STUDY OF HIGGS BOSON PRODUCTION AT HIGH TRANSVERSE MOMENTUM IN THE B-QUARK PAIR DECAY MODE

By

José Gabriel Reyes Rivera

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Physics—Doctor of Philosophy

2023

ABSTRACT

This document presents constraints on Higgs boson production at high transverse momentum using the $b\bar{b}$ channel. The study is based on data collected at $\sqrt{s} = 13$ TeV with the ATLAS detector, corresponding to an integrated luminosity of 136 fb^{-1}. The events of interest consist of two large radius jets recoilling against each other. The Higgs boson decaying to b-quarks is identified using b-tagging techniques, exploiting the experimental signature of b-hadron decays while the other jet is a fully hadronic system. $Z \to b\bar{b}$ events are used to validate experimental techniques. Upper limits at the 95% confidence level on the Higgs boson production cross section are established for transverse momenta above 450 GeV and above 1 TeV. Studies related to possible improvements of these results, by reducing the uncertainties are also discussed, such as the use of modern jet definitions like UFO jets and the development of jet substructure taggers using machine learning techniques.

ACKNOWLEDGMENTS

I would like to start by thanking my advisor Joey Huston, who has always been a source of guidance and support in all the avenues of research I wanted to explore. Thank you for the valuable feedback as well as the great stories. I would like to also thank the MSU ATLAS team of professors and postdocs for always providing me with feedback and suggestions on how to proceed with my projects. I am also thankful of the many friends I've made throughout my life. Appreciate all the support and thank you for the good times.

Termino agradeciendo a mis seres queridos. A toda mi familia, todo lo que soy y he logrado se lo debo a ustedes. Por último, a María, sin tu apoyo durante este proceso, completar esta meta no hubiese sido posible. Me inspiras a ser un mejor ser humano, a creer en mi, y a no rendirme nunca.

TABLE OF CONTENTS

Chapter 1

Introduction

Humanity's quest to understand our place in the universe has lead us to the study of what is the most fundamental representation of our reality, particles. Even though the idea of elementary particles has existed for millenia, it wasn't until the 19th century that a "modern" view of particles was defined with the discovery of atoms. Atoms are the basic particle of chemical elements and it didn't take phycisists too long to discover that atoms are in fact composite particles themselves made of protons, neutrons and electrons. With the developments of quantum physics to explain nuclear phenomena coupled with technogical advancements in acceleration physics and particle colliders, we soon found ourselves within a "particle zoo" of supossedly elementary particles during the 1950's. It wasn't until the 1960's, when physicists formulated what we call today the Standard Model (SM), that the origin of so many particles was explained as combinations of a smaller amount of true fundamental particles.

High energy physics (also known as particle physics) attempts to create a robust mathematical framework that models all the fundamental interactions observed in nature through experimental observations. The SM is constantly being tested and re-tested by continuous analysis of particle collisions produced on the largest and most complex machines ever built by humanity. Teams of scientists and engineers perform a multitude of studies to confirm with greater accuracy the established SM and test theories beyond the Standard Model

(BSM). This document presents one of those measurements in one of those experiments for one specific particle, the Higgs boson. The Higgs boson plays a fundamental role in the SM, as it is the particle responsible for the generation of the W, Z and fermion masses.

The discovery of a particle with the properties of the Higgs boson in 2012 by ATLAS [1] and CMS [2] concluded one of the main goals of the Large Hadron Collider (LHC) program [3]. In subsequent years, with a larger dataset, more studies have been made putting the measured resonance on more solid grounds [4][5]. One of those was the measurement of $H \to b\bar{b}$ published in 2018 [6]. This measurement uses final states that limit a specific Higgs boson production channel: gluon-gluon fusion (ggF), which in itself is a window to BSM physics effects [7][8].

This document explores a fully inclusive Higgs boson production using the $H \to b\bar{b}$ decay mode at very high energies. A boosted all-hadronic $H \to b\bar{b}$ search requires an extra hadronic jet. Therefore, the analysis will focus on $H(\to b\bar{b}) + j$ where both jets must have a boosted topology. It is the first study in the ATLAS collaboration that targets Higgs boson production cross-sections with transverse momenta above 1 TeV.

An analysis in a particle collision experiment requires multiple steps as a setup. You need a collider to create the collisions, a detector to collect signals and procedures to turn those signals into representations of physical objects. At the same time you cannot extract much information without simulating first what process is being produced and how it interacts with the detector. Only after that we can then define the scope of the measurement and complete an analysis that measures a physical observable. Figure 1.1 presents a diagramatic representation of the steps required to perform a collider experiment analysis by the combination of collider data and simulations.

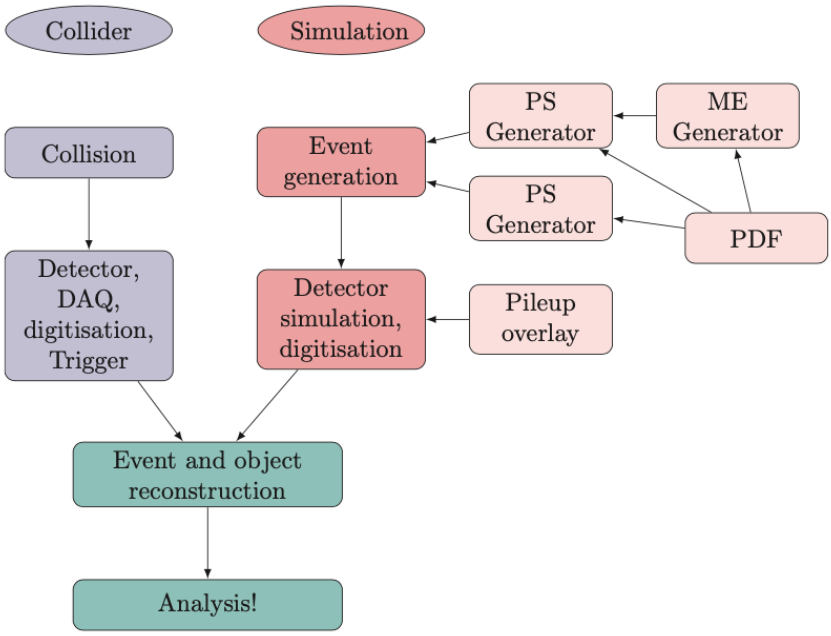


Figure 1.1: Diagram that shows the steps required to perform an analysis in a collider experiment [9].

This thesis is divided into three parts. The first part, starts with the theoretical background behind the SM, the Higgs boson, jets and hadronic collisions in Chapter 2. Then, a description of the experimental apparatus: the Large Hadron Collider and the ATLAS detector is explored in Chapter 3. Chapter 4 describes the reconstruction algorithms used to define the physics objects used within the analysis, mainly Large-R jets and the techniques to identify b-hadrons (b-tagging).

The second part of this thesis, composed of Chapters 5 and 6, presents the boosted all-hadronic $H \to b\bar{b}$ analysis and the results obtained, where the author contributed to various studies, including uncertainties of the signal modeling, multijet background modeling and combination studies. The analysis and the results shown on this document were published on the paper: "Constraints on Higgs boson production with large transverse momentum using $H \to b\bar{b}$ decays in the ATLAS detector" by the ATLAS Collaboration [10].

Finally, the third part of this thesis, consisting of Chapter 7, presents the studies performed by the author as a member of the ATLAS Jet Tagging and Scale Factor Derivation group. This chapter explores modern jet definitions and tagging techniques using Unified Flow Objects [11]. These projects support ATLAS efforts to reduce the systematic uncertainties associated with jet reconstruction as well as the development of tools to identify boosted hadronic jets, which will be useful not just for the next iterations of the analysis presented in this thesis, but for any analysis that aims to make measurements using hadronic jets in the boosted regime.

Chapter 2

Theoretical Background

To give a theoretical description of the Higgs boson, the subject of this thesis, we must first explore the underlying framework that is used to describe fundamental interactions. The framework is based on relativistic quantum mechanics and quantum field theories (QFT) that respect certain symmetry transformations. The introduction given here for these subjects is based on "The Quantum Theory of Fields" by Steven Weinberg [12], for a more detailed approach refer to the original source.

After the introduction, a summary of the QFTs that compose the Standard Model is presented. The specifics of the Electroweak (EW) interaction, where the introduction of the Higgs field becomes a neccesity to explain experimental observations of certain physical procesess, is explored. To finalize, Quantum Chromodynamics is discussed to be able to understand the origin of hadronic jets and how we model hadronic collisions.

2.1 Relativistic Quantum Mechanics

Any physical state is represented by rays in a finite complex vector space known as Hilbert space [13]. A ray $\mathcal{R}$ is a set of normalized state vectors where two states Ψ,Ψ' belong to the same ray if $\Psi' = \xi\Psi$ with ξ being an arbitrary complex number that satisfies $|\xi| = 1$.

An observable is represented by a Hermitian operator that satisfies the reality condition $A^\dagger = A$. They represent mappings $\Psi \to A\Psi$ of Hilbert space into itself. An observable

represented by A acting on state represented by ray $\mathcal{R}$ must have a definitie value α for the observable if its vectors are eigenvectors of A with eigenvalue α:

$$A\Psi = \alpha\Psi \quad \text{for } \Psi \text{ in } \mathcal{R}. \tag{2.1}$$

If a system is in a state represented by $\mathcal{R}$ and an experiment is performed to test if it is in one of the different states represented by mutually orthogonal rays $\mathcal{R}_n$ then the probability of finding it is given by

$$P(\mathcal{R} \to \mathcal{R}_n) = |(\Psi, \Psi_n)|^2 \tag{2.2}$$

Symmetries

A symmetry transformation is a change of point of view that does not change the results of an experiment [12]. That is, two observers $\mathcal{O}$, $\mathcal{O}'$, looking at the same system represented by rays $\mathcal{R}, \mathcal{R}'$ must find the same probabilities

$$P(\mathcal{R} \to \mathcal{R}_n) = P(\mathcal{R}' \to \mathcal{R}'_n). \tag{2.3}$$

Any transformation $(\mathcal{R} \to \mathcal{R}')$ is defined by an operator U on Hilbert space such that if a state Ψ is in ray $\mathcal{R}$ then $U\Psi$ is in ray $\mathcal{R}'$. The operator U must satisfy $U^\dagger = U^{-1}$.

Symmetry transformations have certain properties that define them as mathematical groups [14]. A group is a set and an operation such that any two elements of the set produce a third element of the same set. The operation must be associative, the set must have an identity element and every element of the group has an inverse. For particle physics the symmetry groups of interest for this thesis are the Lie groups. In particular, SU(n), the Lie

group of $n \times n$ unitary matrices with determinant 1, and the Poincaré group, the Lie group of Minkowski spacetime isometries.

Lie Group

A Lie group is a group of transformations $T(\theta)$ that can be described by a finite set of real continuous parameters θ^a. On the Hilbert space the unitary operator $U(T(\theta))$ can be represented by a power series

$$U\big(T(\theta)\big) = 1 + i\theta^a t_a + \frac{1}{2}\theta^b\theta^c\, t_{bc}..., \tag{2.4}$$

where t_a, t_{bc} are Hermitian operators. The operator t_a is known as the generator of the group. Higher order terms of the expansion are related to the generator by the equation

$$t_{bc} = -t_b t_c - i f^a_{bc}\, t_a. \tag{2.5}$$

It is required that the generators satisfy a set of commutation relations known as the Lie algebra:

$$[t_a, t_c] = i C^a_{bc}\, t_a \tag{2.6}$$

where $C^a_{bc} \equiv -f^a_{bc} + f^a_{cb}$ are known as structure constants. For the special case where the generators commute, the group goes from non-abelian to abelian. The unitary operator can be expressed as simply

$$U\big(T(\theta)\big) = \exp(i t_a \theta^a). \tag{2.7}$$

Poincaré Group

The Poincaré group, also known as the inhomogeneous Lorentz group is a 10-dimensional non-Abelian Lie group represented by the set of transformations with the form $T(\Lambda, a)$. A Lorentz transformation connects coordinate systems in different intertial frames in a linear form:

$$x'^{\mu} = \Lambda^{\mu}_{\nu} x^{\nu} + a^{\mu}. \tag{2.8}$$

The constant matrix Λ satisfies

$$\eta_{\mu\nu}\Lambda^{\mu}_{\rho}\Lambda^{\nu}_{\sigma} = \eta_{\rho\sigma} \tag{2.9}$$

where $\eta_{\mu\nu}$ is a Minkowski metric tensor. In Hilbert space, for an infinitesimal Lorentz transformations, the unitary operator $U(\Lambda, a)$ can be expanded in the form

$$U(1+\omega, \epsilon) = 1 + \frac{1}{2} i\omega_{\rho\sigma} J^{\rho\sigma} - i\epsilon_{\rho} P^{\rho}. \tag{2.10}$$

Commutation relations between the combinations of $J^{\mu\nu}$ and P^{μ} with themselves and each other define the Lie algebra of the Poincaré group.

The Hamiltonian operator is given by P^0 and is the generator of time translations. The momentum three-vector $\boldsymbol{P} = \{P^1, P^2, P^3\}$ is the generator of space translations. These form a subgroup of the Poincaré group. In Hilbert space pure translations are represented by

$$U(1, a) = \exp(-iP^{\mu} a_{\mu}), \tag{2.11}$$

In the same fashion, the angular-momentum three-vector $J = \{J^{23}, J^{31}, J^{12}\}$ is the generator

of rotations

$$U(R_\theta, 0) = \exp(i\boldsymbol{J} \cdot \theta). \tag{2.12}$$

Finally the other generators form the boost three-vector $K = \{J^{10}, J^{20}, J^{30}\}$.

Particles

A general one particle state $\Psi_{p,\sigma}$, with momentum p, and degrees of freedom σ, under any Lorentz transformation Λ is given by

$$U(\Lambda)\Psi_{p,\sigma} = \frac{N(p)}{N(\Lambda p)} \sum_{\sigma'} D_{\sigma',\sigma}(W(\Lambda, p))\Psi_{\Lambda p,\sigma'}, \tag{2.13}$$

where N is a normalization factor, W is a Lorentz transformation that leaves the momentum invariant (known as the little group) and $D(W)$ are the coefficients that form a representation of the little group.

Finding irreducible representations of the little group is how we classify physical states and thus how we define the different types of particles. For example, for particles with positive-definite mass, an irreducible representation $D^{(j)}_{\sigma\sigma'}$ of dimensionality $2j+1$ with $j = 0, 1/2, 1, \cdots$, can be built using the standard rotation matrices. For this case σ runs over the values $j, j-1, \cdots, -j$ for a particle with spin j.

Experimental Observables

The Hamiltonian (H) can be divided into two terms, a free-particle Hamiltonian H_0 and an interaction term V:

$$H = H_0 + V. \tag{2.14}$$

The free-particle Hamiltonian has eigenstates Φ_α, with eigenvalue E_α and the full Hamiltonian has eigenstates $\Psi_\alpha^\pm$ with the same eigenvalue as the free-particle Hamiltonian. These are known as "in"$(+)$ and "out" $(-)$ states and can be written in terms of the free-particle eigenstates:

$$\Psi_\alpha^\pm = \Omega(\mp\infty)\Phi_\alpha, \tag{2.15}$$

where

$$\Omega(\tau) = \exp(+iH\tau)\exp(-iH_0\tau). \tag{2.16}$$

The in and out states contain the particles described by the label α if observations are made at $\tau \to \pm\infty$.

The probability amplitude for a transition of states $\alpha \to \beta$ is governed by the scalar product of the "in"$(+)$ and "out"(-) states known as the S-matrix:

$$S_{\beta\alpha} = (\Psi_\beta^-, \Psi_\alpha^+) \equiv (\Phi_\beta, S\Phi_\alpha). \tag{2.17}$$

Where S is the S-operator defined as

$$S = \Omega(\infty)^\dagger\Omega(-\infty). \tag{2.18}$$

The master formula to interpret calculations of S-matrix elements in terms of predictions for actual experiments is

$$S_{\beta\alpha} \equiv -2\pi i\delta^4(p_\beta - p_\alpha)M_{\beta\alpha}. \tag{2.19}$$

Here the delta function ensures the conservation of total energy and momentum and $M_{\beta\alpha}$ represents the non-trivial scattering matrix elements.

Decay Rate

The decay rate for a single particle state α into a general multi-particle state β is given by

$$d\Gamma(\alpha \to \beta) = 2\pi |M_{\beta\alpha}|^2 \delta^4(p_\beta - p_\alpha) d\beta. \tag{2.20}$$

When multiple decay modes are available to a specific particle the total decay rate will be the sum of all of the individual modes

$$\Gamma_{\text{total}} = \sum_{i=1}^{n} \Gamma_i. \tag{2.21}$$

Then, it is useful to define the branching fractions to quantify the probability of each specific decay mode. The branching fraction of mode i is given by

$$B_i = \frac{\Gamma_i}{\Gamma_{\text{total}}}. \tag{2.22}$$

Cross-Section

When α is a two particle state, we can calculate the transition rate per flux, known as the differential cross section, using the decay rate

$$d\sigma(\alpha \to \beta) = d\Gamma(\alpha \to \beta)/\Phi_\alpha = (2\pi)^4 u_\alpha^{-1} |M_{\beta\alpha}|^2 \delta^4(p_\beta - p_\alpha) d\beta, \tag{2.23}$$

where Φ_α is defined as the product of the density and the relative velocity u_α between the two particles in state α. When the differential cross section is integrated over all the possible

configurations i we call it the total or inclusive cross-section

$$\sigma_{\text{total}} = \sum_{i=1}^{n} \sigma_i. \tag{2.24}$$

Interactions

To consider the interaction term we rewrite the S-operator as a Dyson series for the time ordered interaction Hamiltonian density $\mathcal{H}(x)$ defined as:

$$S = 1 + \sum_{n=1}^{\infty} \frac{(-i)^n}{n!} \int d^4x_1 \cdots d^4x_n \, T\{\mathcal{H}(x_1) \cdots \mathcal{H}(x_n)\}, \tag{2.25}$$

where

$$V(\tau) \equiv \exp(H_0\tau) V \exp(-iH_0\tau) = \int d^3x \mathcal{H}(\boldsymbol{x}, t). \tag{2.26}$$

Then it is possible to write an asymptotic expansion of the S-operator in whatever coupling constant factors appear in the inveraction terms of the Hamiltonian density. This technique is known as perturbation theory.

2.2 Quantum Field Theory

For the Hamiltonian density to satisfy both Lorentz invariance and the cluster decomposition principle, the Hamiltonian density must be constructed as a function of creation and annihilation fields. A creation field is defined to have a creation operator ($a^{c\dagger}(p, \sigma)$) that adds a particle to the list of particles in a physical state. The annihilation field contains the annihilation operator ($a(p, \sigma)$) and does the opposite, it removes a particle from any state in which it acts. A general quantum field in the irreducible (A, B) representation of the ho-

mogeneous Lorentz group, where A and B are the spins, is defined by a linear combination of creation and annihilation fields:

$$\psi_{ab}(x) = (2\pi)^{-3/2} \sum_{\sigma} \int d^3p \left[u_{ab}(p,\sigma) a(p,\sigma) e^{ip\cdot x} + (-)^{2B} v_{ab}(p,\sigma) a^{c\dagger} e^{-ip\cdot x} \right]. \tag{2.27}$$

where the coefficients a, b are integers or half-integers running over the values

$$a = -A,\, -A+1,\, \cdots,\, +A \quad \text{and} \quad b = -B,\, -B+1,\, \cdots,\, +B. \tag{2.28}$$

A field according to the (A, B) representation has components that rotate like objects of spin j with

$$j = A+B,\, A+B-1,\, \cdots,\, |A-B|. \tag{2.29}$$

Fields with integer values for the spin commute with each other and are classified as bosons. Bosons do not obey the Pauli exclusion principle and thus are described by Bose-Einstein statistics. On the other hand, half-integer spin fields, known as fermions, anticommute with each other. Fermions obey the Pauli exclusion principle and therefore, a system of fermions follow Fermi-Dirac statistics.

Lagrangians

In practice it is preferable to work with Langrangians (L) instead of the Hamiltonians (H). These two quantities are related to each other by taking the Legendre transformation:

$$H = \sum_{l} \int d^3x\, \Pi_l(\boldsymbol{x}, t) \dot{\Psi}^l(\boldsymbol{x}, t) - L[\Psi(t), \dot{\Psi}(x)], \tag{2.30}$$

where Ψ is a set of generic fields, Π are the conjugate fields and the dotted variables represent time-derivatives. The conjugate field is defined using variational derivatives:

$$\Pi_l(\boldsymbol{x}, t) = \frac{\delta L[\Psi(t), \dot{\Psi}(t)]}{\delta \dot{\Psi}^l(\boldsymbol{x}, t)}. \tag{2.31}$$

These are knwon as the Euler-Lagrange equations, and their time derivatives $\dot{\Pi}_l(\boldsymbol{x}, \boldsymbol{t})$ are the equations of motion. Defining the Lagrangian density $\mathcal{L}$,

$$L[\Psi(t), \dot{\Psi(t)}] = \int d^3x \, \mathcal{L}(\Psi(\boldsymbol{x}, t), \nabla\Psi(\boldsymbol{x}, t), \dot{\Psi}(\boldsymbol{x}, t)) \tag{2.32}$$

we can express the Euler-Lagrange equations in their usual form

$$\partial_\mu \frac{\partial \mathcal{L}}{\partial(\partial_\mu \Psi^l)} = \frac{\partial \mathcal{L}}{\partial \Psi^l}. \tag{2.33}$$

Scalar Fields

A scalar field is a field of type (0,0) in the irreducible representation of the homogeneous Lorentz group and therefore are spin 0 fields. A general Langrangian density $\mathcal{L}$ for a massive free scalar field Φ is given by

$$\mathcal{L} = -\frac{1}{2}\partial_\mu \Phi \partial^\mu \Phi - \frac{m^2}{2}\Phi^2. \tag{2.34}$$

The Euler-Lagrange equation is then

$$(\partial_\mu \partial^\mu - m^2)\Phi = 0, \tag{2.35}$$

which is the usual Klein-Gordon equation.

Vector Fields

A vector field is a field of type $(\frac{1}{2}, \frac{1}{2})$ in the irreducible representation of the homogenous Lorentz group. Therefore they can be spin 0 or spin 1. The spin 0 vector field is just the derivative of a spin 0 scalar particle $(\partial_\mu \Phi)$. For a massive spin 1 vector field A_μ and no external currents the Langrangian density $\mathcal{L}$ is

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - m^2 A_\mu A^\mu, \tag{2.36}$$

where $F_{\mu\nu} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu$ is the field strength tensor. In cojunction with $\partial_\mu A^\mu = 0$, the Euler-Lagrange equation takes the form:

$$(\partial_\mu \partial^\mu - m^2) A^\mu = 0, \tag{2.37}$$

which is known as the Proca equation. This implies that each component of the field fulfils the Klein-Gordon equation.

Dirac Fields

Dirac fields represent particles of spin 1/2 and are of the type $(\frac{1}{2}, 0) \bigoplus (0, \frac{1}{2})$ in the irreducible representation of the homegeneous Lorentz group. A general Langrangian density for Dirac fields is of the form

$$\mathcal{L} = -\bar{\psi}(\gamma^\mu \partial_\mu + m)\psi, \tag{2.38}$$

where $\bar{\psi} = \psi^\dagger \gamma^0$ is the Dirac adjoint. The Euler-Lagrange equations for ψ is known as the Dirac equation

$$(i\gamma^\mu \partial_\mu + m)\psi(x) = 0. \tag{2.39}$$

Taking the hermitian conjugate of the Dirac equation and multiplying on the right by γ^0, the adjoint Dirac equation can be derived. When both solutions are combined we arrive at

$$(\partial_\mu \partial^\mu - m^2)\psi^\mu = 0. \tag{2.40}$$

Therefore, each component of the Dirac field satisfies the Klein-Gordon equation.

2.3 Standard Model

The Standard Model (SM) is the collection of quantum field theories (QFTs) that describe three of the four fundamental forces of the universe and classifies all the elementary particles currently known. The three forces it describes are: the strong interaction, the weak interaction and electromagnetism. The strong interaction is described by quantum chromodynamics (QCD). Electromagnetism (EM) and the weak interaction are unified into the same theory, called the electroweak interaction (EW). On its entirety the SM respects the symmetry under the non-abelian $\mathrm{SU}(3)_C \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ gauge group.

The SM contains both types of particles, fermions and bosons. The fermions can be divided into quarks and leptons and they exist in three generations; each one with increasing mass. There are six types of quarks: up, down, strange, charm, bottom and top. There are also six leptons: electron, muon, tau and their respective neutrinos. The bosons are divided into the spin 1 (vector) force carriers and the spin 0 (scalar) Higgs boson. The force carriers

are the photon γ for the electromagnetic interaction, the gluon g for the strong interaction and the $W^{\pm}$, Z bosons that mediate the weak interaction. Figure 2.1 summarizes all the SM fundamental particles.

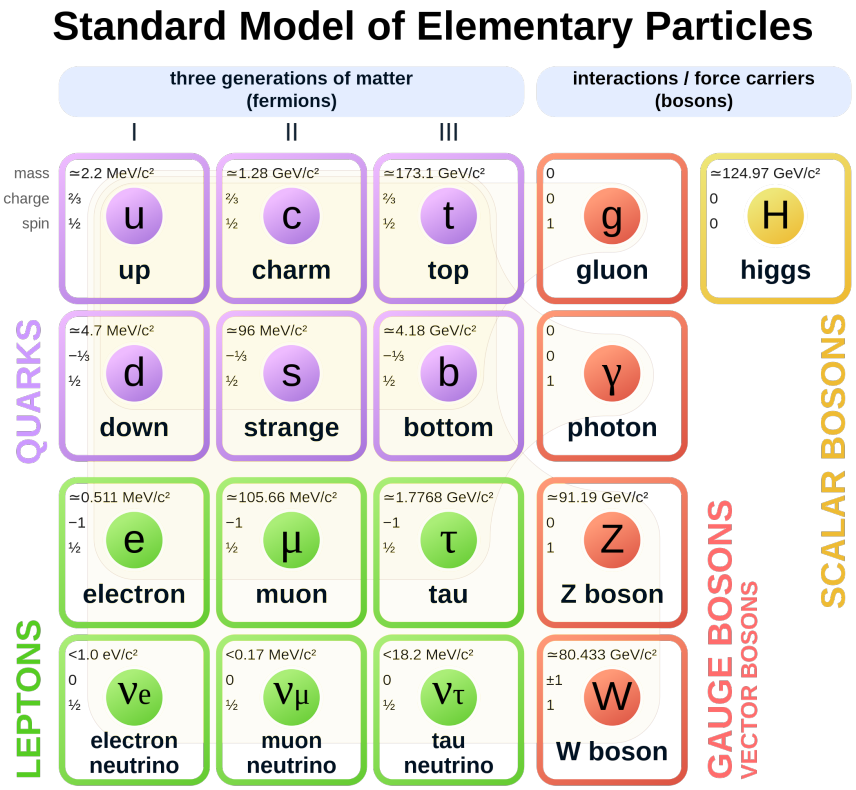


Figure 2.1: Standard Model of elementary particles showing the twelve fundamental fermions and five fundamental bosons [15].

2.4 Quantum Chromodynamics

Quantum Chromodynamics (QCD) is the non-abelian gauge field theory that describes the strong interaction, a force only felt by quarks and gluons. Quarks are in the fundamental representation of the SU(3) color group. They are represented by quark field spinors ψ_i, where i the color-index that goes from 1 to $N_c = 3$ (the number of colors). The gluon is a vector field, $\mathcal{A}^a_\mu$, where a runs from 1 to $N_c^2 - 1 = 8$. Gluons transform under the adjoint

representation of SU(3) color group. The eight 3×3 matrices t^a_{ij} are the generators of SU(3), which can be represented explicitly by the Gell-Mann matrices (λ^a) given by $t^a = \lambda^a/2$. The QCD Lagragian is given by

$$\mathcal{L} = \bar{\psi}_i(i\gamma^\mu \partial_\mu \delta_{ij} - g_s \gamma^\mu t^a_{ij} \mathcal{A}^a_\mu - m\delta_{ij})\psi_j - \frac{1}{4} G^a_{\mu\nu} G^{a\mu\nu}. \tag{2.41}$$

The QCD coupling constant is $\alpha_s = g_s^2/4\pi$. The gluon field strength tensor $G^a_{\mu\nu}$ is given by

$$G^a_{\mu\nu} = \partial_\mu \mathcal{A}^a_\nu - \partial_\nu \mathcal{A}^a_\mu - g_s f_{abc} \mathcal{A}^b_\mu \mathcal{A}^c_\nu \tag{2.42}$$

where f_{abc} are the structure constants of the SU(3) group. The coupling constant α_s is a function of the scale at which the process happens. Figure 2.2 shows different experimental measurements of the coupling constant α_s as a function of energy scale Q. Quarks and gluons cannot be isolated, only color-singlet (color neutral) combinations of them can be observed as free particles. Given that the coupling is really strong at low energies, it leads to the confinement of quarks and gluons into hadrons, a non-perturbative process called hadronization. On the other hand, for hard processes the strong coupling is weak and the theory becomes suitable to perturbative theory techniques, a phenomenon known as asymptotic freedom.

Before hadronization occurs, a hard scattering event involving a QCD interaction starts with the interacting particles radiating more gluons and quarks (parton showering) until the parton energy gets to the hadronization scale (Λ). This leads to the formation of collimated sprays of energetic hadrons, which we call jets [17]. In Chapter 4 we will discuss more about the particular type of jets used in this analysis, how we define them, the rules to group

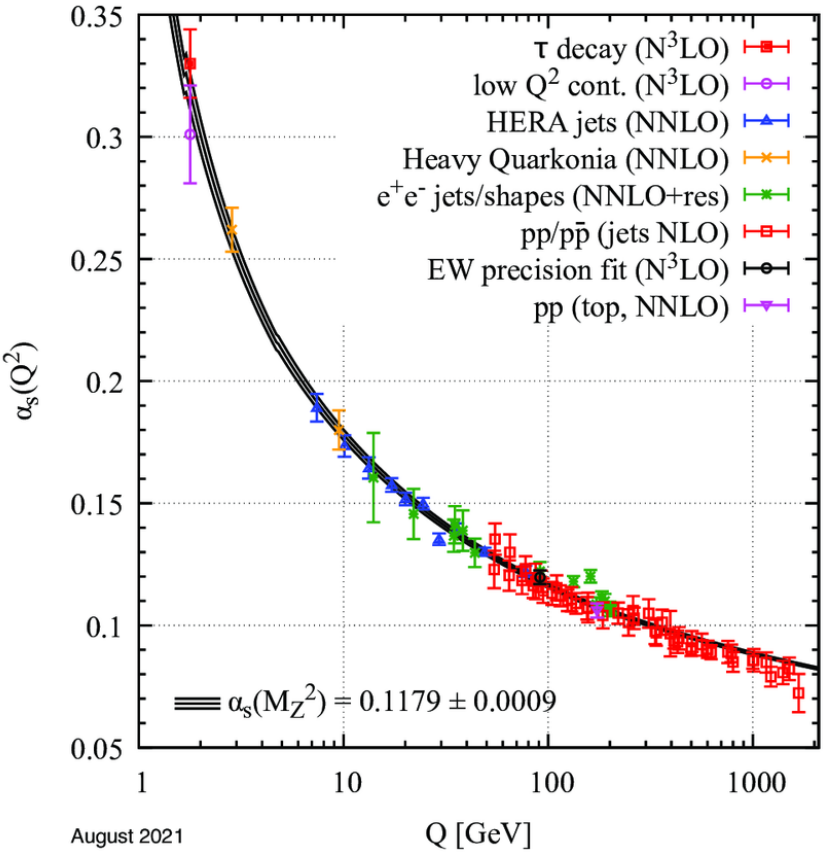


Figure 2.2: Measurements of α_s as a function of the energy scale Q. The degree of QCD perturbation theory used in the extraction of α_s is indicated in parentheses [16].

particles and how to determine their momenta.

2.5 Electroweak Theory

The standard model of electroweak interactions is based on the gauge field theory that respects the symmetry of the $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ product group, with gauge bosons W^i_μ, $i = 1, 2, 3$ and B_μ, and their corresponding gauge coupling constants g and g'. The fundamental quantities are the $\mathrm{SU}(2)_L$ weak isospin and the $\mathrm{U}(1)_Y$ weak hypercharge. The right-handed fields are singlets in SU(2). On the other hand, the left-handed fermion fields transform as

doublets

$$\Psi_i = \begin{pmatrix} \nu_i \\ l_i^- \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} u_i \\ d_i' \end{pmatrix} \tag{2.43}$$

where $d_i' = \sum_j V_{ij} d_j$ and V is the Cabibbo-Kobayashi-Maskawa (CKM) mixing matrix [18]. The CKM matrix is a unitary 3×3 matrix and it describes the quark flavor mixing in weak interactions. In one of the standard parametrizations it can be expressed as

$$V_{CKM} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & 0 & s_{13}e^{-i\delta} \\ 0 & 1 & 0 \\ -s_{13}e^{i\delta} & 0 & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.44}$$

where $s_{ij} = \sin\theta_{ij}$, $c_{ij} = \cos\theta_{ij}$ and δ is the phase responsible for CP-violating phenomena in flavor-changing processes.

Higgs Mechanism

For the EW theory to be consistent with observations we require a mechanism that makes the W and Z bosons massive to render the weak interaction short range. This can be achieved by the introduction of a scalar field Φ, called the Higgs field, that causes a spontaneous breaking of the electroweak gauge symmetry (EWSB) [16]. The Higgs potential is of the form:

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2 \tag{2.45}$$

The Higgs field is a self-interacting SU(2) complex doublet with a weak hypercharge $Y = 1$ normalized such that it has a neutral charge ($Q = T_{3L} + Y/2$):

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{2}\phi^+ \\ \phi^0 + ia^0 \end{pmatrix} \tag{2.46}$$

where ϕ^0 and a^0 are the CP-even and CP-odd neutral components and ϕ^+ is the complex charged component. If the quadratic term of $V(\Phi)$ is negative the neutral component of the scalar doublet acquires a non zero vacuum expectation value (VEV)

$$\langle \Phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \tag{2.47}$$

with $\phi^0 = H + \langle \phi^0 \rangle$ and $\langle \phi^0 \rangle = v$, inducing the spontanueos breaking of the gauge symmetry $\mathrm{SU(3)}_C \times \mathrm{SU(2)}_L \times \mathrm{U(1)}_Y$ into $\mathrm{SU(3)}_C \times \mathrm{U(1)}_{\mathrm{em}}$. Three of the four generators of the $\mathrm{SU(2)}_L \times \mathrm{U(1)}_Y$ are spontaneously broken; this imples the existence of three massless Goldstone bosons which can be identified as three of the four Higgs field degrees of freedom. Figure 2.3 illustrates the fact that the Higgs field VEV is not a single state with an energy of 0 and instead it has degenerate vacua with a VEV of v.

The kinetic term of the Higgs Lagrangian shows how the Higgs field couples to the W_μ and B_μ gauge fields of the $\mathrm{SU(2)}_L \times \mathrm{U(1)}_Y$ local symmetry:

$$\mathcal{L}_{\mathrm{Higgs}} = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi) \tag{2.48}$$

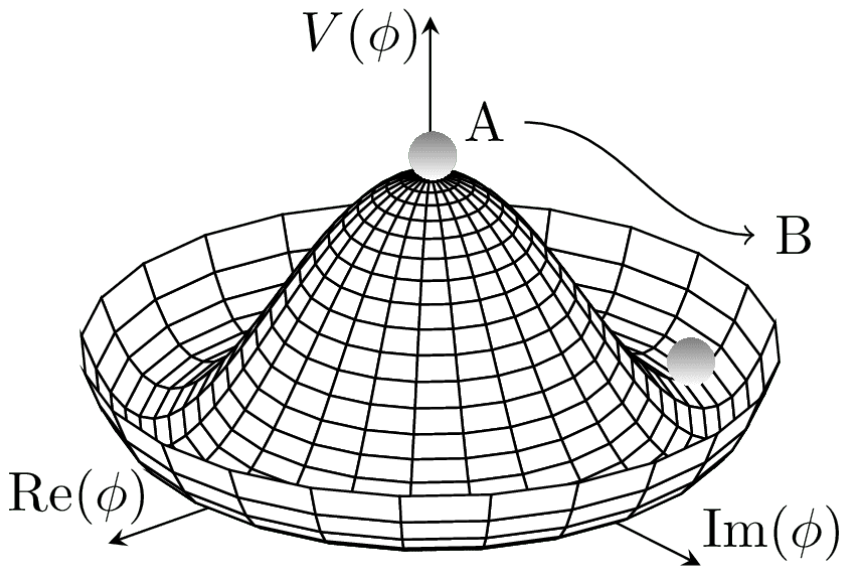


Figure 2.3: Illustration of the Higgs potential $V(\Phi)$. After EWSB, the Higgs field VEV is not a single state with an energy of 0 (represented by A), it has degenerate vacua with a VEV of v (represented by B) [19].

The covariant derivative is:

$$D_\mu = \partial_\mu + ig\sigma^a W_\mu^a/2 + ig' Y B_\mu/2 \tag{2.49}$$

where g and g' are the $\mathrm{SU}(2)_L$ and $\mathrm{U}(1)_Y$ couplings and σ^a are the Pauli matrices. Expanding the kinetic term and rearranging then we can see that the presence of the Higgs field gives mass to the gauge bosons. Examining the mass term of the Lagrangian:

$$\mathcal{L}_m = \frac{g^2 v^2}{8}(W_1^2 + W_2^2) + \frac{v^2}{8}(gW_3^\mu - g' B^\mu) \tag{2.50}$$

we can see that the physical massive bosons are combinations of the original gauge bosons. Rewriting in terms of the physical fields:

$$\mathcal{L}_m = \frac{1}{2} m_W^2 W_\mu^+ W^{-\mu} + \frac{1}{2} m_Z^2 Z_\mu Z^\mu \tag{2.51}$$

where

$$W_\mu^\pm = \frac{W_\mu^1 \mp iW_\mu^2}{\sqrt{2}}, \quad Z_\mu = \frac{1}{\sqrt{g^2+g'^2}}(gW_\mu^3 - g'B_\mu) \tag{2.52}$$

and

$$m_W^2 = \frac{g^2v^2}{4}, \quad m_Z^2 = \frac{(g'^2+g^2)v^2}{4} \tag{2.53}$$

There is one combination of W_3 and B, orthogonal to Z, that is not present in the mass Lagrangian; it corresponds to the photon:

$$A_\mu = \frac{1}{\sqrt{g^2+g'^2}}(g'W_3^\mu + gB^\mu) \tag{2.54}$$

Of the initial four degrees of freedom of the Higgs field, three of them were absorbed by the $W^\pm$ and Z bosons and the remaining degree of freedom, H, becomes the physical Higgs boson. The Higgs boson is a CP-even spin 0 (scalar) particle with a mass given by $m_H = \sqrt{2\lambda}v$, where λ is the self coupling parameter. The Higgs field expectation value is fixed by the Fermi coupling costant G_F: $v = (\sqrt{2}G_F)^{-1/2} \approx 246$ GeV.

Fermions acquire mass through interactions with the Higgs field, also known as Yukawa interactions. Yukawa couplings respect all the symmetries of the SM but generates fermion masses after the EWSB occurs.

$$\mathcal{L}_{\text{Yukawa}} = -\hat{h}_{d_{ij}}\bar{q}_{L_i}\Phi d_{R_j} - \hat{h}_{u_{ij}}\bar{q}_{L_i}i\sigma_2\Phi^* u_{R_j} - \hat{h}_{d_{ij}}\bar{l}_{L_i}\Phi e_{R_j} + h.c. \tag{2.55}$$

After the Higgs field acquires a VEV the fermions acquire a mass in the form: $m_{f_i} = h_{f_i}v/\sqrt{2}$, where h_{f_i} is the Yukawa coupling and $i = 1, 2, 3$ refer to the three families of the up-quark, down-quark or charged lepton sectors.

The coupling of the Higgs boson to other fundamental particles is dictated by how massive the particle is. The interaction is strongest with particles such as the W/Z bosons and to top quarks. For fermions the coupling is linearly proportional to the fermion mass ($g_{Hf\bar{f}} = m_f/v$) and for bosons it is proportional to the square of the boson masses ($g_{HVV} = 2m_V^2/v$).

2.6 Higgs Boson Phenomenology

This section explores the Higgs boson production modes and the branching ratios for all of its decay channels [20]. We will finish the section by exploring how a boosted Higgs boson can be used as a probe for beyond the Standard Model (BSM) physics.

Production and Decays

Experimentally, the Higgs boson mass is measured to be $m_H = 125.25 \pm 0.17$ GeV. To produce a Higgs boson we require a collider experiment with a large center-of-mass (CoM) energy such as the Fermilab Tevatron [21] or the CERN LHC [3]. Given that we are studying the Higgs boson at the LHC, we need to first understand the production mechanisms in hadron (on this case: proton-proton (pp)) collisions. The principal production mode at the LHC is the gluon-gluon fusion (ggF) process, followed by weak-boson (vector-boson) fusion (VBF). Other production modes include associated production with a gauge boson (VH), associated production with $t\bar{t}$ quark pair ($t\bar{t}H$) or associated production with a single top quark (tHq). Figure 2.4 illustrates the leading order Feynman diagrams for some of the Higgs boson production modes at the LHC.

Figure 2.5 shows the Higgs boson production cross section for pp collisions at a CoM energy of $\sqrt{s} = 13$ TeV as a function of Higgs boson mass. For a Higgs boson mass of $m_H =$

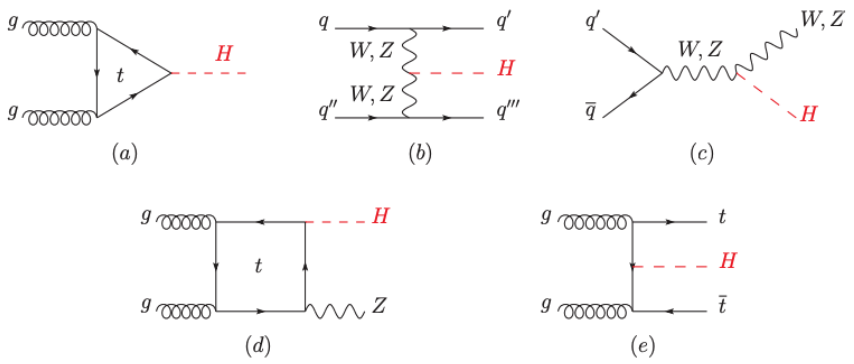


Figure 2.4: Leading order Feynman diagrams that contribute to Higgs boson production in (a) ggF, (b) VBF, (c) Higgs-strahlung, (d) associated production with a gauge boson and (e) associated production with a pair of top-quarks [16].

125 GeV, the total production cross section is 55.1 pb. The production mode breakdown as a percentage is as follows: ggF is the largest contribution with 88% of the total production cross section, VBF accounts for around 7%, VH (WH and ZH combined) sum to 4% and $t\bar{t}H$ is close to 1%.

Detecting the Higgs boson requires an understanding of all the relevant decay channels. The Higgs boson has a natural width of 4 MeV, meaning it has a lifetime of the order of 10^{-22} seconds. Figure 2.6 shows the Higgs boson decay branching ratios as a function of Higgs boson mass. The dominant decay mode is $H \to b\bar{b}$, with a branching fraction of about 58%. This decay mode is the focus of the study presented in this thesis. Even though it is the most common decay mode, the channel suffers from large backgrounds, primarily from $b\bar{b}$ production. To measure its mass the two high mass-resolution sensitive channels are $H \to \gamma\gamma$ and $H \to ZZ \to 4l$, which despite having low branching ratios, have clean signals. These two channels were used for the original discovery of the Higgs boson in 2012 [23][24].

$H \to b\bar{b}$ is a promising channel to study the Higgs field coupling to quarks. For the direct observation of the Higgs boson decaying to a pair of b-quarks, the production mode used in the original studies was the VH channel. The presence of a vector boson reduces the

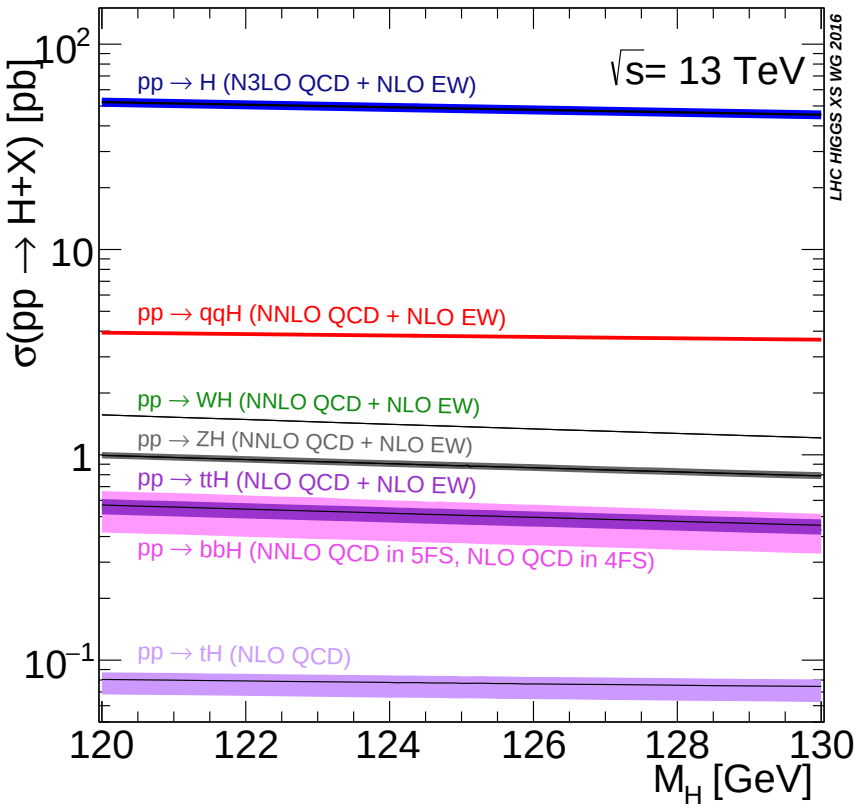


Figure 2.5: Standard Model Higgs boson production cross sections at $\sqrt{s} = 13$ TeV as a function of Higgs boson mass for pp collisions [22].

relative background because the leptonic decay of the W and Z enable efficient triggering and a significant reduction of the multijet background. The Higgs candidate was reconstructed from two b-tagged jets in the event. Both ATLAS and CMS observed a significance of the excesses greater than 5σ when combining Run 1 and Run 2 data [26][27].

Sensitivity for an inclusive search for $H \to b\bar{b}$ in the ggF production mode is limited because of the large amount of background from the inclusive production of $pp \to b\bar{b} + X$. From the Run 1 dataset, no meaningful results exist. The analysis presented in this document is the first ever performed by the ATLAS collaboration that attempts to do this with the full Run 2 data, with the sensitivity increased by focusing on Higgs boson production at high transverse momentum.

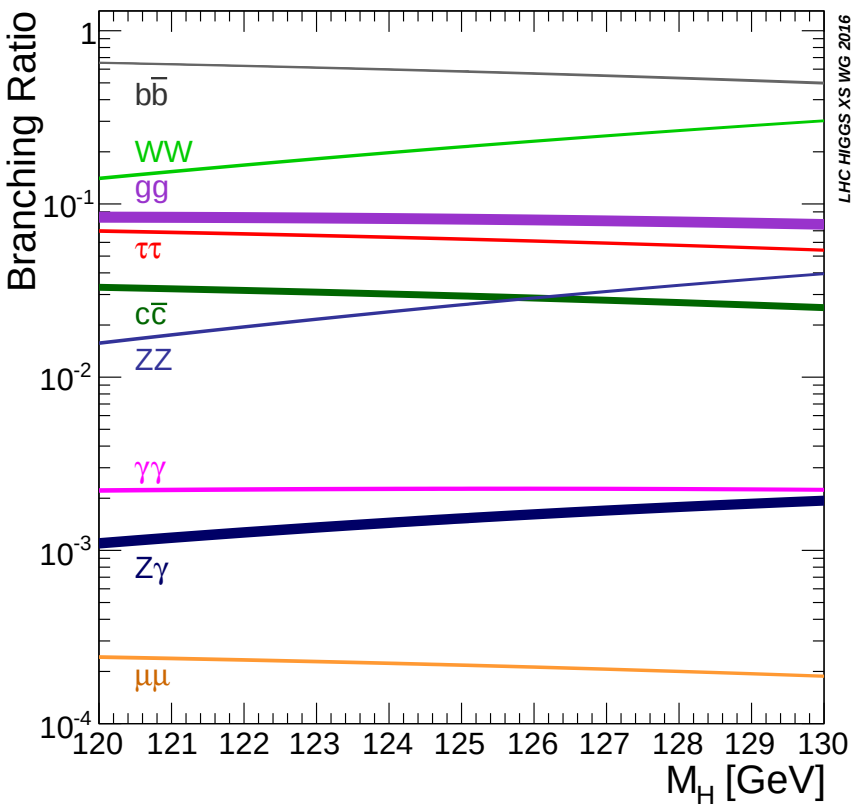


Figure 2.6: Standard Model Higgs boson decay branching ratios as a function of Higgs boson mass [25].

Boosted Higgs boson

To boost a Higgs boson to high momenta, it is required to have an extra jet in the event for the Higgs boson to recoil against. Figure 2.7 contains a couple examples of diagrams that contribute to the $H + j$ production cross section.

A Higgs boson with high transverse momentum can be used to set constraints for beyond the Standard Model (BSM) [29][30]. The inclusion of a set of dimension-six operators [31] in the SM lagrangian that describe physics at a scale Λ above the EW scale, modify the Yukawa operator, provide a contact interaction of the Higgs boson with gluons and introduce the chromomagnetic dipole moment operator [32]. All of these interactions have an impact in the Higgs boson p_{T} distribution. In particular, when considering the chromomagnetic dipole

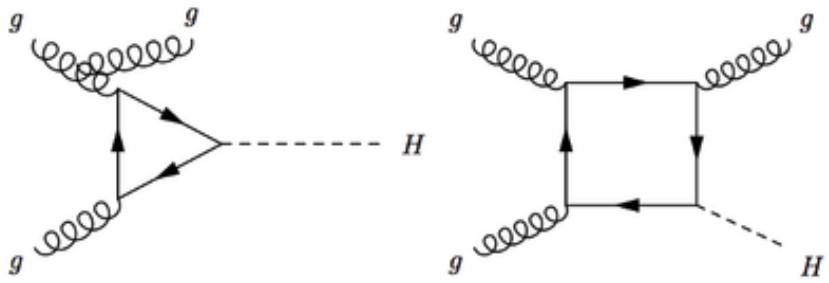


Figure 2.7: Examples of gluon-gluon fusion Feynman diagrams that contribute to the $H + j$ process [28].

moment operator in the case of single Higgs production, it has been shown to have a large impact at high p_{T} [33]. Figure 2.8 illustrates these results. The extra term related to the chromomagnetic dipole moment in this context is of the form

$$\frac{c_3}{\Lambda^2}\mathcal{O}_3 = c_3\frac{g_s m_t}{2v^3}(v+h)G^A_{\mu\nu}(\bar{\psi}_L\sigma^{\mu\nu}t^A\psi_R + h.c.) \tag{2.56}$$

where c_3 is the Wilson coefficient, $\sigma^{\mu\nu}$ are the Pauli matrices and ψ is the spinor representing the top quarks.

2.7 Simulation of proton-proton Collisions

Any cross section that involves QCD interactions of initial-state hadrons is inherently not calculable in perturbative QCD. Structure functions are needed to describe these complex objects. The structure functions are given in terms of non-perturbative parton distribution functions (PDFs) . A PDF $f_{q/p}(x)$ represents the number density of quarks of type q inside a hadron that carry a fraction x of its longitudinal momentum. A typical hadron-hadron

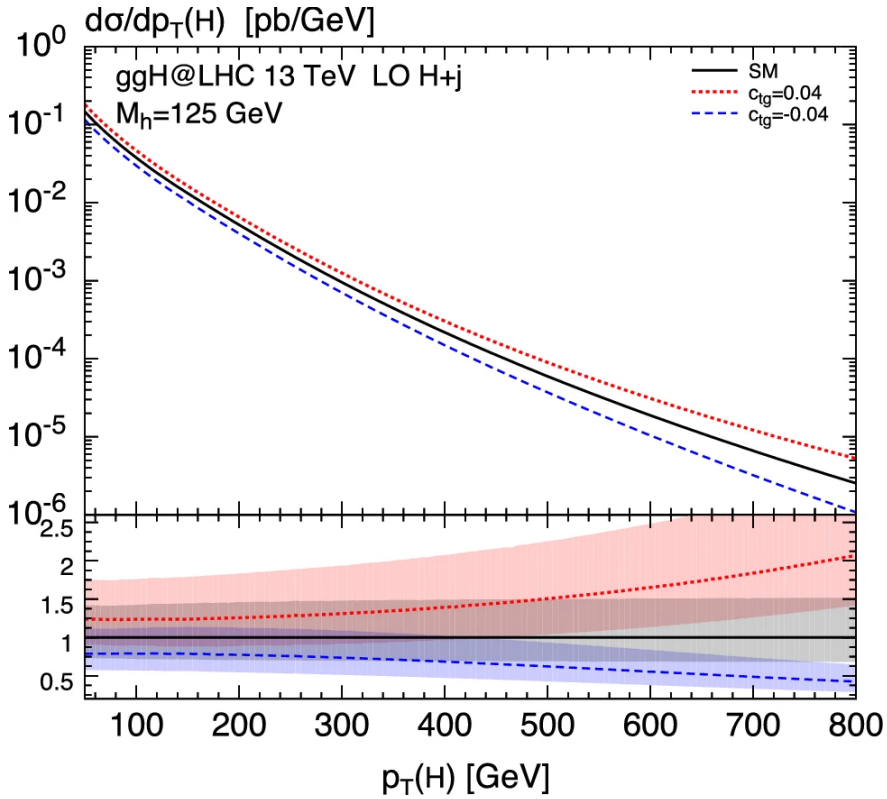


Figure 2.8: Impact of the chromomagnetic operator on the p_{T} spectrum of the Higgs boson. The bottom panel shows the ratio with respect to the SM prediction [33].

(h_1,h_2) collision cross-section is of the form

$$\sigma(h_1,h_2\to X)=\sum_{n=0}^{\infty}\alpha_s^n(\mu_R^2)\sum_{i,j}\int dx_1dx_2\, f_{i/h_1}(x_1,\mu_F^2)f_{j,h_2}(x_2,\mu_F^2)\times\hat{\sigma}^{(n)}_{ij\to X} \qquad (2.57)$$

where s is the squared center-of-mass energy of the collision, μ_R is the renormalization scale and μ_F is the factorization scale, the scale at which emissions with transverse momenta below it are accounted for within the PDFs. The parton level cross-section $\hat{\sigma}_{ij\to X}(x_1x_2s,\mu_R^2,\mu_F^2)$ can be calculated using perturbative QCD.

PDFs are determined empirically by fitting a large number of cross section data points from many experiments, including Deep Inelastic Scattering experiments (DIS) and hadron collider experiments. To evolve those functions to different energy scales, the Dokshitzer-

Gribov-Lipatov-Altarelli-Parisi (DGLAP) [34] equation is employed. Usually the default choice of the scales is $\mu_R = \mu_F = Q$. Figure 2.9 shows the CT18 parton distribution functions at different energy scales.

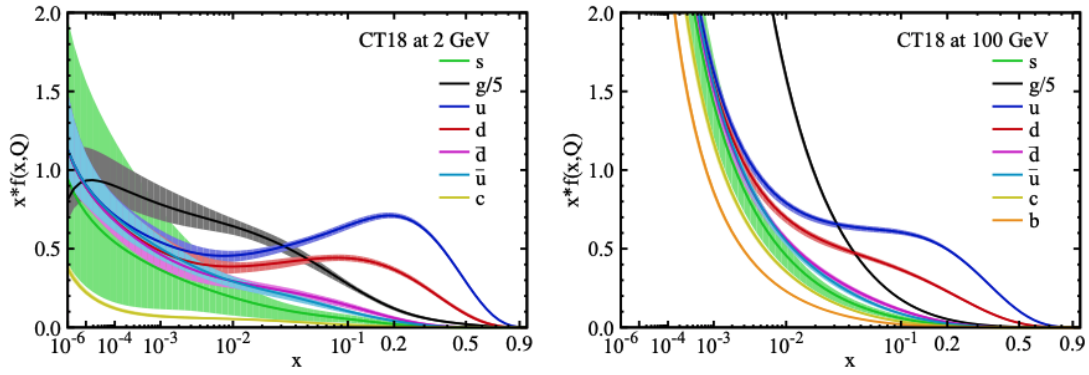


Figure 2.9: The CT18 parton distribution functions at $Q = 2$ GeV and $Q = 100$ GeV for $u, \bar{u}, d, \bar{d}, s = \bar{s}$, and g [35].

The parton-hadron transition is non-perturbative, so it is not possible to calculate quantities like the energy spectrum of hadrons in high-energy collisions. Nevertheless it is possible to factorize the perturbative and non-perturbative behaviours using the concept of fragmentation functions. Similarly to PDFs, they depend on a factorization scale and satisfy the DGLAP evolution equation.

To create simulations of this entire process, we use parton-shower Monte Carlo (MC) event generators such as Pythia [36], Herwig [37] and Sherpa [38]. They provide a full simulation of QCD events at the level of measurable particles. Figure 2.10 shows a sketch of a pp collision as simulated by a MC generator. There are MC generators that only produce the matrix elements, such as MadGraph5 aMC@NLO [39], which are then passed to a shower/hadronization program such as Pythia. The parton shower MC programs model the gluon emissions and gluon splittings simulating a cascade of particles. Each emission is generated at a lower scale, with the emissions stopping at a scale of the order of 1 GeV.

At this point a hadronization model is used to combine the resulting particles into hadrons. There are different hadronization/shower models which might have slight differences in the end result. In practice multiple programs are considered when generating MC predictions for an analysis and the differences are quantified as a source of uncertainty.

The remnants of hadron collisions also have to be modeled; this is refered to as the underlying event (UE). The UE is usually implemented by introducing multiple parton interactions (MPI) at a scale of a few GeV. Similarly, pile-up also has to be simulated. Pile-up refers to any other *pp* collisions in addition to the collision of interest.

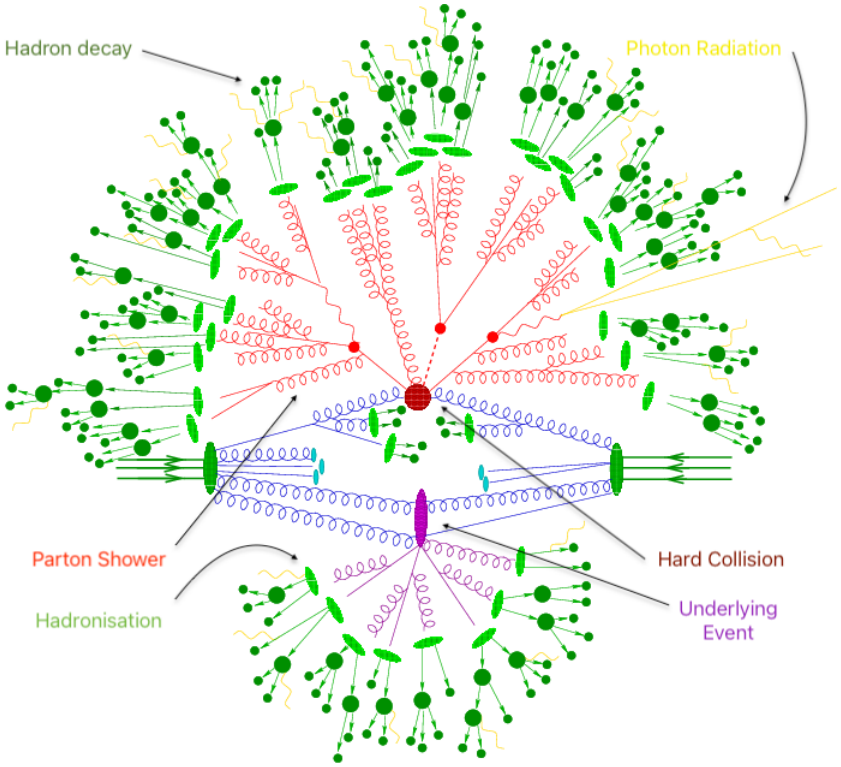


Figure 2.10: Sketch of a proton-proton collision as simulated by a multi-purpose Monte Carlo event generator [40].

As a last step all of the particles/jets generated and their kinematic variables, at "truth-level", are subjected to a detector simulation. All of the particle interactions with the different detector modules are done with GEANT4 [41]. GEANT4 is a toolkit for simulating

the passsage of particles through matter. After the detector simulation is completed all the kinematic variables modified are then refered to as being at "reconstructed-level".

Chapter 3

Experimental Apparatus

3.1 Large Hadron Collider

To study physics at small scales it is necesary to accelerate particles to high energies and have them interact, that is, make them collide. This is done by the Large Hadron Collider (LHC) [3] at the European Organization for Nuclear Research, known as CERN located at the French-Swiss border of Geneva. The LHC is the largest and most powerful particle accelerator ever built and is part of the CERN accelerator complex shown in Figure 3.1. The process starts with a cylinder of hydrogen gas. The hydrogen atoms are ionized to obtain protons. These protons are then accelerated in bunches by using a series of accelerators, first a linear accelerator (LINAC), then the proton synchrotron (PS), the super proton synchrotron (SPS) and finally the LHC. The collider itself consists of two rings with a circumference of approximately 26.7 km, where the two counter-rotating proton beams are accelerated to a momentum of 6.5 TeV per beam, leading to a center of mass energy of $\sqrt{s} = 13$ TeV. To maintain the beams along the trajectory, the LHC uses superconducting dipole magnets which are cooled to a temperature below 2 K using superfluid helium. The superconducting magnets produce magnetic fields with a strength of about 8 T. Quadrupole magnets are used to squeeze the beams as they enter the interaction points. The LHC is designed to run with 2808 bunches per beam separated by a 25 ns gap with each bunch con-

taining 100 billion (10^{11}) protons. This translates to a crossing rate of 40 MHz with typicaly 50 collisions per crossing. There are four distinct interaction points where the beams cross and the protons collide. On these sites the main detectors are placed: ALICE (A Large Ion Collider Experiment) [42], LHCb (LHC-beauty) [43], CMS (Compact Muon Solenoid) [44] and ATLAS (A Toroidal Large ApparatuS) [45].

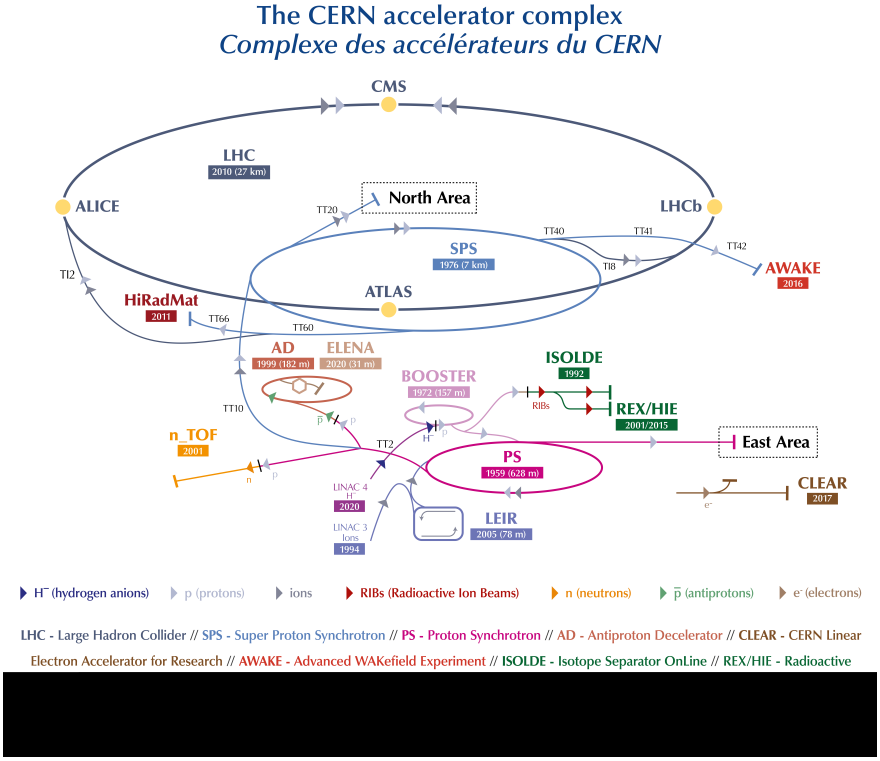


Figure 3.1: Illustration of the CERN accelerator complex. The LHC is the last ring in a complex chain of particle accelerators [46].

The number of events per second generated in the LHC can be described by the equation:

$$N_{event} = L\sigma_{event} \tag{3.1}$$

where σ_{event} is the cross section for a certain process under study and L is the machine luminosity. Given that the luminosity depends only on the beam parameters it can be used

as a measure of the performance of the collider. The full definition for a gaussian beam distribution is:

$$L = \frac{N_b^2 n_b f_{rev} \gamma}{4\pi\epsilon_n \beta} F \tag{3.2}$$

where N_b is the number of particles per bunch, n_b the number of bunches per beam, f_{rev} the revolution frequency, γ is the relativistic gamma factor, ϵ_n the normalized transverse beam emittance (area occupied by the beam), β the beta function (function of the transverse size of the beam) at the collision point, and F the geometric luminosity reduction factor due to the crossing angle at the interaction point. Integrating (with respect to time) the luminosity over the different runs would then give us a measure of the amount of data that was delivered by the LHC. Figure 3.2 shows the total integrated luminosity for Run 2 (2015-2018) of the LHC, as well as the data recorded by the ATLAS detector that was deemed good for physics.

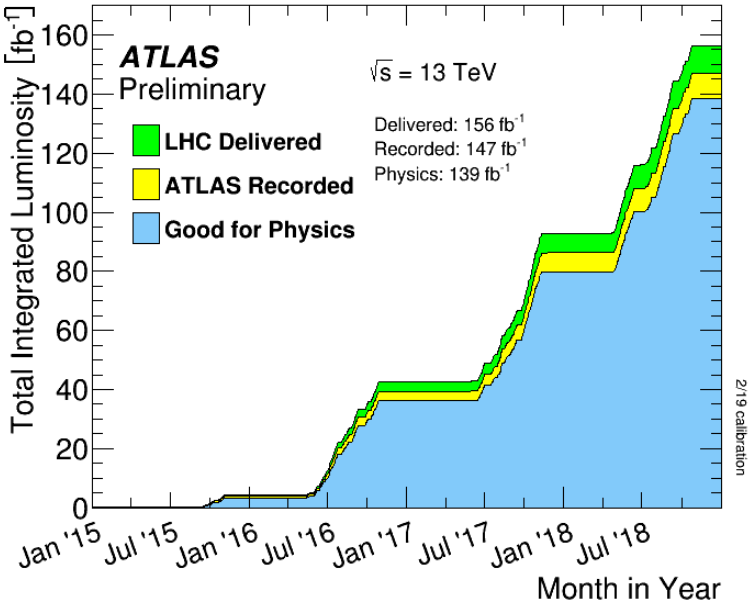


Figure 3.2: Total Integrated Luminosity and Data Quality of the LHC during Run 2 (2015-2018) [47].

3.2 ATLAS Detector

ATLAS [45] is a multi-purpose particle detector of 25 meters in height and 44 meters in length that weighs about 7000 tonnes. It consists of various layers that perform specific measurements of the particles from the collision. ATLAS is located 100 meters below the surface at the CERN LHC Point 1. The detector was designed to have forward-backward symmetry along the beam pipe with a large azimuthal angle coverage. It contains a superconductiong solenoid that surrounds the inner detector, inmersing it in a 2 T solenoid field. ATLAS also contains electromagnetic and hadronic calorimeters that are surrounded by superconducting air-core toroids arranged with an azimuthal symmetry. A muon spectrometer is located within the toroids. Figure 3.3 shows an ATLAS schematic of the different detector modules.

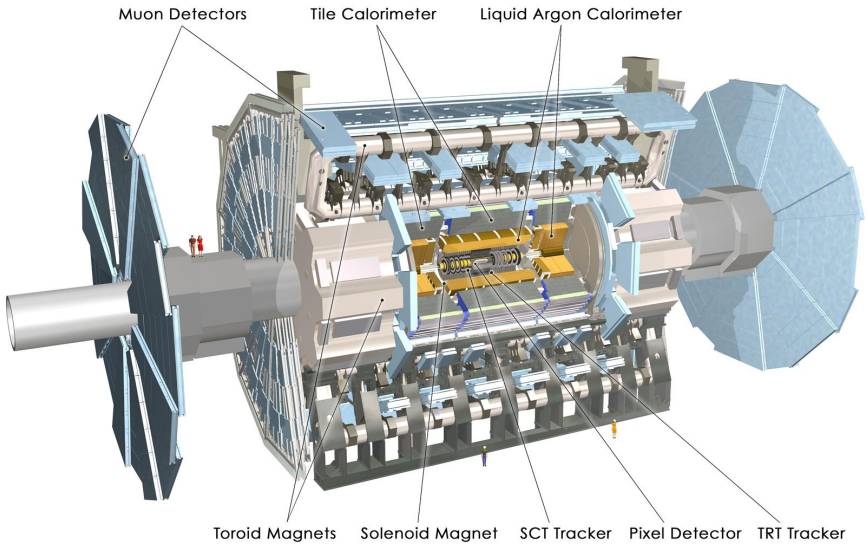


Figure 3.3: Schematic of the ATLAS detector showing its main sub-components. The people in the diagram indicate the scale of the detector [48].

3.3 Coordinate System

To describe the ATLAS detector in detail, we must first describe the conventions regarding the coordinate system used. The clockwise direction of the beam defines the z-axis while the x-y plane lies transversal to the beam direction. The positive x-axis points towards the center of the LHC ring and the positive y-axis points upwards. The azimuthal angle ϕ is defined around the beam axis and the polar angle θ is the angle from the beam axis. Rapidity then is defined as $y = 1/2 \ln (E + p_z)/(E - p_z)$, which for massless particles becomes the pseudorapidity $\eta = -\ln \tan(\theta/2)$. With these quantities the distance in the pseudorapidity-azimuthal angle space can be defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$. Other quantities of interest are the kinematic variables defined on the transverse (x-y) plane, the transverse momentum p_{T}, transverse energy E_{T} and missing transverse energy E_{T}^{miss}.

3.4 Tracking

Because of the large number of particles that emerge from the collision point, the inner detector (ID) must have fine granularity in order to make high precision measurements. It also is designed to provide hermetic and robust pattern recognition. The ID achieves this with its 3 sub-detectors: the pixel detector [49], the silicon microstrip trackers also known as the semiconductor Tracker (SCT) and the transition radiation tracker (TRT). The ID covers the region $\eta < 2.5$, extends to 1.15 m radially and has a length of 6.2 m. It is contained in a solenoid that inmerses its 3 sub-detectors in a 2 T magnetic field which allows charge and momentum measurements. With its track reconstruction capabilities, the ID is the main system used to construct primary and secondary interaction vertices as well as identifying heavy-flavor jets (i.e. b-tagging). Figure 3.4 shows the overall layout of the inner detector.

A schematic view of the overrall path of a charged particle in the inner detector is shown in Figure 3.5.

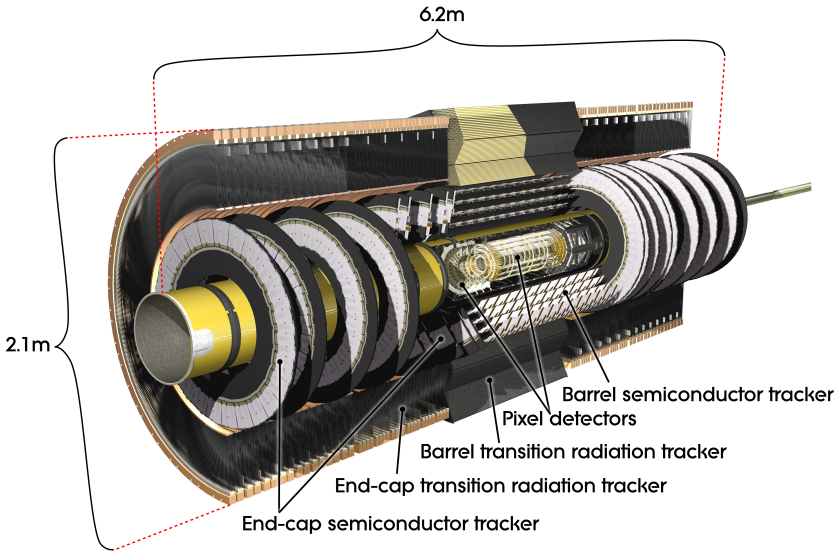


Figure 3.4: Schematic diagram of the ATLAS inner detector showing its different sub-components [50].

Pixel detector

The innermost part of the ID is the pixel detector [49]. The pixel detector provides the highest granularity around the vertex region with a total of 1744 modules. Each module is composed of oxygenated silicon sensors, front-end electronics and flex-hybrids with control circuits. The silicon sensors are the sensitive part of the pixel detector and function as a solid-state ionization chambers. The pixel sensor is an array of bipolar diodes placed on a silicon wafer. The p-n junctions operate under a reverse bias. Ionizing particles passing through the active volume create drifting electron-hole pairs that produce electrical signal that can be measured. The bulk contains oxygen impurities to increase tolerance of the silicon against damage caused by charged hadrons [51].

The modules are arranged in three concentric cylinders around the beam axis and as three disks in the end-cap regions. Three layers allow an effective reconstruction of tracks by requiring a minimum of 3 hits. An Insertable B-Layer (IBL) [52] was installed between the beam pipe and the pixel detector during the 2016 LHC shutdown to maintain robust tracking in the presence of increased pileup and radiation, while also providing improved precision for vertexing and tagging.

Semiconductor Tracker

After the pixel detector we have the semiconductor tracker (SCT), arranged in four concentric cylinders around the beam axis and nine disks in the end-cap regions. Instead of pixels the SCT contains silicon strip sensors. Each module is composed of two sensors glued together. Eight strip layers are crossed by each track. Small-angle stereo strips consisting of two 6.4 cm long daisy-chained sensors measure both coordinates in the barrel region . In the end-cap region, the strips run radially with a set of stereo strips at an angle of 40 mrad. The SCT has a resolution of 16 μm in ϕ and 580 μm in z.

Transition Radiation Tracker

At a larger radius, the straw tubes of the Transition Radiation Tracker (TRT) provide information on particle tracking and identification. The TRT consists of gas-filled (Xe,$C0_2$,O_2) drift tubes with a gold plated tungsten wire inside. In the barrel region, these straws are parallel to the beam axis, while in the end-cap region, they are arranged radially in wheels. In the barrel region the TRT achieves a resolution of 130 μm while in the end-cap region it provides an accuracy of 30-50 μm. Each of the 3 cylidrincal layers contains 32 modules,

and each module is composed of a carbon-fiber laminated shell with an internal array of the straws embedded in a matrix of polypropylene fibers that serve as the transition radiation material.

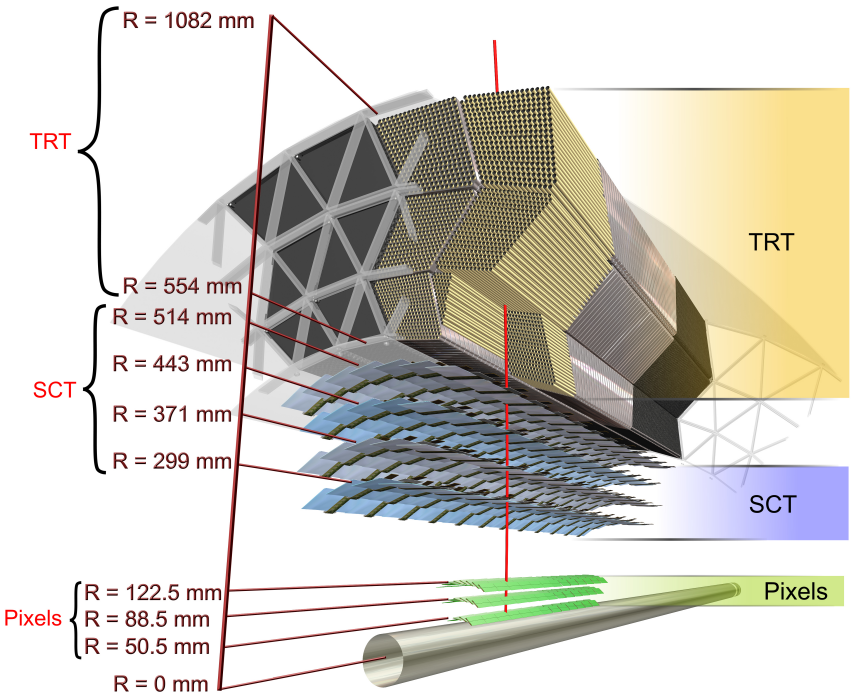


Figure 3.5: Schematic diagram of the structural elements traversed by a charged particle in the barrel inner detector [53].

