SEARCH FOR VECTOR-LIKE QUARKS AT $\sqrt{S} = 13$ TeV USING THE ATLAS DETECTOR

By

Carlos Josué Buxó Vázquez

A DISSERTATION

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

Physics — Doctor of Philosophy

2023

ABSTRACT

SEARCH FOR VECTOR-LIKE QUARKS AT $\sqrt{S} = 13$ TeV USING THE ATLAS DETECTOR

By

Carlos Josué Buxó Vázquez

This dissertation presents two research topics. The first topic focuses on the tagging of jets to hadronically decaying top quarks and W bosons in the ATLAS detector. Two jet tagging algorithm optimization studies are described. The second topic focuses on the search for vector-like top quarks (T), which are predicted by beyond the Standard Model theories that aim to solve the Hierarchy Problem, using a dataset of proton-proton collisions with a center of mass energy of $\sqrt{s} = 13$ TeV collected with the ATLAS detector. Two search analyses that probe different production mechanisms of the T are performed towards this goal.

The first jet tagging algorithm study focuses on the optimization of two deep neural network (DNN) top taggers and a three-variable W tagger, all of which use information from the substructure of jets. The tagging signal efficiency is extracted for each tagger both in Standard Model (SM) Monte Carlo (MC) simulations and in the data that was collected by the ATLAS detector in 2015-2017. The performance of the taggers in MC is calibrated to that of the data with the derivation of tagging signal efficiency scale factors. Additionally, uncertainties are derived for this measurement, which take into account effects from the MC modeling of the SM processes considered and the reconstruction and calibration of the different physics objects used in this measurement.

The second jet tagging algorithm study consists of a topological data analysis (TDA) of jets that analyzes their simplicial homology. A framework that applies a persistent homology

analysis and the Mapper algorithm to jets is devised. The information obtained from this framework is applied in the design of a DNN and convolutional graph neural network (GNN) top tagging algorithms. Optimization studies were performed in which these two taggers achieved a comparable performance to the substructure-based DNN top taggers.

Two search analyses for a T are performed, one targeting the single production mechanism of a T and the other targeting the pair production mechanism of $T\bar{T}$. Both analyses focus on the decay topologies $T \to Ht$ and $T \to Zt$ in final states that include a single electron or muon. A search strategy is devised for each analysis that takes advantage of several experimental features that are unique to each production mechanism. The tagging of jets to hadronically decaying top quarks and W, Z, and Higgs bosons is a cornerstone of the search strategies. A statistical analysis is performed for both searches to test for the presence of potential T production events in the data. No significant excesses over the SM prediction are observed in both searches, and 95% CL upper limits are set on the T and $T\bar{T}$ production cross sections. These limits are interpreted as exclusion limits on the T mass and other theory parameters that vary depending on the signal benchmark considered.

Dedicado a mi mamá y la cariñosa memoria de mi papá.
(Dedicated to my mom and the loving memory of my dad.)

ACKNOWLEDGMENTS

The journey to my Ph.D. has been a long and arduous one that wouldn't have been possible without the influence of many people. Thank you all for making this odyssey possible.

I would like to give a huge thank you to my advisor, Wade Fisher. I'm especially grateful for accepting me as his graduate student and for the wonderful experiences that I have had during my graduate studies under his guidance. I feel lucky to have an advisor who is a walking encyclopedia of particle physics and data science knowledge, who is passionate about teaching, and for the freedom and encouragement he has given me to explore different ideas in my research. Thank you for the support, guidance, instilled knowledge, and friendship you have provided me during this journey.

I'm also thankful to the people I met and collaborated with while working on my jet tagging research and VLQ search analyses for the guidance and imparted knowledge they have provided me. A special thank you goes to Trisha Farooque, whom I've worked with throughout most of my graduate career. Your mentoring, knowledge, and friendship have become invaluable during my growth over these years. Thank you. I would also like to thank Casey Bellgraph, who worked under my guidance. Thank you for teaching me how to be a mentor and allowing me the opportunity to impart my knowledge to you.

I would like to give thanks to the group of ATLAS post-docs here at MSU, who have also provided me with their guidance and knowleged. I appreciate how they have made our research group feel more like a group of friends, especially during my visit to CERN.

To all the friends I have made during these years, thank you for your emotional support and all the good moments we have shared that have made this journey an enjoyable one.

Finalmente, estoy eternamente agradecido a mis padres por ser los pilares que me alzan y

haberme criado para ser la persona que soy hoy en día. Papá, aunque nuestro tiempo junto haya sido corto, estoy agradecido por el amor y las enseñanzas que me brindastes, las cuales son más que suficientes para toda la vida. Mamá, gracias por todo el soporte emocional y amor incondicional que siempre me has brindado, y por haberme inculcado la curiosidad de saber como funciona el mundo. Tus palabras sabias me han ayudado a tener presente que soy humano y que debería de estar orgulloso de mis logros a pesar de mis fallas. Gracias por todos tus sacrificios los cuales me han forjado para ser quien soy hoy. Espero que ambos estén orgullosos de lo que he logrado.

TABLE OF CONTENTS

Chapter 1

Introduction

Over the past few decades the field of elementary particle physics has been the stage of numerous advances both in theoretical and experimental physics. Being the fundamental building blocks at the smallest distance scales that form our perception of the universe through their energetic interactions, elementary particles are inherently both quantum and relativistic objects. In order to give a mathematical description of the nature of elementary particles, one would need to reconcile the disparate theories of quantum physics and relativity. This resulted in one of the most elegant theoretical frameworks to date, known as Quantum Field Theory (QFT), which serves as the rigorous foundation of the Standard Model (SM) of particle physics. The SM describes the different symmetries and interactions that particles obey in nature, which form the basis of our understanding of modern particle physics. Like all physical theories that we use to describe different aspects of the universe we live in, the SM has provided us with several predictions that have been experimentally verified. These predictions range from simplistic facts such as the existence of particle-antiparticle pairs, to more insightful predictions such as the existence of gauge bosons, a special family of particles that are responsible for mediating the interactions between particles that form the ordinary matter we observe in nature.

Experimental physics saw rapid advancements to provide the empirical evidence that bridges QFT and the SM. Many engineering feats were made in the creation of the machin-

ery necessary to recreate the energetic conditions needed to study the simplest of particle interactions. At present, the machinery required for particle physics experiments has become sophisticated. Circular colliders, such as the Large Hadron Collider (LHC), employ a wide array of technologies that allow for the confinement and acceleration to relativistic speeds of particle beams in order to study the outcome of their energetic collisions. However, recreating the conditions to study particle interactions and colliding the particles is only part of the job, as one needs to be able to detect and identify what comes out from these interactions. Modern-day particle detection and identification has evolved from painstakingly analyzing individual photographs of tracks traced by particles in cloud chambers to multi-tiered detector systems, which are designed to detect large numbers of particles simultaneously. The technology behind modern-day detectors is designed to elicit tailored interactions between different types of particles and detector components in order to detect the particles. This makes modern-day detectors analogous to high-resolution cameras that allow us to capture the fine details of nature at the sub-atomic scale.

All this amalgamation of knowledge reached its current pinnacle with the discovery of the Higgs boson in 2012, finalizing the current formulation of the SM. However, all is not well in the SM because there are several open questions remaining, and the SM falls short in providing explanations. Some examples of these open questions are the existence of dark matter (DM), the abundance asymmetry between matter and antimatter in the universe, and the Hierarchy Problem. The Hierarchy Problem, which is related to the research topics presented in this thesis, can be stated as the relative lightness of the Higgs boson mass in comparison to the energy scales, such as the Planck mass scale, at which new physics is expected to emerge. Currently, the experiments at the LHC are providing precision measurements to further test the validity of the SM and to search for new physics beyond the

Standard Model (BSM) that could help explain some of these unanswered questions.

