arg\max_{y \in \mathcal{Y}} P_{Y|X}(Y = y|X = x), \text{ where } P_{Y|X}(Y = y|X = x) = \frac{p_{XY}(x, y)}{\sum_{y \in \mathcal{Y}} p_{XY}(x, y)} \tag{6.3}$$

The optimal decision function $\hat{y}^*(X)$ is also known as *maximum a posteriori* (MAP) rule. From Bayes theorem, we have that:

$$P_{Y|X}(y|x) = \frac{P_Y(y)}{p_X(x)} p_{X|Y}(x|y) \tag{6.4}$$

It follows that if all classes $y \in \mathcal{Y}$ have uniform prior, i.e. $P_Y(y) = \frac{1}{M}\ \forall y \in \mathcal{Y}$, the MAP rule becomes the *maximum likelihood* decision rule. In the simplest case of binary classification, i.e $M = 2$, the maximum likelihood solution will be a particular one from a likelihood ratio test [2].

Since the true $p_{XY}(x, y)$ is not known in practice, otherwise no machine learning algorirthm would be needed at all, the solution to the problem corresponds to the determination of $M - 1$ functions $\eta_i^* : \mathcal{X} \to [0, 1]$ defined as $\eta_i^*(x) := P_{Y|X}(y = i|x)$ with $i = 1, \ldots, M$. For the $M = 2$

case, the optimal rule is determined by $\eta^*(x) = P_{Y|X}(y = 1|x)$. The decision problem can be cast as:

$$\text{If } \eta^*(x) > \frac{1}{2} \text{ decide for } y = 1. \text{ Otherwise decide for } y = 0. \tag{6.5}$$

The curve defined by $\eta^*(x) = \frac{1}{2}$ is the *decision boundary* in $\mathcal{X}$.

In order to find these functions, a parametric model is used:

$$\alpha(x;\theta) = \sum_{l=1}^{L} \theta_l f_l(x) + \theta_{L+1} = \boldsymbol{\theta}^T \boldsymbol{f}(x), \boldsymbol{\theta} \in \mathbb{R}^{L+1} \tag{6.6}$$

Where $\eta^*(x) = \sigma(\alpha^*(x))$ and $\sigma(x) := \frac{1}{1+\exp(-x)}$ is the *logistic sigmoid* function. The functions $\{f_l(x)\}_{l=1}^{L}$ are some fixed basis (chosen in advance) where in general $L \geq K$. With this parametric model, known as *logistic regression*, one looks for the best function $\alpha(x, \hat{\theta})$ that explains the training data. In this sense $\hat{\theta}$ is a function of the training data $(x_i, y_i)_{i=1}^{n}$. By setting $L = K$ and $f_l(x) = x_l$ for each $l = 1, \ldots, K$, with some proper rule for selecting the coefficients $\{\theta_i\}_{i=1}^{K+1}$ in terms of the training set, we obtain the different *linear discriminant models*, e.g., Fisher discriminant rule.

Given a training set $(x_i, y_i)_{i=1}^{n}$ and functions $\{f_l(x)\}_{l=1}^{L}$, a natural approach to find the coefficients $\{\theta_i\}_{i=1}^{L+1}$ would be to minimize the *empirical risk* defined as:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^{L+1}} \sum_{i=1}^{n} \mathbb{1}(\hat{y}(x_i;\theta) \neq y_i) \tag{6.7}$$

where for binary classification $\hat{y}$ is defined, in accordance to the decision boundary 6.5, as:

$$\hat{y}(x_i;\theta) = \begin{cases} 1 \text{ if } \eta(x;\theta) > \frac{1}{2} \\ 0 \text{ otherwise} \end{cases} \tag{6.8}$$

This criterion look for the parameters that minimize the missclassification error in the training set. Theoretical analysis of this criterion is very important. However, from the practical point of view it is seldom used because of its computational complexity specially when L and n are large. For this reason, in practice a *surrogate loss* function is used. Assuming the training set to be independent and identically distributed random variables according to some $p_{X,Y}(x, y)$, the true conditional probability of $\{y_i\}_{i=1}^{n}$, where y_i can takes only two possible outcomes 1 or 0, given $\{x_i\}_{i=1}^{n}$ can be written as:

$$P_{Y_1,\ldots,Y_n|x_1,\ldots,x_n}(\{y_i\}_{i=1}^{n}|\{x_i\}_{i=1}^{n}) = \prod_{i=1}^{n} [P_{Y|X}(1|x_i)]^{y_i}[1 - P_{Y|X}(1|x_i)]^{1-y_i} \tag{6.9}$$

As we look for $\eta(x;\theta)$ to be a good approximation to the true $P_{Y|X}(1|x)$ we can set the following as the surrogate criterion:

$$\hat{\theta} = \arg\max_{\theta \in \mathbb{R}^{L+1}} \prod_{i=1}^{n} \Big[\sigma\big(\boldsymbol{\theta}^T \boldsymbol{f}(x_i)\big)\Big]^{y_i} \Big[1 - \sigma\big(\boldsymbol{\theta}^T \boldsymbol{f}(x_i)\big)\Big]^{1-y_i} \tag{6.10}$$

By taking the *log* and multiplying by -1, it is obtained the *empirical risk function*, also known as the *binary cross-entropy loss*:

$$\mathcal{L}\Big(\{(x_i, y_i)\}_{i=1}^{n}; \boldsymbol{\theta}\Big) = -\sum_{i=1}^{n} y_i \log\Big(\sigma\big(\boldsymbol{\theta}^T \boldsymbol{f}(x_i)\big)\Big) + (1 - y_i)\log\Big(1 - \sigma\big(\boldsymbol{\theta}^T \boldsymbol{f}(x_i)\big)\Big) \tag{6.11}$$

The loss function in this case is differentiable (in fact infinitely dfferentiable) and can be optimized efficiently using iterative methods (as *gradient descent* or the *Newton-Rapshon* methods). Notice that the functions $\{f_l(x)\}_{l=1}^{L}$ are a fixed basis predefined beforehand. In the next section we will see how this basis can also be adapted with training data using neural nets.

Why Neural Nets?

From the mathematical point of view, neural nets (NNs) became so popular because its capacity of estimating or learning very complex functions, as it usually happen in real optimization problems. In the previous section we considered parametric families of the form:

$$\mathcal{F} = \left\{ \sum_{l=1}^{L} \theta_l f_l(x) + \theta_{L+1} : (\theta_1, \ldots, \theta_{L+1}) \in \mathbb{R}^{L+1} \right\} \tag{6.12}$$

where $\{f_l(x)\}_{l=1}^{L}$ are some fixed functions (chosen in advance). As we consider only linear combinations of functions $\{f_l(x)\}_{l=1}^{L}$, $\mathcal{F}$ is the subspace generated by those functions. If the true objective function is not in this subspace then it will not be reconstructed exactly. A possible solution would be to also learn from the data the basis functions $\{f_l(x)\}_{l=1}^{L}$. As it will seen shortly, this is one of the things NNs do. Lets show this in a *multilayer perceptron*, also known as *feed-forward* neural-network. We will make the functions $\{f_l(x)\}_{l=1}^{L}$ to depend on parameters and allow to adjust them along with $\{\theta_i\}_{l=1}^{L+1}$ during training. This can be described by a series of functional operations. We start from the input $\boldsymbol{x} \in \mathbb{R}^M$ and construct L_1 linear combinations from them:

$$u_{1i} = \boldsymbol{W}_{1i}^T \boldsymbol{x} + b_{1i},\ i = 1, \ldots, L_1 \tag{6.13}$$

where $\{\boldsymbol{W_{1i}}\}_{i=1}^{L_1} \in \mathbb{R}^M$ and $\{b_{1i}\}_{i=1}^{L_1} \in \mathbb{R}$ are *weights* and *biases* from the first layer of the network to be optimized. The quantities u_{1i} are known as the *activations*. Each of the activations are transformed using a differentiable, **nonlinear** *activation function* $g : \mathbb{R} \to \mathbb{R}$ to give:

$$x_{2i} = g(u_{1i}),\ i = 1, \dots, L_1 \tag{6.14}$$

These outputs are known as the *hidden* units. This values are again linearly combined to obtain the activations of the second layer:

$$u_{2i} = \boldsymbol{W}_{2i}^T \boldsymbol{x_2} + b_{2i},\ i = 1, \dots, L_2,\ \{\boldsymbol{W_{2i}}\}_{i=1}^{L_2} \in \mathbb{R}^{L_1},\ \{b_{1i}\}_{i=1}^{L_2} \in \mathbb{R} \tag{6.15}$$

In a NN with just one hidden layer the values $\{u_{2i}\}_{i=1}^{L_2}$ are called the *output units activations* and conform the last layer of the network. Finally, this units are transformed by another activation function $h : \mathbb{R} \to \mathbb{R}$ to give the set of network predictions y_k with $k = 1, \dots, K$. For the binary classification problem the function h is a sigmoid. The output of the network $\hat{y}$ as a function of the inputs $\boldsymbol{x}$, for the simple case where $K = L_2 = 1$, is then:

$$\hat{\boldsymbol{y}} = \sigma\left(\sum_{j=1}^{L_1} w_{2j}\, g\Big(\sum_{i=1}^{M} w_{1ji} x_i + b_{1j}\Big) + b_2\right) \tag{6.16}$$

Comparing this expression with the previous one in 6.12 it can be realized that NNs induce also a parametrization on the basis functions.

The case explained before is known as the *single hidden layer network*. An illustrative workflow diagram can be shown in Figure 6.1.

From this figure, it can be seen how the information is propagated *forward* from the input layer vector $\boldsymbol{x}$ to the linear combination resulting in the vector $\{u_{1i}\}_{i=1}^{L_1}$ passing through the activation function to get the hidden units vector $\boldsymbol{x_2}$ and so on to get finally the outputs $\{y_k\}_{k=1}^{K}$. When $K = 1$ we are in the binary-classification case, otherwise this belongs to the multi-classification problem. The same procedure can be repeated through many hidden layers which results in a *deep neural network*. An intuition of why deep nets have become much more succesfull than the single layer *shallow* nets will be shown in the following section.

Why "Deep" Nets?

A deep neural network is one in which there is more than one hidden layer. A workflow representation is shown in Figure 6.2.

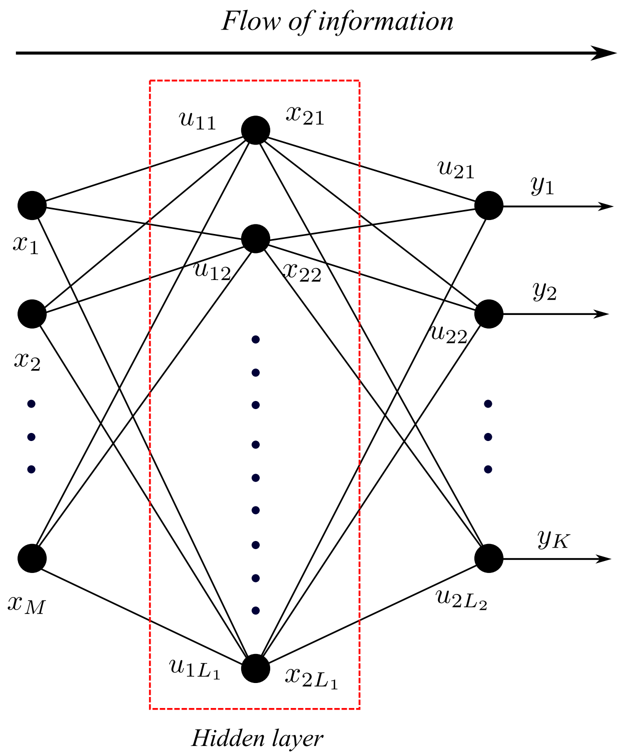
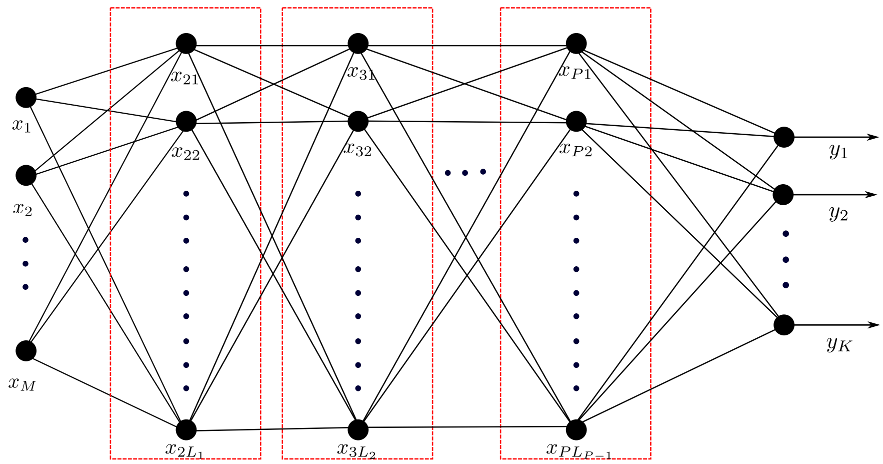


Figure 6.1: Workflow of a single hidden layer feed-forward network for a multi-classification problem.



Network with P hidden layers.

Figure 6.2: Workflow of a deep (P hidden layers) feed-forward network for a multi-classification problem.

As one can expect, the output of a deep net will be a more complex function of the inputs, given by many nested nonlinear functions:

$$\hat{\boldsymbol{y}} = h(g(g(\dots g(\boldsymbol{x}; \boldsymbol{W_1}) \dots ; \boldsymbol{W_{P-1}}); \boldsymbol{W_P}); \boldsymbol{W_{out}}) \tag{6.17}$$

One of the most success of deep representations is its flexibility to learn very complex functions. Lets take as an example the problem of imaging recognition. Intuitively, one can think of earlier layers of the NN as detecting simple features in small portions of the image, like edges or orientation, and then composing them together in the later layers so that it can learn more and more complex features to build up a fully reconstructed image. This visualization is indeed what happens in a *Convolutional Net* used for imaging recognition. This type of simple to complex hierarchical or composition representation applies in other types of data than images and it was the main quality which made deep nets succeed. It is worth to add that in order to achieve this, the activation functions must be nonlinear.