3.5 Calorimetry

Calorimeters are used to measure the energy of particles. Due to their segmented nature they also can provide directional information about energetic charged leptons and hadrons and even neutral particles that don't interact with the trackers. When particles enter the calorimeter they initiate a particle shower. These lower energy particles are absorbed by the

calorimeters and consequently produce a signal that allows the measurement of the energy deposited. The ATLAS calorimeters cover the range $|\eta| < 4.9$. They can be divided in two main categories: the EM calorimeters and hadronic calorimeters. A diagram of the calorimeter system is shown in Figure 3.6. Both calorimeter systems must provide good containment for electromagnetic and hadronic showers and also limit the punch-through into the muon system. The EM calorimeter has a finer granularity that is suited for precision measurements of electrons and photons. The hadronic calorimeter has coarser granularity and is sufficient to satisfy the physics requirements for jet reconstruction and E_{T}^{miss} measurements.

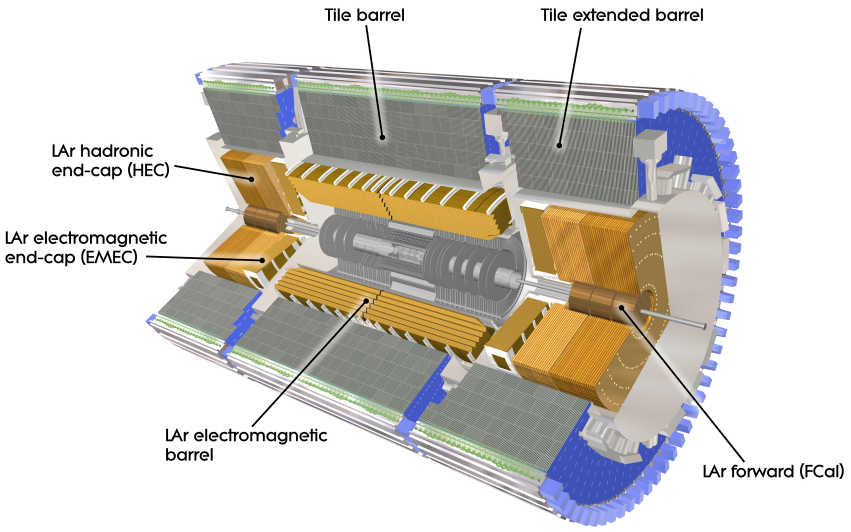


Figure 3.6: Schematic diagram of the ATLAS calorimeter system showing all of its subcomponents [54].

The energy resolution of a sampling calorimeter is parametrized as

$$\frac{\sigma}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{E} \oplus c, \tag{3.3}$$

where a is the stochastic term, b is the noise term and c corresponds to a constant term. The stochastic term represents the random nature of the showering process and is dependent

of the active and absorber materials in the calorimeter as well as the number of layers and their thickness. The noise term describes the electronic noise of the readout system. The constant term reflects local non-uniformities in the response of the calorimeter.

The EM calorimeter was measured [55] to have an energy resolution of

$$\frac{\sigma}{E} = \frac{2.8\%}{\sqrt{E}} \oplus \frac{0.12 \text{ GeV}}{E} \oplus 0.3\%. \tag{3.4}$$

For the hadronic calorimeter the electronic noise was found to be negligible and thus not included. The energy resolution measured [55] was

$$\frac{\sigma}{E} = \frac{52.9\%}{\sqrt{E}} \oplus 5.7\%. \tag{3.5}$$

LAr Electromagnetic Calorimeter

The EM calorimeter has a barrel part ($|\eta| < 1.475$) and two end-cap components ($1.375 < |\eta| < 3.2$). It is a liquid argon (LAr) detector with accordion-shaped electrodes and lead absorber plates. This geometry provides a full ϕ symmetry without azimuthal cracks and lead to an uniform performance in terms of linearity and resolution as a function of ϕ. The absorbers have two stainless-steel sheets glued on either side using a resin-impregnated glass-fiber fabric to provide mechanical strength. The readout electrodes, consisting of conductive copper layers separated by insulating polyminide sheets, are located in the gaps between the absorbers.

The barrel EM calorimeter is composed of two half-barrels, each with a length of 3.2m and a weight of 57 tonnes. One half-barrel consists of 1024 accordion-shaped absorbers interleaved with readout electrodes. For the EM calorimeter, one parameter of interest is

the radiation length X_0, defined as the mean distance a particle can travel before its energy is reduced by a factor of $1/e$. Each half-barrel is divided in 16 modules, each with a total thickness of a minimum of 22 X_0 and cover $\Delta\phi = 22.5°$. These modules have three layers of depth. The front layer is read out at the low-radius side of the electrode while the middle and back layers are read out at the high-radius side of the electrode. A sketch of the different layers of the EM barrel module is provided in Figure 3.7.

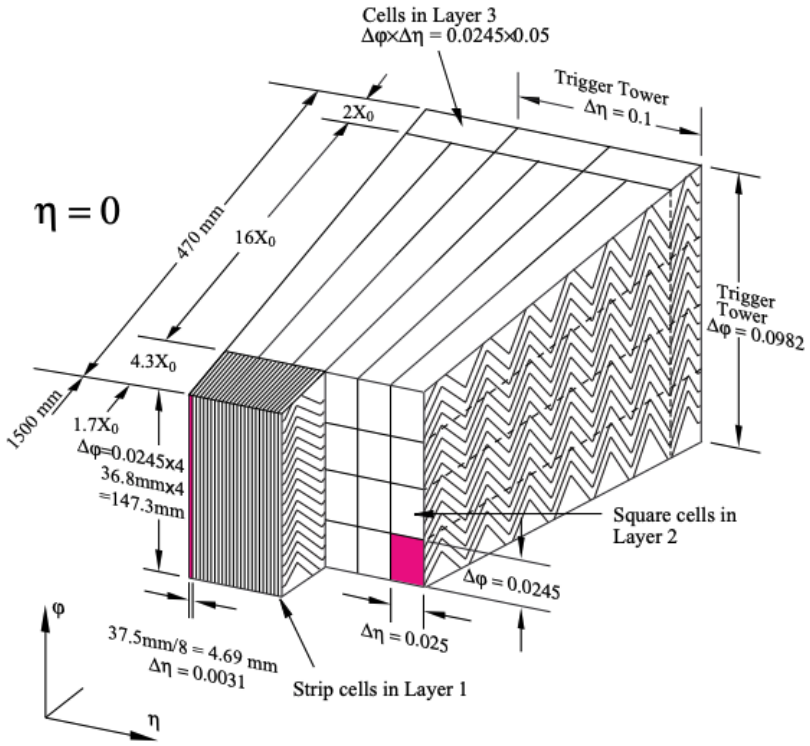


Figure 3.7: Schematic diagram showing the different layers of the EM calorimeter barrel module [45].

Hadronic Calorimeter

The hadronic calorimeter system consists of the tile calorimeter (TileCal), the hadronic end-cap calorimeter (HEC) and the forward calorimeter (FCal).

TileCal is located outside the EM calorimeter envelope. The barrel covers $|\eta| < 1.0$ and the extended barrels cover the range $0.8 < |\eta| < 1.7$. It is a sampling calorimeter that uses scintillating tiles as the active medium and steel as the aborber. The tile calorimeter extendeds radially from an inner radius of 2.28 m to an outer radius of 4.25 m. Azimuthally, TileCal is divided into 64 modules. It is segmented in three longitudinal layers with different interaction lengths. The interaction length λ is defined as the mean free path of a hadronic particle before undergoing an inelastic nuclear interaction. The three segments have 1.5, 4.1 and 1.8 λ for the barrel, and 1.5, 2.6 and 3.3 λ for the extended barrel. The scintillating tiles are read out by wavelength shifting fibers into two photomultiplier tubes (PMT). When an ionising particle crosses the tiles, they induce the production of blue scintillation light that is then converted to green light by the wavelength-shifting fluors in the fibers. A schematic drawing of a TileCal module with its components is shown in Figure 3.8.

The HEC module is a copper/liquid-argon sampling calorimeter with a flat-plate design that covers the range $1.5 < |\eta| < 3.2$. It consists of two cylindrical wheels, each with two longitudinal sections. HEC shares the end-cap cryostats with FCal and the electromagnetic end-cap calorimeter. Each of the HEC wheels is constructed of 32 identical wedge-shaped modules. The modules of the front wheels have 24 copper plates with a thickness of 25 mm. For the rear wheels, the modules are made of 16 copper plates with a thickness of 50 mm making its sampling fraction coarser. Figure 3.9 depicts the HEC module views from different angles. Seven stainless-steel tie-rods provide the structural strength of the modules. Honeycomb sheets are used to fill the space between three electrodes that divide the gaps into four separate LAr drift zones. Each of these drift zones is supplied with a high voltage. The middle electrode serves as the readout electrode and the other two carry surfaces of high resistivity to which high voltage is applied, forming an electrostatic transformer.

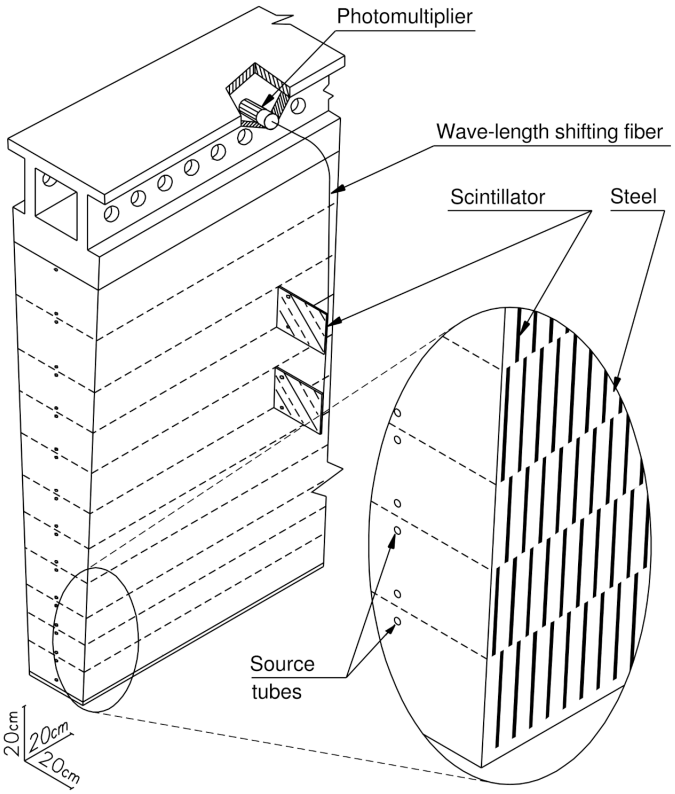


Figure 3.8: Schematic diagram of a TileCal module, showing the slots in the steel for scintillating tiles and the method of light collection by wavelength-shifting fibers to PMTs [56].

The FCal system provides coverage over $3.1 < |\eta| < 4.9$. The FCal modules are located at a distance of 4.7 m from the interaction point and are exposed to high particle fluxes. To avoid ion build-up problems it is designed with very small liquid-argon gaps. These gaps are constructed by using an electrode structure of small-diameter rods centered in tubes that are oriented parallel to the beam direction. Three modules make up the FCal: an electromagnetic module (FCal1) and two hadronic modules (FCal2, FCal3). Figure 3.10 provides a schematic diagram of the FCal modules. FCal1 uses copper as an absorber to optimize resolution and heat removal. FCal2 and FCal2, on the other hand, use mainly tungsten. Extra shielding behind FCal3 is employed to reduce backgrounds in the end-cap muon system.

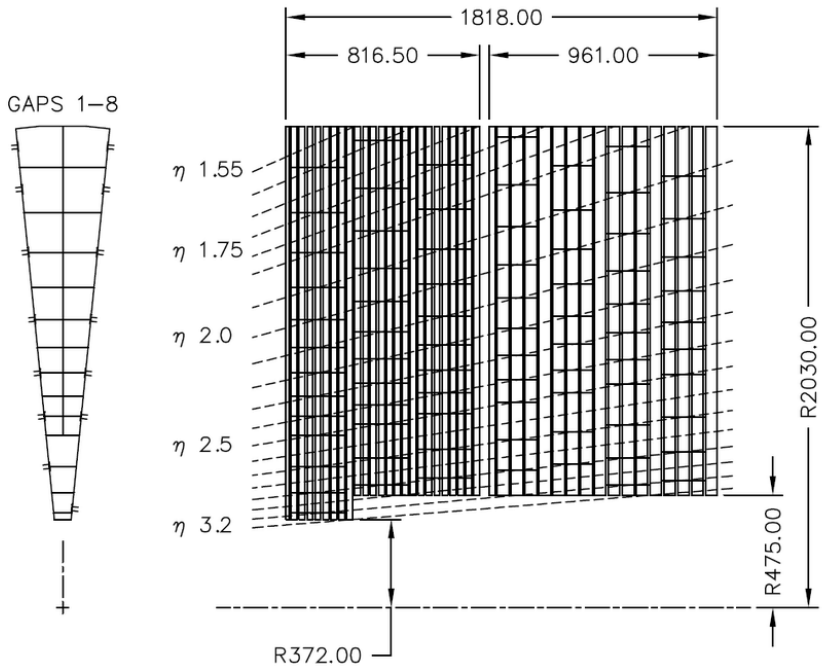


Figure 3.9: Schematic of the R-ϕ (left) and R-z (right) views of the hadronic end-cap calorimeter (HEC) module [57].

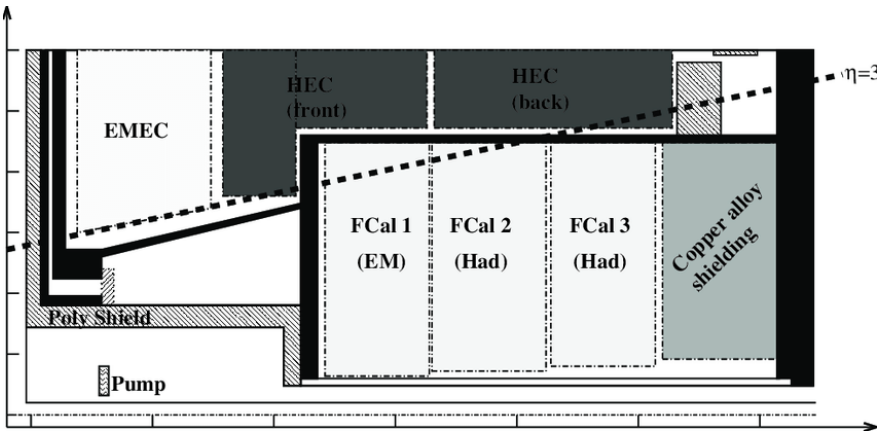


Figure 3.10: Schematic diagram showing the three forward calorimeter (FCal) modules [58].

3.6 Muon system

The outer part of the ATLAS detector is the muon spectrometer. It is designed to detect muons exiting the barrel and end-cap calorimeters in the range $|\eta| < 2.7$. Most muons pass through the inner detector without much interaction and thus it is neccesary to have a dedicated system for them. The muon spectrometer is based on the magnetic deflection of the muon tracks in the large superconducting air-core toroid magnets. Magnetic bending is provided by the large barrel toroid for $|\eta| < 1.4$. Two smaller end-cap magnets inserted at the ends of the barrel toroid provide the track bending for $1.6 < |\eta| < 2.7$. A combinantion of these two fields provide the magnetic deflection in the transisiton region $1.4 < |\eta| < 1.6$. In the barrel region, located between the eight coils of the superconducting barrel toroid magnet, there are eight precision-tracking chambers. In the end-cap, they are in front and behind the two end-cap toroid magnets. Each octant is divided in the azimuthal direction in two sectors (a large and a small sector). The chambers are arranged in three concentric cylindrical shells around the beam axis at a radius of 5, 7.5 and 10 m. In the two end-cap regions, the muon chambers form large wheels that are perpendicular to the z-axis at a distance of $|z| = 7.4$, 10.8, 14 and 21.5 m.

The momentum measurement is performed by the Monitored Drift Tube chambers (MDT) that cover the range $|\eta| < 2.7$. The chambers consist of three to eight layers of drift tubes with an average resolution of 80 μm per tube (35 μm per chamber). In the forward region ($2 < |\eta| < 2.7$), the Cathode-Strip Chambers (CSC) are used due to their higher rate capability and time resolution. The CSCs are multiwire chambers with cathode planes segmented into strips in orthogonal directions. This configurations allows the measurement of both coordinates using the induced-charge distribution. The resolution of these chambers is 40 μm

in the bending plane and 5 mm in the transverse plane.

The muon system also has the capability to trigger on muon tracks. The precision-tracking chambers have a system of fast trigger chambers capable of delivering track information in nanoseconds after the passage of a particle. Resistive Plate Chambers (RPC) and Thin Gap Chambers (TGC) were chosen for this, in the barrel and end-cap respectively. Both chamber types deliver signals with a spread below 25 ns, thus they provide the ability to tag the beam crossing. They also measure both coordinates of the track, one in the bending (η) plane and one in the non-bending (ϕ) plane. Muons can be measured in the inner detector and in the muon system. Figure 3.11 shows the elements of the muon system as they are arranged in the ATLAS detector.

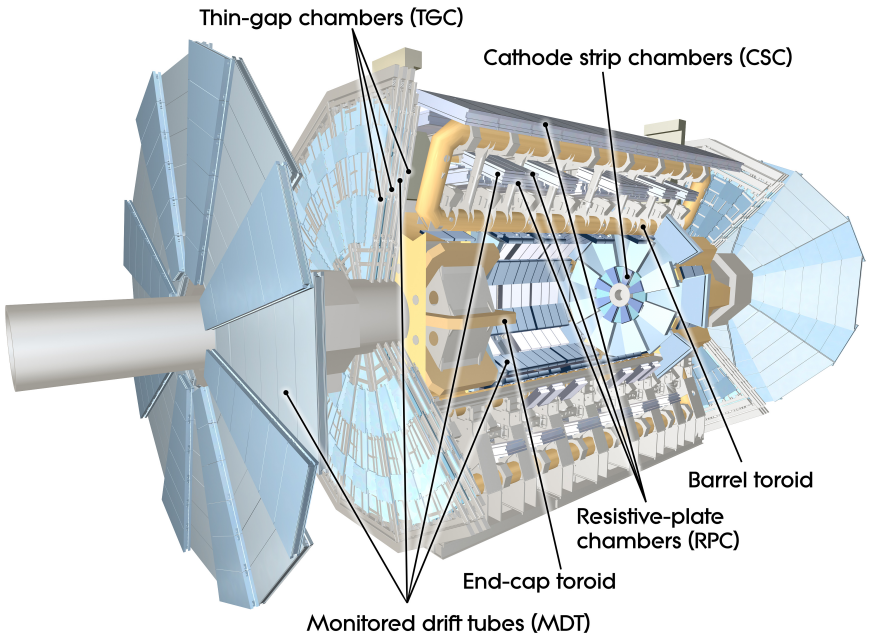


Figure 3.11: Schematic view of the ATLAS muon spectrometer system showing its sub-components [59].

3.7 Magnet system

ATLAS features a unique system of four large superconducting magnets. This system consists of a solenoid, a barrel toroid and two end-cap toroids. Figure 3.12 depicts the magnet system layout in the detector. The powerful magnetic fields produced enable the momentum measurement of electrically charged particles generated in the collisions.

The solenoid is aligned on the beam axis and provides a 2 T magnetic field for the inner detector. It was designed to to keep the material thickness in front of the calorimeter as low as possible. It has an inner diameter of 2.46 m, an outer diameter of 2.56 m and an axial length of 5.8 m. The material used is Al-stabilised NbTi conductor, which achieves a high field with a reduced thickness.

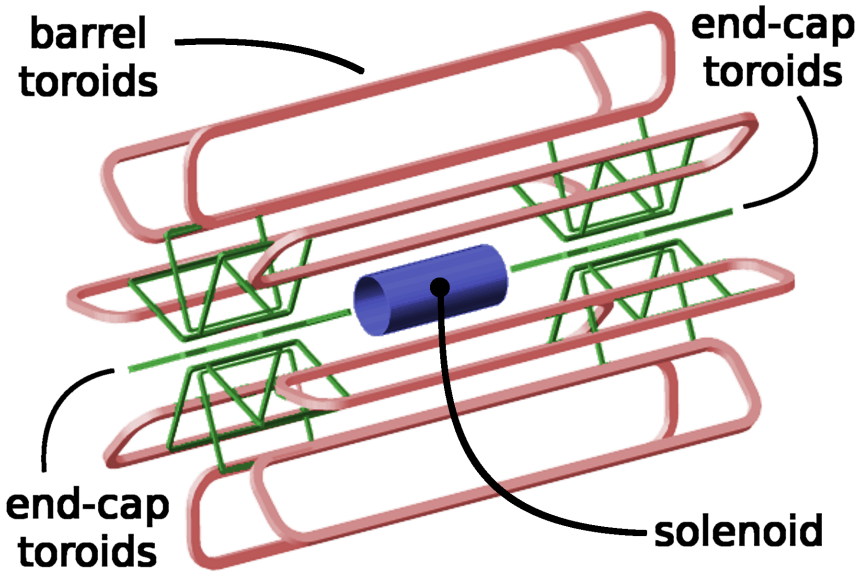


Figure 3.12: Schematic representation of the ATLAS magnet system, showing the central solenoid and the toroids [60].

The barrel toroid system produces a magnetic field that fills the cylindrical volume surrounding the calorimeter and both end-cap toroids. It consists of eight coils encased individually in racetrack-shaped stainless-steel vacuum vessels. It has a length of 25.3 m with

an inner and outer diameters of 9.4 m and 20.1 m respectively. The techology used for the all the toroid system is based on using a conductor of pure Al-stabilised Nb/Ti/Cu reshaped into "pancakes" followed by vacuum impregnation.

3.8 Trigger system

Only a fraction of all the events that ATLAS detects contain interesting and useful information. For this reason a system is needed to ensure a proper selection of events for study. The Trigger and Data Acquisition (TDAQ) systems, the timing and trigger-control logic and the Detector Control System (DCS) achieves this goal [61]. The trigger system [62][63] has three distict levels: L1, L2 and the event filter. Each one refines the decisions the previous trigger made by applying additional selection criteria. The LHC has a collision rate of 40 MHz and ATLAS collects about 60 TB/s of data. The TDAQ and DCS systems reduce the rate of events to the order of 1 kHz and saves to permanent storage around 1.5 GB/s.

The L1 trigger searches for high transverse momentum muons, photons and jets. Its selection is based on information from a subset of detectors. Muons are identified using trigger chambers in the barrel and end-cap regions of the muon spectrometer. Calorimeter selections are based on reduced-granularity information from all the calorimeters. L1 also identifies Regions-of-Interest (RoI) where the detector has identified interesting features. This first level of triggers makes a decision in less than 2.5 μs and reduces the rate from 40 MHz to 75 kHz. Events passing the L1 selection are transferred to the next stages of the detector-specific electronics and to the data acquisition. The L2 selection uses all the available detector data within the RoI's. It is designed to reduce the trigger rate to approximately 3.5 kHz, processing an event in about 40 ms. The final stage is the event

selection carried out by the event filter. The event filter reduces the event rate to about 200 Hz and it is implemented offline, with an average event processing time of four seconds.

Chapter 4

Object Reconstruction

Before we can perform any type of analysis we have to transform the electrical signals recorded by the TDAQ system from particle interactions with the detector to actual physical objects. This thesis is focused in the identification of high momentum $H \to b\bar{b}$ decays. To study this specific process we have to discuss how we reconstruct hadronic jets and how we identify b-hadrons. Muon reconstruction will also be discussed as muons are used in the analysis to perform corrections to the mass spectrum of the Higgs as well as triggers to fill a control region for $t\bar{t}$ events.

4.1 Track and Vertex Reconstruction

Tracking is performed by the inner detector, except for muons where the outer detector may also be involved. Track reconstruction using the ID covers two sequences, a main inside-out track reconstruction and a consecutive outside-in track reconstruction [64]. The pattern recognition sequence (inside-out) starts with the formation of a seed from at least 3 hits in the inner silicon tracker. This is done with the creation of three-dimensinal representations of the silicon detector measurements. From these, track seeds are built. Then, through a window search, using the seed direction, the track candidates are built. Kalman filtering [65] and smoothing are applied to the nearby hits from the detector elements to decide if they are added or rejected to the track candidates. There is a dedicated module for resolving and

cleaning the initial track collection to avoid ambiguity due to the presence of fake tracks or overlapping track segments with shared hits. The ambiguity solving module is based on a scoring algorithm that is optimised for each sub-detector. After this, two modules perform a track extension from the silicon detectors to the TRT. The extension to the TRT improves momentum resolution and particle identification. The final fit of the track is done using a maximum likelihood approach that involves minimizing a global χ^2.

Not all tracks can be found using an inside-out procedure. Some ambiguos hits survive the ambiguity solving process and also tracks coming from secondary decay vertices may not have any silicon hits for the inside-out sequence to proceed. This could occur due to kaons (K_s) decaying or from photon conversions. The outside-in procedure starts with the identification of tracks in the TRT using a Hough transform mechanism [66]. An association tool prevents double counting of hits that have been assigned already to tracks in the inside-out procedure. The TRT segments are then traced back into the silicon detector, which allows one to find small track segments that were missed in the initial inside-out stage. Figure 4.1 provides an example of an event showing the two track reconstruction methods.

The primary vertex is reconstructed by using an iterative vertex finding algorithm [67]. Looking at the reconstructed tracks, vertex seeds are obtained. A χ^2 fit is made using the seeds and nearby tracks. Each possible track gets a weight associated with it which quantifies the compatibility with the fitted vertex. Any track that has a displacement larger than 7σ from the vertex is used to seed a new vertex. The algorithm is iterated until no more vertices are found.

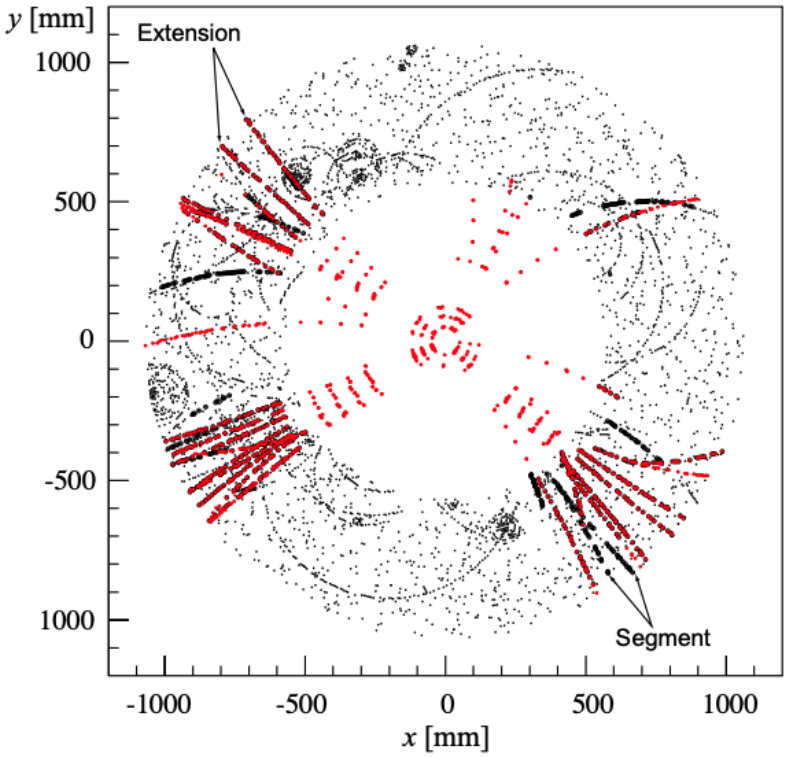


Figure 4.1: Example of an event showing the two possible TRT hit associations. Red shows extensions using the inside-out method and black shows extensions using the outside-in method [64].

4.2 Jets

Jets are a collimated spray of particles coming from a single hard interaction. A jet, in an experimental context, will be detected by its interaction with the different detector modules, creating tracks in the inner detector and depositing energy in the calorimeters. To define a jet, we need a set of rules for grouping these particles and the calorimeter deposits. Jet clustering algorithms are the main way of performing this task. For ATLAS, and this thesis, the clustering is done by using the anti-k_t algorithm [68] an algorithm in the same family as the k_t and Cambridge/Aachen sequential recombination algorithms.

To define the clustering algorithms, we consider two distances: d_{ij} beween entities i and j, and d_{iB} between entity i and the beam B. Entities refer to particles, energy deposits or

pseudo-jets. The distance metrics are defined as

$$d_{ij} = \min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2}, \tag{4.1}$$

$$d_{iB} = k_{ti}^{2p}, \tag{4.2}$$

where $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, k_t is the transverse momentum, y_i is the rapidity, ϕ is the azimuthal angle, R is the jet radius and p is a parameter that governs the relative power of the energy. A value of $p = -1$ results in the anti-k_t algorithm, while p = 1 is the usual k_t algorithm and $p = 0$ corresponds to the Cambridge/Aachen algorithm. The algorithm proceeds by identifying the shortest between the distance measures. If it is d_{ij}, then i and j get combined into one pseudo-jet. If the shortest distance is d_{iB}, then the entity i is classified as a jet and removed from the list. This procedure ends when every entity has been combined and eventually classified as a jet. Soft particles will cluster with hard particles before they cluster among themselves. A hard particle that doesn't have another hard particle close to it, will just accumulate all the soft particles within the radius R, resulting in a conical jet. Figure 4.2 presents how a particular event is clustered into jets with four different jet algorithms.

From these algorithms, only the anti-k_t algorithm is simple, yet Infrared-Collinear (IRC) safe, and soft-resilient in terms of shape. Infrared safety is used to describe algorithms that are robust under the addition of soft radiation. Collinear satefy describes the fact that the result from the algorithm is not changed irregardless if the the particles are collinear (moving together in the same direction). The impact of the underlying event (UE) and pile-up on the momentum resolution for jets is close to zero, which is crucial for high luminosity experiments, like at the LHC. This can be observed by looking at the average jet area at a

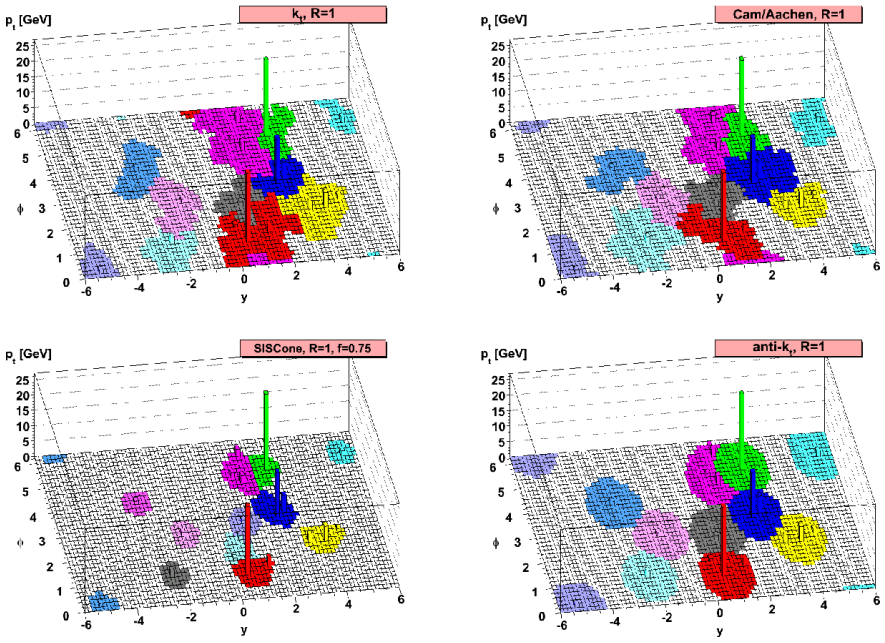


Figure 4.2: Sample parton-level event clustered with four different jet algorithms with a radius parameter value of $R = 1$ [68].

given p_{T} for dijet events clustered using different algorithms when including the underlying event and pile-up. When the ratio of the jet area and πR^2 is calculated as a function of p_{T} using different jet clustering algorithms, only the anti-k_t clustered jets stay close to 1 [68].

Topological Clustering

Jets deposit their energies into the calorimeters. Before appyling the anti-k_t algorithm we must first find the energy clusters deposited in the detector. There are many algorithms that have been used to construct the clusters. The fixed-sized sliding window algorithm [69] was used in the early years of the ATLAS experiment but currently a more complex dynamical topological cell clustering approach is employed [70].

Topological clustering consists of finding topologically connected calorimeter signals due to a specific collision event in an attempt to extract a significant signal from a noisy background. The metric used for the formation of topo-clusters is the cell signal significance σ_{cell},

defined as the ratio of the cell signal E_{cell} to the average noise in the cell $\sigma_{\text{noise,cell}}$:

$$\sigma_{\text{cell}} = \frac{E_{\text{cell}}}{\sigma_{\text{noise,cell}}}. \tag{4.3}$$

Topo-clusters are then formed starting from a calorimeter cell with a highly significant seed signal. Three parameters (S,N,P) control how the algorithm evolves and define signal thresholds for seeding, growth and boundary features of the topological clustering. To begin, proto-cluster seeds from calorimeter cells with $\sigma_{\text{cell}} > S$ are selected. Then all the neighboring cells satisfying $\sigma_{\text{cell}} > N$ around the seeds are added. Finally the neighboring cells with $\sigma_{\text{cell}} > P$ are also added to the cluster. The optimised configuration for ATLAS is: (S=4,N=2,P=0) making the resulting clusters 4-2-0 topo-clusters. Figure 4.3 shows an example of the stages of topo-cluster formation.

Large Radius Jets

When highly boosted massive particles decay, their decay products tend to become collimated, resulting in high levels of overlap between them. For a quasi-collinear splitting [17] into two objects i and j, the total mass is given by $m^2 \simeq p_{\text{T}i} p_{\text{T}j} \Delta R_{ij}^2$. Defining the total momentum $p_{\text{T}} = p_{\text{T}i} + p_{\text{T}j}$ and $z = p_{\text{T}j}/p_{\text{T}}$, then

$$m^2 \simeq z(1-z) p_{\text{T}}^2 \Delta R_{ij}^2. \tag{4.4}$$

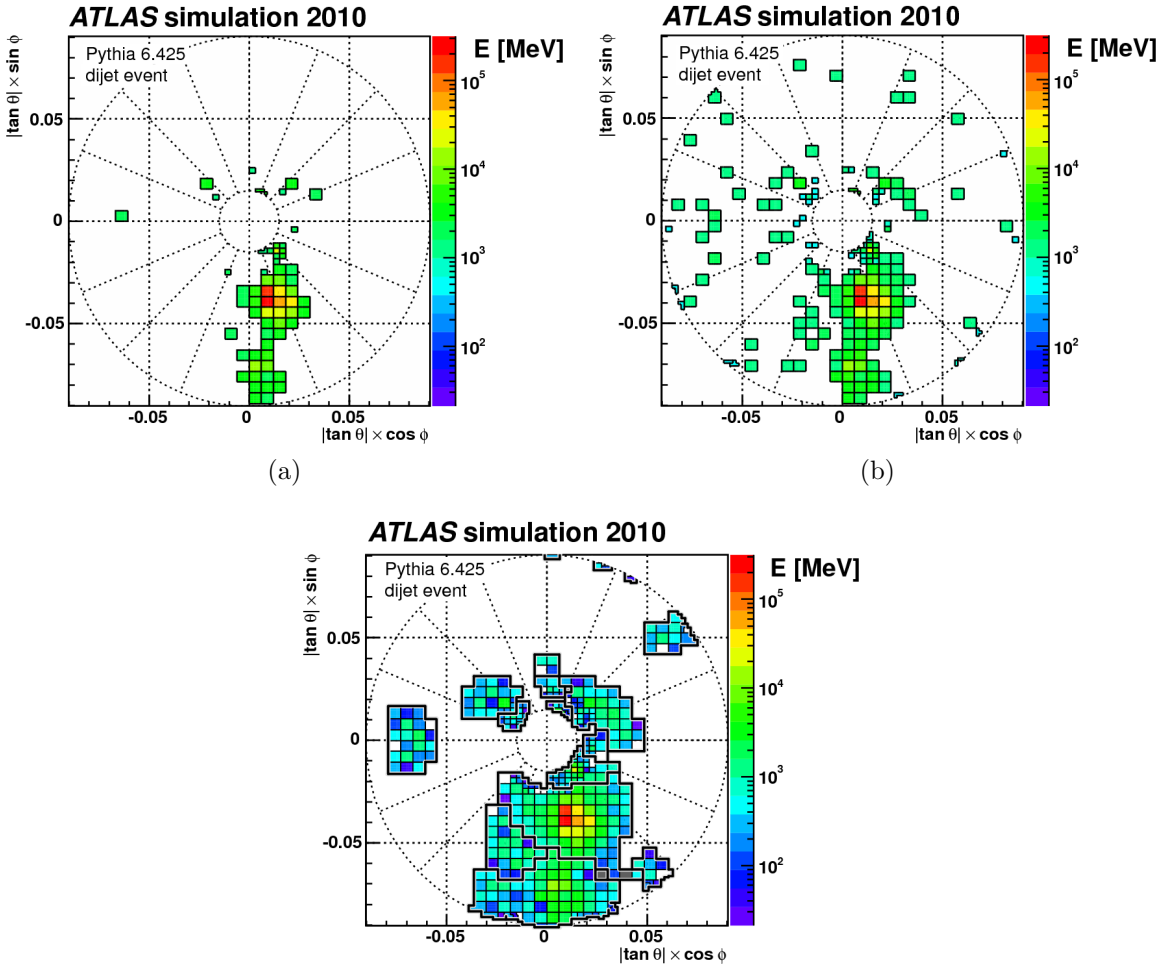


Figure 4.3: Stages of topo-cluster formation in the first module of the FCAL calorimeter for a simulated dijet event. Shown in (a) are the cells with signal significance $\sigma_{\text{cell}} > 4$ that can seed topo-clusters, in (b) cells with $\sigma_{\text{cell}} > 2$ controlling the topo-cluster growth, and in (c) all clustered cells and the outline of topo-clusters in this module [71].

In the case of a Higgs boson decaying to a b-quark pair, the momentum fraction is uniform ($z = 0.5$). Therefore the angular separation of its decay products is approximately

$$\Delta R \simeq \frac{2m_H}{p_{\mathrm{T}}}. \tag{4.5}$$

For massive particles with high p_{T}, the ability to resolve individual hadronic decay products using standard narrow-radius jets begins to degrade. The b-quark pair coming from a Higgs boson at a $p_{\mathrm{T}} \simeq 250$ GeV would be separated by approximately $\Delta R \simeq 1$. Reconstructing these objects in a single large-radius (large-R) jet is advantageous in order to maximize efficiency [72]. Figure 4.4 contains an illustration of the degree of collimation of the decay products of a massive Z' boson when the p_{T} increases.

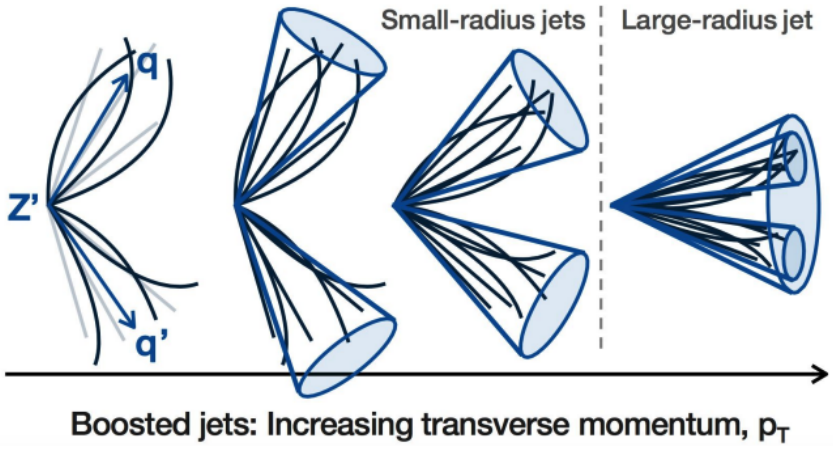


Figure 4.4: Diagram showing the degree of collimation of the decay products of massive particle decaying as p_{T} increases [73].

A single jet containing all the decay products of a massive particle has different properties than a jet originating from a light quark. These large-R jets are rich with multi-pronged substructure, properties that are absent in jets formed from gluons and light quarks. In ATLAS, large-R jets are reconstructed using the anti-k_t clustering algorithm with a radius

parameter $R = 1.0$.

Jet Trimming

When a hard scattering event occurs, the detector records more than just the final states. Initial state radiation (ISR), multiple parton interactions (MPI), underlying event (UE) remnants and pile-up all contribute to the final state. This complicates the jet definition as it is often important to discriminate between these types of energy and the jet of interest. In the case of large-R jets, the subtle substructure differences of jets formed from a massive particle decay products and jets coming from quarks and gluons can be resolved more clearly by removing soft QCD radiation from them [72]. The process of removing soft radiation during the jet reconstruction is referred to as jet grooming.

One of these grooming procedures is known as jet trimming [74]. The trimming algorithm starts by clustering cells into jets with any clustering algorithm, for example, the anti-k_t algorithm, and calling them seed jets. For each seed jet, all of its constituents are then reclustered using another jet algorithm into subjects with a characteristic radius R_{sub}. Subjets from the original seed jet are discarded if they have $p_{\text{T}i} < f_{\text{cut}} \cdot \Lambda_{\text{hard}}$, where f_{cut} is a fixed dimensionless parameter and Λ_{hard} is a hard scale chosen depending on the kinematics of the event. Finally, the remaining subjets are assembled into the new trimmed jet. Figure 4.5 contains a diagramatic representation of how the trimming procedure is performed.

The analysis presented in this document uses trimmed large-R jets with parameters $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 0.05$. The scale Λ_{hard} chosen is the original jet p_{T}, and therefore the subjets with a p_{T} of less than 5% of the original jet p_{T} are removed.

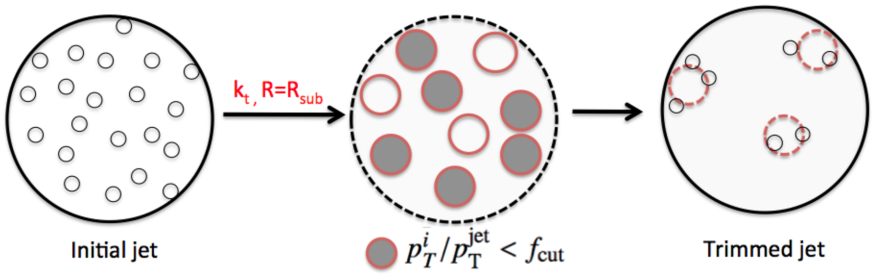


Figure 4.5: Diagram that depicts the jet trimming procedure employed in this analysis [75].

Jet Calibration

Before applying a jet clustering algorithm, cell clusters need calibration to correct for the effects of a non-compensating calorimeter response to hadrons, to accidental signal losses and to energy lost in the inactive material. The calibration strategy is referred to as "local hadronic cell weighting" (LCW) [70]. Topo-clusters calibrated using this method are transformed to be at the LCW scale. After the calibration a large-R jet is defined with the topo-clusters using the anti-k_t algorithm and subsequently trimmed.

The energy, pseudorapidity and mass calibration of the LCW jets are corrected for residual detector effects, using energy and pseudorapidity dependent calibration factors derived from simulation [76]. The correction restores the average reconstructed calorimeter jet energy scale (JES) to that of particle-level jets. These scale factors are applied as multiplicative weights that correct the distributions to the proper scales. The reconstructed large-R jet energy, mass, pseudorapidity and transverse momentum become

$$E_{\text{reco}} = c_{\text{JES}}\, E_0, \quad m_{\text{reco}} = c_{\text{JES}}\, m_0, \tag{4.6}$$

$$\eta_{\text{reco}} = \eta_0 + \Delta\eta, \quad p_{\text{T}}^{\text{reco}} = c_{\text{JES}}\, |\vec{p}_0| \cosh(\eta_0 + \Delta\eta), \tag{4.7}$$

where $E_0, m_0, \eta_0, \vec{p}_0$ are the jet properties before any calibration. The correction factors

c_{JES} and $\Delta\eta$ are smooth functions of the large-R jet kinematics. The JES factor c_{JES} is parametrized by a Gaussian fit of the average jet energy response $R_E = \langle E_{\mathrm{reco}}/E_{\mathrm{truth}} \rangle$.

An extra jet mass scale (JMS) calibration is performed after the energy scale calibration. The correction factor c_{JMS} is applied as a function of E_{reco}, η and $\log(m_{\mathrm{reco}}/E_{\mathrm{reco}})$. The definition of the correction factor is determined using the same procedure as the jet energy calibration but using the jet mass response, $R_m = \langle m_{\mathrm{reco}}/m_{\mathrm{truth}} \rangle$, instead. The reconstructed kinematic variables corrected are then

$$m_{\mathrm{reco}} = c_{\mathrm{JES}}\, c_{\mathrm{JMS}}\, m_0, \quad p_{\mathrm{T}}^{\mathrm{reco}} = c_{\mathrm{JES}} \sqrt{E_0^2 - c_{\mathrm{JMS}}^2 m_0^2} \,/\, \cosh(\eta_0 + \Delta\eta). \tag{4.8}$$

The final step of the calibration is the in situ calibration method [77] to bring data to agreement with MC using response measurements in pp collision data of well known objects, such as dijet events, that work as a reference. Scale factors are derived in the same fashion as the JES and JMS calibration but the response is defined by the ratio of jet properties of data and MC. At the end the groomed jets should be at the proper jet energy scale (JES) and jet mass scale (JMS). Uncertainties are also derived through this process. An overview of all the calibration steps is shown in Figure 4.6.

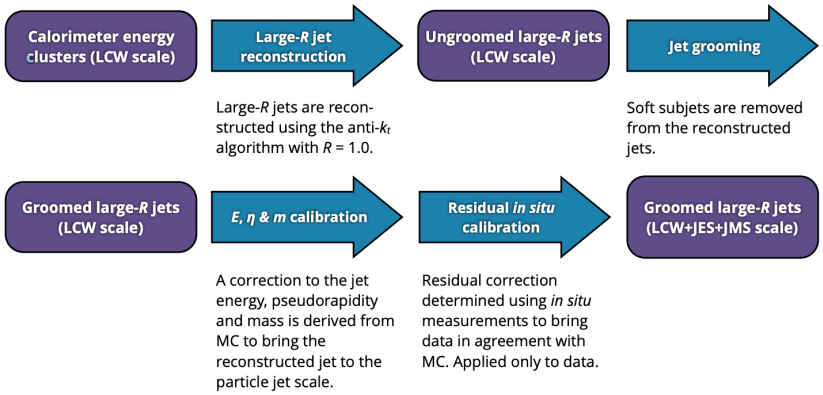


Figure 4.6: Overview of the reconstruction and calibration procedure for large-R jets [77].

Jet Mass

One of the most powerful tools to distinguish jets that contain the decay products of massive particles from the multijet background is its mass. Jet mass is considered one of the most important jet substructure (JSS) variables for large-R jets. The jet mass used in the analysis is referred to as combined jet mass m^{comb} [77]. It is called the combined jet mass because it is a smooth interpolation between two other mass definitions: a calorimeter-based jet mass m^{calo} and track-assisted jet mass m^{TA} [78].

For a large-R jet J with calorimeter-cell cluster constituents i the m^{calo} is defined as:

$$m^{\text{calo}} = \sqrt{\left(\sum_{i \in J} E_i\right)^2 - \left(\sum_{i \in J} \vec{p}_i\right)^2}. \tag{4.9}$$

Given that the angular spread of the decay products of a boosted massive particle scales as $1/p_{\text{T}}$, the spread is comparable with the calorimeter granularity at high values of p_{T}. It is possible to include tracking information to maintain performance at high p_{T}. The track-assisted mass is defined as:

$$m^{\text{TA}} = \frac{p_{\text{T}}^{\text{calo}}}{p_{\text{T}}^{\text{track}}} \times m^{\text{track}} \tag{4.10}$$

where $p_{\text{T}}^{\text{calo}}$ is the transverse momentum of the large-R calorimeter jet, $p_{\text{T}}^{\text{track}}$ is the transverse momentum of the four-vector sum of tracks associated to the large-radius calorimeter jet, and m^{track} is the invariant mass of the four-vector sum. This mass measurement has a better resolution for high-p_{T} jets with low values of m/p_{T}.

The combined jet mass smoothly interpolates between m^{calo} at low p_{T} and m^{TA} at high

p_{T}. A weighted least-squares combination is performed to define m^{comb}:

$$m^{\mathrm{comb}} = w_{\mathrm{calo}}\, m^{\mathrm{calo}} + w_{\mathrm{TA}}\, m^{\mathrm{TA}} \tag{4.11}$$

where the weights are determined by the mass resolutions σ_{calo}, σ_{TA} of the calorimeter and track-assisted measurements. These are derived using the jet mass response distribution in dijet events. They are defined as

$$w_{\mathrm{calo}} = \frac{\sigma_{\mathrm{calo}}^{-2}}{\sigma_{\mathrm{calo}}^{-2} + \sigma_{\mathrm{TA}}^{-2}} \quad w_{\mathrm{TA}} = \frac{\sigma_{\mathrm{TA}}^{-2}}{\sigma_{\mathrm{calo}}^{-2} + \sigma_{\mathrm{TA}}^{-2}} \tag{4.12}$$

with the constraint $w_{\mathrm{calo}} + w_{\mathrm{TA}} = 1$.

Variable Radius Track Jets

Track jets are formed by applying jet clustering on tracks in the inner detector from charged particles originating from the hard scattering vertex. They are crucial for finding b-hadrons and are used in this analysis to integrate b-tagging methods with large-R groomed jets.

Using a fixed radius size approach for reconstructing track jets from highly boosted massive particles presents problems in the identification of more than one charged jet given the high degree of collimation. A variable radius (VR) approach is neccesary to maintain acceptable levels of efficiency.

A VR jet is a jet that has been reconstructed with the use of a p_{T} dependent effective radius R_{eff} [79]. It requires the defnition of a parameter ρ that controls how the radius

changes as a function of p_{T},

$$R_{\mathrm{eff}}(p_{\mathrm{T}}) = \min\left[\frac{\rho}{p_{\mathrm{T}}},\, R_{\mathrm{max}}\right]. \tag{4.13}$$

In this analysis the VR track jets are reconstructed using the anti-k_t algorithm, with $\rho = 30$ GeV and the lower and upper bounds of the track-jet radius being $R_{\mathrm{min}} = 0.02$ and $R_{\mathrm{max}} = 0.4$. The value of R_{min} is dictated by the detector resolution. These provide the optimal performance for high-p_{T} Higgs jets decaying to b-quarks [80]. Figure 4.7 shows the efficiency of subjet double b-labelling of track jets associated with a Higgs jet (in MC) using different track jet clustering algorithms. The efficiency using the standard track jets with $R = 0.2$ degrades sharply for Higgs jets with $p_{\mathrm{T}} > 1$ TeV.

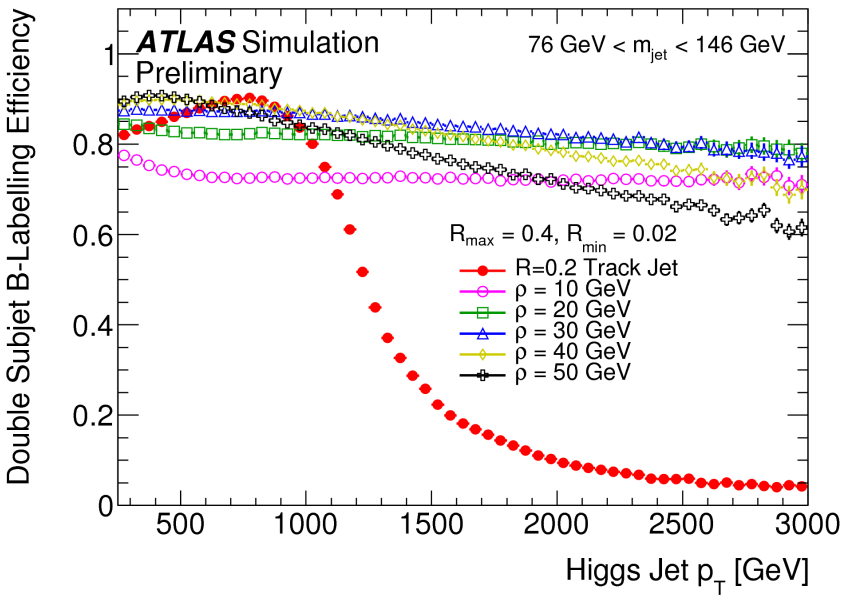


Figure 4.7: Efficiency of subjet double b-labelling at the truth level of a Higgs jet as a function of p_{T} using VR track jets with $R_{\mathrm{min}} = 0.02$ and $R_{\mathrm{max}} = 0.4$ for different values of ρ [80].

Track jets are matched to large-R jets using a process called ghost-association [81]. This procedure consists of treating track jets as infinitely soft particles by setting their p_{T} to 1 eV.

This is done to not affect the reconstruction of the calorimeter jets. The jets are added to the list of inputs of the jet finding algorithm which makes it possible to identify which tracks were clustered in which subjets. This technique facilitates the measurement of the ghost area, the effective area of a jet. Instead of identifying tracks associated with the resulting jets, the number of ghost particles present in the jet after recontruction defines the effective area of that jet.

4.3 b-hadron Identification

ATLAS uses various tagging algorithms to identify b-jets [82]. These are referred to as b-tagging algorithms [83], and they exploit the long lifetime of b-hadrons as well as the properties of the b-quark fragmentation. Measureable b-hadrons have a significant mean flight length in the detector before decaying. This leads to an extra vertex displaced from the hard-scatter collision point. An illustration of a track jet with a displaced secondary vertex is shown in Figure 4.8.

b-tagging

Taggers can be divided into two main categories, low-level taggers and high-level taggers. Low level taggers are traditional track-based impact parameter taggers. Examples of these include IP2D and IP3D, SV1 [85] and JetFitter [86]. They are based on a log-likelihood ratio (LLR) discriminant that separates tracks associated to jets according to their compatibility to the primary vertex. IP2D and IP3D use the transverse impact parameter significance as discriminating variables. The other low-level taggers, SV1 and JetFitter, are secondary vertex-based b-tagging algorithms. All the discriminating variables produced by the low-level

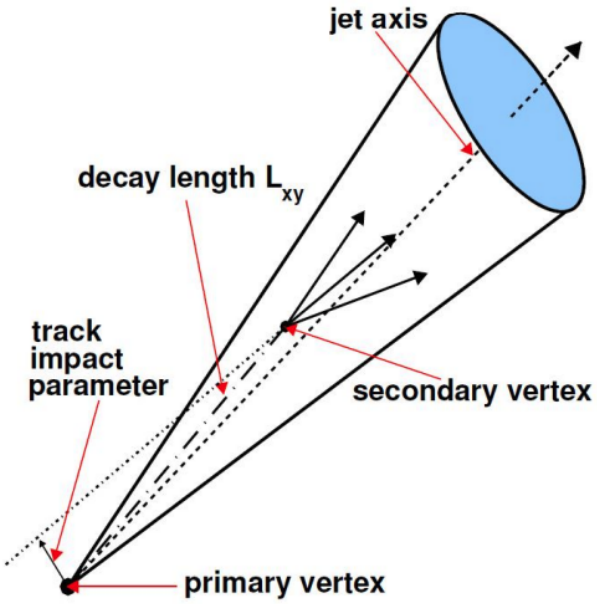
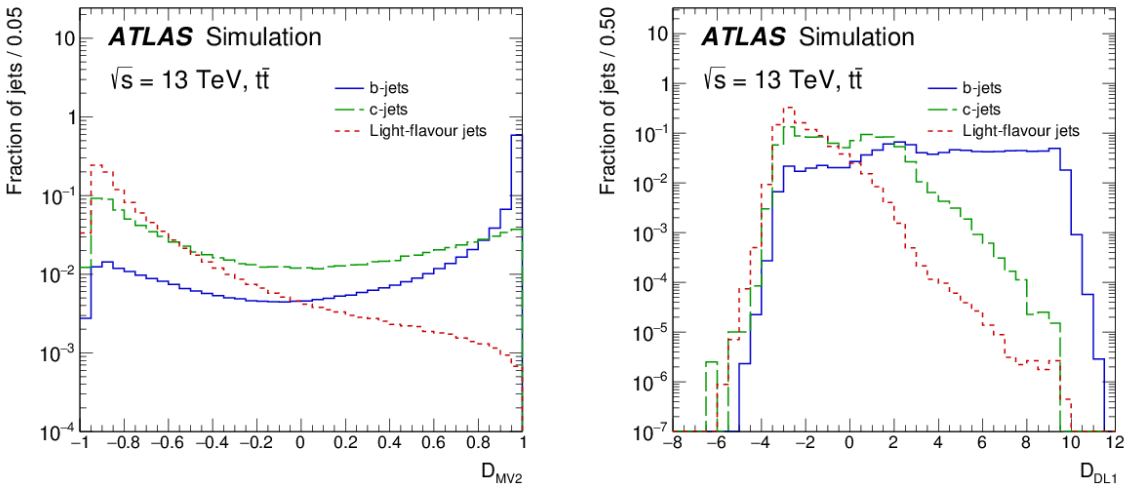


Figure 4.8: Diagram of a track jet with displaced secondary vertex from the primary vertex [84].

taggers are used as inputs for the high-level algorithms. The high-level taggers used in this thesis are MV2 [87] and DL1 [87].

MV2 consists of a boosted decision tree (BDT) algorithm. It is trained using the ROOT Toolkit for Multivariate Data Analysis (TMVA) [88] on a hybrid $t\bar{t} + Z'$ sample. The kinematic properties of the jets are included in the training in order to exploit the correlations with the other input variables. However, for b-jets and c-jets, p_{T} and $|\eta|$ are reweighted to match the spectrum of the *light*-flavor jets.

DL1 is based on an Artificial Deep Neural Network (DNN) trained using Keras [89] with a Theano [90] backend. DL1 has a multidimensional output corresponding to the probabilities for a jet to be a b, c or *light*-flavor jets. Similar to MV2, a reweighting of p_{T}



(a) (b)

Figure 4.9: Distribution of the output discriminant for (a) MV2 and (b) DL1 b-tagging algorithms [83] .

and $|\eta|$ is performed. DL1's final discriminant is defined as:

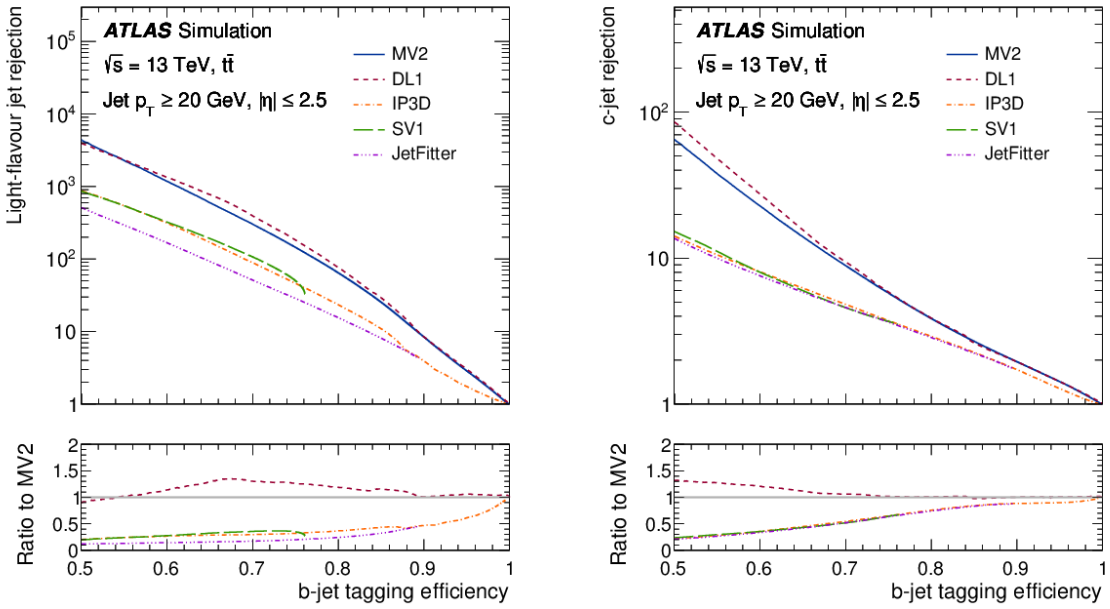
$$D_{\text{DL1}} = \ln\left(\frac{p_b}{f_c \cdot p_c + (1 - f_c) \cdot p_{light}}\right) \tag{4.14}$$

where p_b, p_c, p_{light} and f_c are the b-jet, c-jet, $light$-flavor jet probabilities, and the effective c-jet fraction in the background training sample. Figure 4.9 the values of the MV2 and DL1 discriminant is shown for the different types of jets.

The b-tagging algorithms are calibrated in terms of their efficiency working points (WP) by making a cut on the discriminant values. The b-tagging efficiency is defined as

$$\epsilon_{b\text{tagging}} = \frac{N_{b\text{tagged}\,b\text{jets}}}{N_{\text{total}\,b\text{jets}}}. \tag{4.15}$$

A cut on the corresponding discriminant is done such that the overall efficiency WP at a desired kinematic range stays constant. The WP chosen for the study shown in this



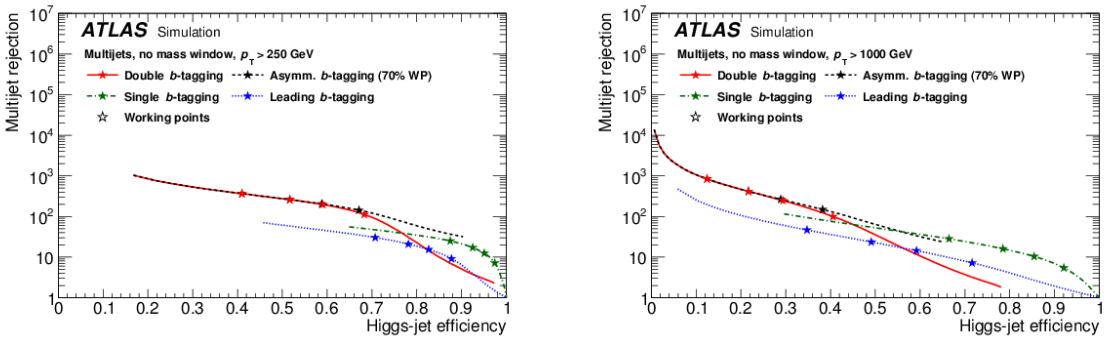
(a) (b)

Figure 4.10: The (a) *light*-flavor jet (b) c-jet rejections versus the b-tagging efficiency [83] .

thesis is the 77% efficiency WP. For every efficiency, there is also a background rejection rate associated with it. The background rejection is calculated as the inverse of the b-mistag rate ($1/\epsilon_{\text{bkg}}$). The *light*-flavor and c-flavor rejections are shown in Figure 4.10 as a function of b-tagging efficiency for multiple b-taggers.

Double *b*-tagging

Given that we are looking for $H \rightarrow b\bar{b}$ we require the identification of 2 b-jets. When tagging 2 b-jets, multiple schemes have been developed [91]. The benchmarks are: double, asymmetric, single and leading single b-tagging. Double b-tagging takes the two highest-p_{T} track jets that pass the b-tagging requirement. For asymmetric b-tagging, the track jet that is more consistent with the interpretation of being a b-jet must pass a fixed efficiency working point, while the b-tagging requirement on the second track jet is varied. In single b-tagging



(a) (b)

Figure 4.11: The multijet rejection as a function of the Higgs tagging efficiency for large-R jets with p_{T} above (a) 250 GeV and above (b) 1000 GeV [91] .

at least one of the two highest-p_{T} track jets must pass the b-tagging requirement, while on leading single b-tagging only the highest-p_{T} track jet must pass the b-tagging requirement.

The scheme adopted by the study performed in this thesis is double b-tagging as it is the scheme with the largest multijet and top-jet rejection across a large range of Higgs efficiencies. This can be seen in Figure 4.11.

4.4 Muons

Muon reconstruction [92] is performed both in the Inner Detector (ID) and in the Muon Spectrometer (MS). This information is then combined to form the muon tracks that are used in the analyses. The Monitored Drift Tubes (MDT) segments are reconstructed with a straight-line fit to the hits of each layer. The Resistive Plate Chambers (RPC) and Thin Gap Chambers (TGC) are used to measure the coordinate orthogonal to the bending plane. For segments in the Cathode Strip Chambers (CSC), a separate combinatorial search in the η, ϕ detector planes is performed. Muon track candidates are then built by combining together all the segments in the different layers.

There are four muon types defined depending on which subdetectors are used in their reconstruction. They are: Combined (CB) muons, Segment-tagger (ST) muons, Calorimeter-tagger (CT) muons and Extrapolated (ME) muons. Combinations of these are used to select and identify muons.

Muon identification selection is divided into four categories: *Medium*, *Loose*, *Tight* and *High-*p_{T}. In this analysis, *Medium* muons are used, the default ATLAS selection. This selection uses only tracks from CB muons and ME muons and it minimizes the systematic uncertainties associated with reconstruction and calibration.

For CB muons, reconstruction follows an outside-in pattern recognition; the muons are first reconstructed in the MS and then are extrapolated inward by matching to an ID track. The independent tracks from the ID and MS are combined using a global fit. ME muon reconstruction is based only on the MS track and a requirement on compatiblity with the interaction point (IP). ME muons are mainly used to extend the acceptance into the region not covered by the ID ($2.5 < \eta < 2.7$).

Muon isolation criteria optimized for different analyses are also defined [92]. Two variables are used to make the isolation cuts, $p_{\mathrm{T}}^{\mathrm{varcone30}}$ and $E_{\mathrm{T}}^{\mathrm{topocone20}}$. The track-based variable, $p_{\mathrm{T}}^{\mathrm{varcone30}}$, is the scalar sum of the transverse momenta of the tracks with $p_{\mathrm{T}} > 1$ GeV in a cone size $\Delta R = \min(10\,\mathrm{GeV}/p_{\mathrm{T}}^{\mu}, 0.3)$ around the muon of transverse momentum p_{T}^{μ}. The calorimeter-based isolation variable, $E_{\mathrm{T}}^{\mathrm{topocone}}$, is the sum of the transverse energy of the topological clusters in a cone size $\Delta R = 0.2$ around the muon direction. The contribution of the muon itself to the energy deposits is subtracted.

The isolation working point (WP) used in this analysis is referred to as the *FixedCutLoose*

WP. The cuts on the discriminating variables are

$$p_{\mathrm{T}}^{\mathrm{varcone30}}/p_{\mathrm{T}}^{\mu} < 0.15 \quad \text{and} \quad E_{\mathrm{T}}^{\mathrm{topocone20}}/p_{\mathrm{T}}^{\mu} < 0.30. \tag{4.16}$$

Figure 4.12 shows the efficiency for the *FixedCutLoose* muon isolation working point as a function of p_{T}.

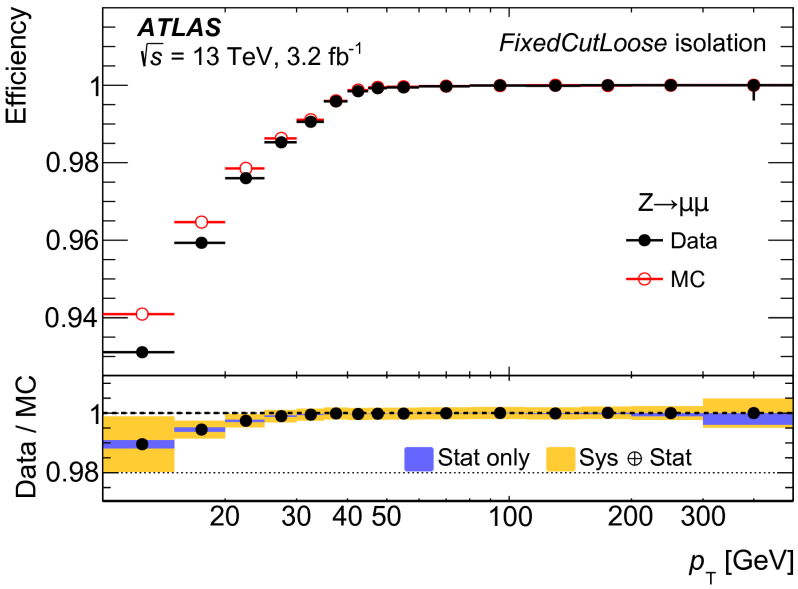


Figure 4.12: Isolation efficiency for the *FixedCutLoose* muon isolation working point [92].

Chapter 5

Boosted $H \to b\bar{b}$ Analysis

5.1 Introduction

After describing how the different physics objects are constructed from collisions in the detectors we can now perform the analysis. The analysis has the goal of extracting information by examining the events from real data gathered by the ATLAS detector using guidance from MC based collision simulations. For this particular analysis, the goal is to measure the Higgs boson production cross section inclusively, in the fiducial volume and differentially (in p_{T}), using its $H \to b\bar{b}$ decay. The measurement of the Higgs boson cross section for $p_{\mathrm{T}} > 1$ TeV is of particular interest and is the first of its kind made by the ATLAS Collaboration.

Before this study, the measurement of the Higgs boson cross section using its $H \to b\bar{b}$ decay exploited the semileptonic decays of vector bosons in the VH production modes by requiring additional triggers involving muons and missing energy [26], or in the VBF production mode by requiring additional jets [93]. In this analysis, we do not impose any restrictions on the production mode and simply require the presence of an extra jet accompanying the Higgs. This selection allows for access to a large Higgs boson p_{T} range using triggering with large-R jets. Without this requirement, the large QCD backgrounds would require single b-jet triggering which limits the mass range for which the two jets can be reconstructed [94], eliminating the possibility of measuring the 125 GeV Higgs boson.

Signal Measurement

The measurements presented in this thesis uses the signal strength μ_H extracted through a binned likelihood fit. The signal strength is defined as the ratio of the observed yield of the signal (N_H) over the Standard Model prediction (N_{SM}):

$$\mu_H = \frac{N_{\text{H}}}{N_{\text{SM}}}. \tag{5.1}$$

The measurement of the signal strength μ_H is then used to calculate the Higgs boson production cross-sections for different phase space regions. The measured cross-section is given by

$$\sigma_H = \frac{N_H}{c \times L}, \tag{5.2}$$

where c is a correction factor, L is the integrated luminosity and $N_H = \mu_H N_{\text{SM}}$ is the yield of signal events. The correction factor is defined as $c = A\epsilon$ where A is the fiducial acceptance and ϵ is the selection efficiency. By this definition, the signal strength can also be defined as the ratio between the measured and the expected signal cross sections $\mu_H = \sigma_H/\sigma_{\text{SM}}$.

Maximum Likelihood Method

Suppose you have a set of random variables $\boldsymbol{x}$ that are distributed following a probability density function (p.d.f.) $f(\boldsymbol{x}; \boldsymbol{\Theta})$ with unknown parameters $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_m)$. By choosing a specific functional form of $f(\boldsymbol{x}; \boldsymbol{\Theta})$ we can use the maximum likelihood method to estimate the values of the parameters given a finite sample of data. The likelihood function is defined as

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\Theta}) \tag{5.3}$$

where x_i are the different outcomes of repeated measurements. Then the estimators for the parameters are those that maximize the likelihood function,

$$\frac{\partial L}{\partial \Theta_i} = 0, \quad i = 1, \ldots, m. \tag{5.4}$$

In our analysis the random variable x_i is the number of events in a specific mass bin given the signal and background models. Therefore $x_i \equiv x_i(\boldsymbol{\mu}, \boldsymbol{\theta})$ where $\boldsymbol{\mu}$ are the set of all the signal strengths for signal or scale factors for the backgrounds and $\boldsymbol{\theta}$ are nuisance parameters used for uncertainty estimation. In each bin, the expected number of events is given by

$$x_i(\boldsymbol{\mu}, \boldsymbol{\theta}) = \sum_{s \in \text{sig}} \mu_s \, x_{i,s}(\boldsymbol{\theta}) + \sum_{b \in \text{bkg}} \mu_b \, x_{i,b}(\boldsymbol{\theta}). \tag{5.5}$$

The quantities $x_{i,s}$ and $x_{i,b}$ are taken from the Asimov datasets constructed using MC mass templates of the resonances, or in the case of the QCD background, a parametric function. When a maximum likelihood fit is performed on an Asimov dataset, the results are the expected signal strengths and their expected uncertainties $\hat{\mu}_H = 1 \pm \hat{\sigma}(\mu_H)$. The hat on the parameters is used to identify them as estimators (parameter values for which the likelihood function is a maximum) of the real parameters. The specifics of the MC templates for the signal and backgrounds are explored in the modeling sections 5.5 and 5.6.

The functional form of the p.d.f. is a Poisson distribution. The Poisson distribution expresses the probability of a given number of events ocurring in a fixed interval. The binned likelihood function therefore takes the following form:

$$L(\boldsymbol{\mu}, \boldsymbol{\theta}) = \prod_{i \in \text{bins}} \frac{(x_i(\boldsymbol{\mu}, \boldsymbol{\theta}))^{N_i}}{N_i!} e^{x_i(\boldsymbol{\mu}, \boldsymbol{\theta})}, \tag{5.6}$$

where N_i is the number of data events in bin i.

The nuisance parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ encode the dependence on systematic uncertainties. The prior knowledge on these parameters is used to make the displacement from the nominal value disfavored. They enter the equation as multiplicative factors in the likelihood function and are Gaussian distributions centered at 0 and a standard deviation of 1:

$$g(\theta_m) = \frac{1}{\sqrt{2\pi}} e^{-\theta_m^2/2}. \tag{5.7}$$

Confidence Levels

To test the compatibility of a background only model with the data or derive confidence limits on our observations we use a test statistic. According to the Neyman-Pearson lemma, we can consider the function constructed from a ratio of likelihood functions for a set of parameters that describe the null hypothesis H_0 and a set of parameters that describe the alternate hypothesis H_1 as the best hypothesis discriminator

$$q = -2 \ln \frac{L(H_0)}{L(H_1)}. \tag{5.8}$$

This reduces to a χ^2 distribution in the large sample limit. The fit result is obtained by maximizing the log-likelihood function with respect to all the parameters. For a discovery, the null hypothesis H_0 is the likelihood function where $\boldsymbol{\mu} = 0$ (no signal). More specifically in our case, where the observed signal is small, we can build confidence limits using the CL_s [95] method based on the test statistic

$$q_\mu = -2 \ln \frac{L(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}))}{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})}, \tag{5.9}$$

where $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}$ are the parameters that maximize the constrained ($0 \leq \hat{\mu} \leq \mu$) likelihood and $\hat{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu})$ are the nuisance parameters which maximize the likelihood function for a given value of $\boldsymbol{\mu}$. The variance of $\hat{\mu}$, can be directly calculated using the test statistic

$$\sigma^2_{\hat{\mu}} = \frac{(\mu - \hat{\mu})^2}{q_\mu}. \tag{5.10}$$

The p-value of an hypothesized μ and the corresponding significance is given by

$$p_\mu = 1 - \Phi(\sqrt{q_\mu}) \quad \text{and} \quad Z_\mu = \sqrt{q_\mu}, \tag{5.11}$$

where Φ is the cumulative distribution function for the standard Gaussian. Finally the upper limit of an estimator at a $1 - \alpha$ confidence level is given by

$$\mu_{\text{up}} = \hat{\mu} + \sigma_{\hat{\mu}} \Phi^{-1}(1 - \alpha). \tag{5.12}$$

Therefore, for a 95% upper limit, α would be 0.05.

5.2 Samples

Data

The data used was collected by the ATLAS detector at $\sqrt{s} = 13$ TeV during Run 2 (2015-2018) of the LHC. All the events come from the Good Runs Lists (GRL) [96]. The events in the GRL have collisions with bad detector performance removed. Table 5.1 lists all the GRL datasets used.

The GRL list of events sums up to an ingrated luminosity of 139 fb^{-1} [97]. This quantity

Table 5.1: Summary of the Good Run List (GRL) datasets used in this analysis. The data was collected during Run 2 of the LHC.