This thesis describes two research topics. The first topic aims to improve particle identification in the ATLAS detector using collimated sprays of particle decays in the detector, known as jets, which is an important process in nearly all precision measurements and BSM searches. The second topic is a search for Vector-Like Quarks (VLQs), which are particles predicted by some BSM theories seeking to resolve the Hierarchy Problem. The theoretical background of particle physics needed to understand and motivate these analyses is described in Chapter 2 by providing an introduction to the SM, its shortcomings, and a brief overview of the BSM VLQ theory. The LHC and the ATLAS detector, which are the experimental devices that provided the data utilized in the analyses described in the latter chapters, are presented in Chapter 3. An overview of the different detector components, the interactions between particles and the detector, and the process of reconstructing physics objects from detector data are discussed in this chapter. Chapter 4 describes the particle physics processes of interest in the studies presented in this thesis, as well as the requirements made to select events from the detector data and Monte Carlo (MC) simulations that provide a relatively pure selection of these processes. Chapter 5 presents the first research topic of this thesis through the concept of tagging a jet to a particle. Some particles, such as the top quark and the Higgs boson, decay into other particles before interacting with the detector due to their short lifetime. If these decays are fully hadronic, then the source particle can be reconstructed as a jet. Several particles that are predicted by BSM theories, such as VLQs, can decay into these particles with short lifetimes. Jet tagging can aid in the reconstruction of events where these yet undiscovered particles are produced by identifying jets with their decays. Two jet tagging studies are presented in this chapter. The first study consists of utilizing information from the substructure of jets to improve top quark and W boson jet

tagging. The second study consists of utilizing topological data analysis techniques as an alternative for top quark jet tagging. These techniques have not been used in the context of jet tagging previously. The potential improvement that these techniques bring over the traditional taggers utilized in ATLAS is assessed in this chapter. Chapter 6 presents the second research topic of this thesis through two search analyses of a vector-like top quark (T). The first search analysis focuses on the single production of a T, while the second analysis focuses on the pair production $T\bar{T}$. Both analyses target the decay channels $T \to Ht$ and $T \to Zt$ in final states associated with exactly one electron or muon. The search strategy, statistical analysis, and results of both searches are discussed in this chapter. Finally, Chapter 7 gives an overall summary and concluding remarks, as well as potential research outlooks, for both research topics presented in this thesis.

Chapter 2

Theory

In this chapter, the theoretical framework that is necessary to understand and motivate the subsequent chapters of this thesis, the Standard Model (SM) of particle physics, is presented. The chapter starts with the introduction of the different particles that form the SM and a brief overview of their properties. This is followed by an exposure to the group theory framework, which is used as the mathematical foundation of the SM, emphasizing the importance of Lie groups and their role in describing the symmetries of the SM. The following sections are dedicated to the construction of the SM Lagrangian using symmetry arguments as the initial motivation. This will be presented in parts, starting with the electroweak interactions of the SM, followed by the spontaneous symmetry breaking mechanism from which the Higgs boson originates, and culminating with the strong force interactions.

The next section of this chapter focuses on the BSM aspect of the theoretical framework. First, motivation for the necessity of extending the current SM is given through phenomenological examples that the SM is unable to explain. More emphasis will be placed on the Hierarchy Problem, which serves as the theoretical motivation for many BSM searches performed at ATLAS, such as the vector-like quark (VLQ) analyses presented in Chapter 6. Finally, a brief overview of the Composite Higgs model, which aims to solve the Hierarchy Problem, is presented. Some realizations of the Composite Higgs model predict the existence of VLQs; thus, a discovery of VLQs could serve as validation of this extension to the SM.

The description provided for the theoretical background of the SM of elementary particle physics is based on [1] with some elements of group theory from [2]. The discussion of the BSM theory of Composite Higgs models and VLQs is based on [3]. In the following discussion, all mathematical expressions that have repeating indices imply a sum of the indexed terms following Einstein's summation convention. Furthermore, the system of natural units will be used throughout the discussion and the remainder of this thesis, as it is the convention used in particle physics. The natural units are defined by setting Planck's reduced constant and the value of the speed of light in vacuum to:

$$\begin{aligned} \hbar = \frac{h}{2\pi} = 1.055 \times 10^{-34}\ \mathrm{J \cdot s} \rightarrow \hbar = 1 \\ c = 2.998 \times 10^{8}\ \mathrm{m/s} \rightarrow c = 1 \end{aligned} \tag{2.1}$$

Under this convention, quantities such as energy, momentum, and mass are measured in electronvolts (eV), which is the energy of a single electron accelerated through a potential of 1 Volt, and quantities such as distance and time are measured in eV^{-1}. Since quantities of interest in particle physics are small, it is common practice to use gigaelectronvolts ($1\ \mathrm{GeV} = 1.6 \times 10^{-10}\ \mathrm{J}$) instead. Additionally, all electric charges are expressed in terms of the fundamental charge e, which is the charge of the proton. As an example, the electron has a charge of -1, while the up quark has a charge of $2/3$. Finally, the spin angular momentum of all particles is measured in units of $\hbar$.

2.1 The Standard Model of Particle Physics

2.1.1 Particles of the Standard Model

All elementary particles that have been predicted by the SM of particle physics and experimentally observed are summarized in Figure 2.1. These particles can be classified into two groups based on their intrinsic spin quantum number: fermions, defined by having half-integer spin, and bosons, defined by having integer spin. The atoms that form the ordinary

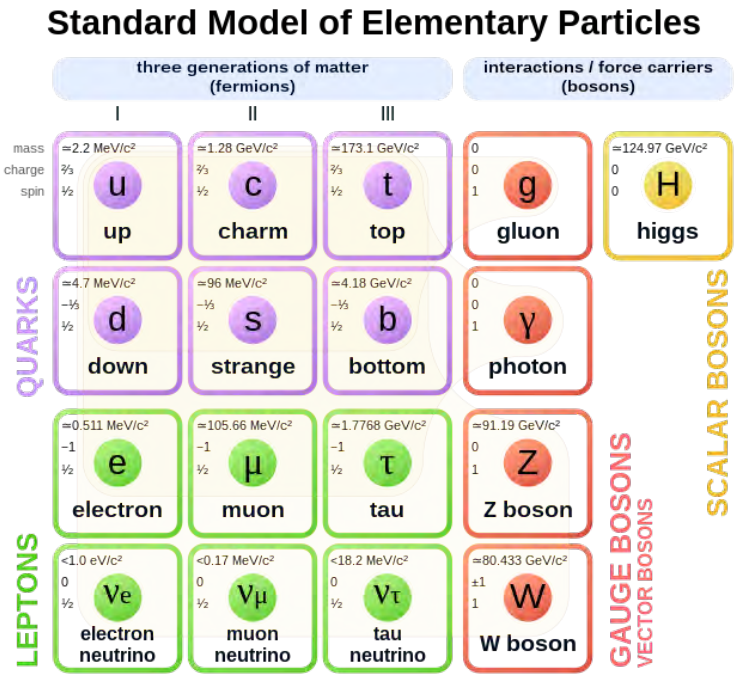


Figure 2.1: Summary of all elementary particles in the SM and their properties. This figure is taken from [4].

matter in the universe are composed of fermions that are bound together through the fundamental forces that are mediated by the gauge bosons. The gravitational force is the only exception that has evaded prediction from the SM, thus lacking a mediator particle. All fermions have an associated antiparticle that shares the same mass value but differs in quantum numbers, which dictates how these particles interact in the SM. Fermions are arranged

into three generations, which are characterized by their mass value and flavor quantum number. Each subsequent generation contains a heavier particle of a given flavor. The ordinary matter we observe in the universe is only composed of first generation fermions. Particles in the other generations are not stable enough to compose matter due to their large masses and consequently short lifetimes. Fermions are further subdivided into quarks and leptons based on the types of interactions in which these particles can participate.

The leptons consist of the electron (e), muon (μ), and tau (τ), which carry electric charge $Q = -1$ and can interact through the electromagnetic force, and their associated neutrinos, ν_e, ν_μ, and ν_τ, which are electrically neutral and thus cannot interact electromagnetically. Both charged and neutral leptons contain an additional intrinsic quantum number known as the weak isospin, which allows them to interact through the weak force. It should be noted that neutrinos are predicted to be massless in the minimal formulation of the SM, which disagrees with recent experimental evidence. The SM can be extended without significant effort to incorporate neutrino masses; however, the details of the mechanism required to do so cannot be explained solely by the SM.