Another motivation of deep nets comes from circuit theory, about what types of functions one can compute with different logic gates. In short, there are functions which can be computed with "small" (number of units) L-layer deep net that shallower nets would require exponentially more hidden units to compute in order to reproduce the same function. As an example lets suppose one want to compute the exclusive OR (XOR) of the parity of all the input features: x_1 XOR $x_2 \ldots$ XOR x_n. A deep net arquitecture, i.e a tree directed structure, would require $\mathcal{O}(\log n)$ number of nodes among all hidden layers. On the contrary, a single layer net would need $\mathcal{O}(2^n)$ units in the hidden layer to exhaustively enumerate all 2^n possible configurations of the input bits.

6.0.21 Analysis Introduction

Hadronic decay of heavy particles, such as top quarks or W bosons, are very important processes for searches of new physics. As pointed in Sec. 4.0.10, such resonances result in collimated decay products reconstructed as a large-radius jet in the calorimeter. Multijet production (QCD) represent a vast background which makes the identification of jets originating from heavy particle decay a crucial and a non-trivial task to do.

Multivariate analysis (MVA) approaches using jet substructure variables , in particular boosted decision trees (BDT) and neural networks (NN) have been shown improvementsin the identification of W bosons with respect to single jet substructure variables [34, 10]. Nevertheless, jet substructure variables exhibit non-linear correlations with the reconstructed large-R jet mass, allowing MVA classifiers to learn the mass as a powerful feature for discriminating against the non-resonant background without even using it as an input. As a result, fixed-value cuts on such MVA classifier observables tend to distort the shape of the large-R jet mass distribution for non-resonant jet production, making it resemble the resonance jet mass distribution [34]. This effect leads to systematic uncertainties which cannot be simply characterized or controlled and results in a reduced of discovery significance. The desire for optimal discrimination and optimal decorrelation are naturally at tension with each other.

A novel technique based in adversarial neural networks (ANN) is presented as a way to leverage classification power while reducing their correlation with the jet mass. This method is compared against analytical, single-variable decorrelated taggers as well as MVA-based ones. In particular, designed decorrelated taggers (DDT) [56], fixed-efficiency k-nearest neighbours (k-NN) regression [57], and convolved substructure (CSS) [80] provide analytical methods for decorrelating a single jet substructure observable from the jet mass. Similarly, adaptive boosting for uniform efficiency (uBoost) [96] is presented as an extension of standard BDT, as a MVA technique to manage decorrelation power as well.

The input preparation is described briefly in Sec. 6.0.22. In Sec. 6.0.23, the metrics chosen to quantify tagger classification power and mass-decorrelation are detailed. Sec. 6.0.24 focuses in describing the adversarial analysis technique, the rest decorrelated methods are briefly introduced in the Appendix. D. Finally, results for the various techniques are presented in Sec. 6.0.26, evaluated according to multiple relevant metrics.

6.0.22 Sample preparation

The study was done in MC simulation using hadronic decay of W bosons coming from exotic Z' particles as the signal and multijets QCD as background. The analysis was restricted to the range 50 GeV $< m <$ 300 GeV and a reconstructed transverse momentum in the range $p_{\mathrm{T}} \in [200, 2000]$ GeV, where boosted final states is guaranteed and resonances searches with masses below 500 GeV are being performed [11]. Additional kinematic cuts were applied to ensure that all substructure moments are well-defined. After this baseline selection, 10^6 signal and 10^6 background jets each are retained for training, ensuring balance between the classes. In addition, a separate set of 10^6 signal and 10^7 background jets are used for final performance evaluation, in order to ensure sufficient statistics for differential studies.

6.0.23 Evaluation Metrics

Classification

Classification performance is measured by the relative selection efficiency of signal jets $\varepsilon^{\mathrm{rel}}_{\mathrm{sig}}$ and the associated relative background rejection factor $1/\varepsilon^{\mathrm{rel}}_{\mathrm{bkg}}$ for cuts on the classification variables.

These measures are defined as:

$$\varepsilon^{\text{rel}}_{\text{sig}} = \frac{N^{\text{tagged}}_{\text{sig}}}{N^{\text{total}}_{\text{sig}}}, \qquad \frac{1}{\varepsilon^{\text{rel}}_{\text{bkg}}} = \frac{N^{\text{total}}_{\text{bkg}}}{N^{\text{tagged}}_{\text{bkg}}} \tag{6.18}$$

where N^{total} refers to the total number of events of a given class passing the baseline selection and N^{tagged} refers to the number of events passing both the baseline selection and the tagging cut(s). The (relative) background rejection at $\varepsilon^{\text{rel}}_{\text{sig}} = 50\%$ is used as a summary metric to evaluate the classification performance, where 50% is chosen as a typical working point for jet tagging studies [34, 10] which is also indicative of the level of selection efficiency used by searches in this kinematic regime [48, 35].

Decorrelation

The linear correlations between the jet mass, substructure variables, and MVA jet classifier outputs were studied in Ref. [34]. However, a linear correlation coefficient has insufficient capacity to express the highly non-linear correlations observed between the jet mass and existing MVA jet taggers [34, 10]. Instead, a figure of merit which directly quantifies the sculpting of the background jet mass distribution, resulting from a cut on a jet tagger variable, is proposed.

The Kullback-Leibler divergence [72] for discrete probability distributions P, Q is defined as:

$$\text{KL}(P \parallel Q) = -\sum_i P_i \log_n \left(\frac{Q_i}{P_i} \right) \tag{6.19}$$

where i enumerates the discrete bins of the distributions. The Kullback-Leibler divergence measures the relative entropy of P with respect to Q, and can therefore be used to quantify the similarity between discrete distributions.

To avoid divergences and to symmetrise the metric with respect to P and Q, the Jensen-Shannon divergence [75] is used:

$$\text{JSD}(P \parallel Q) = \frac{1}{2}\big(\text{KL}(P \parallel M) + \text{KL}(Q \parallel M)\big), \qquad \text{with} \quad M = \frac{P+Q}{2} \tag{6.20}$$

In this study, the Jensen-Shannon divergence is used to measure the difference between the normalised mass distributions of the background jets passing and failing, respectively, a given jet tagger cut:

$$\text{JSD} \equiv \text{JSD}\left(N^{\text{pass}}_{\text{bkg}}(m) \,/\, \textstyle\sum_i N^{\text{pass}}_{\text{bkg},i}(m) \,\big|\big|\, N^{\text{fail}}_{\text{bkg}}(m) \,/\, \textstyle\sum_i N^{\text{fail}}_{\text{bkg},i}(m)\right) \tag{6.21}$$

For brevity, this metric will be referred to simply as JSD for any given signal or background selection efficiency. Using a logarithm base $n = 2$ in Equation (6.19), JSD will be in the range

[0, 1] with smaller values indicating a smaller relative entropy between the jet mass distributions and therefore a greater degree of mass-decorrelation; and *vice versa*. For summary performance evaluation, JSD or 1/JSD at $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$ will be used to quantify the mass-decorrelation of a given jet classifier.

6.0.24 Adversarial Neural Networks

It was proposed in [91] the use of adversarial neural networks to build a decorrelated discriminant with the jet mass. This technique is based on *Generative Artificial Networks* (GANs) paradigm, a supervised learning algorithm created by Google in 2014 [62]. It was already proposed in ATLAS as a way to reduce the sensitivity with systematic uncertainties of a given predictive model [76].

GANs have become a very powerful technique within artificial intelligence and its application is growing very fast within high energy physics (HEP) and beyond. A particular example to HEP is the speed up of MC detector simulations, in order to meet the growing analysis demands, without major losses in precision [83]. It was found recently to have promising results also in signal processing, such as deconvolution of astrophysical images, demonstrating the possibility of going beyond the deconvolution limit imposed by the Shannon–Nyquist sampling theorem [71]. GANs has been shown to succesfully deal with inverse problems as the unfolding approach described in Sec. 5.0.16.

The use of GANs as a decorrelated method is the first time tried in ATLAS. As it will be seen shortly, this technique is a generalization of a standard feedforward neural network used for classification. What is more, the problem can be stated formally as follows:

Given the data $x \in \mathcal{X}$, $y \in \mathcal{Y}$ the target, and $d \in \mathcal{D}$ an observable in which the *mutual information* shared with x is different from zero. In other words, $p(x, d) \neq p(x)p(d)$, where $p(x, d)$ is the joint distribution of x and d. One wish to learn a predictive function $f : \mathcal{X} \to \mathcal{Z}$ with parameters θ_f, where $z \in \mathcal{Z} = \mathbb{R}^{|\mathcal{Y}|}$ corresponds to the categorical scores used in mapping predictions $y \in \mathcal{Y}$, such that:

$$p(f(x; \theta_f) = z|d) = p(f(x; \theta_f) = z|d')$$

for all d, $d' \in \mathcal{D}$ and all values $z \in \mathcal{Z}$ of $f(x; \theta_f)$. In words, one looks for a predictive function f which is a *pivotal quantity* with respect to the observable d. In the particular case of our problem, x represents the jet substructure features, y the label of a jet ($y = 1$ for signal and $y = 0$ for background), z the predictions and d the jet mass observable m.

To find such pivotal quantity in terms of a configuration in θ_f parameters, two feedforward neural networks, the second conditioned to the first one, are used. The firts net is a standard classifier parametrised in θ_f. It receives N substructure variables as input and outputs a jet tagging variable in the range $[0, 1]$. Its purpose is to minimise the classification loss L_{clf} with respect to θ_f, which introduces the correlation with the jet mass. The second net, called the adversary aims to infer the jet mass from the output of the classifier, minimising its own loss L_{adv} with respect to the adversary network weights θ_{adv}. If the adversary is able to infer the jet mass from the classifier output beyond random guessing, some generally non-linear correlation must exist between the two. Since we are interested in a mass-decorrelated tagger for background jets, the adversary loss L_{adv} is evaluated only for multijets, as it is the case for the mass-decorrelation metric in Eq.[].

The combined neural network is trained by min-max optimisation of the joint objective:

$$\min_{\theta_{\mathrm{clf}}} \max_{\theta_{\mathrm{adv}}} L_{\mathrm{clf}}(\theta_{\mathrm{clf}}) - \lambda L_{\mathrm{adv}}(\theta_{\mathrm{clf}}, \theta_{\mathrm{adv}}) \tag{6.22}$$

balancing the classification task (first term) with the decorrelation task (second term). The adversary loss L_{adv} depends on the weights of both networks, since the adversary's task is conditional on the classifier configuration. The inner optimisation ($\max_{\theta_{\mathrm{adv}}}$) leads the adversary to minimise $L_{\mathrm{adv}}(\theta_{\mathrm{clf}}, \theta_{\mathrm{adv}})$ with respect to the adversary network weights θ_{adv}. The outer optimisation ($\min_{\theta_{\mathrm{clf}}}$) leads the classifier to minimise the effective loss $L_{\mathrm{effective}}(\theta_{\mathrm{clf}}, \theta_{\mathrm{adv}}) = L_{\mathrm{clf}}(\theta_{\mathrm{clf}}) - \lambda L_{\mathrm{adv}}(\theta_{\mathrm{clf}}, \theta_{\mathrm{adv}})$ with respect to the classifier network weights θ_{clf}.

The difference in sign between the two terms means that the classifier is trained to maximise L_{adv}, making it harder for the adversary to infer the jet mass. The trade-off between the two competing objectives is controlled by the regularization parameter λ: For $\lambda \to \infty$, the classifier is allowed only to retain information "orthogonal" to the jet mass; for $\lambda \to 0$ the standard neural network classifier is recovered. The adversarial neural net architecture is in this way a generalization of a traditional classifier. The parameter λ can be considered as an additional hyperparameter of the adversarial neural network, to be optimised according to some use case–specific figure of merit, such as a search sensitivity metric which would depend on a specific analysis.

Implementation

The classifier objective, parametrised by weights θ_{clf}, is defined as a binary cross-entropy loss optimised to perform binary classification of the jet labels Y based on input features X:

$$L_{\mathrm{clf}}(\theta_{\mathrm{clf}}) = \mathbb{E}_{x\sim X}\,\mathbb{E}_{y\sim Y}\left[-y\log p_{\mathrm{clf}}(y\,|\,x,\theta_{\mathrm{clf}}) - (1-y)\log\big(1-p_{\mathrm{clf}}(y\,|\,x,\theta_{\mathrm{clf}})\big)\right] \tag{6.23}$$

On the other hand, the task of the adversary net is to infer the jet mass, which is implemented by parametrising a probability density function (p.d.f.) in m conditioned to the classifier output $z = p_{\mathrm{clf}}(x,\theta_{\mathrm{clf}})$ and an auxiliary variable $a = \log p_{\mathrm{T}}$[1].

The output of the adversary is therefore the conditional probability $p_{\mathrm{adv}}(m\,|\,z,a,\theta_{\mathrm{adv}})$, evaluated at the value m for each jet in the training sample. The adversary, tasked with decorrelating the classifier from the general set of features D, is trained with the heuristic loss [68]:

$$L_{\mathrm{adv}}(\theta_{\mathrm{clf}},\theta_{\mathrm{adv}}) = \mathbb{E}_{z\sim p_{\mathrm{clf}}(X,\theta_{\mathrm{clf}})}\,\mathbb{E}_{d\sim D}\,\mathbb{E}_{a\sim A}\left[-\log p_{\mathrm{adv}}(d\,|\,z,a,\theta_{\mathrm{adv}})\right] \tag{6.24}$$

This loss is only computed for background jets. In practice, the posterior $p_{\mathrm{adv}}(m\,|\,z,a,\theta_{\mathrm{adv}})$ is constructed by specifying the coefficients $\pi_k(x)$, means $\mu_k(x)$ and width $\sigma_k^2(x)$ components of a *Gaussian Mixture Model* (GMM) [39], such that:

$$p(t=m|x) = \sum_{k=1}^{N_{\mathrm{GMM}}} \pi_k(x)\mathcal{N}\big(t=m|\mu_k(x),\sigma_k(x)\big)$$

where $x = (z,a,\theta_{\mathrm{adv}})$ represents the advesarial input.

The classifier and adversary neural networks are connected in a single, feed-forward architecture as shown in Figure 6.3. Both models are constructed as densely connected networks in `keras` 2.1.5 [46] using the `tensorflow` 1.4.1 backend [18].