Year	Dataset Name
2015	data15_13TeV.periodAllYear_DetStatus-v89-pro21-02_Unknow_PHYS_StandardGRL_All_Good_25ns.xml
2016	data16_13TeV.periodAllYear_DetStatus-v89-pro21-01_DQDefects-00-02-04_PHYS_StandardGRL_All_Good_25ns.xml
2017	data17_13TeV.periodAllYear_DetStatus-v99-pro22-01_Unknown_PHYS_StandardGRL_All_Good_25ns_Triggerno17e33prim.xml
2018	data18_13TeV.periodAllYear_DetStatus-v102-pro22-04_Unknown_PHYS_StandardGRL_All_Good_25ns_Triggerno17e33prim.xml

is less than the total integrated luminosity (156 fb^{-1}) that was delivered by the LHC due to the quality requirements and trigger efficiencies. Table 5.2 summarizes the integrated luminosity and its uncertainties for each year of data taking period included in the Run 2 dataset. The uncertainty sources can be correlated between all years, correlated between a subset of years or completely uncorrelated. Therefore the relative error on the total is not a weighted sum of the relative errors on the individual years. It considers the covariance matrix of the absolute luminosity uncertainties for the differrent years where correlated sources are represented by terms with non-zero off-diagonal entries.

Table 5.2: Summary of the integrated luminosities and uncertainties for the Run 2 *pp* data sample at $\sqrt{s} = 13$ TeV [97].

Data sample	Luminosity	Uncertainty
2015-2016	36.2 fb^{-1}	2.1 %
2017	44.3 fb^{-1}	2.4 %
2018	58.5 fb^{-1}	2.0 %
Total	139.0 fb^{-1}	1.7 %

In this analysis, the data amounts to an integrated luminosity of 136 fb^{-1} due to the specific Large-R jet and muon trigger requirements. Each event satisfies a trigger that requires a large-R ($R = 1.0$) jet reconstructed using the anti-k_t algorithm. For each year of data taking, the jet p_{T} and mass thresholds for triggers differ due to changes in luminosity profiles, inclusion of new techniques [98] and generally different beam conditions. The jet p_{T} thresholds go from 360 to 460 GeV and the mass trigger is either 0, 30 or 35 GeV. The

additional muon triggering is used to fill a control region for top quarks that requires a muon with $p_{\mathrm{T}} > 50$ GeV [99]. Table 5.3 summarizes the triggers used in this analysis. Plots for the efficiency curves for the triggers used in this analysis are presented in Appendix B.

Table 5.3: Summary of the triggers used in this analysis. Large-R jet triggers are used to identify candidate jets, while the muon trigger is used to fill the $t\bar{t}$ control region.

Year	Trigger	Treshhold	Luminosity [fb^{-1}]
Large-R jet Triggers			
2015	HLT_j360_a10_lcw_sub_L1J100	$p_{\mathrm{T}} > 360$ GeV	3.2
2016	HLT_j420_a10_lcw_L1J100	$p_{\mathrm{T}} > 420$ GeV	33.0
2017	HLT_j390_a10t_lcw_jes_30smcINF_L1J100	$p_{\mathrm{T}} > 390$ GeV, $m_J > 30$ GeV	41.0
	HLT_j440_a10t_lcw_jes_L1J100	$p_{\mathrm{T}} > 440$ GeV	41.2
2018	HLT_j420_a10t_lcw_jes_35smcINF_L1J100	$p_{\mathrm{T}} > 420$ GeV, $m_J > 35$ GeV	58.5
	HLT_420_a10t_lcw_jes_35smcINF_L1SC111	$p_{\mathrm{T}} > 420$ GeV, $m_J > 35$ GeV	55.4
	HLT_460_a10t_lcw_jes_LJ100	$p_{\mathrm{T}} > 460$ GeV	58.5
Muon triggers			
All years	HLT_mu50_L1_MU20	$p_{\mathrm{T}}^{\mu} > 50$ GeV	139 fb^{-1}

Simulated Samples

Monte Carlo (MC) programs are used to simulate events which then are used to model the signal and backgrounds pertinent to this analysis. The signal consists of the Higgs production processes ggF, VBF, VH and $t\bar{t}H$. The background samples include W+jets, Z+jets, top quark production and dijets events. Table 5.4 has a summary of all the simulated samples used in the analysis.

Signal Samples

Higgs production through ggF is simulated at NLO QCD accuracy including, the finite top mass effects with the Hj-MINLO (Multi-Scale Improved NLO) [100] prescription using

POWHEG BOX v2 [101] [102]. In a similar manner, NLO accuracy in QCD is achieved for VBF and $t\bar{t}H$ production [103] [104]. $gg \to VH$ production at LO accuracy is calculated using POWHEG BOX v2 as calculations at NLO required new developments [105] not published at the time this analysis was performed. $qq \to VH$ [106] is calculated at NLO accuracy using GoSam [107]. Electroweak (EW) NLO corrections are also applied as a function of the Higgs boson momentum for the VBF, VH and $t\bar{t}H$ production modes using HAWK [108] (see Appendix B). Finally the branching ratios are calculated using HDECAY [109] and PROPHECY4F [110].

Background Samples

Vector boson + jets was simulated using Sherpa [111] with NLO QCD accuracy. NLO EW approximate corrections were applied which reduced the predicted yield by $10 - 20\%$. The NNLOJET group provided NNLO QCD custom corrections as a function of the generated vector-boson momentum (p_T^V) that are then applied on top of the NLO EW corrections.

Every top quark production mode was modeled using POWHEG BOX v2 at NLO QCD. Top quark pair production, tW and single-top t-channel and s-channel are all included [112] [113].

QCD multijet events were modeled using a parametric model. The MC used to study the model was generated by PYTHIA 8.230 [114].

After generation, hadronization and showering every event is put through an ATLAS detector simulation that is based on GEANT4 [41]. Pileup and multi-particle interactions were also modeled using PYTHIA 8.186 with the A3 tuning [115] and were also fed through the same ATLAS detector simulation.

Table 5.4: Summary of the simulated samples for the signal and background processes [10].

Process	ME generator	ME PDF	PS and hadronization	UE model tune	Cross-section order
Higgs Boson					
$gg \to H \to b\bar{b}$	Powheg Box v2 + MINLO	NNPDF3.0$_{NNLO}$	Pythia 8.212	AZNLO	NLO(QCD) + LO(EW)
$qq \to H \to q'q'b\bar{b}$	Powheg Box v2	NNPDF3.0$_{NLO}$	Pythia 8.230	AZNLO	NLO(QCD) + NLO(EW)
$qq \to WH \to q'q'b\bar{b}$	Powheg Box v2 + MINLO + GoSam	NNPDF3.0$_{NLO}$	Pythia 8.240	AZNLO	NNLO(QCD) + NLO(EW)
$qq \to WH \to l\nu b\bar{b}$	Powheg Box v2 + MINLO + GoSam	NNPDF3.0$_{NLO}$	Pythia8.212	AZNLO	NNLO(QCD) + NLO(EW)
$qq \to ZH \to q\bar{q}b\bar{b}$	Powheg Box v2 + MINLO + GoSam	NNPDF3.0$_{NLO}$	Pythia 8.240	AZNLO	NNLO(QCD) + NLO(EW)
$qq \to ZH \to \nu\nu b\bar{b}$	Powheg Box v2 + MINLO + GoSam	NNPDF3.0$_{NLO}$	Pythia 8.212	AZNLO	NNLO(QCD) + NLO(EW)
$qq \to ZH \to llb\bar{b}$	PowhegBox v2 + MINLO + GoSam	NNPDF3.0$_{NLO}$	Pythia 8.212	AZNLO	NNLO(QCD) + NLO(EW)
$gg \to ZH \to q\bar{q}b\bar{b}$	Powheg Box v2	NNPDF3.0$_{NLO}$	Pythia 8.240	AZNLO	LO +NLL(QCD)
$gg \to ZH \to \nu\nu b\bar{b}$	Powheg Box v2	NNPDF3.0$_{NLO}$	Pythia 8.212	AZNLO	LO +NLL(QCD)
$gg \to ZH \to llb\bar{b}$	Powheg Box v2	NNPDF3.0$_{NLO}$	Pythia 8.212	AZNLO	LO +NLL(QCD)
$gg \to t\bar{t}H \to all$	Powheg Box v2	NNPDF3.0$_{NLO}$	Pythia 8.230	AZNLO	NLO(QCD) + NLO(EW)
$gg \to t\bar{t}H \to all$	Powheg Box v2	NNPDF3.0$_{NLO}$	Pythia 8.230	AZNLO	NLO(QCD) + NLO(EW)
Vector Boson + jets					
$W \to q\bar{q}$	Sherpa 2.2.8	NNPDF3.0$_{NNLO}$	Sherpa 2.2.8	Default	NNLO(QCD) + approx NLO(EW)
$Z \to q\bar{q}$	Sherpa 2.2.8	NNPDF3.0$_{NNLO}$	Sherpa 2.2.8	Default	NNLO(QCD) + approx NLO(EW)
Top quark					
$t\bar{t}$ →all	Powheg Box v2	NNPDF3.0$_{NLO}$	Pythia 8.230	A14	NNLO + NNLL
tW	Powheg Box v2	NNPDF3.0$_{NLO}$	Pythia 8.230	A14	NLO
t t-channel	Powheg Box v2	NNPDF3.0$_{NLO}$	Pythia 8.230	A14	NLO
t s-channel	Powheg Box v2	NNPDF3.0$_{NLO}$	Pythia 8.230	A14	NLO
Multijet					
Dijets	Pythia 8.230	NNPDF2.3.$_{LO}$	Pythia 8.230	A14	LO

5.3 Object Definition

Object Reconstruction

A Lorentz boosted Higgs boson event has a topology of the form: $pp \to H(\to b\bar{b}) + j$. Therefore the events of interest are better described by two large-R ($R = 1.0$) jets in which one of them contains the decay products of two b-hadrons. The large-R jets are defined by applying the anti-k_t algorithm, using the software package Fastjet [116], to topological clusters of calorimeter energy deposits. A jet trimming procedure is employed with parameters $R = 0.2$ and $p_{\mathrm{T}}^{\mathrm{subjet}}/p_{\mathrm{T}}^{\mathrm{jet}} < 0.05$. The jet mass m_J is defined as the combined mass, a weighted combination of the calorimeter based mass and the track-assisted jet mass.

Variable radius (VR) track jets are formed using the anti-k_t algorithm with parameters $R_{\mathrm{eff}} = \rho/p_{\mathrm{T}}$ where $\rho = 30$ GeV and have an upper bound of $R_{\mathrm{max}} = 0.4$. Ghost association is used to match large-R jets (before trimming) to VR track jets. For simulation events,

track jets are labeled as having b,c or *light* (u,d,c and g) flavor by truth matching hadrons with $p_{\mathrm{T}} > 5$ GeV within $\Delta R = 0.3$ of the jet axis [87].

The b-tagger MV2 is used to tag VR track jets containing a b-hadron decay. Track jets must have $p_T > 10$ GeV and $|\eta| < 2.5$. At least two track jets per event are considered. The working point is tuned to have an average b-tagging efficiency of 77% for b-jets in simulated $t\bar{t}$ events. The misidentification efficiencies are 0.9% for *light*-jets and 25% for c-jets. To prevent overlap between VR track jets, if the ΔR between any two track jets with $p_{\mathrm{T}} > 5$ GeV associated with a large-R jet [91] is less than their respective radii, then the jet is not considered for b-tagging.

Muons satisfy the *Medium* quality criterion. Muons have to satisfy $|\eta| < 2.5$ and $p_{\mathrm{T}}^{\mu} > 10$ GeV. Isolated muons also have to satisfy loose track and calorimeter based isolation conditions [92].

For a more detailed view of the object reconstruction, algorithms, calibrations and references used refer to Chapter 4.

Analysis Object Definitions

Reconstructed jets that have the properties compatible with a $H \to b\bar{b}$ decay are labeled *candidate jets*. Theoretically the Higgs boson and the hadronic recoil system both have the same p_{T}. In practice, the reconstructed jet p_{T} is affected by final state radiation, jet resolution and any other activity outside the jet cone like pile-up. From simulation it is estimated that roughly 50% of the Higgs jets are leading jets (jet with the highest p_{T}) and 47% are subleading jets (jet with the second largest p_{T}). Figure 5.1 illustrates this phenomenon using simulated ggF events. For this reason, candidate jets are defined as either the leading or sub-leading jet that has a $p_{\mathrm{T}} > 250$ GeV, has $|\eta| < 2.0$ and satisfies the

boosted condition: $2m_J/p_{\mathrm{T}} < 1$. Each candidate jet must contain at least two track jets. The candidate jets are classified as *double-tagged* if its two leading track jets are b-tagged or *anti-tagged* if neither of them are b-tagged.

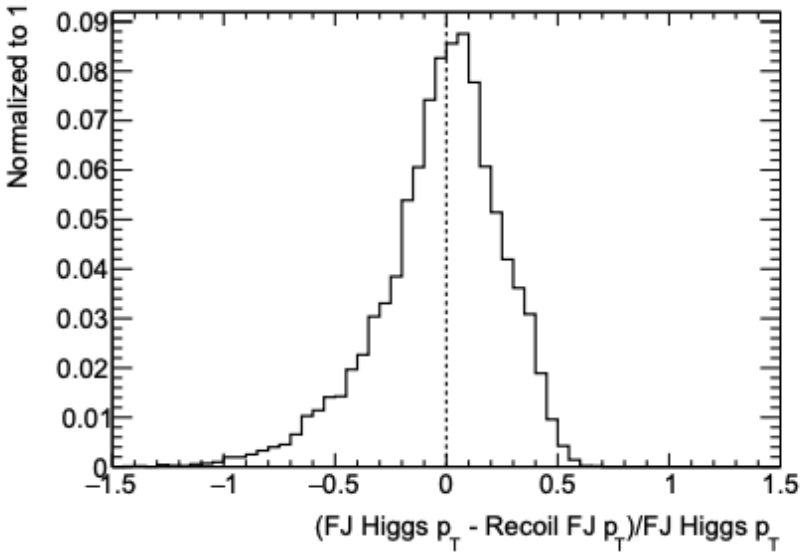


Figure 5.1: Difference between the p_{T} of the Higgs matched large-R jet and the recoil jet. Higgs jets are leading p_{T} if the difference is positive (right of the dashed line) or subleading if the difference is negative (left of the dashed line).

The presence of semileptonic b-hadron decays motivates the application of a correction to candidate jets. The 'muon in jet' correction uses the leading-p_{T} muon found within $\Delta R = \min(0.4, 0.4 + 10/p_{\mathrm{T}}^{\mu})$ of a b-tagged VR track jet. The correction consists of removing the energy deposited by the muon in the calorimeter and adding its four-momenta to the trimmed large-R jet. The corrrection is of the order of 13% for leading Higgs jets and 33% for subleading Higgs jets in simulated ggF events. The m_J width is reduced by 5% and 12% respectively. The uncorrected momentum and mass in candidate jets are denoted as p_{T}^0 and m_J^0 respectively.

5.4 Event Selection

Events are classified into three separate regions: a signal region (SR), a control region ($CR_{t\bar{t}}$) and a validation region (VR). The SR is used to calculate the signal strength. The $CR_{t\bar{t}}$ is used as a control region to study top quark events. The VR is used to test the multijet and V+jets models.

For both the SR and the VR at least one jet with $p_{\mathrm{T}}^{0} > 450$ GeV and $m_{J}^{0} > 60$ GeV is required. The second jet required has to have a $p_{\mathrm{T}}^{0} > 200$ GeV. From these two jets at least one of them have to satisfy the candidate jet criteria. The SR and VR are subdivided into subregions according to the p_{T} ordering of the candidate jets. Figure 5.2 summarizes the regions used for event categorization in the analysis.

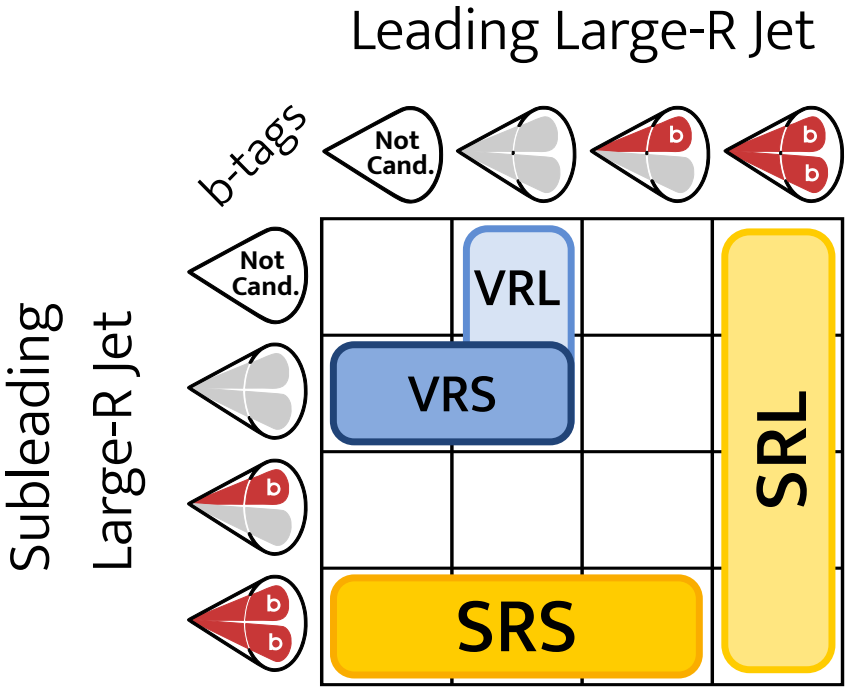
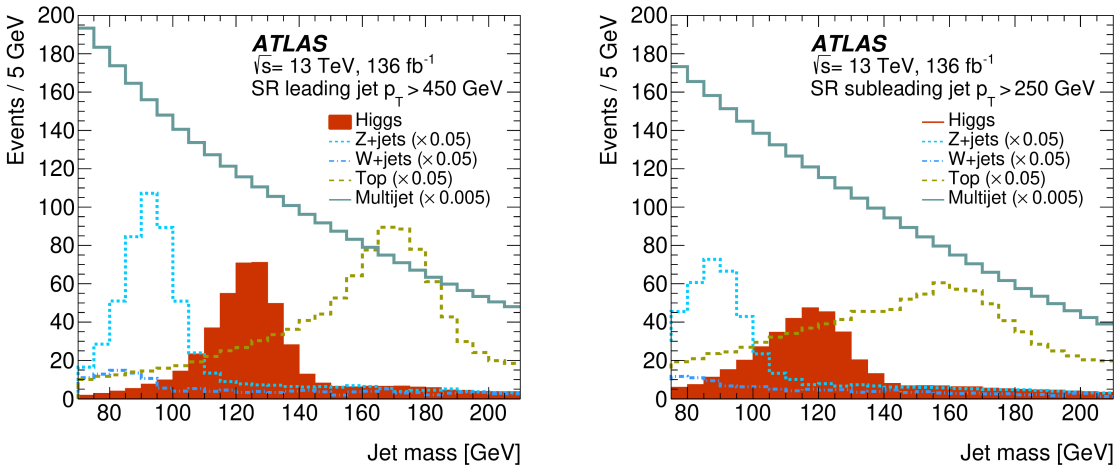


Figure 5.2: Diagram showing the event categorization criteria. Candidate jets are categorized by the number of ghost-associated b-tagged track jets as well as its p_{T} ordering in the event.

Signal Region

For an event to be assigned to the SR it must be a double-tagged candidate jet. If the event is the leading jet then it will populate the leading jet signal region (SRL). If the double-tagged jet is not the leading jet, but the subleading jet, it will populate the subleading-jet signal region (SRS). Figure 5.3 shows the mass distributions for all the processes that contribute to the signal region.



(a) (b)

Figure 5.3: Jet mass distributions for the Higgs boson, Z+jets, W+jets, and top quark contributions from the SM prediction as well as the multijet jet mass distribution extracted from data in the signal region (SR) defined by the leading (a) and subleading (b) jets [10].

SR Configurations

The signal is extracted in three different SR configurations, providing three measurements of the Higgs cross-section. First, for the inclusive measurement, the Higgs boson signal strength μ_H is extracted from the signal region containing candidate jets with $p_{\mathrm{T}} > 250$ GeV. Second, a fiducial measurement is performed on the fiducial volume where the candidate jets have $p_{\mathrm{T}} > 450$ GeV and $|y_H| < 2$ defined by the acceptance cuts of this analysis. Finally, the

differential measurement, where the signal strength extraction is performed for candidate jets in the p_{T} ranges 250-450 GeV, 450-650 GeV, 650-1000 GeV and > 1 TeV. The bin with $250 < p_{\mathrm{T}} < 450$ GeV is populated only by candidate jets from the subleading signal region (SRS).

Table 5.5: Summary of the candidate jet p_{T} requirements for the three Signal Region configurations [10].

Region	Candidate jet p_{T} (GeV) SRL	SRS
Inclusive	> 450	> 250
Fiducial	> 450	> 450
Differential	$450 - 650$, $650 - 1000$, > 1000	$250 - 450$, $450 - 650$, $650 - 1000$

Validation Region

Similarly to the SR region the VR is subdivided into the leading jet validation region (VRL) and the subleading jet validation region (VRS). The main difference lies in the b-tagging requirements. Every event of the VR must be an anti-tagged candidate jet.

Control Region

To constrain the $t\bar{t}$ background a dedicated control region $\mathrm{CR}_{t\bar{t}}$ was defined. The high purity of top quark pair events is achieved using the muon-trigger to choose events in which one of the tops decays semiletopnically ($t \to l\nu b$) while the other decays hadronically ($t \to qq'b$). Each large-R jet in this region must have at least one b-tagged VR track-jet associated to it. The large-R jet that has a close isolated muon with $p_{\mathrm{T}} > 52.5$ GeV is labeled as J_b and is associated to the semileptonically decaying top quark. On the other hand, for the

hadronically decaying top, the large-R jet J_t requires at least 3 associated VR track-jets. These two jets (J_b, J_t) must have an angular separation of at least $\Delta\phi > 2\pi/3$ to ensure a back-to-back topology. Table 5.6 summarizes the selection criteria used for the $t\bar{t}$ control region $\mathrm{CR}_{t\bar{t}}$.

Table 5.6: Summary of the $\mathrm{CR}_{t\bar{t}}$ selection criteria for the semileptonically decaying top quark (J_b) and the hadronically decaying top quark (J_t) [10].

Jet	N track-jets	N b-tags	Angular selection	Jet mass [GeV]
J_b	≥ 1	1	$0.04 + 10/p_{\mathrm{T}}^{\mu} < \Delta R(\mu, J^b) < 1.5$	–
J_t	≥ 3	1	$\Delta\phi(J^b, J^t) > 2\pi/3$	$140 - 200$

5.5 Higgs Boson Modeling

The selection criteria chosen in this analysis provides an inclusive view of the Higgs boson in terms of its four main production modes. Therefore the production modes considered are ggF, VH, VBF and $t\bar{t}H$. Considering Higgs bosons near the mass peak ($105 < m_J < 140$ GeV) and $p_{\mathrm{T}} > 450$ GeV, the largest contribution for Higgs production comes from the ggF process. On the other hand, in the SRS ($p_{\mathrm{T}} < 450$ GeV) the largest production mode is $t\bar{t}H$. In this bin, highly energetic hadronic tops can satisfy the trigger requirements in events where the Higgs boson has low p_{T} due to the nature of three body decays. Table 5.7 shows the relative contribution of the main production modes in the SR. The Higgs boson also contributes to the VR to a lesser extent. The breakdown of the contribution of the Higgs boson to the SR and VR as a function of mass is shown in Figure 5.4 and as a function of p_{T} in Figure 5.5.

Multiple modeling systematics were considered for the Higgs boson. These are: factorization and renormalization scale variations, cross-section and acceptance, PDF uncertainty,

Table 5.7: Fractional contribution of each production mode to the SR configurations around the Higgs boson mass peak ($105 < m_J < 140$ GeV) [10].

Process	Jet p_{T} range (GeV)			
	$250 - 450$	$450 - 650$	$650 - 1000$	> 1000
		SRL		
ggF	–	0.56	0.50	0.39
VBF	–	0.17	0.16	0.17
VH	–	0.14	0.18	0.25
$t\bar{t}H$	–	0.13	0.16	0.19
		SRS		
ggF	0.28	0.46	0.43	–
VBF	0.07	0.19	0.21	–
VH	0.26	0.24	0.26	–
$t\bar{t}H$	0.39	0.11	0.10	–

jet shower systematics and EW correction uncertainties. For the factorization and renormalization scale variations a 7-point scale variation on $\mu_{R/F}$ was performed. The variation was found to show a flat effect in the cross section of 2% for ggF, 0.5% for VBF, 5% for VH and 13% for $t\bar{t}H$. For the shower systematics, Pythia and Herwig samples at the truth level were compared. From these samples, p_{T} and m_J dependent reweighting maps were constructed, applied to the recontructed level samples and used to estimate the uncertainty. The shower systematics were found to be negligible, as no substantial differences were seen in the m_J shape in the resonance peak region. Appendix A contains plots related to the studies of the jet parton showers as well as the EW correction systematics [117].

Higgs Boson Resolution

The p_{T} and mass resolution of the Higgs jet is studied by truth matching candidate jets to a Higgs boson. A Gaussian is then fitted to the difference of the reconstructed jet and the truth jet values. The Gaussian resolution is therefore taken as the standard deviation of the

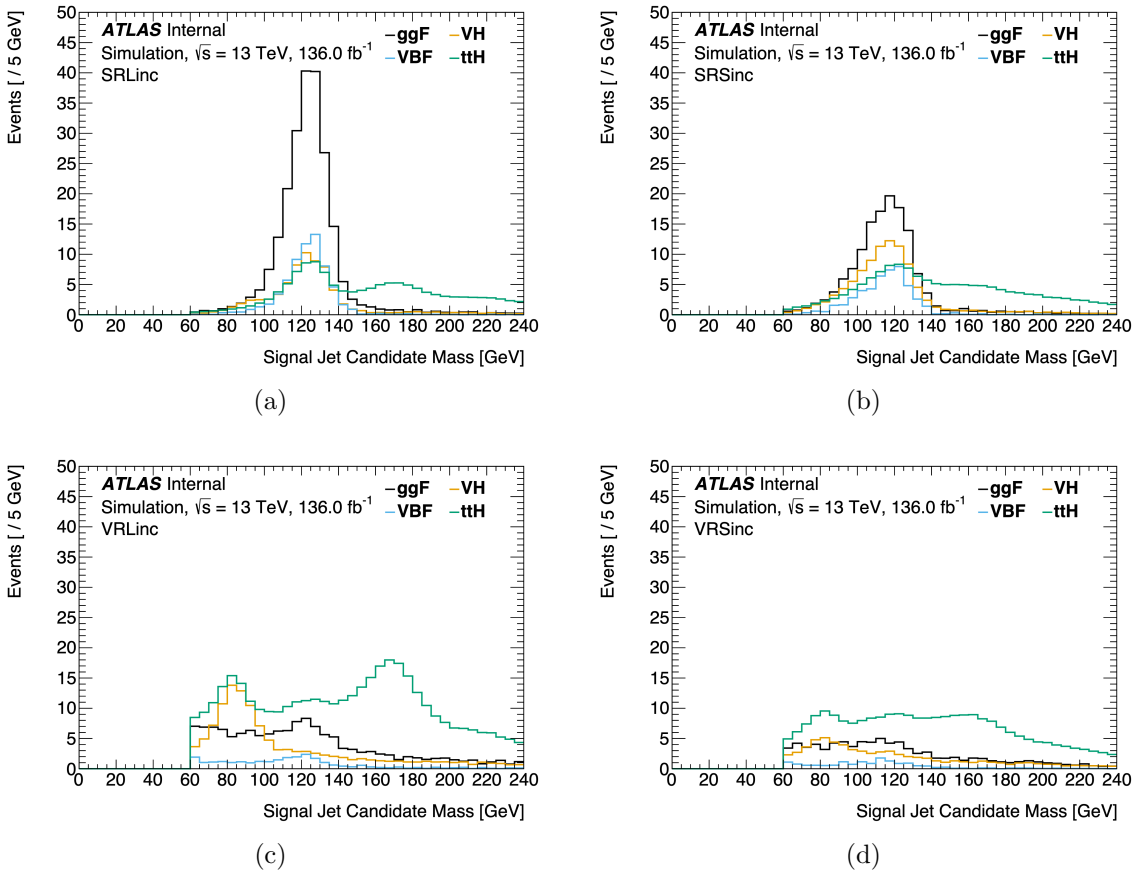
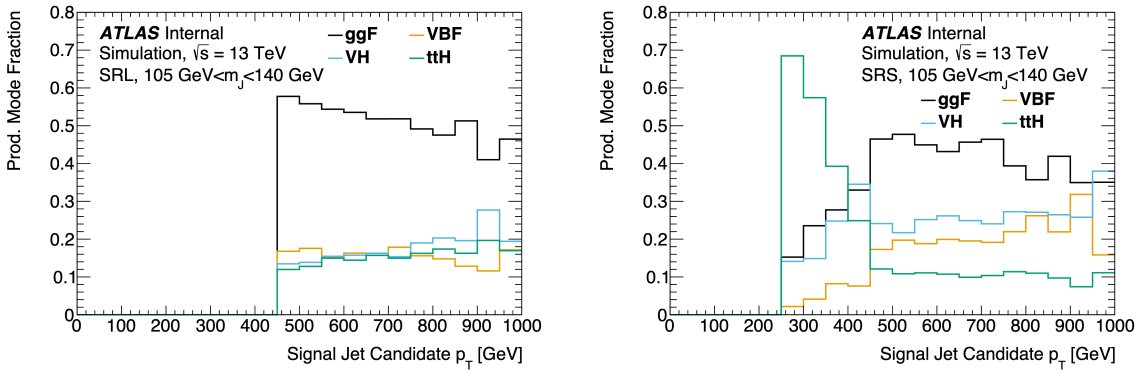


Figure 5.4: Breakdown of the Higgs boson contributions to the mass peak for the different production modes for signal (a,b) and validation regions (c,d). The plots on the left (a,c) show the contribution to the leading regions and on the right (b,d) to the subleading regions.

fitted Gaussian. The impact of pile-up (PU) on the mass resolution was also studied and shown to be small, given that the large-R jets are subject to a trimming procedure. The p_{T} resolution is shown in Table 5.8 and the mass resolution is shown in Table 5.9.

5.6 Background Process Modeling

The dominant background process is QCD multijet production which presents itself as a non-resonant monotonically decreasing spectrum. The resonant backgrounds, the V+jets process and the top quark, peak outside the Higgs boson signal window but still contribute



(a) (b)

Figure 5.5: Breakdown of the contributions for the different Higgs boson production modes for the (a) SRL and (b) SRS as a function of p_{T}.

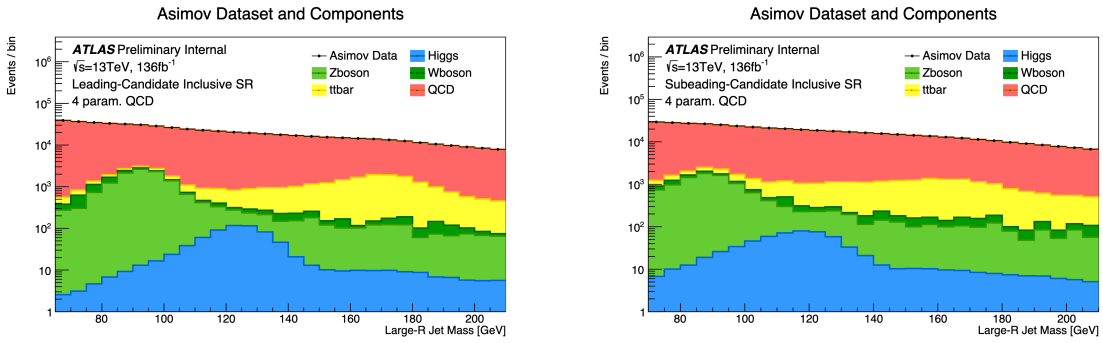
Table 5.8: Momentum resolution of the candidate jets truth matched to a Higgs boson for ggF events.

p_{T} [GeV]	p_{T} Resolution [GeV] Leading	Subleading
$250 < p_{\mathrm{T}} < 450$	–	38.3 ± 1.2
$450 < p_{\mathrm{T}} < 650$	29.3 ± 0.3	39.9 ± 1.2
$650 < p_{\mathrm{T}} < 1000$	38.3 ± 0.9	56.1 ± 2.1

in a minor way. Within the Higgs mass peak ($105 < m_j < 140$ GeV), the V+jets process, represents approximately 1% of the total background, top quarks represent about 3%, and the rest of the background is due to QCD multijets. Figure 5.6 shows the mass distributions of the expected MC estimates (Asimov datasets) of the signal and backgrounds for the signal regions.

Table 5.9: Mass resolution of the candidate jets truth matched to a Higgs boson for ggF events.

p_{T} [GeV]	Mass Resolution [GeV] Leading	Subleading
$250 < p_{\mathrm{T}} < 450$	–	17.7 ± 0.7
$450 < p_{\mathrm{T}} < 650$	11.3 ± 0.1	13.4 ± 0.2
$650 < p_{\mathrm{T}} < 1000$	10.8 ± 0.3	13.5 ± 0.4



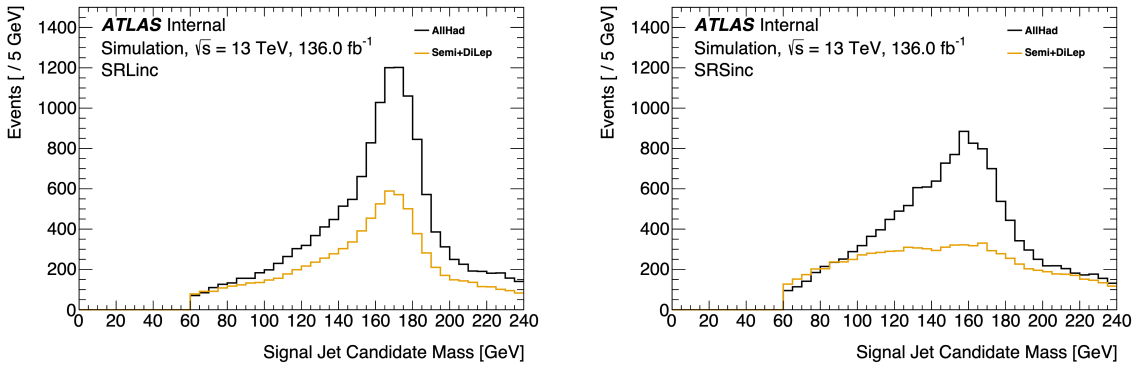
(a) (b)

Figure 5.6: Mass distributions in the (a) SRL and (b) SRS for the MC estimates of the signal and backgrounds [10].

Top Quark Modeling

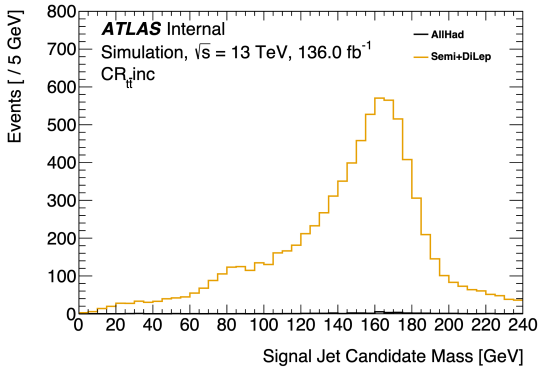
A top quark event is characterized by the presence of a b-quark and two hadronic decay products of a W boson. The $t\bar{t}$ control region $\mathrm{CR}_{t\bar{t}}$ was defined to estimate the top quark contributions to the signal regions. The jet mass distributions in both the $\mathrm{CR}_{t\bar{t}}$ and the SR are comparable given that both regions probe a similar phase space. Figure 5.7 shows the breakdown of the $t\bar{t}$ contribution to the candidate jet mass distribution in the signal and control regions. The shape of the spectrum is taken from MC but the normalization is extracted from data. Adjustments performed on the $\mathrm{CR}_{t\bar{t}}$ are directly applied to the SR by including it in the global likelihood fit. To extract the scale factor, a simultaneous fit is performed on the $\mathrm{CR}_{t\bar{t}}$ and SR together. Given that the $\mathrm{CR}_{t\bar{t}}$ has a $t\bar{t}$ purity of 97% (with similar levels in the fiducial and differential regions), the normalization was determined from data with better than or equal to 10% precision.

In the SR, single top events contribute between 2-3% (tW) and 1-5% (t-channel) of the candidate jets relative to the total $t\bar{t}$ yield. These events have a candidate jet mass distribution similar to that of the $t\bar{t}$ events. The s-channel contribution was found to be negligible. To account for this contribution to the likelihood fits the $t\bar{t}$ MC was scaled



(a)

(b)



(c)

Figure 5.7: Breakdown of the $t\bar{t}$ contribution to the candidate jet mass for the inclusive (a) SRL, (b) SRS and (c) $\mathrm{CR}_{t\bar{t}}$.

accordingly to match number of events in tW and t-channel MC samples for each p_{T} bin. A 50% normalization uncertainty was applied to the estimated number of single top quark events due to comparisons between diagram subtraction and diagram removal schemes [118] in tW events.

Standalone fits were also performed for the $\mathrm{CR}_{t\bar{t}}$ using Asimov datasets. The measurement of the scale factor had greater uncertainty in the high p_{T} region due to the lower number of events. It was found that the scale factor between the data and the MC was about 0.8. For the global fit of the SR with the control region, the data and simulation seem to agree. Figure 5.8 shows the results of the fit for the differential analysis regions.

Systematic uncertainty estimates were calculated using simulated samples for alternative parton shower models (POWHEG vs HERWIG) finding 6-19% difference in yield across the analysis regions. Similarly, uncertainties due to matrix element calculations (MADGRAPH5 vs POWHEG BOX V2) were performed and found to have a 1-19% difference in yield. Weight variations on the nominal sample associated with initial and final state radiation (ISR and FSR) produced uncertainties between 1-7%. Renormalization and factorization scale variations were found to be negligible. The two largest uncertainties on the $t\bar{t}$ normalization were from b-tagging efficiency and JMS.

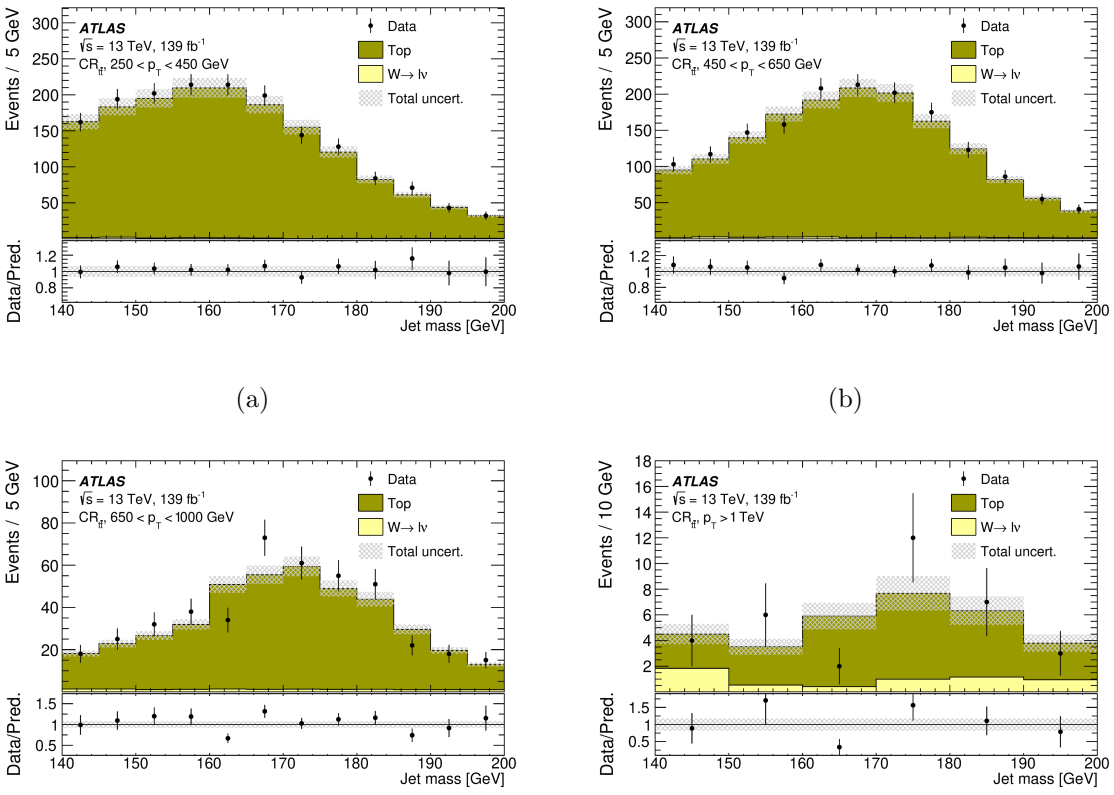


Figure 5.8: The post-fit CR$_{t\bar{t}}$ J_t mass distribution in the four p_{T} regions used in the global likelihood of the differential fit. The $W(l\nu)$ contribution is flat in jet mass and for events with $p_{\mathrm{T}} < 1$ TeV it is estimated to be 1-3% of the total. The $p_{\mathrm{T}} > 1$ TeV region is shown in 10 GeV jet mass bins. The ratio of the data to the background prediction is shown in the lower panel. The shaded areas indicate the 68% CL for all background processes [10].

V + jets Modeling

Vector boson production offers a unique opportunity to validate the signal measurement procedure for the Higgs given that their decay structure, mass peak and resolution are similar. The Z boson, especially, given that it populates the signal region with around 20 times more events than the Higgs boson, it can be used to study experimental effects that would not be apparent with the statistically limited Higgs boson measurement. Therefore, to have a proper measurement of H, a well understood Z background is neccesary.

In the SR, the number of Z+jets events is more than 3 times that of W+jets because of the large branching ratio of its b-quark pair decay ($Z \to b\bar{b}$) coupled with our b-tagging selection criteria. For approximately 90% of the candidate jets in Z+jets events, the decay products of the Z bosons are fully contained within the jet. On the other hand, only 40% of the candidate jets in W+jets events contain its decay products. This is due to the low misidentification rate for b-tagging. The remaining candidates from W+jets events come from the recoiling hadronic system resulting in a broader m_J distribution in the SR.

In the VR, due to the requirement that the candidate jets must be anti-tagged, W+jets are three times larger than Z events. They both have comparable acceptance but the W has a larger cross-section. The decay products of the vector bosons are reconstructed within the candidate jets only in 60% of events. This results in a non-resonant mass distribution, with a shape that is similar to the QCD multijet background.

Given that the Z+jets normalization is directly extracted from the data with the global likelihood fit, the systematic uncertainties of the modeling are limited to changes in acceptance in the different regions and to the mass distribution shape. For the W+jets cross section, a 10% uncertainty in the signal region is assigned [119]. The semi-leptonic W+jets

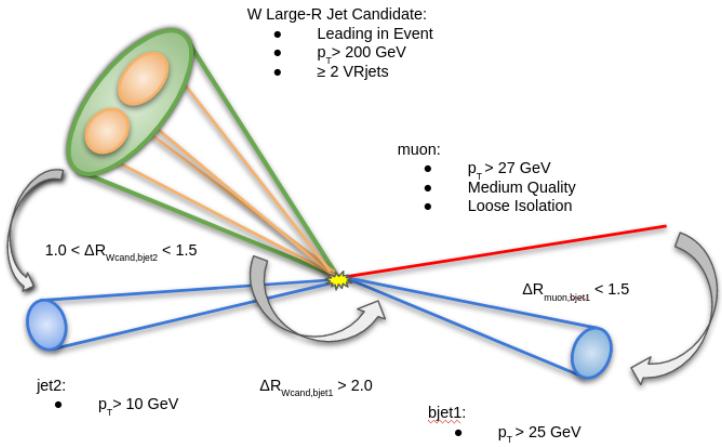
decays ($W \to l\nu$) contribution in the $\mathrm{CR}_{t\bar{t}}$ has a total uncertainty of 30%. Systematic uncertainties due to renormalization and factorization scale variation represent a 3-20% error to the acceptance across the different regions. Other variations were studied, but found to have a negligible impact. These include an alternative PDF ($\mathrm{MMHT2015_{NLO}}$ [120]), α_s variations in the nominal PDF and alternative cluster fragmentation modeling (Lund string model [121]). For the normalization, the largest experimental uncertainties are associated with the JMR and JMS. Agreement between simulation and data in the leading jet VR is shown in Figure 5.10.

V + jets Resolution

It was found that the fitted Z+jets normalization in the SR had a correlation with the reconstructed mass resolution. This is due to the flexibility of the Z+jets template and the multijet model (discussed in the next section). In some cases, the best value of the JMR parameter broadened the Z+jets peak, which corresponds to a increase of Z+jets normalization and a decrease of the contribution of multijets compared to the expected values. A dedicated control region rich in large-R jets containing W bosons from the decay products of semileptonic $t\bar{t}$ decays was created to constrain the JMR systematic in conjunction with the VRL. This control region is denoted as the $\mathrm{WCR_{t\bar{t}}}$.

The $\mathrm{WCR}_{t\bar{t}}$ requires the presence of two top quarks in different hemispheres where one top quark decays leptonically while the other top quark decays hadronically. The decay products of the W boson from the hadronically decaying top quark must be isolated in the large-R jet. This region provides a high purity reconstructed W peak with p_{T} from 200-600 GeV. Similar to the $\mathrm{CR}_{t\bar{t}}$, an isolated medium quality muon is used. The selection requires at least one large-R jet (leading will be the W candidate) with $p_{\mathrm{T}} > 200$ GeV and

at least 2 VR track jets with $p_{\mathrm{T}} > 10$ GeV. Both VR track jets must pass the b-tagging requirements. One of the b-tagged VR track jets has to be close to the muon by satisfying: $0.04 + 10/p_{\mathrm{T}}^{\mu} < \Delta R_{\mathrm{btag1,muon}} < 1.5$ and to also have a $p_{\mathrm{T}} > 25$ GeV. This b-tagged jet must be well separated from the W candidate ($\Delta R_{\mathrm{btag1,Wcand}} > 2.0$). The second b-tagged VR track jet is required within $1.0 < \Delta R_{\mathrm{btag2,Wcand}} < 1.5$ of the W candidate. Figure 5.9 shows a diagram of the topology of the $\mathrm{WCR}_{\mathrm{t\bar{t}}}$ events. The mass and p_{T} distributions of the inclusive $\mathrm{WCR}_{\mathrm{t\bar{t}}}$ can be seen on Figure 5.11.

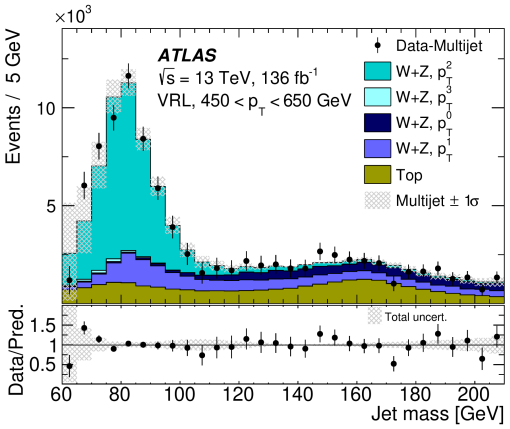


(a)

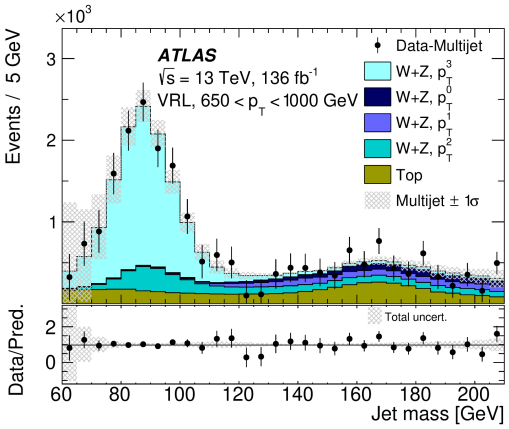
Figure 5.9: Diagram depicting the topology of a $\mathrm{WCR}_{\mathrm{t\bar{t}}}$ event. The hadronically decaying W boson must be isolated in the large-R jet.

The $\mathrm{WCR}_{\mathrm{t\bar{t}}}$ provides a good source of W bosons in the p_{T} range below 600 GeV. The VRL provides a clear peak in p_{T} ranges above 450 GeV but with more multijet background. The jet mass width of the W and Z resonances show a slow evolution from low p_{T} in the $\mathrm{WCR}_{\mathrm{t\bar{t}}}$ to high p_{T} in the VRL. This can be seen in Figure 5.12. The results from of the jet mass width using the $\mathrm{WCR}_{t\bar{t}}$ have around 1/5 of the original JMR uncertainty after the contraint transfer to the $Z \to b\bar{b}$ dominated V+jets sample in the SR. The correlation

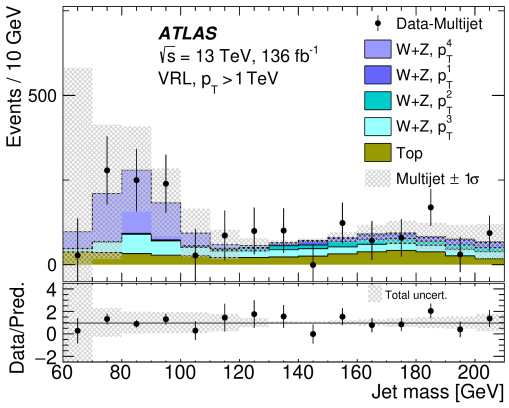
between the Z+jets normalization and the JMR is reduced when included in the global likelihood fit. In the inclusive signal region, the correlation reduces from around 90% to 30% when this auxiliary mass measurement is considered.



(a)

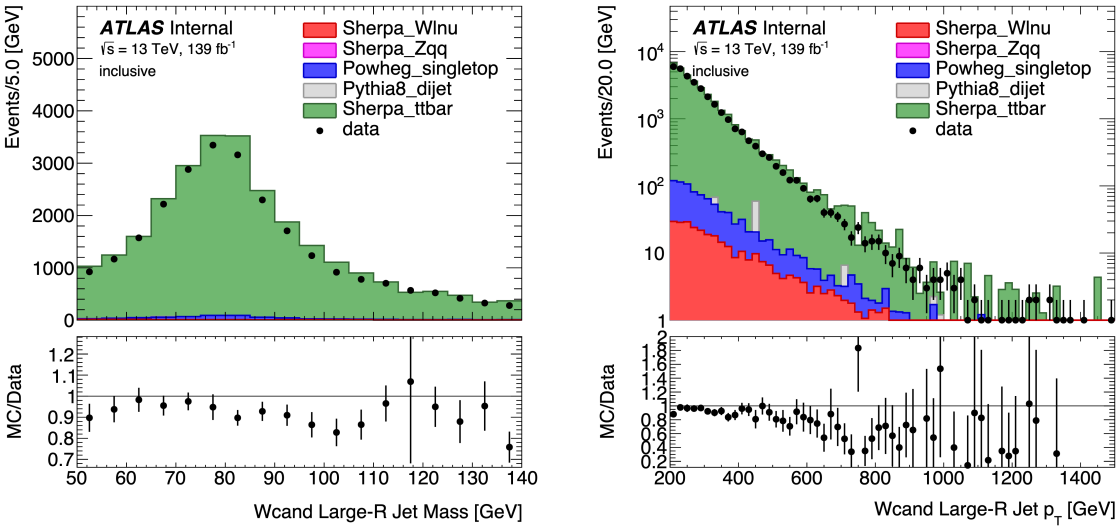


(b)



(c)

Figure 5.10: Post-fit leading-jet invariant mass distributions after the multijet background was subtracted in the validation region for data and the V+jets (W+Z) and top quark components for (a) $450 < p_{\mathrm{T}} < 650$ GeV, (b) $650 < p_{\mathrm{T}} < 1000$ GeV, and (c) $p_{\mathrm{T}} > 1$ GeV shown in wider 10 GeV jet mass bins. The V+jets contribution is split into five generator 'truth' p_{T}^{V} volumes labeled p_{T}^{0}–p_{T}^{4} for $p_{\mathrm{T}}^{V} < 300$ GeV, 300–450 GeV, 450–650 GeV, 650–1000 GeV, and > 1000 GeV, respectively. The $t\bar{t}$ normalization and its uncertainty are set to the corresponding values from the $\mathrm{CR}_{t\bar{t}}$. The m_J range has been extended down to 60 GeV for only this fit to show the level of agreement along the rising edge of the V+jets m_J distribution. The ratio of the data to the background prediction is shown in the lower panel. The shaded areas indicate the 68% CL for all background processes [10].



(a) (b)

Figure 5.11: Inclusive $\mathrm{WCR}_{t\bar{t}}$ (a) mass distribution and (b) p_{T} distribution of the W candidates [122].

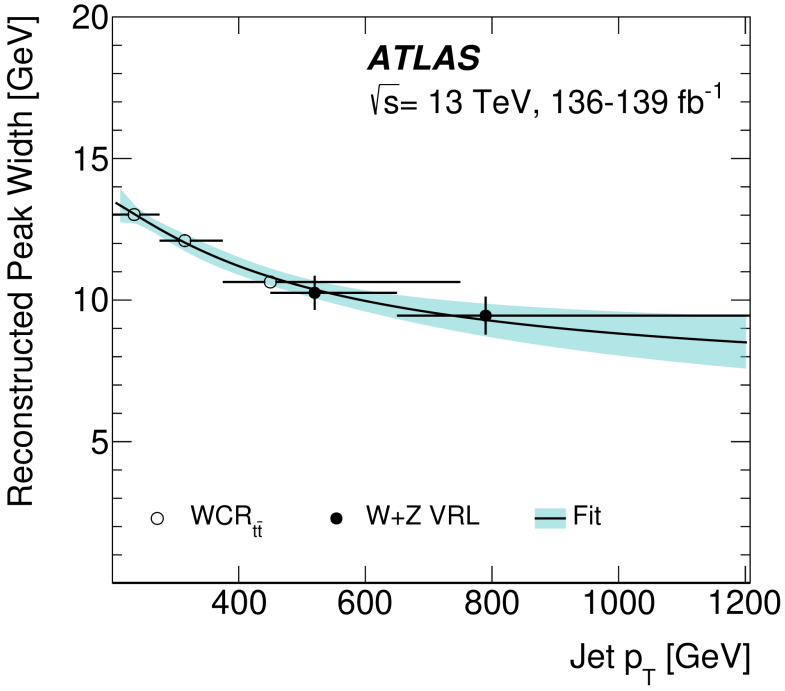


Figure 5.12: A summary of the Z and W resonance peak reconstructed-width measurements as a function of the jet p_{T} using a resolved W boson in top quark decays in the $\mathrm{WCR}_{t\bar{t}}$ region and the combined W and Z boson mass distribution in the validation region. The continuous black curve is a fit to the measurements with resultant errors shown as a cyan band [122].

Multijet Modeling

The QCD multijet background has a monotonically decreasing mass spectrum. It is modeled using an exponential function of a polynomial of degree N with the form:

$$f_N(x|\mathbf{\Theta}) = \Theta_0 \exp\left(\sum_{i=1}^{N} \Theta_i x^i\right), \tag{5.13}$$

where Θ_i are the parameters of the fit and $x = (m_J - 140)/70$ GeV. The parameters are simultaneously determined during the signal extraction fit independently for each region. The number of events and the shape of the spectrum has an impact on the optimal degree of the polynomial of the model. Small values of N make the function too rigid and therefore prone to bias in the resonant process yields. On the other hand, large values of N decrease the statistical significance of the resonant process models, due to the increased correlation which can create or absorb the resonances. Modified VR (which we call hybrid VR) ensembles are used to study the optimal values for N given that they contain more than 50 times the amount of data than the SR. Ten of these ensemble (VR slices), with roughly the same amount of data as the SR, are used to find the optimal parameters of the fit by taking an average for each of the regions.

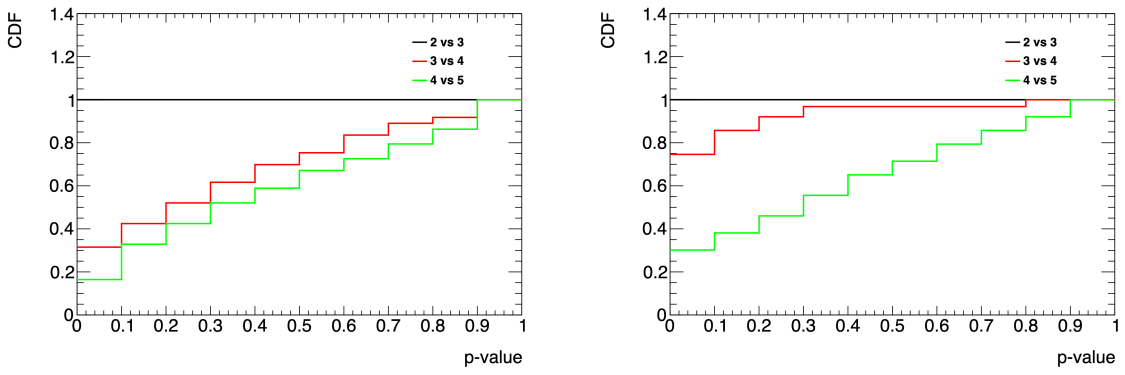
The hybrid VR is constructed by replacing the VR resonance peaks with the SM prediction in the SR while correcting the mass spectrum to match the SR. A shape correction factor, defined as the ratio of the SR multijet estimate ($\mathrm{MJ_{SR}}$) with the VR model ($\mathrm{MJ_{VR}}$), is applied to the VR slices. The values of $\mathrm{MJ_{SR}}$ are obtained from the likelihood fit of the SR and $\mathrm{CR_{t\bar{t}}}$ while including all the systematic uncertainties. The $\mathrm{MJ_{VR}}$ is taken as the average of likelihood fits of 10 random orthogonal subsets of the VR while including all systematic uncertainties. The resonant peak estimates for $V+$ jets and Top (V_{VR} and $\mathrm{Top_{VR}}$)

are extracted from the average post-fit contributions of the same 10 VR fits. Each hybrid VR ($\mathrm{VR}^i_{\mathrm{hyb}}$) is defined as

$$\mathrm{VR}^i_{\mathrm{hyb}} = (\mathrm{VR}^i - V_{\mathrm{VR}} - \mathrm{Top}_{\mathrm{VR}}) \times \frac{\mathrm{MJ}_{\mathrm{SR}}}{\mathrm{MJ}_{\mathrm{VR}}} + V_{\mathrm{SR}} + t\bar{t}_{\mathrm{SR}} + H_{\mathrm{SR}}, \tag{5.14}$$

where VR^i is the mass distribution of the data events in the VR slice and the variables with subscript SR are the nominal MC predictions for the resonant sources in the SR.

The log-likelihood ratio (LLR) is used to test the results between different values of the polynomial degree N in each $\mathrm{VR}_{\mathrm{hyb}}$ but without the injected resonances. The null hypothesis is defined as the fit using a polynomial of degree N, while the alternate hypothesis is the fit using a polynomial of degree $N+1$. Wilk's theorem [123] relates the log-likelihood ratio to a χ^2 distribution with $N+1-N=1$ degrees of freedom (d.o.f.). The smallest value of N that yields a uniform distribution of p-values is selected as the optimal model. A uniform distribution is represented by a linear increase in the corresponding cumulative distribution function (CDF). Figure 5.13 shows the CDF as a function of p-value for the VR.



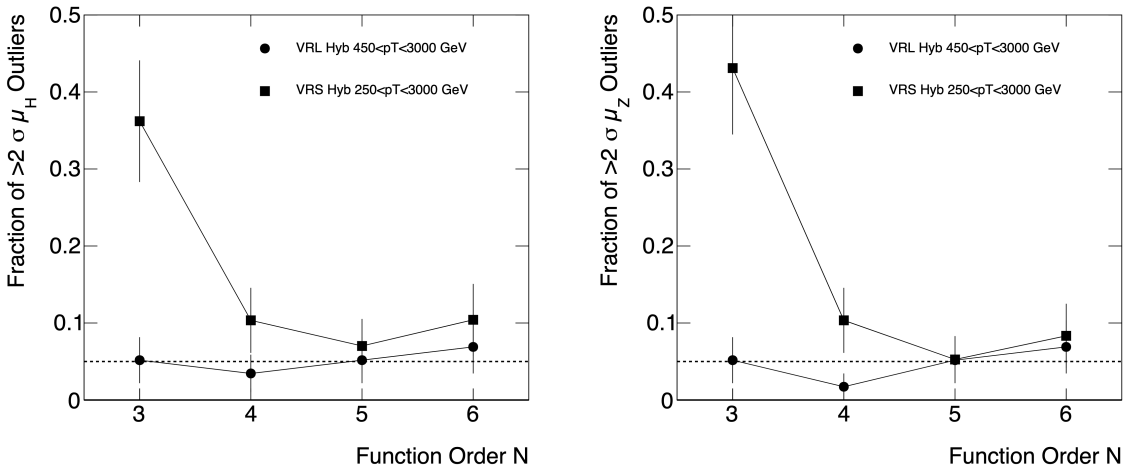
(a) (b)

Figure 5.13: Cumulative distribution function (CDF) of the p-values of the log-likelihood ratio of the exponential polynomial of degrees N and $N+1$. Plots correspond to the (a) VRL and (b) VRS [122].

To look for local effects due to the resonances, tests were performed by including a free normalization parameter $\mu_{\mathrm{VR}} \pm \sigma^{\mathrm{VR}}_{\mathrm{stat}}$ for either the Z + jets process or the Higgs boson in a fit for the $\mathrm{VR}_{\mathrm{hyb}}$ by doing an artificial signal injection. The quantity $F_{2\sigma}$ was used to estimate the probability that the multijet model allows artificial excesses or deficits. $F_{2\sigma}$ is defined as the fraction of fitted Z and H signal in excess of twice its error $\sigma^{\mathrm{VR}}_{\mathrm{stat}}$:

$$|\mu^{\mathrm{VR}} - 1| > 2\sigma^{\mathrm{VR}}_{\mathrm{stat}}. \tag{5.15}$$

The average ratio $\overline{\mu/\sigma} = (\mu^{\mathrm{VR}} - 1)/\sigma^{\mathrm{VR}}_{\mathrm{stat}}$ quantifies the bias in the signal strength determination and can be used to estimate the spurious signal systematic uncertainty when applied to $\mathrm{VR}_{\mathrm{hyb}}$ without any signal. The value of N is chosen so that $F_{2\sigma}$ is compatible with a value of 0.05 and $\overline{\mu/\sigma}$ is stable for both Z + jets and Higgs production. The spurious signal systematic uncertainties range from 0.01-0.33 for H and 0.15-0.65 for Z. Figure 5.14 shows the values of $F_{2\sigma}$ for the Higgs and the Z bosons.



(a) (b)

Figure 5.14: Fraction of fitted signal in excess of twice its error for (a) H and (b) Z as a function of the exponential polynomial degree N [122].

The optimal values for N were found to be $N = 5$ in the inclusive region and between 4 and 5 for the differential p_{T} bins. The results are summarized in Table 5.10, where the differential bins are labeled as p_{T}^{0} ($250 < p_{\mathrm{T}} < 450$ GeV), p_{T}^{1} ($450 < p_{\mathrm{T}} < 650$ GeV), p_{T}^{2} ($650 < p_{\mathrm{T}} < 1000$ GeV) and p_{T}^{3} ($p_{\mathrm{T}} > 1$ TeV). A comparison of the QCD multijet fits for all the p_{T} binned analysis regions is shown on Figure 5.15.

Table 5.10: Optimal degree N of the exponential polynomial used to model the QCD multijet background for all the analysis regions

Candidate jet	Inclusive	Differential			
		p_{T}^{0}	p_{T}^{1}	p_{T}^{2}	p_{T}^{3}
Leading	5	–	5	4	4
Subleading	5	5	4	5	4

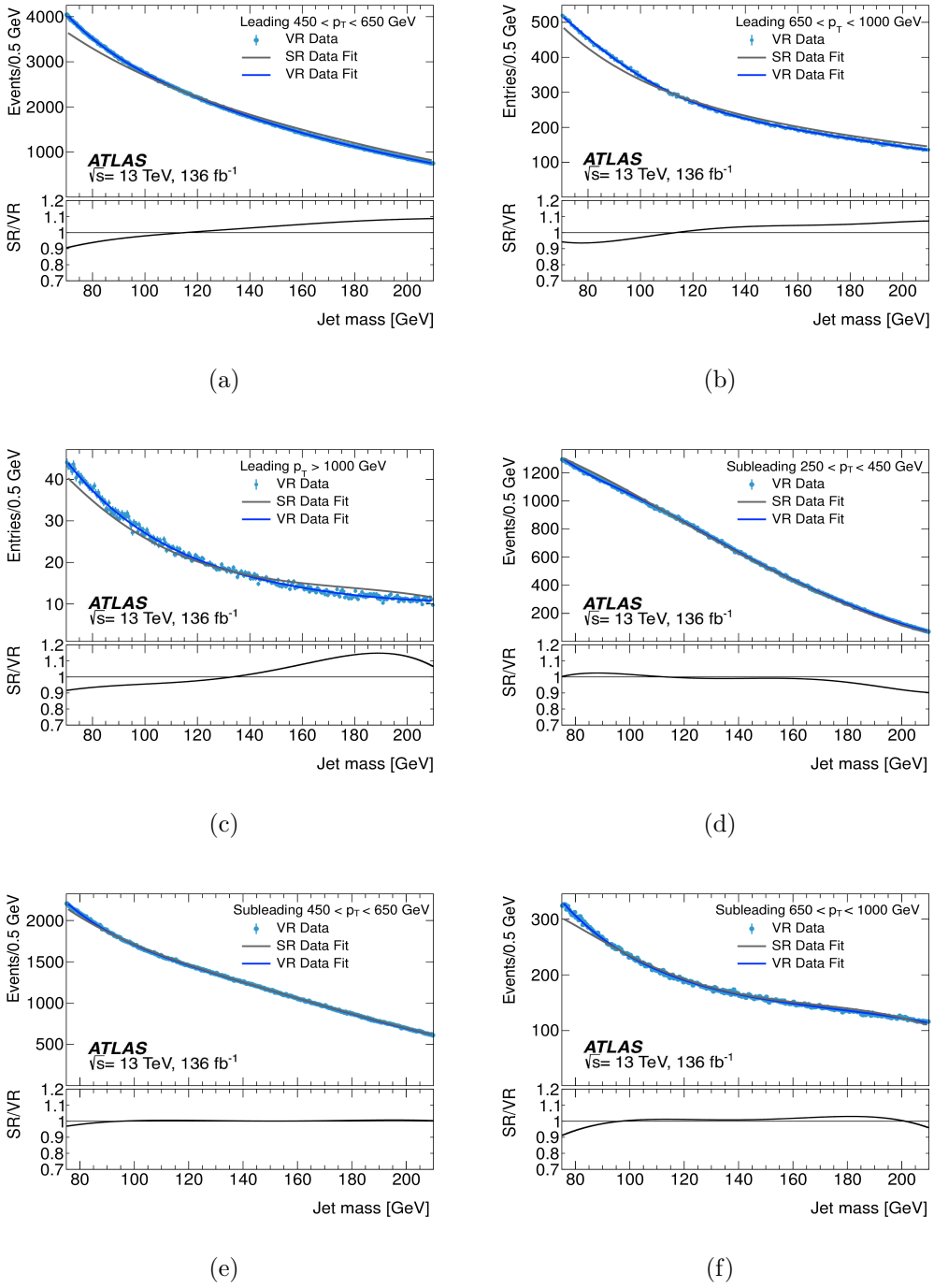


Figure 5.15: Multijet jet-mass distribution from the different p_{T}-binned analysis regions. The solid lines show the multijet function after a fit to the SR data (gray) and VR data (blue). The solid points are the data from VR slices with the same number of events as the SR after the SM resonances are subtracted. The bottom panel shows the ratio of the SR data fit to the VR data fit [10].

5.7 Statistical Analysis

The signal extraction is performed using the maximum likelihood method. In practice, this is achieved through the minimization of the negative log likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\theta})$ using the RooStats framework [124] and the RooFit library [125]. The likelihood function is defined as a product of Poisson probability density functions as described in Sec 5.1. One of these terms is defined for each m_J bin of the SRL, SRS and $\mathrm{CR}_{t\bar{t}}$. The bin width for the mass distribution was chosen to be 5 GeV. A recent RooFit extension [126] was needed to remove an existing bias for wide binned datasets. The nuisance parameters $\boldsymbol{\theta}$ represent the systematic uncertainties and are constrained with Gaussian or log-normal probability density functions. The V+jets JMR contraints obtained from the $\mathrm{WCR}_{t\bar{t}}$ and VRL are implemented as Gaussian p.d.f priors. The normalization of the MC templates is controlled by free parameters for each p_{T} region or the truth-based volume common to the SRL, SRS and $\mathrm{CR}_{t\bar{t}}$. For the multijet model, both the normalization and the polynomial coefficients are treated as free parameters independent from each jet mass distribution. The yield of each of the signals is given by the signal strength μ. The signal stregnth μ is defined as ratio between the fitted number of signal events and the corresponding SM prediction. Upper limits on the Higgs boson signal strength μ_H and production cross section σ_H are obtained using the CL_s method, where the expected limits are determined by assuming no Higgs boson contribution.

Systematic Uncertainties

Uncertainties related to a low number of events in MC samples for the background predictions were parametrized with the Beeston-Barlow technique [127]. A smoothing procedure was

also used to remove large variations with a threshold for pruning of only 2% [128]. Table 5.11 summarizes all the systematic uncertainties considered in the likelihood fit for the H and Z signal strength extraction.

Table 5.11: Summary of the systematic uncertainties included in the proifle likehood fit for the signal strength extaction. The second column states the processes for which an independent nuisance parameter is considered. The third column indicates the regions for which the systematic uncertainty is correlated. The fourth column describes the effect of the systematic uncertainty induced by the parameter: N denotes a normalization change and S represents an impact to the shape. (*) $t\bar{t}$ and V+jets events have two extra minor components only applied to them. ($\star$) This uncertainty only covers relative acceptance across regions instead of the absolute cross section uncertainty. ($\bullet$) Only applied to Z+jets when the signal extraction performed on truth-based volumes is tested using the SR [10].

Description	Processes	Category	Effect
Reconstructed object systematic uncertainties			
JMR	$t\bar{t}$, V+ jets, H	p_{T}	N+S
JMS (dominant)	$t\bar{t}$, V+ jets, H	p_{T}	N+S
JMS (rest)	$t\bar{t}$,V+ jets $+H$	all	N+S
Jet Energy Scale	all$^{(*)}$	all	N+S
Jet Energy Resolution	all	all	N+S
b-tag efficiency for b-jets	all	all	N+S
b-tag efficiency for c-jets	all	all	N+S
b-tag efficiency for $light$-jets	all	all	N+S
Process modeling systematic uncertainties			
Renormalization and factorization scale	V+ jets	all	N+S
Cross section	W+ jets	all	N
Cross section and acceptance	$W(l\nu)$	all	N
Parton shower model	$t\bar{t}$	all	N+S
Matrix element calculation	$t\bar{t}$	all	N+S
Initial and final state radiation	$t\bar{t}$	all	N+S
Cross section and acceptance	t	all	N
Cross section and acceptance$^{(\star)}$	H	all	N
NLO EW corrections	VBF $+VH+t\bar{t}H$	all	N
	H	p_{T}^{H} bins $\times$ LS	N
Spurious signal	Z+ jets$^{(\bullet)}$	p_{T}^{Z} bins $\times$ LS	N

Chapter 6

Boosted $H \to b\bar{b}$ Results

Three different configurations are used to study Higgs boson production at high p_{T}. The inclusive region is used to measure the H signal strength, the fiducial region is used to measure the fiducial cross section and the differential regions are used to measure the cross section for four different p_{T} bins.

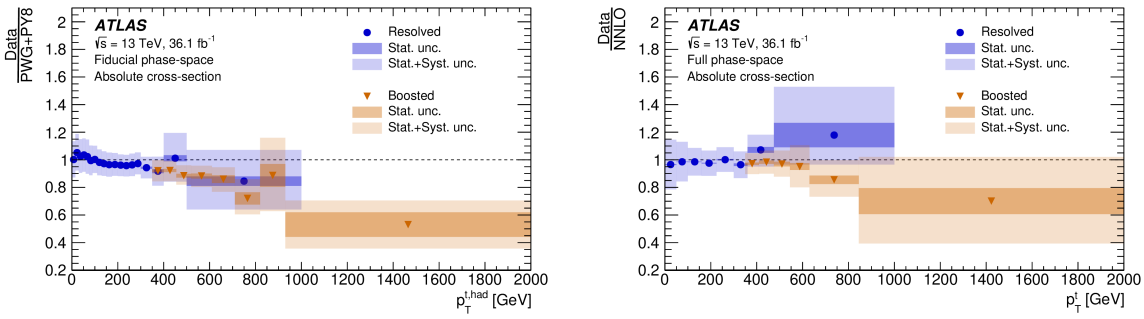
The signal strength extraction in the fiducial region considers the events on the fiducial volume defined by the requirements applied to the truth Higgs boson transverse momentum p_{T}^{H} and its rapidity y_H. The same truth information is used for the differential regions. The p_{T}-y volume bins are based on the simplified template cross-section (STXS) framework [129][130] for ggF production, with the modification of a tighter y_H requirement and the inclusion of all production modes. The STXS framework was developed to maximize the sensitivity of the Higgs boson measurements, while at the same time minimizing the theory dependence of their determination. The same p_{T} boundaries are used for the V+jets production cross section measurements in the VRL, and for Z+jets production in the SR, and are used to validate the method. The fiducial and STXS volumes for these are defined by requirements on the generator truth vector boson transverse momentum p_{T}^{V}. The summary of the fiducial and differential region volumes is given in Table 6.1.

Table 6.1: Summary of the fiducial and STXS volumes used to determine the signal events considered for the signal strength measurement [10].

Volume	p_{T}^{H}	$\lvert y_H \rvert$
Fiducial	> 450	< 2
STXS	$300 - 450$, $450 - 650$, $650 - 1000$, > 1000	< 2

6.1 Inclusive Region

The inclusive region is the signal region containing candidate jets with $p_{\mathrm{T}} > 250$ GeV. The extraction of the Higgs boson signal strength in the inclusive region yields a value of $\mu_H = 0.8 \pm 3.2$ when combining the SRL, SRS and $\mathrm{CR}_{t\bar{t}}$ information. The breakdown of the uncertainty was found to be $\pm 3.2\,(\mathrm{total}) = \pm 3\,(\mathrm{stat}) \pm 1.1\,(\mathrm{syst})$. The measurement is dominated by the statistical uncertainty (size of the data sample) which limits the sensitivity of the signal. The observed signal strength corresponds to a sensitivity of 0.29σ (0.36σ was expected). The largest contributions to the systematic uncertainty are from the jet mass resolution (JMR) and the jet mass scale (JES). For the $t\bar{t}$ contribution, the value was found to be $\mu_{t\bar{t}} = 0.80 \pm 0.06$. The poor modeling seen for $t\bar{t}$ agrees with previous published results where the boosted top-quark pair differential cross-section in the l+jets channel was measured [131]. The Z+jets process had a signal strength value of $\mu_Z = 1.29 \pm 0.22$. These results are summarized in Table 6.2. The yields in the three regions pertinent to the Higgs boson signal extraction are presented in Table 6.3.



(a) (b)

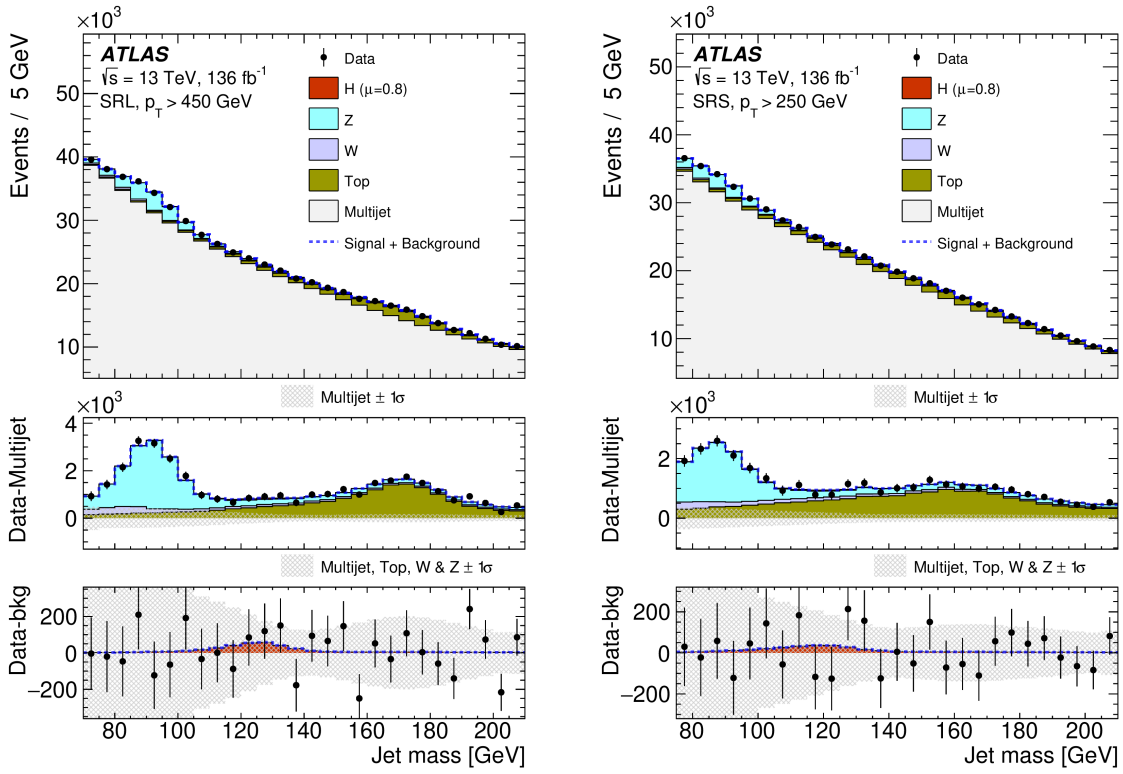
Figure 6.1: The ratios of the measured fiducial phase-space absolute differential cross-sections to the predictions obtained to (a) the Powheg+Pythia8 MC generator and to (b) the NNLO predicitions, in the resolved and boosted topologies as a function of the top quark p_{T} [131].