The quarks consist of the up (u), down (d), charm (c), strange (s), top (t), and bottom (b). All quarks carry electric charge, $Q = 2/3$ in the case of u, c, and t, and $Q = -1/3$ in the case of d, s, and b, which allows them to interact electromagnetically. Like leptons, quarks also carry weak isospin allowing them to interact through the weak force. What sets quarks apart from leptons is that quarks carry an additional quantum number known as color charge, which allows them to participate in strong force interactions. Unlike electric charge, which is characterized by a single value that can be either positive or negative, color is characterized by three values that are labeled as red, green, and blue, along with their corresponding anti-values: anti-red, anti-green, and anti-blue. All quarks carry a single unit

of color. Due to a phenomenon known as color confinement, no single quark can constitute ordinary matter; only bound states of multiple quarks with certain combinations of color can constitute ordinary matter. One combination, known as baryons, consists of having a quark of each color charge, or anti-color in the case of an antiparticle, in equal parts. An example of a baryon is the proton, which distributes a single unit of red, green, and blue charge across its quark constituents. The other possible combination, known as mesons, consists of arrangements that contain at least one color anti-color pair.

The interactions between particles can be interpreted as the interacting particles exchanging the corresponding mediator gauge boson of an interaction force. The photon (γ), which is massless and electrically neutral, mediates the electromagnetic force between particles that carry electric charge. The gluon (g), which is massless, charge-neutral, and carries a unit of color and anti-color charge, mediates the strong interaction between particles that carry color charge. The weak vector bosons $W^{\pm}$ and Z mediate the weak interaction between particles that carry weak isospin. The Z is electrically neutral, while the W^{+} and W^{-} carry electric charges $Q = 1$ and $Q = -1$, respectively, allowing them to interact electromagnetically.

Finally, to complete the overview of particles in the SM, there is a single scalar boson, which is the Higgs boson. The Higgs boson is both charge and color neutral; however it has a weak isospin of $-1/2$. Unlike the gauge bosons in the SM, the Higgs boson does not mediate a force from nature but plays an important role in the Higgs mechanism, which is the process through which the weak vector bosons acquire their mass.

2.1.2 Symmetries and Lie Groups

The mathematical formulation of the SM is based on the Lagrangian formulation of QFT, which is used to describe the different interactions between particles as discussed in the

following sections. This formulation is attractive in particle physics for several reasons. First, the Lagrangian is a scalar function, which implies that it must remain invariant under transformations such as Lorentz transformations. This means that the description of physical processes provided by the Lagrangian must be independent of the frame of reference of an observer. This is a desired property when describing the interactions of elementary particles, which are usually relativistic in nature. Another benefit of the Lagrangian formulation is its ability to encode conservation laws through symmetries in the Lagrangian. Conservation of momentum and energy manifests when the Lagrangian is invariant under the Lorentz transformations. Similarly, the conservation of quantum numbers such as electric charge, weak isospin, and color charge is manifested when the Lagrangian is invariant under gauge transformations, which alter the internal properties of particles. This is formally stated in Noether's Theorem, which states that any continuous local transformation that leaves invariant the action of a Lagrangian is associated with a conserved quantity.

These continuous transformations are described with Lie groups, of which the most relevant in particle physics are the Poincaré group, the unitary group $U(1)$, and the special unitary groups $SU(2)$ and $SU(3)$. The Poincaré group represents the symmetries of the Lorentz transformations. The unitary group $U(1)$ represents the symmetries associated with the choice of the electromagnetic potential. The special unitary groups $SU(2)$ and $SU(3)$ represent the rotation symmetries on the internal weak isospin and color charge spaces, respectively. In the following discussion, more attention will be given to the groups that describe the conservation of quantum numbers of a particle since they are more essential in the construction of the SM Lagrangian. It should be noted that there are multiple representations of Lie groups; however, the properties and results discussed in this section are independent of the representation used. The matrix representation of Lie groups will be used

in order to maintain consistency with the SM Lagrangian formulation. Formally, a matrix Lie group G is a group whose matrix elements $M(\theta)$ depend continuously on a set of real parameters $\theta \in \mathbb{R}^m$ [5]. This dependence on continuous parameters endows the Lie group G with the additional structure of a topological manifold. Matrix multiplication can be viewed as a continuous function $f : G \times G \to G$ from the product manifold $G \times G$ to the manifold G such that $M(\theta) = M(\theta_1)M(\theta_2)$ if $\theta = f(\theta_1, \theta_2)$. The choice of parameters is made so that the identity matrix I coincides with $\theta = 0$. For each matrix Lie group G, there is an associated Lie algebra $\mathfrak{g} = \text{Span}\{v_1, \cdots, v_n\}$, which is a vector space spanned by the matrices in the tangent space to the identity matrix of G when viewed as a topological manifold. The Lie algebra is equipped with an additional operation known as the Lie bracket, which is analogous to a commutator of its elements in the case of matrix Lie groups:

$$\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g} : [v_i, v_j] = v_i v_j - v_j v_i \tag{2.2}$$

The basis matrices v_i of the tangent space are known as the generators of the Lie algebra. A Lie group is associated to its Lie algebra through the exponential map:

$$\exp : \mathfrak{g} \to G : M(\theta) = e^{\sum_i \theta_i v_i} = \lim_{N \to \infty} \left(\mathrm{I} + \frac{1}{N} \sum_i \theta_i v_i \right)^N \tag{2.3}$$

The exponential map is only valid for matrices that are path-connected to the identity matrix of G. For this reason, the special unitary groups cannot be generalized to the unitary groups $U(2)$ and $U(3)$ since unitary matrices are characterized by having their determinant equal to ± 1. This essentially splits the manifold structure of the groups $U(2)$ and $U(3)$ into two path-connected components based on the sign of the determinant. The matrices with a

negative determinant act as reflections in the internal spaces of the weak isospin and color charge; thus, they would include terms in the Lagrangian that do not conserve these quantum numbers. The structure of a Lie group G can be fully described near its identity matrix with the generators of the Lie algebra and the structure constants that are obtained from the Lie bracket.

A particular representation of the group $U(1)$ is given by the set of 2×2 matrices with determinant 1 over the real numbers that depend on a single real parameter θ:

$$M(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad \theta \in \mathbb{R} \tag{2.4}$$

From this representation, it is understood that the group structure satisfies:

$$M(\theta) = M(\theta_1)M(\theta_2) = M(\theta_1 + \theta_2) \tag{2.5}$$

and is subject to the periodicity condition $\theta + 2\pi = \theta$. This implies that the manifold structure of this group is the unit circle, which is a compact and connected space as shown in Figure 2.2. By differentiating Equation 2.5 with respect to θ_1 and applying the chain rule, the following expression is obtained:

$$\frac{dM(\theta)}{d\theta_1} = \frac{dM(\theta)}{d\theta}\frac{d(\theta_1 + \theta_2)}{d\theta_1} = \frac{dM(\theta)}{d\theta} = \frac{dM(\theta_1)}{d\theta_1}M(\theta_2) \tag{2.6}$$

Evaluating this expression at $\theta = 0$, near the identity element, the matrices in the tangent

space satisfy:

$$\left.\frac{dM(\theta)}{d\theta}\right|_{\theta=0} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} M(\theta_2 = 0) = \mathbb{J}\mathrm{I} = \mathbb{J} \tag{2.7}$$

which yields the following relation:

$$\left.\frac{d^n M(\theta)}{d\theta^n}\right|_{\theta=0} = \mathbb{J}^n \tag{2.8}$$

By expanding an arbitrary element of $U(1)$ in this representation into a power series, the following expression is obtained:

$$M(\theta) = \sum_{n=0}^{\infty} \frac{1}{n!} \left.\frac{d^n M(\theta)}{d\theta^n}\right|_{\theta=0} \theta^n = e^{\theta\mathbb{J}} \tag{2.9}$$

which is the exponential map that maps the Lie algebra $\mathfrak{u}(1)$ to $U(1)$. From this expression, $\mathfrak{u}(1)$ is identified as the one-dimensional vector space generated by the matrix $\mathbb{J}$, which is isomorphic to the line $i\mathbb{R}$. Although the actual tangent line in the manifold structure of $U(1)$ is $1 + i\mathbb{R}$, this line can be parametrized by the vector space $i\mathbb{R}$ by treating the point of tangency as the origin of $\mathfrak{u}(1)$. Under this isomorphism, $U(1)$ is now represented by complex numbers with modulus 1 under multiplication. The exponential map takes the familiar form of Euler's identity, which maps the line $1 + i\mathbb{R}$ to the unit circle in the complex plane.