The classifier network consists in three hidden layers, each with 64 nodes with rectified linear unit (ReLU) activation, and a single output node with sigmoid activation which represents the classification probability.

The adversary consists in a single hidden layer comprising 64 nodes with ReLU activation, parametrising a GMM posterior p.d.f. $N_{\mathrm{GMM}} = 20$ components was found to have suffcient capacity to perform the decorrelation. To model the mixing coefficients $\pi_k(x)$ a *softmax* activation function was used, which ensures that $\sum_k \pi_k(x) = 1$. For the means and widths, the

[1] This parametrisation yields robust results as a function of jet p_{T} as well as ensuring a better mass-decorrelation.

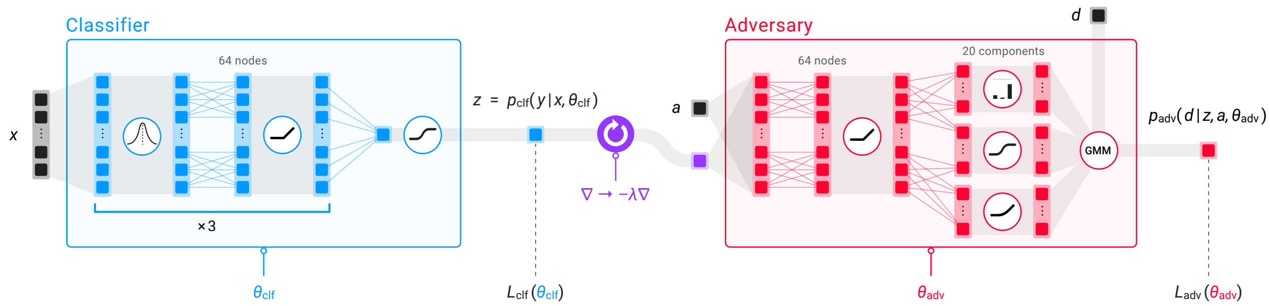


Figure 6.3: Adversarial neural network architecture. The classifier network is tasked with predicting jet labels (y) based on jet substructure variable inputs (x), outputting a tagger variable (z). The adversary network is tasked with inferring the value(s) of the variables from which the classifier is to be decorrelated (d; here the jet mass m), optionally aided by auxiliary features (a; here $\log p_{\mathrm{T}}$), by parametrising a posterior p.d.f. as a Gaussian mixture model (GMM). The adversarial training is implemented using a gradient reversal layer, the trade-off between L_{clf} and L_{adv} controlled by the parameter λ.

tanh and *softplus* activation functions were used, both make sure that $-1 < \mu_k(x) < 1$ and $\sigma_k(x) > 0$, respectively.

For the implementation of the joint optimisation, a gradient reversal layer [67] is used. It consists in applying a gradient scaling operation to the connection between the classifier and the adversary. In the feed-forward mode it acts as the identify operations and during back-propagation the gradient from the adversary is scaled by $-\lambda$. The minus sign means that the gradient flowing backwards from adversary has the inverse effect for the classifier weights as for the adversary. That is, whereas the gradient acts on θ_{adv} to minimise L_{adv}, it will have the effect of maximising L_{adv} for θ_{clf}.

Feature Engineering

The ten jet substructure variables used in the training of the neural network–based jet classifier are listed in Table 4.1. This choice was motivated from optimisation studies on feature selection performed in Ref. [10] due to the similarity of the problem and the datasets used. In contrast to Ref. [10], the jet mass and p_{T} are not used as input features. Using only substructure moments as input features is intended to provide a more direct comparison with the analytical taggers described in the Appendix. D, all of which are based on single substructure moment as their base observable.

To validate this selection of inputs features, a comparison in terms of classification was performed between a neural network classifier trained on the features in Table 4.1 and one trained on an expanded set of 20 substructure variables, including the nominal ten. No significant improvement in performance was found. Similarly, no improvements were found by performing a principal component analysis (PCA) [84] on either set of inputs.

Hyperparameters Tuning

The optimal architecture and learning configuration, such as learning rate, hidden layers, nodes per hidden layers, etc. for both the classifier and adversary (collectively referred to as hyperparameters) are chosen using Bayesian optimisation with `Spearmint` [93] based on *Gaussian Processes* [85].

The classifier is optimised according to the classification loss L_{clf} using 3-fold stratified cross-validation, training for 50 epochs (pass-throughs of the entire training dataset), and shuffling the order of the training samples between each. In order to ensure stability of the result, the optimisation metric is chosen to be the mean classification loss across validation splits ($L_{\text{clf}}^{\text{val}}$) plus one standard deviation.

The hyperparameters of the adversary network, in contrast with the stand-alone classifier optimisation, cannot be meaningfully optimised according to the loss in Equation 6.24 alone. This loss measures only the capacity of the adversary to construct the posterior p.d.f. for the jet mass, not the quality of the resulting decorrelated tagger. Therefore, to tune the adversarially trained neural network tagger according to expected performance, the optimisation is performed by maximising the metric $1/\varepsilon_{\text{bkg}}^{\text{rel}} + \lambda/\text{JSD}$ computed at $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$ for a fixed λ, chosen to be $\lambda = 10$ since this value is found to yield robust results for mass-decorrelation, as discussed in Section[]. The choice of this metric allows to capture the joint optimisation nature of the problem. Throughout the optimisation, a classifier pre-trained with optimal hyperparameters is used.

Bayesian optimisation has no convergence criteria and is not guaranteed to yield the true, optimal configuration in any finite number of iterations. However, it is capable to find efficiently an optimal configuration with few evaluations at locations carefully selected based on a probability of improvement acquisition function. A tradeoff between *exploitation* (evaluating at configuration points with low mean, sensible to local minimas) and *exploration* (evaluating at points with high variance, leads to under/over-fitting of training data) exist. Therefore, the chosen hyperparameter configuration is deemed to be highly performant but is not guaranteed to be optimal.

The optimisation process, run for 100 Bayesian hyperparameter configuration evaluations in order to check stability for extended training duration, is shown in Figure 6.4.

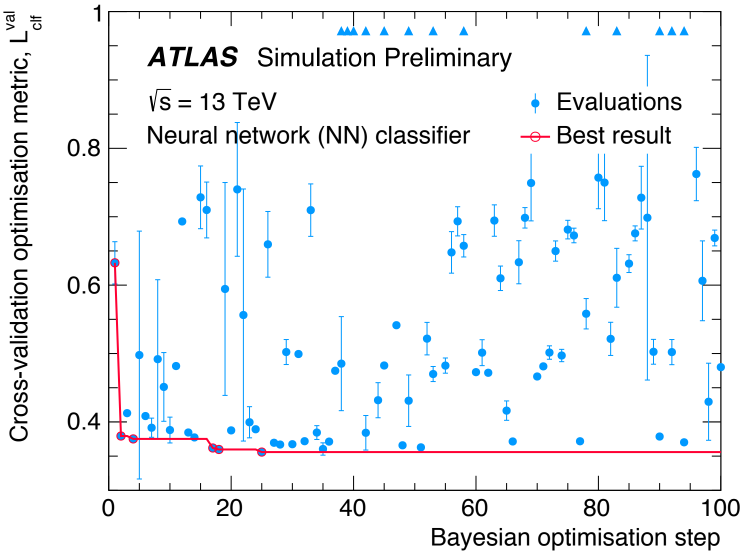


Figure 6.4: Neural network classifier Bayesian hyperparameter optimisation. Blue markers indicate the mean and standard deviation for each evaluation. The red line indicates the running optimisation metric minimum. Open red markers indicate improvements. Triangular markers indicate evaluation metrics outside of the axis range.

6.0.25 Training Procedure

Pre-training

The training of both classifier and adversary was performed on a cluster of NVIDIA Tesla K80 GPUs. To give adequate initial conditions to the adversarial joint optimisation, the classifier and adversary are first pre-trained stand-alone with the chosen hyperparameter configuration using 3-fold cross-validation.

First, the classifier is pre-trained for 200 epochs. The evolution of the classifier loss L_{clf} as a function of the number of training epochs is shown in Figure 6.5.

L_{clf} decreases monotonically during training, for both training and cross-validation splits, indicating no over-training given the chosen hyperparameter configuration. In addition, the final classifier loss is comparable to Ref. [34], suggesting that the classifier network architecture is performant.

The adversary is then pre-trained for 10 epochs at a fixed θ_{clf} configuration found previously, resulting in a minimal perturbation around the optimal neural network classifier. The adversary is allowed in this way to condition its posterior on the pre-trained classifier. During the

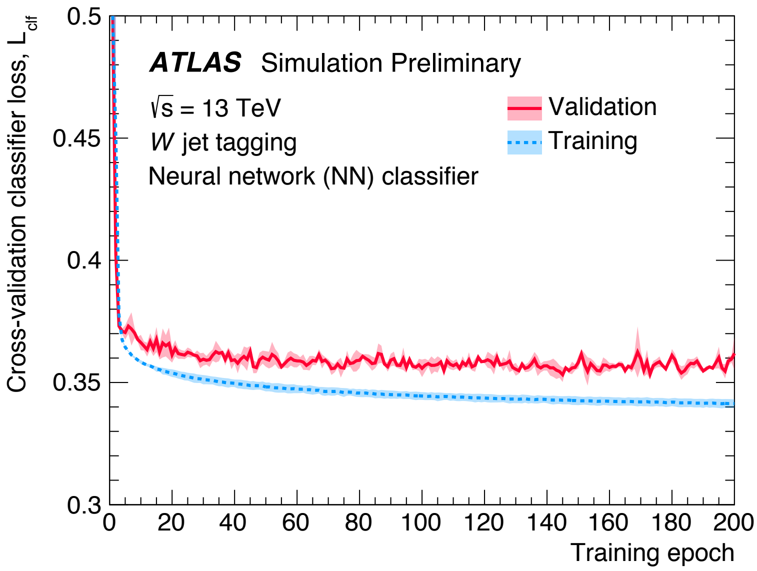


Figure 6.5: Stand-alone neural network classifier loss during 3-fold cross validation training as a function of the number of training epochs, for training and validation splits. Lines indicate mean loss across folds. Shaded bands indicate standard deviation of losses across folds.

adversary pre-training, the adversary loss is expected to converge to a minimum which is the conditional entropy H of the random variables m and (z, a):

$$L_{\mathrm{adv}}(\theta_{\mathrm{clf}}, \theta_{\mathrm{adv}})\Big|_{\theta_{\mathrm{clf}}} \xrightarrow{\theta_{\mathrm{adv}}} H(m|z, a)$$

Joint Optimisation

The training dynamics of the classifier, adversary, and effective losses is shown in Figure 6.6. As it can be seen, during the adversary pre-training, the classifier is kept fixed, and thus the classifier loss L_{clf} remains constant, while the L_{adv} decreases monotically from 0 (corresponding to random guessing of the mass) up to a minimum. After the adversary pre-training, the two networks are trained simultaneously. The L_{clf} gets worse as a way to make the adversary be unable to infer the jet mass and so rising at the same time L_{adv}. In other words, the classifier makes the variable z to be more pivotal with m at the cost of loosing classification power. The net effect results in a minimisation of the effective loss $L_{\mathrm{clf}} - \lambda L_{\mathrm{adv}}$ as seen in the last panel.

Ideally, for every parameter update of the classifier, the adversary should be allowed to fully converge. In practice, the optimisation is typically done using alternating [62] or simultaneous [81] gradient updates of the classifier and adversary weights. Since the latter has been found to lead to better convergence, this is chosen as the training method for the joint neural network.

In the limit of full mass-decorrelation, the variable z will be pivotal with m and so the adversary posterior will be equal to the prior, i.e the mass distribution $p(m)$. In other words, updates in θ_{clf} will produce the conditional entropy $H(m|z,a)$ to reach its maximum $H(m)$.

The optimal classifier $f(x, \theta_{\text{clf}})$, would correspond to fulfill both classifier and adversary ideal situations, where f is either an optimal classifier, i.e $L_{\text{clf}} = H(y|x)$ and pivotal, i.e $L_{\text{adv}} = H(m)$. In theory, we have that:

$$H(y|x) - H(m) \leq L_{\text{clf}} - H(m|z,a)$$

and the equality holds for an optimal θ_{clf}. In practice, the assumption of existence of an optimal and pivotal classifier may not hold, resulting in a strict lower bound. In particular, the deviation between the final L_{adv} and $H(\text{prior})$ could be due to either a trade-off between L_{adv} and L_{clf} that is beneficial to the overall loss and/or a lack of capacity in the parametrisation of adversary posterior p.d.f. as a Gaussian Mixture Model. Nevertheless, a situation very close to the ideal case is attained.

6.0.26 Results

The various mass-decorrelated jet taggers, treated in Sec. 6.0.24 and Ap. D) result in a single discriminant variable which classifies jets as either W jets or non-resonant multijets while keeping the shape of the background jet mass distribution unaltered at a certain degree.

As a representative example, the distributions of the NN and ANN tagger discriminants are shown in Figure 6.7. It can be seen that the NN tagger powerfully separates the two classes of jets, while for the ANN it happens in a less level as a result of the mass-decorrelation.

The performance of each tagger is evaluated for both classification and mass-decorrelation. The applications for which mass-decorrelated taggers are immediately useful are typically characterised by p_{T}given the jet and photon trigger thresholds of around 500 and 200 GeV, respectively. Therefore, results are studied in two different p_{T}bins: $[200, 500]$ GeV and $[500, 1000]$ GeV.

Classification

The classification performance, measured in terms of receiver operating characteristic (ROC) curves, is shown in Figure 6.8 for each tagger. This clearly shows a deterioration in the discrimination power of the ANN with respect to the NN tagger.