Table 6.2: Expected and observed values of the signal strengths for the H, Z, and $t\bar{t}$ components in the inclusive fit [10].

Result	μ_H	μ_Z	$\mu_{t\bar{t}}$
Expected	1.0 ± 3.2	1.00 ± 0.17	1.00 ± 0.07
Observed	0.8 ± 3.2	1.29 ± 0.22	0.80 ± 0.06

Table 6.3: Event yields and associated uncertainties after the global likelihood fit in the inclusive region [10].

Process	SRL	SRS	$\mathrm{CR}_{t\bar{t}}$
Multijet	$590\,700 \pm 4200$	$529\,300 \pm 3500$	–
$Z+$ jets	$16\,100 \pm 2800$	$12\,000 \pm 2100$	–
$W+$ jets	3050 ± 720	2510 ± 500	–
Top	$16\,300 \pm 1900$	$15\,900 \pm 2000$	3737 ± 68
$W(l\nu)$	–	–	53 ± 16
H	400 ± 1500	300 ± 1300	–
Total	$626\,530 \pm 820$	$560\,090 \pm 770$	3790 ± 66
Data	626 532	560 083	3791



(a) (b)

Figure 6.2: Post-fit jet mass distributions for the various components in the inclusive SRL (left) and SRS (right) regions. In the middle panels the shaded areas indicate the 68% CL for the multijet background from the fitted parameters and normalizations of the exponentiated polynomials. In the lower panels the shaded areas indicate the 68% CL for all background processes [10].

6.2 Fiducial Region

In the fiducial region, the Higgs boson yield is determined using the fiducial volume defined by the Higgs boson transverse momentum ($p_{\mathrm{T}}^{H} > 450$) GeV and rapidity ($|y_H| < 2.0$). The transverse momentum cut was chosen to ensure an unbiased truth spectrum due to the trigger turn on. Therefore, this measurement doesn't include the SRS region below 450 GeV. The signal acceptance times efficiency in the fiducial volume is presented in Table 6.4.

Table 6.4: Signal acceptance times efficiency within the fiducial volume used in the fiducial region [10].

Process	$p_{\mathrm{T}}^{H} > 450$ GeV $\lvert y_H \rvert < 2$
All	0.24
ggF	0.26
VBF	0.22
VH	0.27
$t\bar{t}H$	0.20

The signal outside the fiducial region is set to the SM value within uncertainties. Therefore the fit considers two Higgs boson mass templates in each SR. The component from the fiducial volume accounts for more than 80% of the Higgs boson signal. The component from outside the fiducial volume has a broader mass spectrum shifted to higher values. This procedure was tested with $W \to qq'$ and $Z \to q\bar{q}$ in the VR and with $Z \to b\bar{b}$ in the SR. By fixing the Higgs signal to the SM values, in both in and out the fiducial region, the signal strength for V+jets was found to be $\mu_V = 1.01 \pm 0.09$. Similarly, for Z events in the SR the signal strength was found to be $\mu_Z = 1.35 \pm 0.25$, both being in agreement with the SM.

For the Higgs boson signal strength, the likelihood fit yields a value of $\mu_H = -0.1 \pm 3.5$. The results are summarized in Table 6.5. Post-fit mass distributions are shown in Figure 6.3. No signal of the Higgs boson is shown given that the signal strength was found to be

below 0. The SM prediction for the Higgs boson production with $p_{\mathrm{T}}^{H} > 450$ GeV is 18.4 fb. Our measurement corresponds to a 95% CL upper limit on the observed (expected) Higgs boson production cross section of

$$\sigma_H(p_{\mathrm{T}}^{H} > 450\,\mathrm{GeV}) < 115\,(128)\,\mathrm{fb}. \tag{6.1}$$

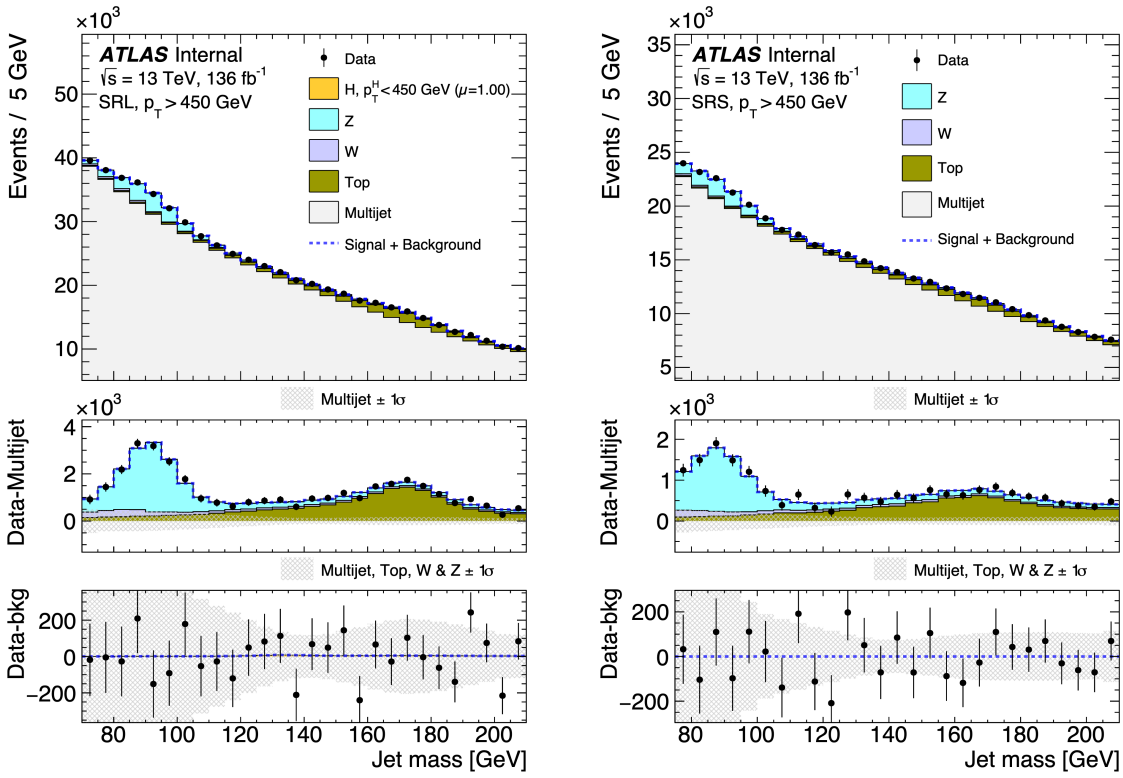
Table 6.5: Expected and observed values of the signal strengths for the H, Z and $t\bar{t}$ components in the fiducial fits [10].

Result	μ_H	μ_Z	$\mu_{t\bar{t}}$
Expected	1.0 ± 3.4	1.00 ± 0.18	1.00 ± 0.08
Observed	-0.1 ± 3.5	1.30 ± 0.22	0.75 ± 0.06

The statistical uncertainty is the largest contributor to the total uncertainty of the signal strength, with the systematic uncertainty being somewhat smaller. The largest component of the systematic uncertainty, with almost a 80% contribution, is the jet systematics driven by the jet mass scale (JMS) effects. The JMS uncertainty comes from both background (V+jets and $t\bar{t}$ contribute 50%) and the reconstructed Higgs bosons (which contributes the other 50%). The breakdown is shown in Table 6.6.

Table 6.6: Contributions to the systematic uncertainties for the measurement of the fiducial volume signal strength [10].

Uncertainty Contribution	$p_{\mathrm{T}}^{H} > 450$ GeV
Total	3.5
Statistical	2.6
Systematic	2.3
Jet systematic uncertainties	2.2
Modeling and theory systs.	0.8
Flavor-tagging systs.	0.2



(a) (b)

Figure 6.3: Post-fit jet mass distributions for the various components in the fiducial SRL (left) and SRS (right) regions. In the middle panels the shaded areas indicate the 68% CL for the multijet background from the fitted parameters and normalizations of the exponentiated polynomials. In the lower panels the shaded areas indicate the 68% CL for all background processes [10].

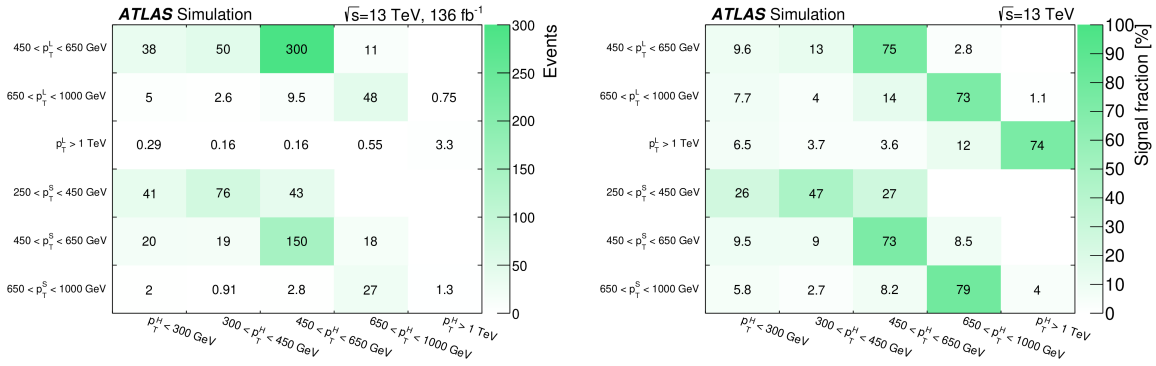
6.3 Differential Regions

The differential region measurements aims to extract the Higgs boson transverse momentum spectrum in four p_{T}^{H} volumes. These are based o the STXS template and consist of the p_{T}^{H} volumes with 300-450 GeV, 450-650 GeV, 650-1000 GeV and above 1 TeV. The same procedure established for the Higgs boson measurement for the fiducial region is employed. The Higgs boson mass template for each p_{T}^{H} volume is used within each p_{T} region in the global likelihood. Only the SRL and $\mathrm{CR}_{t\bar{t}}$ regions are included for measurements above 1 TeV, given that the SRS expected sensitivity in this region is marginal. Outside the volumes ($p_{\mathrm{T}}^{H} < 300$ GeV) the components were fixed to their SM expectations. The signal acceptance times the efficiency for the STXS volumes is shown in Table 6.7. The expected yield and percentage contributions of the Higgs boson subprocesses is shown in Figure 6.4.

Table 6.7: Signal acceptances times efficiency for the STXS volumes in the differential measurement [10].

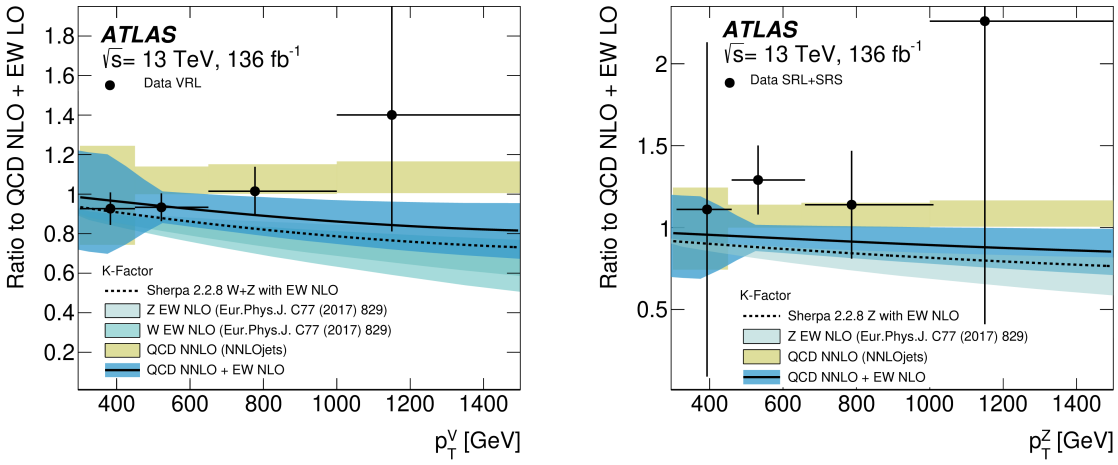
Process	$300 < p_{\mathrm{T}}^{H} < 450$ GeV	$450 < p_{\mathrm{T}}^{H} < 650$ GeV	$650 < p_{\mathrm{T}}^{H} < 1000$ GeV	$p_{\mathrm{T}}^{H} > 1$ TeV
All	1.3×10^{-2}	0.23	0.31	0.23
ggF	0.7×10^{-2}	0.25	0.35	0.28
VBF	0.4×10^{-2}	0.21	0.32	0.25
VH	1.7×10^{-2}	0.26	0.30	0.20
$t\bar{t}H$	4.7×10^{-2}	0.19	0.24	0.19

Similar to what was done to the fiducial measurement, the signal determination method was tested with $W \to qq'$ and $Z \to q\bar{q}$ in the VR, and with $Z \to b\bar{b}$ in the SR. The VRL is divided into 5 slices with the fit being performed independently in each slice. The results are then averaged. The Z fit is performed in the SRL, SRS, $\mathrm{CR}_{\mathrm{t\bar{t}}}$ regions with the Higgs boson contribution fixed to the SM prediction. The results of the differential fit signal strengths for V+jets in the VRL and Z+jets in the SR are shown in Figure 6.5.



(a) (b)

Figure 6.4: For each of the p_{T}^{H} differential volumes (x-axis), the expected signal event yield for all Higgs boson events (left) and the fraction of signal in percent (right) in each reconstructed jet p_{T} region (y-axis) is shown. The leading jet p_{T} in the SRL is denoted by p_{T}^{L} and the subleading jet p_{T} in the SRS is denoted by p_{T}^{S} [10].



(a) (b)

Figure 6.5: Comparison of differential fit signal strengths for (a) V+jets in the VRL and (b) Z+jets in the SR. The signal strength within the STXS volumes is calculated relative to the prediction at NLO QCD and LO EW accuracy. They are compared with the NLO EW correction provided by SHERPA, the NNLO QCD correction provided by the NNLOJET group, and their product. The points are located at the weighted center of the bin considering the underlying p_{T}^{V} or p_{T}^{H} spectrum [10].

The Higgs boson signal strengths in the STXS volumes is extracted by fitting simultaneously the ten differential SR and CR regions defined in Tables 5.5 and 6.1. The results are summarized in Tables 6.8 and 6.9. The four Higgs boson signal strengths are compatible with the SM and have a p-value of 0.53.

Table 6.8: Expected and observed values of the signal strengths for the H component in the differential fits [10].

p_{T}^{H} [GeV]	μ_H	
	Exp.	Obs.
$300-450$	1.0 ± 18	-6 ± 18
$450-600$	1.0 ± 3.3	-3 ± 5
$650-1000$	1.0 ± 6	5 ± 7
>1000	1.0 ± 30	18 ± 32

Table 6.9: Expected and observed values of the signal strengths for the Z and $t\bar{t}$ components in the differential fits [10].

Jet p_{T} [GeV]	μ_Z		$\mu_{t\bar{t}}$	
	Exp.	Obs.	Exp.	Obs.
$300-450$	1.0 ± 1.1	1.8 ± 1.1	1.0 ± 0.07	0.85 ± 0.06
$450-600$	1.0 ± 0.17	1.28 ± 0.22	1.0 ± 0.07	0.76 ± 0.06
$650-1000$	1.0 ± 0.33	1.4 ± 0.4	1.0 ± 0.09	0.74 ± 0.08
>1000	1.0 ± 1.6	2.4 ± 1.7	1.0 ± 0.22	0.57 ± 0.18

The Higgs boson production cross section for $p_{\mathrm{T}}^{H} > 1$ TeV was found to be

$$\sigma_H(p_{\mathrm{T}}^{H} > 1\,\mathrm{TeV}) = 2.3 \pm 3.9\,(\mathrm{stat}) \pm 1.3\,(\mathrm{syst}) \pm 0.5\,(\mathrm{theory})\,\mathrm{fb}. \tag{6.2}$$

The SM prediction for this quantity is 0.13 fb. Because the sensitivity was low, upper limits were calculated. The 95% CL upper limits on the Higgs boson differential production cross

section were found to be

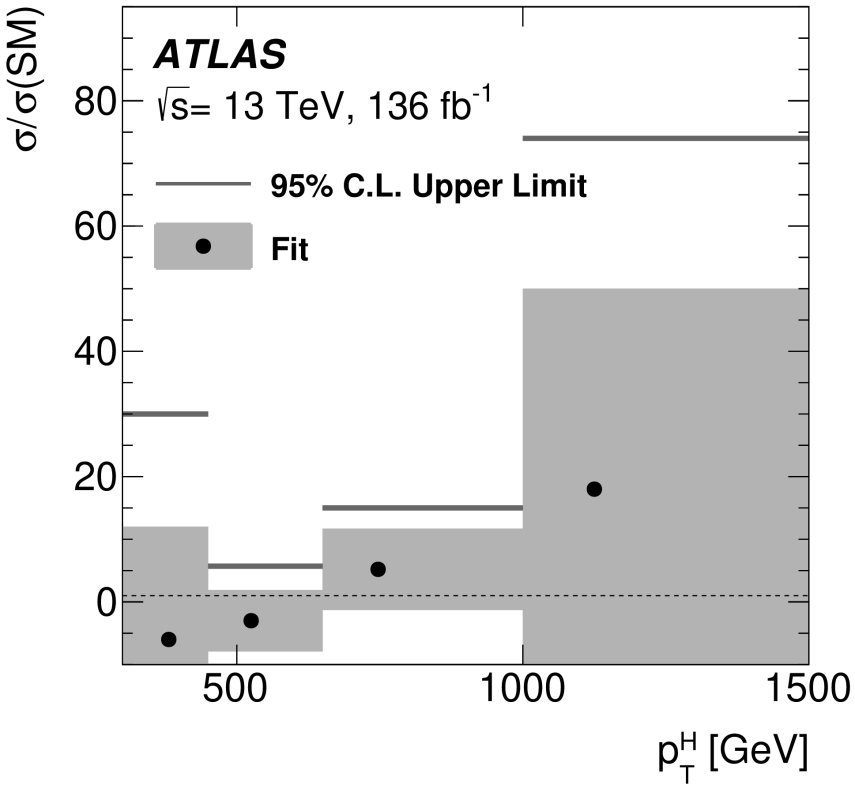
$$\begin{aligned}
\sigma_H(300 < p_{\mathrm{T}}^H < 450\,\mathrm{GeV}) &< 2.9\,(3.1)\,\mathrm{pb},\\
\sigma_H(450 < p_{\mathrm{T}}^H < 650\,\mathrm{GeV}) &< 89\,(102)\,\mathrm{fb},\\
\sigma_H(650 < p_{\mathrm{T}}^H < 1000\,\mathrm{GeV}) &< 39\,(34)\,\mathrm{fb},\\
\sigma_H(p_{\mathrm{T}}^H > 1000\,\mathrm{GeV}) &< 9.6\,(7.4)\,\mathrm{fb}.
\end{aligned} \tag{6.3}$$

These results are shown in Figure 6.6. As for the first two results of this analysis, the largest source of uncertainty is of statistical nature given that the sample is small. Table 6.10 summarizes the breakdown of the uncertainties associated with the measurement. The largest contribution to the systematic uncertainties is the JMS uncertainty.

Table 6.10: Contributions to the systematic uncertainties for the differential measurements of the signal strength [10].

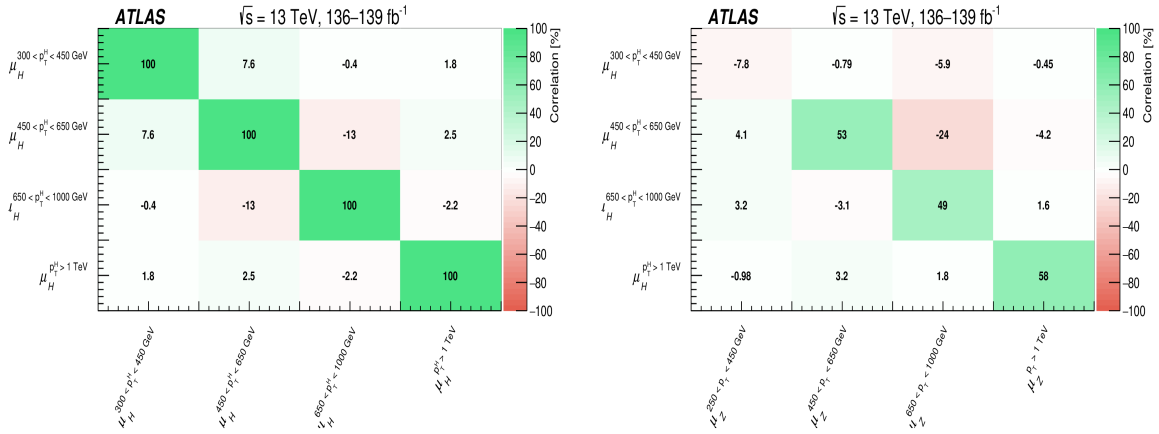
Uncertainty Contribution	$300 < p_{\mathrm{T}}^H < 450$ GeV	$450 < p_{\mathrm{T}}^H < 650$ GeV	$650 < p_{\mathrm{T}}^H < 1000$ GeV	$p_{\mathrm{T}}^H > 1$ TeV
Total	18	5.0	6.5	32
Statistical	16	3.0	5.5	30
Systematic	7	3.9	3.4	10
Jet systematic uncertainties	6	3.8	3.4	9.5
Modeling and theory systs.	4	0.7	0.7	2
Flavor-tagging systs.	0.2	0.4	0.4	2

The correlations between the differential Higgs boson signal strength measurements in p_{T}^H bins were found to be small. This implies a low number of events migrating from the analysis bins and the STXS truth Higgs bins. Figure 6.7 shows the correlations of the μ_H and μ_Z measurements. A post-fit mass distribution of all the components in the differential leading jet signal region is shown in Figure 6.8.



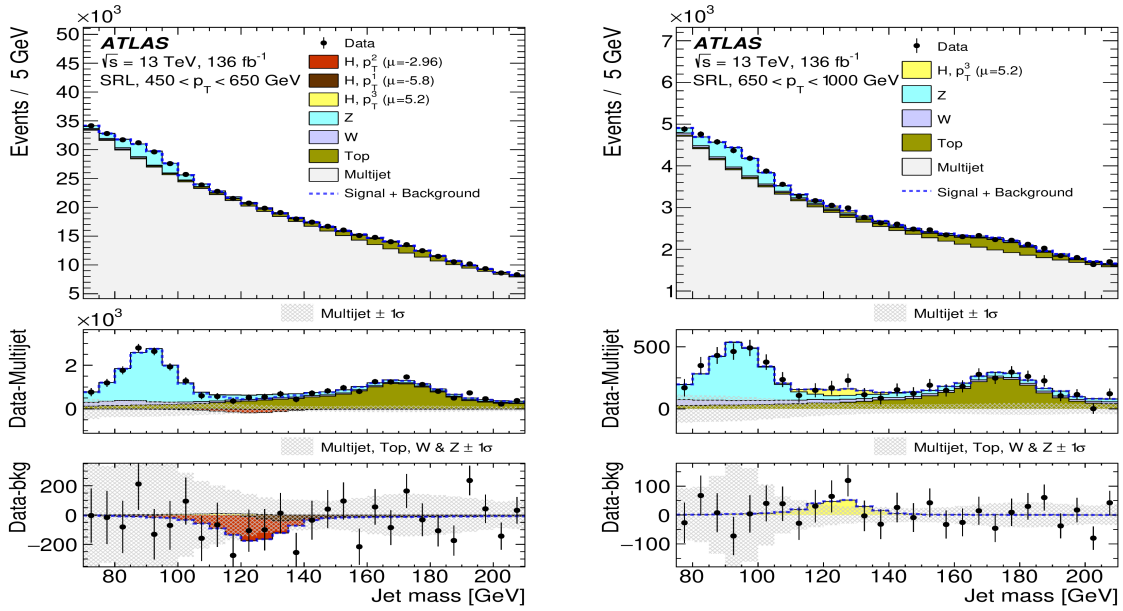
(a)

Figure 6.6: Summary of the STXS volume signal strengths measured using the differential signal regions. Within the same kinematic regimes, measurements of the $Z \to b\bar{b}$ process agree with the Standard Model predictions, validating the methods. The points are located at the weighted center of the bin considering the underlying p_{T}^{H} spectrum [10].

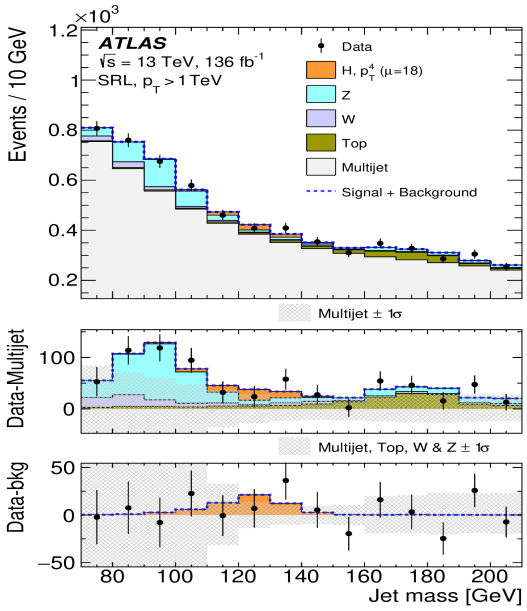


(a) (b)

Figure 6.7: Correlations among the four Higgs boson signal strengths, and between the four Higgs boson and Z+jets signal strengths. The Higgs boson signal strengths μ_H are labeled with the corresponding p_{T}^{H} range as a superscript. The Z+jets signal strengths μ_Z are labeled with the corresponding jet p_{T} range as a superscript [10].



(a) (b)



(c)

Figure 6.8: Post-fit jet mass distributions of the various components in the differential leading-jet signal region defined by the selected candidate jet with (a) $450 < p_{\mathrm{T}} < 650$ GeV, (b) $650 < p_{\mathrm{T}} < 1000$ GeV, and (c) $p_{\mathrm{T}} > 1000$ GeV shown in wider 10 GeV bins [10].

Chapter 7

Unified Flow Objects

7.1 Introduction

From the results presented in Chapter 6, it is evident that the optimization of large-radius jet definitions could result in considerable gains for our measurement in terms of precision. Approximately 90% of the systematic uncertainties result from the jet definitions, in particular the jet mass resolution and jet mass scale. In the analysis presented, large-R jet reconstruction was based on topological cluster inputs reconstructed using calorimeter-based energy measurements. A trimming procedure was performed and a combined mass scheme was employed. Even though a good energy resolution is achieved, in the high p_{T} regime, the resulting showers are so collimated that the calorimeter's granularity is not sufficient to spatially resolve individual particles in the jet. For this reason, the use of jet substructure variables (JSS) is limited with these type of jet definitions. As a step to reduce these limitations during Run 2, particle-flow (PFlow) [134] algorithms were implemented to improve performance at low p_{T}. For high p_T, on the other hand, Track-CaloClusters (TCCs) [135] were designed in order to reconstruct jet substructure (JSS) variables. A new type of jet input, called "unified flow object" (UFO) [11] was then developed using both particle-flow objects (PFOs) and TCCs. This object combines calorimeter and inner detector based signals in order to achieve optimal performance across a wide kinematic range. This new

definition, combined with better pile-up mitigation techniques, such as Constituent Subtraction (CS) [136] and SoftKiller (SK) [137], as well as grooming algorithms, such as Soft-Drop [138], motivated the re-optimization of the large-R jet definitions used by ATLAS. The ATLAS Jet Tagging and Scale Factor Derivation group has been developing UFO reconstructed large-R jet dedicated taggers for hadronically decaying boosted objects. In this chapter we explore the reconstruction algorithm behind UFO jets, the different dedicated taggers developed and the main ideas behind their development, such as jet substructure variables and machine learning. The strategy used to extrapolate their scale factors to higher momenta and the manner in which we estimate the uncertainties associated with the extrapolation procedure, will be presented. The chapter contains final results of the extrapolation uncertainties for the already calibrated UFO taggers supported by the ATLAS Collaboration. Finally, studies regarding a multiclass tagger (MCT) that includes Higgs boson identification will also be explored. These projects include work that started as the author's "qualification task" (authorship project) and the subsequent collaboration with the ATLAS Jet Tagging and Scale Factor Derivation group.

Particle-flow Objects

Particle-flow (PFlow) [134] reconstruction combines both track and calorimeter based measurements. Particle-flow objects (PFOs) themselves improve pile-up stability relative to jets reconstructed from topo-clusters. The PFlow algorithm matches each selected track to a single topo-cluster. For each track/topo-cluster system the probability that the particle's energy was deposited in more than one topo-cluster is evaluated. Then, the expected energy deposited in the calorimeter by the particle that produced the track is subtracted. Any topo-cluster that is not matched to a track is considered to be produced by a neutral particle

and is left unmodified. The subtraction is gradually disabled for tracks with $p_{\mathrm{T}} < 100$ GeV if the energy deposited (E_{clus}) in a cone of size $\Delta R = 0.15$ by the extrapolated track satisfies

$$\frac{E_{\mathrm{clus}} - \langle E_{\mathrm{dep}} \rangle}{\sigma(E_{\mathrm{dep}})} > 33.2 \times \log_{10}(40\,\mathrm{GeV}/p_{\mathrm{T}}^{\mathrm{trk}}), \tag{7.1}$$

where E_{dep} is the expected energy deposition. Any charged PF0 that is not matched to the primary vertex is removed to reduce the contribution from pile-up. This procedure is known as "Charged Hadron Subtraction" [139].

Track-CaloClusters

Track-CaloClusters (TCCs) [135] are optimized to perform jet substructure reconstruction for very high p_{T} jets. TCCs use energy scale information from topo-clusters and angular information from tracks. The algorithm matches a loose track in a particular event to topo-clusters that have been calibrated to the local hadronic scale. When a track is matched to a topo-cluster, the p_{T} is determined using the p_{T} associated with the topo-cluster while its angular coordinates (η,ϕ) are taken from the track. If a topo-cluster is not matched to any track then the TCC is created only using the topo-clusters 4-vector directly. Similarly, for an independent track that is not matched to any topo-cluster, the track information is used directly to create the TCC. If multiple tracks are matched to a single topo-cluster, each TCC object is given a fraction of the total p_{T} of the topocluster. The momentum fraction is determined using the ratios of momenta of the matched tracks. Any unmatched topo-clusters are included as unmodified neutral objects.

Pile-Up Mitigation and Grooming Algorithms

Before jet reconstruction, pile-up mitigation techniques are employed. For topo-clusters, the techniques are applied to the entire set of inputs. On the other hand, for charged PFOs, only the CHS method is employed, while neutral PFOs and TCCs are subject to the same preprocessing techniques used for topo-clusters. During reconstruction, grooming algorithms are applied to reduce contamination from soft radiation originating from the underlying event (UE), pile-up and initial state radiation (ISR).

Constituent Subtraction

Constituent Subtraction (CS) [136] is a method that performs area subtraction on jet input objects. Therefore, it is a local subtraction of pile-up at the level of individual jet constituents. Each input is defined using ghost association, a process where massless particles (ghosts) with low momentum are overlaid uniformly in the event. Ghosts have to satisfy

$$p_{\mathrm{T}}^{g} = A_g \times \rho, \tag{7.2}$$

where A_g is the area of the ghosts and p_{T}^{g} is the expected contribution from pile-up radiation in a small angular area. The pile-up energy density ρ is assumed to have a weak dependence in azimuthal angle ϕ and rapidity y. ρ is estimated as the median of p_{T}/A distribution of $R = 0.4$ k_t jets in each event.

The distance $\Delta R_{i,k}$ between a cluster i and ghost k is given by

$$\Delta R_{i,k} = \sqrt{(\eta_i - \eta_k)^2 + (\phi_i - \phi_k)^2}. \tag{7.3}$$

The algorithm proceeds iteratively through each cluster-ghost pair, after sorting in order of ascending $\Delta R_{i,k}$, by modifying the p_T of each pair as follows

$$\begin{aligned} \text{if} \quad p_{\mathrm{T},i} \geq p_{\mathrm{T},k}: \quad & p_{\mathrm{T},i} \to p_{\mathrm{T},i} - p_{\mathrm{T},k}, \\ & p_{\mathrm{T},k} \to 0; \\ \text{otherwise}: \quad & p_{\mathrm{T},k} \to p_{\mathrm{T},k} - p_{\mathrm{T},i}, \\ & p_{\mathrm{T},i} \to 0. \end{aligned} \tag{7.4}$$

The procedure continues until $\Delta R_{i,k} > \Delta R_{\max}$, where $\Delta R_{\max}$ is a free parameter. The particles with zero transverse momentum are then discarded.

In the original formulation of this procedure, a similar modification was performed for the mass, but this is ignored in the latest ATLAS implementation given that all neutral ATLAS jet inputs are defined to be massless.

SoftKiller

SoftKiller (SK) [137] is an algorithm that applies a p_{T} cut to input objects. The p_{T} cutoff, $p_{\mathrm{T}}^{\mathrm{cut}}$, is chosen such that the value of ρ is approximately zero. The event is broken into square patches in the rapidity-azimuth plane. The parameter ρ is the event-wide estimate of the transverse momentum density in an area patch

$$\rho = \underset{i \in \text{patches}}{\text{median}} \left\{ \frac{p_{\mathrm{T},i}}{A_i} \right\}. \tag{7.5}$$

The cut is determined such that when the cut is applied, half of the grid spaces are empty.

Computationally, this is given by the next formula

$$p_{\mathrm{T}}^{\mathrm{cut}} = \underset{i \in \mathrm{patches}}{\mathrm{median}} \{p_{\mathrm{T},i}^{\mathrm{max}}\}, \tag{7.6}$$

where $p_{\mathrm{T},i}^{\mathrm{max}}$ is the p_{T} of the hardest particle in patch i. This implies that half the patches will contain only particles that have $p_{\mathrm{T}} < p_{\mathrm{T}}^{\mathrm{cut}}$. Therefore, after applying the cut, the value of ρ will be zero.

The best performance is achieved when SK is applied after the CS algorithm.

Soft-Drop

Soft-Drop (SD) [138] is a grooming technique that removes soft and wide-angle radiation from jets. For this procedure, the constituents of the large-R jets are reclustered using the C/A algorithm. Using the angle-ordered jet clustering history, determined from the C/A algorithm, the clustering sequence is traversed from the widest-angled radiation iterating to the smallest-angle radiation. The condition

$$\frac{\min(p_{\mathrm{T},1}, p_{\mathrm{T},2})}{p_{\mathrm{T},1} + p_{\mathrm{T},2}} < z_{\mathrm{cut}} \left(\frac{\Delta R_{12}}{R} \right)^{\beta}, \tag{7.7}$$

is evaluated for each splitting, where 1 and 2 represent the harder and softer branches of the splitting respectively. The parameters z_{cut} and β dictate how aggresive the removal of soft and wide-angle radiation is and its dependence on the distance parameter. When a splitting fails the condition, the lower p_{T} branch is removed. If the condition is satisfied, the process ends and the consituents remaining form the groomed jet. The use of the SD grooming algorithm allows the calculation of certain jet substructure observables to beyond leading-logarithm (LL) accuracy, while other trimming algorithms do not [140].

Unified Flow Objects

The fact that no single jet definition is optimal according to all metrics motivated the development of a new jet input that combines all the desirable aspects of both particle-flow objects (PFOs) and Track-CaloClusters (TCCs) reconstruction. TCCs improve tagging performance at high p_{T} but their performance is worse than with baseline trimmed topo-cluster based jets at low p_{T}, as they are more sensitive to pile-up. PFOs on the other hand can improve on the baseline definition for the entire p_{T} range but their tagging performance is worse than with TCCs at high p_{T}. Combining both approaches by defining a new jet input object, called Unified Flow Object (UFO) [11], we can achieve optimal performance across the full kinematic range.

Figure 7.1 contains an illustration of how the UFO reconstruction algorithm is performed. The process starts by the application of the standard particle-flow (PFlow) algorithm. Any charged PFO that is matched to a pile-up vertex is removed. The remaining PFOs are then divided into three categories: neutral PFOs, charged PFOs that were used to subtract energy from a topo-cluster, and the charged PFOs that were not used for energy subtraction. At this point jet input pile-up mitigation algorithms (i.e. CS+SK) are applied. Then a modified TCC splitting algorithm is applied. Tracks that have been used for the PFlow subtraction are not considered as they have already been subtracted from the energy of the topo-clusters. The TCC algorithm proceeds with the remaining collection of tracks to split neutral and unsubtracted charged PFOs.

The use of UFOs improves jet mass resolution (JMR) relative to topo-cluster-based jets, by 40% for high p_{T} hadronically decaying W bosons and by 26% for hadronically decaying high p_{T} top quarks. Figure 7.2 shows the JMR relative performance of different UFO

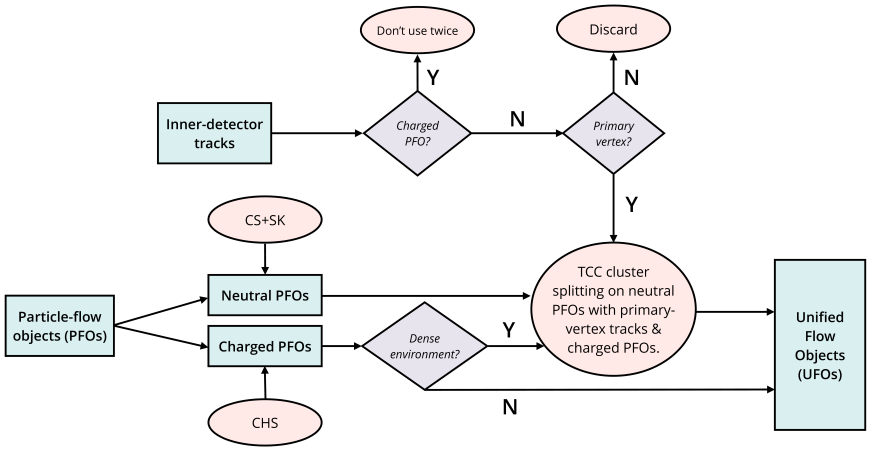


Figure 7.1: Illustration of the UFO reconstruction algorithm. The procedure starts with the identification of particle-flow objects (PFOs) and inner-detector tracks [11].

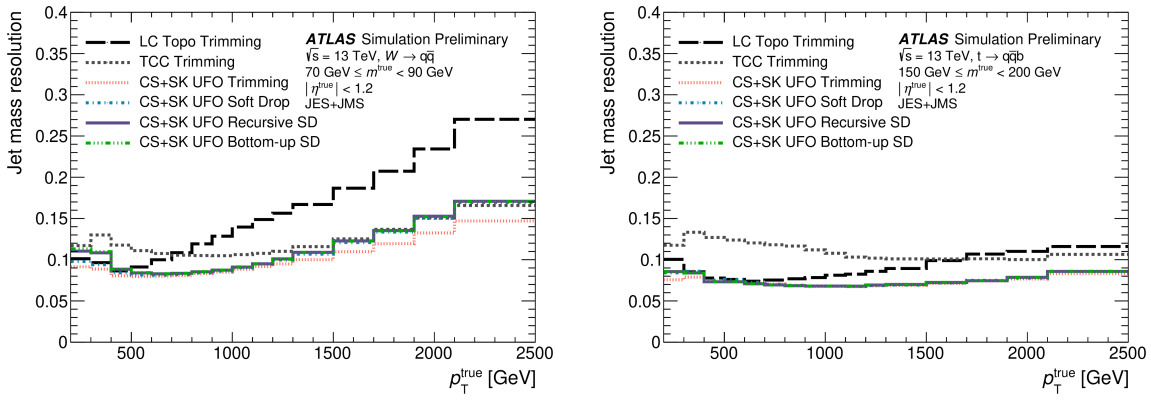
definitions compared to the ATLAS baseline jet definitions.

For tagging, UFOs bring significant improvements over the usual topo-cluster and TCC definitions. Some studies show an increase of 120% in background rejection at a fixed signal effifiency of 50% at high p_{T}. One example of the performance of UFOs for tagging is shown in Figure 7.3. In this example, using UFOs for hadronically decaying top quarks at high p_{T}, improves the background rejection by 135% when compared with the baseline jet definitions.

The p_{T} resolution is degraded for large-R jets coming from UFOs compared to the baseline topo-cluster and TCC definitions, but given the improvements in jet mass resolution and jet tagging at high p_{T}, it is worthwhile to proceed with the development and study of these jet definitions and to define taggers for future ATLAS analyses.

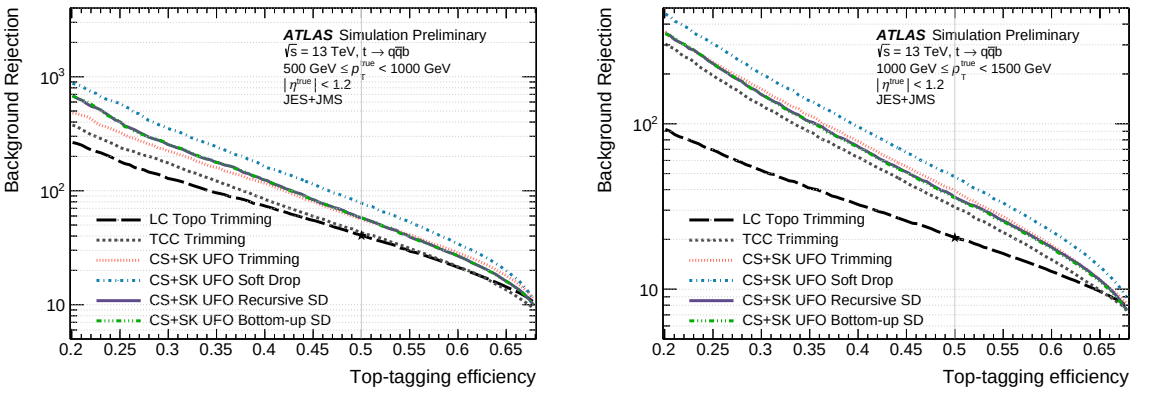
7.2 Jet Substructure Variables

The jet substructure techniques [141] can be summarized as a set of tools to exploit the radiation pattern inside hadronic jets. These correlations are quantified by a set of variables



(a) (b)

Figure 7.2: Jet mass resolution for (a) W boson jets and (b) top quark jets as a function of p_{T}. The relative performance of the studied UFO definitions compared to the current ATLAS baseline topo-cluster and TCC jets is shown [11].



(a) (b)

Figure 7.3: Background rejection as a function of signal efficiency for a tagger using the jet mass and τ_{32} for top quark jets at (a) low p_{T} and (b) high p_{T}. The relative performance of different UFO definitions are compared with the current ATLAS baseline topo-cluster and TCC jets [11].

which in principle are infrared and collinear (IRC) safe [142], i.e. they are insensitive to infinitesimally soft or collinear emissions, the presence of which presents difficulties for higher order QCD predictions.

Angularities

The generalized angularities [143] are a family of jet shapes defined as

$$a_{\beta}^{\kappa} = \sum_{i \in J} z_i^{\kappa} \left(\frac{\Delta R_{i,J}}{R} \right)^{\beta}, \tag{7.8}$$

where z_i is the jet transverse momentum fraction carried by the constituent i of jet J, and $\Delta R_{i,J}$ its distance to the jet axis. Only angularities with $\kappa = 1$ are IRC safe. Angularities can be seen as a measure of QCD radiation around the jet axis.

N-subjettiness

N-subjettiness [144] is an angularity-type observable. Within a jet, N-subjettiness identifies N subjet axes, calculates the jet thrust about each and sums all of them together. For a jet J, with transverse momentum $p_{\mathrm{T}J}$ and particles i each with $p_{\mathrm{T}i}$, the N-subjettiness is given by

$$\tau_N^{(\beta)} = \frac{1}{p_{\mathrm{T}J}} \sum_{i \in J} p_{\mathrm{T}i} \min\{R_{i,1}, R_{i,2}, \cdots R_{i,N}\}^{\beta}, \tag{7.9}$$

where $R_{i,n}$ is the distance between particle i and the closest subjet axis in the η-ϕ plane. The angular exponent β controls the sensitivity to collinear radiation. By taking the ratio of multiple N-subjettiness variables, new dimensionless quantities can be derived, which aid

in the discrimination of multi-pronged objects. For example, the ratio

$$\tau_{N,N-1}^{(\beta)} = \frac{\tau_N^{(\beta)}}{\tau_{N-1}^{(\beta)}}, \tag{7.10}$$

can be used to identify the presence of N subjets within a jet.

Generalized Energy Correlation Functions

The N-point energy correlation function (ECF) [145] is defined as

$$\mathrm{ECF}(n,\beta) = \sum_{i_1<i_2<\cdots<i_n\in J} \left(\prod_{a=1}^{n} p_{\mathrm{T}i_a}\right)\left(\prod_{b=1}^{n-1}\prod_{c=b+1}^{n} R_{i_b i_c}\right)^{\beta}, \tag{7.11}$$

where n denotes the number of particles to be correlated, R_{ij} is the distance between particle i and particle j in the η-ϕ plane and β is used to adjust the weighting of the distance between particles. A dimensionless definition for the energy correlation functions can be constructed using a ratio

$$e_n^{\beta} \equiv \frac{\mathrm{ECF}(n,\beta)}{\mathrm{ECF}(1,\beta)^n}. \tag{7.12}$$

Combinations of these can be used to define other observables that have been found to be very useful for identifying multi-pronged structure. Some examples used for 2-prong discrimination in boosted jets include D_2 and C_2, which are defined as

$$C_2^{(\beta)} = \frac{e_3^{(\beta)}}{(e_2^{(\beta)})^2} \quad \text{and} \quad D_2^{(\beta)} = \frac{e_3^{(\beta)}}{(e_3^{(\beta)})^3}. \tag{7.13}$$

These functions probe multiple angular scales simultaneously. To isolate different physics effects, a more general definition that identifies one scale at a time was developed. These

are the generalized energy correlation functions [146]

$$ {}_{v}e_{n}^{(\beta)} = \sum_{i_1<i_2<\cdots<i_n\leq n_J} z_{i_1} z_{i_2} \cdots z_{i_n} \prod_{m=1}^{v} \min_{s<t\in\{i_1,i_2,\cdots,i_n\}}^{(m)} \{R_{st}^{(\beta)}\}, \tag{7.14} $$

where $z_i \equiv p_{\mathrm{T}i}/\sum_{j\in J} p_{\mathrm{T}j}$ and $\min^{(m)}$ denotes the m-th smallest element in the list. The subscript v represents the number of pair-wise distances entering the product. Using these definitions new substructure discriminants have been defined and found to be useful for boosted top tagging. Some examples of these are L_2 and L_3 defined as

$$ L_2 = \frac{{}_{3}e_{3}^{\beta=1}}{({}_{1}e_{2}^{\beta=2})^{3/2}} \quad \text{and} \quad L_3 = \frac{{}_{1}e_{2}^{\beta=1}}{({}_{3}e_{3}^{\beta=1})^{1/3}}. \tag{7.15} $$

Jet Charge

The jet charge is an energy weighted sum of the electric charges of the hadrons in a jet. For a jet with energy E_J it is defined as

$$ Q_W = \sum_{i\in J} \left(\frac{E_i}{E_J}\right)^W q_i, \tag{7.16} $$

where E_i is the energy of particle i and charge q_i. The parameter W controls how strong is the weighting for each energy fraction.

Planar Flow

Planar flow [147] is a jet shape observable that distinguishes planar from linear configurations. To define it, we first construct a matrix I_w for a given jet as

$$I_w^{kl} = \frac{1}{m_J} \sum_i \frac{p_{i,k}}{w_i} \frac{p_{i,l}}{w_i}, \tag{7.17}$$

where m_J is the jet mass, w_i is the energy of particle i in the jet, and $p_{i,k}$ is the k^{th} component of its transverse momentum relative to the axis of the jet's momentum. Then, the planar flow $\mathcal{P}$ is defined as

$$\mathcal{P} = \frac{4 \det(I_w)}{\text{tr}(I_w)^2}. \tag{7.18}$$

7.3 Machine Learning

Machine learning is an umbrella term to describe a framework that automates statistical models, through a set of algorithms, to make better predictions. More formally, "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E." [148], as defined by Tom Mitchell. Therefore, machine learning problems can be described as an optimization process where a performance measure is maximized. In practice, this is achieved by the minimization of a *loss function*, which measures the model's prediction error. The experience part of the definition comes from the fact that you need data to train your model to predict the outcome of interest. One pass of a training dataset through the learning algorithm in the training process is referred to as an *epoch*. The number of epochs is considered a hyperparameter that can be tuned, as once during an epoch the internal

model's parameters are updated as part of the loss function minimization process.

Three categories are usually used to describe machine learning approaches, depending on the nature of the task and structure of the dataset. These are, supervised learning, unsupervised learning and reinforcement learning. For supervised learning, the computer is shown examples and the desired outputs with the goal of creating a map between them. Unsupervised learning involves leaving the computer on its own to find structure between the inputs. For reinforcement learning on the other hand, the computer is provided with feedback while it is navigating the problem space.

Tagging jets is a classification problem. Classification is a suitable candidate to use the supervised learning approach. A supervised learning problem neccesitates the inclusion of 'labels' associated with the data sample that the algorithm has to learn from by exploring a set of features (referred to as 'predictors'). In our particular case the 'label' would be the type of particle truth matched to a jet. This information is known because we use Monte Carlo programs to simulate the events, therefore access to the truth information is available.

Deep Neural Networks as Classifiers

Neural networks [149] are a set of machine learning models based on a collection of connected nodes (artificial neurons) inspired by biological neurons. An artificial neuron receives one or more separately weighted inputs that pass through a non-linear function known as the activation function. A neural network is composed of groups of these nodes, referred to as layers. Each layer of nodes is fully connected to the layer preceding it and to the layer following it. The layer that initiates the network and receives the data is known as the input layer. The last layer is known as the output layer and it will produce the final result of the network. Any layers between the input and output are known as the hidden layers. Any

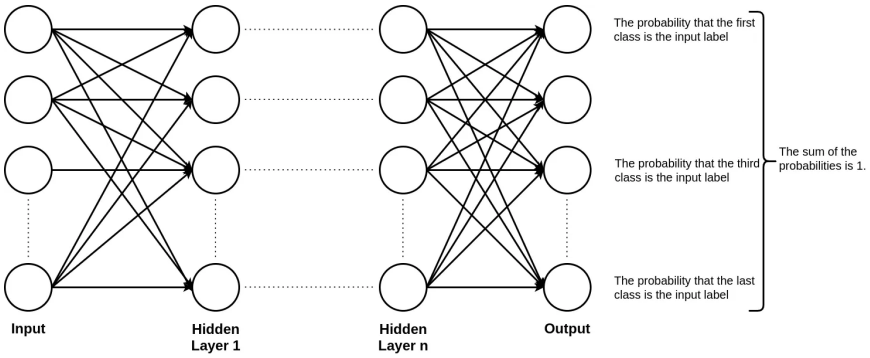


Figure 7.4: Fully connected neural network with a categorical output. Circular nodes represent the artificial neurons and the arrows represent a connection from the output of one neuron to the input of another [151].

networks with at least two hidden layers is considered to be a deep neural network [150].

To construct a classifier from a deep neural network, a softmax (normalized exponential) activation function has to be used for the output layer. The softmax function converts a vector of values to a probability distribution. The elements of the layer are in the range between 0 and 1 and will have a total sum of 1. The output of each node in the layer will be associated with the probability that the inputs came from a certain class. A binary classifier distinguishes between two classes, while a multiclass classifier can distinguish between more than two. Figure 7.4 contains an illustration of a general deep neural network implemented as a classifier of n classes. The labels of the classes have to represented as numerical values through a technique called one-hot encoding. A binary classification problem just requires two labels, 0 and 1. For multiple classes, the one-hot encoding consists of constructing unique vectors or matrices for each class where only one of the values is 1 and all the others are 0.

7.4 Binary Taggers for Boosted UFO jets

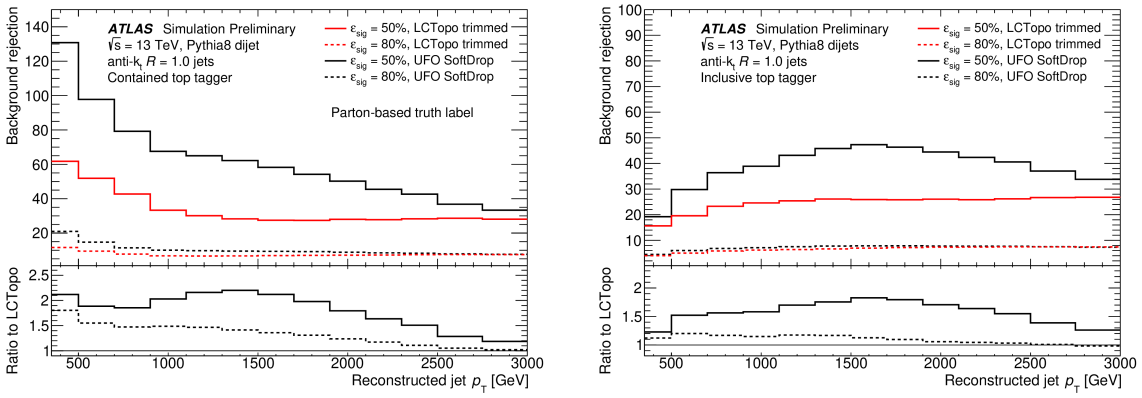
With the introduction of UFO large-R jets, development of dedicated tagging algorithms aiming at identifying boosted hadronic objects using this new reconstruction technique has

been achieved with very promising results. These taggers exploit jet substructure (JSS) variables within hadronic large-R jets making use of machine learning techniques to classify between the desired signal and the QCD multijet background.

DNN Top Taggers

The UFO top taggers [152] consist of two taggers based on a deep neural network (DNN) [150]. One tagger is optimized to identify inclusive top quark jets while the other is optimized for identifying fully contained top quark jets. A contained top quark jet is a jet where the decay products of the top quark ($t \rightarrow Wb \rightarrow qq'b$) are inside the large-R jet. The variables found to be optimal for discriminating between hadronically decaying top-quarks and quark/gluon-initiated jets are the JSS variables: N-subjettiness (τ_1, τ_2, τ_3, τ_4), the generalized energy correlation functions and their ratios (ECF_1, ECF_2, ECF_3, C_2, D_2, L_2, L_3), Q_W and thrust major T_M [153]. Other variables used are the splitting scales [154] $\sqrt{d_{12}}$ and $\sqrt{d_{23}}$. The splitting scale $\sqrt{d_{ij}}$ is defined as the k_t scale from the $i \rightarrow j$ splitting of a reconstructed jet.

The ungroomed truth labeling strategy based on particle-level large-R jets was used to distinguish between signal and background jets while training these taggers. A full description of the strategy is included in Section 7.5. As shown in Figure 7.5, the multijet rejection is improved by a factor of 30% to 100% when comparing UFO SD jets with LCTopo trimmed jets.



(a) (b)

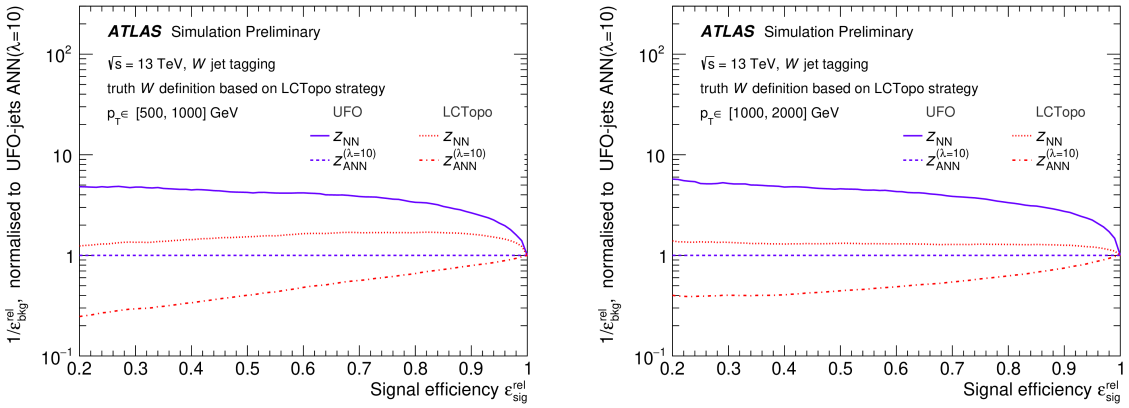
Figure 7.5: Comparison of the background rejection of the (a) contained and (b) inclusive top tagging algorithms as a function of p_{T} for a fixed signal efficiency of 50% and 80% [152].

W/Z boson taggers

The W/Z UFO taggers [155] are divided into three categories: a simple three variable (3-var) tagger that uses rectangular cuts on JSS variables; a deep neural network (DNN) W tagger and an adversarial deep neural network (ANN [156]) W/Z tagger.

The cut-based 3-var W/Z tagger uses rectangular cuts (cut on the variables per p_{T} bin) on the jet mass m_J (both upper and lower cuts), the energy correlation function ratio D_2 with $\beta = 1$ and n_{trk}, defined as the number of ID tracks with $p_{\mathrm{T}} > 500$ MeV ghost-associated with the ungroomed jet.

The DNN W tagger uses 3 fully connected layers of 32 nodes. The variables used are JSS variables D_2, C_2, τ_{21}, Fox-Wolfram moment R_2^{FW} [157], planar flow $\mathcal{P}$ [158], angularity a_3 [159], aplanarity A [160], Z_{cut} [161], the splitting scale $\sqrt{d_{12}}$ and $k_t \Delta R$ [162]. The improved JSS variable reconstruction for UFO jets improves the background rejection by a factor of 2 to 4 when comparing the DNN tagger on UFO jets with its counterpart for LCTopo jets. The main issue with this tagger is a strong sculpting of the background jet shape in the regions close to the signal W jets. This was resolved by uncorrelating the mass from the



(a) (b)

Figure 7.6: Comparison of the background rejection as a function of signal efficiency of the DNN and ANN taggers between UFO jets and LCTopo jets in (a) $500 < p_{\mathrm{T}} < 1000$ GeV and (b) $1000 < p_{\mathrm{T}} < 2000$ GeV [155].

DNN by the implementation of an ANN.

The ANN W/Z tagger consists of a trained adversarial network that competes with the DNN tagger described before. The adversarial network is trained to infer the jet mass from the outputs of the DNN. A modified loss function is used to minimize the ANN-DNN system. This mass decorrelation makes it possible for this network to be trained for identifying both W and Z bosons. Figure 7.6 shows a comparison of the performance of the W/Z taggers for different p_{T} ranges.

7.5 High p_{T} Scale Factor Extrapolation

To correct the tagging efficiencies in simulation to match those observed in data, it is necessary to define scale factors. The scale factors, defined as the ratio of the tagging efficiency in data with the tagging efficiency in simulation, can be determined up to a p_{T} where the calibration dataset has enough statistics. Therefore, the data-to-simulation scale factors are

limited and alternative methods to explore the high p_{T} regime must be implemented. A simulation-based method [163] was used to extrapolate the scale factors to a p_{T} up to 3 TeV for all the supported UFO taggers. Uncertainties were also estimated by recomputing the MC-to-MC scale factors using alternative datasets.

Strategy

The scale factor for any tagger at a specific working point is defined as

$$\mathrm{SF}(p_{\mathrm{T}}) = \frac{\epsilon^{\mathrm{data}}}{\epsilon^{\mathrm{MC}}}, \tag{7.19}$$

where ϵ^{data} is the tagging efficiency in data and ϵ^{MC} is the efficiency from simulation. Every scale factor has an associated uncertainty $\sigma(\mathrm{SF}(p_{\mathrm{T}}))$. The scale factors are available for jets with $p_{\mathrm{T}} < p_{\mathrm{T,ref}}$, where $p_{\mathrm{T,ref}}$ is defined as the "reference momentum". For the top taggers the reference momentum is 1 TeV and for the W/Z taggers the reference momentum is 450 GeV. To extrapolate the scale factors to higher momenta ($p_{\mathrm{T}} > p_{\mathrm{T,ref}}$) we define the multiplicative factor $\mathcal{R}^{MC}(p_{\mathrm{T}}; p_{\mathrm{T,ref}})$. The model used is

$$\mathrm{SF}(p_{\mathrm{T}}) = \mathrm{SF}(p_{\mathrm{T}}) \cdot \mathcal{R}^{\mathrm{MC}}(p_{\mathrm{T}}; p_{\mathrm{T,ref}}) \quad \text{for} \quad p_{\mathrm{T}} > p_{\mathrm{T,ref}}. \tag{7.20}$$

The explicit definition of the multiplicative factor is

$$\mathcal{R}^{\mathrm{MC}}(p_{\mathrm{T}}; p_{\mathrm{T,ref}}) = \frac{\epsilon^{\mathrm{data}}(p_{\mathrm{T}})/\epsilon^{\mathrm{data}}(p_{\mathrm{T,ref}})}{\epsilon^{\mathrm{MC}}(p_{\mathrm{T}})/\epsilon^{\mathrm{MC}}(p_{\mathrm{T,ref}})}. \tag{7.21}$$

Assuming a weak p_{T} dependence on the scale factors in the extrapolation region, the extrapolation scale factor can be estimated as

$$\mathrm{SF}(p_{\mathrm{T}}) := \mathrm{SF}(p_{\mathrm{T,ref}}) \Rightarrow \mathcal{R}^{\mathrm{MC}}(p_{\mathrm{T}}; p_{\mathrm{T,ref}}) = 1 \quad \text{for} \quad p_{\mathrm{T}} > p_{\mathrm{T,ref}}. \tag{7.22}$$

The extrapolation uncertainty then is composed of two terms,

$$\sigma^2(\mathrm{SF}(p_{\mathrm{T}})) = \sigma^2(\mathrm{SF}(p_{\mathrm{T,ref}})) + \sigma^2_{\mathrm{extrap}}(\mathcal{R}^{\mathrm{MC}}(p_{\mathrm{T}}; p_{\mathrm{T,ref}})) \quad \text{for} \quad p_{\mathrm{T}} > p_{\mathrm{T,ref}}. \tag{7.23}$$

Extrapolation Uncertainty

To calculate the extrapolation uncertainty, we use alternative samples (labeled MC_i) and recalculate the efficiency of each tagger ($\epsilon^{\mathrm{MC}i}$). The uncertainty is then defined as a function of the sum in quadrature of the difference between extrapolation factor evaluated for the nominal MC (equal to 1) and the factor evaluated for every systematic variation ($\mathcal{R}_i^{\mathrm{MC}}(p_{\mathrm{T}}; p_{\mathrm{T,ref}})$):

$$\sigma^2_{\mathrm{extrap}}(\mathcal{R}_i^{\mathrm{MC}}(p_{\mathrm{T}}; p_{\mathrm{T,ref}}))/\mathrm{SF}^2(p_{\mathrm{T,ref}}) \propto \sum_i \left(\Delta\mathcal{R}_i^{\mathrm{MC}}(p_{\mathrm{T}}; p_{\mathrm{T,ref}})\right)^2, \tag{7.24}$$

where

$$\mathcal{R}_i^{\mathrm{MC}}(p_{\mathrm{T}}; p_{\mathrm{T,ref}}) = \frac{\epsilon^{\mathrm{MC}i}(p_{\mathrm{T}})/\epsilon^{\mathrm{MC}i}(p_{\mathrm{T,ref}})}{\epsilon^{\mathrm{MC}}(p_{\mathrm{T}})/\epsilon^{\mathrm{MC}}(p_{\mathrm{T,ref}})}. \tag{7.25}$$

This proportionality can also be expressed as the difference between the relative impact of a systematic uncertainty on the efficiency at $p_{\mathrm{T,ref}}$. The relative difference between the

nominal efficiency and the efficiency of an alternative sample is given by

$$\delta\epsilon^{\mathrm{MC}i} = \frac{\epsilon^{\mathrm{MC}i}(p_{\mathrm{T}}) - \epsilon^{\mathrm{MC}}(p_{\mathrm{T}})}{\epsilon^{\mathrm{MC}}(p_{\mathrm{T}})} \tag{7.26}$$

Therefore the extrapolation uncertainty can be quantified by

$$\sigma^2_{\mathrm{extrap}}(p_{\mathrm{T}}; p_{\mathrm{T,ref}}) = \sum_i \max_{<p_{\mathrm{T}}} \left[\delta\epsilon^{\mathrm{MC}i}(p_{\mathrm{T}}) - \delta\epsilon^{\mathrm{MC}i}(p_{\mathrm{T,ref}}) \right]^2. \tag{7.27}$$

The $\max_{<p_{\mathrm{T}}}$ represents an ad-hoc procedure in which the maximum uncertainty value observed up until the jet p_{T} is chosen. This procedure was implemented to ensure that the extrapolation uncertainty is a monotonically increasing function.

For each source of uncertainty, the quantites σ^+ and σ^- are calculated using the upper and lower envelope values of efficiency. This is performed because the total uncertainty is asymmetric and therefore must be considered separately.

Samples and Event Selection

The datasets consist of samples of BSM processes. For the top taggers, $Z' \to t\bar{t}$ events are used while for the W/Z taggers a $W' \to WZ$ sample was produced. The cross-section of the events is reweighted to produce a flat p_{T} distribution in order to populate the region 200 GeV $< p_{\mathrm{T}} <$ 3 TeV. Only events with at least one UFO jet with $p_{\mathrm{T}} > 200$ GeV, a reconstructed $m > 40$ GeV, and $|\eta| < 2.0$ are selected.

Top Quarks

$Z' \to t\bar{t}$ events are simulated using PYTHIA8 with the NNPDF2.3Lo PDF set using the A14 tuning for the parton shower and MPI interactions. Only hadronically decaying top quarks are considered. To prevent the inclusion of high p_{T} gluons recoiling against the $t\bar{t}$ system, the selection strategy requires that the $t\bar{t}$ pairs are well separated; they must satisfy the condition $\Delta R(t,\bar{t}) = \sqrt{\Delta\phi^2_{t\bar{t}} + \Delta\eta^2_{t\bar{t}}} > 2.0$.

The truth labeling procedure starts by geometrically matching detector-level large-R jet (J) to a particle-level jet (J_{truth}) by requiring $\Delta R(J, J_{\mathrm{truth}}) < 0.75$. Then J_{truth} is matched to a top quark by requiring $\Delta R(J_{\mathrm{truth}}, \mathrm{top}) < 0.75$. If these two requirements are satisfied then the large-R jet is labeled as *inclusive* top.

For *contained* tops , all the decay products of the top quark ($t \to Wb \to qq'b$) must be included inside the large-R jet. The ungroomed truth labeling strategy is employed. The strategy requires that at least one b-hadron is ghost-associated to the ungroomed truth jet (J_{truth}). A cut on the ungroomed mass ($m_{\mathrm{ungroomed}} > 140$ GeV) is also required. Finally it also has to satisfy the ungroomed p_{T}-dependent k_t splitting scale ($\sqrt{d_{32}}$) cut. The splitting scale cut is

$$\sqrt{d_{23}} > \exp\big(3.3 - \frac{6.98 \cdot 10^{-4}}{\mathrm{GeV}} p_{\mathrm{T,ungroomed}}\big)\ \mathrm{GeV}. \tag{7.28}$$

W/Z Bosons

The $W' \to WZ$ events are also simulated using PYTHIA8 with the NNPDF2.3Lo PDF set and using the A14 tuning for the parton shower and MPI interactions. Only hadronically decaying W/Z bosons are considered. To prevent overlap, the W and Z bosons must satisfy the condition $\Delta R(W,Z) = \sqrt{\Delta\phi^2_{WZ} + \Delta\eta^2_{WZ}} > 2.0$.

The ungroomed truth labeling strategy is also applied for the W/Z boson sample. There-

fore large-R jets must satisfy $\Delta R(J, J_{\text{truth}}) < 0.75$ and $\Delta R(J_{\text{truth}}, \text{W/Z}) < 0.75$. For W-labeled jets, in order to remove contamination from top-quark decays it is required that the number of ghost associated b-hadrons is zero. The ungroomed truth mass ($m_{\text{ungroomed}}$) is required to be above 50 GeV and a cut on the energy scale of the first k_t-declustering ($\sqrt{d_{12}}$) must also be satisfied. The cut is defined as:

$$\sqrt{d_{12}} > 55.25 \cdot \exp\Big(\frac{-2.34 \cdot 10^{-3}}{\text{GeV}} p_{\text{T,ungroomed}}\Big)\,\text{GeV}. \tag{7.29}$$

Uncertainty Sources

Shower Variations

The variation of the unphysical scales that arise in fixed order QCD calculations can be used to estimate the theoretical uncertainties associated with showering models.

The Pythia event generator provides a way to evaluate these variations of the renormalization scale (μ_R) and splitting kernel in the showering process. Standard parton shower algorithms generate the scale of the next branching by solving an equation that is a function of the differential branching probability [164]. Pythia uses transverse momentum-ordered showers, where the differential branching probability is given by

$$P(t, z) = \frac{\alpha_s(t)}{2\pi} \frac{P(z)}{t}, \tag{7.30}$$

where $t = p_\perp^2$, z is the momentum fraction carried by a parton after the splitting and $P(z)$ the DGLAP splitting kernel. The splitting kernel is the function that represents the probability of a particular parton splitting into two daughter partons with specified momentum fractions. For a baseline gluon-emission with a NLO compensating term and a renormalization scale

variation $\mu_R = p_\perp \to \mu'_R = kp_\perp$, the probability density is given by

$$P'(t,z) = \frac{\alpha_s(\mu'_R)}{2\pi}\left(1 + \frac{\alpha}{2\pi}\beta_0 \ln k\right)\frac{P(z)}{t}, \tag{7.31}$$

where $\beta_0 = (11N_C - 2n_F)/3$ with $N_C = 3$ and n_F the number of active flavors at the scale $\mu = p_\perp$. The renormalization scale (μ_R) variation is performed independently for ISR and FSR branchings given that their kernels receive different NLO corrections. The scale is varied by factors of 2 and 1/2. A 7-point scale variation prescription is employed to calculate the uncertainty:

$$\left(\frac{\xi}{\xi_0}\right) = \big\{(0.5,0.5),(1,0.5),(0.5,1),(1,1),(1,2),(2,1),(2,2)\big\}, \tag{7.32}$$

where $\xi = (\mu_R^{\mathrm{ISR}}, \mu_R^{\mathrm{FSR}})$ and ξ_0 represents the nominal (central) value.

In addition to the 7-point scale variation two extra variations related to non-singular finite terms in the splitting functions are included. In the context of a DGLAP approach, the shower splitting kernels are modified in the following way

$$\frac{P(z)}{Q^2}dQ^2 \to \left(\frac{P(z)}{Q^2} + \frac{c_{\mathrm{NS}}}{m^2_{\mathrm{dip}}}\right)dQ^2 = \left(P(z) + \frac{c_{\mathrm{NS}}Q^2}{m^2_{\mathrm{dip}}}\right)\frac{dt}{t}, \tag{7.33}$$

where m_{dip} is the invariant mass of the dipole in which the splitting occurs and c_{NS} is a dimensionless constant that parametrizes the amount of non-singular splitting kernel variation. These two splitting kernel variations are labeled as "hardHi" and "hardLo" and correspond to the values $c_{\mathrm{NS}} = \pm 2$. The 7(+2) point variation envelope (up and down) around the nominal value is then chosen to estimate the uncertainties. Two variations due to the PDF set chosen are also included as an independent source of uncertainty.

Detector Geometry and Nuclear Interaction Models

Several alternative MC samples were produced to determine the uncertainties associated with nuclear interactions and detector geometry variations. The samples are part of the GEANT4 "Physics List" [165] and thus modify observables at the reconstruction level. All the samples are summarized with a short description in Tables 7.2 and 7.1.

The detector geometry variations simulate misaligned detector geometry or other imperfections that could have an impact on how the detector records the collision events. The variations include LAr distorted geometry, inner detector variations and material scaling.

The nuclear interaction modeling involves the consideration of different models for elastic, inelastic, capture and fission processes. Some of the nuclear interaction models used in these alternative samples are the quark gluon string model (QGS) [166], the Bertini intranuclear cascade model (BERT) [167], the Fritiof model of string excitation (FTF) [168] and the chiral invariant phase space model (CHIPS) [169]. The quark gluon string model (QGS) is a model that uses color flow between partons to simulate reactions between high energy hadrons with nuclei and high energy electro-nuclear reactions. The Bertini intranuclear cascade model (BERT) is a classical model that solves the Boltzmann equation for the transport of a particle through a "gas" of nucleons. It is valid for protons, neutrons, pions, kaons and hyperons in the kinetic range up to 10 GeV. The Fritiof model of string excitation (FTF) is a high energy string model that simulates hadron-hadron, hadron-nucleus and nucleus-nucleus interactions. This model includes elastic scattering and a separate simulation of single diffractive and non-diffractive events. It is meant to work with energies between 3 GeV and 1 TeV. The chiral invariant phase space model (CHIPS) is used to approximate the Drell-Yan process in hadron-nucleon interactions. In GEANT4 it is used for γ-nuclear interactions,

nuclear capture of negatively charged hadrons and quasi-elastic scattering processes.

Table 7.1: Summary of the "Physics List" samples used to estimate the high p_{T} extrapolation uncertainties related to detector geometry variations.

s-tag	Description
3126	Nominal sample - Nominal detector geometry
3373	LAr distorted geometry with material between LAr and Tile
3374	LAr distorted geometry with all distortions before LAr EM
3375	Inner detector systematic geometry variations
3376	Inner detector with +5% overall material scaling

Table 7.2: Summary of the "Physics List" samples used to estimate the high p_{T} extrapolation uncertainties related to nuclear interaction models.

s-tag	Description
3126	Nominal sample - QGS, BERT
3295	FTF, BERT with no diffraction
3296	QGS for high energies, FTF for lower energies, BERT
3297	FTF, BERT with no re-scattering of the final state
3298	FTF, BERT, CHIPS for nuclear capture
3299	FTF, BERT with a high precision data driven neutron transport package

Results

The total uncertainties for the high p_{T} extrapolation of the scale factors due to scale variations, PDFs, nuclear interaction modeling and detector geometry variations are summarized in Figures 7.7 - 7.15. Appendix B contains all the effiiciency plots as well as the efficiency envelopes used to calculated the uncertainties.

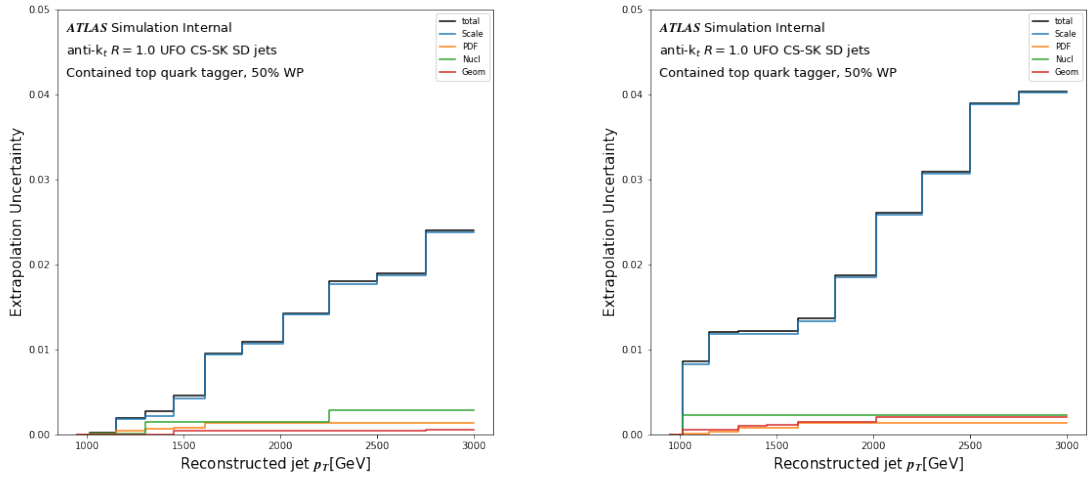
For the top taggers, the total uncertainty ranges from 0% at $p_{\mathrm{T,ref}} = 1000$ GeV (by

construction) to 1-4% at $p_{\mathrm{T}} = 3$ TeV. The 'down' component of the total uncertainty is larger across all taggers and all working points. Shower scale variations dominate as the largest source of uncertainty while nuclear interaction modeling, detector geometry and PDF variation are negligible ($< 1\%$).