The group $SU(2)$ can be represented as the set of all 2×2 unitary matrices with determinant 1 that have the form:

$$M = \begin{pmatrix} a + ib & -c + id \\ c + id & a - ib \end{pmatrix} \quad a, b, c, d \in \mathbb{R} \tag{2.10}$$

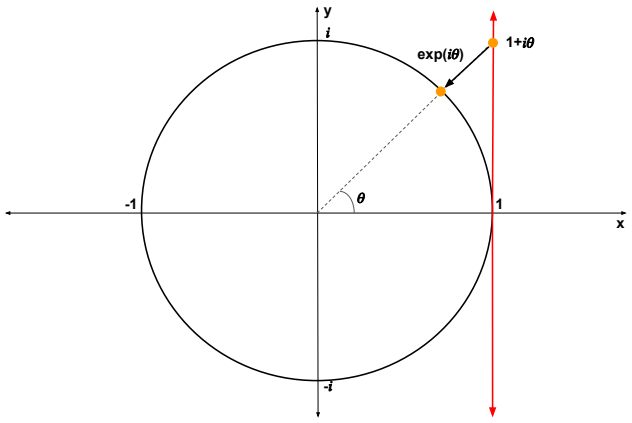


Figure 2.2: Topological representation of the group $U(1)$. Each element of this group represents a point on the unit circle in the complex plane. The red line is the tangent line to the identity element of this group, which corresponds to the associated Lie algebra $\mathfrak{u}(1)$. Any point on this line can be mapped to the unit circle via the exponential map. Figure adapted from [6].

The condition on the determinant $a^2 + b^2 + c^2 + d^2 = 1$ implies that this group has the manifold structure of the boundary of a 4-dimensional sphere of radius 1. In this case, it is more instructive to show how the Lie algebra $\mathfrak{su}(2)$ is constructed based on the properties of the matrices in $SU(2)$. Using the fact that any complex matrix σ satisfies the identity

$$\det e^{\sigma} = e^{\mathrm{Tr}(\sigma)} \tag{2.11}$$

it follows from the exponential map that if $M \in SU(2)$ and $\sigma \in \mathfrak{su}(2)$, then the Lie algebra $\mathfrak{su}(2)$ consists of all 2×2 traceless complex matrices, which can be spanned by the set of Pauli matrices:

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{2.12}$$

Unlike $U(1)$, the group $SU(2)$ is a non-abelian group, which is what originates the self-interaction terms of the electroweak bosons in the SM Lagrangian.

Finally, the group $SU(3)$ is similar to $SU(2)$ but with 3×3 matrices instead. The non-

abelian nature of $SU(3)$ gives rise to the gluon self-interaction terms in the SM Lagrangian. The procedure to obtain the associated Lie algebra $\mathfrak{su}(3)$ is similar to the one used for $\mathfrak{su}(2)$, with the spanning set being the Gell-Mann λ matrices:

$$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \qquad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix} \qquad \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{2.13}$$

$$\lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \qquad \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}$$

2.1.3 Quantum Electrodynamics

To start the description of the SM Lagrangian, we begin with Quantum Electrodynamics (QED), which describes the electromagnetic interactions between electrically charged fermions. All fermions that are not subject to an external potential can be described with Dirac's equation

$$i\gamma^\mu \partial_\mu \psi - m\psi = 0 \tag{2.14}$$

where $\psi(x)$ is a Dirac spinor representing the wave function of the fermion field, m is the mass of the fermion, and the γ^μ are the 4×4 matrices

$$\gamma^0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \quad \gamma^i = \begin{pmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{pmatrix} \tag{2.15}$$

where σ_i are the Pauli matrices in Equation 2.12 for $i = 1, 2, 3$. Dirac's equation can be obtained as the equation of motion of the Lagrangian

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi \tag{2.16}$$

From this Lagrangian it can already be seen that it is invariant under Lorentz transformation, which is one of the desired symmetries in the SM. Furthermore, upon closer inspection, this Lagrangian is also invariant under the $U(1)$ global gauge transformation

$$\psi \to \psi' = \psi e^{i\theta} \tag{2.17}$$

Although in its global form, it is not yet clear that this gauge transformation is associated with the choice electromagnetic potential and the conservation of electric charge. However, as stated in subsection 2.1.2, in order to have a conserved quantity, there must be an associated local gauge transformation. Promoting Equation 2.17 to a local gauge transformation by introducing a space-time dependence on the phase $\theta(x) = q\beta(x)$

$$\psi \to \psi' = \psi e^{iq\beta(x)} \tag{2.18}$$

we see that the Lagrangian is no longer invariant under this transformation. The transformed Lagrangian $\mathcal{L}^{'}$ takes the form

$$\mathcal{L}^{'} = \mathcal{L} - q\bar{\psi}(\partial_{\mu}\beta(x))\psi \tag{2.19}$$

This is remedied using the minimal coupling rule, which introduces a minimal number of new fields to the Lagrangian so that it remains locally invariant under the gauge transformation. For electromagnetic interactions, it is only necessary to introduce a single vector field $A_{\mu}(x)$ such that it transforms under the local gauge transformation as:

$$A_{\mu}(x) \rightarrow A^{'}_{\mu}(x) = A_{\mu}(x) - \partial_{\mu}\beta(x) \tag{2.20}$$

and to redefine the covariant derivative term as:

$$\partial_{\mu} \rightarrow \partial_{\mu} + iqA_{\mu}(x) \tag{2.21}$$

With these minimal coupling changes, the Lagrangian takes the form:

$$\mathcal{L} = \bar{\psi}(i\gamma^{\mu}\partial_{\mu} - m)\psi - q\bar{\psi}\gamma^{\mu}\psi A_{\mu} \tag{2.22}$$

making it invariant under the local $U(1)$ transformations described in Equation 2.18 and Equation 2.20. It should be noted that the local $U(1)$ transformation also allows to accommodate a kinetic term for the field A_{μ} of the form:

$$\mathcal{L}_{\text{kin.}} = -\frac{1}{4}A^{\mu\nu}A_{\mu\nu}, \quad A^{\mu\nu} = \partial^{\mu}A^{\nu} - \partial^{\nu}A^{\mu} \tag{2.23}$$

However, A_μ must be massless since introducing a mass term that is proportional to $A_\mu A^\mu$ breaks the local $U(1)$ invariance. Based from this description, the field A_μ represents the photon in the SM. Thus, the Lagrangian

$$\mathcal{L}_{\text{QED}} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi - q\bar{\psi}\gamma^\mu \psi A_\mu - \frac{1}{4}A^{\mu\nu}A_{\mu\nu} \tag{2.24}$$

is the quantum description of electromagnetic interactions in the SM. The second term in $\mathcal{L}_{\text{QED}}$ indicates that the fermion fields couple to the photon field with the coupling strength being proportional to the electric charge q of the fermion. This is represented in the Feynman diagram shown in Figure 2.3.

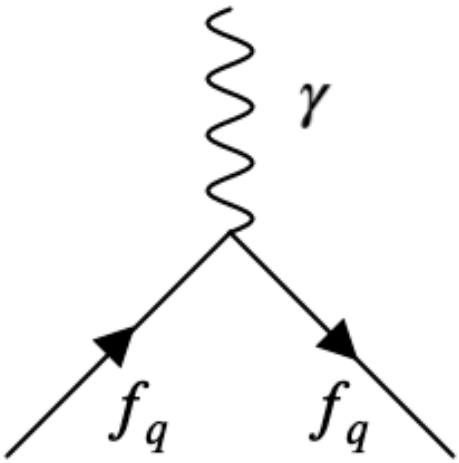


Figure 2.3: Vertex interaction from the QED Lagrangian, which can be combined to described processes like particle-antiparticle annihiliation or scattering.