For 200 GeV $< p_{\text{T}} <$ 500 GeV shown in Figure 6.8a, among all taggers, the best discrimination power is achieved by the standard MVA taggers, NN and AdaBoost, as measured using

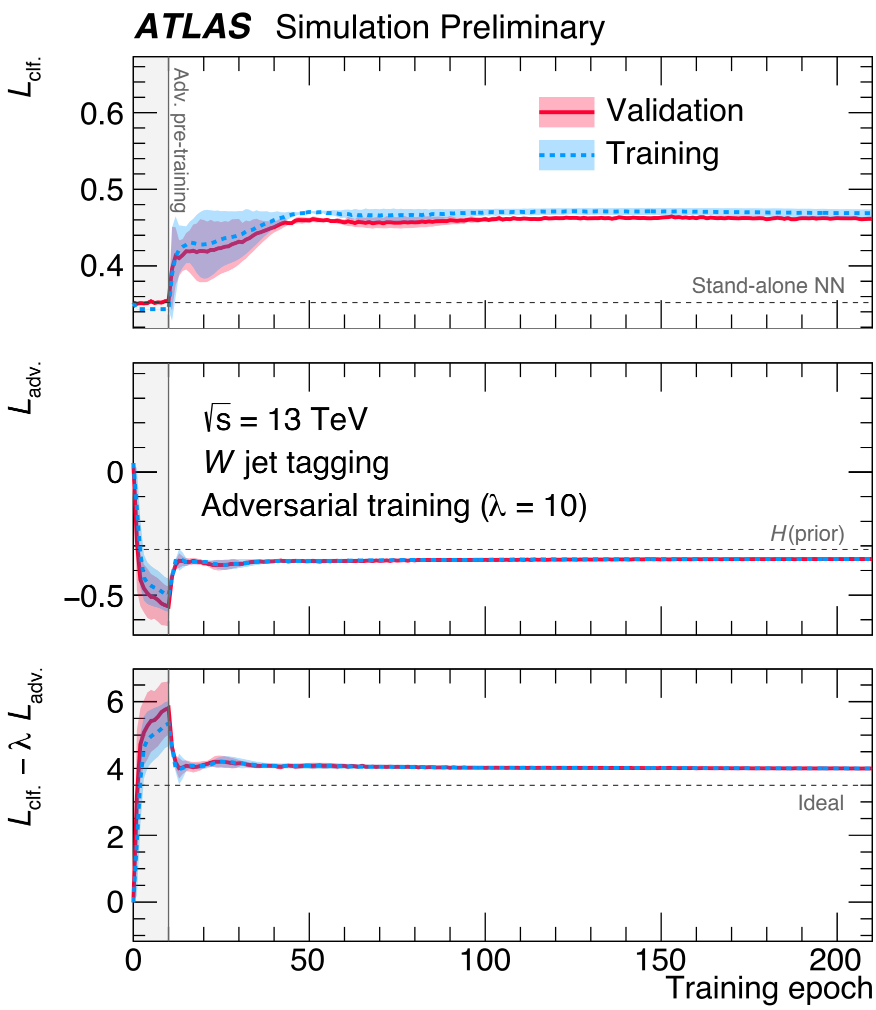
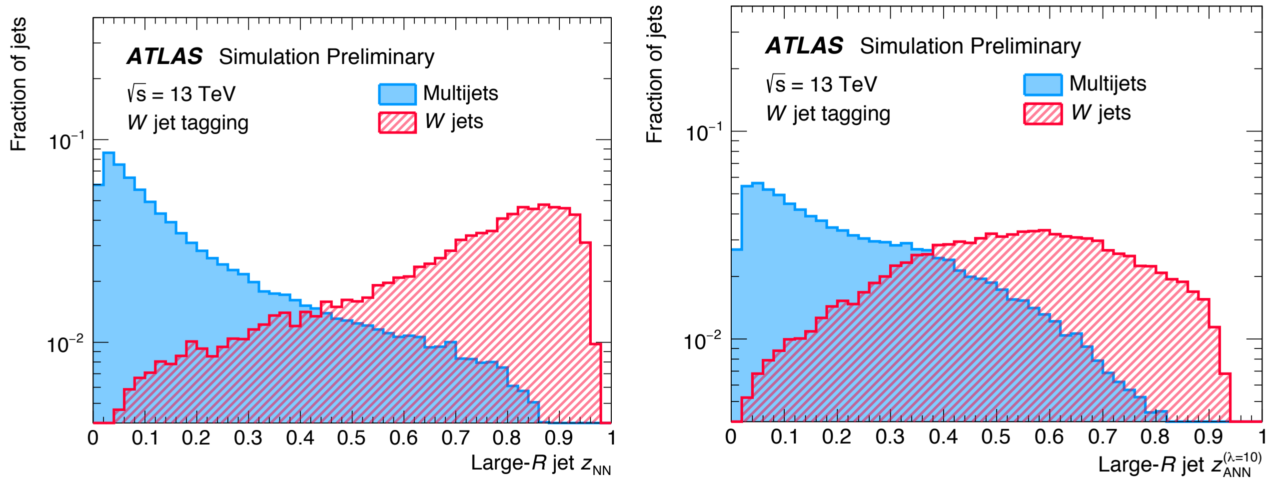


Figure 6.6: Classification (top), adversary (middle), and effective classifier losses (bottom) associated with the adversarial neural network tagger during 3-fold cross validation training as a function of the number of epochs, for training and validation splits. The first 10 epochs are spent on pre-training the adversary network (grey band). Lines indicate mean loss across folds, shaded bands indicate standard deviation of losses across folds. References for the optimal (stand-alone) NN classifier loss, the entropy H of the adversary prior, and the ideal, effective loss are shown by the grey dashed line.

the background rejection at fixed signal efficiency at $\varepsilon^{\text{rel}}_{\text{sig}} = 50\%$. The two standard MVA taggers are seen to yield background rejections which are up to four times those of the standard, single-variable taggers. In this kinematic region, for a particular choice of decorrelation parameters ($\lambda = 10, \alpha = 0.3$), all mass-decorrelated taggers are seen to perform almost identically in terms of classification.

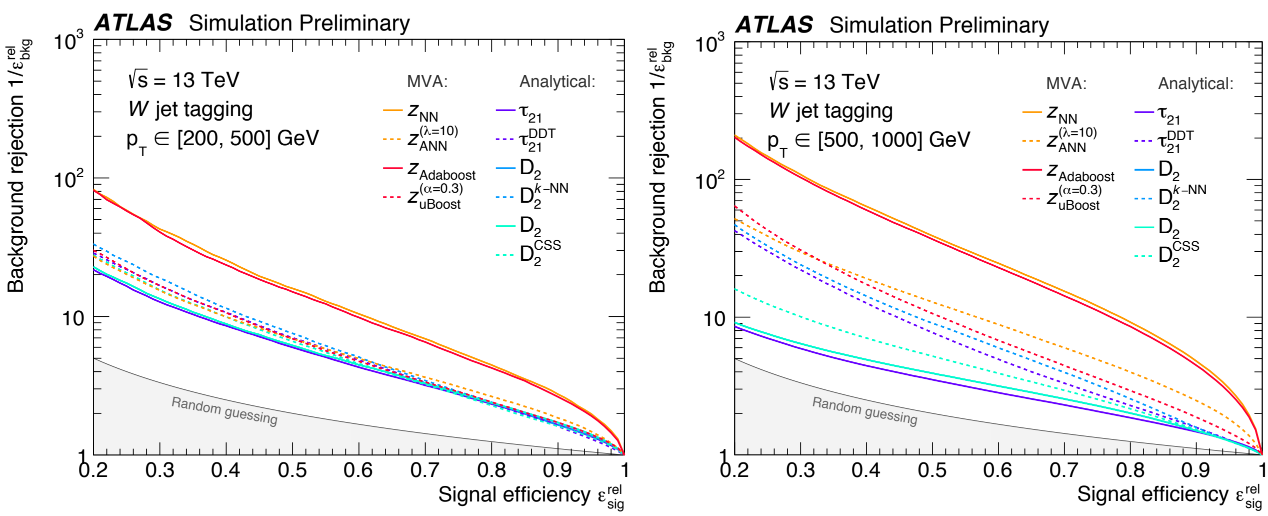
Among the analytical mass-decorrelation methods, k-NN is seen to provide the best classification; better than CSS, with both based on the D_2 substructure variable.

For all single-variable taggers, the observed background rejection after mass-decorrelation is greater than for the original taggers, an effect which is particularly appreciated in the high-



(a) Neural network classifier (b) Adversarial neural network classifier

Figure 6.7: Distributions of the jet tagger variables for the standard neural network classifier and the adversarially trained neural network classifier, for multijets and W jets.



(a) Jet $p_T \in [200, 500]$ GeV (b) Jet $p_T \in [500, 1000]$ GeV

Figure 6.8: Rejection of multijets (background) as a function of W jet (signal) selection efficiency, for standard and mass-decorrelated version of analytical and multivariate (MVA) jet taggers in two regions of jet p_T, without the addition of a jet mass-window cut.

$p_T$bin. The standard, single-variable taggers are seen to lose classification power with p_T, which is an expression of relative shifts of the substructure variable distributions for multijets and W jets as a function of p_T. On the other hand, since the DDT and k-NNregression methods both decorrelate the substructure variable not only from mass, but also from $p_T$through ρ and ρ^{DDT}, the fact that these methods recover lost performance exactly by removing p_T-dependence is not surprising. Indeed, the k-NNmethod removes the mass and p_{T}-dependence by construction exactly at $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$.

The same effect is also seen in the low-p_{T}bin, although to a much smaller extent. In the case of DDT, an instructive parallel is provided by the observation that the linear transform acts similarly to a Fisher discriminant, extracting additional information from ρ^{DDT} which slightly improves classification.

In contrast to the mass-decorrelated single-variable taggers, the mass-decorrelated MVA taggers exhibit a considerable decrease in classification performance relative to their standard variants. This effect is expected and acceptable as part of a trade-off between the two. In addition, the greater initial degree of correlation with the jet mass observed for standard MVA taggers compared to standard single-variable taggers means that there is more potential for degradation of classification power.

Decorrelation

The most direct measure to see the mass-decorrelation effect produced for each tagger, is the inspection of the normalised multijet mass distribution before and after the cut on the discriminant variable. Such comparisons, together with the signal mass distribution before tagging, are shown in Figure 6.9.

Cutting on each of the standard taggers sculpts the background jet mass distribution, thereby introducing artificial structures not present in the original spectrum. Such sculpting may directly impact the sensitivity of physics searches by reducing the ability to constrain systematic uncertainties on e.g. multijet backgrounds. In particular, the standard MVA taggers sculpt the background jet mass distribution to resemble the W jet mass peak.

Each of the mass-decorrelation procedures serves to mitigate such sculpting. For the mass-decorrelated single-variable taggers, CSS and in particular k-NNregression are both seen to remove most sculpting effects. The DDT transform removes the sculpting in the lower jet mass region, while some disagreement persists at higher masses as a result of the limitations of the assumption of linearity underlying the method.

The mass-decorrelation of both MVA taggers is considerably improved following the application of their respective training methods for decorrelation. In particular, the ANN tagger yields a largely smooth, un-sculpted distribution of multijet events passing the cut. On the other hand, the mass distribution after the cut on $z_{\mathrm{uBoost}}^{(\alpha=0.3)}$ retains some residual sculpting, particularly around $m \approx 100$ GeV.

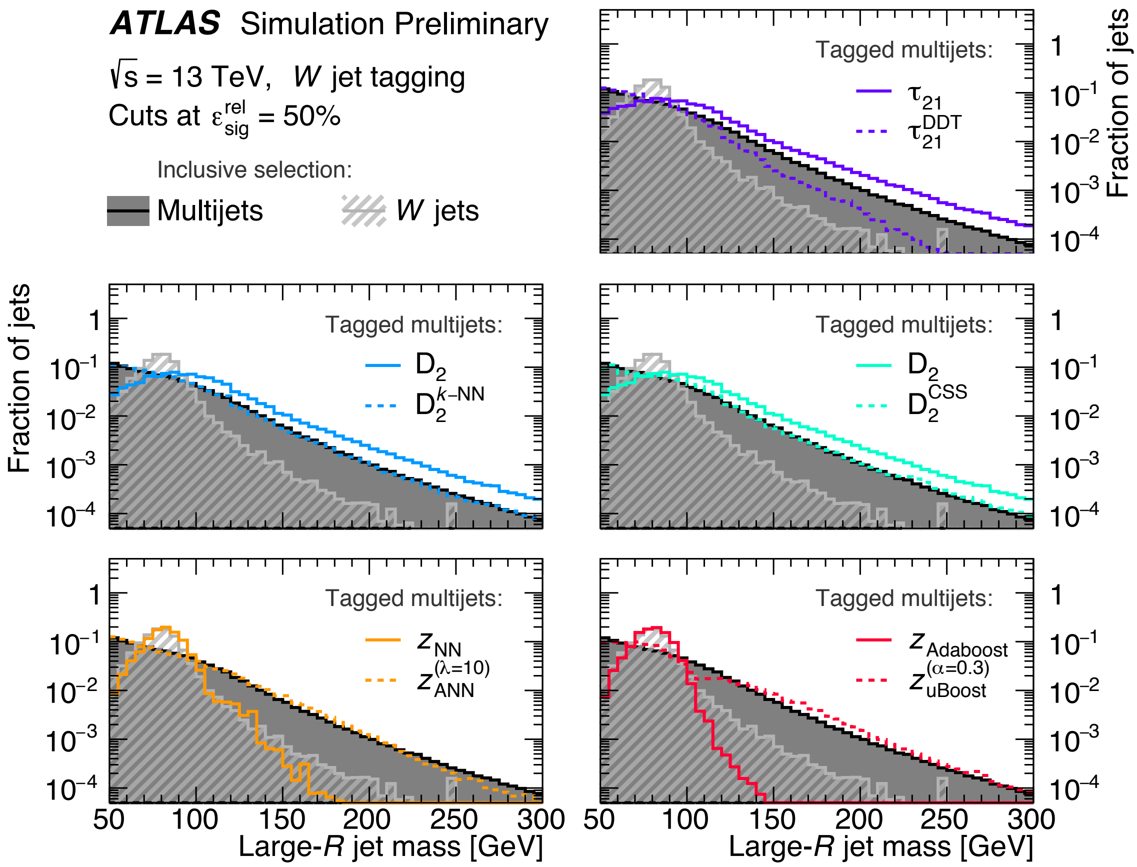


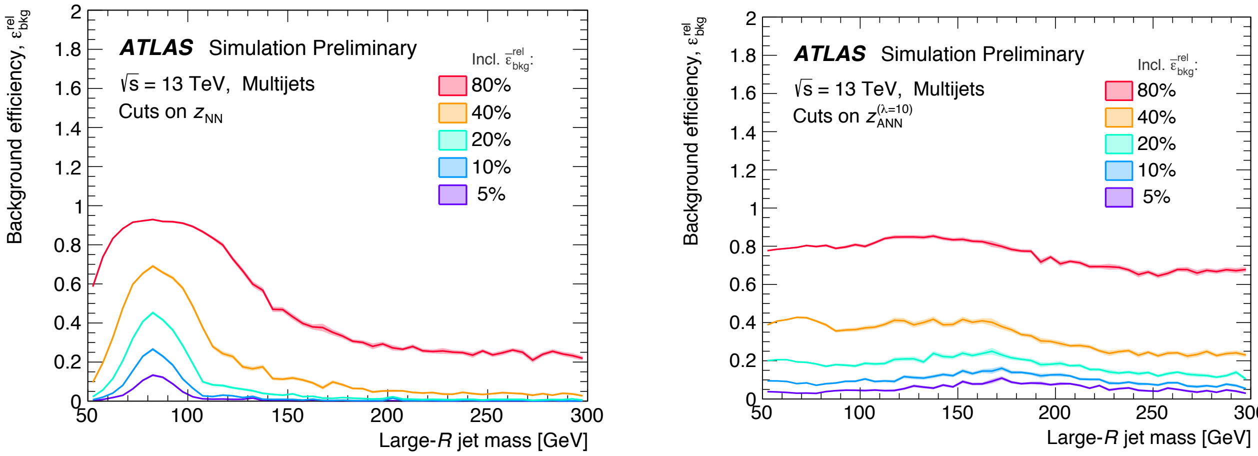
Figure 6.9: Normalised jet mass distribution for inclusive multijets before cuts, compared to the same distributions after cuts on the studied jet discriminants. Also shown for reference is the jet mass distribution for W jets before tagging. Cuts are chosen to correspond to a W jet (signal) selection efficiency of $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$.