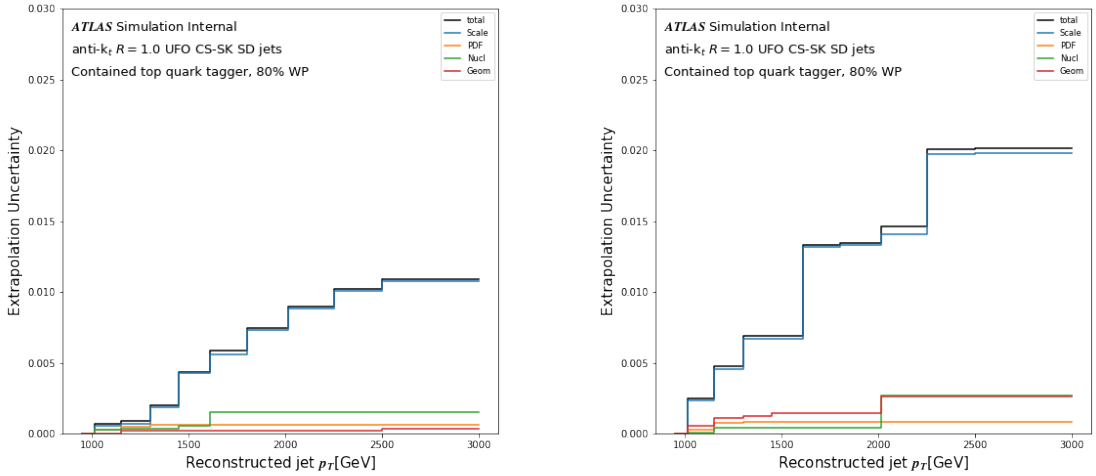
The 3-var W/Z tagger 'up' uncertainties range from 0.5 - 1.50% at $p_{\mathrm{T}} = 3$ TeV. On the other hand, for the same range, the 'down' uncertainties go from 1.75 - 2.25%. Similar to the top taggers, the largest source of uncertainty is the scale variations.

For the DNN W tagger, the total uncertainties for the 50% WP are much larger than the uncertainties for the 80% WP. The 'up' uncertainties for jets with $p_{\mathrm{T}} = 3$ TeV go from 1% for the 80% WP to 3% for the 50% WP. Similarly, the 'down' components are 0.5% and 1.5% for the 80% WP and 50% WP respectively.

The ANN W/Z taggers contain overall the lowest extrapolation uncertainties. Scale variations still dominate as the largest source but not in all cases. The largest uncertainties come from the 'down' components of the W tagger calibrated with the 50% WP with a value of around 1.75% at a p_{T} of 3 TeV.

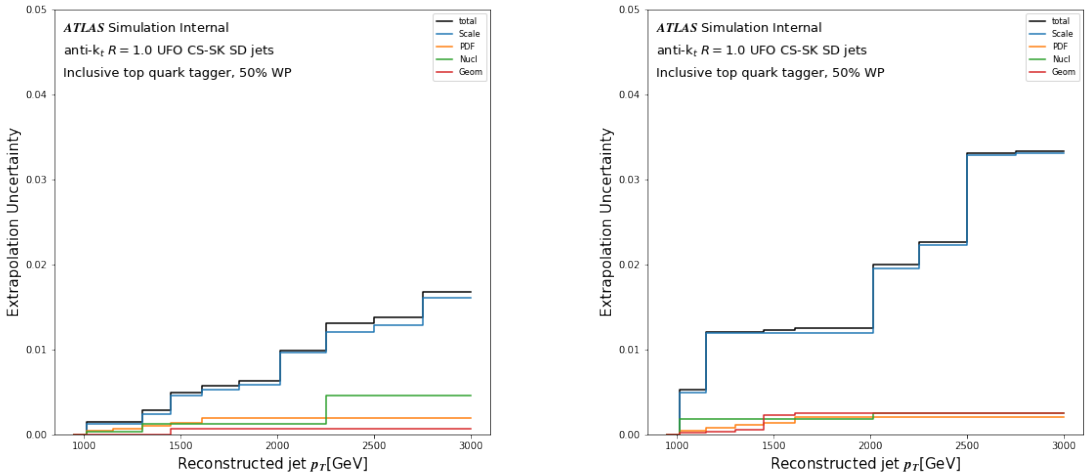


(a) (b)

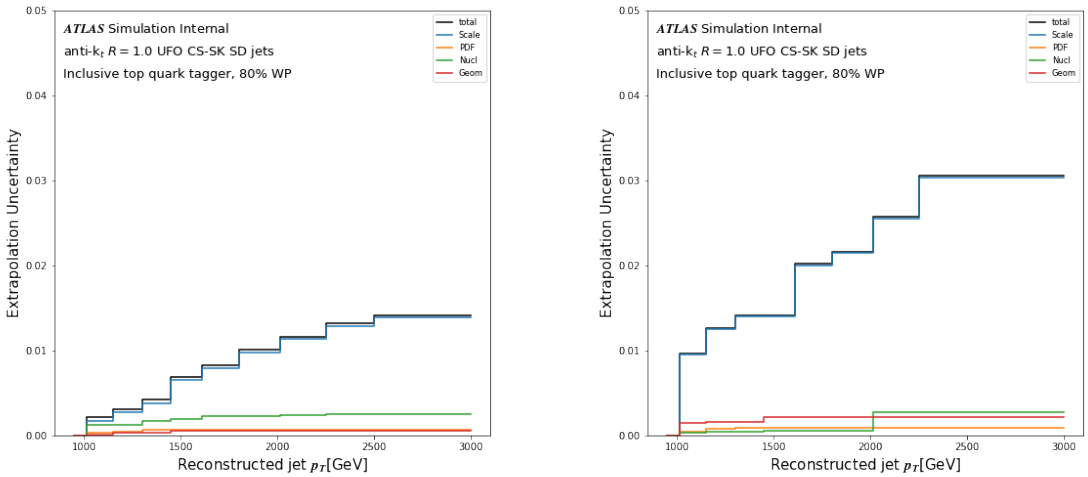


(c) (d)

Figure 7.7: Total scale factor extrapolation uncertainty for the contained top DNN tagger at the (a,b) 50% WP and (c,d) 80% WP as a function of p_{T}. Plots (a,c) show the 'up' (σ^{+}) uncertainty component and plots (b,d) show the 'down' (σ^{-}) uncertainty component.

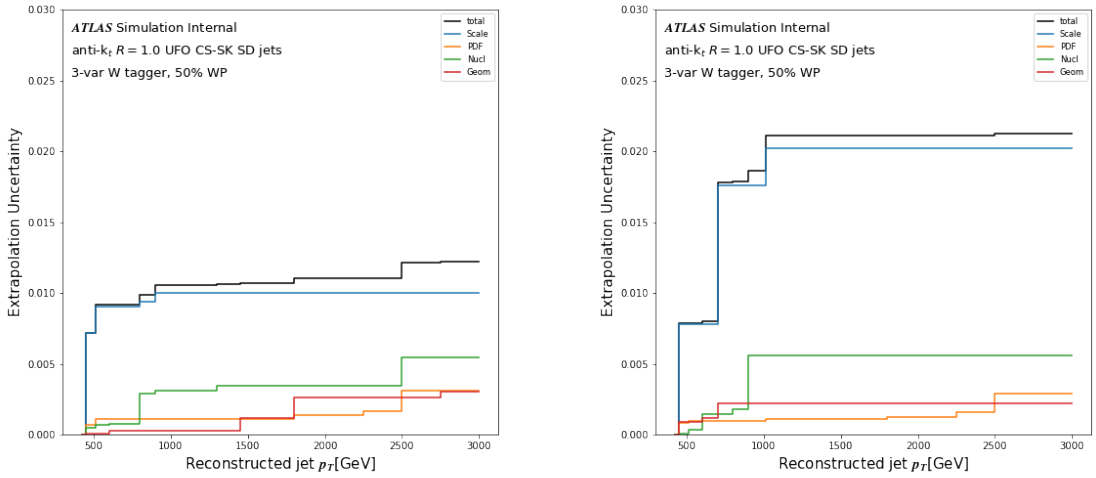


(a) (b)



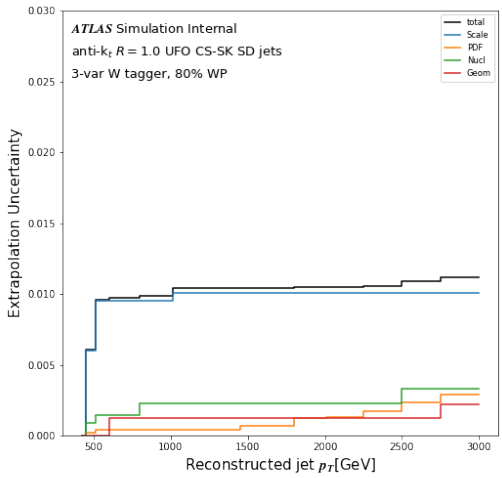
(c) (d)

Figure 7.8: Total scale factor extrapolation uncertainty for the inclusive top DNN tagger at the (a,b) 50% WP and (c,d) 80% WP as a function of p_{T}. Plots (a,c) show the 'up' (σ^{+}) uncertainty component and plots (b,d) show the 'down' (σ^{-}) uncertainty component.

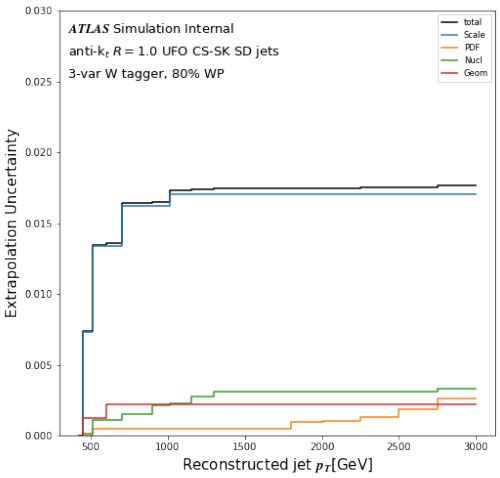


(a)

(b)

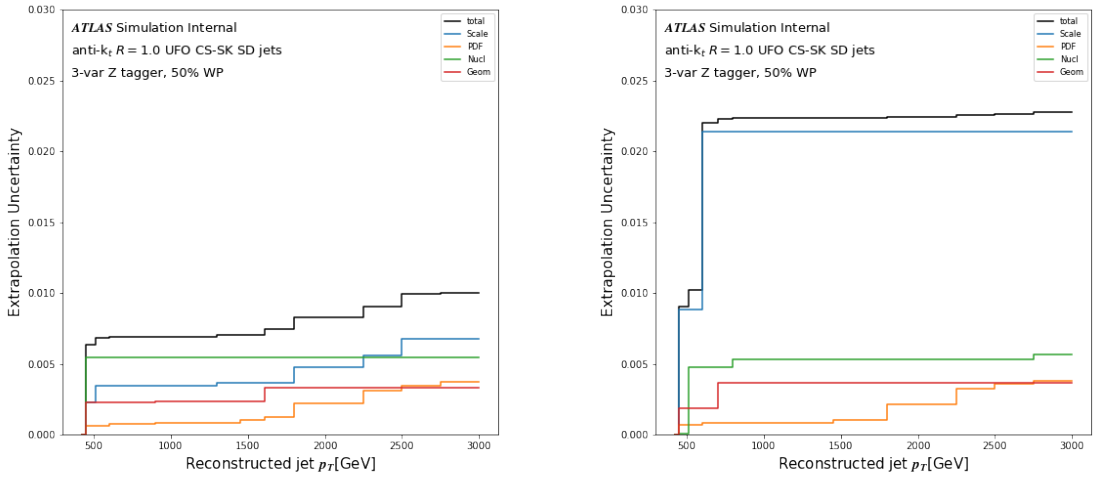


(c)

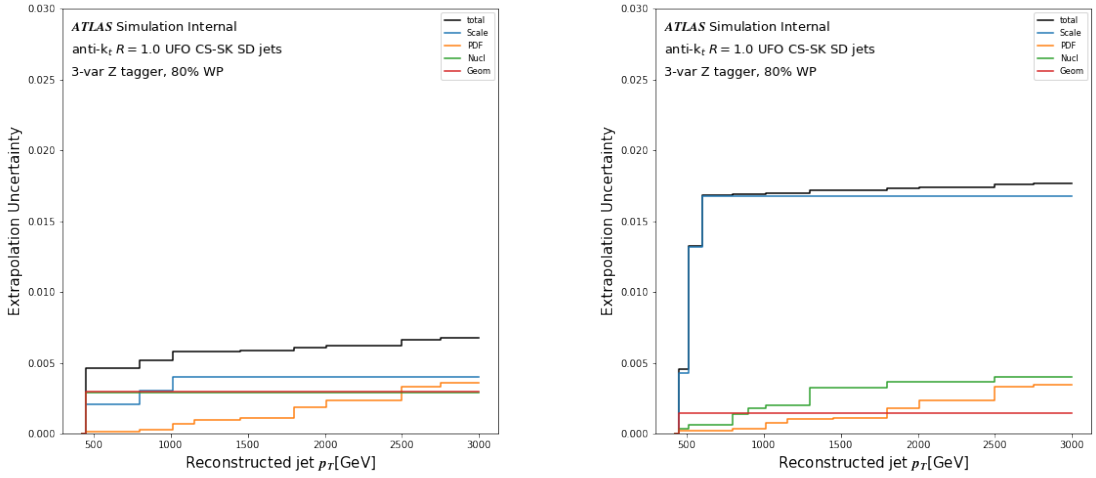


(d)

Figure 7.9: Total scale factor extrapolation uncertainty for the 3-var W tagger at the (a,b) 50% WP and (c,d) 80% WP as a function of p_{T}. Plots (a,c) show the 'up' (σ^{+}) uncertainty component and plots (b,d) show the 'down' (σ^{-}) uncertainty component.

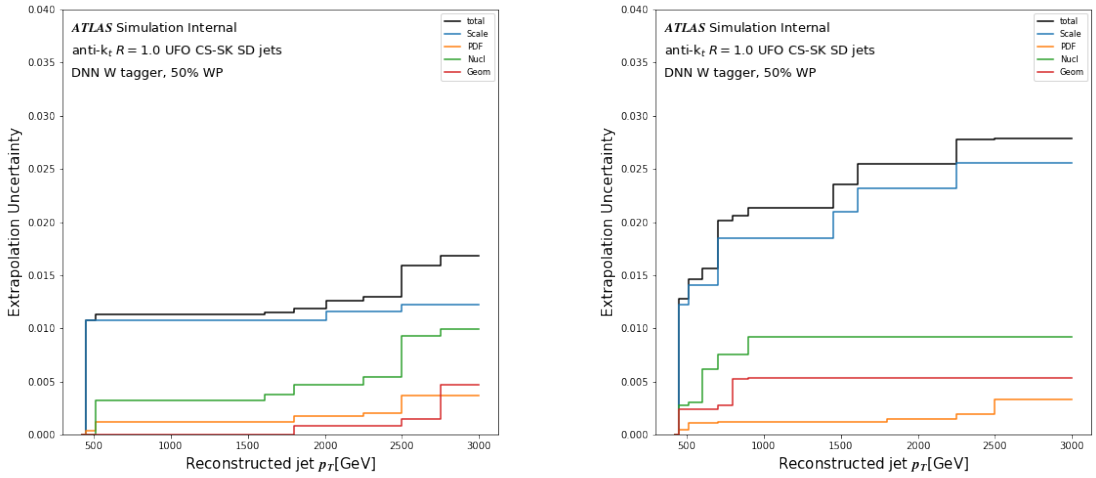


(a) (b)



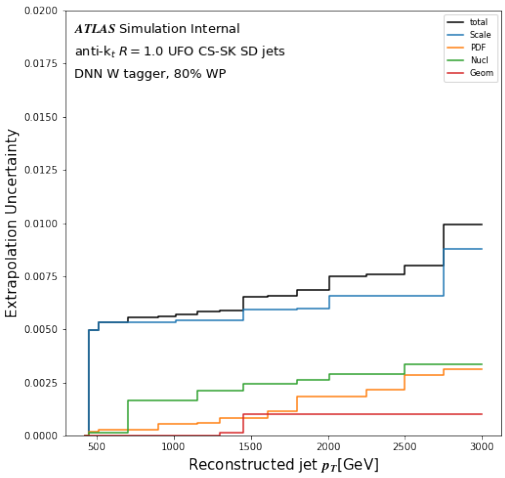
(c) (d)

Figure 7.10: Total scale factor extrapolation uncertainty for the 3-var Z tagger at the (a,b) 50% WP and (c,d) 80% WP as a function of p_{T}. Plots (a,c) show the 'up' (σ^{+}) uncertainty component and plots (b,d) show the 'down' (σ^{-}) uncertainty component.

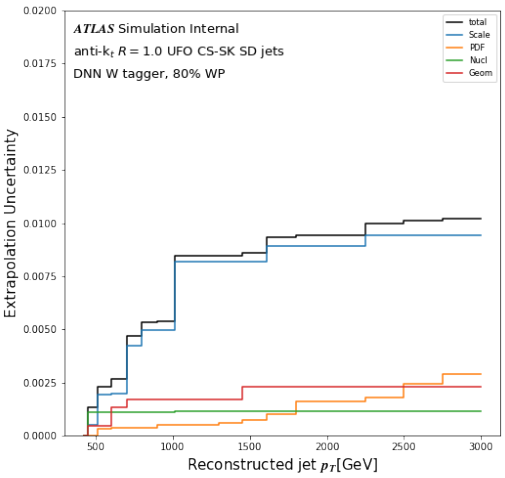


(a)

(b)

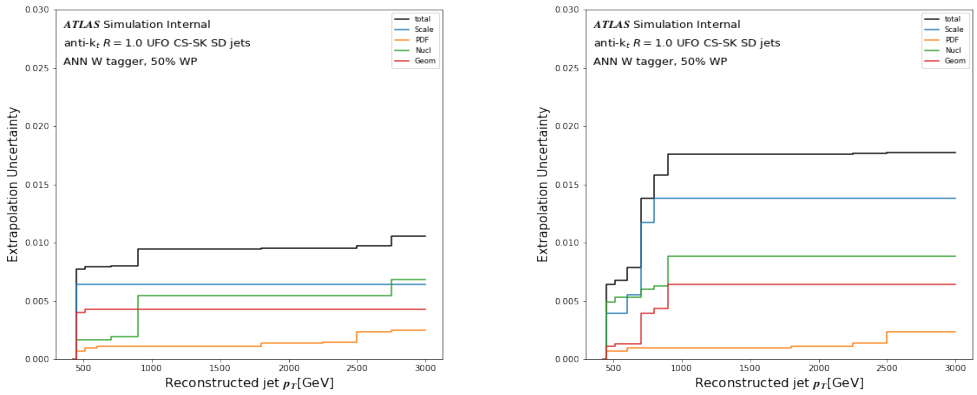


(c)



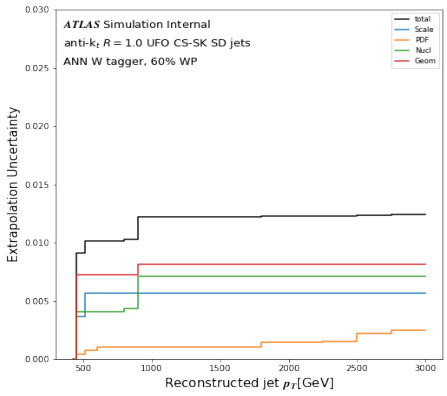
(d)

Figure 7.11: Total scale factor extrapolation uncertainty for the DNN W tagger at the (a,b) 50% WP and (c,d) 80% WP as a function of p_{T}. Plots (a,c) show the 'up' (σ^{+}) uncertainty component and plots (b,d) show the 'down' (σ^{-}) uncertainty component.

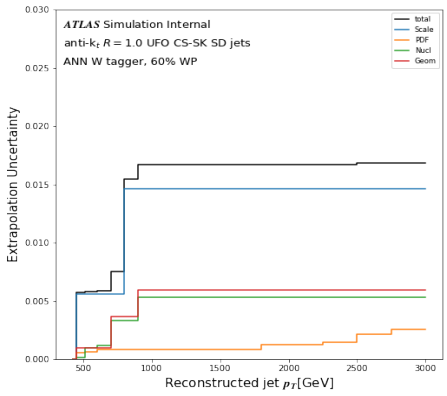


(a)

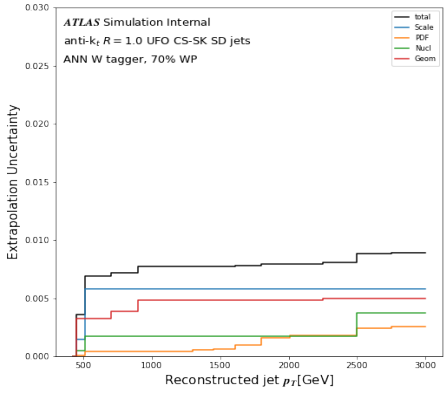
(b)



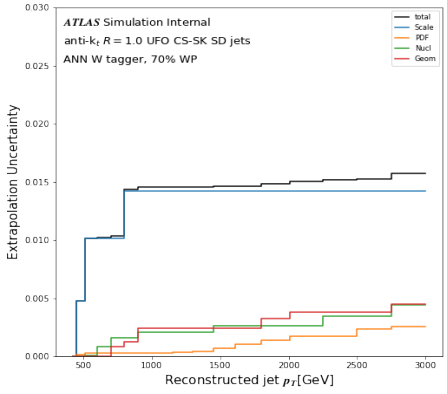
(c)



(d)

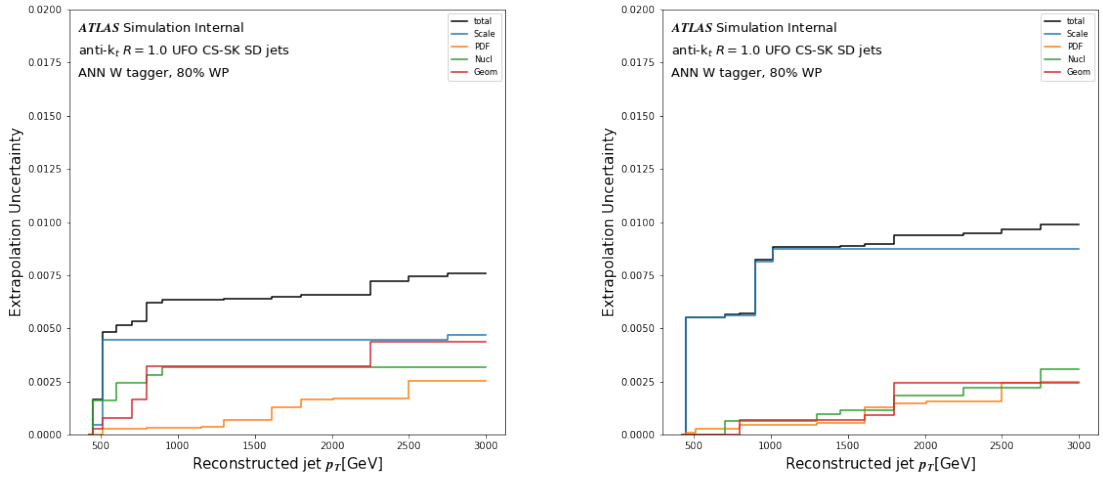


(e)



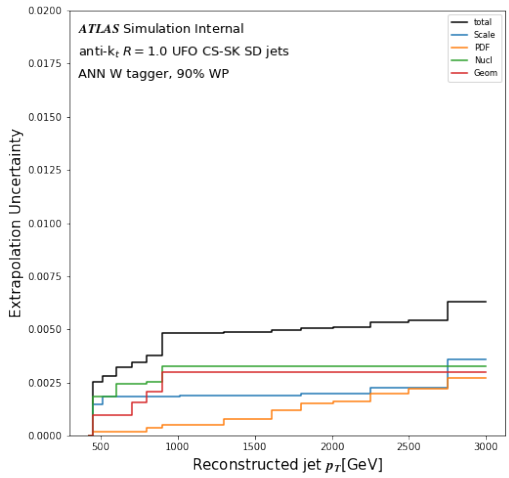
(f)

Figure 7.12: Total scale factor extrapolation uncertainty for the ANN W tagger at the (a,b) 50% WP, (c,d) 60% WP and (e,f) 70% as a function of p_{T}. Plots (a,c,e) show the 'up' (σ^{+}) uncertainty component and plots (b,d,f) show the 'down' (σ^{-}) uncertainty component.

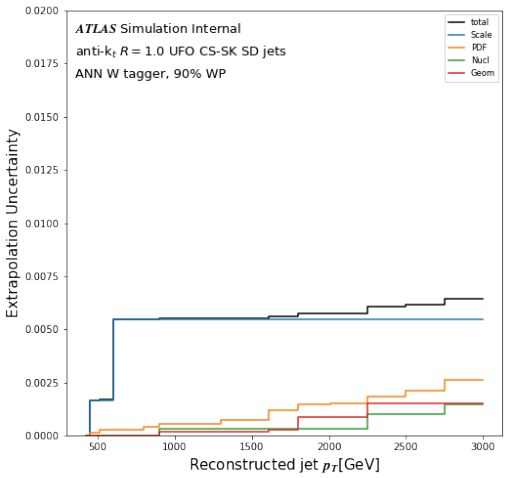


(a)

(b)

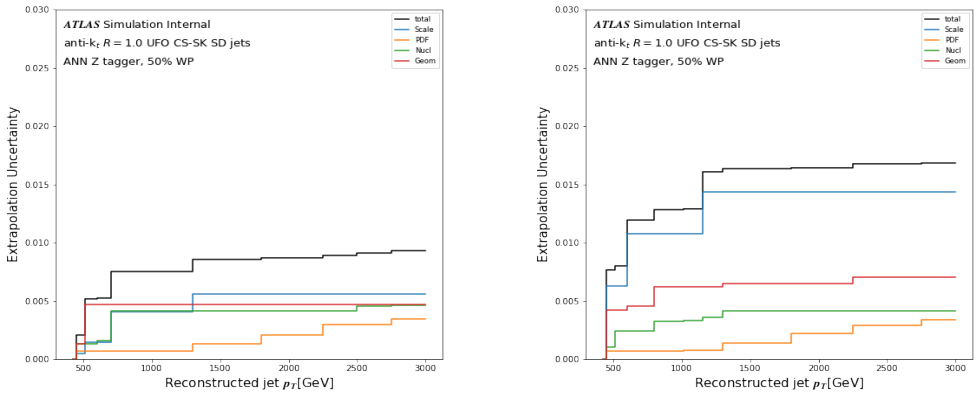


(c)



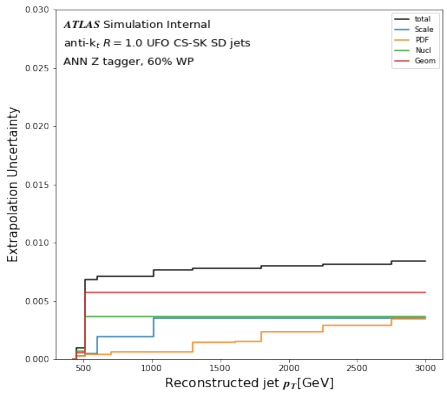
(d)

Figure 7.13: Total scale factor extrapolation uncertainty for the ANN W tagger at the (a,b) 80% WP, (c,d) and 90% WP as a function of p_{T}. Plots (a,c) show the 'up' (σ^{+}) uncertainty component and plots (b,d) show the 'down' (σ^{-}) uncertainty component.

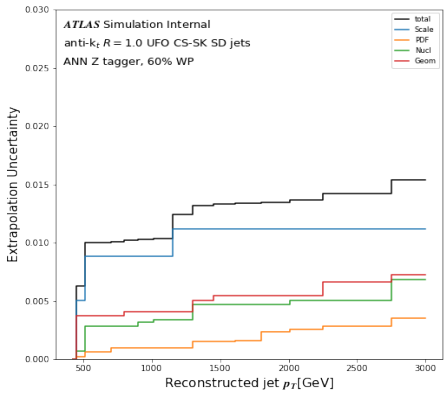


(a)

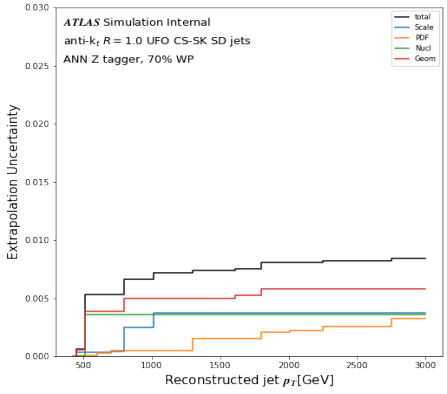
(b)



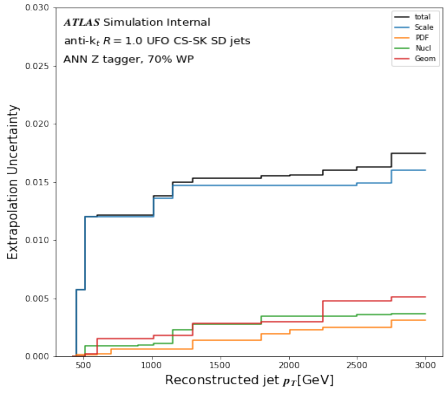
(c)



(d)

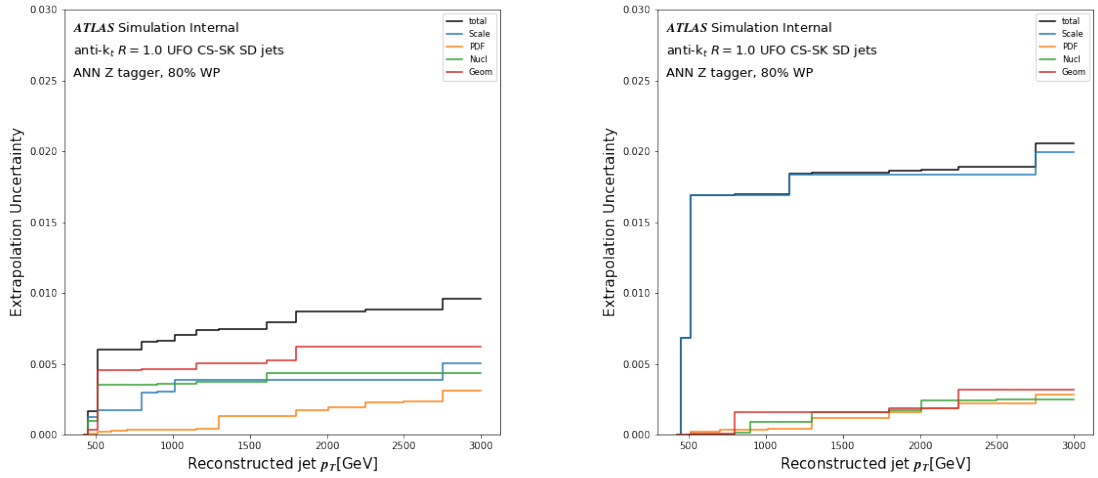


(e)



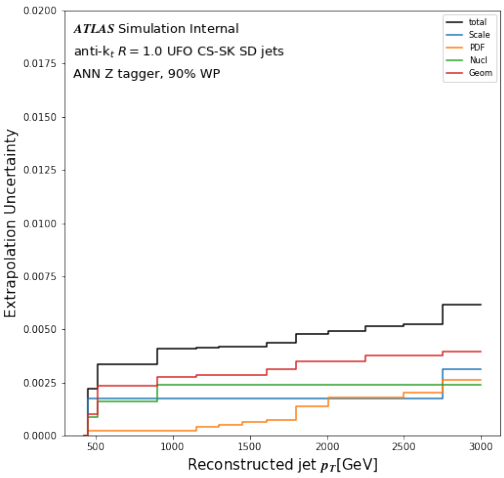
(f)

Figure 7.14: Total scale factor extrapolation uncertainty for the ANN Z tagger at the (a,b) 50% WP, (c,d) 60% WP and (e,f) 70% as a function of p_{T}. Plots (a,c,e) show the 'up' (σ^{+}) uncertainty component and plots (b,d,f) show the 'down' (σ^{-}) uncertainty component.

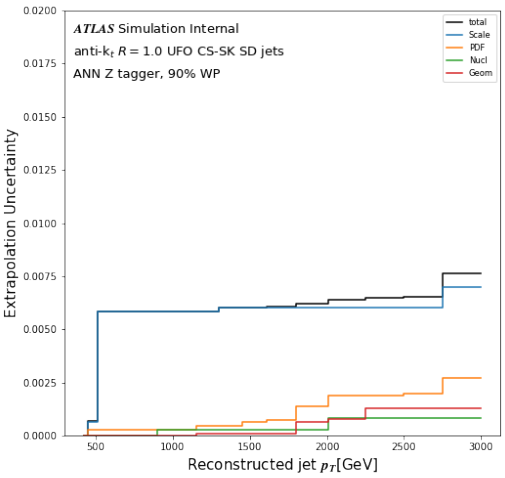


(a)

(b)



(c)



(d)

Figure 7.15: Total scale factor extrapolation uncertainty for the ANN Z tagger at the (a,b) 80% WP, (c,d) and 90% WP as a function of p_{T}. Plots (a,c) show the 'up' (σ^{+}) uncertainty component and plots (b,d) show the 'down' (σ^{-}) uncertainty component.

7.6 Multiclass Tagger for Boosted UFO jets

Multiclass problems have more structure than binary classification problems because the output relates classes with each other. This section will explore studies related to building a multiclass tagger (MCT) based on a deep neural network that classifies between 5 classes: top quarks, W bosons, Z bosons, Higgs bosons and QCD jets. This kind of tagger has already been found to be effective by the CMS collaboration [170]. The aim is to gain insight on the feasability of a unified tagger for multiple types of particles as well as the performance difference between taggers developed for UFO jets and taggers developed for LCTopo jets.

To train the deep neural network, Monte Carlo samples were produced. For top quarks and W/Z bosons the same samples and selection criteria described in Section 7.5 were employed. To produce Higgs bosons, a new sample of the BSM process $G \to HH$ events was used. Here G refers to Randall-Sundrum gravitons [171]. The events are simulated using Pythia8 with the NNPDF2.3Lo PDF set using the A14 tuning. An ungroomed truth labeling strategy was applied to select the events where the truth Higgs boson was associated with the reconstructed large-R jet. Two b-hadrons ghost-associated to the ungroomed truth jet were also required. The events are reweighted so that the p_{T} distribution for each sample is the same.

The initial set of features used as input included kinematical variables, N-subjettiness variables and their ratios, generalized energy correlation functions and their ratios, jet charge, splitting scales, number of constituents, Fox-Wolfram moment, angularity, aplanarity, planar flow, kinematic variables of up to three track jets associated with the large-R jet and finally the output scores of the DL1r b-tagger. The total amount of features was 74 and thus a strategy to reduce significantly this number was required.

Before any study, the input variables were standardized so they demonstrate features of a standard Gaussian distribution,

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \tag{7.34}$$

where μ and σ are the mean and standard deviation of the feature respectively. This procedure reduces the impact of outliers and also ensures the dataset is on the same scale, avoiding the domination of features with larger values. Class imbalances are reduced by weighting the events such that the total number of events per class is of a similar value. Larger weights are given to the classes that have fewer events in the samples.

Feature Selection

The permutation feature importance algorithm was performed on a simple deep neural network with 4 hidden layers. The permutation feature importance [172] is a model inspection technique that can be used to study the decrease in a model score when a single feature value is randomly shuffled. Let s be the reference score (i.e. accuracy) of the model on our testing dataset. For each feature j, we perform k in $1, \cdots, K$ repeated permutations (shuffling) of the feature j and recompute the score $s_{k,j}$ of the model. The importance of a feature j can be then calculated from the importance, defined as

$$i_j = s - \frac{1}{K} \sum_{k=1}^{K} s_{k,j}. \tag{7.35}$$

The permutation feature importance tests revealed rapidly declining and overall low values for the importance scores for the original 74 features. A false lower importance score can

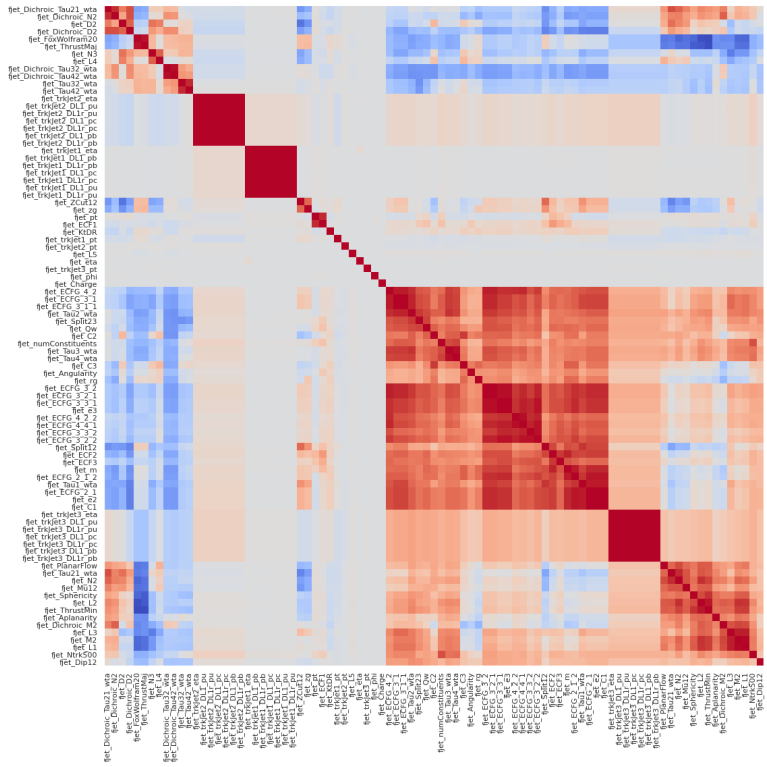


Figure 7.16: Correlation matrix of the initial set of features used as inputs for the MCT. Red represents high correlation while blue represents low correlation.

be seen when there are features with strong correlations between each other. The model still has access to the feature through its correlated features and falsely concludes that it is not important. Therefore, it is important to understand the correlations between features as the set of features selected may contain redundant information. Figure 7.16 shows the correlation matrix of the set of initial features used as inputs for the deep neural network. It is evident that there are high correlations between some features which can be seen as clusters in the correlation matrix.

The high degree of correlation can be exploited to perform dimensionality reduction methods by employing clustering algorithms. Hierarchical clustering is a method to build nested clusters at different resolutions. This method requires a distance measure between the

variables $X = \{X_1, \cdots, X_n\}$. In this case, the Ward's linkage [173] is used as our distance metric d. The algorithm proceeds as follows:

Let $T_n = \{C_1, C_2, \cdots, C_n\}$ where $C_i = \{X_i\}$, for $j = n - 1$ to 1:

(a) Find j, k to minimize $d(C_j, C_k)$ over all $C_j, C_k \in T_{j+1}$.

(b) Let T_j be the same as T_{j+1} except that C_j and C_k are replaced with $C_j \cup C_k$

Finally, return the sets of clusters $T_1, \cdots T_n$. The results can then be represented as a dendrogram. Inspecting the dendrogram we can then choose which feature from each cluster to keep as a representative variable. A threshold of 2 was used for the cophenetic distance, defined as the height of the dendrogram where two branches that include the two objects merge into a single branch. The resulting dendrogram is shown in Figure 7.17. The selection of features within each group are required to have generally larger importance, as well as good agreement with data within modeling uncertainties. Appendix B contains plots of data/MC comparisons of most of the jet substructure variables distributions. This method reduced the number of features from 74 to 26 without a significant loss in the accuracy of the classifier.

The final features used as input for the MCT are n_{trk}, p_T, m, τ_{32}, e_3^2, C_2, D_2, L_2, L_3, L_4, A, a_3, R_2^{FW}, $\sqrt{d_{12}}$, $\sqrt{d_{23}}$, Z_{cut}, charge and the 3 DL1r outputs used to define the discriminant (p_{light}, p_b, p_c) for the three leading track jets.

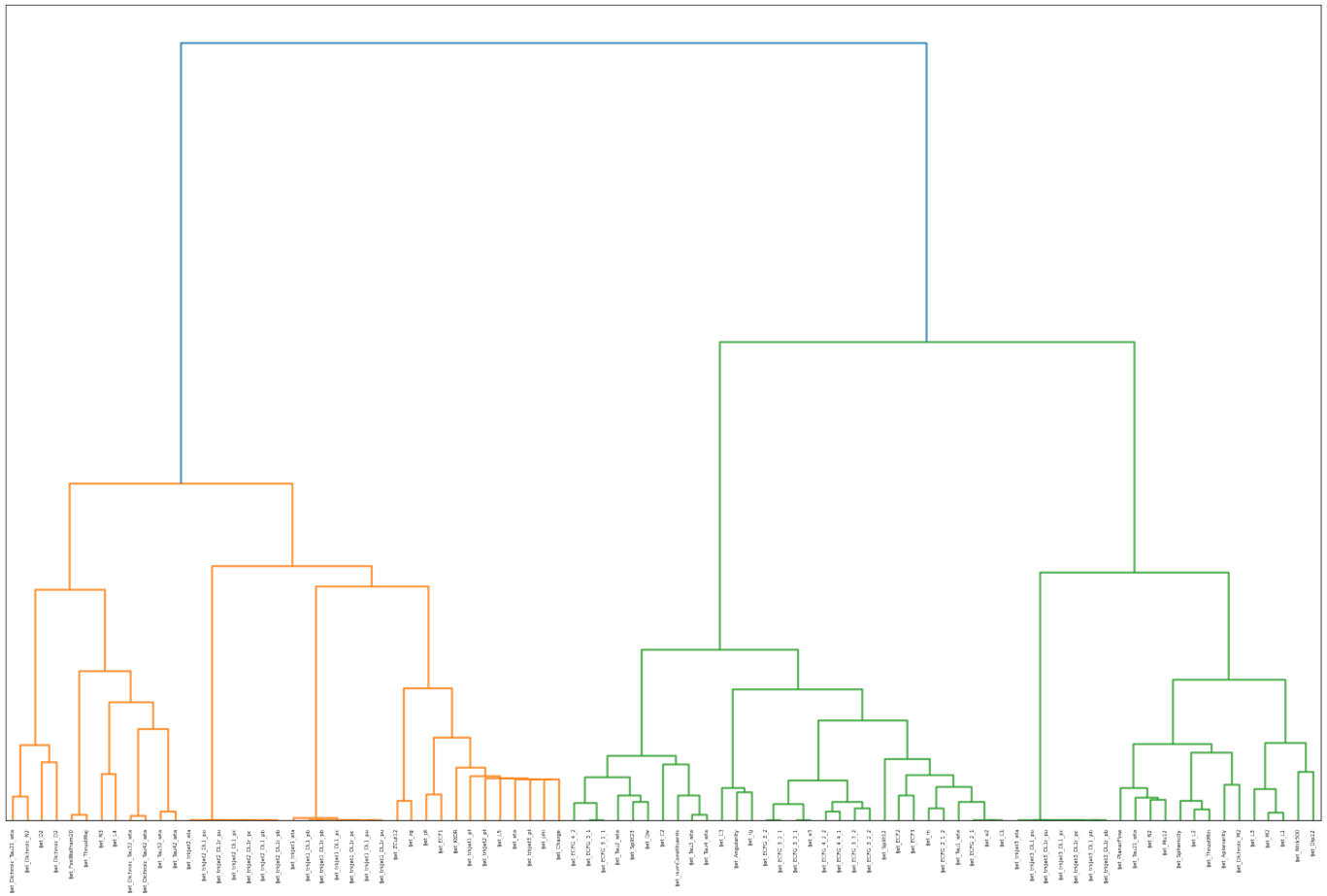


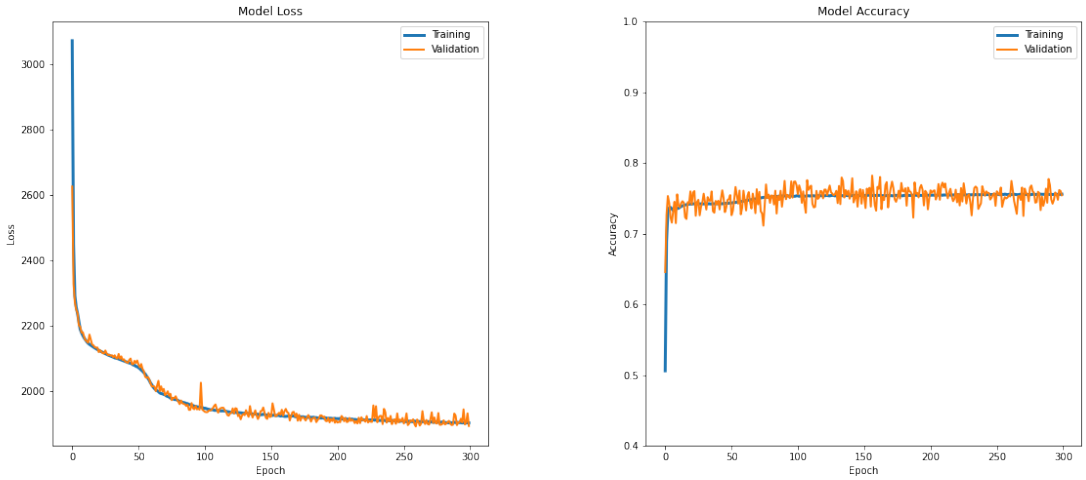
Figure 7.17: Dendrogram representation of the grouping of the input features correlations using hierarchical clustering with Ward's linkage as a distance metric.

Deep Neural Network Model

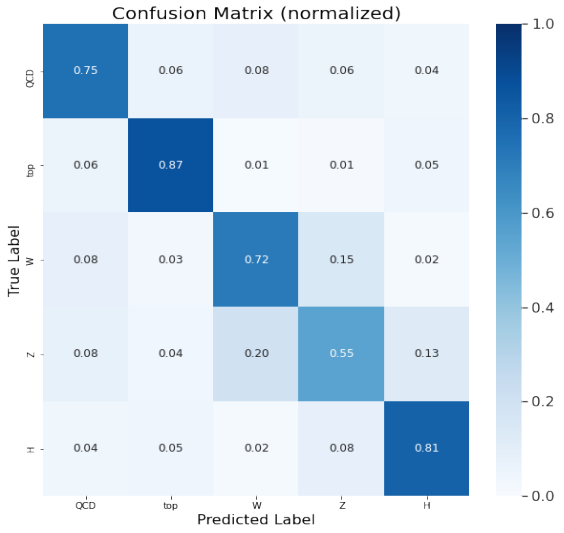
Tensorflow [174] was used as a backend to construct the model. For training the model, 75% of the total events are used and trained for 300 epochs. The other 25% is divided equally to use for validation and testing. The network contains four hidden layers, each with 10 to 20 nodes and a rectified linear unit (ReLU) [175] activation function. The output layer uses a softmax activation function and outputs five scores, each one associated to the probability of the jet being from each class ($s_{QCD}, s_{top}, s_W, s_Z, s_H$). The largest score is taken as the predicted class. The loss function used was the categorical cross-entropy with a L2 (Ridge regression) [176] regularization and was optimized using Adam [177], an extended version of stochastic gradient descent. Results for the network are shown in Figure 7.18. The regularization technique introduces a penalty term into the loss function to prevent overfitting. The model loss as a function of epoch shows a steady decrease in value for both the training loss and validation loss. Conversely, the model accuracy increases slowly as a function of epoch. This shows that the model still has potential to learn (by training for more epochs) as no signs of overfitting were detected.

The average accuracy considering all classes is 74%. Per class, the tagger achieves an accuracy of 75% for QCD multijets, 87% for top quarks, 72% for W bosons, 55% for Z bosons and 81% for Higgs bosons. The MCT doesn't perform well for identifying Z bosons, with only a 55% accuracy; most of the true Z bosons are being identified as W bosons (20%) and Higgs bosons (13%). The ANN W/Z tagger has shown a good performance when introducing an adversarial term to the loss function as a mass decorrelation measure. Such an extension to the MCT could alleviate the misidentification rate between these two classes.

These results are compatible with other observations from a multiclass tagger on LCTopo



(a) (b)



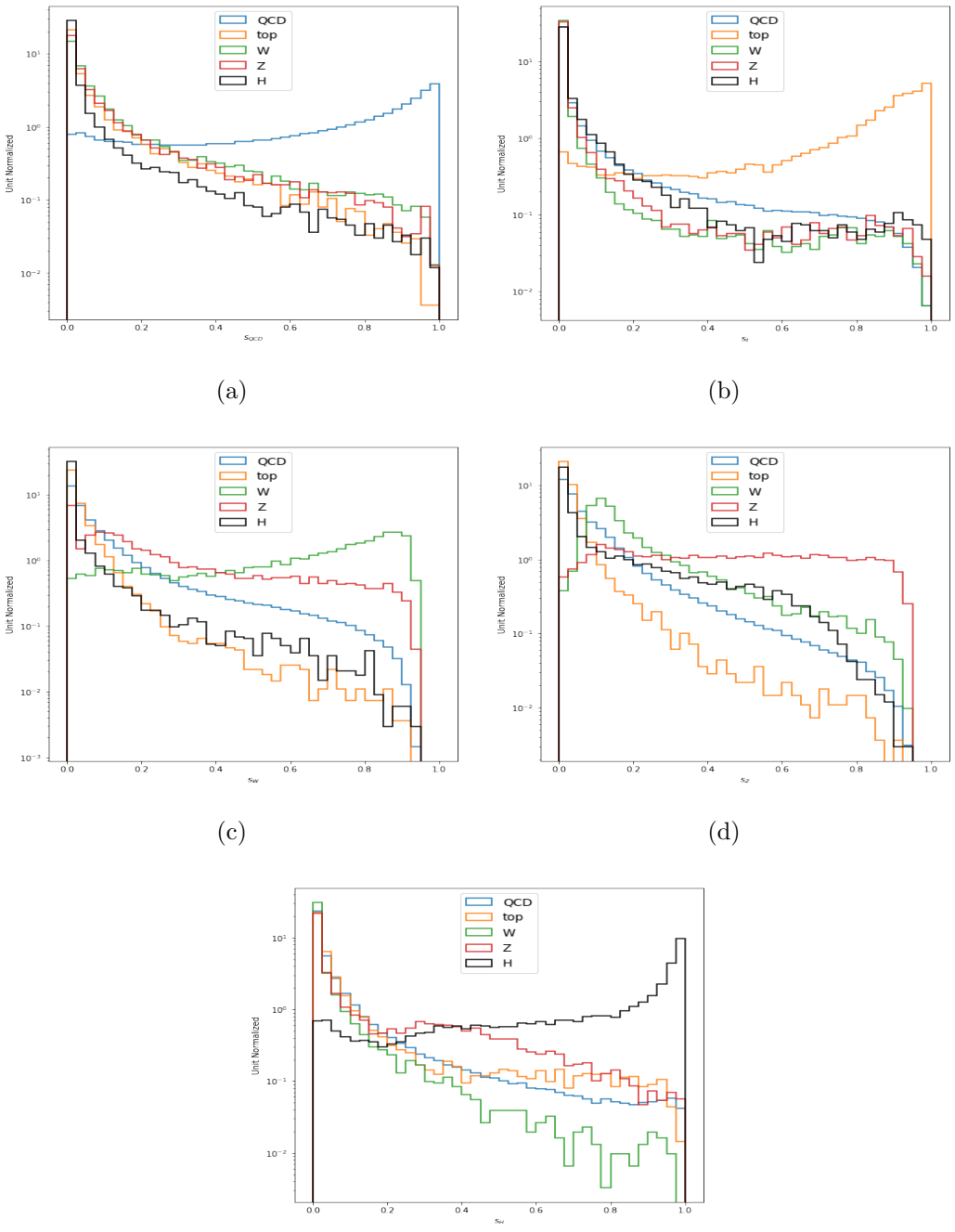
(c)

Figure 7.18: Final results for the DNN model used for the MCT. The training results are presented as the (a) loss function and the (b) accuracy as a function of training epoch. (c) Confusion matrix derived from the application of the MCT to the testing dataset.

large-R jets [178]. Other than for vector bosons, the accuracy of every class shows an improvement of 2-4% in accuracy when trained using UFO jets. Further optimization, in the form of a deeper net as well as hyperparameter tuning, has the potential to increase the performance of the UFO MCT.

MCT Output Scores

Even though the maximum output score indicates what class the jet is being identified as, the taggers can be calibrated in such a way that the efficiency of the tagger is fixed at a value. For a multiclass tagger, this can be done by placing cuts on the score of the class (particle) of interest. Distributions of the scores for the 5 classes are shown in Figure 7.19. For a class i, predicted as a signal, a single cut on the score $s_i > s_{i,\text{cut}}$ fixes the signal efficiency and background rejection. Figure 7.20 shows the background (QCD, top, W, Z) rejection as a function of the Higgs signal efficiency. Comparing with 4.11, we see that the unoptimized MCT, as a single Higgs tagger, is comparable in performance to the double b-tagging approach employed for the prior boosted $H \to b\bar{b}$ analysis. Recent studies [179] have shown that using the log of the ratio of the MCT output scores, inspired by log-likelihood ratios, as a discriminator instead of the score itself provides better background rejection at all signal efficiencies.



(a) (b)

(c) (d)

(e)

Figure 7.19: Multiclass tagger outputs scores for UFO SD CS-SK large-R jets truth matched to (a) QCD multijets, (b) top quarks, (c) W bosons, (d) Z bosons and (e) Higgs bosons.

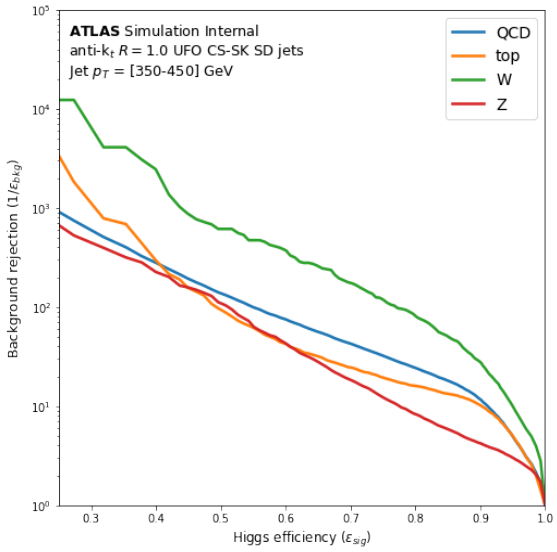
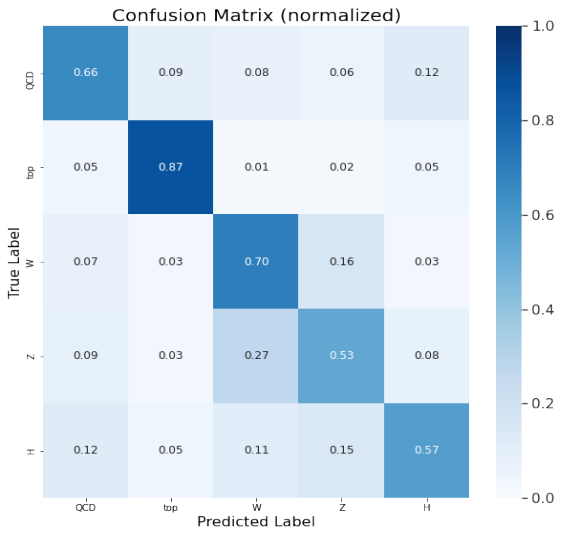


Figure 7.20: Background rejection for all classes (QCD, top quarks, W bosons and Z bosons) as a function of the Higgs boson signal efficiency.

b-tagging Importance

Most of these jet substructure variables are already being used for the binary UFO taggers, providing supporting evidence of their usefulness for the identification of boosted jets reconstructed with the UFO algorithm. For UFO taggers the only variables not currently used for the binary taggers are the DL1r scores of the track jets associated with UFO large-R jets. The DL1r scores are the output of a deep neural network that has been fed all the low-level b-taggers outputs. The low-level algorithms include track jets kinematics, number of tracks, energy fractions of the tracks, number of displaced vertices as well as their distances, invariant masses and more [87]. Given the rich structure provided by that information it is important to study their impact for boosted object tagging. Currently, only the Xbb tagger [180] currently uses the discriminants of the DL1r tagger as an input for their neural network and has shown very promising results as a tool for $H \to b\bar{b}$ tagging aimed at large-R jets constructed from topological clusters. Using the MCT we can measure the impact of including the DL1r scores on our UFO tagger by comparing the confusion matrices resulting



(a)

Figure 7.21: Confusion matrix derived from the application of the MCT to the testing dataset were the MCT was trained without DL1r output scores.

from applying the model with and without including them on the testing dataset. Figure 7.21 shows the confusion matrix resulting from training the MCT without any of the DL1r output scores as input. Negligible impact is seen for the top quark tagging accuracy and for the vector bosons only a 2% decrease in accuracy was observed. The largest impact occurs on the identification of Higgs boson where the accuracy falls from 81% to 57% and for QCD jets where the accuracy falls from 75% to 66%. The false positive rate of QCD jets mistaken as Higgs bosons increases from 4% to 12%. This shows that the information coming from b-tagging the track jets is crucial for building a Higgs tagger. Similarly 11% and 15% of jets associated to a Higgs boson are classified incorrectly as W and Z bosons respectively. A random guessing approach would achieve roughly 20% accuracy for each class (total of 5 classes), therefore, with a 57% accuracy on only relying on jet substructure variables we can conclude that they also provide rich and useful information for boosted jets identification.

For Further Research

The combination of the b-tagger DL1r scores, along with jet substructure variables provide the opportunity to build a versatile and powerful tool for tagging UFO large-R jets. Even with the unoptimized small model studied and described in this document, comparable performance to other taggers was achieved. It is possible that with proper hyperparameter tuning and hidden layers with hundreds of nodes (as the state of the art taggers) that considerable gains in performance could be achieved.

The rate of confusion between the vector bosons could be alleviated by a modification of the model to perform mass decorrelation. This could also help with the mass sculpting that is usually seen in the predicted QCD jets mass distribution. Following what was implemented for the ANN W/Z tagger, the loss fuction should be modified to be of the form

$$L = L_C - \lambda L_{MP} \tag{7.36}$$

where L_C is the original loss function (categorical cross-entropy), L_{MP} is the loss function of a mass predictor network and λ quantifies how strong the penalty for the mass correlation is. The mass predictor network should be trained with the same inputs as the classifier and has to predict the mass of the jet. When minimizing, the new joint loss will improve classification accuracy while preventing mass correlation.

The exploration of discriminants constructed from the MCT outputs should also be studied further. Using the logarithms of the ratios of different scores has been shown to provide better background rejections than just using the raw scores. Using this method, the MCT framework is currently being tested for a VV/VH semi-leptonic analysis by the ATLAS Collaboration [179].

Chapter 8

Conclusions

Boosted $H \to b\bar{b}$ Results

Higgs boson production was studied using its $b\bar{b}$ decay channel at high transverse momenta. The measurements are based on pp collisions at a center-of-mass of $\sqrt{s} = 13$ TeV using the ATLAS detector with a total integrated luminosity of 136 fb^{-1}. Large-R jets reconstructed from topological clusters were used to recontruct the Higgs boson and b-tagging techniques were employed for its identification. Signal strengths of the hadronic decays of the W and Z were used in the validation region and $Z \to b\bar{b}$ in the signal region to validate the experimental techniques with results that agree with the Standard Model.

Upper limits for the Higgs boson cross section at 95% CL were obtained for the fiducial region ($p_{\mathrm{T}} > 450$ GeV and $|y| < 2$) and for four differential regions (250-450, 450-650, 650-1000 and > 1000 GeV).

The Higgs boson signal strength was measured inclusively for jets with $p_{\mathrm{T}} > 250$ GeV by a simultaneous fit to the SRL, SRS and CR$_{t\bar{t}}$ yielding a value of $\mu_H = 0.8 \pm 3.2$.

For the fiducial region, the observed (expected) 95% CL limit was found to be:

$$\sigma_H(p_{\mathrm{T}} > 450\mathrm{GeV}) < 115\,(128)\,\mathrm{fb} \tag{8.1}$$

For the differential region bins, the observed (expected) 95% CL limits were found to be

$$\begin{aligned}
\sigma_H(300 < p_{\mathrm{T}} < 450\,\mathrm{GeV}) &< 2.9\,(3.1)\,\mathrm{pb},\\
\sigma_H(450 < p_{\mathrm{T}} < 650\,\mathrm{GeV}) &< 89\,(102)\,\mathrm{fb},\\
\sigma_H(650 < p_{\mathrm{T}} < 1000\,\mathrm{GeV}) &< 39\,(34)\,\mathrm{fb},\\
\sigma_H(p_{\mathrm{T}} > 1000\,\mathrm{GeV}) &< 9.6\,(7.4)\,\mathrm{fb}.
\end{aligned} \tag{8.2}$$

For $p_{\mathrm{T}} > 1$ TeV the Higgs boson cross section was found to be:

$$\sigma_H(p_{\mathrm{T}} > 1\,\mathrm{TeV}) = 2.3 \pm 3.9\,(\mathrm{stat}) \pm 1.3\,(\mathrm{syst}) \pm 0.5\,(\mathrm{theory})\,\mathrm{fb}. \tag{8.3}$$

All of the results are consistent with the Standard Model predictions. This analysis established a template for future analyses with larger datasets (i.e. Run 3), has produced results for cross sections for Higgs boson production in the high p_{T} regions providing useful inputs for combination analyses and constraints on beyond the standard model physics. The work was published in the ATLAS publication "Constraints on Higgs boson production with large transverse momentum using $H \to b\bar{b}$ decays in the ATLAS dectector" [10] and supported through the internal documentation in the ATLAS note "Study of Higgs boson production at High p_{T}^{H} in the $H \to b\bar{b}$ Decay Mode" [122].

UFO Jets Tagging

Studies supporting ATLAS efforts to develop new tagging frameworks using unified flow object (UFO) jets were performed. First, for the currently supported and already implemented UFO taggers, a strategy to extrapolate the scale factors to higher p_{T} regimes innaccesible with calibration datasets was devised and implemented. Multiple sources were used to estimate the uncertainty associated with the extrapolation procedure. These include renormalization scale variations for ISR and FSR branchings, alternative PDFs, detector geometry variations and alternate nuclear interaction models. The uncertainties were found range from negligible for some cases, to up to 4% at the highest p_{T} accessible (3 TeV). The taggers and working points for which this method was performed were: the DNN inclusive top tagger (50%, 80%), the DNN contained top tagger (50%, 80%), the DNN W tagger (50%, 80%), the ANN W/Z tagger (50%, 60%, 70%, 80%, 90%) and the 3-variable W/Z taggers (50%, 80%).

Finally, studies laying the ground for a dedicated UFO jet multi-class tagger for jets resulting from QCD, W bosons, Z bosons, top quarks and Higgs bosons were performed. The tagger developed was compared to an earlier topocluster-based tagger, showing overall better or equal accuracies for all classes. The tagger is a classifier built as a multi-output deep neural network. Hierarchical clustering based on correlations in conjunction with a permutation importance approach was used to establish the optimal number of features that serve as input. The impact of b-tagging and track jet information was shown to be crucial, specifically for Higgs boson tagging. Even though the calibration of the tagger was out of the scope of this project, it is part of the many results that support the inclusion of jet substructure variables in conjunction with track jet information to achieve better tagging

performance.

The move from trimmed topological clustered jets to soft drop groomed CS-SK UFO jets will improve any analysis that focuses on boosted objects reconstructed as large-R jets due to their improved mass resolution and jet substructure reconstruction capabilities. This new jet input definition, coupled with all the taggers in development, will aid in future iterations of multiple analyses, including the one presented in this thesis (boosted hadronic $H \rightarrow b\bar{b}$), by providing better jet systematics and higher sensitivity by enhancing background rejection. Furthermore, by the end of the LHC's Run 3, it is expected that a total integrated luminosity of 350 fb^{-1} will be delivered, reducing even further the statistical limitations of the analysis.

BIBLIOGRAPHY

[1] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (Sept. 2012), pages 1–29. DOI: 10.1016/j.physletb.2012.08.020. URL: https://doi.org/10.1016%2Fj.physletb.2012.08.020 (cited on page 2).

[2] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (Sept. 2012), pages 30–61. DOI: 10.1016/j.physletb.2012.08.021. URL: https://doi.org/10.1016%2Fj.physletb.2012.08.021 (cited on page 2).

[3] L. Evans and P. Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08001–S08001. DOI: 10.1088/1748-0221/3/08/s08001. URL: https://doi.org/10.1088/1748-0221/3/08/s08001 (cited on pages 2, 24, 33).

[4] CMS Collaboration. "Combined measurements of Higgs boson couplings in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *The European Physical Journal C* 79.5 (May 2019). DOI: 10.1140/epjc/s10052-019-6909-y. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-019-6909-y (cited on page 2).

[5] ATLAS Collaboration. "Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment". In: *Physical Review D* 101.1 (Jan. 2020). DOI: 10.1103/physrevd.101.012002. URL: https://doi.org/10.1103%2Fphysrevd.101.012002 (cited on page 2).

[6] ATLAS Collaboration. "Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector". In: *Physics Letters B* 786 (Nov. 2018), pages 59–86. DOI: 10.1016/j.physletb.2018.09.013. URL: https://doi.org/10.1016%2Fj.physletb.2018.09.013 (cited on page 2).

[7] R. V. Harlander and T. Neumann. "Probing the nature of the Higgs-gluon coupling". In: *Phys. Rev. D* 88 (7 Oct. 2013), page 074015. DOI: 10.1103/PhysRevD.88.074015. URL: https://link.aps.org/doi/10.1103/PhysRevD.88.074015 (cited on page 2).

[8] K. Mimasu, V. Sanz, and C. Williams. "Higher order QCD predictions for associated Higgs production with anomalous couplings to gauge bosons". In: *Journal of High Energy Physics* 2016.8 (Aug. 2016). DOI: 10.1007/jhep08(2016)039. URL: https://doi.org/10.1007%2Fjhep08%282016%29039 (cited on page 2).

[9] D. Kar. *Experimental Particle Physics*. 2053-2563. IOP Publishing, 2019. ISBN: 978-0-7503-2112-9. DOI: 10.1088/2053-2563/ab1be6. URL: https://dx.doi.org/10.1088/2053-2563/ab1be6 (cited on page 3).

[10] ATLAS Collaboration. "Constraints on Higgs boson production with large transverse momentum using $H \to b\bar{b}$ decays in the ATLAS detector". In: *Physical Review D* 105.9 (May 2022). DOI: 10.1103/physrevd.105.092003. URL: https://doi.org/10.1103%2Fphysrevd.105.092003 (cited on pages 3, 81, 85–88, 91, 93, 98, 104, 106, 108–120, 170).

[11] ATLAS Collaboration. *Optimisation of large-radius jet reconstruction for the ATLAS detector in 13 TeV proton-proton collisions*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2020-021. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2723736 (cited on pages 4, 121, 127–129).

[12] S. Weinberg. *The Quantum Theory of Fields*. Press Syndicate of the University of Cambridge, 1995 (cited on pages 5, 6).

[13] M. Loaiza. "A Short Introduction to Hilbert Space Theory". In: *Journal of Physics: Conference Series* 839.1 (May 2017), page 012002. DOI: 10.1088/1742-6596/839/1/012002. URL: https://dx.doi.org/10.1088/1742-6596/839/1/012002 (cited on page 5).

[14] W. Tung. *Group Theory in Physics*. World Scientific Publishing, 1985 (cited on page 6).

[15] Wikimedia Commons. *Standard Model of Elementary Particles*. 2019. URL: https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg (cited on page 17).

[16] R. L. Workman et al. "Review of Particle Physics". In: *PTEP* 2022 (2022), page 083C01. DOI: 10.1093/ptep/ptac097 (cited on pages 19, 20, 25).

[17] G. P. Salam. "Towards jetography". In: *The European Physical Journal C* 67.3-4 (May 2010), pages 637–686. DOI: 10.1140/epjc/s10052-010-1314-6. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-010-1314-6 (cited on pages 18, 57).

[18] N. Cabibbo. "Unitary Symmetry and Leptonic Decays". In: *Phys. Rev. Lett.* 10 (12 June 1963), pages 531–533. DOI: 10.1103/PhysRevLett.10.531. URL: https://link.aps.org/doi/10.1103/PhysRevLett.10.531 (cited on page 20).

[19] J. Riebesell. *Higgs mechanism TikZ graphic*. 2021. URL: https://tikz.net/higgs-potential/ (cited on page 22).

[20] CERN. *CERN Yellow Reports: Monographs, Vol 2 (2017): Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector*. en. 2017. DOI: 10.23731/CYRM-2017-002. URL: https://e-publishing.cern.ch/index.php/CYRM/issue/view/32 (cited on page 24).

[21] None. "Design Report Tevatron 1 Project (Second Printing November 1982)". In: (Oct. 1982). DOI: 10.2172/1413194. URL: https://www.osti.gov/biblio/1413194 (cited on page 24).

[22] LHC Higgs Working Group. *Higgs boson cross-sections as a function of mass*. URL: https://twiki.cern.ch/twiki/pub/LHCPhysics/LHCHWGCrossSectionsFigures/plot_13tev_H_sqrt.pdf (cited on page 26).

[23] ATLAS Collaboration. "Measurement of the Higgs boson mass in the $H \to ZZ* \to 4l$ and $H \to \gamma\gamma$ channels with $\sqrt{s} = 13$ TeV pp collisions using the ATLAS detector". In: *Physics Letters B* 784 (2018), pages 345–366. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2018.07.050. URL: https://www.sciencedirect.com/science/article/pii/S0370269318305884 (cited on page 25).

[24] CMS Collaboration. "A measurement of the Higgs boson mass in the diphoton decay channel". In: *Physics Letters B* 805 (2020), page 135425. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2020.135425. URL: https://www.sciencedirect.com/science/article/pii/S037026932030229X (cited on page 25).

[25] LHC Higgs Working Group. *Higgs boson decay branching ratios as a function of mass*. URL: https://twiki.cern.ch/twiki/pub/LHCPhysics/LHCHWGCrossSectionsFigures/SMHiggsBR.YR4-square.pdf (cited on page 27).

[26] ATLAS Collaboration. "Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector". In: *Physics Letters B* 786 (Nov. 2018), pages 59–86. DOI: 10.1016/j.physletb.2018.09.013. URL: https://doi.org/10.1016%2Fj.physletb.2018.09.013 (cited on pages 26, 73).

[27] CMS Collaboration. "Observation of Higgs Boson Decay to Bottom Quarks". In: *Physical Review Letters* 121.12 (Sept. 2018). DOI: 10.1103/physrevlett.121.121801. URL: https://doi.org/10.1103%2Fphysrevlett.121.121801 (cited on page 26).

[28] M. Buschmann et al. "Mass effects in the Higgs-gluon coupling: boosted vs. off-shell production". In: *Journal of High Energy Physics* 2015 (Feb. 2015), page 38. DOI: 10.1007/JHEP02%282015%29038 (cited on page 28).

[29] A. Banfi et al. "Digging for top squarks from Higgs data: from signal strengths to differential distributions". In: *Journal of High Energy Physics* 2018.11 (Nov. 2018). DOI: 10.1007/jhep11(2018)171. URL: https://doi.org/10.1007%2Fjhep11%282018%29171 (cited on page 27).

[30] C. Grojean et al. "Very boosted Higgs in gluon fusion". In: *Journal of High Energy Physics* 2014.5 (May 2014). DOI: 10.1007/jhep05(2014)022. URL: https://doi.org/10.1007%2Fjhep05%282014%29022 (cited on page 27).

[31] A. Biekötter et al. "Vices and virtues of Higgs effective field theories at large energy". In: *Physical Review D* 91.5 (Mar. 2015). DOI: 10.1103/physrevd.91.055029. URL: https://doi.org/10.1103%2Fphysrevd.91.055029 (cited on page 27).

[32] M. Grazzini et al. "Modeling BSM effects on the Higgs transverse-momentum spectrum in an EFT approach". In: *Journal of High Energy Physics* 2017.3 (Mar. 2017). DOI: 10.1007/jhep03(2017)115. URL: https://doi.org/10.1007%2Fjhep03%282017%29115 (cited on page 27).

[33] M. Grazzini, A. Ilnicka, and M. Spira. "Higgs boson production at large transverse momentum within the SMEFT: analytical results". In: *The European Physical Journal C* 78,808 (Oct. 2018). DOI: 10.1140/epjc/s10052-018-6261-7. URL: https://doi.org/10.1140/epjc/s10052-018-6261-7 (cited on pages 28, 29).

[34] G. Altarelli and G. Parisi. "Asymptotic freedom in parton language". In: *Nuclear Physics B* 126.2 (1977), pages 298–318. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(77)90384-4. URL: https://www.sciencedirect.com/science/article/pii/0550321377903844 (cited on page 30).

[35] T.-J. Hou et al. "New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC". In: *Physical Review D* 103.1 (Jan. 2021). DOI: 10.1103/physrevd.103.014013. URL: https://doi.org/10.1103%2Fphysrevd.103.014013 (cited on page 30).

[36] T. Sjöstrand et al. "High-energy-physics event generation with Pythia 6.1". In: *Computer Physics Communications* 135.2 (Apr. 2001), pages 238–259. DOI: 10.1016/s0010-4655(00)00236-8. URL: https://doi.org/10.1016%2Fs0010-4655%2800%2900236-8 (cited on page 30).

[37] G. Corcella et al. “HERWIG 6: an event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)”. In: *Journal of High Energy Physics* 2001.01 (Jan. 2001), pages 010–010. DOI: 10.1088/1126-6708/2001/01/010. URL: https://doi.org/10.1088%2F1126-6708%2F2001%2F01%2F010 (cited on page 30).

[38] T. Gleisberg et al. “Event generation with SHERPA 1.1”. In: *Journal of High Energy Physics* 2009.02 (Feb. 2009), pages 007–007. DOI: 10.1088/1126-6708/2009/02/007. URL: https://doi.org/10.1088%2F1126-6708%2F2009%2F02%2F007 (cited on page 30).

[39] J. Alwall et al. “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”. In: *Journal of High Energy Physics* 2014.7 (July 2014). DOI: 10.1007/jhep07(2014)079. URL: https://doi.org/10.1007%2Fjhep07%282014%29079 (cited on page 30).

[40] S. Hoche. “Introduction to parton-shower event generators”. In: 2014 (cited on page 31).

[41] S. Agostinelli et al. “Geant4—a simulation toolkit”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pages 250–303. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8. URL: https://www.sciencedirect.com/science/article/pii/S0168900203013688 (cited on pages 31, 80).

[42] ALICE Collaboration. *ALICE Electromagnetic Calorimeter Technical Design Report*. Technical report. 2008. URL: https://cds.cern.ch/record/1121574 (cited on page 34).

[43] LHCb Collaboration. “The LHCb Detector at the LHC”. In: *JINST* 3 (2008). Also published by CERN Geneva in 2010, S08005. DOI: 10.1088/1748-0221/3/08/S08005. URL: https://cds.cern.ch/record/1129809 (cited on page 34).

[44] CMS Collaboration. *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical design report. CMS. There is an error on cover due to a technical problem for some items. Geneva: CERN, 2006. URL: https://cds.cern.ch/record/922757 (cited on page 34).

[45] ATLAS Collaboration. “The ATLAS Experiment at the CERN Large Hadron Collider”. In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08003–S08003. DOI: 10.1088/1748-0221/3/08/s08003. URL: https://doi.org/10.1088/1748-0221/3/08/s08003 (cited on pages 34, 36, 43).

[46] E. Mobs. “The CERN accelerator complex - 2019. Complexe des accélérateurs du CERN - 2019”. In: (July 2019). General Photo. URL: https://cds.cern.ch/record/2684277 (cited on page 34).

[47] ATLAS Collaboration. *Luminosity Public Results Run 2*. 2018. URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2 (visited on 2021) (cited on page 35).

[48] J. Pequenao. "Computer generated image of the whole ATLAS detector". Mar. 2008. URL: https://cds.cern.ch/record/1095924 (cited on page 36).

[49] ATLAS Collaboration. "ATLAS pixel detector electronics and sensors". In: *Journal of Instrumentation* 3.07 (July 2008), P07007. DOI: 10.1088/1748-0221/3/07/P07007. URL: https://dx.doi.org/10.1088/1748-0221/3/07/P07007 (cited on pages 37, 38).