2.1.4 The Weak Force and Electroweak Unification

As mentioned in subsection 2.1.1, the weak force is mediated by three gauge bosons: the electrically neutral Z boson and the two electrically charged W^+ and W^- bosons. Since

the Z boson is electrically neutral, all interactions mediated by it are similar to the electromagnetic interactions that are mediated by the photon. Since electric charge is a conserved quantity, all SM weak interactions that are mediated by the $W^{\pm}$ result in the interacting fermion changing to the other fermion from the same generation, a process known as flavor-changing current. For example, an up quark interacting with a W^{-} can turn into a down quark, while an electron interacting with a W^{+} can turn into an electron neutrino. If each generation of fermions is thought of as a two-dimensional space with each axis corresponding to a fermion flavor, then the W boson acts as a rotation operator in this space.

Experimentally, it is known that the SM weak interactions violate the discrete symmetry of parity [7]. If a given particle interaction is physically valid, then the parity symmetry dictates that reversing the interaction is also a physically valid process. This violation is manifested in the chirality of neutrinos. The chirality of a particle is a property that measures the orientation of the spin of a particle relative to its momentum. Particles that have their spin parallel to their momentum are referred to as right-handed, while those that are anti-parallel are referred to as left-handed. Neutrinos are restricted to being left-handed, while anti-neutrinos are restricted to being right-handed, as a consequence of the SM predicting the neutrinos as massless particles. Thus, the W boson only interacts with left-handed particles as a consequence of this. The Lagrangian that describes the weak interactions must take into account the chirality distinction of particles.

Instead of only describing the weak interaction Lagrangian, it will be more instructive to consider the combination of the electromagnetic force and the weak force in a single force, known as the electroweak (EWK) force. This is in part motivated by the fact that the Z boson behaves like a photon with mass. With all these considerations, we can split the SM

Lagrangian into two sectors: a left-chirality sector and a right-chirality sector.

$$\mathcal{L} = \mathcal{L}_L + \mathcal{L}_R \tag{2.25}$$

$\mathcal{L}_L$ will contain the interactions between left-handed fermions that are mediated by the photon and all three weak gauge bosons, while $\mathcal{L}_R$ will contain the interactions between right-handed fermions that are mediated by the photon and the Z boson. Both $\mathcal{L}_L$ and $\mathcal{L}_R$ are given by the free fermion Lagrangian in Equation 2.16, with the field ψ_L in $\mathcal{L}_L$ representing the left-handed isospin doublets:

$$\begin{pmatrix} u \\ d \end{pmatrix}_L, \begin{pmatrix} c \\ s \end{pmatrix}_L, \begin{pmatrix} t \\ b \end{pmatrix}_L, \begin{pmatrix} e \\ \nu_e \end{pmatrix}_L, \begin{pmatrix} \mu \\ \nu_\mu \end{pmatrix}_L, \begin{pmatrix} \tau \\ \nu_\tau \end{pmatrix}_L \tag{2.26}$$

while the field ψ_R in $\mathcal{L}_R$ represents the right-handed isospin singlets:

$$u_R,\, b_R,\, c_R,\, s_R,\, t_R,\, b_R,\, e_R,\, \mu_R,\, \tau_R \tag{2.27}$$

Both sectors transform identically under the global $U(1)_Q$ transformations, while the left-chirality sector contains the additional global symmetry of $SU(2)_{I_3}$ transformation invariance. The labels Q and I_3 represent the charge and weak isospin, which are to be conserved when these transformations are promoted to local transformations. The combination of the electromagnetic force and weak force symmetries can be represented with the cartesian product of these two groups, $SU(2) \times U(1)_Y$, where the quantity $Y = 2(Q - I_3)$, known as the weak hypercharge, is to be conserved under EWK interactions. Following a similar argument

as used in subsection 2.1.3, we let the left-handed isospin doublets transform as:

$$\psi_L \rightarrow \psi_L^{'} = e^{ig_1 Y\beta(x)} e^{ig_2 \alpha^a(x)\sigma_a} \psi_L \tag{2.28}$$

where σ_a are the Pauli matrices in Equation 2.12 that generate the $SU(2)$ rotations, while the right-handed isospin singlet transform as:

$$\psi_R \rightarrow \psi_R^{'} = e^{ig_1 Y\beta(x)} \psi_R \tag{2.29}$$

The Lagrangian is again no longer invariant under these transformations, taking the form:

$$\mathcal{L}^{'} = \mathcal{L}_L - \bar{\psi}_L(g_1 Y\gamma^\mu \partial_\mu \beta(x) + g_2 \gamma^\mu \partial_\mu \alpha^a(x)\sigma_a)\psi_L + \mathcal{L}_R - g_1 Y \bar{\psi}_R \gamma^\mu \partial_\mu \beta(x)\psi_R \tag{2.30}$$

Invoking the minimal coupling rule, four vector fields, B_μ and W_μ^a, are introduced such that they transform under the local $SU(2) \times U(1)_Y$ transformation as:

$$\begin{aligned} B_\mu(x) &\rightarrow B_\mu^{'}(x) = B_\mu(x) - \partial_\mu \beta(x) \\ W_\mu^a(x) &\rightarrow W_\mu^{'a}(x) = W_\mu^a(x) - \partial_\mu \alpha^a(x) \end{aligned} \tag{2.31}$$

and the left-chirality and right-chirality covariant derivative terms transform as:

$$\partial_\mu \rightarrow \partial_\mu + ig_1 Y B_\mu + ig_2 W_\mu^a \sigma_a \tag{2.32}$$

$$\partial_\mu \rightarrow \partial_\mu + ig_1 Y B_\mu \tag{2.33}$$

With these minimal coupling changes the Lagrangian describing EWK interactions takes the

form:

$$\begin{aligned}
\mathcal{L}_{\text{EWK}} &= \mathcal{L}_R + \mathcal{L}_R + \mathcal{L}_{\text{kin.}} \\
\mathcal{L}_L &= \bar{\psi}_L(i\gamma^\mu\partial_\mu - m)\psi_L - g_1 Y \bar{\psi}_L\gamma^\mu B_\mu\psi_L - g_2\bar{\psi}_L\gamma^\mu W^a_\mu\sigma_a\psi_L \\
\mathcal{L}_R &= \bar{\psi}_R(i\gamma^\mu\partial_\mu - m)\psi_R - g_1 Y \bar{\psi}_R\gamma^\mu B_\mu\psi_R \\
\mathcal{L}_{\text{kin.}} &= -\frac{1}{4}B^{\mu\nu}B_{\mu\nu} - \frac{1}{4}W^i_{\mu\nu}W_i^{\mu\nu}
\end{aligned} \tag{2.34}$$

where the fields B and W^a couple to the fermion fields with strengths g_1 and g_2, respectively. The kinetic term $B^{\mu\nu}$ is analogous to the photon kinetic term introduced in the previous section, while the kinetic term $W^i_{\mu\nu}$ is given by:

$$W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu + g_2\epsilon^{ijk}W^j_\mu W^k_\nu \tag{2.35}$$

The last term in this expression arises from the non-abelian nature of $SU(2)$ and is responsible for the self-interaction terms of the vector boson fields. Introducing mass terms for these fields in Equation 2.34 breaks the local $SU(2) \times U(1)_Y$ gauge invariance, which is a glaring issue since the weak vector bosons are experimentally known to have mass. This issue can be remedied with a spontaneous symmetry breaking process known as the Higgs mechanism, which will be discussed in the next section. Additionally, based on the shapes of the Pauli matrices, the fields W^1 and W^2 will mix together to form the $W^\pm$ boson fields, while the B and W^3 fields will mix together to form the Z boson and the photon fields. The Feynman diagrams of the weak interactions introduced by $\mathcal{L}_{\text{EWK}}$ are shown in Figure 2.4.