Similar residual sculpting is not observed for the ANN tagger, which is evident from the local background efficiency as a function of the jet mass, for a range of inclusive background efficiencies, as shown in Figure 6.10 for the neural network taggers.

Figure 6.10a shows the sculpting of the multijet background around the W jet mass peak, and the mass-decorrelation effect is evident in Figure 6.10b, where the background efficiency profiles are roughly uniform as a function of the jet mass. The deviation from uniformity in Figure 6.10b, most evident between jet masses of 150 and 200 GeV, reflects the residual correlation with the jet mass, which may be reduced by more fine-tuned optimisation and longer adversarial training.

Another advantage of ANN is that, in contrast with uBoost, it provides uniform selection efficiencies for all inclusive selections, in particular at $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$.

As the quantitative summary metric for the mass-decorrelation, Figure 6.11 shows the JSDas a function of the background efficiency at which the cut is performed. As expected,



(a) Neural network classifier

(b) Adversarial neural network classifier

Figure 6.10: Jet mass-dependent multijet selection efficiencies for various inclusive efficiencies for the standard neural network tagger and the adversarially trained neural network tagger.

the JSD values for the mass-decorrelated taggers are consistently and considerably lower — i.e. have lower degrees of correlation with the jet mass — than the standard ones.

k-NN regression is seen to lead to the greatest degree of mass-decorrelation, especially in the vicinity of the background efficiency percentile at which the regression is performed (16%). The other methods exhibit more uniform mass-decorrelation across $\varepsilon^{\text{rel}}_{\text{bkg}}$. Among the MVA taggers, the ANN performs considerably better than uBoostfor the chosen values of λ and α, for the chosen metric.

The fact that the JSD is computed from histograms with finite statistics imposes a lower bound for the mass-decorrelation given the chosen testing dataset. This statistical limit on JSD can be estimated using bootstrap sampling [59], treated in Section 5.0.17, which is shown as a function of the background selection efficiency in Figure 6.11 as a smoothed, dashed line and shaded band, indicating the mean and standard deviation, respectively, of the bootstrap sampling. Additionally, for a given physics task, full mass-decorrelation might not be necessary or optimal. Depending e.g. on the level of systematic uncertainty, some intermediate level of correlation with the jet mass might provide a balance with classification power which is better suited for the physics task at hand than full mass-decorrelation.

Combined Performance

A combined metric, reflecting both classification performance and mass-decorrelation, is necessary to assess the trade-offs balanced by each of the mass-decorrelation procedures. A more complete picture of the two metrics of performance is found by plotting the two metrics to-

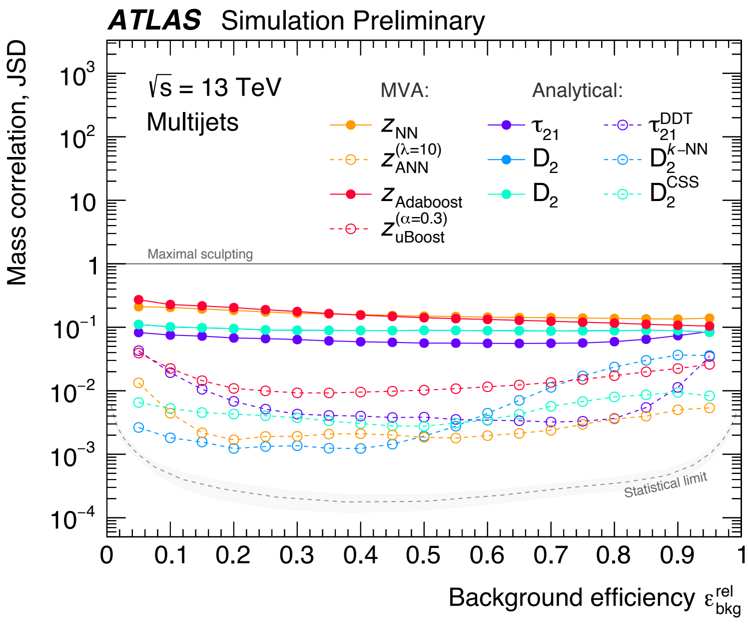


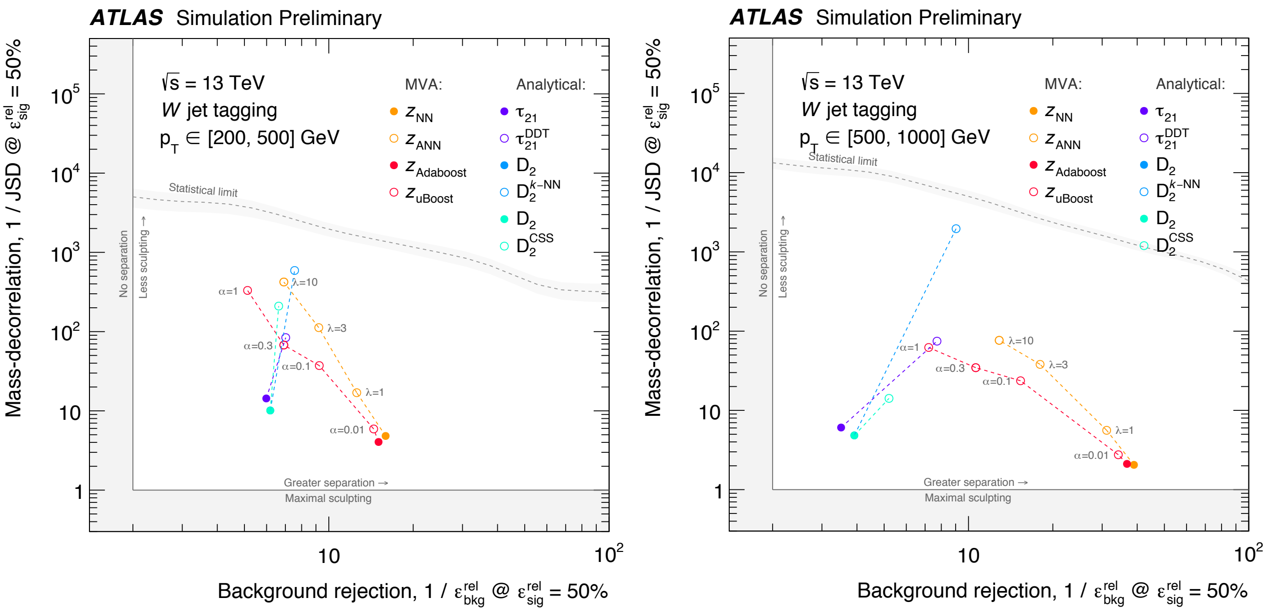
Figure 6.11: Profiles of the Jensen-Shannon divergence (JSD) for cuts corresponding to various multijet (background) selection efficiencies. Standard classifiers are indicated with filled markers. Mass-decorrelated classifiers indicated with open markers. The shaded grey band indicates the statistical limit on JSD from the finite number of simulated jets.

gether. Figure 6.12 shows the mass-decorrelation (1/JSD) versus the background rejection ($1/\varepsilon^{\rm rel}_{\rm bkg}$) for tagger cuts at $\varepsilon^{\rm rel}_{\rm sig} = 50\%$, in two $p_{\rm T}$bins. The x-axis measures classification power and the y-axis measures mass-decorrelation, with larger values along each indicating better performance. Depending on the given task, a specific direction in the plane of Figure 6.12 (with respect to the origin) will correspond to the best trade-off.

For each of the mass-decorrelated MVA taggers, several working points are evaluated, by scanning λ for the ANN tagger and α for uBoost. For high values of λ ($\gtrsim 10$), the ANN method starts to saturate given the chosen network configurations, training procedures, and datasets.

The dashed line and shaded band at high 1/JSD indicate the statistical limit of the mass-decorrelation, estimated using bootstrap sampling.

Figure 6.12 shows that for equal levels of mass-decorrelation, the ANN tagger generally provides the greatest background rejection. The BDT-based MVA taggers have comparable performance to the NN-based taggers for the standard variants, but the adversarial training mass-decorrelation method is seen to perform better than the uBoostmethod for the chosen configurations. From Figure 6.12b, the effect of the analytical, single-variable mass-decorrelation methods on improving the classification power while similarly decorrelation from the jet mass, as discussed above, is particularly evident.



(a) Jet $p_\mathrm{T} \in [200, 500]$ GeV

(b) Jet $p_\mathrm{T} \in [500, 1000]$ GeV

Figure 6.12: Unified plot of the metrics for classification (background rejection, $1/\varepsilon^{\mathrm{rel}}_{\mathrm{bkg}}$) and mass-decorrelation (inverse Jensen-Shannon divergence? @ Z 1/JSD), for cuts corresponding to $\varepsilon^{\mathrm{rel}}_{\mathrm{sig}} = 50\%$, in two $p_\mathrm{T}$bins. Greater values along each axis indicate better performance. Standard classifiers are indicated with filled markers. Mass-decorrelated classifiers are indicated with open markers, with parameter scans traced out by dashed lines. The shaded grey band indicates the statistical limit on 1/JSD from the finite number of simulated jets.

The k-NN regression method is seen to be the most effective of the analytical decorrelation methods, leading to a tagger variable which is close to fully decorrelated from the jet mass within statistical uncertainties. The saturation of the ANN tagger at high λ means that an upper limit on 1/JSD exists for $\lambda \gtrsim 10$ with the chosen configuration. This is a reasonable observation, considering the complexity of the ANN training procedure and the fact that the tagger is evaluated on a dataset 10 times larger than the training dataset. This discrepancy in data sample sizes means that complex decorrelation methods like the ANN are likely decorrelated at, or close to, the statistical limit of the training dataset, but not necessarily at the statistical limit for the much larger evaluation dataset. The simplicity of e.g. the k-NN method means that it will be more robust to such differences in statistics. For these reasons, raising the upper limit on 1/JSD for ANN will likely be possible by increasing training statistics, performing a more fine-tuned model architecture optimisation, and similar.

Chapter 7

Conclusions

In the presented thesis the measurement of the dijet cross-section in proton-proton collisions at 13 TeV was performed. The measurement used data collected at the LHC with the ATLAS detector during 2015. A quantitative comparison of the measurements to fixed-order NLO QCD calculations, corrected for non-perturbative and electroweak effects. Good agreement is observed (with p-values in the percent range) in all the kinematic regions studied, validating the SM description of the QCD dijet production at an unprecedented energy scale of 13 TeV. Several techniques used through the measurement were described, together with a careful description of the jet energy Monte Carlo based calibration, one of the main tasks performed during the PhD.

Finally, a novel machine learning technique based on adversarial neural-networks was studied for the identification of jets coming from a top quark or W boson against the large background QCD. Standard classifiers based on jet substructure variables exhibit non-linear correlations with the jet mass, since the classifiers learn that the mass is a powerful feature for discriminating against the non-resonant background. Therefore, fixed-value cuts on such classifier observables tend to distort the shape of the jet mass distribution for non-resonant jet production, resulting in a reduction of the statistical significance for some physics analyses. Results based in Monte Carlo simulation provide an encouraging first look at the prospect for using adversarial neural-networks for mass-decorrelated substructure tagger in physics searches.

Appendix A

Jet Response

The jet response grows with the energy mainly due to the increase of nuclear interactions as the energy increases. In order to view this, let us suppose a jet composed by 3 pions (π^-, π^0, π^+). Before interacting with any nuclei (0th generation) the electromagnetic (em) and hadronic (had) components are 1/3 and 2/3 respectively, the response will be then $\sim 1/3$ if one supposed there are not energy losses in em interactions. As far as π^- and π^+ particles find a nuclei, each of them will peform a nuclear interaction and generate new pions (π^-, π^0, π^+) equally probable (1/3 each) to produce (on average) and so on (1st ... Nth generation).[1] As the number of generatios grows, the production of π^0 increases, resulting in a higher em fraction ($\mathrm{f_{em}}$) of the jet as can be seen in Figure [?]. It can be clearly realized from the diagram that $\mathrm{f_{em}}$ follows a potential law: $\mathrm{f_{em}} = \frac{1}{3}\sum_{k=0}^{n}(1-\frac{1}{3})^k = 1-(1-\frac{1}{3})^{n+1}$. Here we are making an important assumption on the fact that every nuclear interacion produces 3 pions, which in reality a lot of more particules (as kaons) are produced. As a result, $\mathrm{f_{em}}$ will grow with the number of generation (n) actually slower than a power law.