[50] J. Pequenao. "Computer generated image of the ATLAS inner detector 1". 2008. URL: https://cds.cern.ch/record/1095926 (cited on page 38).

[51] G. Lindström et al. "Developments for radiation hard silicon detectors by defect engineering—results by the CERN RD48 (ROSE) Collaboration". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 465.1 (2001). SPD2000, pages 60–69. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(01)00347-3. URL: https://www.sciencedirect.com/science/article/pii/S0168900201003473 (cited on page 38).

[52] M. Capeans et al. *ATLAS Insertable B-Layer Technical Design Report*. Technical report. 2010. URL: https://cds.cern.ch/record/1291633 (cited on page 39).

[53] J. Pequenao. "Computer generated image of the ATLAS inner detector 2". 2008. URL: https://cds.cern.ch/record/1095926 (cited on page 40).

[54] J. Pequenao. "Computer Generated image of the ATLAS calorimeter". 2008. URL: https://cds.cern.ch/record/1095927 (cited on page 41).

[55] F. Cavallari. "Performance of calorimeters at the LHC". In: *Journal of Physics: Conference Series* 293.1 (Apr. 2011), page 012001. DOI: 10.1088/1742-6596/293/1/012001. URL: https://dx.doi.org/10.1088/1742-6596/293/1/012001 (cited on page 42).

[56] ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (Aug. 2008), page 122. DOI: 10.1088/1748-0221/3/08/s08003. URL: https://doi.org/10.1088/1748-0221/3/08/s08003 (cited on page 45).

[57] ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (Aug. 2008), page 127. DOI: 10.1088/1748-0221/3/08/s08003. URL: https://doi.org/10.1088/1748-0221/3/08/s08003 (cited on page 46).

[58] ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (Aug. 2008), page 131. DOI: 10.1088/1748-0221/3/08/s08003. URL: https://doi.org/10.1088/1748-0221/3/08/s08003 (cited on page 46).

[59] J. Pequenao. "Computer generated image of the ATLAS Muons subsystem". 2008. URL: https://cds.cern.ch/record/1095929 (cited on page 48).

[60] J.J. Goodson. "Search for Supersymmetry in States with Large Missing Transverse Momentum and Three Leptons including a Z-Boson". Presented 17 Apr 2012. PhD thesis. Stony Brook University, May 2012 (cited on page 49).

[61] *The ATLAS Collaboration Software and Firmware*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-SOFT-PUB-2021-001. Geneva: CERN, 2021. URL: https://cds.cern.ch/record/2767187 (cited on page 50).

[62] ATLAS Collaboration. "Performance of the ATLAS trigger system in 2015". In: *The European Physical Journal C* 77.5 (May 2017). DOI: 10.1140/epjc/s10052-017-4852-3. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-017-4852-3 (cited on page 50).

[63] T. A. collaboration. "Operation of the ATLAS trigger system in Run 2". In: *Journal of Instrumentation* 15.10 (Oct. 2020), P10004–P10004. DOI: 10.1088/1748-0221/15/10/p10004. URL: https://doi.org/10.1088%2F1748-0221%2F15%2F10%2Fp10004 (cited on page 50).

[64] F. Meloni. *Track and vertex reconstruction in the ATLAS experiment*. Technical report. Geneva: CERN, 2012. URL: https://cds.cern.ch/record/1455427 (cited on pages 52, 54).

[65] N. Braun. "Combinatorial Kalman Filter". In: *Combinatorial Kalman Filter and High Level Trigger Reconstruction for the Belle II Experiment*. Cham: Springer International Publishing, 2019, pages 117–174. ISBN: 978-3-030-24997-7. DOI: 10.1007/978-3-030-24997-7_6. URL: https://doi.org/10.1007/978-3-030-24997-7_6 (cited on page 52).

[66] R. O. Duda and P. E. Hart. "Use of the Hough Transformation to Detect Lines and Curves in Pictures". In: *Commun. ACM* 15.1 (Jan. 1972), pages 11–15. ISSN: 0001-0782. DOI: 10.1145/361237.361242. URL: https://doi.org/10.1145/361237.361242 (cited on page 53).

[67] ATLAS Collaboration. *Performance of the ATLAS Inner Detector Track and Vertex Reconstruction in the High Pile-Up LHC Environment*. Tech-

nical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2012-042. Geneva: CERN, 2012. URL: https://cds.cern.ch/record/1435196 (cited on page 53).

[68] M. Cacciari, G. P. Salam, and G. Soyez. "The anti-kt jet clustering algorithm". In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), pages 063–063. DOI: 10.1088/1126-6708/2008/04/063. URL: https://doi.org/10.1088%2F1126-6708%2F2008%2F04%2F063 (cited on pages 54, 56).

[69] ATLAS Collaboration. "Electron reconstruction and identification efficiency measurements with the ATLAS detector using the 2011 LHC proton–proton collision data". In: *The European Physical Journal C* 74.7 (July 2014). DOI: 10.1140/epjc/s10052-014-2941-0. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-014-2941-0 (cited on page 56).

[70] ATLAS Collaboration. *Electron and photon reconstruction and performance in ATLAS using a dynamical, topological cell clustering-based approach*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2017-022. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2298955 (cited on pages 56, 61).

[71] ATLAS Collaboration. "Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1". In: *The European Physical Journal C* 77.7 (July 2017). DOI: 10.1140/epjc/s10052-017-5004-5. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-017-5004-5 (cited on page 58).

[72] ATLAS Collaboration. *Optimisation of large-radius jet reconstruction for the ATLAS detector in 13 TeV proton-proton collisions*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2020-021. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2723736 (cited on pages 59, 60).

[73] J. Varela. *Boosted jets: Increasing transverse momentum*. 2019. URL: https://www.nevis.columbia.edu/reu/2019/talks/Varela_slides.pdf (cited on page 59).

[74] D. Krohn, J. Thaler, and L.-T. Wang. "Jet trimming". In: *Journal of High Energy Physics* 2010.2 (Feb. 2010). DOI: 10.1007/jhep02(2010)084. URL: https://doi.org/10.1007%2Fjhep02%282010%29084 (cited on page 60).

[75] ATLAS Collaboration. "Performance of jet substructure techniques for large-R jets in proton-proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector". In: *Journal of High Energy Physics* 2013.9 (Sept. 2013). DOI: 10.1007/jhep09(2013)076. URL: https://doi.org/10.1007%2Fjhep09%282013%29076 (cited on page 61).

[76] ATLAS Collaboration. "Jet energy measurement with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7$ TeV". In: *The European Physical Journal C* 73.3 (Mar. 2013). DOI: 10.1140/epjc/s10052-013-2304-2. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-013-2304-2 (cited on page 61).

[77] ATLAS Collaboration. "In situ calibration of large-radius jet energy and mass in 13 TeV proton-proton collisions with the ATLAS detector". In: *The European Physical Journal C* 79.2 (Feb. 2019). DOI: 10.1140/epjc/s10052-019-6632-8. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-019-6632-8 (cited on pages 62, 63).

[78] ATLAS Collaboration. *Jet mass reconstruction with the ATLAS Detector in early Run 2 data*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2016-035. Geneva: CERN, 2016. URL: https://cds.cern.ch/record/2200211 (cited on page 63).

[79] D. Krohn, J. Thaler, and L.-T. Wang. "Jets with variable R". In: *Journal of High Energy Physics* 2009.06 (June 2009), pages 059–059. DOI: 10.1088/1126-6708/2009/06/059. URL: https://doi.org/10.1088%2F1126-6708%2F2009%2F06%2F059 (cited on page 64).

[80] ATLAS Collaboration. *Variable Radius, Exclusive-k_T, and Center-of-Mass Subjet Reconstruction for Higgs($\to b\bar{b}$) Tagging in ATLAS*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2017-010. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2268678 (cited on page 65).

[81] ATLAS Collaboration. "Performance of jet substructure techniques for large-R jets in proton-proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector". In: *Journal of High Energy Physics* 2013.9 (Sept. 2013). DOI: 10.1007/jhep09(2013)076. URL: https://doi.org/10.1007%2Fjhep09%282013%29076 (cited on page 65).

[82] ATLAS Collaboration. "Performance of b-jet identification in the ATLAS experiment". In: *Journal of Instrumentation* 11.04 (Apr. 2016), P04008–P04008. DOI: 10.1088/1748-0221/11/04/p04008. URL: https://doi.org/10.1088%2F1748-0221%2F11%2F04%2Fp04008 (cited on page 66).

[83] ATLAS Collaboration. "ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV". In: *The European Physical Journal C* 79.11 (Nov. 2019). DOI: 10.1140/epjc/s10052-019-7450-8. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-019-7450-8 (cited on pages 66, 68, 69).

[84] J. Varela. *Track jet schematic with a displaced vertex*. 2019. URL: https://www.nevis.columbia.edu/reu/2019/talks/Varela_slides.pdf (cited on page 67).

[85] ATLAS Collaboration. *Secondary vertex finding for jet flavour identification with the ATLAS detector*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2017-011. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2270366 (cited on page 66).

[86] ATLAS Collaboration. *Topological b-hadron decay reconstruction and identification of b-jets with the JetFitter package in the ATLAS experiment at the LHC*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2018-025. Geneva: CERN, 2018. URL: https://cds.cern.ch/record/2645405 (cited on page 66).

[87] ATLAS Collaboration. "ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV". In: *The European Physical Journal C* 79.11 (Nov. 2019). DOI: 10.1140/epjc/s10052-019-7450-8. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-019-7450-8 (cited on pages 67, 82, 166).

[88] A. Hoecker et al. *TMVA - Toolkit for Multivariate Data Analysis*. 2007. DOI: 10.48550/ARXIV.PHYSICS/0703039. URL: https://arxiv.org/abs/physics/0703039 (cited on page 67).

[89] F. Chollet et al. *Keras*. https://keras.io. 2015 (cited on page 67).

[90] The Theano Development Team et al. *Theano: A Python framework for fast computation of mathematical expressions*. 2016. DOI: 10.48550/ARXIV.1605.02688. URL: https://arxiv.org/abs/1605.02688 (cited on page 67).

[91] ATLAS Collaboration. "Identification of boosted Higgs bosons decaying into b-quark pairs with the ATLAS detector at 13 TeV". In: *Eur. Phys. J. C* 79.10 (2019). 54 pages in total, author list starting page 38, 20 figures, 3 tables. All figures including auxiliary figures are available at http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/PERF-2017-04,

page 836. DOI: 10.1140/epjc/s10052-019-7335-x. arXiv: 1906.11005. URL: https://cds.cern.ch/record/2680245 (cited on pages 69, 70, 82).

[92] ATLAS Collaboration. "Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$=13 TeV. Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$=13 TeV". In: *Eur. Phys. J. C* 76 (2016). Comments: 27 pages inlcuding cover page plus author list (45 pages total), 12 figures, 3 tables, submitted to Eur. Phys. J. C. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/PERF-2015-10/, page 292. DOI: 10.1140/epjc/s10052-016-4120-y. arXiv: 1603.05598. URL: https://cds.cern.ch/record/2139897 (cited on pages 70–72, 82).

[93] ATLAS Collaboration. "Measurements of Higgs Bosons Decaying to Bottom Quarks from Vector Boson Fusion Production with the ATLAS Experiment at $\sqrt{s} = 13$ TeV". In: *European Physics Journal* C.81 (Nov. 2020). DOI: 10.1140/epjc/s10052-021-09192-8. URL: https://doi.org/10.48550/arXiv.2011.08280 (cited on page 73).

[94] ATLAS Collaboration. "Search for resonances in the mass distribution of jet pairs with one or two jets identified as b-jets in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Rev. D* 98 (3 Aug. 2018), page 032016. DOI: 10.1103/PhysRevD.98.032016. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.032016 (cited on page 73).

[95] G. Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (Feb. 2011). DOI: 10.1140/epjc/s10052-011-1554-0. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-011-1554-0 (cited on page 76).

[96] *Good Run Lists for Analysis - Run 2*. Dec. 2022. URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/GoodRunListsForAnalysisRun2 (cited on page 77).

[97] ATLAS Collaboration. *Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2019-021. Geneva: CERN, 2019. URL: https://cds.cern.ch/record/2677054 (cited on pages 77, 78).

[98] ATLAS Collaboration. "Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector". In: *The European Physical Journal C* 76.11 (Oct. 2016). DOI: 10.1140/epjc/s10052-016-4395-z (cited on page 78).

[99] ATLAS Collaboration. "Performance of the ATLAS muon triggers in Run 2". In: *Journal of Instrumentation* 15.09 (Sept. 2020), P09015–P09015. DOI: 10.1088/1748-0221/15/09/p09015. URL: https://doi.org/10.1088%2F1748-0221%2F15%2F09%2Fp09015 (cited on page 79).

[100] K. Hamilton, P. Nason, and G. Zanderighi. "MINLO: multi-scale improved NLO". In: *Journal of High Energy Physics* 2012.10 (Oct. 2012). DOI: 10.1007/jhep10(2012)155. URL: https://doi.org/10.1007%2Fjhep10%282012%29155 (cited on page 79).

[101] S. Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *Journal of High Energy Physics* 2010.6 (June 2010). DOI: 10.1007/jhep06(2010)043. URL: https://doi.org/10.1007%2Fjhep06%282010%29043 (cited on page 80).

[102] K. Hamilton, P. Nason, and G. Zanderighi. *Finite quark-mass effects in the NNLOPS POWHEG+MiNLO Higgs generator*. 2015. DOI: 10.48550/ARXIV.1501.04637. URL: https://arxiv.org/abs/1501.04637 (cited on page 80).

[103] P. Nason and C. Oleari. "NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG". In: *Journal of High Energy Physics* 2010.2 (Feb. 2010). DOI: 10.1007/jhep02(2010)037. URL: https://doi.org/10.1007%2Fjhep02%282010%29037 (cited on page 80).

[104] H. Hartanto et al. "Higgs boson production in association with top quarks in the POWHEG BOX". In: *Physical Review D* 91.9 (May 2015). DOI: 10.1103/physrevd.91.094003. URL: https://doi.org/10.1103%2Fphysrevd.91.094003 (cited on page 80).

[105] L. Chen et al. "ZH production in gluon fusion at NLO in QCD". In: *Journal of High Energy Physics* 2022.8 (Aug. 2022). DOI: 10.1007/jhep08(2022)056. URL: https://doi.org/10.1007%2Fjhep08%282022%29056 (cited on page 80).

[106] G. Luisoni et al. "HW ± HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO". In: *Journal of High Energy Physics* 2013.10 (Oct. 2013). DOI: 10.1007/jhep10(2013)083. URL: https://doi.org/10.1007%2Fjhep10%282013%29083 (cited on page 80).

[107] G. Cullen et al. "Automated one-loop calculations with GoSam". In: *The European Physical Journal C* 72.3 (Mar. 2012). DOI: 10.1140/epjc/s10052-012-1889-1. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-012-1889-1 (cited on page 80).

[108] A. Denner et al. "HAWK 2.0: A Monte Carlo program for Higgs production in vector-boson fusion and Higgs strahlung at hadron colliders". In: *Computer Physics Com-*

munications 195 (2015), pages 161–171. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2015.04.021. URL: https://www.sciencedirect.com/science/article/pii/S0010465515001630 (cited on page 80).

[109] A. Djouadi, J. Kalinowski, and M. Spira. "HDECAY: a program for Higgs boson decays in the Standard Model and its supersymmetric extension". In: *Computer Physics Communications* 108.1 (Jan. 1998), pages 56–74. DOI: 10.1016/s0010-4655(97)00123-9. URL: https://doi.org/10.1016%2Fs0010-4655%2897%2900123-9 (cited on page 80).

[110] A. Bredenstein et al. "Precision calculations for the Higgs decays H $\to$ ZZ/WW $\to$ 4 leptons". In: *Nuclear Physics B - Proceedings Supplements* 160 (Oct. 2006), pages 131–135. DOI: 10.1016/j.nuclphysbps.2006.09.104. URL: https://doi.org/10.1016%2Fj.nuclphysbps.2006.09.104 (cited on page 80).

[111] E. Bothmann et al. "Event generation with Sherpa 2.2". In: *SciPost Physics* 7.3 (Sept. 2019). DOI: 10.21468/scipostphys.7.3.034. URL: https://doi.org/10.21468%2Fscipostphys.7.3.034 (cited on page 80).

[112] E. Re. "Single-top Wt-channel production matched with parton showers using the POWHEG method". In: *The European Physical Journal C* 71.2 (Feb. 2011). DOI: 10.1140/epjc/s10052-011-1547-z. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-011-1547-z (cited on page 80).

[113] R. Frederix, E. Re, and P. Torrielli. "Single-top t-channel hadroproduction in the four-flavour scheme with POWHEG and aMC@NLO". In: *Journal of High Energy Physics* 2012.9 (Sept. 2012). DOI: 10.1007/jhep09(2012)130. URL: https://doi.org/10.1007%2Fjhep09%282012%29130 (cited on page 80).

[114] T. Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (June 2015), pages 159–177. DOI: 10.1016/j.cpc.2015.01.024. URL: https://doi.org/10.1016%2Fj.cpc.2015.01.024 (cited on page 80).

[115] *The Pythia 8 A3 tune description of ATLAS minimum bias and inelastic measurements incorporating the Donnachie-Landshoff diffractive model*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2016-017. Geneva: CERN, 2016. URL: https://cds.cern.ch/record/2206965 (cited on page 80).

[116] M. Cacciari, G. P. Salam, and G. Soyez. "FastJet user manual". In: *The European Physical Journal C* 72.3 (Mar. 2012). DOI: 10.1140/epjc/s10052-012-1896-2. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-012-1896-2 (cited on page 81).

[117] Ahmed Tarek. *EW Corrections using HAWK*. Dec. 2019. URL: https://indico.cern.ch/event/870440/contributions/3671095/subcontributions/295099/attachments/1961576/3260333/Hbb_boosted_EW_12dec19.pdf (cited on page 88).

[118] S. Frixione et al. "Single-top hadroproduction in association with a W boson". In: *Journal of High Energy Physics* 2008.07 (July 2008), pages 029–029. DOI: 10.1088/1126-6708/2008/07/029. URL: https://doi.org/10.1088%2F1126-6708%2F2008%2F07%2F029 (cited on page 92).

[119] J. M. Lindert et al. "Precise predictions for $V+$ jets dark matter backgrounds". In: *The European Physical Journal C* 77.12 (Dec. 2017). DOI: 10.1140/epjc/s10052-017-5389-1. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-017-5389-1 (cited on page 94).

[120] L. A. Harland-Lang et al. "Parton distributions in the LHC era: MMHT 2014 PDFs". In: *The European Physical Journal C* 75.5 (May 2015). DOI: 10.1140/epjc/s10052-015-3397-6. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-015-3397-6 (cited on page 95).

[121] T. Sjöstrand. "Jet fragmentation of multiparton configurations in a string framework". In: *Nuclear Physics B* 248.2 (1984), pages 469–502. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(84)90607-2. URL: https://www.sciencedirect.com/science/article/pii/0550321384906072 (cited on page 95).

[122] M. Battaglia et al. *Measurement of Inclusive Higgs Boson Production at High p_T^H in the $H \to b\bar{b}$ Decay Mode*. Technical report. Geneva: CERN, 2019. URL: https://cds.cern.ch/record/2703097 (cited on pages 99, 101, 102, 170).

[123] G. Cowan. *Statistical data analysis*. Oxford University Press, USA, 1998 (cited on page 101).

[124] L. Moneta et al. *The RooStats Project*. 2011. arXiv: 1009.1003 [physics.data-an] (cited on page 105).

[125] W. Verkerke and D. Kirkby. *The RooFit toolkit for data modeling*. 2003. arXiv: physics/0306116 [physics.data-an] (cited on page 105).

[126] V. Gligorov et al. "Avoiding biases in binned fits". In: *Journal of Instrumentation* 16.08 (Aug. 2021), T08004. DOI: 10.1088/1748-0221/16/08/t08004. URL: https://doi.org/10.1088%2F1748-0221%2F16%2F08%2Ft08004 (cited on page 105).

[127] R. Barlow and C. Beeston. "Fitting using finite Monte Carlo samples". In: *Computer Physics Communications* 77.2 (1993), pages 219–228. ISSN: 0010-4655. DOI: https://

doi.org/10.1016/0010-4655(93)90005-W. URL: https://www.sciencedirect.com/science/article/pii/001046559390005W (cited on page 105).

[128] M. Stankaityte. "Probing high energy Higgs bosons using $H \to b\bar{b}$ decays in $\sqrt{s} = 13$ TeV proton-proton collisions at the ATLAS experiment". Presented 03 Dec 2021. 2022. URL: https://cds.cern.ch/record/2800522 (cited on page 106).

[129] CERN. *CERN Yellow Reports: Monographs, Vol 2 (2017): Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector*. en. 2017. DOI: 10.23731/CYRM-2017-002. URL: https://e-publishing.cern.ch/index.php/CYRM/issue/view/32 (cited on page 107).

[130] N. Berger et al. *Simplified Template Cross Sections - Stage 1.1*. 2019. arXiv: 1906.02754 [hep-ph] (cited on page 107).

[131] ATLAS Collaboration. "Measurements of top-quark pair differential and double-differential cross-sections in the l+ jets channel with pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector". In: *The European Physical Journal C* 79.12 (Dec. 2019). DOI: 10.1140/epjc/s10052-019-7525-6. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-019-7525-6 (cited on pages 108, 109).

[132] *W, Z, and Higgs bosons as portals to exotic physics - CMS Experiment*. URL: https://cms.cern/news/w-z-and-higgs-bosons-portals-exotic-physics.

[133] *Hadronic Physics*. URL: https://www.anl.gov/phy/hadronic-physics.

[134] ATLAS Collaboration. "Jet reconstruction and performance using particle flow with the ATLAS Detector". In: *The European Physical Journal C* 77.7 (July 2017). DOI: 10.1140/epjc/s10052-017-5031-2. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-017-5031-2 (cited on pages 121, 122).

[135] ATLAS Collaboration. *Improving jet substructure performance in ATLAS using Track-CaloClusters*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2017-015. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2275636 (cited on pages 121, 123).

[136] P. Berta et al. "Particle-level pileup subtraction for jets and jet shapes". In: *Journal of High Energy Physics* 2014.6 (June 2014). DOI: 10.1007/jhep06(2014)092. URL: https://doi.org/10.1007%2Fjhep06%282014%29092 (cited on pages 122, 124).

[137] M. Cacciari, G. P. Salam, and G. Soyez. "SoftKiller, a particle-level pileup removal method". In: *The European Physical Journal C* 75.2 (Feb. 2015). DOI: 10.1140/epjc/s10052-015-3267-2. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-015-3267-2 (cited on pages 122, 125).

[138] A. J. Larkoski et al. "Soft drop". In: *Journal of High Energy Physics* 2014.5 (May 2014). DOI: 10.1007/jhep05(2014)146. URL: https://doi.org/10.1007%2Fjhep05%282014%29146 (cited on pages 122, 126).

[139] CMS Collaboration. "Pileup mitigation at CMS in 13 TeV data". In: *Journal of Instrumentation* 15.09 (Sept. 2020), P09018–P09018. DOI: 10.1088/1748-0221/15/09/p09018. URL: https://doi.org/10.1088%2F1748-0221%2F15%2F09%2Fp09018 (cited on page 123).

[140] S. Marzani, L. Schunk, and G. Soyez. "A study of jet mass distributions with grooming". In: *Journal of High Energy Physics* 2017.7 (July 2017). DOI: 10.1007/jhep07(2017)132. URL: https://doi.org/10.1007%2Fjhep07%282017%29132 (cited on page 126).

[141] A. J. Larkoski, I. Moult, and B. Nachman. "Jet substructure at the Large Hadron Collider: A review of recent advances in theory and machine learning". In: *Physics Reports* 841 (Jan. 2020), pages 1–63. DOI: 10.1016/j.physrep.2019.11.001. URL: https://doi.org/10.1016%2Fj.physrep.2019.11.001 (cited on page 128).

[142] R. K. Ellis, W. J. Stirling, and B. R. Webber. *QCD and Collider Physics*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 1996. DOI: 10.1017/CBO9780511628788 (cited on page 130).

[143] S. Marzani, G. Soyez, and M. Spannowsky. *Looking Inside Jets*. Springer International Publishing, 2019. DOI: 10.1007/978-3-030-15709-8. URL: https://doi.org/10.1007%2F978-3-030-15709-8 (cited on page 130).

[144] C. F. Berger, T. Kucs, and G. Sterman. "Event shape–energy flow correlations". In: *Physical Review D* 68.1 (July 2003). DOI: 10.1103/physrevd.68.014012. URL: https://doi.org/10.1103%2Fphysrevd.68.014012 (cited on page 130).

[145] A. J. Larkoski, G. P. Salam, and J. Thaler. "Energy correlation functions for jet substructure". In: *Journal of High Energy Physics* 2013.6 (June 2013). DOI: 10.1007/jhep06(2013)108. URL: https://doi.org/10.1007%2Fjhep06%282013%29108 (cited on page 131).

[146] I. Moult, L. Necib, and J. Thaler. "New angles on energy correlation functions". In: *Journal of High Energy Physics* 2016.12 (Dec. 2016). DOI: 10.1007/jhep12(2016)153. URL: https://doi.org/10.1007%2Fjhep12%282016%29153 (cited on page 132).

[147] L. G. Almeida et al. "Substructure of high-p_T jets at the LHC". In: *Physical Review D* 79.7 (Apr. 2009). DOI: 10.1103/physrevd.79.074017. URL: https://doi.org/10.1103%2Fphysrevd.79.074017 (cited on page 133).

[148] T. M. Mitchell. *Machine learning*. Volume 1. 9. McGraw-hill New York, 1997 (cited on page 133).

[149] D. M. D'Addona. *CIRP Encyclopedia of Production Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pages 911–918. ISBN: 978-3-642-20617-7. DOI: 10.1007/978-3-642-20617-7_6563. URL: https://doi.org/10.1007/978-3-642-20617-7_6563 (cited on page 134).

[150] J. Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural Networks* 61 (Jan. 2015), pages 85–117. DOI: 10.1016/j.neunet.2014.09.003. URL: https://doi.org/10.1016%2Fj.neunet.2014.09.003 (cited on pages 135, 136).

[151] C. Zatout. *A brief introduction to neural networks: A classification problem*. Feb. 2023. URL: https://towardsdatascience.com/a-brief-introduction-to-neural-networks-a-classification-problem-43e68c770081 (cited on page 135).

[152] *Identification of hadronically-decaying top quarks using UFO jets with ATLAS in Run 2*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2021-028. Geneva: CERN, 2021. URL: https://cds.cern.ch/record/2776782 (cited on pages 136, 137).

[153] C. Chen. "New approach to identifying boosted hadronically decaying particles using jet substructure in its center-of-mass frame". In: *Physical Review D* 85.3 (Feb. 2012). DOI: 10.1103/physrevd.85.034007. URL: https://doi.org/10.1103%2Fphysrevd.85.034007 (cited on page 136).

[154] ATLAS Collaboration. "Measurement of k_T splitting scales in $W \to l\nu$ events at $\sqrt{s} = 7$ TeV with the ATLAS detector". In: *The European Physical Journal C* 73.5 (May 2013). DOI: 10.1140/epjc/s10052-013-2432-8. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-013-2432-8 (cited on page 136).

[155] ATLAS Collaboration. *Performance of W/Z taggers using UFO jets in ATLAS*. Technical report. All figures including auxiliary figures are available at

https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2021-029. Geneva: CERN, 2021. URL: https://cds.cern.ch/record/2777009 (cited on pages 137, 138).

[156] I. Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Edited by Z. Ghahramani et al. Volume 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf (cited on page 137).

[157] M. Das-Gupta and G. P. Salam. "Event shapes in e+e- annihilation and deep inelastic scattering". In: *J. Phys. G* 30 (2004), page 143. URL: https://cds.cern.ch/record/694105 (cited on page 137).

[158] L. G. Almeida et al. "Top quark jets at the LHC". In: *Physical Review D* 79.7 (Apr. 2009). DOI: 10.1103/physrevd.79.074012. URL: https://doi.org/10.1103%2Fphysrevd.79.074012 (cited on page 137).

[159] ATLAS Collaboration. "ATLAS measurements of the properties of jets for boosted particle searches". In: *Physical Review D* 86.7 (Oct. 2012). DOI: 10.1103/physrevd.86.072006. URL: https://doi.org/10.1103%2Fphysrevd.86.072006 (cited on page 137).

[160] C. Chen. "New approach to identifying boosted hadronically decaying particles using jet substructure in its center-of-mass frame". In: *Physical Review D* 85.3 (Feb. 2012). DOI: 10.1103/physrevd.85.034007. URL: https://doi.org/10.1103%2Fphysrevd.85.034007 (cited on page 137).

[161] J. Thaler and L.-T. Wang. "Strategies to identify boosted tops". In: *Journal of High Energy Physics* 2008.07 (July 2008), pages 092–092. DOI: 10.1088/1126-6708/2008/07/092. URL: https://doi.org/10.1088%2F1126-6708%2F2008%2F07%2F092 (cited on page 137).

[162] S. Catani et al. "Longitudinally-invariant k_t-clustering algorithms for hadron-hadron collisions". In: *Nucl. Phys. B* 406 (1993), pages 187–224. DOI: 10.1016/0550-3213(93)90166-M. URL: https://cds.cern.ch/record/246812 (cited on page 137).

[163] *Simulation-based extrapolation of b-tagging calibrations towards high transverse momenta in the ATLAS experiment*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2021-003. Geneva: CERN, 2021. URL: https://cds.cern.ch/record/2753444 (cited on page 139).

[164] S. Mrenna and P. Skands. "Automated parton-shower variations in pythia 8". In: *Physical Review D* 94.7 (Oct. 2016). DOI: 10.1103/physrevd.94.074005. URL: https://doi.org/10.1103%2Fphysrevd.94.074005 (cited on page 143).

[165] URL: https://geant4-userdoc.web.cern.ch/UsersGuides/PhysicsListGuide/html/index.html (cited on page 145).

[166] G. Folger and J. P. Wellisch. *String Parton Models in Geant4*. 2003. arXiv: nucl-th/0306007 [nucl-th] (cited on page 145).

[167] D. Wright and M. Kelsey. "The Geant4 Bertini Cascade". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 804 (2015), pages 175–188. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2015.09.058. URL: https://www.sciencedirect.com/science/article/pii/S0168900215011134 (cited on page 145).

[168] J. Yarba. "Recent Developments and Validation of Geant4 Hadronic Physics". In: *Journal of Physics: Conference Series* 396.2 (Dec. 2012), page 022060. DOI: 10.1088/1742-6596/396/2/022060. URL: https://dx.doi.org/10.1088/1742-6596/396/2/022060 (cited on page 145).

[169] M. V. Kossov. "Chiral Invariant Phase Space model". In: *Eur. Phys. J. A* 36 (2008), pages 289–293. DOI: 10.1140/epja/i2008-10597-2 (cited on page 145).

[170] CMS Collaboration. "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques". In: *Journal of Instrumentation* 15.06 (June 2020), P06005–P06005. DOI: 10.1088/1748-0221/15/06/p06005. URL: https://doi.org/10.1088%2F1748-0221%2F15%2F06%2Fp06005 (cited on page 157).

[171] L. Randall and R. Sundrum. "Large Mass Hierarchy from a Small Extra Dimension". In: *Physical Review Letters* 83.17 (Oct. 1999), pages 3370–3373. DOI: 10.1103/physrevlett.83.3370. URL: https://doi.org/10.1103%2Fphysrevlett.83.3370 (cited on page 157).

[172] L. Breiman. "Random Forests". English. In: *Machine Learning* 45.1 (2001), pages 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A%3A1010933404324 (cited on page 158).

[173] *Ward's Linkage*. URL: https://www.statistics.com/glossary/wards-linkage/ (cited on page 160).

[174] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/ (cited on page 162).

[175] A. F. Agarap. *Deep Learning using Rectified Linear Units (ReLU)*. cite arxiv:1803.08375Comment: 7 pages, 11 figures, 9 tables. 2018. URL: http://arxiv.org/abs/1803.08375 (cited on page 162).

[176] A. E. Hoerl and R. W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 42.1 (2000), pages 80–86. ISSN: 00401706. URL: http://www.jstor.org/stable/1271436 (visited on 07/17/2023) (cited on page 162).

[177] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG] (cited on page 162).

[178] Garabed Halladjian. *Estimation of top Scale Factors for Boosted Multi-Class Tagger using profile-likelihood fits*. Apr. 2019. URL: https://indico.cern.ch/event/910074/contributions/3828338/attachments/2021352/3379938/slides-2020-04-16.pdf (cited on page 164).

[179] Maria Mazza. *Multi-class W/Z/h jet tagging*. May 2023. URL: https://indico.cern.ch/event/1267254/contributions/5363497/attachments/2656595/4600890/DBLWorkshop_230531_MariaMazza.pdf (cited on pages 164, 168).

[180] *Identification of Boosted Higgs Bosons Decaying Into $b\bar{b}$ With Neural Networks and Variable Radius Subjets in ATLAS*. Technical report. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2020-019. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2724739 (cited on page 166).

[181] M. Murin. *Signal Scale Factors in semileptonic ttbar for the new UFO jets W/top taggers*. Oct. 2021. URL: https://indico.cern.ch/event/1089531/contributions/4580167/attachments/2332414/3975094/2021-10-21_JetEtmiss_WTopTagSF_presentation.pdf (cited on pages 244–250).

APPENDIX A

$H \to b\bar{b}$ Analysis

Large-R Jet Triggers

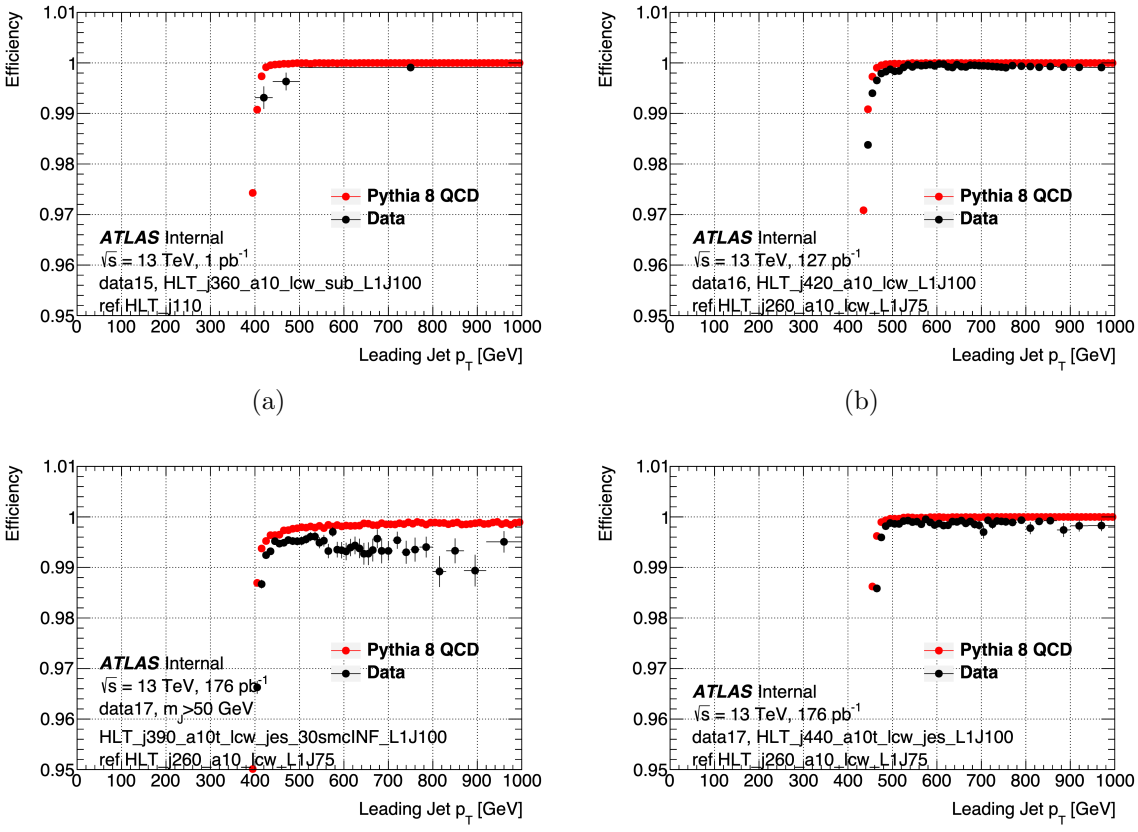


Figure A.1: Large-R jets trigger efficiency curves for the (a) 2015, (b) 2016 and (c,d) 2017 triggers as a function of the leading large-R jet p_{T}.

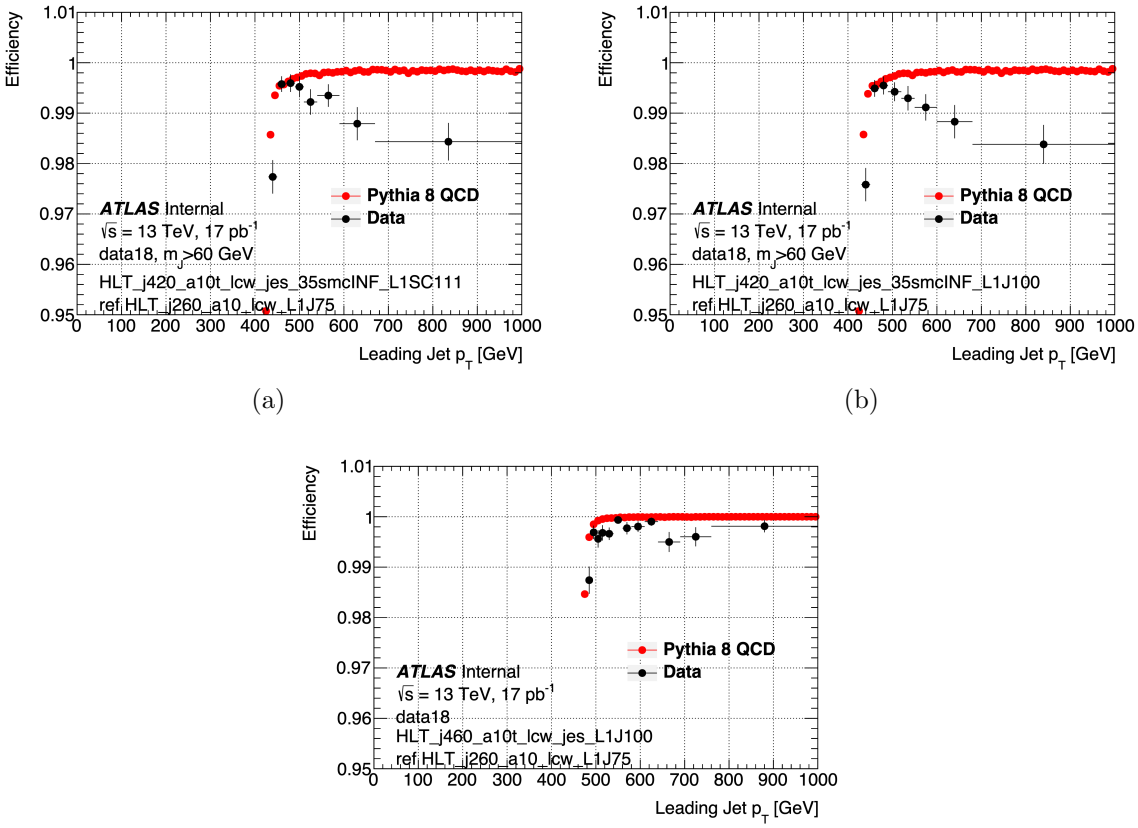


Figure A.2: Large-R jets trigger efficiency curves for the 2018 triggers as a function of the leading large-R jet p_{T}.

Higgs Shower Systematics

Parton showers rely on the factorization of a process into a hard scattering, a perturbative shower describing soft collinear emissions, and a non-perturbative hadronization process into final state hadrons. The two main parton shower models are the angular-ordered model used by PYTHIA, and the dipole showering model used by Herwig. The choice of parton shower model is considered a source of systematic uncertainty for the signal.

To estimate the uncertainty associated with the choice of parton shower model, MC samples using POWHEG+PYTHIA8 (nominal) and POWHEG+Herwig7 (alternate) were produced. The samples included ggF, VBF, WH, $qq \to ZH$ and $gg \to ZH$ with all the possible W and Z decays. For the nominal samples at reconstruction level (full simulation), the samples described in Chapter 5 were used. A private production of Herwig7 samples at truth level was produced as the alternative samples.

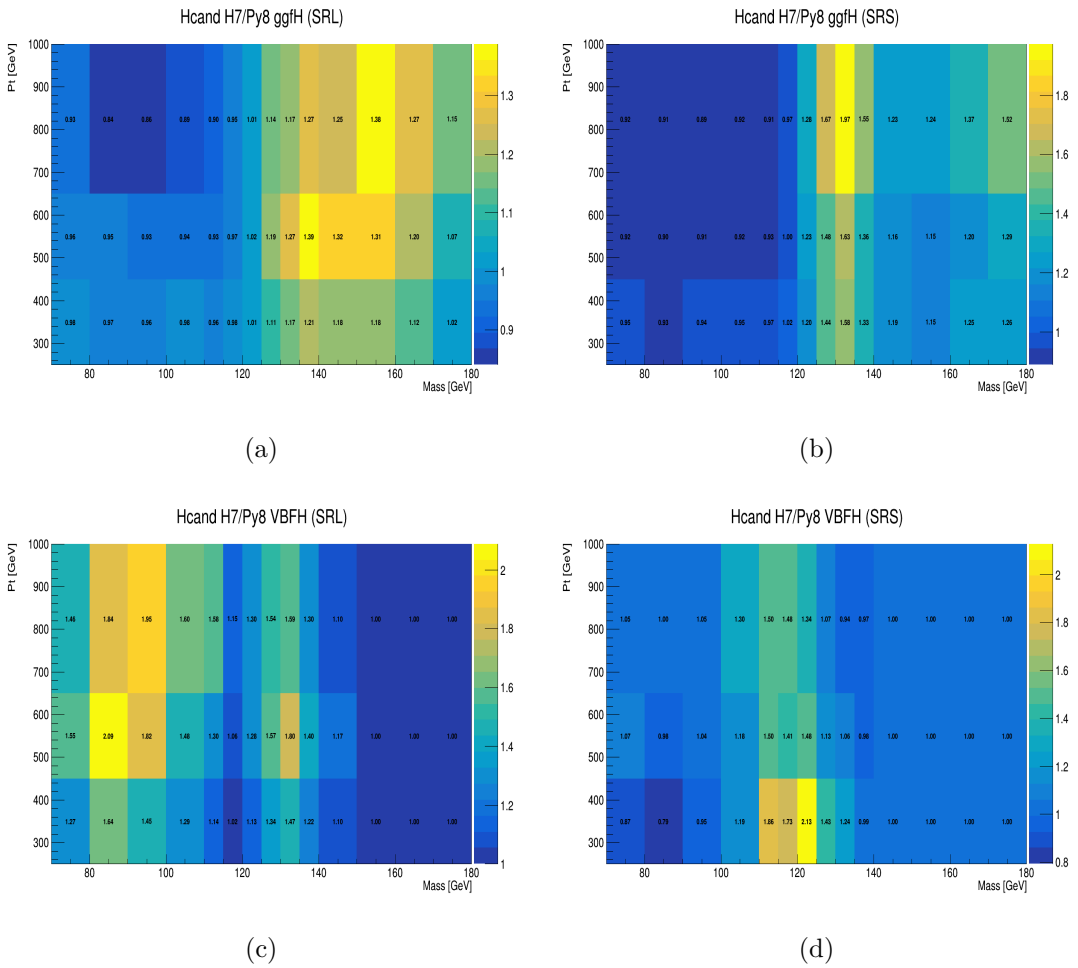
After the selection cuts and truth-matching a Higgs boson to the large-R jets, the samples are compared at truth level. The assumption is that the differences between samples at truth level, translate directly to the differences at reconstructed level. Taking the ratio of the 2D histograms of m_J and p_{T}, a reweighting map w_r is defined.

$$\frac{\mathrm{H7}_{\mathrm{truth}}}{\mathrm{Py8}_{\mathrm{truth}}} = w_r \tag{A.1}$$

After applying the map to the nominal sample, the reweighted PYTHIA mass distributions should model a Herwig generated sample at reconstructed level. These reweighted samples are then used to make comparisons with the unweighted versions.

$$\mathrm{Py8}_{\mathrm{rw}} = w_r \, \mathrm{Py8}_{\mathrm{reco}} \approx \mathrm{H7}_{\mathrm{reco}} \tag{A.2}$$

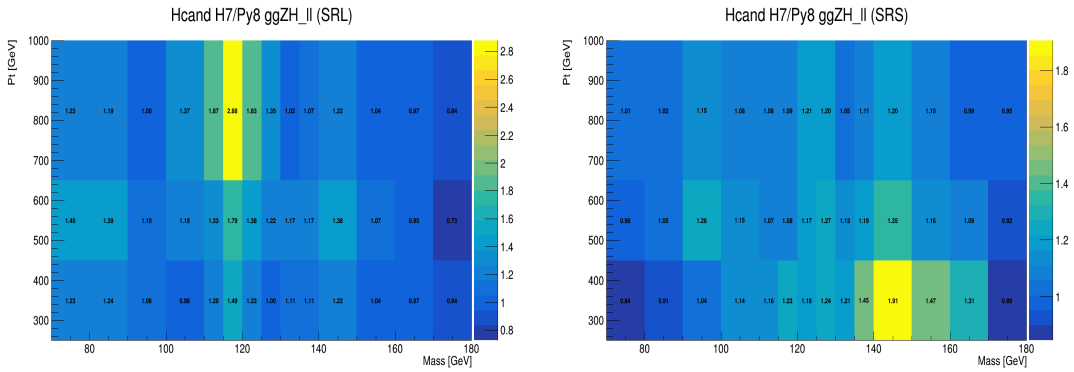
The reweighting maps and the comparisons of the samples at reconstructed level are presented below. A constant fit of the ratio of $\text{Py8}_{\text{rw}}/\text{Py8}_{\text{reco}}$ is shown on the legend of the mass distributions. The average percentage difference is around 5% when considering the entire mass spectrum. Most distributions have approximately 1% difference, but the average is brought up by the samples with low statistics. After inspecting the differences close to the mass peak, where statistics is not a limiting factor, it was decided that the parton shower model differences are negligible and therefore not considered in the likelihood fit as a nuisance parameter.



(a) (b)

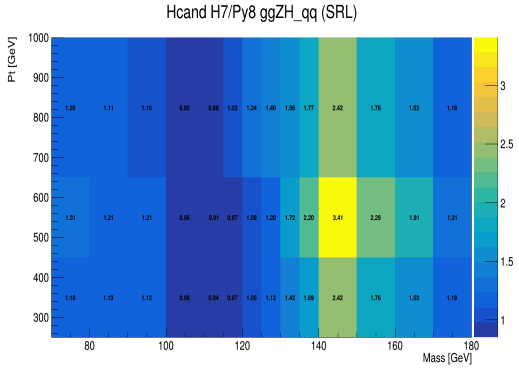
(c) (d)

Figure A.3: Reweighting maps (Herwig/Pythia) used to estimate the Higgs jet shower systematic for leading (left) and subleading (right) large-R jets from (a,b) ggF and (c,d) VBF.

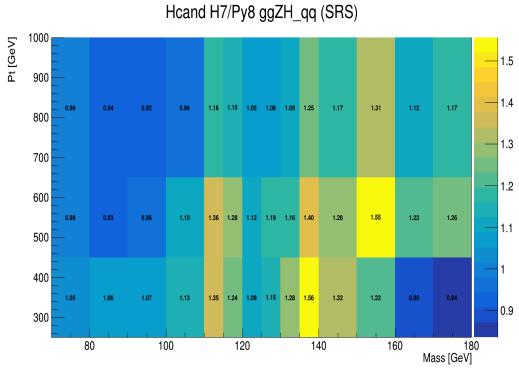


(a)

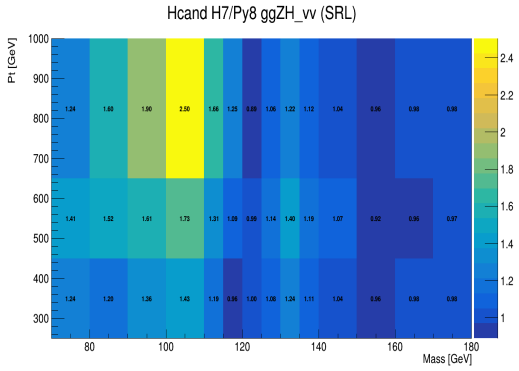
(b)



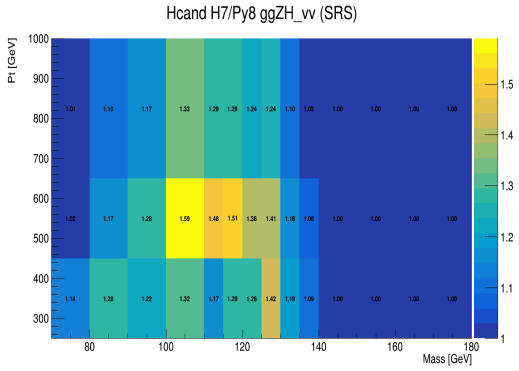
(c)



(d)



(e)



(f)

Figure A.4: Reweighting maps (Herwig/Pythia) used to estimate the Higgs jet shower systematic for leading (left) and subleading (right) large-R jets from $gg \to ZH$ production with subchannels (a,b) $Z(l^+l^-)\,H$, (c,d) $Z(qq)\,H$ and (e,f) $Z(\nu\nu)\,H$.

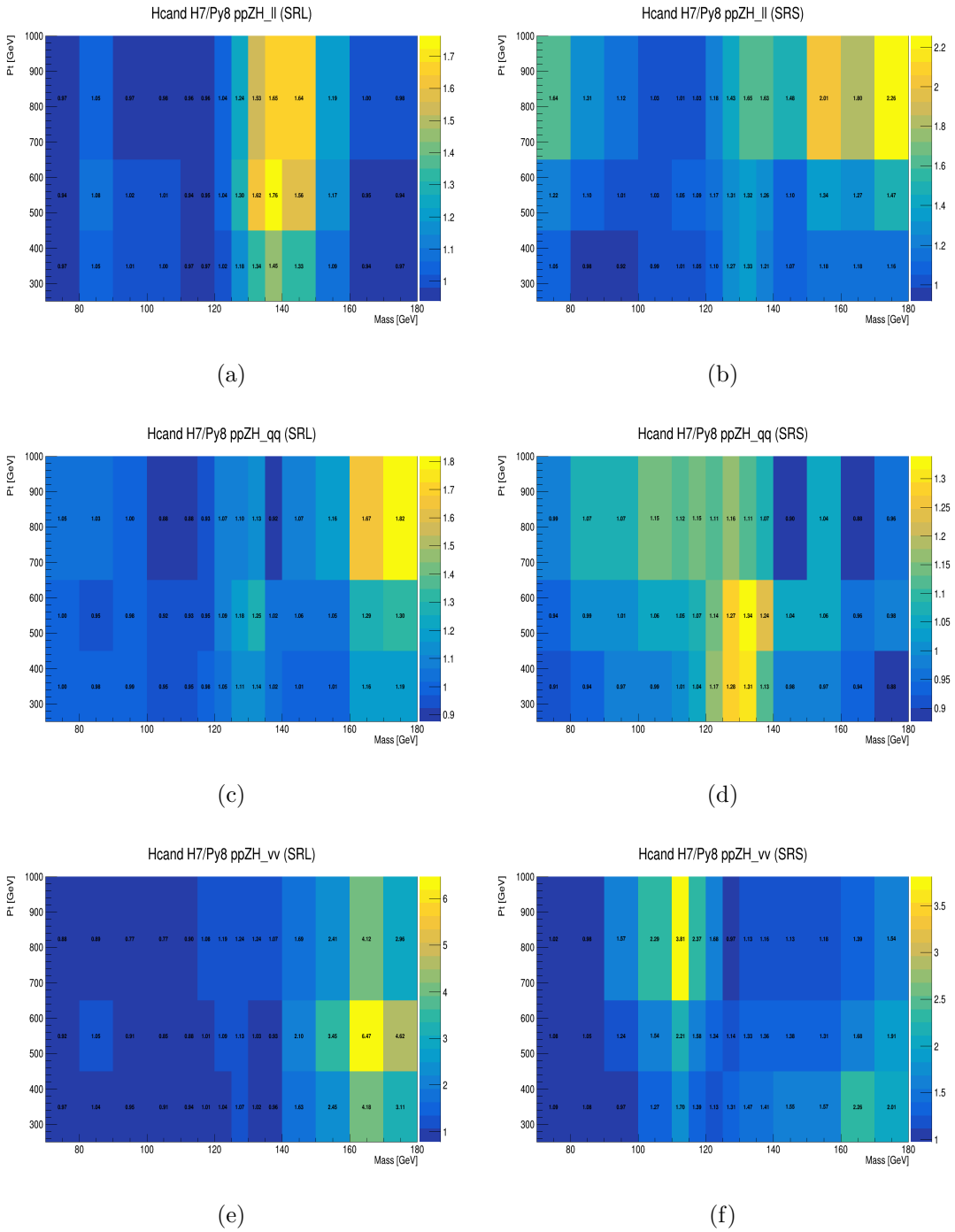
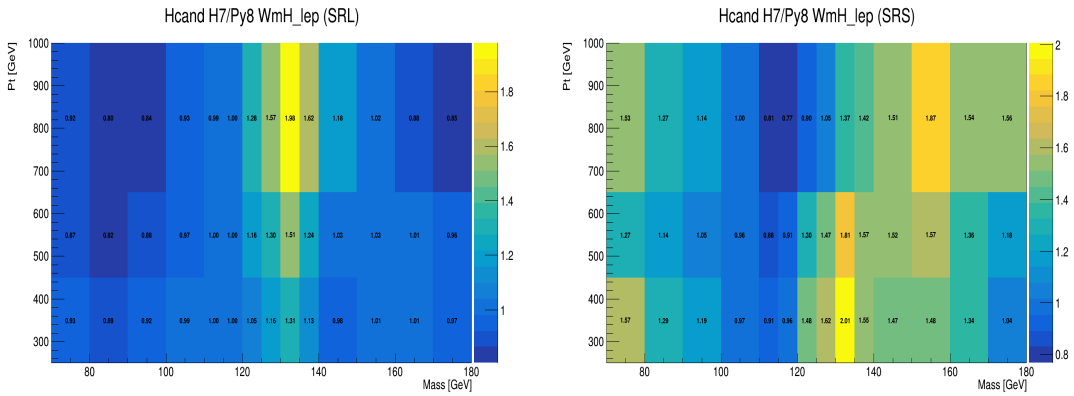
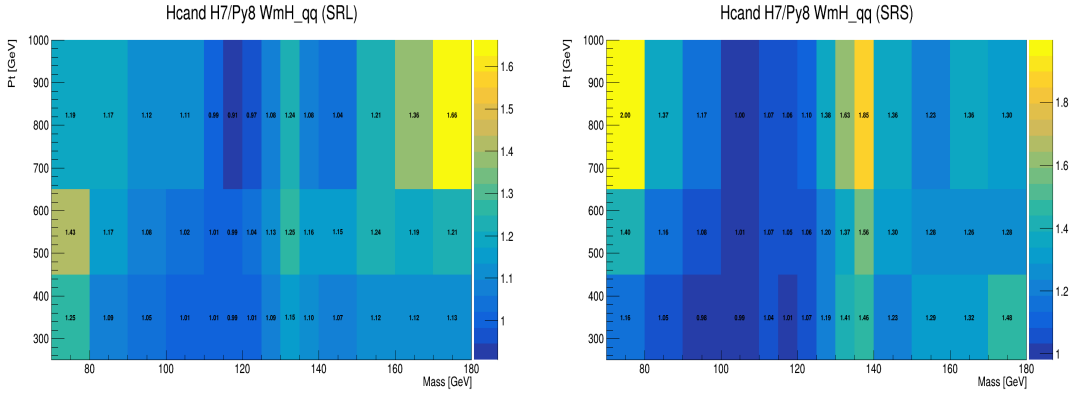


Figure A.5: Reweighting maps (Herwig/Pythia) used to estimate the Higgs jet shower systematic for leading (left) and subleading (right) large-R jets from $pp \to ZH$ production with subchannels (a,b) $Z(l^+l^-)\,H$, (c,d) $Z(qq)\,H$ and (e,f) $Z(\nu\nu)\,H$.

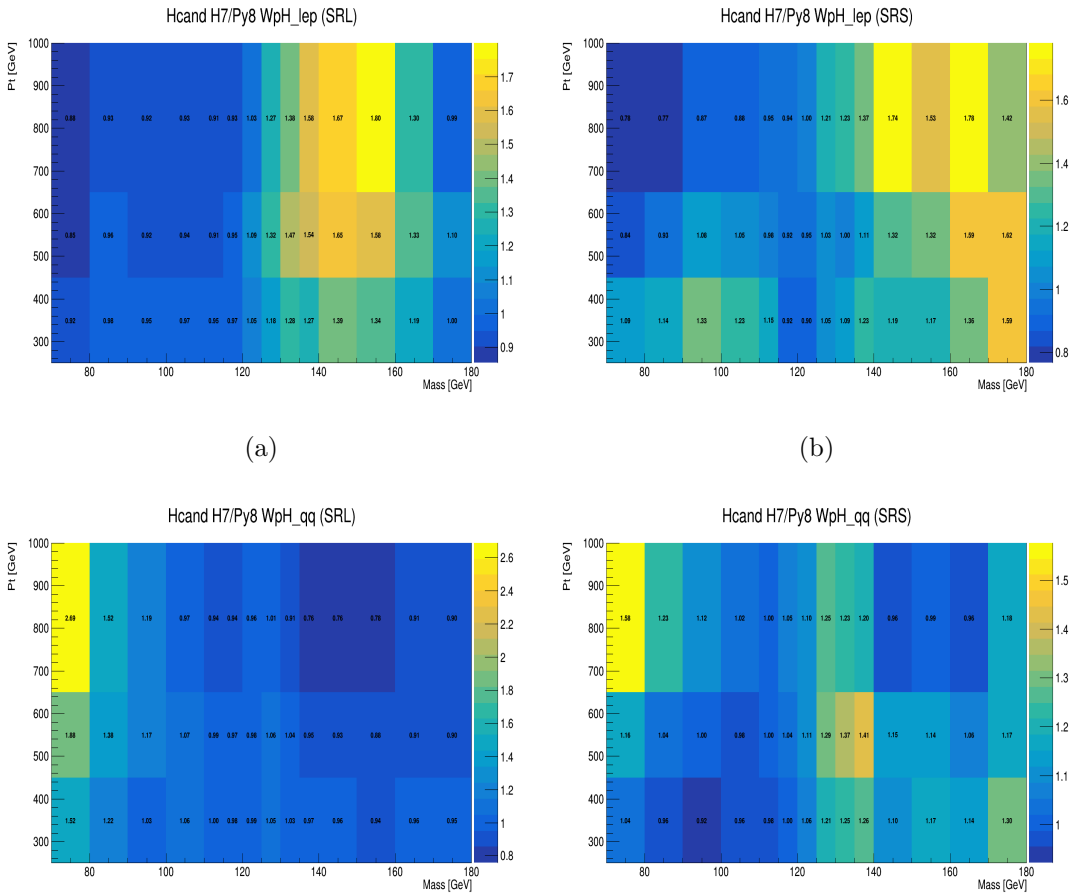


(a) (b)



(c) (d)

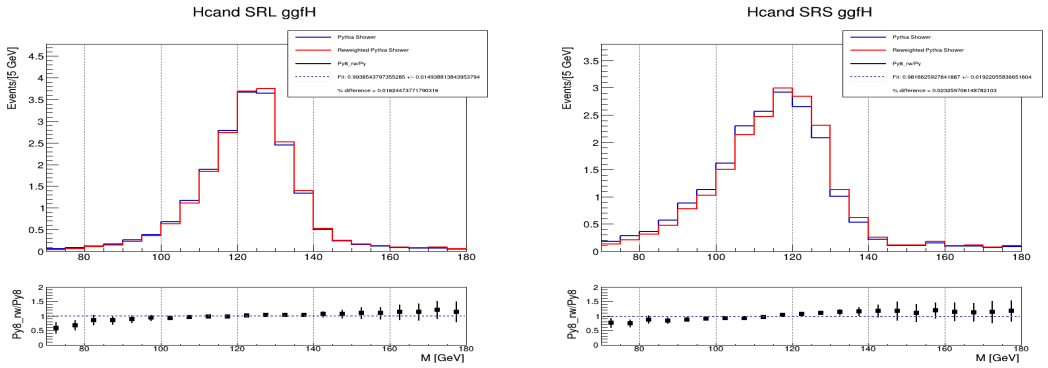
Figure A.6: Reweighting maps (Herwig/Pythia) used to estimate the Higgs jet shower systematic for leading (left) and subleading (right) large-R jets from $gg \to W^{-}H$ production with subchannels (a,b) $W^{-}(l^{-}\nu_{l})\,H$ and (c,d) $W^{-}(qq')\,H$.



(a) (b)

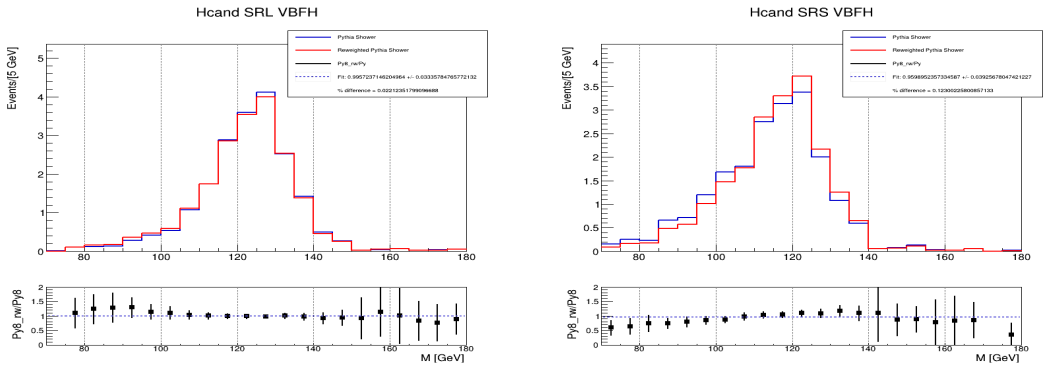
(c) (d)

Figure A.7: Reweighting maps (Herwig/Pythia) used to estimate the Higgs jet shower systematic for leading (left) and subleading (right) large-R jets from $gg \to W^+H$ production with subchannels (a,b) $W^+(l^+\nu_l)\,H$ and (c,d) $W^+(qq')\,H$.



(a)

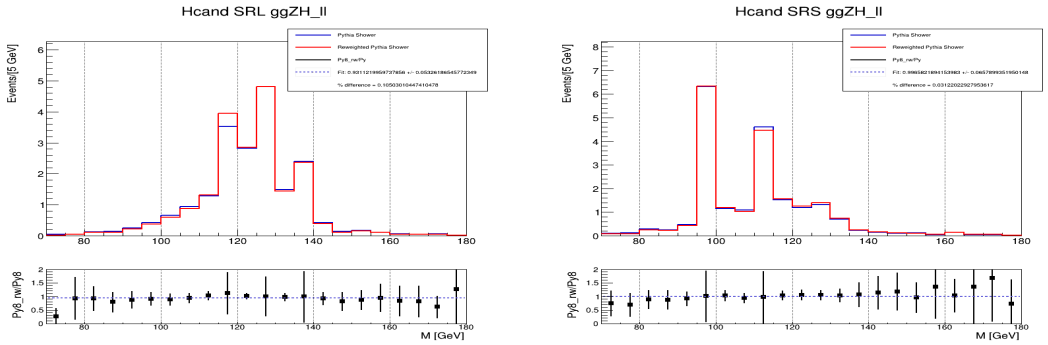
(b)



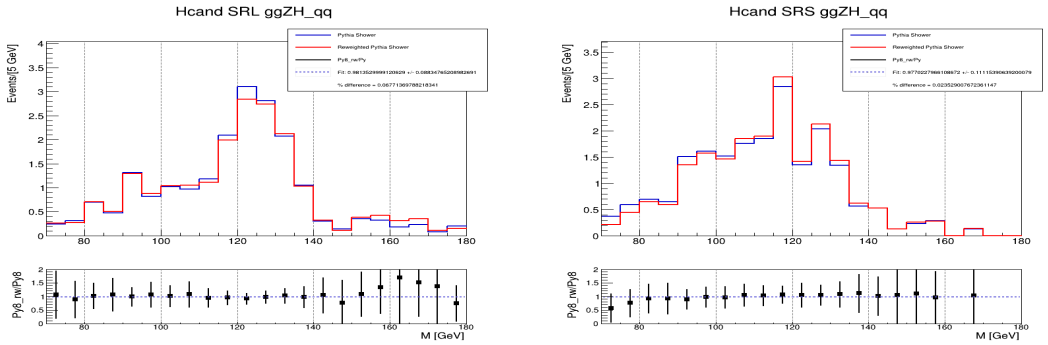
(c)

(d)

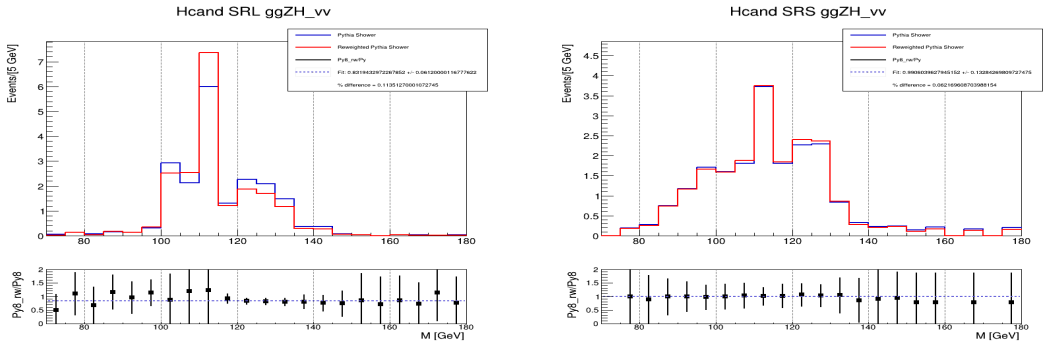
Figure A.8: Comparison of the Higgs jet mass distribution after applying the reweighting maps for leading (left) and subleading (right) large-R jets from (a,b) ggF and (c,d) VBF.



(a) (b)

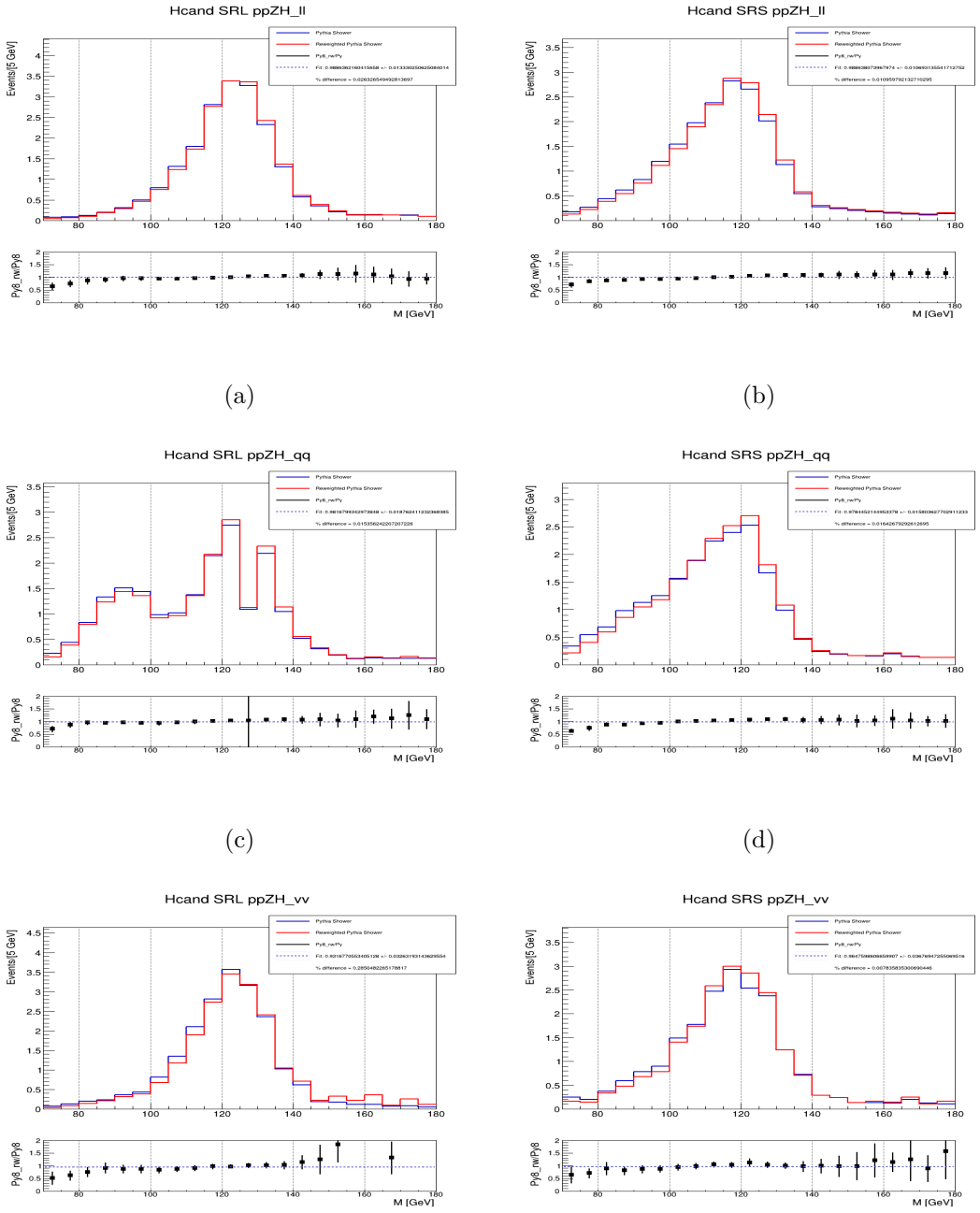


(c) (d)



(e) (f)

Figure A.9: Comparison of the Higgs jet mass distribution after applying the reweighting maps for leading (left) and subleading (right) large-R jets from $gg \to ZH$ production with subchannels (a,b) $Z(l^+l^-)\,H$, (c,d) $Z(qq)\,H$ and (e,f) $Z(\nu\nu)\,H$.

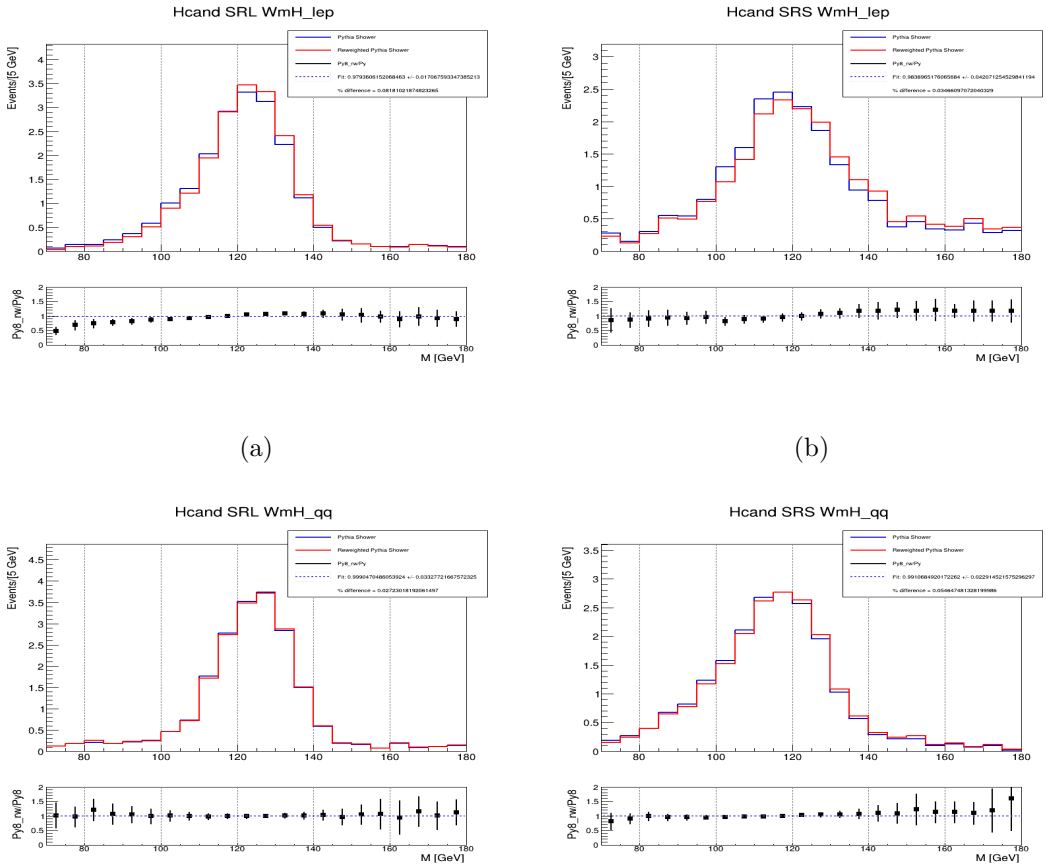


(a) (b)

(c) (d)

(e) (f)

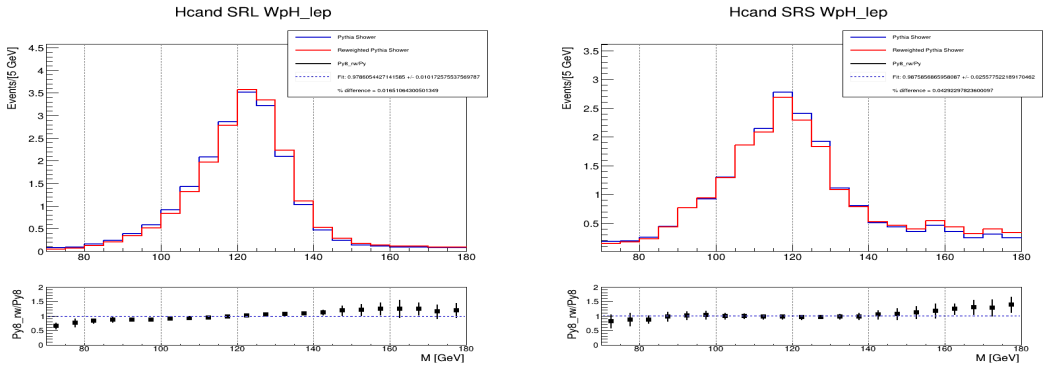
Figure A.10: Comparison of the Higgs jet mass distribution after applying the reweighting maps for leading (left) and subleading (right) large-R jets from $pp \to ZH$ production with subchannels (a,b) $Z(l^+l^-)\,H$, (c,d) $Z(qq)\,H$ and (e,f) $Z(\nu\nu)\,H$.



(a) (b)

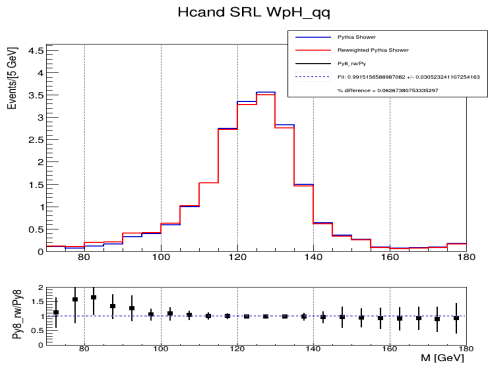
(c) (d)

Figure A.11: Comparison of the Higgs jet mass distribution after applying the reweighting maps for leading (left) and subleading (right) large-R jets from $gg \to W^-H$ production with subchannels (a,b) $W^-(l^-\nu_l)\,H$ and (c,d) $W^-(qq')\,H$.

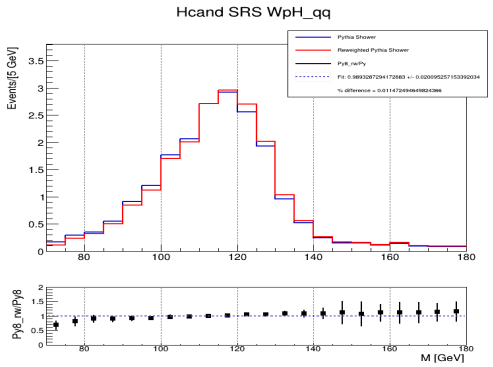


(a)

(b)



(c)



(d)

Figure A.12: Comparison of the Higgs jet mass distribution after applying the reweighting maps for leading (left) and subleading (right) large-R jets from $gg \to W^+H$ production with subchannels (a,b) $W^+(l^+\nu_l)\,H$ and (c,d) $W^+(qq')\,H$.