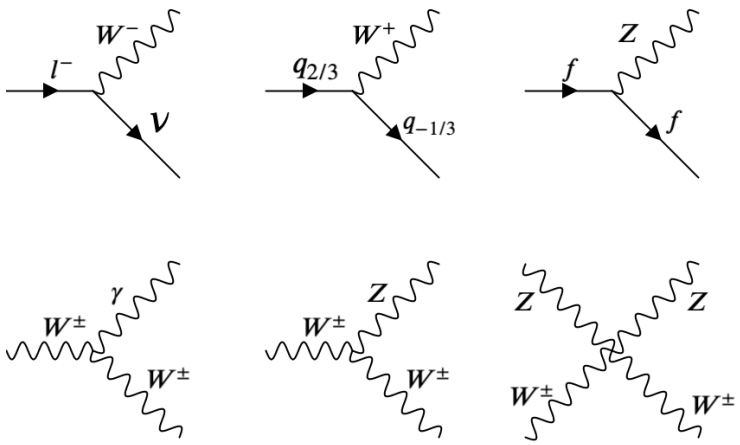


Figure 2.4: Vertex interaction from the weak force in the EWK Lagrangian.

2.1.5 Higgs Mechanism

As discussed in the previous section, four massless vector bosons arise when the Lagrangian describing the EWK interactions of the SM particles is required to be invariant under local $SU(2) \times U(1)_Y$ transformations. From experimental evidence, it is known that three of these vector bosons have non-zero mass. The Higgs mechanism enables the $W^{\pm}$ and Z bosons to acquire their mass in the SM through a spontaneous breaking of the EWK symmetry. This is achieved by introducing a spin-zero field, known as the Higgs field, which has a non-zero vacuum expectation value (VEV) with non-zero $SU(2) \times U(1)_Y$ quantum numbers. This process will leave the Lagrangian invariant under local $SU(2) \times U(1)_Y$ transformations; however, the ground state of the system will no longer be invariant due to it having a non-zero Y quantum number. A way to achieve the spontaneous symmetry breaking is to allow the Higgs field to transform as an $SU(2)$ doublet:

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1(x) + i\phi_2(x) \\ \phi_3(x) + i\phi_4(x) \end{pmatrix} \tag{2.36}$$

where the fields ϕ_{1-4} are real-valued scalar fields. Furthermore, the Higgs field is allowed to interact with itself through the Lagrangian

$$\mathcal{L}_\phi = (\partial_\mu \bar{\phi})(\partial^\mu \phi) - \mu^2 \bar{\phi}\phi - \lambda(\bar{\phi}\phi)^2 \tag{2.37}$$

where μ and λ are the parameters that govern the self-interaction potential of the Higgs field. For now, it suffices to assume that $\lambda > 0$ in order to have the potential bounded from below. To find the VEV of the Higgs field we need to minimize its potential:

$$\frac{\partial V(\phi)}{\partial \phi} = 0 \implies \bar{\phi}(\mu^2 + 2\lambda(\bar{\phi}\phi)) = 0 \tag{2.38}$$

If $\mu^2 > 0$, then the potential has its minimum when $\phi(x) = 0$, meaning there is no Higgs field and thus the vector bosons remain massless. The interesting case is when $\mu^2 < 0$, in which we obtain a non-trivial solution

$$\bar{\phi}\phi = \frac{\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2}{2} = -\frac{\mu^2}{2\lambda} = \frac{v^2}{2} \tag{2.39}$$

where choosing a particular set of values for the fields ϕ_{1-4} will spontaneously break the $SU(2)$ symmetry of the vacuum, as shown in Figure 2.5. An appropriate choice so that the vector bosons $W^\pm$ and Z acquire mass is to set $\phi_3(x) = v$ and the other fields to zero. Thus, with this choice of field values, the Higgs field at the minimum of the potential becomes

$$\phi_0(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \\ v \end{pmatrix} \tag{2.40}$$

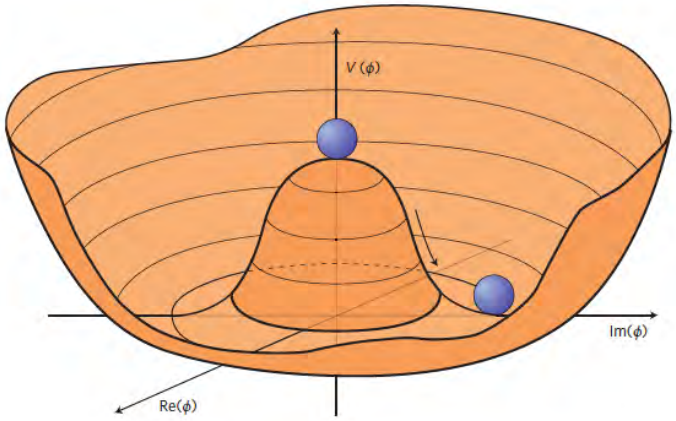


Figure 2.5: Graphical representation of the Higgs interaction potential with $\mu^2 < 0$ and $\lambda > 0$. The EWK symmetry is spontaneously broken when a particular value of the field components ϕ_i is chosen to represent the VEV. This figure is taken from [8].

Since we are interested in breaking the $SU(2)$ component of the $SU(2) \times U(1)_Y$ symmetry, this choice of the field at the minimum corresponds to the Higgs field being electrically neutral, so that charge is conserved at the ground state. This choice also determines the remaining quantum numbers of the field. Since only the down component of the doublet is non-zero, then the weak isospin must be $I_3 = -1/2$, and through $Y = 2(Q - I_3)$, the weak hypercharge is determined to be $Y_h = 1$. To see the Higgs mechanism operate, we consider perturbations around the minimum of the potential of the form

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \\ v + h(x) \end{pmatrix} \tag{2.41}$$

where $h(x)$ is the actual Higgs field. To make the Lagrangian in Equation 2.37 locally invariant under the $SU(2) \times U(1)_Y$ transformation, we set the covariant term to the one obtained in Equation 2.32. Thus, the Lagrangian now takes the form

$$\mathcal{L}_\phi = -\frac{\mu^2 h^2}{2} + \frac{1}{2} \begin{pmatrix} 0 & v \end{pmatrix} (g_1 Y_h B_\mu + g_2 \sigma_a W_\mu^a)(g_1 Y_h B^\mu + g_2 \sigma^a W_\mu^a) \begin{pmatrix} 0 \\ \\ v \end{pmatrix} + \mathcal{O} \tag{2.42}$$

where only the relevant terms for the Higgs mechanism are shown. This can be further simplified by setting $Y_h = 1$ for the Higgs field and writing down the Pauli matrices explicitly:

$$\mathcal{L}_\phi = -\frac{\mu^2 h^2}{2} + \frac{1}{2} v^2 g_2^2 ((W_\mu^1)^2 + (W_\mu^2)^2) + \frac{1}{2} v^2 (g_1 B_\mu - g_2 W_\mu^3)^2 + \mathcal{O} \tag{2.43}$$

From here we can see that the Higgs field has a mass term $m_h^2 = |\mu|^2$. To make the mass terms of the weak vector bosons apparent, we need to change from the W_μ^a basis to the

electric charge basis as follows:

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(-W_\mu^1 \pm iW_\mu^2) \tag{2.44}$$
$$W_\mu^0 = W_\mu^3$$

Under this basis, the Z boson and the photon are represented as

$$Z_\mu = \frac{1}{\sqrt{g_1^2 + g_2^2}}(g_2 W_\mu^3 - g_1 B_\mu) \tag{2.45}$$
$$A_\mu = \frac{1}{\sqrt{g_1^2 + g_2^2}}(g_1 W_\mu^3 + g_2 B_\mu)$$

Thus, the Lagrangian in Equation 2.43 becomes

$$\mathcal{L}_\phi = -\frac{\mu^2 h^2}{2} + \frac{1}{2}v^2 g_2^2 W_\mu^+ W^{-\mu} + \frac{1}{2}v^2\sqrt{g_1^2 + g_2^2} Z_\mu Z^\mu + \mathcal{O} \tag{2.46}$$

By identifying the mass terms

$$m_{W^\pm} = \frac{v g_2}{2}$$
$$m_Z = \frac{m_{W^\pm}}{\cos\theta_W}, \quad \tan\theta_W = \frac{g_1}{g_2} \tag{2.47}$$
$$m_A = 0$$

where θ_W is known as the weak mixing angle, we see that the Higgs mechanism has reconciled the SM theory with experimental observations by providing the mass terms to the weak vector bosons while still requiring the photon to be massless.