We can go deeper with this toy model and relate $\mathrm{f_{em}}$ with the initial energy of the jet (E). For this, we will assume that the number of particles produced per interaction (multiplicity) does not depend on the energy of the parent particle, and that the energy is proportional to the number of π^0 produced: $\frac{E}{E_0} = \#\pi^0$, where E_0 is the energy necessary to produce a π^0, which is around 1.3 GeV [ref].

It can be deduced from the diagram that $n \sim \log_2(\#\pi^0)$. As a result, we get that $n \sim \log(\frac{E}{E_0})$. Finally, we can conclude (at zero order approximation) the following relation:

$$\mathrm{f_{em}} \sim 1 - \left(1 - \frac{1}{3}\right)^{\log\left(\frac{E}{E_0}\right)} \tag{A.0.1}$$

[1] Each pion has the same isospin value ($I = 1$) and so strong interactions do not distinguish between them.

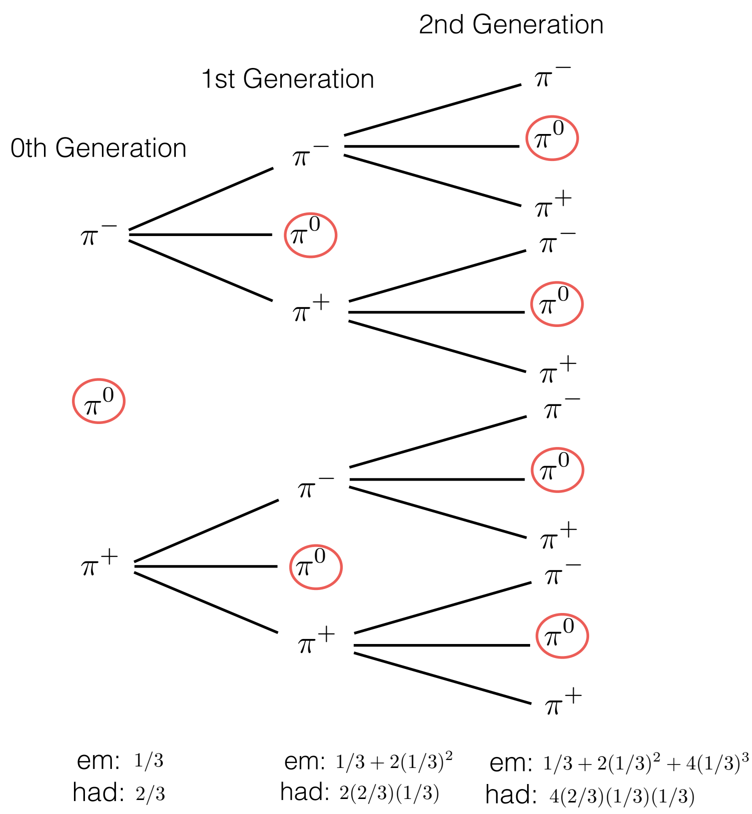


Figure A.1: Diagram of pions generations.

This relation follows a similar pattern as the jet energy response from Figure [?]. Although it is worth to note some differences with the expression we derived. It can be seen that when $E \to \infty$, $\mathrm{f}_{\mathrm{em}} \to 1$, which is not consistent with Figure [?] and cleary does not happen in reality. There are a lot effects we are not taking into account which will result in a decrease of the em response. For instance, we are neglecting the energy lose in breaking the nucleous, which is fundamentally impossible to avoid. A more detailed description of hadronic showers can be found here [ref].

Appendix B

Smoothing statistical fluctuations in systematic uncertainties

To avoid including statistical fluctuations in the reported systematics, a smoothing technique is applied. The process begins with the relative uncertainty for each systematic component which statistical fluctuations calculated using the bootstrapping method 5.0.17.

First, bins are combined iteratively until their weighted significance is $> 2\sigma$ (in some cases $> 1\sigma$ were considered), starting from both the right and left sides of the spectra. The combination with the most bins remaining is chosen.

Second, a rebinned histogram is built in this way from the original one were each bin is now statistical significant and its content (y_j) for each bin j correspond to the weighted average:

$$y_j = \frac{\sum_{i=j}^{j+\Delta j} \frac{x_i}{\sigma_i^2}}{\sum_{i=j}^{j+\Delta j} \frac{1}{\sigma_i^2}}$$

were i labels the original binning and moves from j until $j + \Delta j$ whose width is determined until the bin becomes significant. x_i is the relative uncertainty and σ_i the statistical uncertainty of x_i.

Finally, a smoothed histogram (μ_i) with the original binning is built by applying Gaussian kernel sliding window. Particularly, for each bin i, the following weighted average is computed:

$$\mu_i = \frac{\sum_j y_j K_{ij}}{\sum_j K_{ij}}$$

where j and i labels the original binning and

$$K_{ij} = \exp\Big(-\frac{(x_i - x_j)^2}{\Delta_i^2/4}\Big)$$

where $\Delta_i = a + bx_i$, for $a = 10$ and $b = 0.3$. Given that now j labels the original binning, the following equality was also considered: $y_j = y_{j+1} = \cdots = y_{j+\Delta j}$

The outcome effect is a reduction of statistical fluctuations in the systematic uncertainty components, with minimal added bias. This also has the effect of reducing artificial asymmetries, which are shown to cause problems when performing statistical tests for agreement between data and theory (C).

Results on this technique can be seen in Figure B.1. The plot on the left shows a relative uncertainty that has high statistical fluctuations, as a result, it can be seen that the smoothed result differs in a positive way from the original fluctuated uncertainty. On the other hand, on the right plot all bins are already significant, and so the smoothing does not make any substantial improvement.

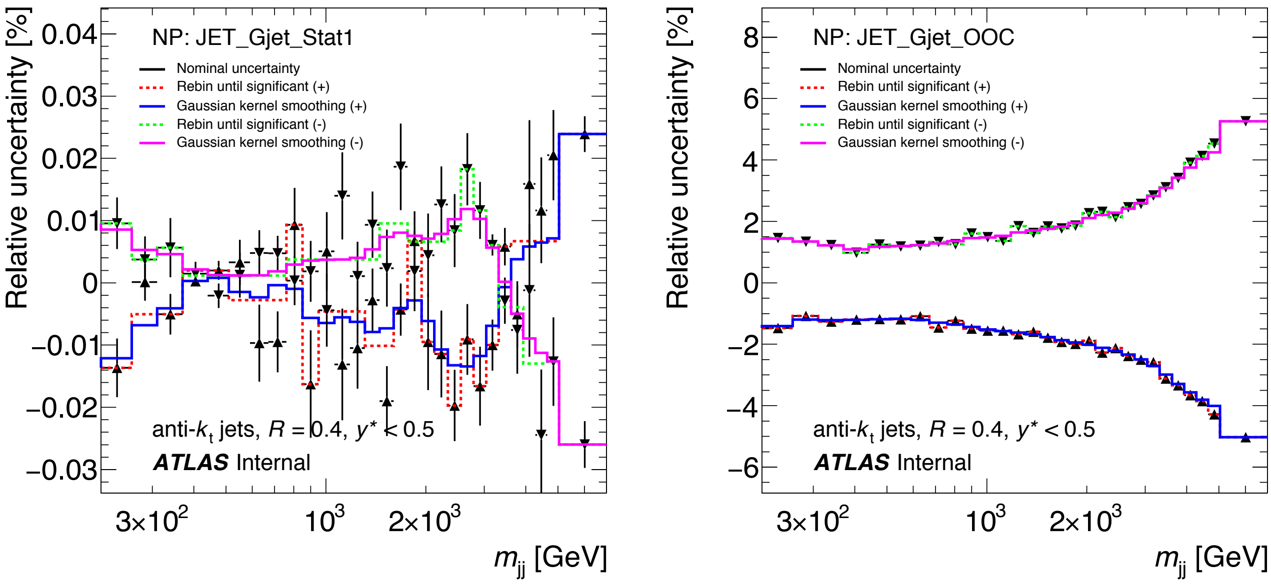


Figure B.1: Relative systematic uncertainty for up (blue) and down (pink) variations. On the left (right) uncertainty components with high (low) statistical fluctuations are shown. Markers corresponds to the nominal uncertainty, dash line the rebinned histogram for the intermediate smoothing step and solid line the resulted smoothed uncertainty.

Appendix C

χ^2 Calculation

A key method to test the agreement of SM predictions to the measured cross-sections is via a χ^2 test.

The test statistic is often constructed from a sum of squared errors representing the degree of deviation of one spectrum from another and is defined as:

$$\chi^2(d;t) = \sum_i \left(\frac{d_i - t_i}{\sigma_i}\right)^2$$

where d_i, t_i and σ_i^2 are the data, theoretical cross-sections and variance on bin i. Under the assumption that the SM describes well the data (null hypothesis is true) and that the data is normally distributed, which is valid in many cases due to the central limit theorem, this test statistic definition follows a χ^2 distribution.

This expression accounts only for the theoretical and experimental uncertainties (including both statistical and systematics) on individual bins (σ_i), which ignores the statistical and, even stronger, systematic correlations between bins. In addition, corrections for resolution effects were applied to the data (unfolding) and so enhancing the correlations effects. As a result, the standard definition results in a less powerful test. To account for this, the covariance matrix (C) is introduced in the χ^2 definition,

$$\chi^2(d;t) = \sum_{i,j} (d_i - t_i) \cdot [C^{-1}(t)]_{i,j} \cdot (d_j - t_i)$$

which also follows a χ^2 distribution. However, the covariance matrix is built from symmetrized uncertainties and thus cannot account for the asymmetries between positive and negative uncertainty components.

An alternative, more general, χ^2 definition, was proposed motivated from refs. [66] and [67]. It accounts for asymmetric uncertainties by separating them from the symmetric ones in the test statistic, which is now:

$$\begin{aligned}\chi^2(d;t) = \min_{\beta_a} \Bigg\{ \sum_{i,j} \Big[d_i - \Big(1 + \sum_a \beta_a \cdot (\epsilon_a^{\pm}(\beta_a))_i\Big) t_i \Big] \cdot [C_{\rm su}^{-1}(t)]_{i,j} \\ \cdot \Big[d_j - \Big(1 + \sum_a \beta_a \cdot (\epsilon_a^{\pm}(\beta_a))_j\Big) t_j \Big] + \sum_a \beta_a^2 \Bigg\}\end{aligned} \tag{C.0.1}$$

where $C_{\rm su}$ is the covariance matrix built using only the *symmetric* uncertainties, and β_a are the profiled coefficients resulted from a fit on the *asymmetric* uncertainties that minimizes the χ^2. Here $\epsilon_a^{\pm}$ is the positive component of the a$^{\rm th}$ asymmetric relative uncertainty if the fitted value of β_a is positive, or the negative component otherwise. For this analysis, an uncertainty component was considered asymmetric if the absolute difference between the magnitudes of the positive and negative portions is larger than 1% of the cross-section in at least one bin. Due to the large asymmetric uncertainties, coming mainly from the theoretical predictions as can be seen in Figure xxx, this χ^2 definition is a better motivated choice. Although it is not probed to be the optimal one, it provides stronger statistical power than the two approaches mentioned above.

As explained in Sec. [?], 1000 replicas of unfolded data was obtained for the estimation of the statistical uncertainty. The statistical covariance matrix is then built from these toys following the standard definition. To built the symmetric systematic covariance, the uncertainties are assumed to be non correlated between each other and fully correlated across m_{jj} bins. As a result, the covariance for a given symmetric component k is then defined as $\text{cov}(x_i^k, x_j^k) = \sigma_i^k \sigma_j^k$, which comes from the fact that they are fully correlated across bins. The assumption that the systematics are independent between each-other implies that one can sum linearly their covariance matrices. The result is added to the statistical covariance to get the total uncertainty matrix: $C_{\rm su} = \sum_k \sigma_i^k \sigma_j^k + \text{stat. cov.}$

The χ^2 distribution of a given theory hypothesis is required in order to calculate the probability of measuring a specific χ^2 value under that theory hypothesis. Since Eq. C.0.1 does not follows a standard χ^2 distribution, an empirical approach based on pseudo-experiments (see Sec. [?]) was performed. They represent fluctuations of the theory hypothesis due to the full set of experimental and theoretical uncertainties. To do this, the statistical and symmetric systematic uncertainties were taken as Gaussian-distributed components, while a two-sided Gaussian distribution is used for the asymmetric ones. For each pseudo-experiment, the χ^2 value is computed between the pseudo-data (theory fluctuated spectra) and the theory hypothesis. In this way, the χ^2 distribution that would be expected for experiments drawn from the theory hypothesis is obtained without making assumptions about its shape.

The observed χ^2 (χ^2_{obs}) value is computed using the data and the theory hypothesis. To quantify the compatibility of the data with the theory, the ratio of the area of the χ^2 distribution with $\chi^2 > \chi^2_{\text{obs}}$ to the total area is used. This fractional area, called a p-value, is the observed probability, under the assumption of the theory hypothesis, to find a value of χ^2 with equal or lesser compatibility with the hypothesis relative to what is found with χ^2_{obs}. If the observed p-value is smaller than 5%, the theoretical prediction is considered to poorly describe the data at the 95% CL.

Appendix D

Decorrelated Jet Tagging Algorithms

A variety of approaches to constructing mass-decorrelated jet taggers, from simple linear transforms to complex MVA architectures, are detailed below.

D.0.27 Designed decorrelated taggers

The original method relies on the jet scaling variable $\rho = \log\left(m^2/p_\mathrm{T}^2\right)$ [56]. It is observed that profiling (averaging) the substructure variable τ_{21} as a function of ρ exposes a linear relationship, which can be exploited to perform a linear transform, decorrelating τ_{21} with respect to ρ.

In practice [47, 35] it is found that the jet scaling variable

$$\rho^\mathrm{DDT} = \log\left(\frac{m^2}{p_\mathrm{T} \times \mu}\right) \tag{D.0.1}$$

more robustly removes residual dependence on the jet p_T, and therefore leads to better decorrelation across the jet kinematic phase space. Here, the parameter μ balances the dimensions and is usually taken to be $\mu = 1\;\;\mathrm{GeV}$.