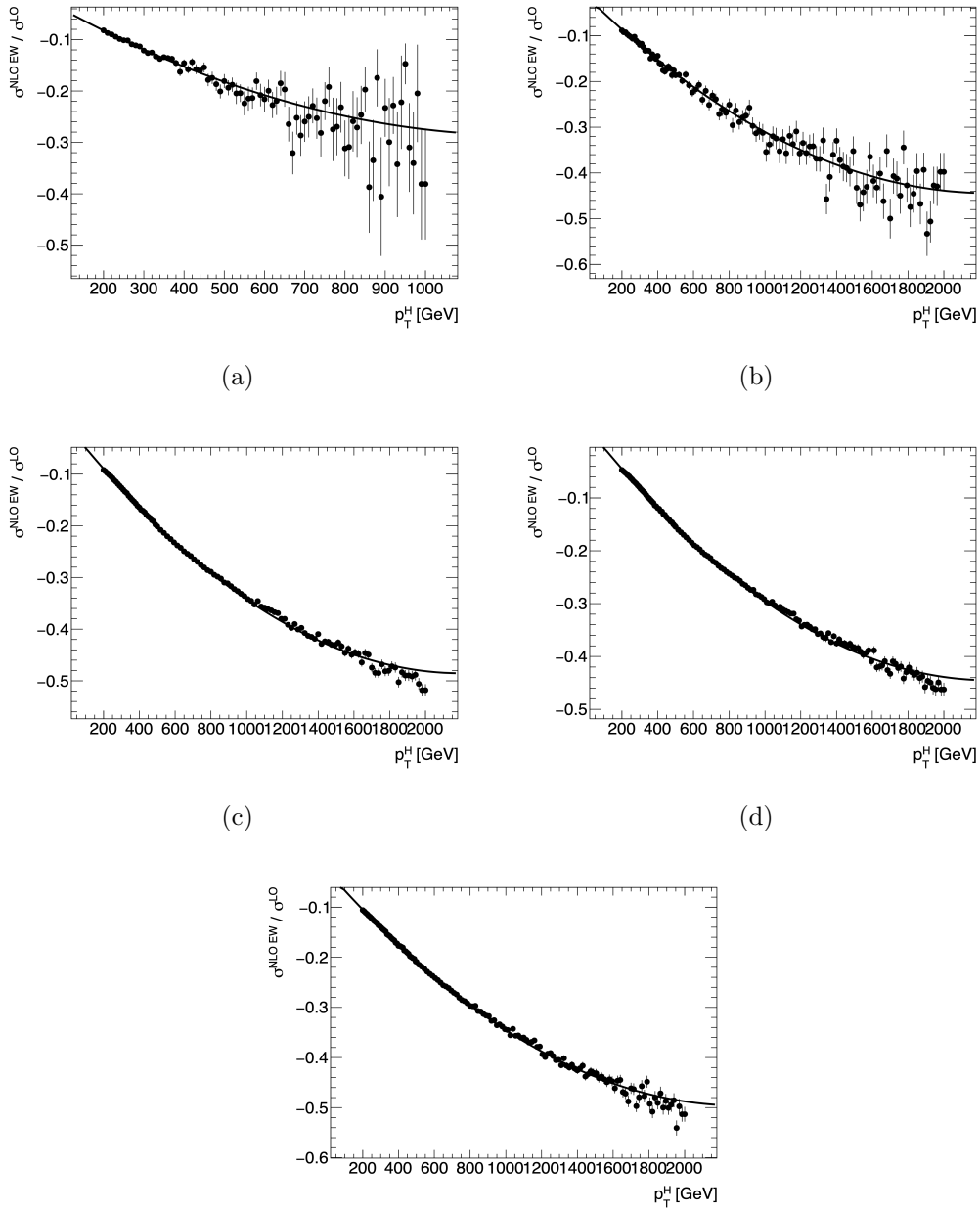


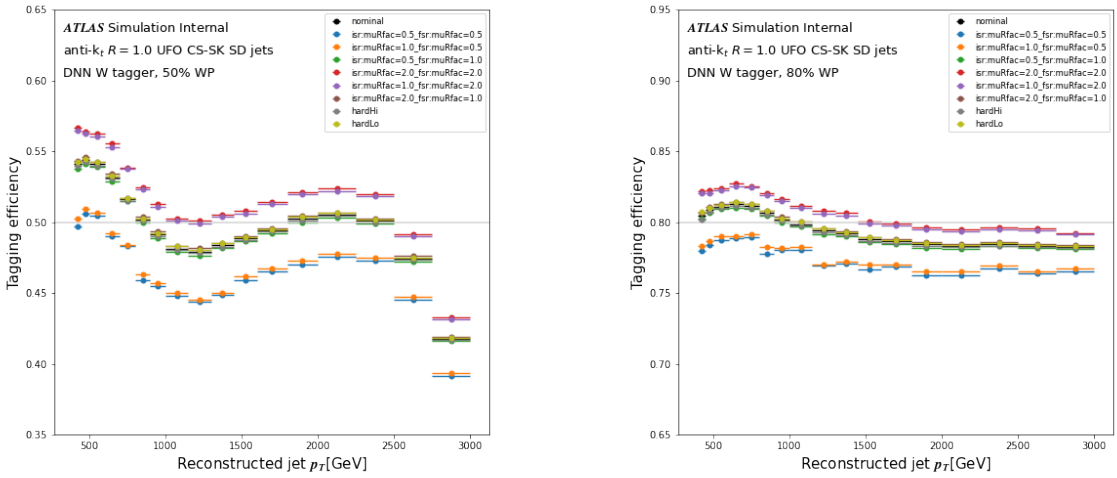
Figure A.13: Systematics of the NLO EW corrections calculated using HAWK as a function of the Higgs p_{T} for (a) VBF and (b) VH production, and the subchannels (c) $Z(l^+l^-)\,H$, (d) $Z(\nu\nu)\,H$ and (e) WH. The y-axis is the value of $\delta_{\mathrm{EW}} = \sigma^{\mathrm{NLO\,EW}}/\sigma^{\mathrm{LO}} - 1$.

APPENDIX B

Unified Flow Objects

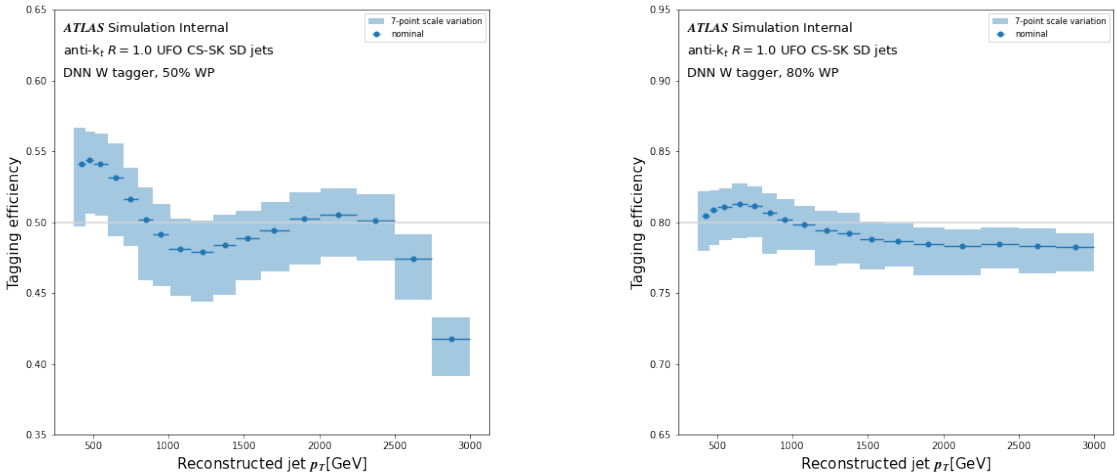
High p_{T} Extrapolation Efficiencies

Shower Variations



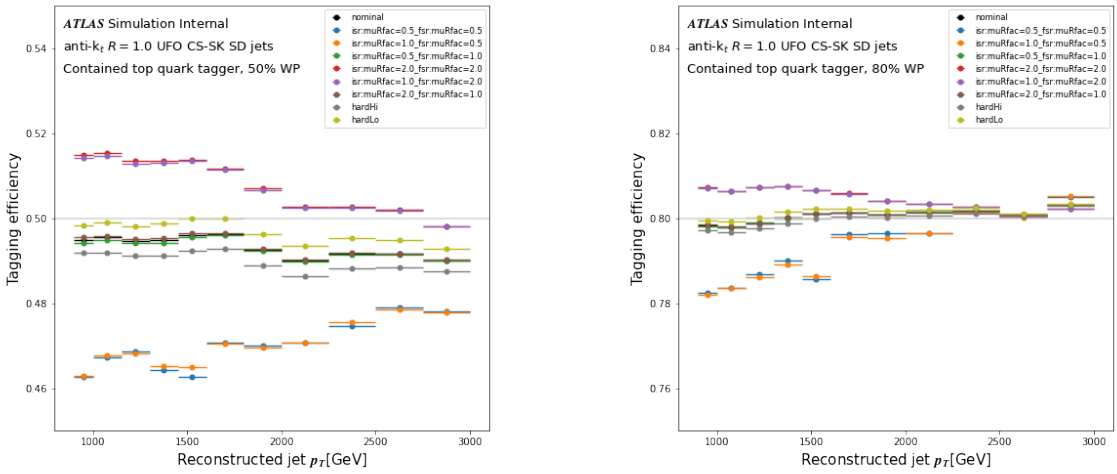
(a) (b)

Figure B.1: DNN W tagger 7(+2) point scale variation efficiencies as a function of p_{T} for W-jets at the (a) 50% and (b) 80% working points.



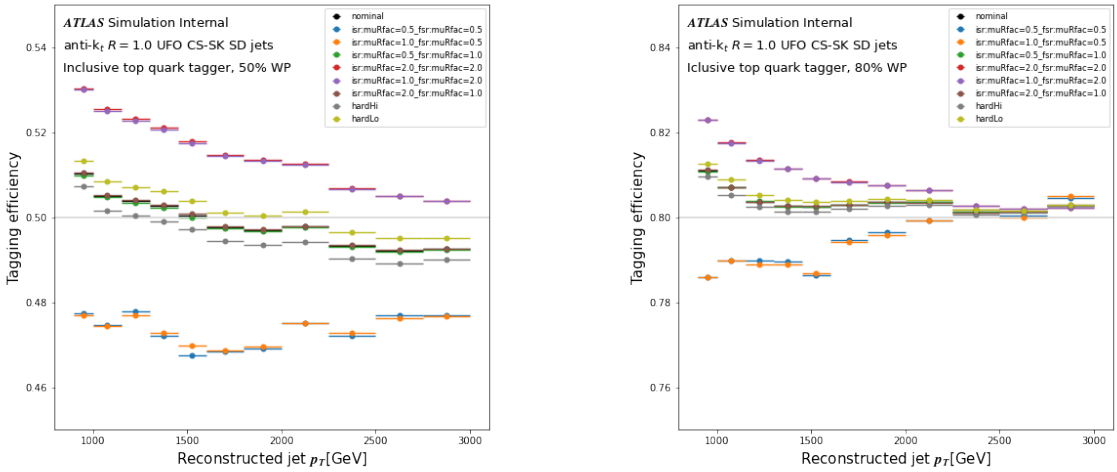
(a) (b)

Figure B.2: DNN W tagger 7(+2) point scale variation efficiency envelope as a function of p_{T} for W-jets at the (a) 50% and (b) 80% working points.



(a)

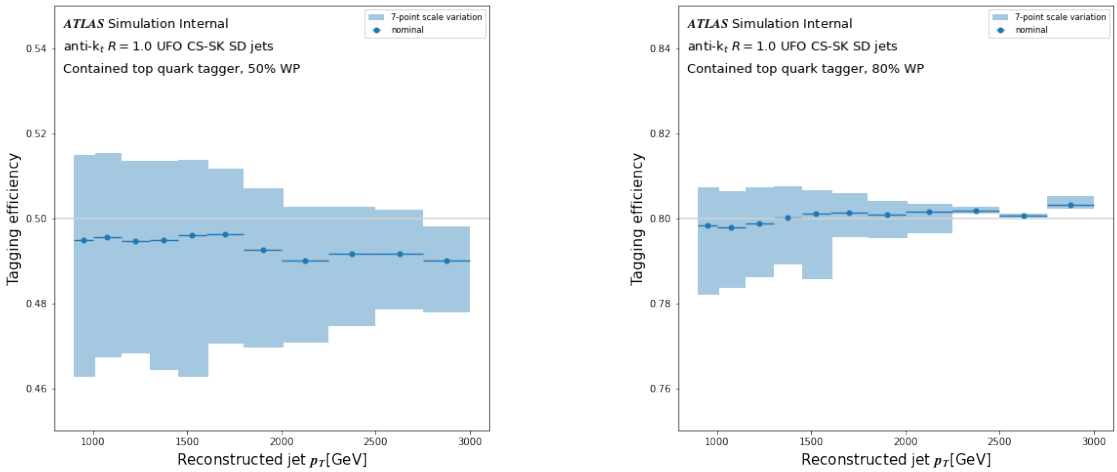
(b)



(c)

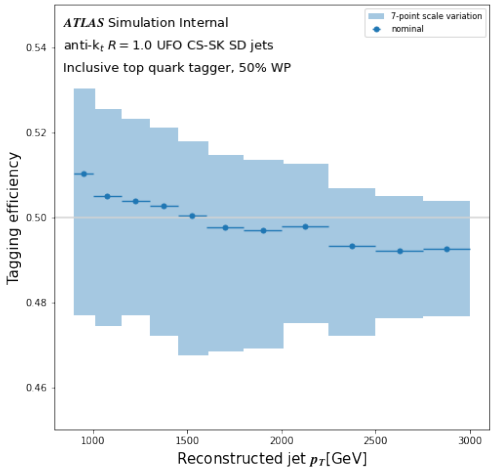
(d)

Figure B.3: DNN top taggers 7(+2) point scale variation efficiencies as a function of p_{T} for contained tops (top) at the (a) 50% and (b) 80% working points and for inclusive tops (bottom) at the (c) 50% and (d) 80% working points.

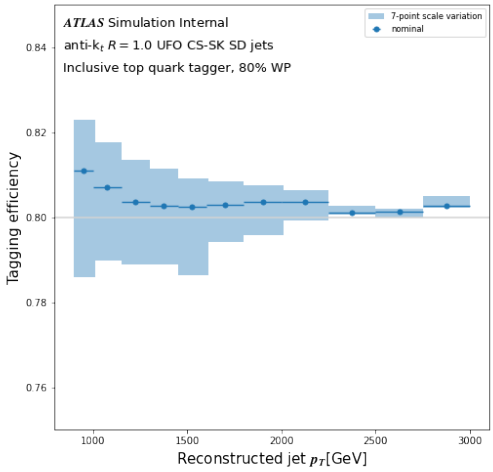


(a)

(b)

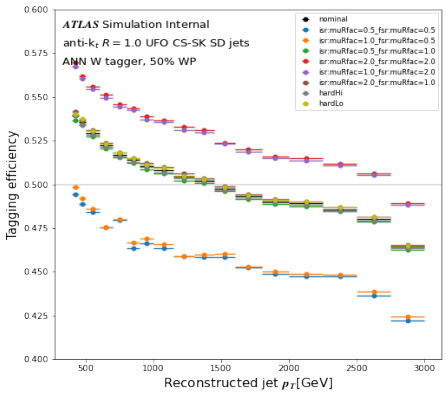


(c)

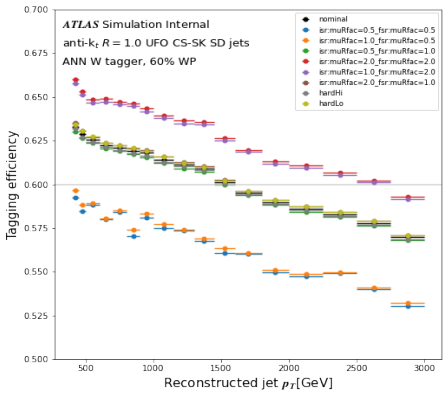


(d)

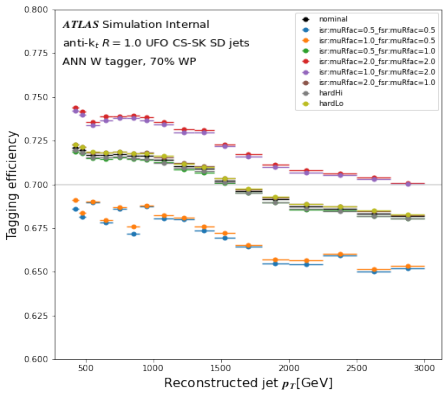
Figure B.4: DNN top taggers 7(+2) point scale variation efficiency envelope as a function of p_{T} for contained tops (top) at the (a) 50% and (b) 80% working points and for inclusive tops (bottom) at the (c) 50% and (d) 80% working points.



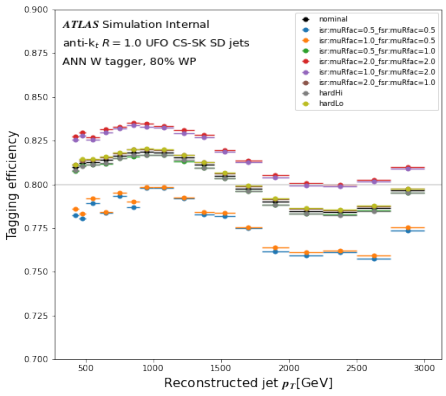
(a)



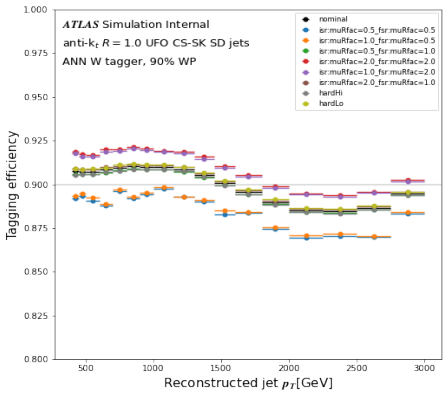
(b)



(c)



(d)



(e)

Figure B.5: ANN W/Z tagger 7(+2) point scale variation efficiencies as a function of p_{T} for W-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

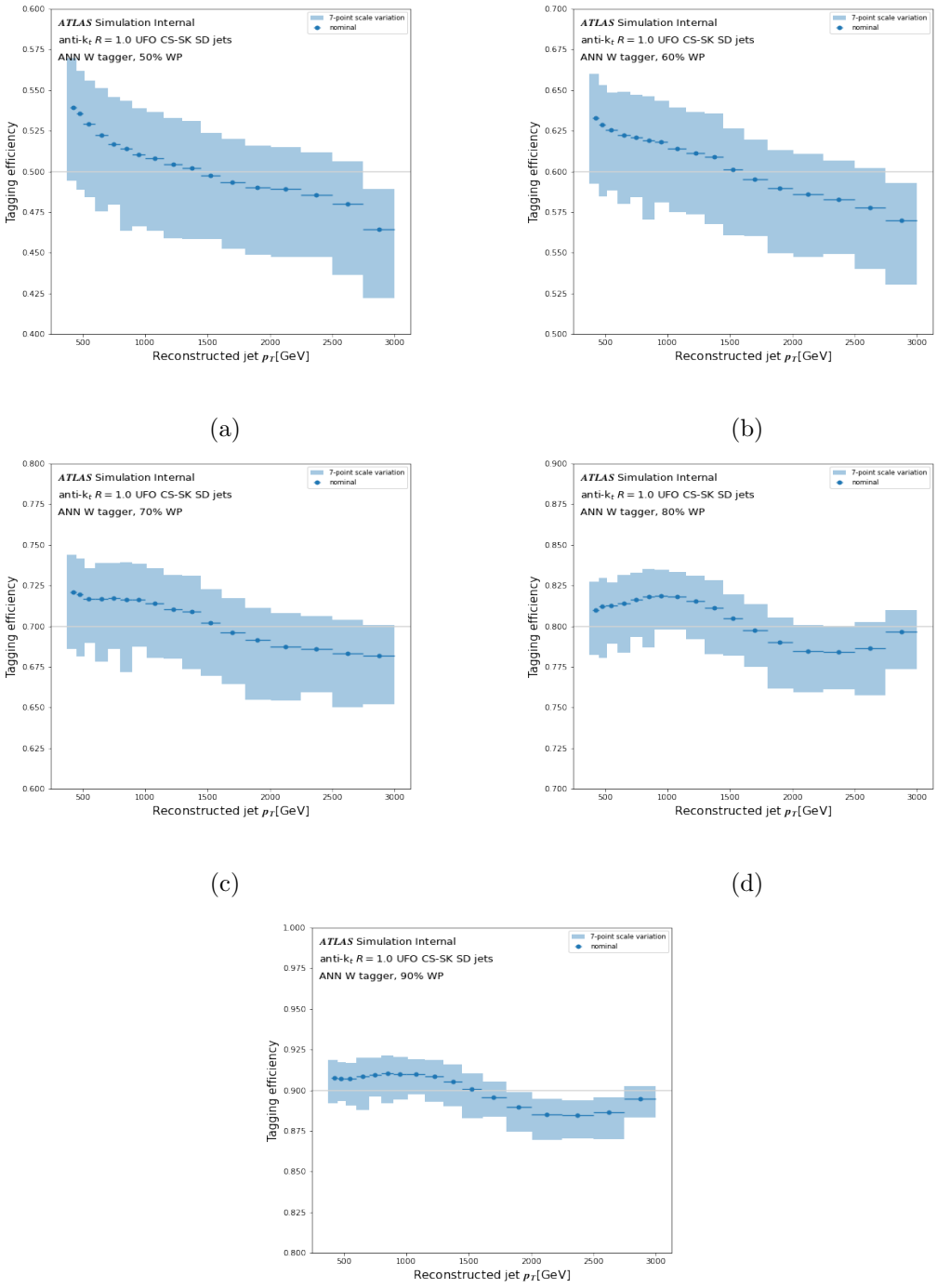
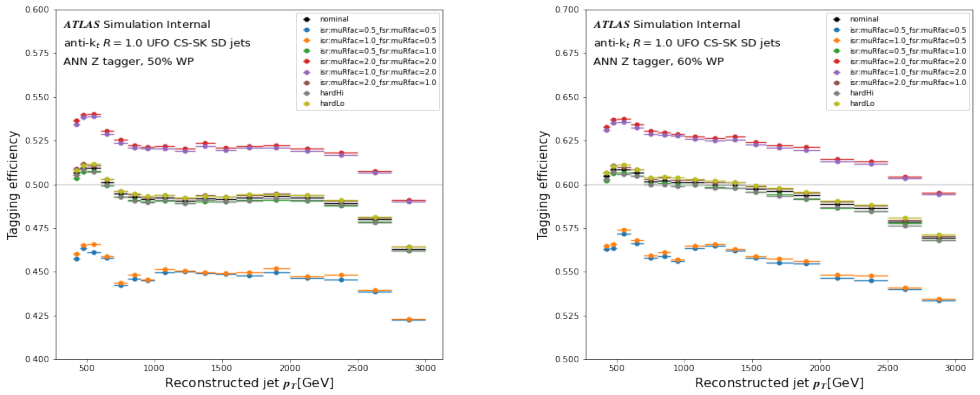
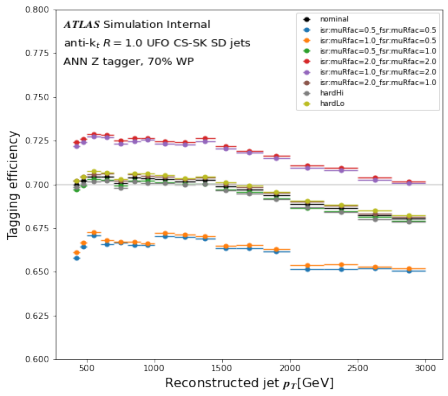


Figure B.6: ANN W/Z tagger 7(+2) point scale variation efficiency envelope as a function of p_{T} for W-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

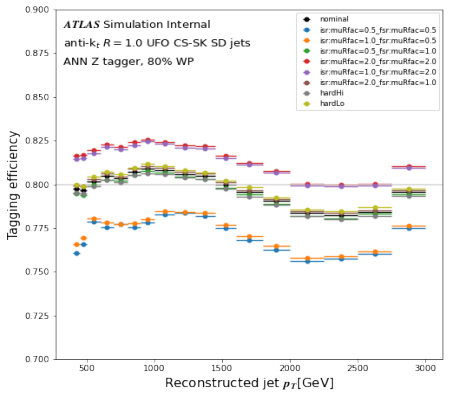


(a)

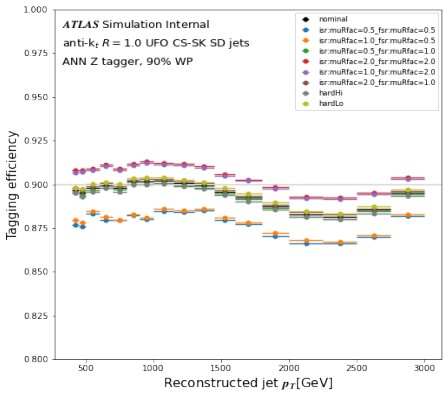
(b)



(c)

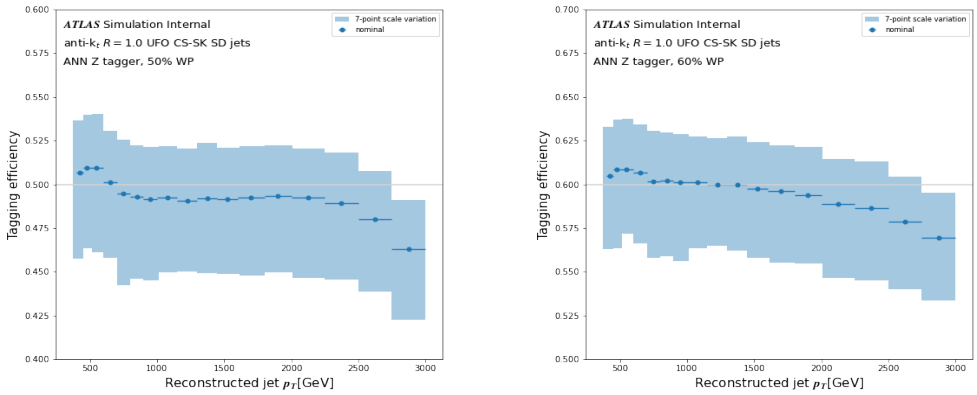


(d)



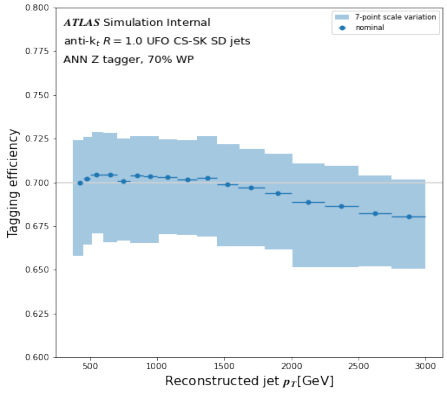
(e)

Figure B.7: ANN W/Z tagger 7(+2) point scale variation efficiencies as a function of p_{T} for Z-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

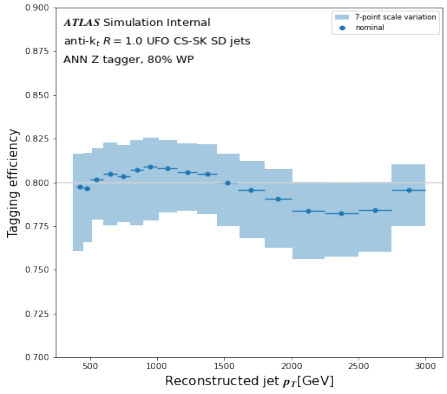


(a)

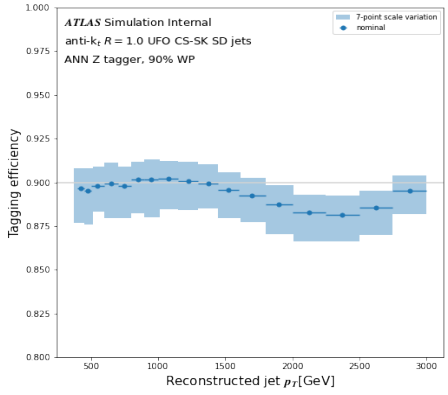
(b)



(c)

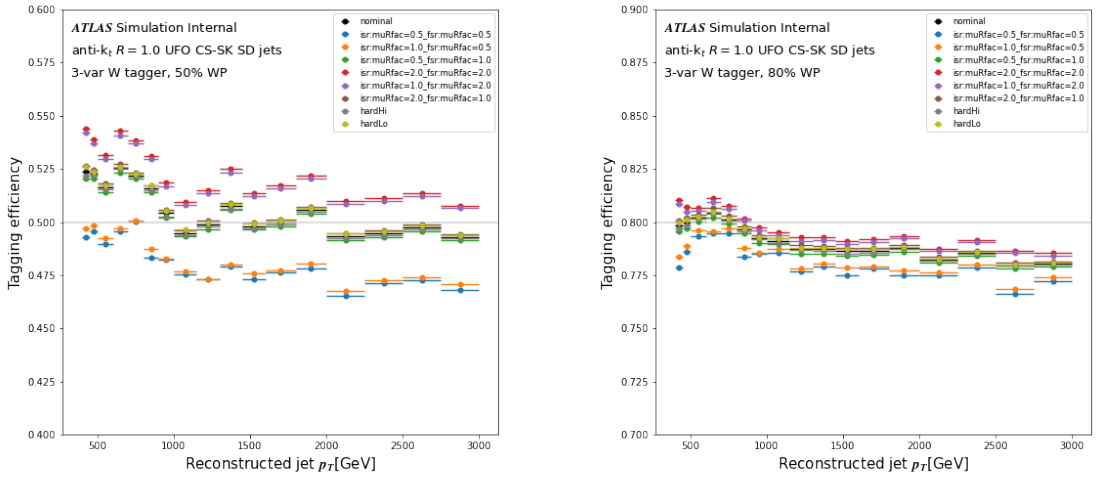


(d)



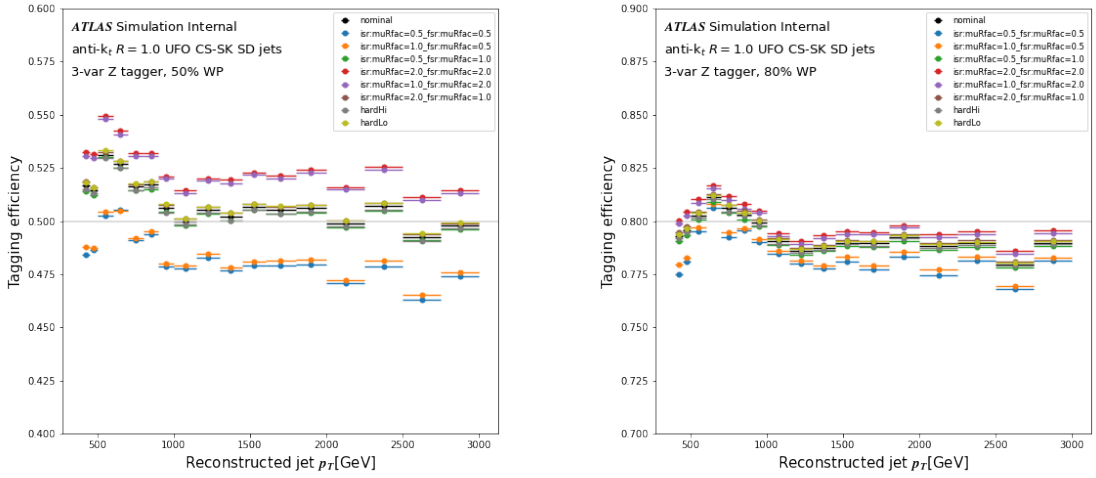
(e)

Figure B.8: ANN W/Z tagger 7(+2) point scale variation efficiency envelope as a function of p_{T} for Z-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.



(a)

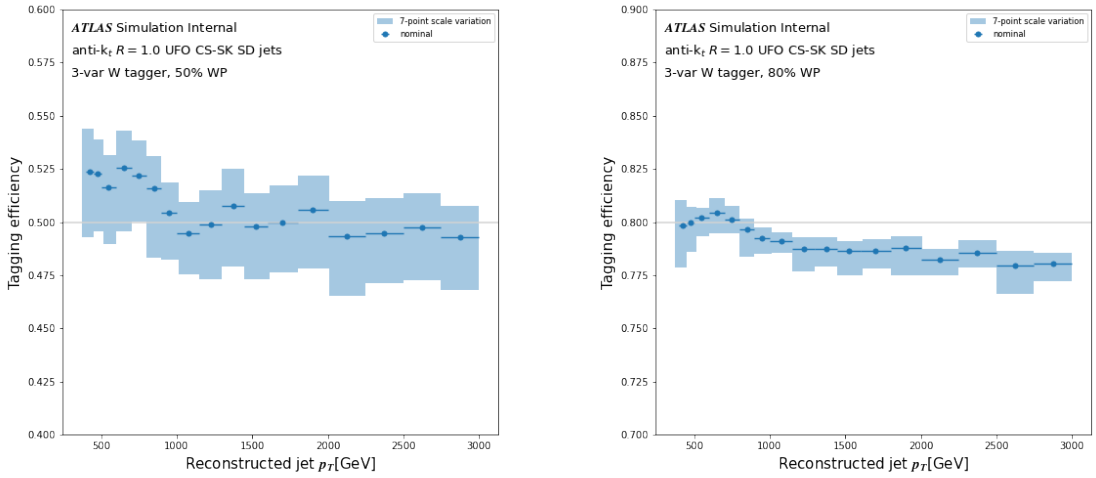
(b)



(c)

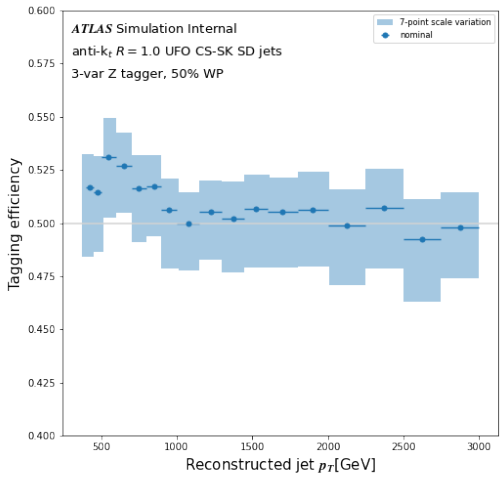
(d)

Figure B.9: Three variable W/Z taggers 7(+2) point scale variation efficiencies as a function of p_{T} for W-jets (top) at the (a) 50% and (b) 80% working points and for Z-jets (bottom) at the (c) 50% and (d) 80% working points.

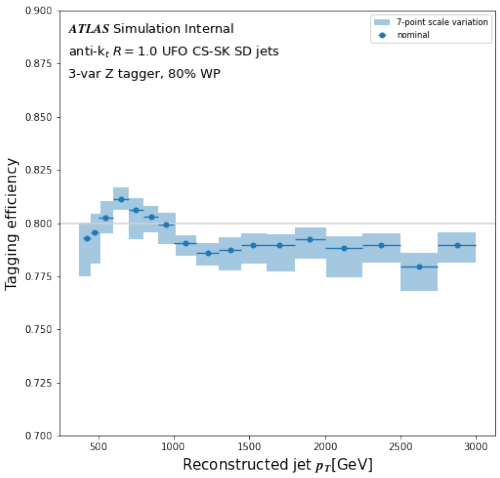


(a)

(b)



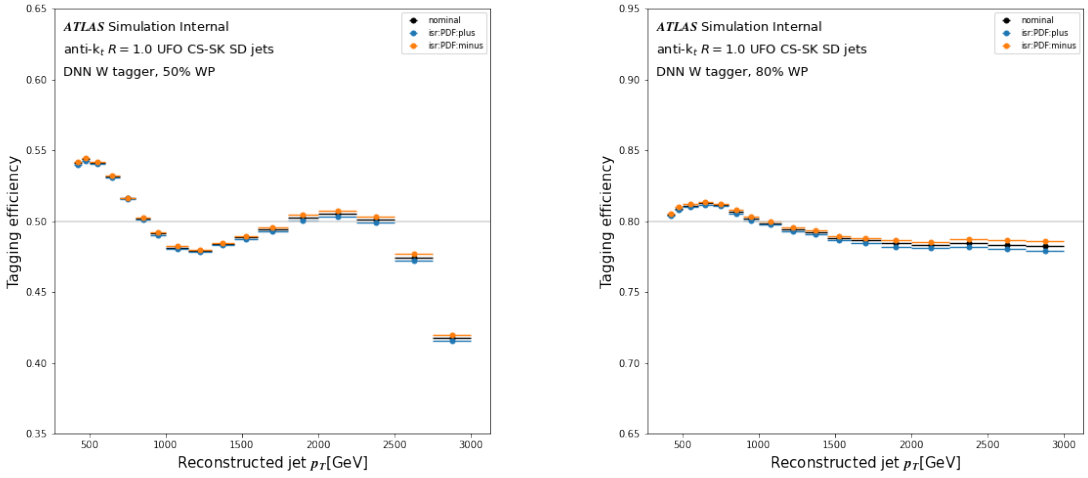
(c)



(d)

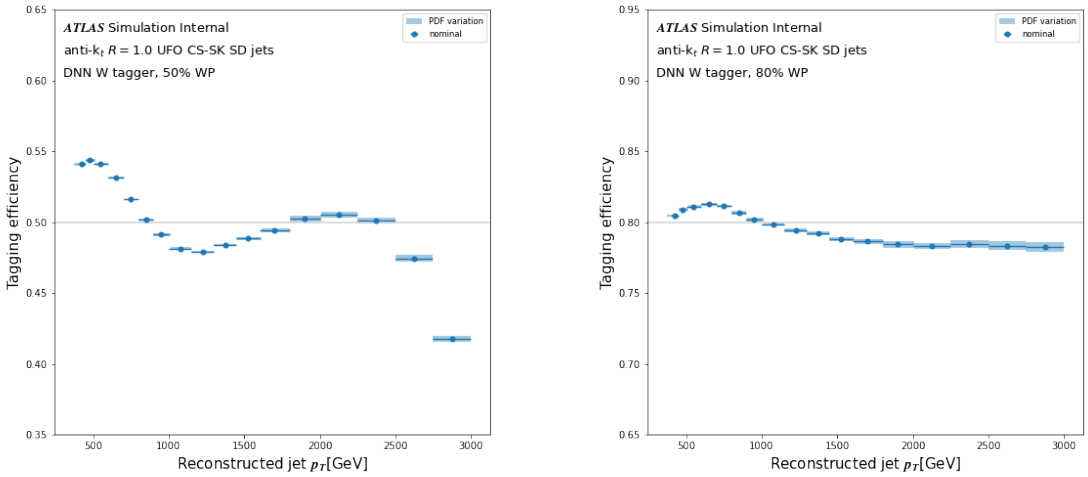
Figure B.10: Three variable W/Z taggers 7(+2) point scale variation efficiency envelope as a function of p_{T} for W-jets (top) at the (a) 50% and (b) 80% working points and for Z-jets (bottom) at the (c) 50% and (d) 80% working points.

PDF Variations



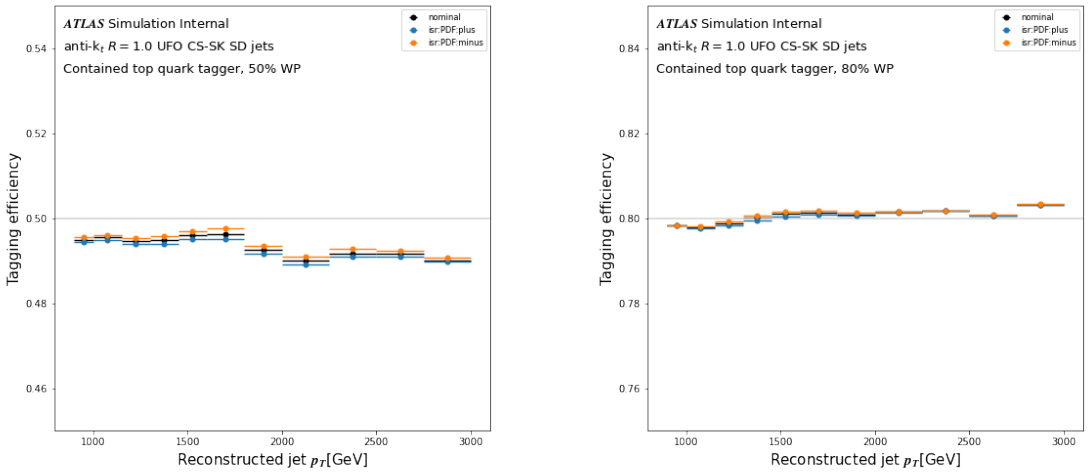
(a) (b)

Figure B.11: DNN W tagger PDF variation efficiencies as a function of p_{T} for W-jets at the (a) 50% and (b) 80% working points.



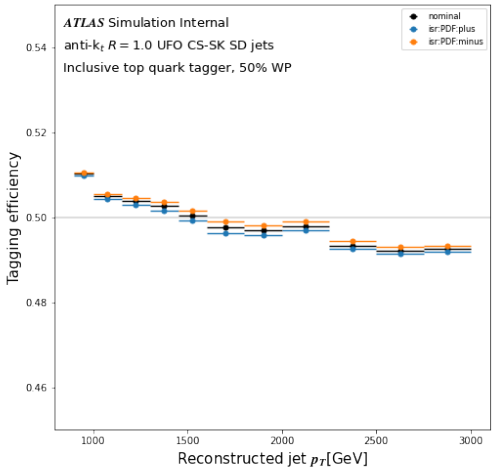
(a) (b)

Figure B.12: DNN W tagger PDF variation efficiency envelope as a function of p_{T} for W-jets at the (a) 50% and (b) 80% working points.

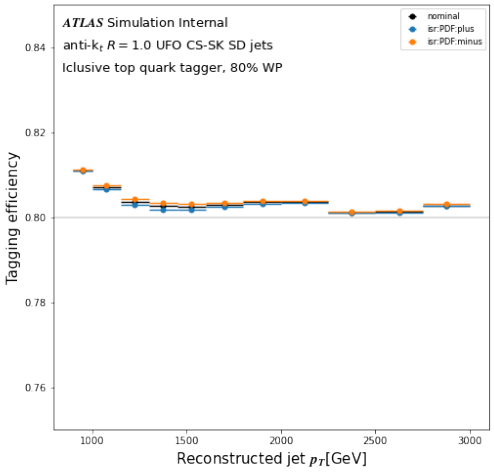


(a)

(b)

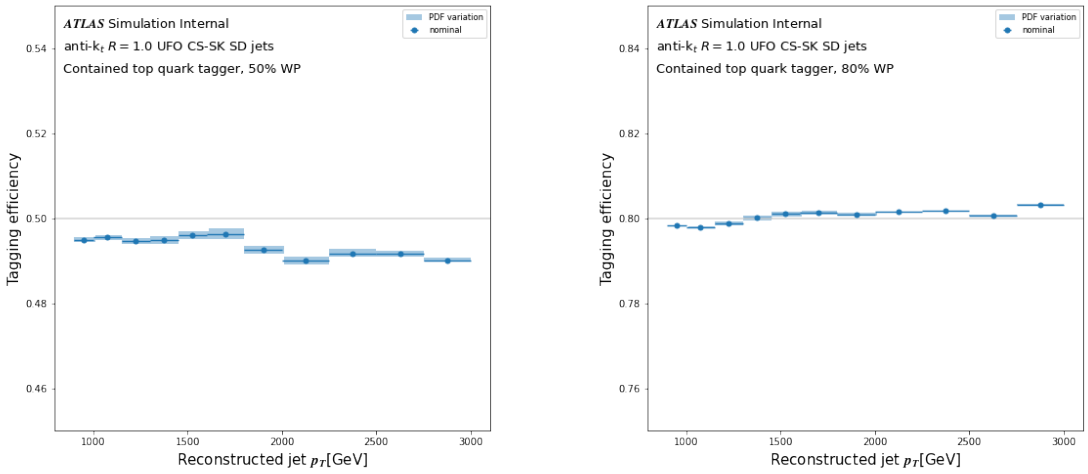


(c)



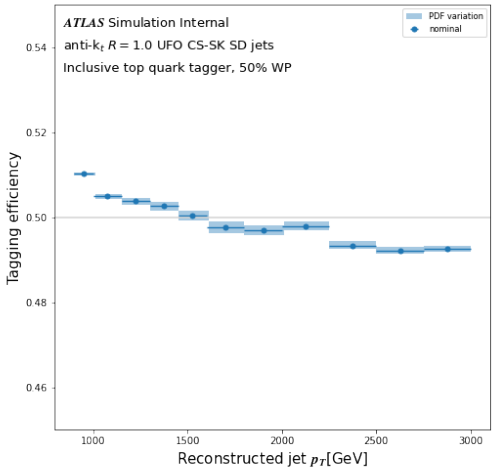
(d)

Figure B.13: DNN top taggers PDF variation efficiencies as a function of p_{T} for contained tops (top) at the (a) 50% and (b) 80% working points and for inclusive tops (bottom) at the (c) 50% and (d) 80% working points.

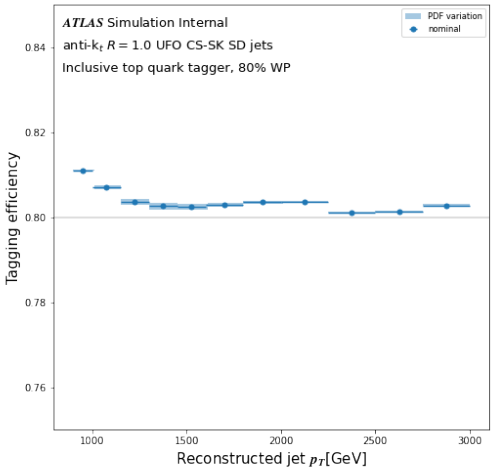


(a)

(b)

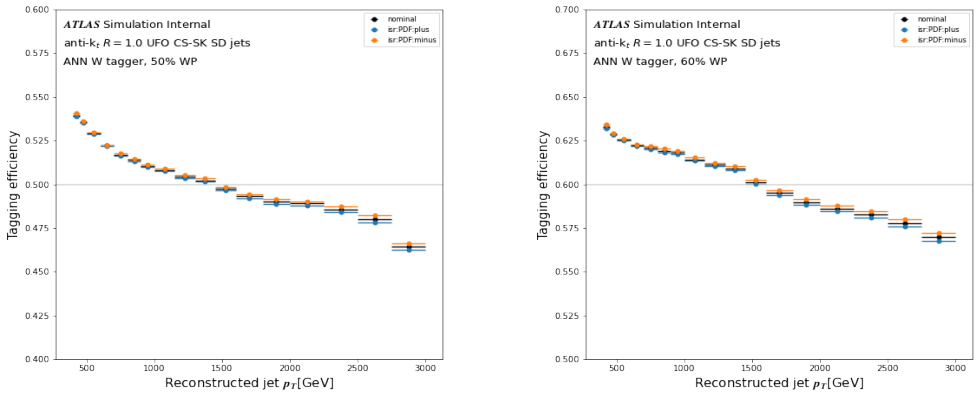


(c)



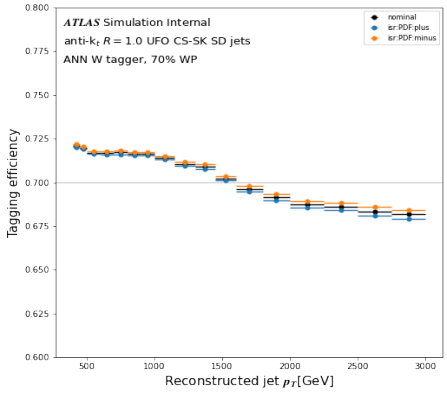
(d)

Figure B.14: DNN top taggers PDF variation efficiency envelope as a function of p_{T} for contained tops (top) at the (a) 50% and (b) 80% working points and for inclusive tops (bottom) at the (c) 50% and (d) 80% working points.

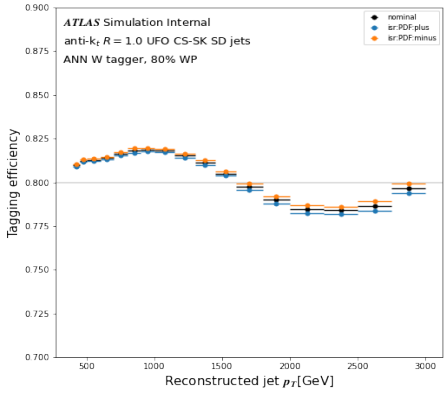


(a)

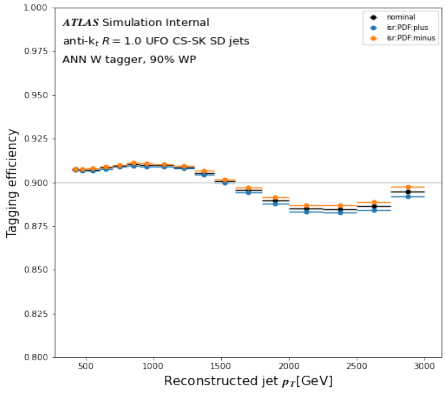
(b)



(c)

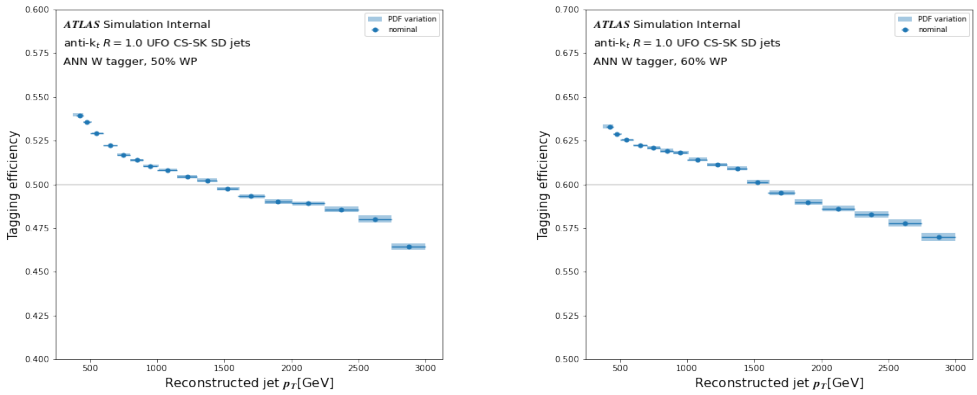


(d)



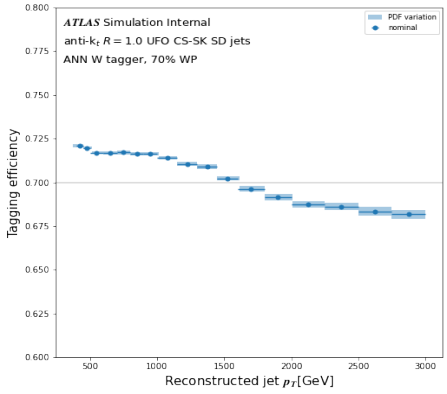
(e)

Figure B.15: ANN W/Z tagger PDF variation efficiencies as a function of p_{T} for W-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

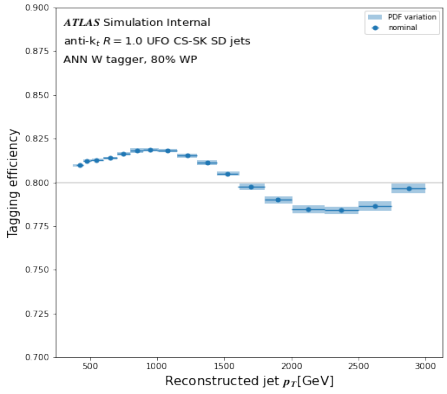


(a)

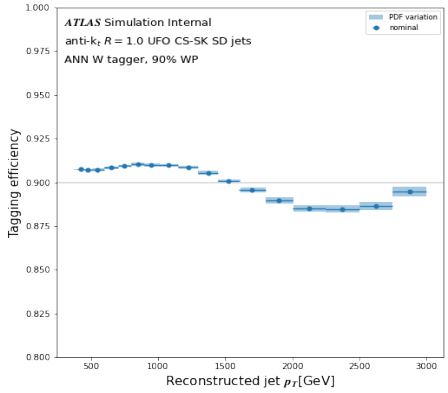
(b)



(c)

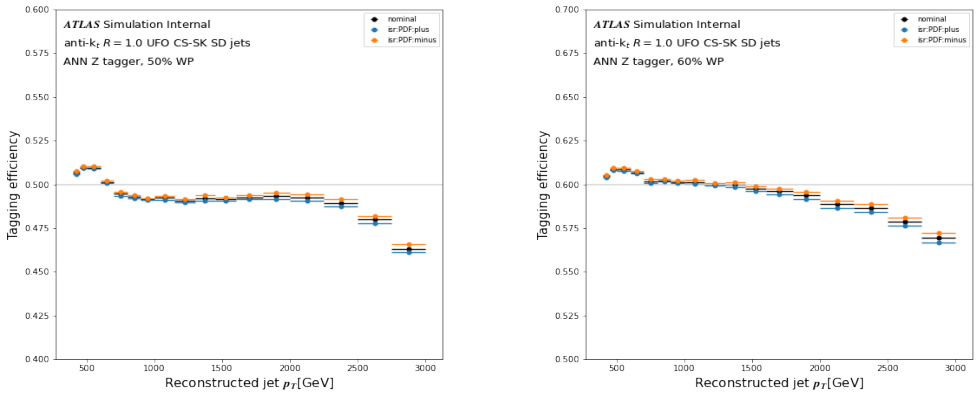


(d)



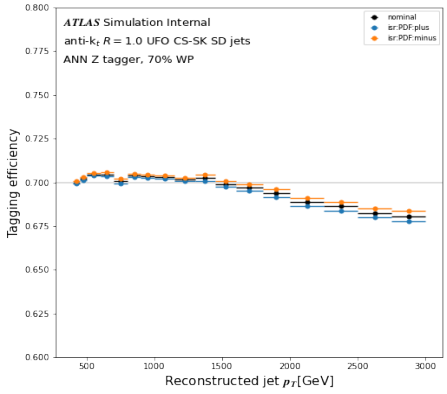
(e)

Figure B.16: ANN W/Z tagger PDF variation efficiency envelope as a function of p_{T} for W-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

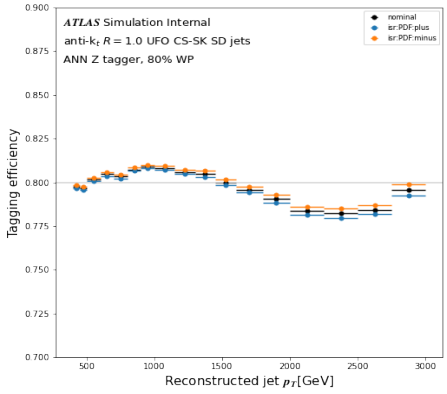


(a)

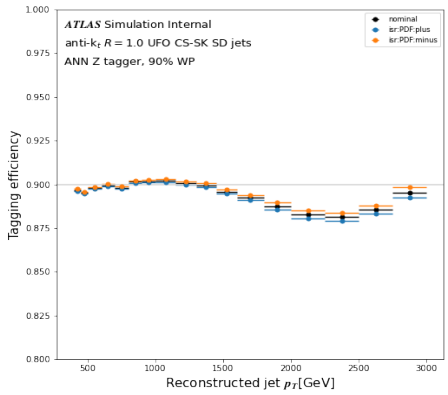
(b)



(c)

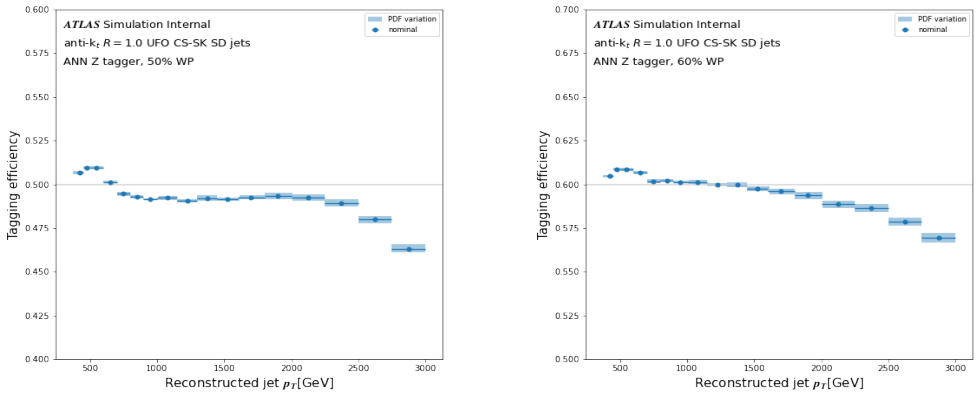


(d)



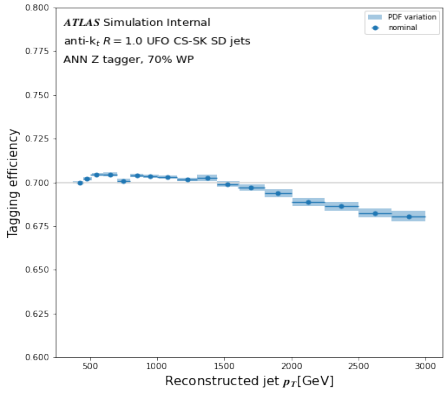
(e)

Figure B.17: ANN W/Z tagger PDF variation efficiencies as a function of p_{T} for Z-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

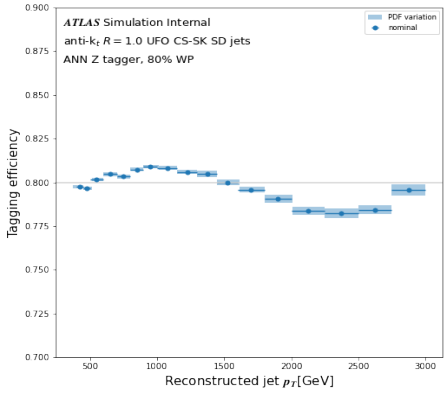


(a)

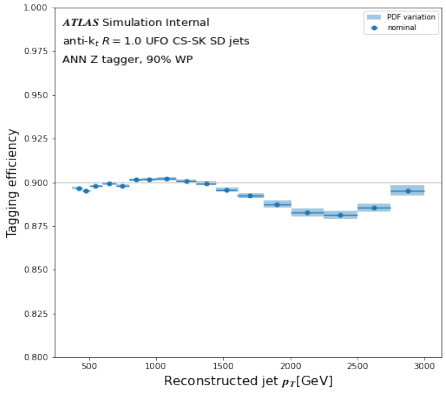
(b)



(c)

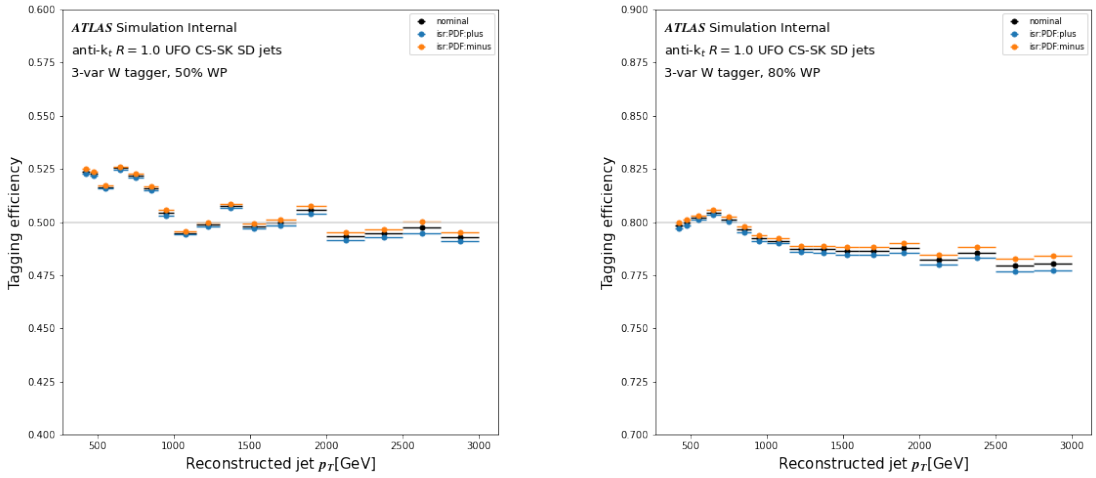


(d)



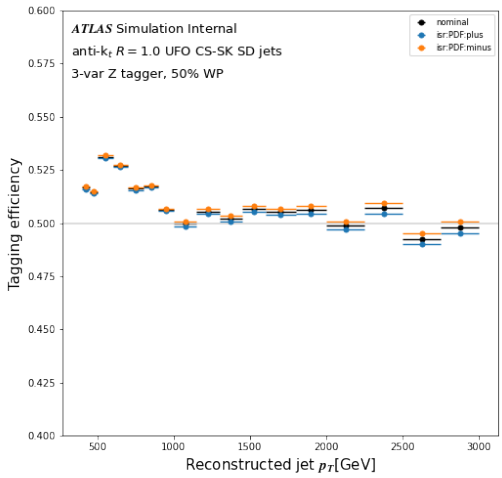
(e)

Figure B.18: ANN W/Z tagger PDF variation efficiency envelope as a function of p_{T} for Z-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

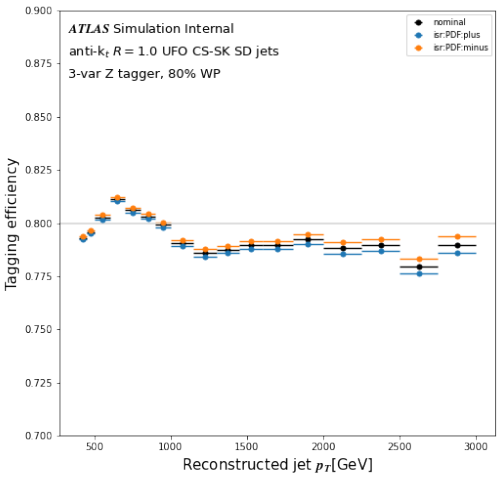


(a)

(b)

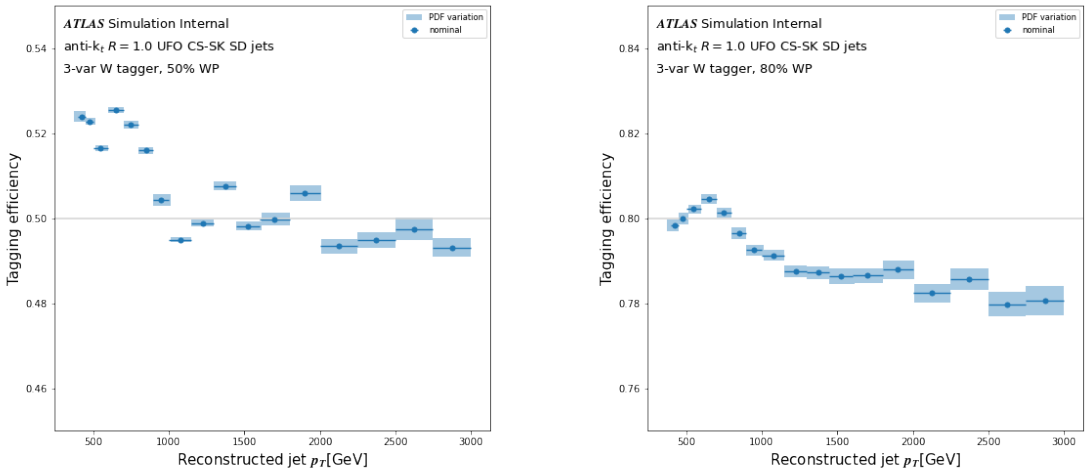


(c)



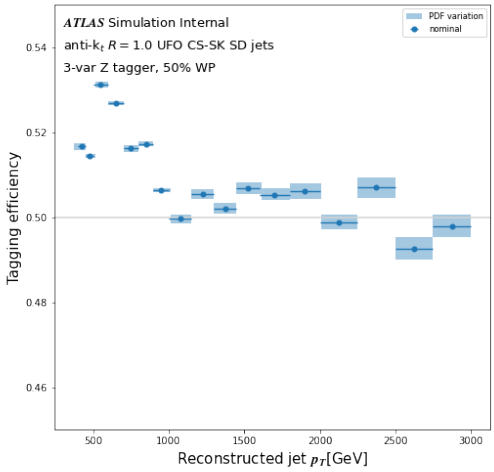
(d)

Figure B.19: Three variable W/Z taggers PDF variation efficiencies as a function of p_{T} for W-jets (top) at the (a) 50% and (b) 80% working points and for Z-jets (bottom) at the (c) 50% and (d) 80% working points.

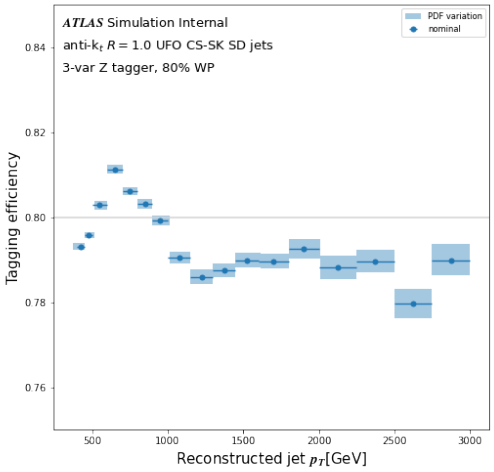


(a)

(b)



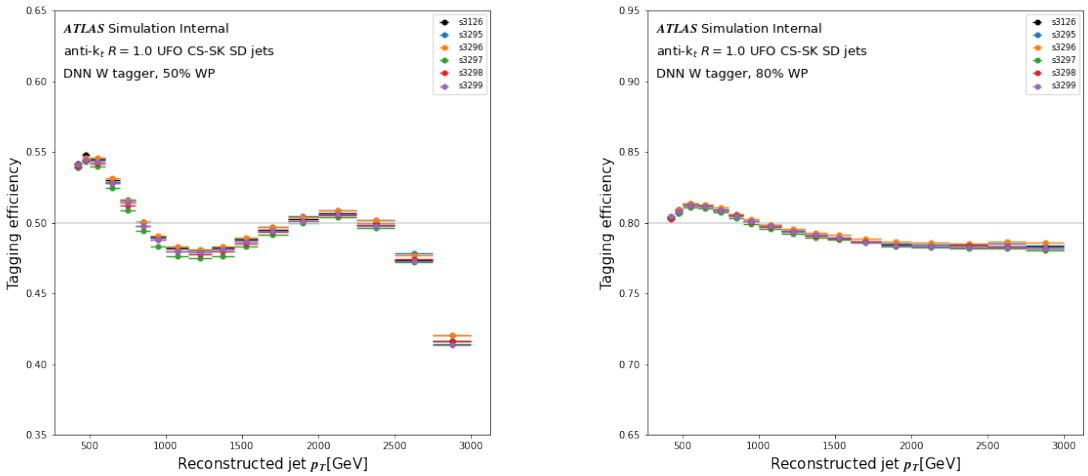
(c)



(d)

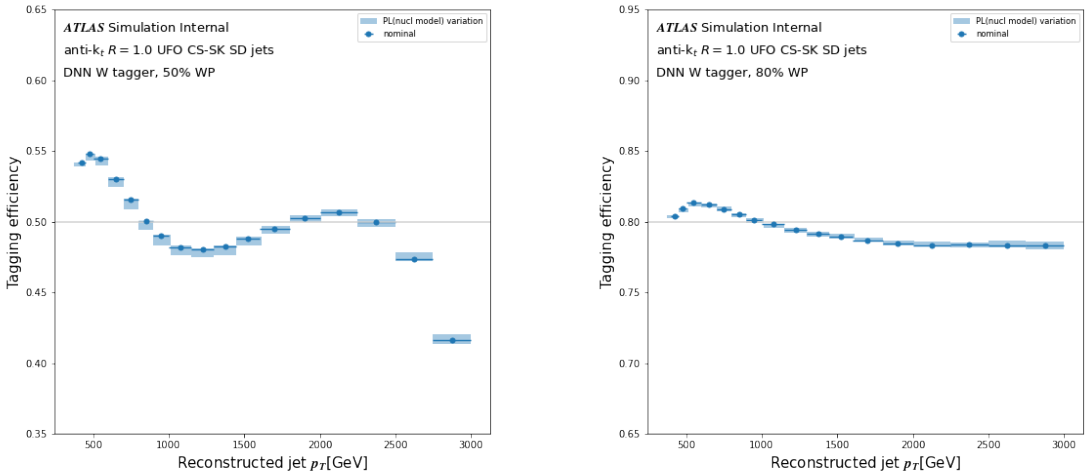
Figure B.20: Three variable W/Z taggers PDF variation efficiency envelope as a function of p_{T} for W-jets (top) at the (a) 50% and (b) 80% working points and for Z-jets (bottom) at the (c) 50% and (d) 80% working points.

Physics List Variations



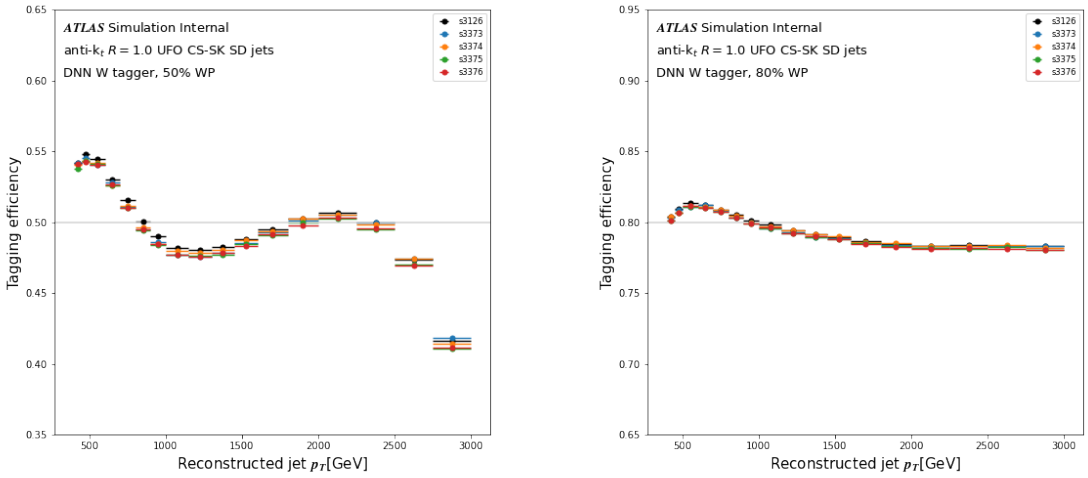
(a) (b)

Figure B.21: DNN W tagger Physics List nuclear model variation efficiencies as a function of p_{T} for W-jets at the (a) 50% and (b) 80% working points.



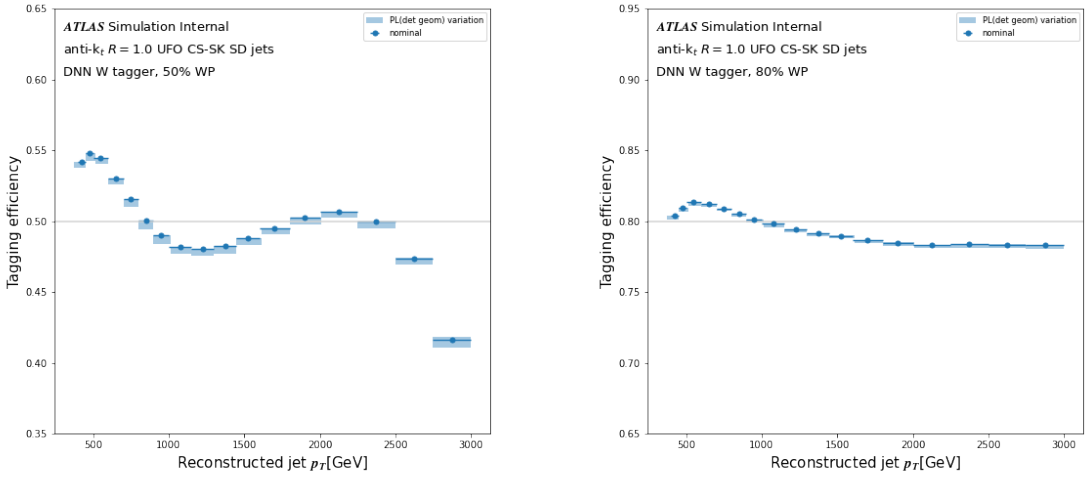
(a) (b)

Figure B.22: DNN W tagger Physics List nuclear model variation efficiency envelope as a function of p_{T} for W-jets at the (a) 50% and (b) 80% working points.



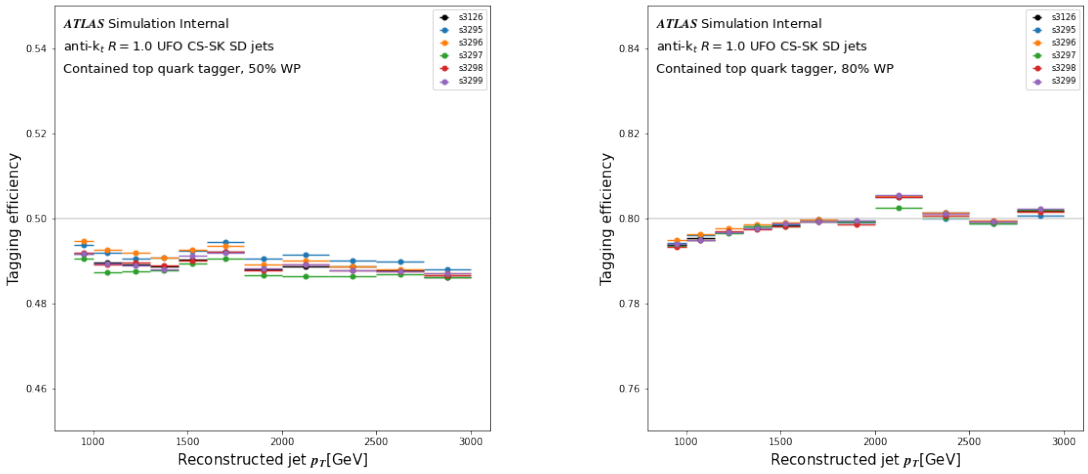
(a) (b)

Figure B.23: DNN W tagger Physics List detector geometry variation efficiencies as a function of p_{T} for W-jets at the (a) 50% and (b) 80% working points.



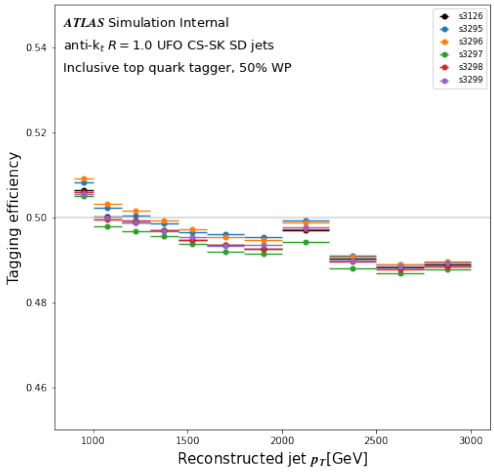
(a) (b)

Figure B.24: DNN W tagger Physics List detector geometry variation efficiency envelope as a function of p_{T} for W-jets at the (a) 50% and (b) 80% working points.