The main purpose of the Higgs mechanism was to incorporate the masses of the weak

vector bosons into the SM; however, for fermions a similar argument can be made to explain why not all fermions are massless. To achieve this, we consider the following $SU(2)$ invariant interaction between fermions and the Higgs doublet, which is added to the Lagrangian in Equation 2.43

$$\mathcal{L}_{\text{int}} = g_f(\bar{\psi}_L \phi \psi_R + \bar{\phi}\bar{\psi}_R \psi_L) \tag{2.48}$$

where ψ_L and ψ_R are the left-handed doublet and right-handed singlet of a fermion that couples to the Higgs field with strength g_f. By perturbing the VEV of the Higgs doublet, the interaction Lagrangian becomes

$$\begin{aligned}\mathcal{L}_{\text{int}} &= \frac{g_f}{\sqrt{2}}\left(\begin{pmatrix}\bar{\psi}_L^1 & \bar{\psi}_L^2\end{pmatrix}\begin{pmatrix}0\\ v+h\end{pmatrix}\psi_R + \begin{pmatrix}0 & v+h\end{pmatrix}\bar{\psi}_R\begin{pmatrix}\psi_L^1\\ \psi_L^2\end{pmatrix}\right) \\ &= \frac{g_f v}{\sqrt{2}}(\bar{\psi}_L^2\psi_R + \bar{\psi}_R\psi_L^2) + \frac{g_f}{\sqrt{2}}(\bar{\psi}_L^2\psi_R + \bar{\psi}_R\psi_L^2)h\end{aligned} \tag{2.49}$$

where the fermion left-handed doublet has been decomposed into its up (ψ_L^1) and down (ψ_L^2) components. From this equation it can be observed that the fermions acquire a mass term $m_f = g_f v/\sqrt{2}$, which is directly proportional to the Higgs coupling. The new interactions included in the SM with the introduction of the Higgs mechanism are shown in Figure 2.6.

2.1.6 Quantum Chromodynamics

The final important piece of the SM Lagrangian to be introduced is Quantum Chromodynamics (QCD), which describes the interactions between particles that carry color charge. Since there are three color charges, the QCD Lagrangian will be required to have local $SU(3)$ symmetry, which will give rise to the gluons of the SM. Using a similar argument as it was done for the EWK Lagrangian, we start with the free fermion Lagrangian in Equation 2.16,

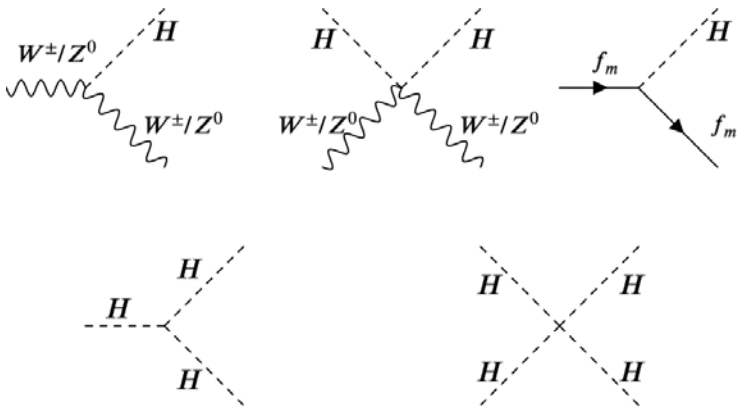


Figure 2.6: Vertices corresponding to the interactions between the Higgs boson and the weak vector bosons $W^{\pm}$ and Z, fermions with non-zero mass f_m, and the self-interactions of the Higgs boson.

where the field ψ now represents the quark $SU(3)$ triplet in color space. We now require the Lagrangian to be invariant under the transformation

$$\psi \rightarrow \psi^{'} = e^{ig_3\alpha^a(x)\lambda_a}\psi \tag{2.50}$$

where λ_a are the eight Gell-Mann matrices introduced in Equation 2.13. To make the QCD Lagrangian invariant under this local transformation, we require that the covariant derivative transforms as

$$\partial_\mu \rightarrow \partial_\mu + ig_3 G^a_\mu \lambda_a \tag{2.51}$$

where eight new vector fields G^a_μ have been introduced through the minimal coupling rule. These fields correspond to the gluon in the SM and must transform as

$$G^a_\mu(x) \rightarrow G^a_\mu(x) - \partial_\mu\alpha^a(x) - f_{abc}G^c_\mu(x)\alpha^b(x) \tag{2.52}$$

where the f_{abc} terms are the structure constants obtained from the Lie brackets of $\mathfrak{su}(3)$ that give rise to the gluon self-interaction terms. With these transformation rules, the part

of the QCD Lagrangian that describes the strong force interaction between quarks is given by

$$\mathcal{L}_{\text{QCD}} = ig_3\bar{\psi}G^a_\mu\lambda_a\psi \tag{2.53}$$

The physical explanation as to why there are eight fields associated with the gluon instead of one is that when gluons mediate the strong force between quarks, in order to conserve color charge, each gluon must carry a unit of color and another unit of anti-color. Naively, one would assume that there would be nine gluon fields since there are three different color charges. To understand why this is not the case, we can represent each single color state as the three axes of the internal color charge space

$$r = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \; b = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \; g = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{2.54}$$

Since each new vector boson corresponds to a generator of the $\mathfrak{su}(3)$ Lie algebra, we can represent each of the Gell-Mann matrix generators as the following exterior products:

$$\begin{aligned}
(r\bar{b} + b\bar{r})/\sqrt{2} &\to \lambda_1/\sqrt{2} & -i(r\bar{b} - b\bar{r})/\sqrt{2} &\to \lambda_2/\sqrt{2} \\
(r\bar{r} - b\bar{b})/\sqrt{2} &\to \lambda_3/\sqrt{2} & (r\bar{g} + g\bar{r})/\sqrt{2} &\to \lambda_4/\sqrt{2} \\
-i(r\bar{g} - g\bar{r})/\sqrt{2} &\to \lambda_5/\sqrt{2} & (b\bar{g} + g\bar{b})/\sqrt{2} &\to \lambda_6/\sqrt{2} \\
-i(b\bar{g} - g\bar{b})/\sqrt{2} &\to \lambda_7/\sqrt{2} & (r\bar{r} + b\bar{b} - 2g\bar{g})/\sqrt{6} &\to \lambda_8/\sqrt{2}
\end{aligned} \tag{2.55}$$

Any other combination of exterior products will result in a linear combination of the existing generators or in a matrix that is not in the Lie algebra $\mathfrak{su}(3)$ since it will not be traceless,

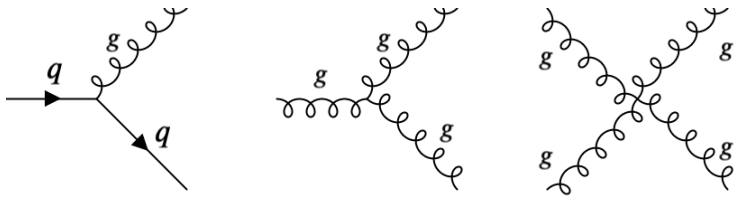


Figure 2.7: Vertices corresponding to the strong interactions between quarks q and self-interaction terms between gluons g.

like the case of the color singlet $(r\bar{r} + b\bar{b} + g\bar{g})/\sqrt{3}$. The new interactions included in the SM through the QCD Lagrangian are shown in Figure 2.7

2.1.7 The Standard Model Lagrangian

After going through the individual components of the Lagrangian that repesent the electroweak and strong interactions and having a mechanism that spontaneously breaks the EWK symmetry, which explains how the weak vector bosons acquire their mass, we are now in a position to give a full description of the most important terms of the SM Lagrangian.