In Figure D.1, a linear relationship between $\langle\tau_{21}\rangle$ and ρ^DDT is observed, roughly in the range $\rho^\mathrm{DDT} \in [1.5, 4.0]$.

Towards higher values, the linearity breaks down due to effects arising from the fixed-radius jet clustering algorithm used, while it breaks down towards low values of ρ^DDT where soft QCD effects become important.

A linear fit is performed to the τ_{21} profile in the range $\rho^\mathrm{DDT} \in [1.5, 4.0]$. From this fit, the transform $\tau_{21} \mapsto \tau_{21}^\mathrm{DDT}$ is defined as

$$\tau_{21}^\mathrm{DDT} = \tau_{21} - a \times \left(\rho^\mathrm{DDT} - 1.5\right) \tag{D.0.2}$$

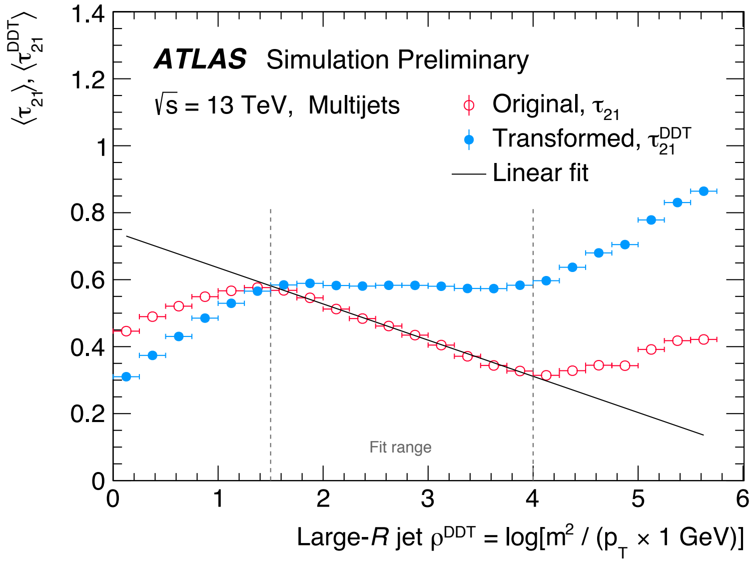


Figure D.1: Mean values of τ_{21} and τ_{21}^{DDT} as functions of ρ^{DDT} for the multijet ba ckground. A linear fit to the τ_{21}-profile, used in the definition of τ_{21}^{DDT}, is performed on the indicated range.

where a is the measured slope of the fit. Figure D.1 shows how the DDT transformation removes the linear correlation of τ_{21} with ρ^{DDT} (blue dots). Consequently, since ρ^{DDT} effectively encodes information about the kinematics of the jet (through m and p_{T}), the DDT transform yields a jet substructure discriminant which is decorrelated from the jet mass.

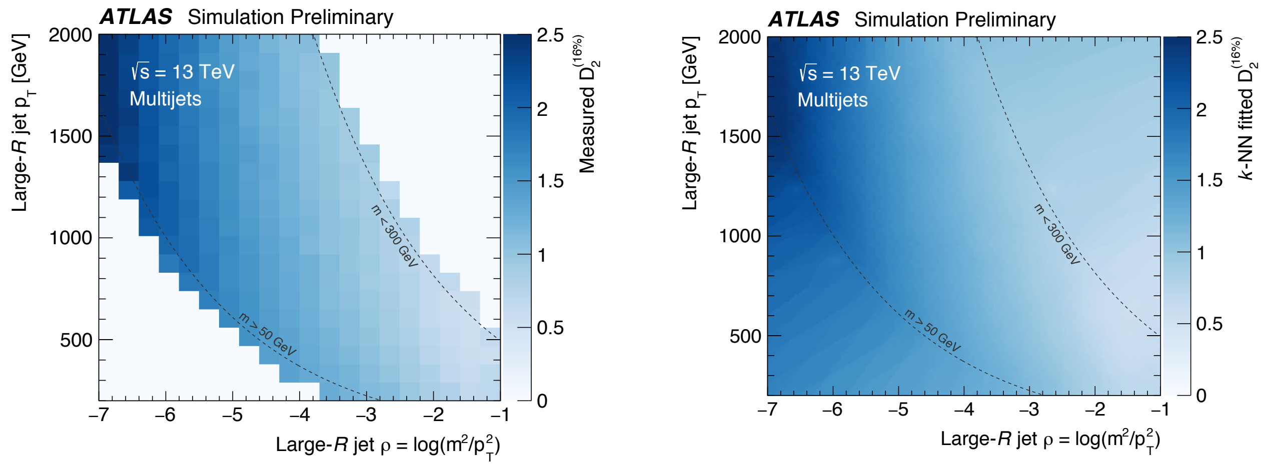
D.0.28 *k*-NN Regression

Whereas the DDT transform removes linear correlations between a substructure variable and kinematic variable(s), more general strategies able to face non-linear relationships exist. This method relies in a k-NN regression and can be summarized as follows:

First, a fixed percentile of inclusive background efficiency $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}}$ for the D_2 distribution is computed for multijets in bins of $\rho = \log(m^2/p_{\mathrm{T}}^2)$ and p_{T}, resulting in a two-dimensional profile. The efficiency selected in this case is $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}} = 16\%$, corresponding roughly to a signal efficiency of $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$. Second, in order to determine its functional dependence, a two-dimensional regression fit using the k-nearest neighbours (k-NN) algorithm [57] is performed, yielding the continuous distribution $D_2^{(16\%)}(\rho, p_{\mathrm{T}})$. Finally, for each jet a new observable $D_2 \mapsto D_2^{k\text{-NN}}$ is constructed by subtracting the predicted values from D_2:

$$D_2^{k\text{-NN}} = D_2 - D_2^{(16\%)} \tag{D.0.3}$$

The profiles of the fixed-efficiency D_2 cut as a function of ρ and p_{T} are shown in Figure D.2. The k-NN technique generalises the central concept behind DDT, making the method admissible to a more general class of substructure variables.



(a) Measured

(b) k-NN fitted

Figure D.2: Profiles of D_2 at $\varepsilon_{\text{bkg}}^{\text{rel}} = 16\%$ for multijets as measured in the training dataset (a) and as fitted using k-NN regression (b). Dashed lines indicate the phase space limits arising from the jet mass selection.

D.0.29 Convolved Substructure

Another analytical, more general, technique than the DDT described above is the convolved substructure (CSS) method [80]. It is capable to remove the dependence with the mass on the entire shape by convolving the distribution of the substructure variable of interest with a shape function:

$$\frac{1}{\sigma}\frac{d\sigma}{dx} \mapsto \frac{1}{\sigma}\frac{d\sigma}{dx_{\text{CSS}}} = \frac{1}{\sigma}\frac{d\sigma}{dx} \otimes F_{\text{CSS}}(x|\alpha, \Omega_D), \tag{D.0.4}$$

where $F_{\text{CSS}}(x|\alpha, \Omega_D)$ is a Gamma distribution.

$$F_{\text{CSS}}(x|\alpha, \Omega_D) = \left(\frac{\alpha}{\Omega_D}\right)^{\alpha} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{\alpha x}{\Omega_D}}. \tag{D.0.5}$$

Equation (D.0.4) is defined only at the level of distributions; in practice, the convolution indicated by $\otimes$ is estimated by transforming $x \mapsto x_{\text{CSS}} = G^{-1}(C(x)|\alpha, \Omega_D)$, where $C(x)$ is the cumulative distribution function of x and G is the cumulative distribution function corresponding to $\frac{1}{\sigma}\frac{d\sigma}{dx_{\text{CSS}}}$.

The parameters α and Ω_D are optimised inclusively and per bin of jet mass, respectively, using a two-step optimisation procedure as described in Ref. [80].

The transformed D_2^{CSS} distribution in the highest p_{T} bin is shown in Figure D.3. The convolution procedure from Equation (D.0.4) evolves the distribution of D_2 at one mass to the distribution at a reference mass. Therefore, using the lowest mass bin as the reference, the D_2^{CSS} in this bin is approximately unchanged. In contrast, the distribution of D_2^{CSS} in the high mass bin visibly differs from D_2 in the same mass bin. Not only has the mean value of D_2 shifted towards the reference mass, but the entire distribution is now coherently closer. Using the CSS technique, the D_2 distribution is transformed to be similar for all jet masses, thereby directly reducing correlation with the jet mass.

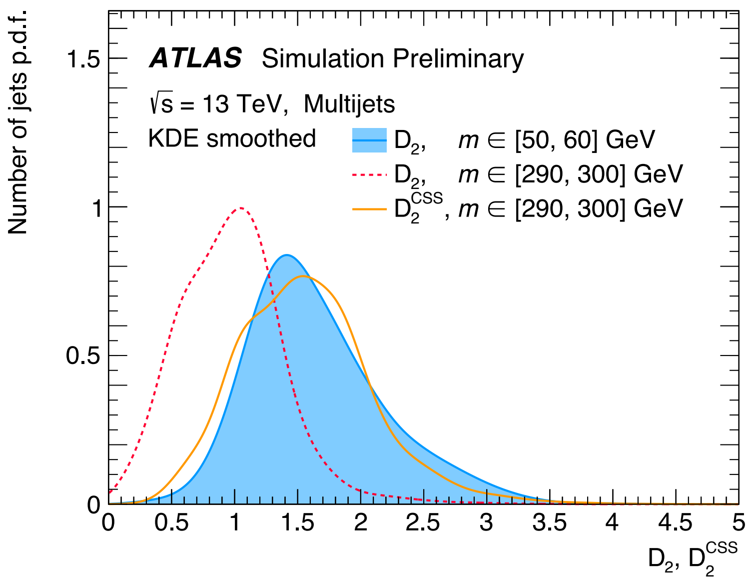


Figure D.3: Distributions of D_2 and D_2^{CSS} for multijets in the highest mass bin, as well as the reference distribution in the lowest mass bin. A kernel-density estimation (KDE) based smoothing B is applied to all training distributions.

D.0.30 Adaptive boosting for uniform efficiency

Boosted decision trees (BDTs) algorithms like AdaBoost (from "adaptive boosting") [65] are playing an increasingly prominent role in HEP physics applications. As an example, the observation of $H \to b\bar{b}$ has recently been achieved thanks to the use of BDTs [9]. They are also being used in object calibration [53], MC reweigthing [17], b-tagging [55], and software triggers. In contrast with the analytical taggers described above, based on a single feature, BDTs are multivariate (MVA) taggers that combine information of multiple features in order to achieve a single predictor. Other examples of MVA based techniques are Neural-Networks. BDTs posses an additional property called *boosting*, where an ensemble of *weak* classifiers are combined in order to obtain a single *strong* learner.

Despite its success, BDTs are also found to yield selection efficiencies which are non-uniform with respect to the jet mass, sculpting the mass background spectra after jet classification. The *adaptive boosting for uniform efficiency* (uBoost) method [96] is proposed to mitigate this non-uniformity behaviour.

Boosting works by assigning training events weights based on classification errors made by previous members of the series; events that are misclassified are given more weight. The uBoost technique augments this procedure by not just considering classification error, but also the uniformity of the selection. Events in regions of m where the selection efficiency is lower (higher) than the mean are given larger (smaller) weights. In this way, uBoost is able to drive the BDT towards a fixed uniform selection efficiency in the m observable at the cost of loosing discrimination power.

The BDT classifiers use the substructure variables listed in Table 4.1, the same used for the neural network classifiers, as input features. The hyperparameter configuration adopted for AdaBoostis the same as the one used for the BDT classifier in Ref. [34]. For the remaining uBoosthyperparameters, the default values in Ref. [86] are used. The degree of mass-decorrelation for uBoostis controlled by a hyperparameter α, called the uniforming rate. For $\alpha \to 0$, the adaptive boosting only takes classification loss into account, and the standard AdaBoostclassifier is retrieved. Conversely, for larger α, the uniform efficiency boosting becomes gradually more important.

The AdaBoostand uBoostclassification objectives during training are shown in Figure D.4.

The AdaBoostclassification loss is seen to decrease monotonically and reach a plateau for the testing dataset after 500 epochs of training. In contrast, the classification objective for uBoostinitially decreases due to improved discriminating power, and then rebounds as the adaptive boosting for uniform efficiency takes effect. The level of mass-decorrelation is given by the degree of divergence at the end of the training duration, which in turn is controlled by rate at which the uniformity boosting takes effect, controlled by α.

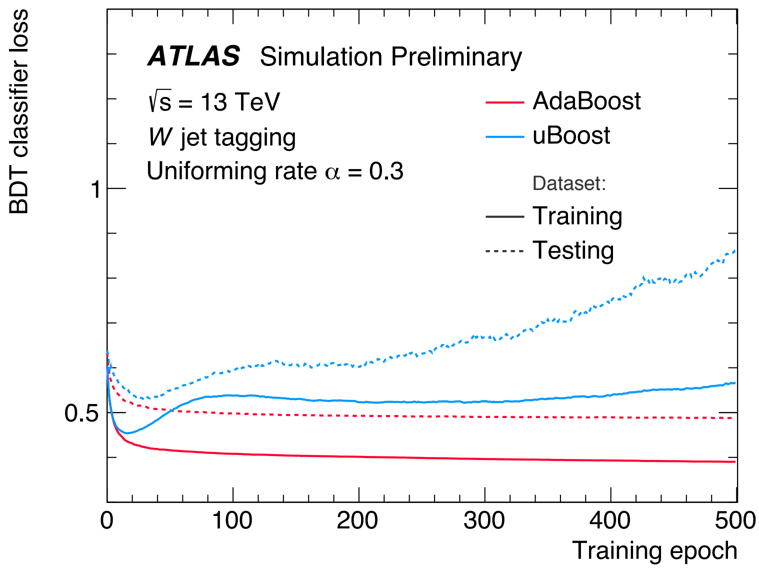


Figure D.4: Classification objective during training of AdaBoostand uBoostclassifiers for training and testing datasets for $\alpha = 0.3$.

Bibliography

[1] (https://home.cern/about/engineering/restarting-lhc-why-13-tev).

[2] https://stats.stackexchange.com/bayes-optimal-classifier-vs-likelihood-ratio.

[3] Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1. *Eur. Phys. J.*, C 77:490, 2017.

[4] M. Aaboud et al. Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev.*, D96(7):072002, 2017.