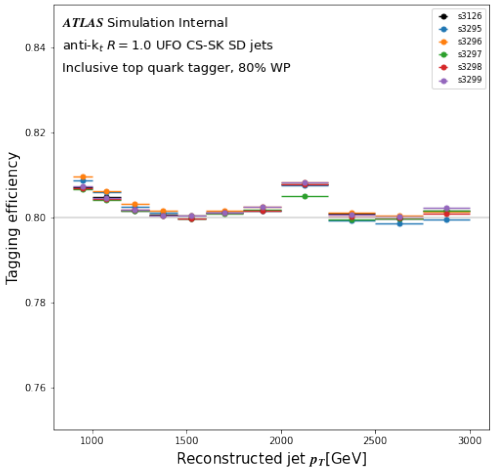


(a)

(b)

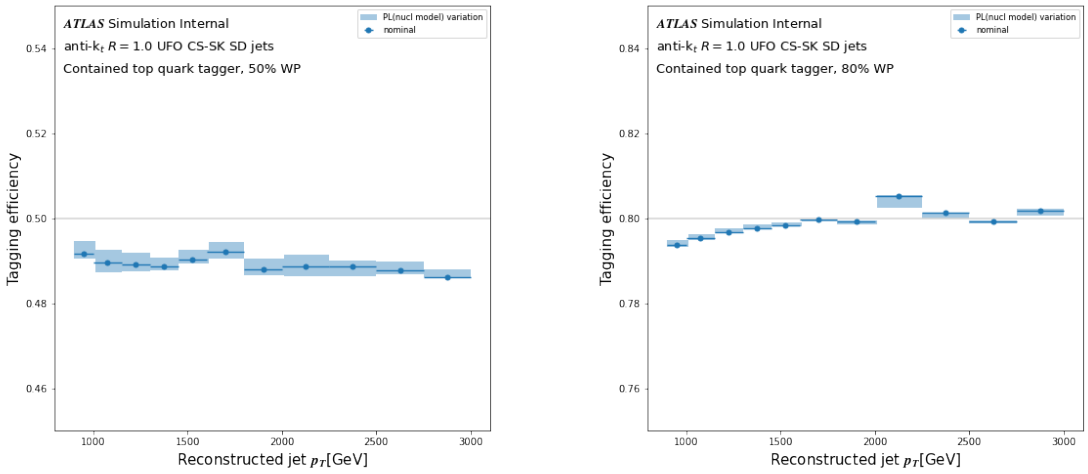


(c)



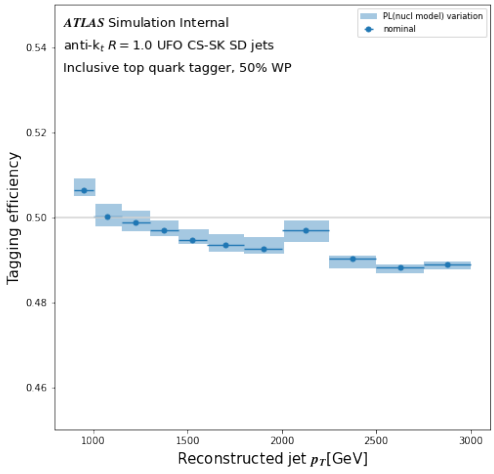
(d)

Figure B.25: DNN top taggers Physics List nuclear model variation efficiencies as a function of p_{T} for contained tops (top) at the (a) 50% and (b) 80% working points and for inclusive tops (bottom) at the (c) 50% and (d) 80% working points.

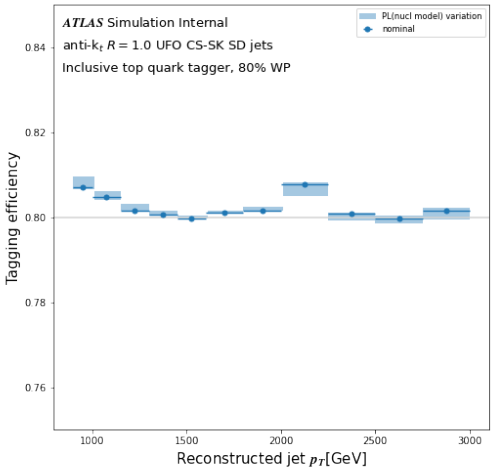


(a)

(b)

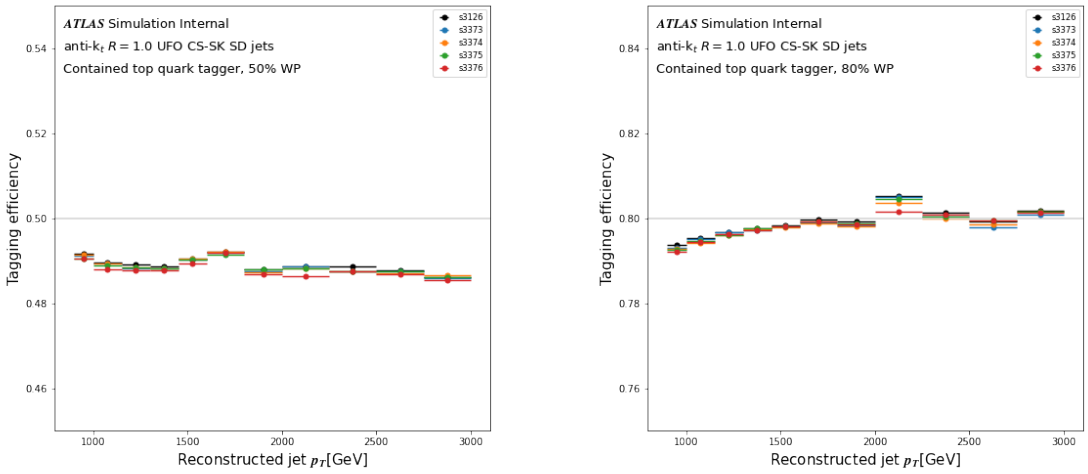


(c)



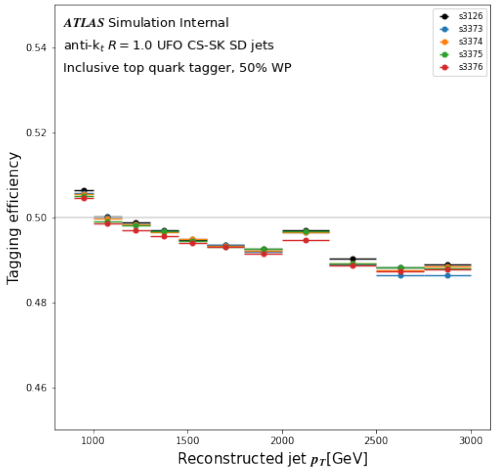
(d)

Figure B.26: DNN top taggers Physics List nuclear model variation efficiency envelope as a function of p_{T} for contained tops (top) at the (a) 50% and (b) 80% working points and for inclusive tops (bottom) at the (c) 50% and (d) 80% working points.

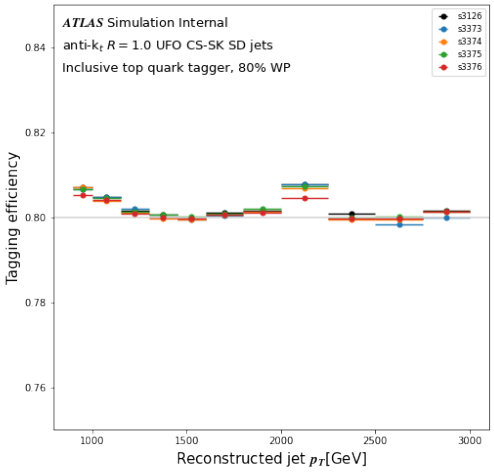


(a)

(b)

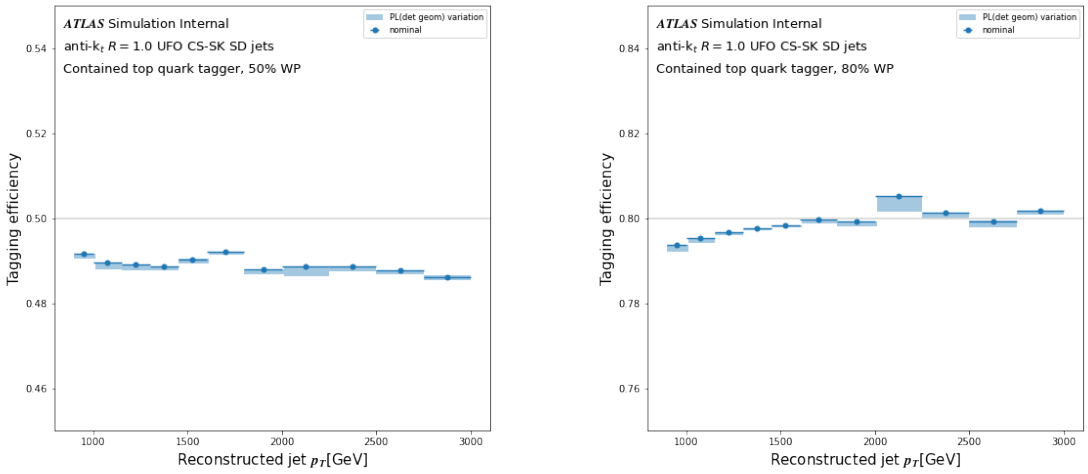


(c)



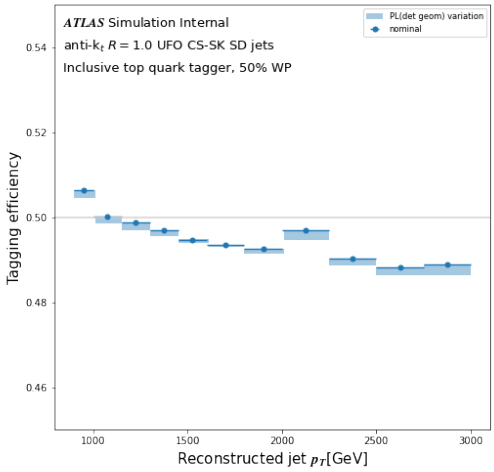
(d)

Figure B.27: DNN top taggers Physics List detector geometry variation efficiencies as a function of p_{T} for contained tops (top) at the (a) 50% and (b) 80% working points and for inclusive tops (bottom) at the (c) 50% and (d) 80% working points.

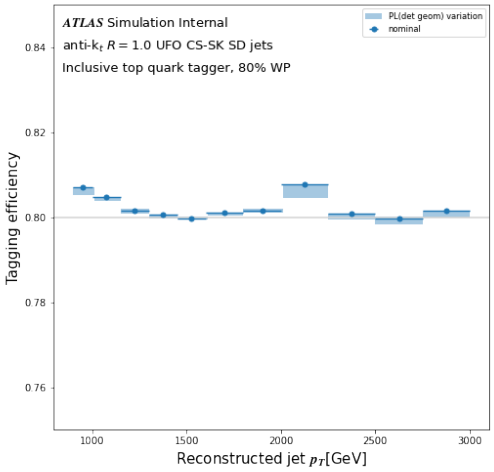


(a)

(b)

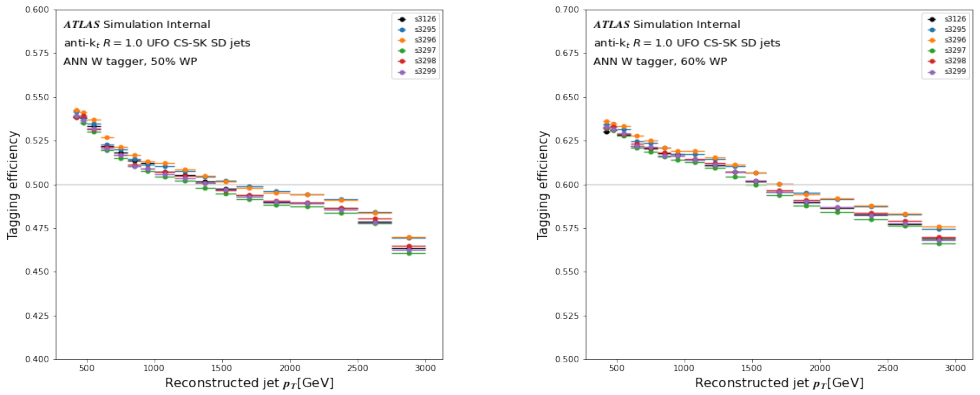


(c)



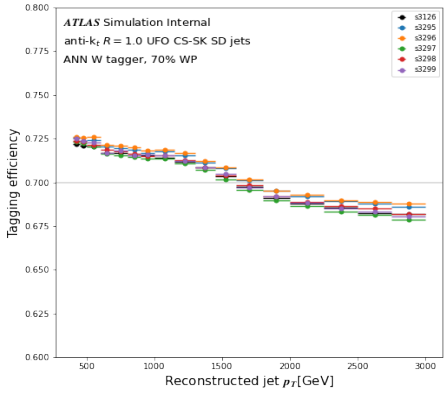
(d)

Figure B.28: DNN top taggers Physics List detector geometry variation efficiency envelope as a function of p_{T} for contained tops (top) at the (a) 50% and (b) 80% working points and for inclusive tops (bottom) at the (c) 50% and (d) 80% working points.

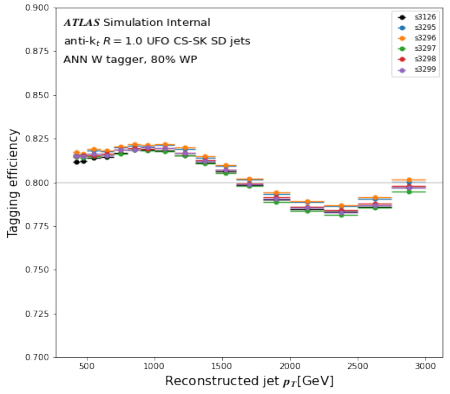


(a)

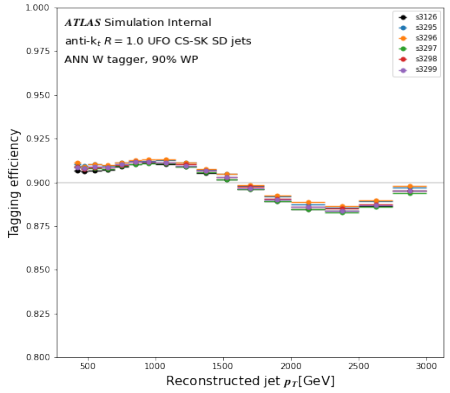
(b)



(c)



(d)



(e)

Figure B.29: ANN W/Z tagger Physics List nuclear model variation efficiencies as a function of p_{T} for W-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

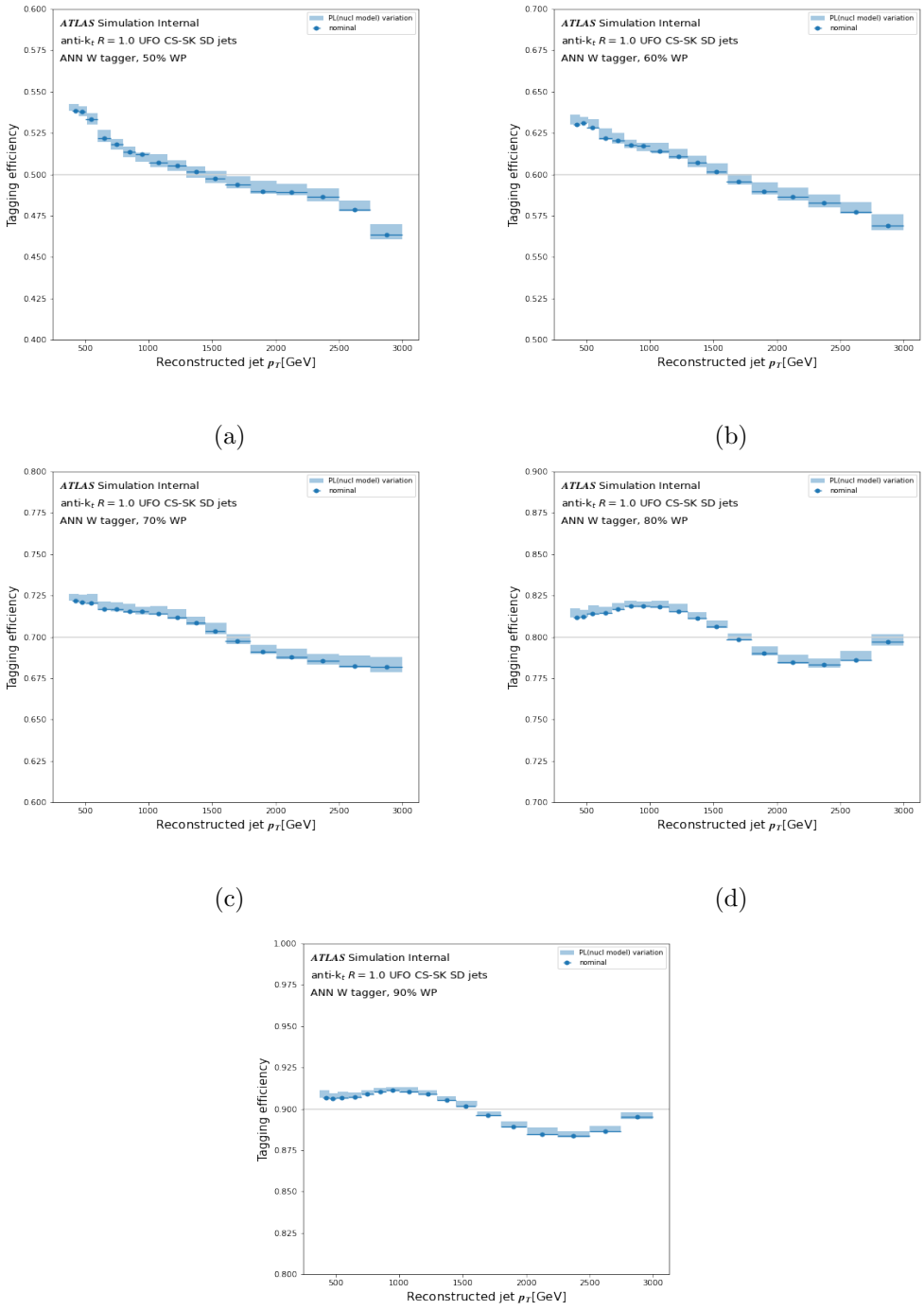


Figure B.30: ANN W/Z tagger Physics List nuclear model variation efficiency envelope as a function of p_{T} for W-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

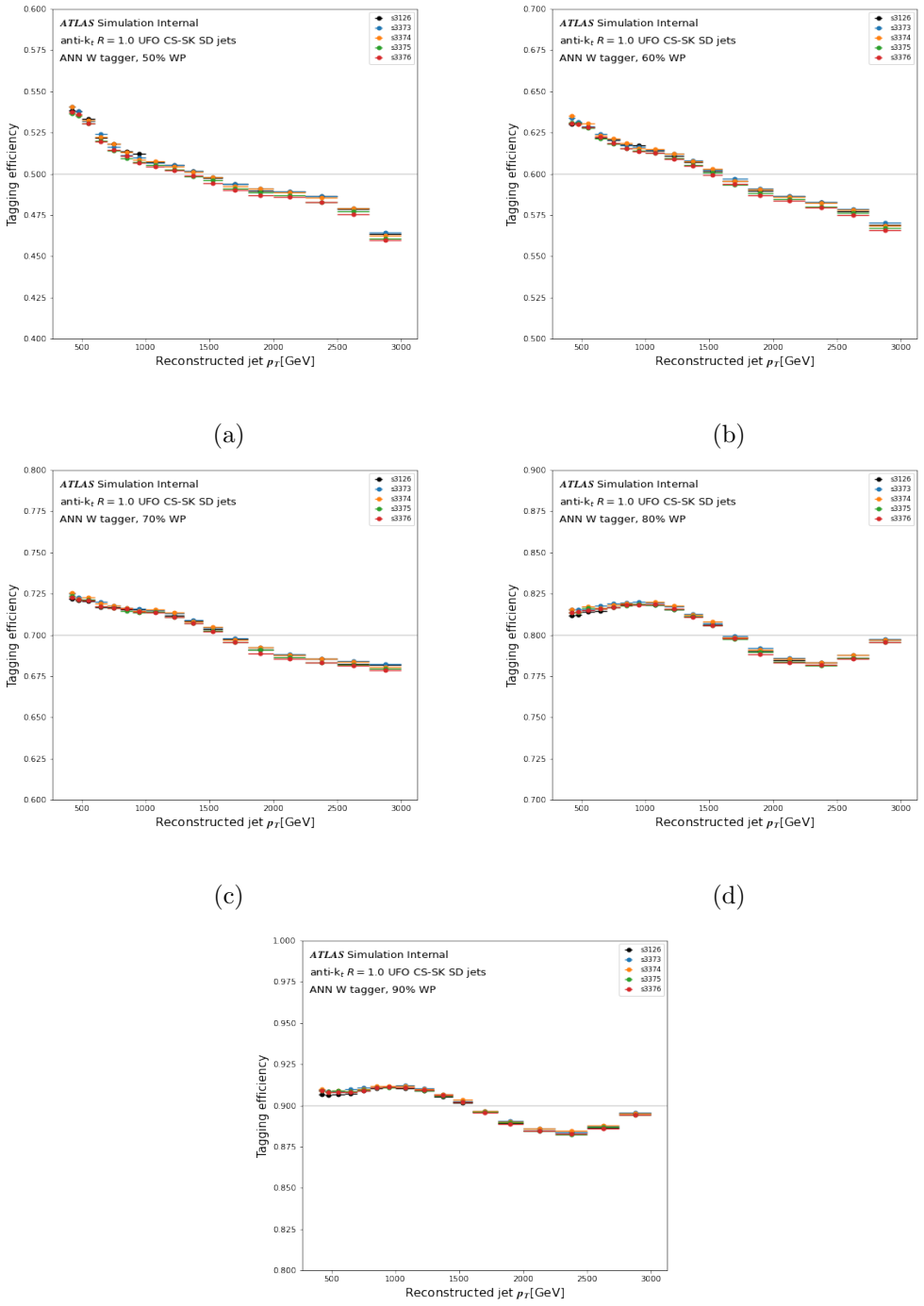
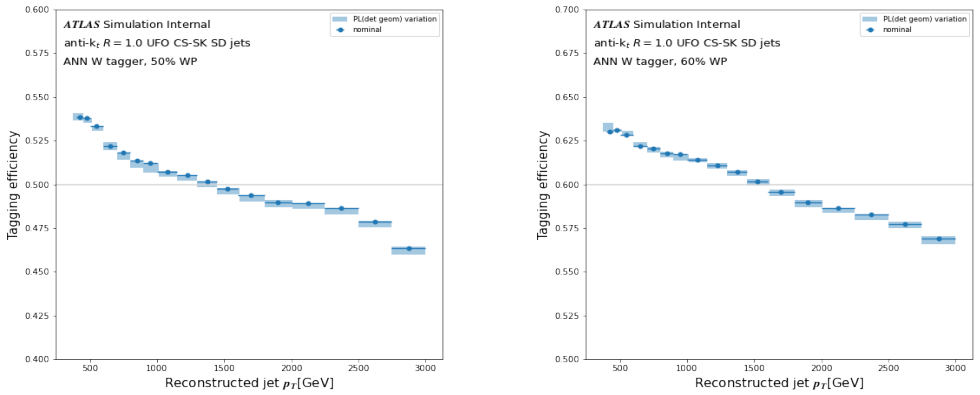
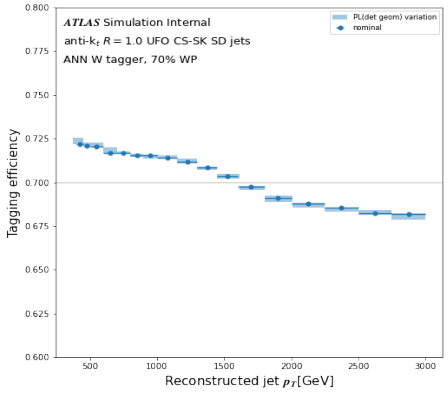


Figure B.31: ANN W/Z tagger Physics List detector geometry variation efficiencies as a function of p_{T} for W-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

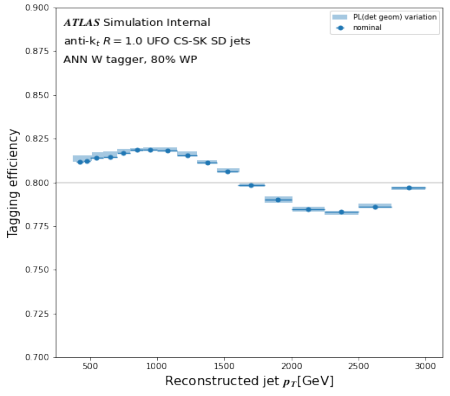


(a)

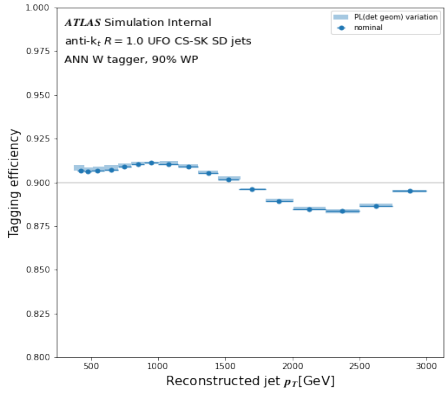
(b)



(c)

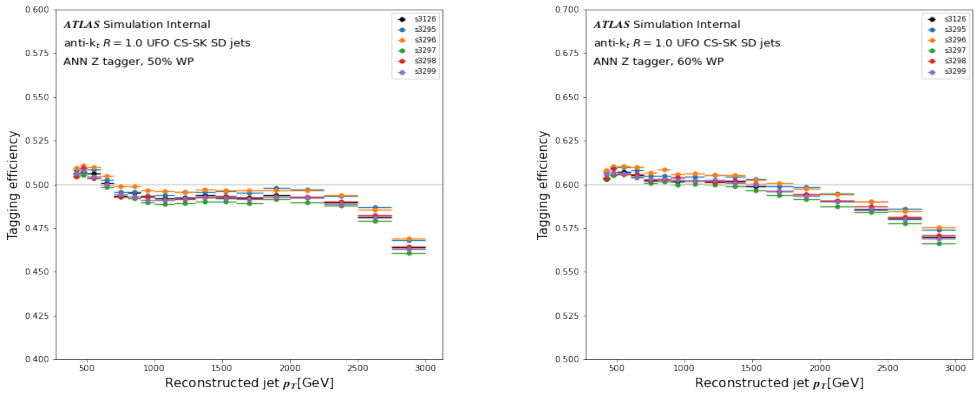


(d)



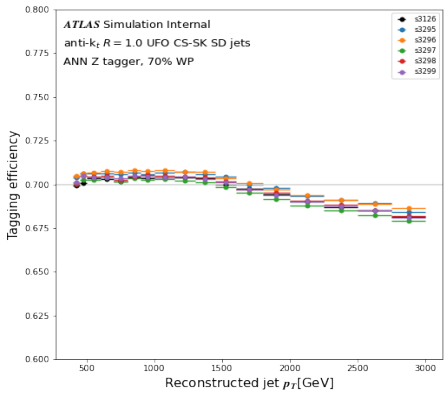
(e)

Figure B.32: ANN W/Z tagger Physics List detector geometry variation efficiency envelope as a function of p_{T} for W-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

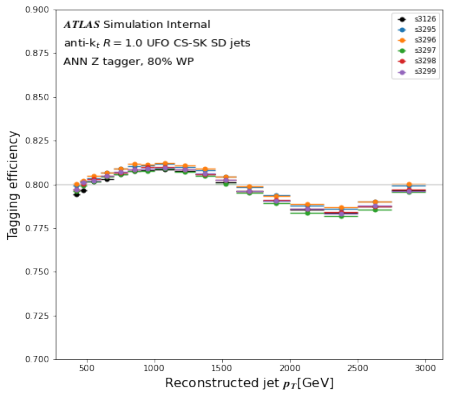


(a)

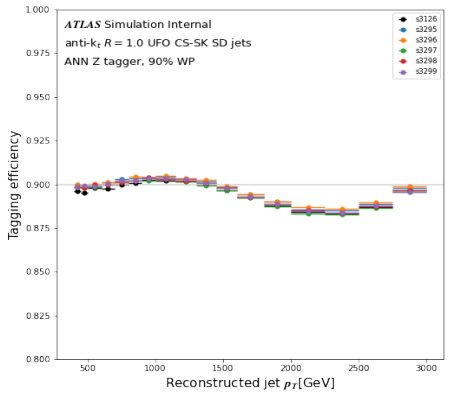
(b)



(c)

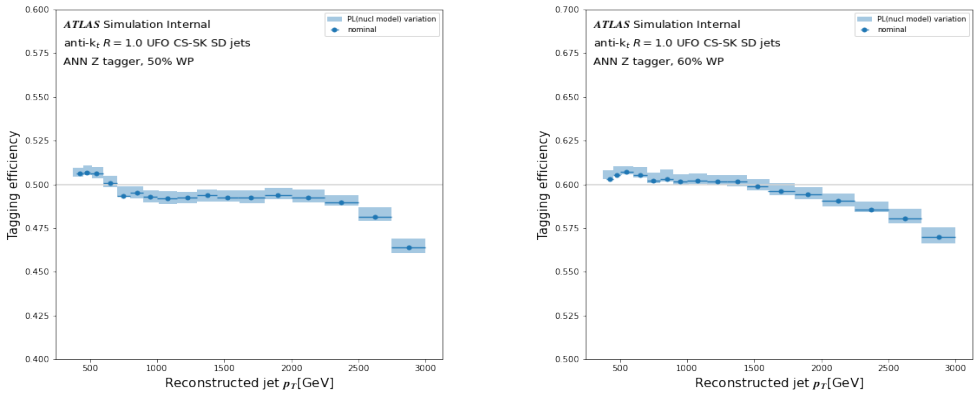


(d)



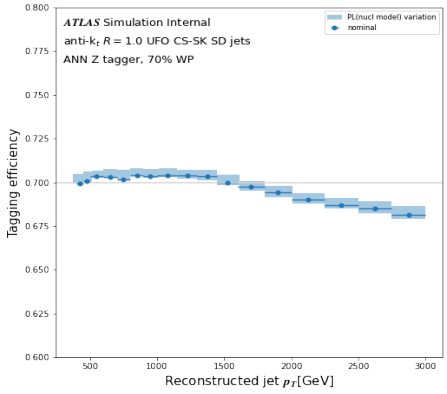
(e)

Figure B.33: ANN W/Z tagger Physics List nuclear model variation efficiencies as a function of p_{T} for Z-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

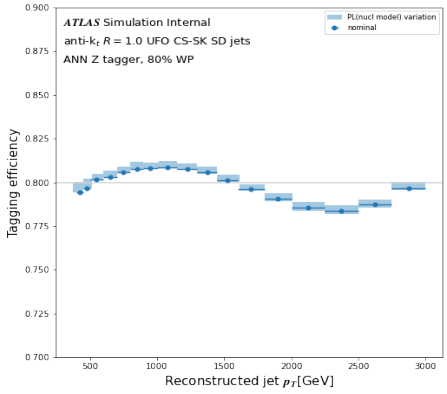


(a)

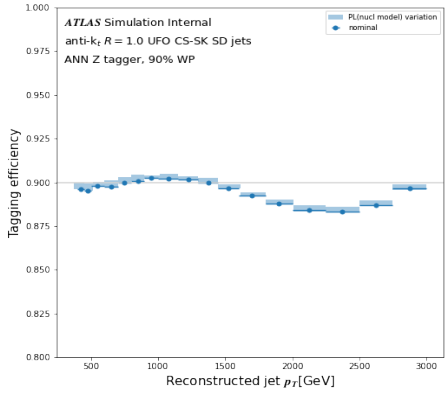
(b)



(c)

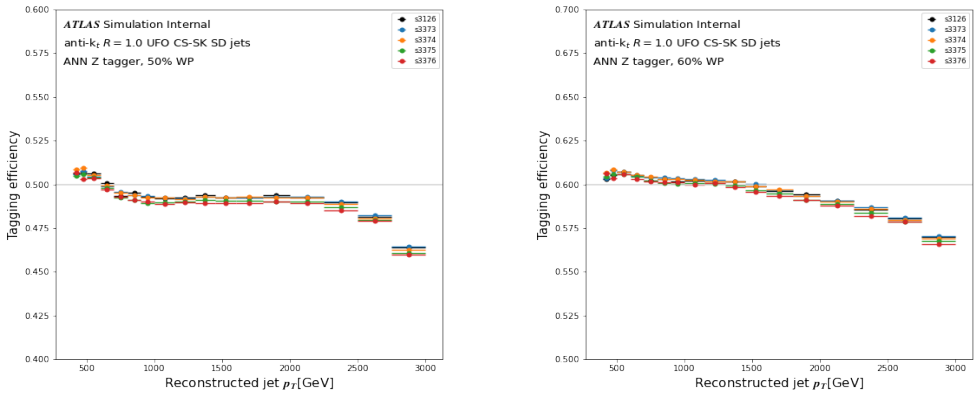


(d)



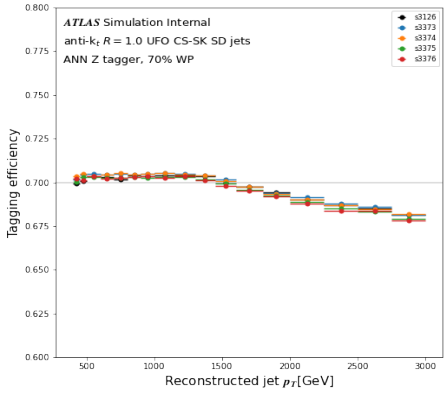
(e)

Figure B.34: ANN W/Z tagger Physics List nuclear model variation efficiency envelope as a function of p_{T} for Z-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

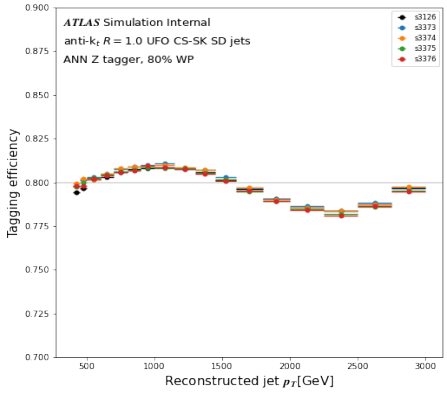


(a)

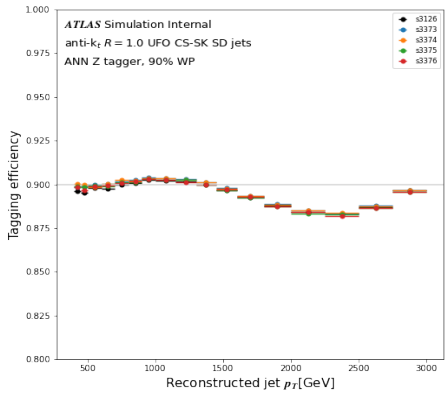
(b)



(c)

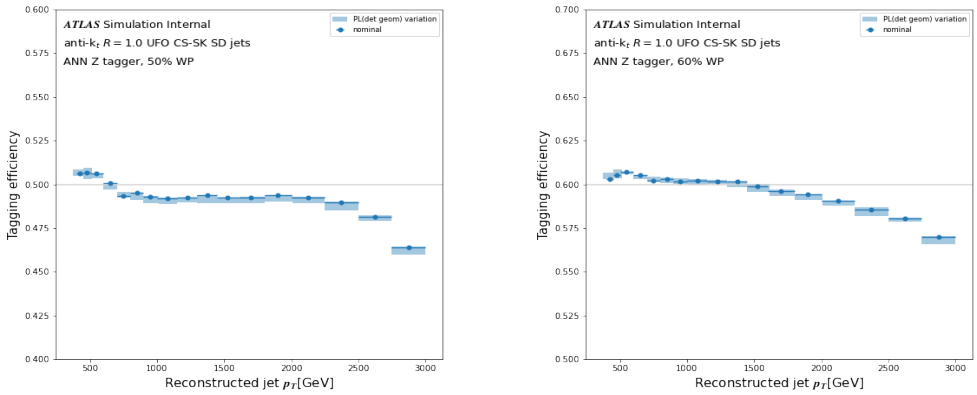


(d)



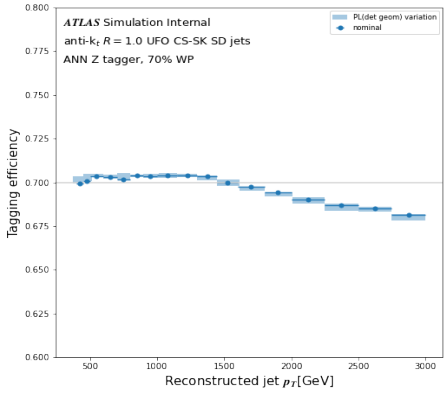
(e)

Figure B.35: ANN W/Z tagger Physics List detector geometry variation efficiencies as a function of p_{T} for Z-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

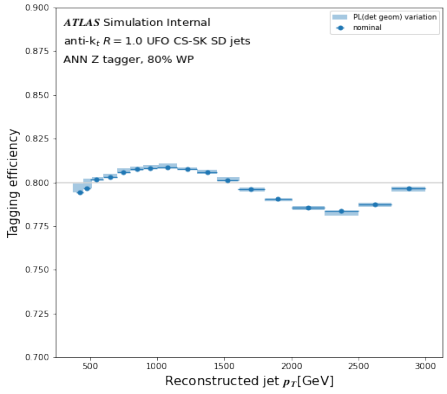


(a)

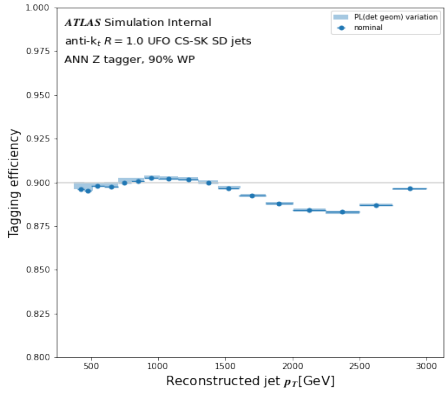
(b)



(c)

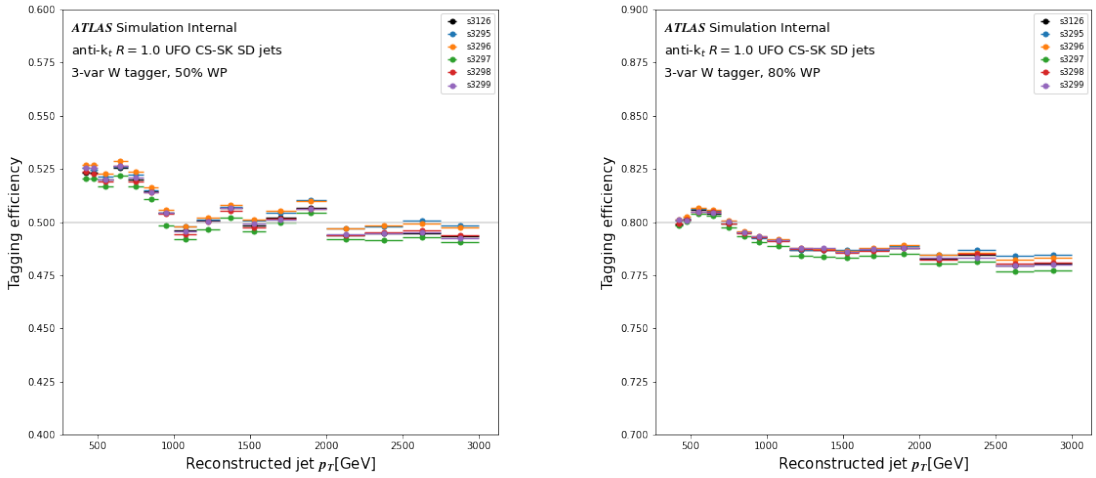


(d)



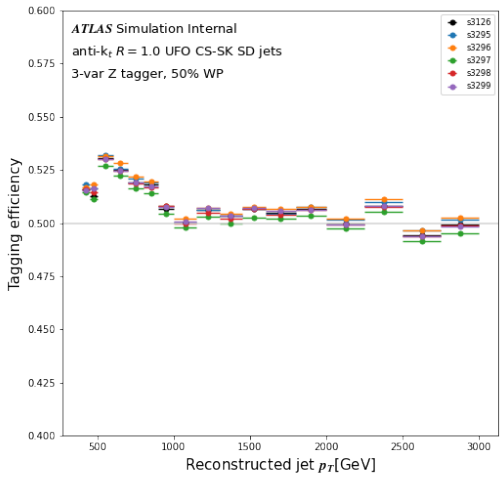
(e)

Figure B.36: ANN W/Z tagger Physics List detector geometry variation efficiency envelope as a function of p_{T} for Z-jets at the (a) 50%, (b) 60%, (c) 70%, (d) 80%, (e) 90% working points.

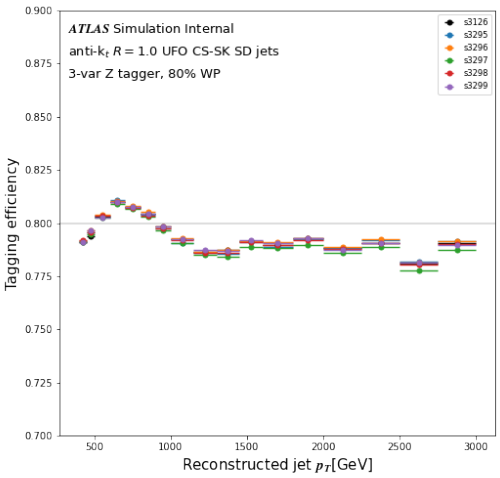


(a)

(b)

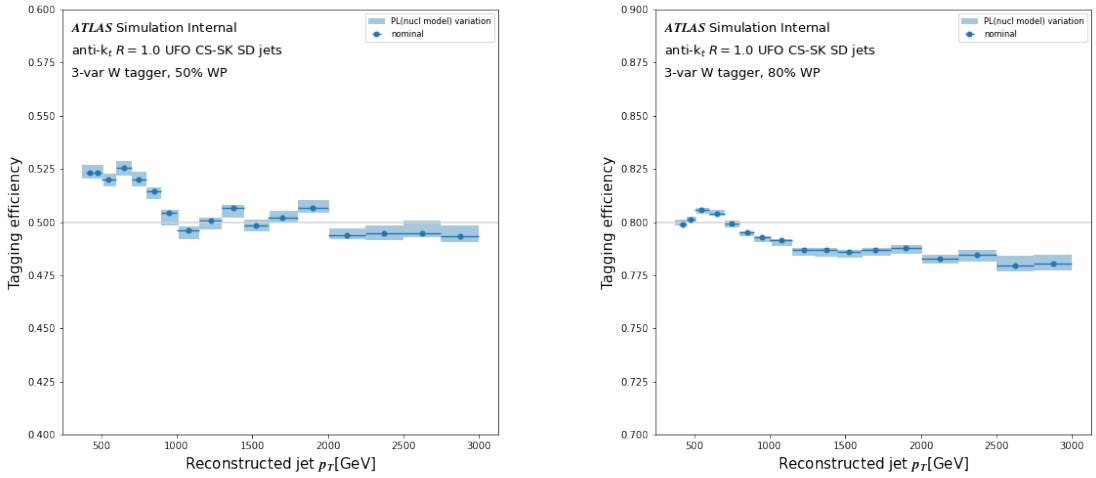


(c)



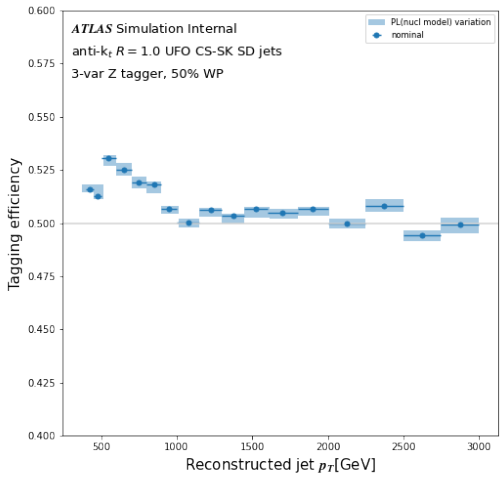
(d)

Figure B.37: Three variable W/Z taggers Physics List nuclear model variation efficiencies as a function of p_{T} for W-jets (top) at the (a) 50% and (b) 80% working points and for Z-jets (bottom) at the (c) 50% and (d) 80% working points.

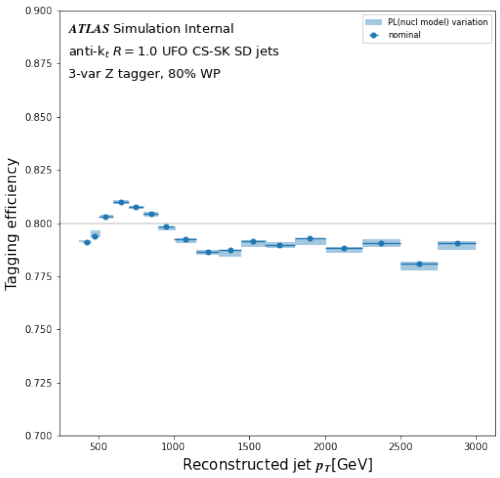


(a)

(b)

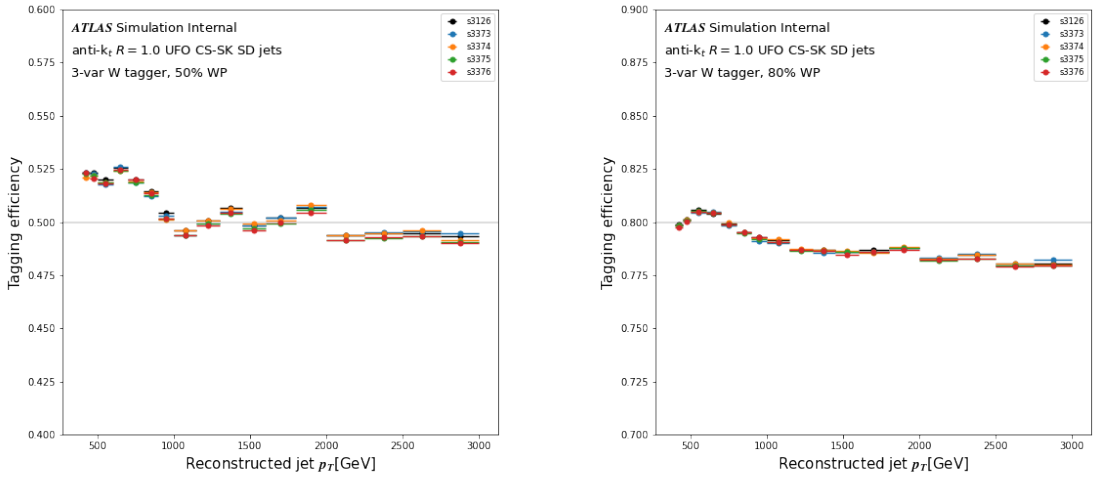


(c)



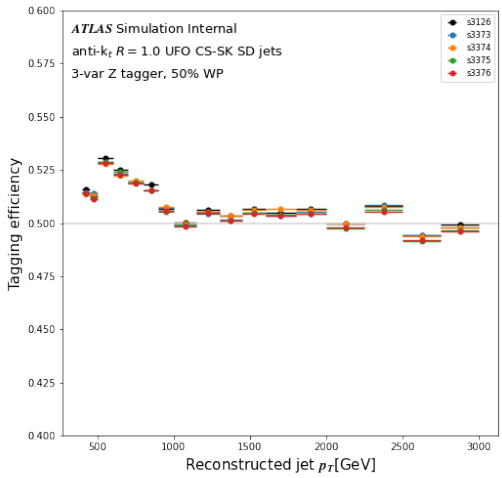
(d)

Figure B.38: Three variable W/Z taggers Physics List nuclear model variation efficiency envelope as a function of p_{T} for W-jets (top) at the (a) 50% and (b) 80% working points and for Z-jets (bottom) at the (c) 50% and (d) 80% working points.

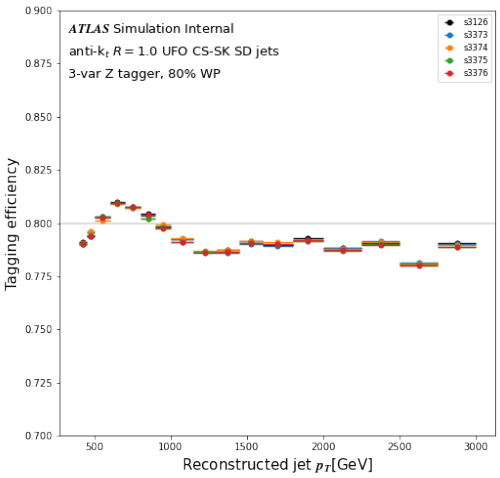


(a)

(b)

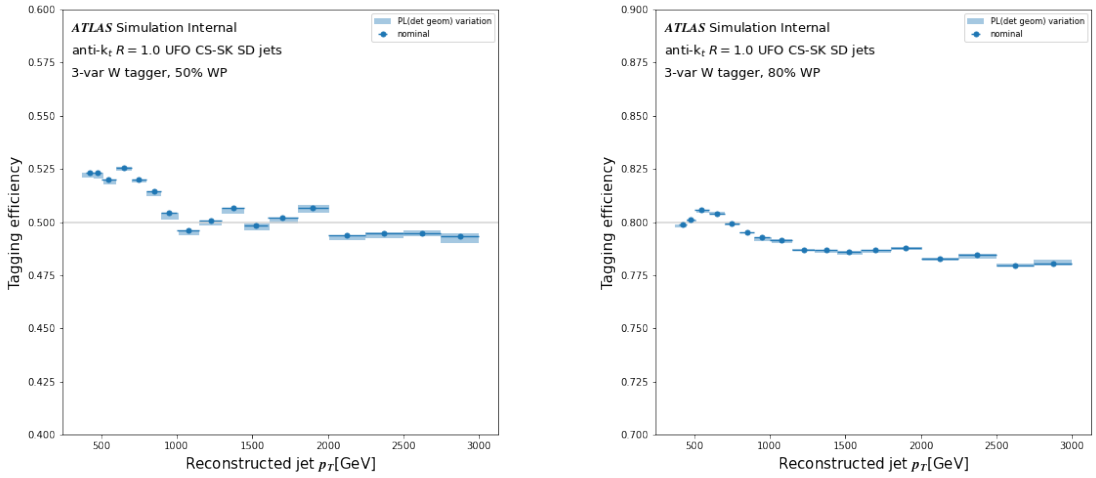


(c)



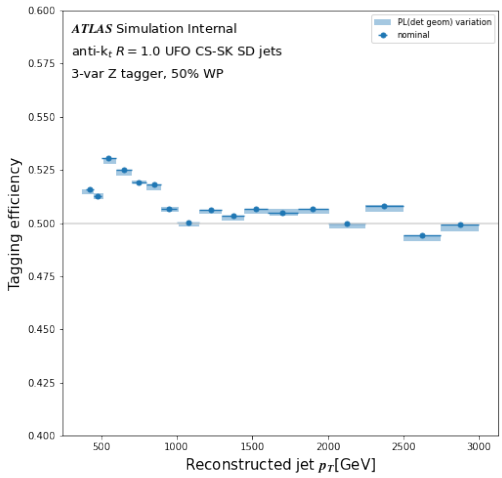
(d)

Figure B.39: Three variable W/Z taggers Physics List detector geometry variation efficiencies as a function of p_{T} for W-jets (top) at the (a) 50% and (b) 80% working points and for Z-jets (bottom) at the (c) 50% and (d) 80% working points.

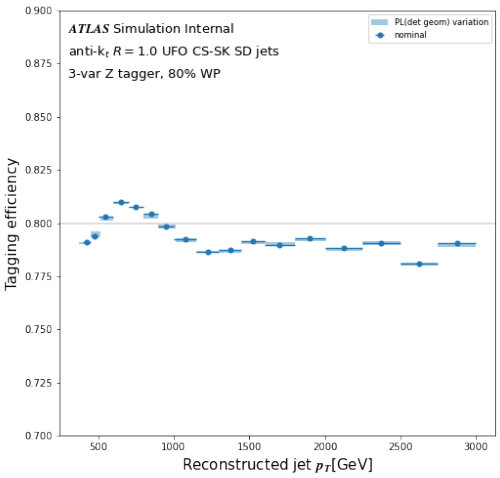


(a)

(b)



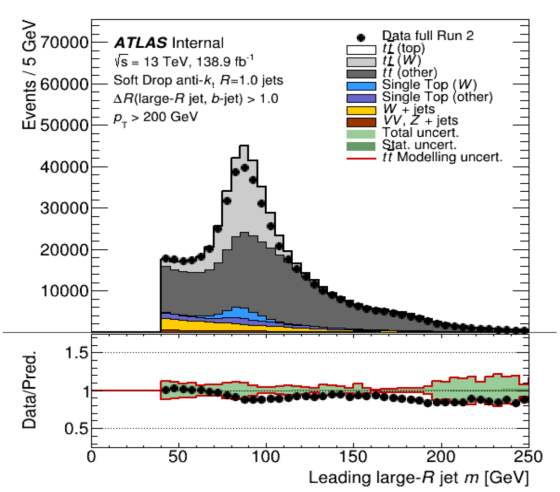
(c)



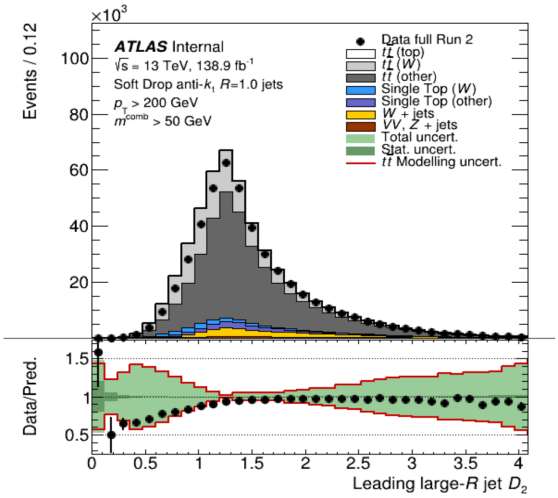
(d)

Figure B.40: Three variable W/Z taggers Physics List detector geometry variation efficiency envelope as a function of p_{T} for W-jets (top) at the (a) 50% and (b) 80% working points and for Z-jets (bottom) at the (c) 50% and (d) 80% working points.

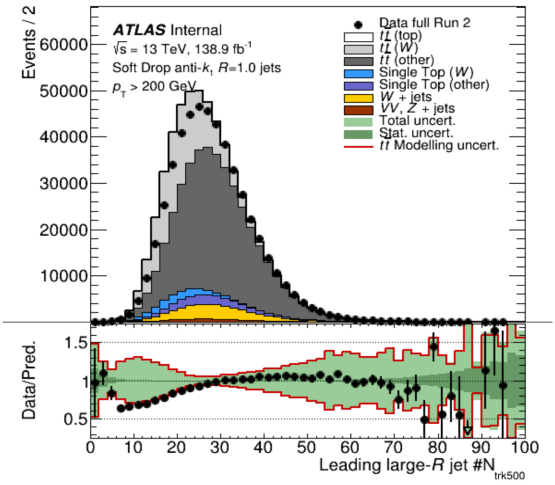
Jet Substructure Modeling



(a)

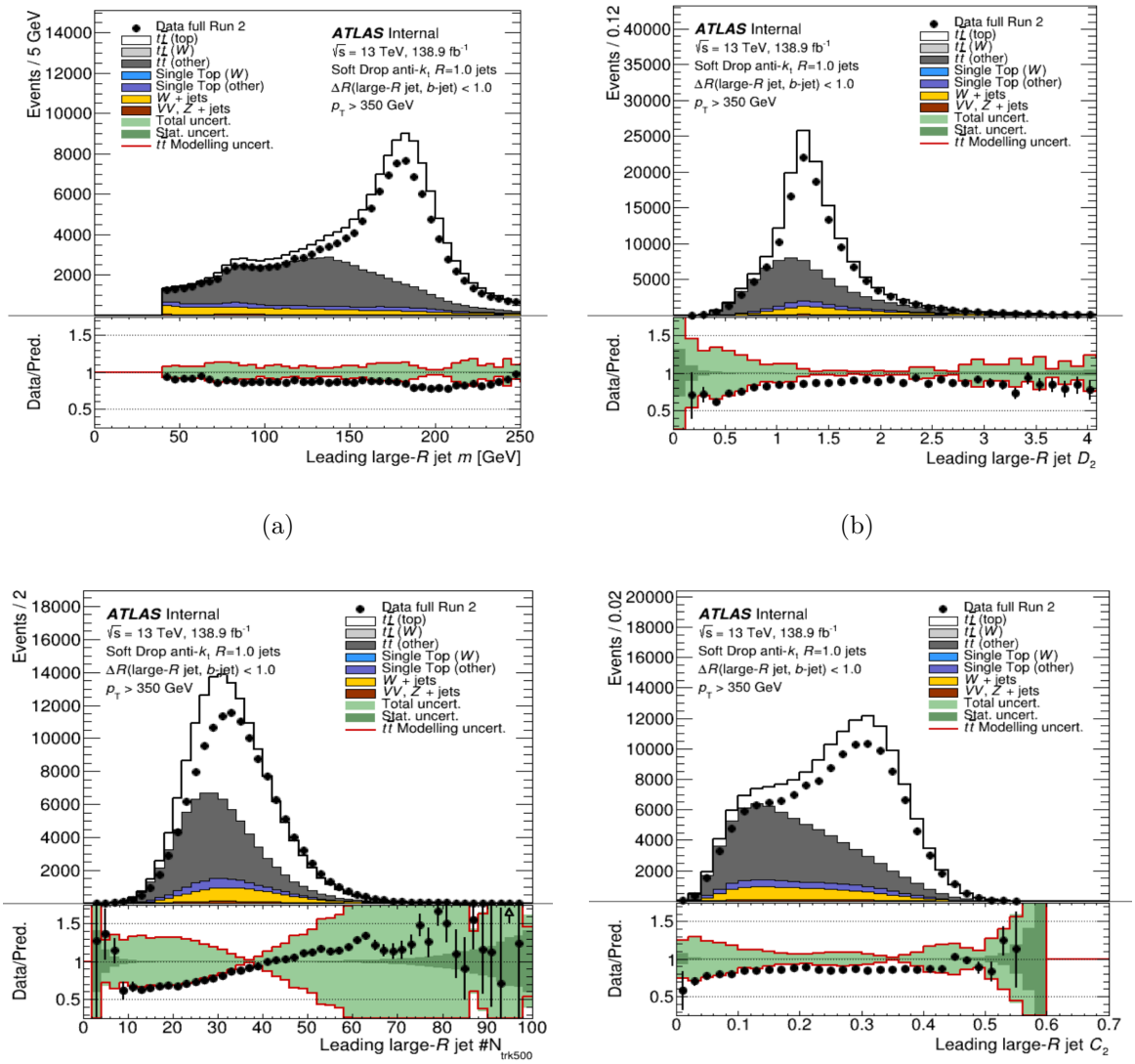


(b)



(c)

Figure B.41: Distributions for UFO SD CS-SK Large-R jets passing the W selection criteria of (a) mass, (b) D_2 and (c) n_{trk} [181].



(a) (b)

(c) (d)

Figure B.42: Distributions for UFO SD CS-SK Large-R jets passing the top quark selection criteria of (a) mass, (b) D_2, (c) n_{trk} and (d) C_2 [181].

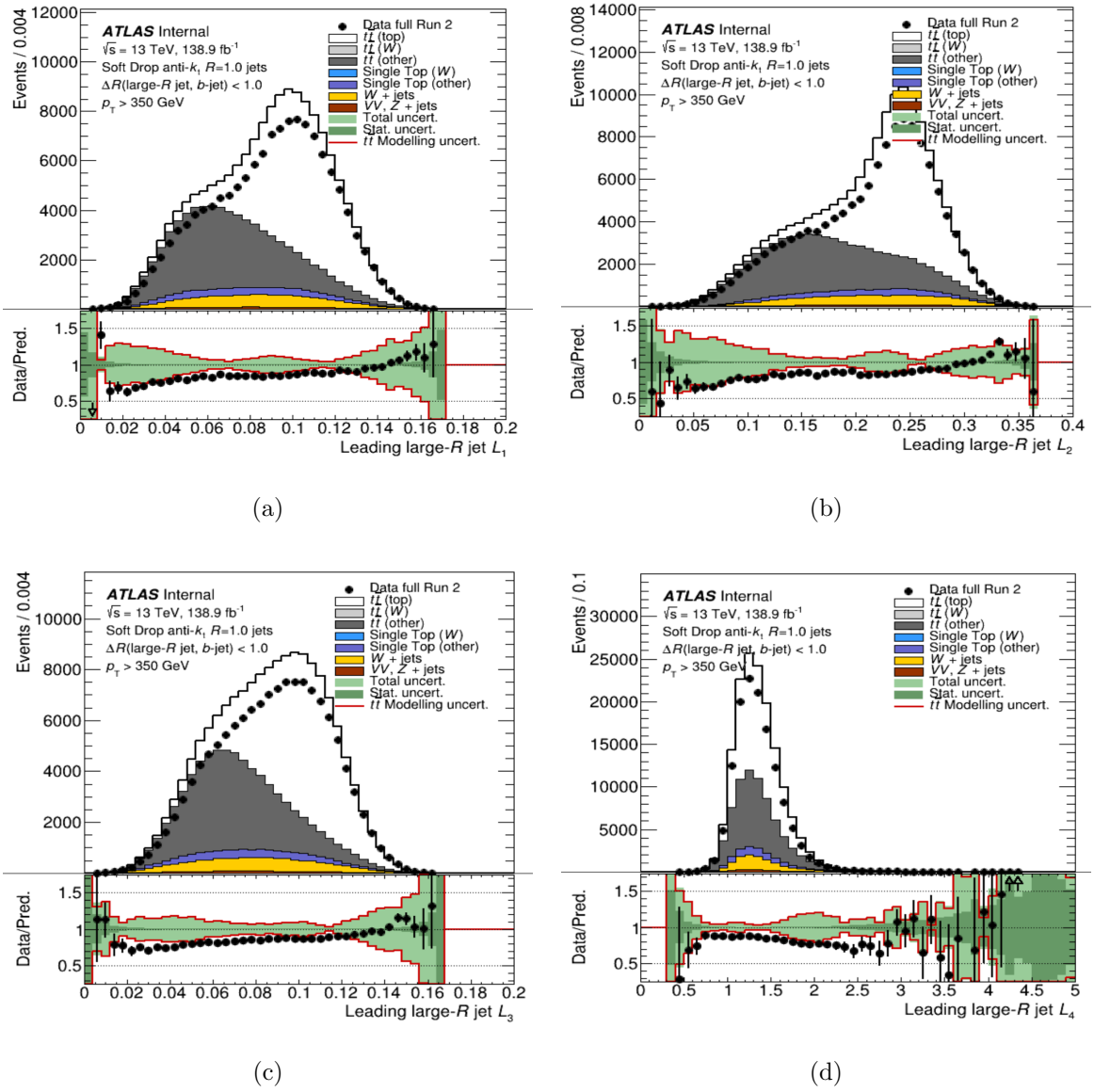
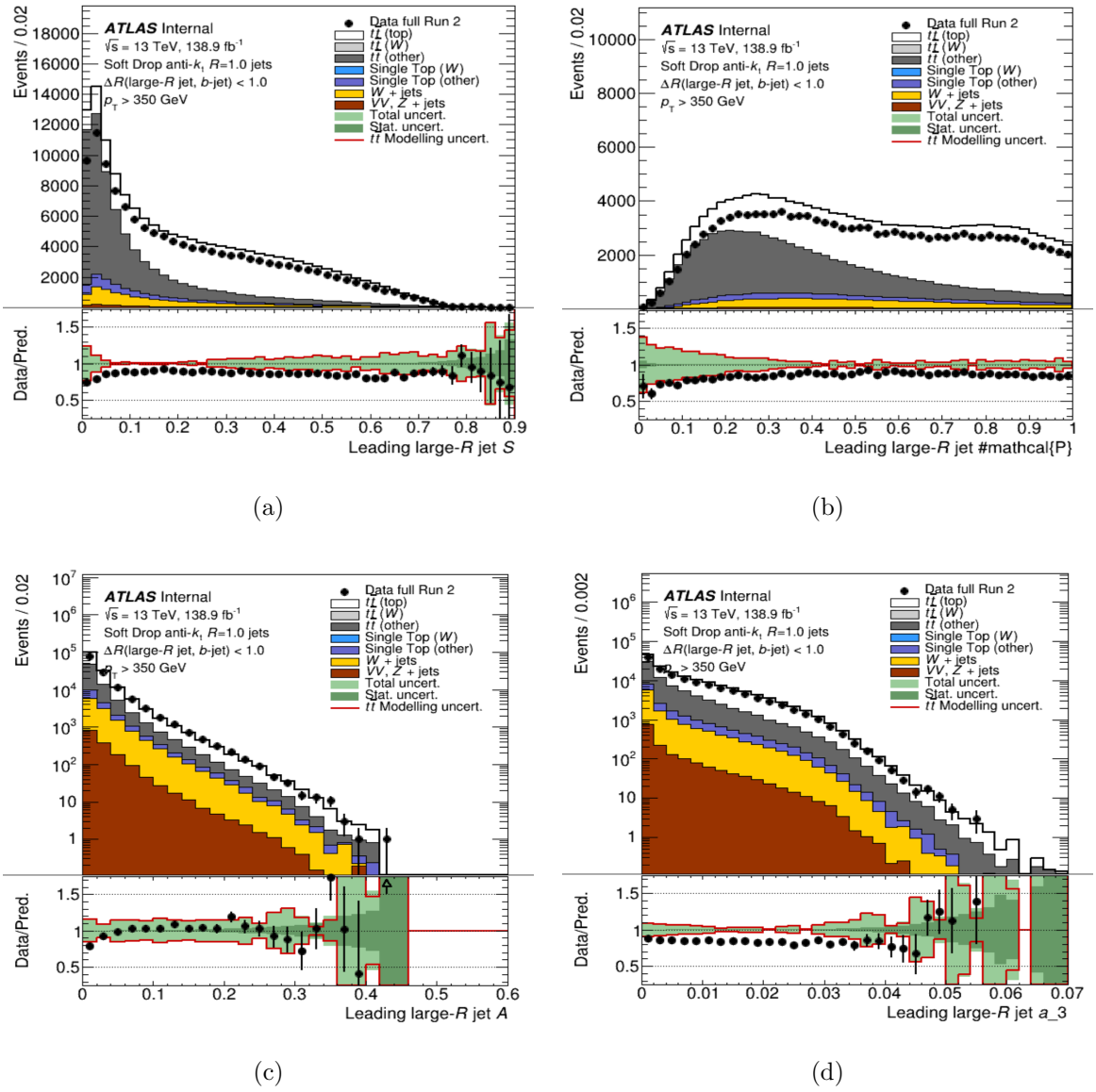


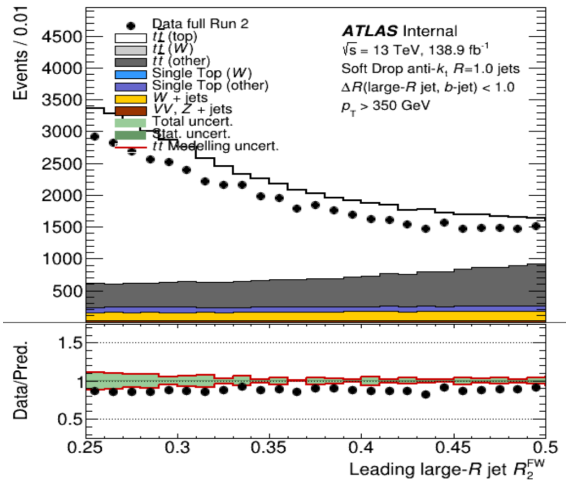
Figure B.43: Distributions for UFO SD CS-SK Large-R jets passing the top quark selection criteria of (a) L_1, (b) L_2, (c) L_3 and (d) L_4 [181].



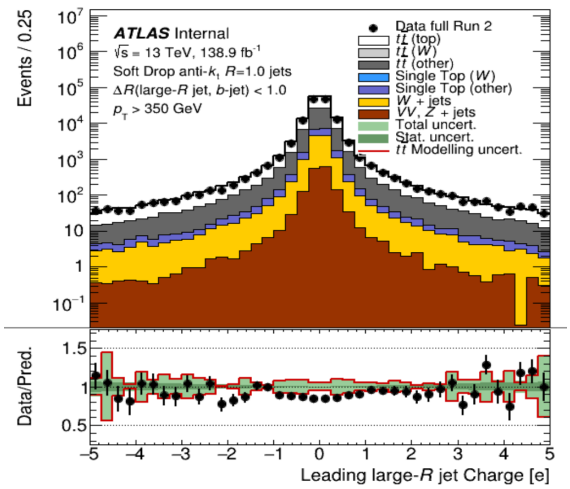
(a) (b)

(c) (d)

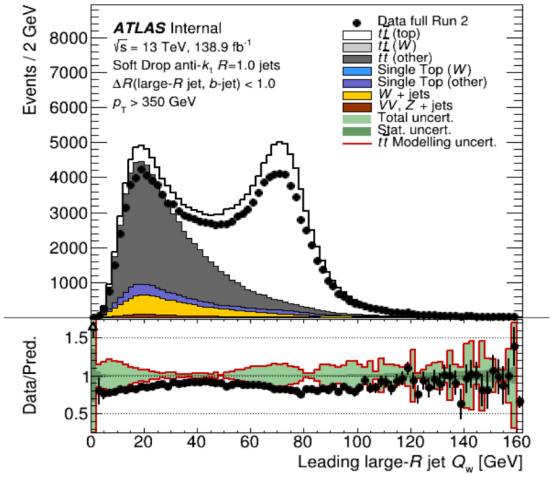
Figure B.44: Distributions for UFO SD CS-SK Large-R jets passing the top quark selection criteria of (a) sphericity, (b) planar flow, (c) aplanarity and (d) angularity [181].



(a)



(b)



(c)

Figure B.45: Distributions for UFO SD CS-SK Large-R jets passing the top quark selection criteria of (a) Fox-Wolfram moment, (b) jet total charge and (c) Q_w [181].

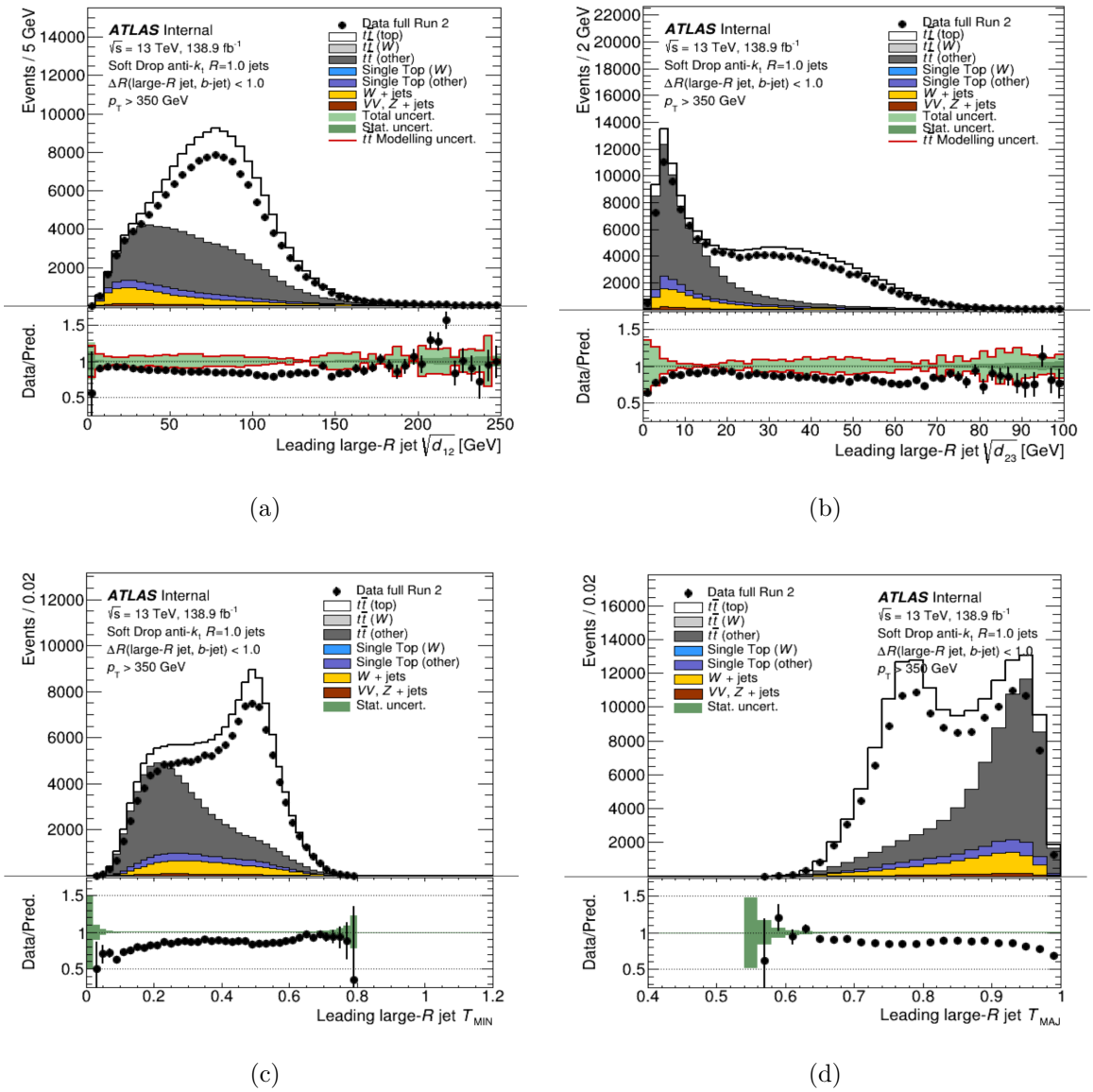
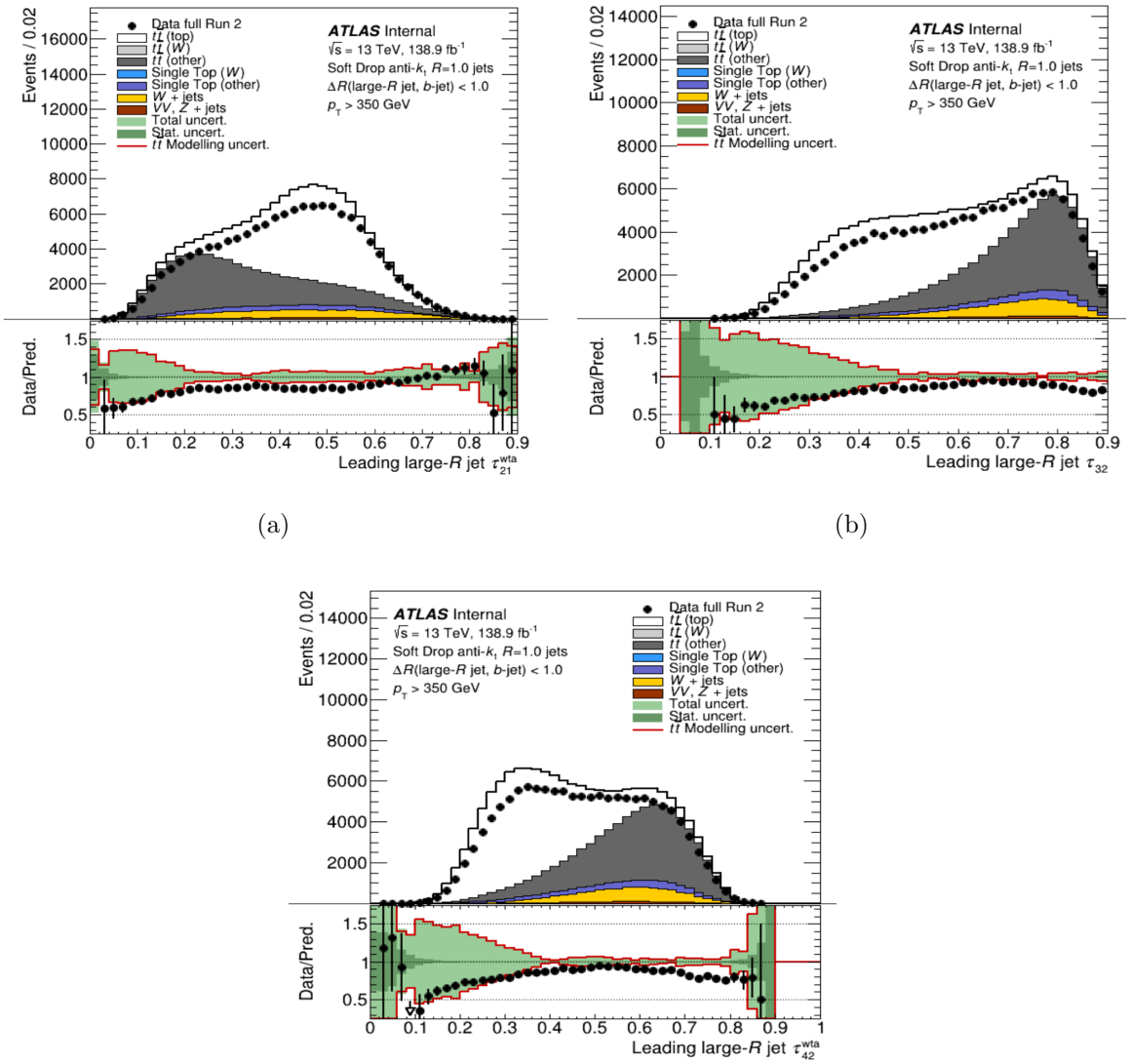


Figure B.46: Distributions for UFO SD CS-SK Large-R jets passing the top quark selection criteria of (a) $\sqrt{d_{12}}$, (b) $\sqrt{d_{23}}$, (c) thrust minor and (d) thrust major [181].



(a) (b)

(c)

Figure B.47: Distributions for UFO SD CS-SK Large-R jets passing the top quark selection criteria of (a) τ_{21}^{wta}, (b) τ_{32} and (c) τ_{42}^{wta} [181].