[5] Morad Aaboud et al. Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC. *Eur. Phys. J.*, C76(12):653, 2016.

[6] Morad Aaboud et al. Determination of the strong coupling constant α_s from transverse energy–energy correlations in multijet events at $\sqrt{s} = 8$ TeV using the ATLAS detector. *Eur. Phys. J.*, C77(12):872, 2017.

[7] Morad Aaboud et al. Jet reconstruction and performance using particle flow with the ATLAS Detector. *Eur. Phys. J.*, C77(7):466, 2017.

[8] Morad Aaboud et al. Performance of the ATLAS Trigger System in 2015. *Eur. Phys. J.*, C77(5):317, 2017.

[9] Morad Aaboud et al. Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector. 2018.

[10] Morad Aaboud et al. Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC. 2018.

[11] Morad Aaboud et al. Search for low-mass dijet resonances using trigger-level jets with the ATLAS detector in pp collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.*, 121(8):081801, 2018.

[12] Georges Aad et al. Luminosity Determination in pp Collisions at $\sqrt{s} = 7$ TeV Using the ATLAS Detector at the LHC. *Eur. Phys. J.*, C71:1630, 2011.

[13] Georges Aad et al. Performance of the ATLAS Trigger System in 2010. *Eur. Phys. J.*, C72:1849, 2012.

[14] Georges Aad et al. Jet energy measurement with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7$ TeV. *Eur. Phys. J.*, C73(3):2304, 2013.

[15] Georges Aad et al. Measurement of dijet cross sections in pp collisions at 7 TeV centre-of-mass energy using the ATLAS detector. *JHEP*, 05:059, 2014.

[16] Roel Aaij et al. Observation of $J/\psi p$ Resonances Consistent with Pentaquark States in $\Lambda_b^0 \to J/\psi K^- p$ Decays. *Phys. Rev. Lett.*, 115:072001, 2015.

[17] Roel Aaij et al. Measurement of angular and CP asymmetries in $D^0 \to \pi^+\pi^-\mu^+\mu^-$ and $D^0 \to K^+K^-\mu^+\mu^-$ decays. *Phys. Rev. Lett.*, 121(9):091801, 2018.

[18] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA, 2016. USENIX Association.

[19] F. Abe et al. Measurement of the dijet mass distribution in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV. *Phys. Rev.*, D48:998–1008, 1993.

[20] H. Abramowicz et al. Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data. *Eur. Phys. J.*, C75(12):580, 2015.

[21] C. Adloff et al. Measurement of dijet cross-sections in photoproduction at HERA. *Eur. Phys. J.*, C25:13–23, 2002.

[22] S. Agostinelli et al. GEANT4: A simulation toolkit. *Nucl. Instrum. Meth.*, A 506:250–303, 2003.

[23] S. Alekhin, J. Blümlein, S. Moch, and R. Placakyte. Parton distribution functions, α_s, and heavy-quark masses for LHC Run II. *Phys. Rev.*, D96(1):014011, 2017.

[24] Simone Alioli, Keith Hamilton, Paolo Nason, Carlo Oleari, and Emanuele Re. Jet pair production in POWHEG. *JHEP*, 04:081, 2011.

[25] Leandro G. Almeida, Seung J. Lee, Gilad Perez, Ilmo Sung, and Joseph Virzi. Top Jets at the LHC. *Phys. Rev.*, D79:074012, 2009.

[26] ATLAS Collaboration. Measurement of inclusive jet and dijet cross sections in proton–proton collisions at 7 TeV centre-of-mass energy with the ATLAS detector. *Eur. Phys. J. C*, 71:1512, 2011.

[27] ATLAS Collaboration. ATLAS measurements of the properties of jets for boosted particle searches. *Phys. Rev. D*, 86:072006, 2012.

[28] ATLAS Collaboration. Measurement of k_{T} splitting scales in $W \to \ell\nu$ events at $\sqrt{s} =$ 7 TeV with the ATLAS detector. *Eur. Phys. J. C*, 73:2432, 2013.

[29] ATLAS Collaboration. Boosted hadronic top identification at ATLAS for early 13 TeV data. ATL-PHYS-PUB-2015-053, 2015.

[30] ATLAS Collaboration. Identification of Boosted, Hadronically-Decaying W and Z Bosons in $\sqrt{s} = 13$ TeV Monte Carlo Simulations for ATLAS. ATL-PHYS-PUB-2015-033, 2015.

[31] ATLAS Collaboration. A new method to distinguish hadronically decaying boosted Z bosons from W bosons using the ATLAS detector. *Eur. Phys. J. C*, 76:238, 2016.

[32] ATLAS Collaboration. Identification of Boosted, Hadronically Decaying W Bosons and Comparisons with ATLAS Data Taken at $\sqrt{s} = 8$ TeV. *Eur. Phys. J. C*, 76:154, 2016.

[33] ATLAS Collaboration. Identification of high transverse momentum top quarks in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *JHEP*, 06:093, 2016.

[34] ATLAS Collaboration. Identification of Hadronically-Decaying W Bosons and Top Quarks Using High-Level Features as Input to Boosted Decision Trees and Deep Neural Networks in ATLAS at $\sqrt{s} = 13$ TeV. ATL-PHYS-PUB-2017-004, 2017.

[35] ATLAS Collaboration. Search for light resonances decaying to boosted quark pairs and produced in association with a photon or a jet in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. 2018.

[36] M. Bahr et al. Herwig++ Physics and Manual. *Eur. Phys. J.*, C58:639–707, 2008.

[37] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Nathan P. Hartland, and Juan Rojo. Parton distributions with QED corrections. *Nucl. Phys.*, B877:290–320, 2013.

[38] Richard D. Ball et al. Parton distributions for the LHC Run II. *JHEP*, 04:040, 2015.

[39] Christopher M. Bishop. Mixture density networks, 1994.

[40] Oliver Sim Brüning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock. *LHC Design Report*. CERN, Geneva, 2004.

[41] Andy Buckley, James Ferrando, Stephen Lloyd, Karl Nordström, Ben Page, Martin Rüfenacht, Marek Schönherr, and Graeme Watt. LHAPDF6: parton density access in the LHC precision era. *Eur. Phys. J.*, C75:132, 2015.

[42] Matteo Cacciari and Gavin P. Salam. Dispelling the n^3 myth for the k_t jet-finder. *Phys. Lett. B*, 641(1):57–61, 2006.

[43] Matteo Cacciari and Gavin P. Salam. Pileup subtraction using jet areas. *Phys. Lett.*, B659:119–126, 2008.

[44] S. Catani, Yu.L. Dokshitzer, M.H. Seymour, and B.R. Webber. Longitudinally-invariant k_t-clustering algorithms for hadron-hadron collisions. *Nuclear Physics B*, 406(1):187 – 224, 1993.

[45] Chunhui Chen. New approach to identifying boosted hadronically-decaying particle using jet substructure in its center-of-mass frame. *Phys. Rev.*, D85:034007, 2012.

[46] François Chollet et al. Keras. https://github.com/fchollet/keras, 2018.

[47] CMS Collaboration. Search for Low Mass Vector Resonances Decaying to Quark-Antiquark Pairs in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.*, 119(11):111802, 2017.

[48] CMS Collaboration. Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair. *Phys. Rev. Lett.*, 120(7):071802, 2018.

[49] ATLAS Collaboration. Atlas pythia 8 tunes to 7 tev data, atl-phys-pub-2014-021. 2014.

[50] The ATLAS Collaboration. ATLAS Detector and Physics Performance Technical Design Report. Technical report, CERN/LHCC, 1999.

[51] The ATLAS Collaboration. The ATLAS experiment at the CERN Large Hadron Collider. *Jinst*, 2008.

[52] The ATLAS collaboration. Monte Carlo Calibration and Combination of In-situ Measurements of Jet Energy Scale, Jet Energy Resolution and Jet Mass in ATLAS. 2015.

[53] The ATLAS collaboration. Measurement of the tau lepton reconstruction and identification performance in the ATLAS experiment using pp collisions at $\sqrt{s} = 13$ TeV. 2017.

[54] James Currie, Aude Gehrmann-De Ridder, Thomas Gehrmann, E. W. N. Glover, Alexander Huss, and Joao Pires. Precise predictions for dijet production at the LHC. *Phys. Rev. Lett.*, 119(15):152001, 2017.

[55] Francesco Armando Di Bello. Optimisation of the ATLAS b-tagging algorithms for the 2017-2018 LHC data-taking. *PoS*, EPS-HEP2017:733, 2017.

[56] James Dolen, Philip Harris, Simone Marzani, Salvatore Rappoccio, and Nhan Tran. Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure. *JHEP*, 05:156, 2016.

[57] S. A. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(4):325–327, April 1976.

[58] Sayipjamal Dulat, Tie-Jiun Hou, Jun Gao, Marco Guzzi, Joey Huston, Pavel Nadolsky, Jon Pumplin, Carl Schmidt, Daniel Stump, and C. P. Yuan. New parton distribution functions from a global analysis of quantum chromodynamics. *Phys. Rev.*, D93(3):033006, 2016.

[59] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979.

[60] F. Abe et al. Observation of top quark production in pp collisions with the collider detector at fermilab. *Phys. Rev. Lett.*, 74:2626–2631, Apr 1995.

[61] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. 2012.

[62] Ian J. Goodfellow et al. Generative adversarial networks.

[63] S. Abachi et al. Search for high mass top quark production in pp collisions at $sqrts = 1.8tev$. *Phys. Rev. Lett.*, 74:2422–2426, Mar 1995.

[64] Geoffrey C. Fox and Stephen Wolfram. Observables for the analysis of event shapes in e^+e^- annihilation and other processes. *Phys. Rev. Lett.*, 41:1581–1585, Dec 1978.

[65] Yoav Freund and Robert E Schapire. A short introduction to boosting. 14:771–780, 10 1999.

[66] Stefano Frixione and Giovanni Ridolfi. Jet photoproduction at HERA. *Nucl. Phys.*, B507:315–333, 1997.

[67] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[68] Ian J. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *CoRR*, 2017.

[69] L. A. Harland-Lang, A. D. Martin, P. Motylinski, and R. S. Thorne. Parton distributions in the LHC era: MMHT 2014 PDFs. *Eur. Phys. J.*, C75(5):204, 2015.

[70] P.W. Higgs. Broken symmetries, massless particles and gauge fields. *Physics Letters*, 12(2):132–133, 1964.

[71] Hantian Zhang Lucas Fowler Gokula Krishnan Santhanam Kevin Schawinski, Ce Zhang. Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit.

[72] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

[73] Andrew J. Larkoski, Ian Moult, and Duff Neill. Power Counting to Better Jet Observables. *JHEP*, 12:009, 2014.

[74] Victor Lendermann, Johannes Haller, Michael Herbst, Katja Kruger, Hans-Christian Schultz-Coulon, and Rainer Stamen. Combining Triggers in HEP Data Analysis. *Nucl. Instrum. Meth.*, A604:707–718, 2009.

[75] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, Jan 1991.

[76] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 981–990. Curran Associates, Inc., 2017.

[77] Bogdan Malaescu. An Iterative, Dynamically Stabilized(IDS) Method of Data Unfolding. In *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN,Geneva, Switzerland 17-20 January 2011*, pages 271–275, 2011.

[78] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k t jet clustering algorithm. *JHEP*, page 04:063, 2008.

[79] M. Cacciari, G. P. Salam and G. Soyez. The FastJet Project. http://fastjet.fr/.

[80] Ian Moult, Benjamin Nachman, and Duff Neill. Convolved Substructure: Analytically Decorrelating Jet Substructure Observables. *JHEP*, 05:002, 2018.

[81] V. Nagarajan and J. Z. Kolter. Gradient descent gan optimization is locally stable.

[82] Zoltan Nagy. Next-to-leading order calculation of three jet observables in hadron hadron collision. *Phys. Rev.*, D68:094002, 2003.

[83] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Phys. Rev.*, D97(1):014021, 2018.

[84] 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*, 2(11):559–572, 1901.

[85] Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian processes for machine learning, 2006.

[86] Alex Rogozhnikov et al. hep_ml: Machine learning for high energy physics. https://github.com/arogozhnikov/hep_ml, 2017.

[87] S. Catani, Y.L. Dokshitzer, H. Seymour, and B.R. Webber. Longitu- dinally invariant k t clustering algorithms for hadron hadron collisions. *Nucl. Phys*, B406:187, 1993.

[88] Gavin P. Salam. Elements of QCD for hadron colliders. In *High-energy physics. Proceedings, 17th European School, ESHEP 2009, Bautzen, Germany, June 14-27, 2009*, 2010.

[89] Gavin P. Salam. Towards Jetography. *Eur. Phys. J.*, C67:637–686, 2010.

[90] Steffen Schumann and Frank Krauss. A Parton shower algorithm based on Catani-Seymour dipole factorisation. *JHEP*, 03:038, 2008.

[91] Chase Shimmin, Peter Sadowski, Pierre Baldi, Edison Weik, Daniel Whiteson, Edward Goul, and Andreas Søgaard. Decorrelated Jet Substructure Tagging using Adversarial Neural Networks. *Phys. Rev.*, D96(7):074034, 2017.

[92] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.

[93] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2951–2959. Curran Associates, Inc., 2012.

[94] John C. Collins Davison E. Soper and George F. Sterman. Factorization of Hard Processes in QCD. *Adv.Ser.Direct.High Energy Phys.*, 5:1–91, 1988.

[95] Stephen D. Ellis and Davison E. Soper. Successive combination jet algorithm for hadron collisions. *Phys. Rev*, D48:3160–3166, 1993.

[96] Justin Stevens and Mike Williams. uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers. *JINST*, 8:P12013, 2013.

[97] Jesse Thaler and Ken Van Tilburg. Identifying Boosted Objects with N-subjettiness. *JHEP*, 03:015, 2011.

[98] Jesse Thaler and Lian-Tao Wang. Strategies to Identify Boosted Tops. *JHEP*, 07:092, 2008.

[99] S. van der Meer. Calibration of the Effective Beam Height in the ISR. 1968.

[100] Yu.L. Dokshitzer and G.D. Leder and S. Moretti and B.R. Webber. Better jet clustering algorithms. *Journal of High Energy Physics*, 1997.