The SM Lagrangian can be summarized as:

$$
\begin{aligned}
\mathcal{L} = & \; i \sum_f \bar{\psi}_f \gamma^\mu \partial_\mu \psi_f && \text{fermion kinetic terms} \\
& + \sum_f v g_f (\bar{\psi}^f_R \psi^f_L + \bar{\psi}^f_L \psi^f_R)/\sqrt{2} && \text{fermion mass terms} \\
& + \sum_f \bar{\psi}_f (i g_1 Y \gamma^\mu B_\mu + i g_2 \gamma^\mu W^a_\mu \sigma_a) \psi_f && \text{fermion-EWK boson interaction terms} \\
& + \sum_q \bar{q} (i g_3 G^a_\mu \lambda_a) q && \text{quark-gluon interaction terms} \\
& + (\partial^\mu \bar{\phi})(\partial_\mu \phi) && \text{Higgs kinetic term} \\
& - (\mu^2 \bar{\phi}\phi + \lambda(\bar{\phi}\phi)^2) && \text{Higgs potential term} \\
& + \bar{\phi}(i g_1 Y B_\mu + i g_2 W^a_\mu \sigma_a)\phi + \mathcal{O}
\end{aligned}
\tag{2.56}
$$

where $\mathcal{O}$ includes additional kinetic terms of the vector bosons, chirality terms and higher order terms.

2.2 Shortcomings of the Standard Model

The SM has proven to be a successful theory that describes the interactions of elementary particles and has gone through a battery of experimental tests to validate its predictions. However, there are certain phenomena that we observe in the universe for which the SM is unable to provide an explanation, suggesting that the SM is in fact an effective theory that may be valid up to a certain energy scale. This is not the first time that a physical theory that has provided several verifiable predictions falls short when accommodating new observations. In fact one could argue that this is a desirable thing to happen, for it is a

sign that new physics is to be discovered. In this section, a limited exposure to some of these phenomenological issues that may be hinting at new physics will be presented. More attention will be given to the Hierarchy Problem, which is related to the VLQ searches presented in this thesis.

Gravity

As discussed in section 2.1, the SM predicts the existence of four gauge bosons, which are the quantizations of the EWK and strong fundamental forces. Gravity is the remaining fundamental force that has evaded a quantum description of its interactions. Fundamentally, this issue can be explained as trying to reconcile QFT and general relativity (GR) into a single theory. Although QFT incorporates special relativity, as can be seen in Dirac's equation, the full effects of curved space-time are not taken into account in QFT. This would render the theory nonrenormalizable due to the self-interaction terms that a mediator particle of gravity would have. This is a puzzling phenomenon, as gravitational interactions are very weak at short-distance scales that are characteristic of the SM interactions but become dominant at astronomical scales. This suggests that the SM is an effective theory that is not able to fully resolve all degrees of freedom at smaller length scales, which correspond to more energetic interactions.

Baryogenesis

The fact that the universe exists with matter predominantly occupying space and is not a vacuum occupied by energy is a phenomenon that the SM cannot explain. Although the SM does provide the interactions to produce matter and anti-matter pairs from energy, there

is no built-in mechanism that favors the production of one over the other, which resulted in the excess of ordinary matter after the big bang. As ordinary matter is composed of baryons, this phenomenon, known as Baryogenesis, indicates that the SM does not conserve the number of baryons at a fundamental level, hinting at a possible breaking of an unknown symmetry.

Dark Matter

The existence of dark matter has been used to explain astronomical phenomena such as the discrepancy in rotational curves of galaxies [9], which measure the velocity distribution of stars in galaxies as a function of their distance to the center of the galaxy. Without dark matter, the rotational curves are expected to decrease at larger distances from the center since there is less matter to provide gravitational pull to the stars. Instead, an increase in velocity is observed at larger distances, as shown in Figure 2.8. Another phenomenon

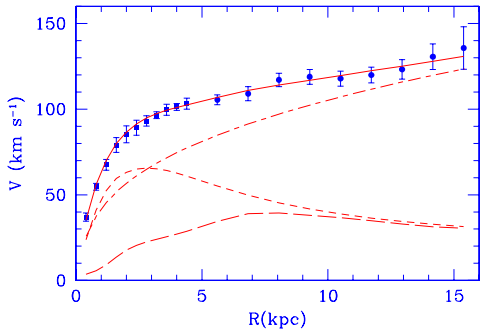


Figure 2.8: Rotational curve of the spiral galaxy Messier 33 (M33) broken down into individual contributions. The contribution to the velocity distribution from gas in the galaxy is shown by the long dashed line, from the stellar disk by the short dashed line, and from the dark matter halo by the dashed-dotted line. The continuous line is the best fit model which incorporate dark matter to explain the observed velocity data points. This figure is taken from [10].

that can be explained by dark matter is gravitational lensing in regions where there is not enough visible matter after the collision of galaxy clusters [11]. In both examples, dark

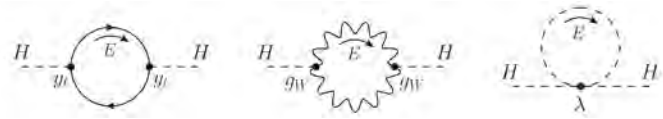


Figure 2.9: Loop corrections to the Higgs boson mass from interactions with the top quark, EWK bosons, and self-interactions. Figure taken from [3].

matter would explain these observations by being an abundant source of dense matter that does not interact electromagnetically with ordinary matter. The phenomenological aspect of dark matter in the SM is the existence of new particles that form dark matter, which could have their own set of interactions that are not part of the SM.

Hierarchy Problem

The Hierarchy Problem has its origins in the large discrepancy between the EWK energy scale $v \approx 256$ GeV, which is related to the Higgs boson mass through the EWK spontaneous symmetry breaking, and the Planck energy scale $M_P = \sqrt{\hbar/(8\pi G_N)} \approx 2.4 \times 10^{18}$ GeV, at which gravitational effects in particle interactions need to be taken into consideration. The Planck energy scale can be taken as the cutoff energy scale Λ_{SM} at which the SM loses its predictive power as an effective theory. Given that the EWK and Planck scales are separated by approximately 17 orders of magnitude and the Higgs boson mass sits at the lower end of the spectrum, this indicates that there is a wide range of energy scales with no physics that is described by the SM. To fully appreciate the phenomenological issue that is the Hierarchy Problem, one needs to take into account the contributions that the particles of the SM that interact with the Higgs have on its mass. These contributions arise from loop corrections, such as the ones shown in Figure 2.9. If all fundamental parameters of the theory that

describe the Higgs mass are known, then the Higgs mass can be generalized as

$$m_h^2 = \int_0^\infty dE \frac{dm_h^2}{dE}(E; p_{\text{true}}) \tag{2.57}$$

where the integrand contains all loop corrections to the Higgs mass that originate from the SM particles. This motivates to split the integral into two regions that are defined by the cutoff energy scale Λ_{SM} as follows:

$$m_h^2 = \int_0^{\Lambda_{\text{SM}}} dE \frac{dm_h^2}{dE}(E; p_{\text{true}}) + \int_{\Lambda_{\text{SM}}}^\infty dE \frac{dm_h^2}{dE}(E; p_{\text{true}}) = \delta_{\text{SM}} m_h^2 + \delta_{\text{BSM}} m_h^2 \tag{2.58}$$

where $\delta_{\text{SM}} m_h^2$ are the contributions to the Higgs mass that are attributed to the SM interactions, and $\delta_{\text{BSM}} m_h^2$ are unknown contributions from BSM physics. The SM term can be roughly estimated from the main loop correction contributions as

$$\delta_{\text{SM}} m_h^2 = \frac{3y_t^2}{4\pi^2}\Lambda_{\text{SM}}^2 - \frac{3g_2^2}{8\pi^2}\left(\frac{1}{4} + \frac{1}{8\cos^2\theta_W}\right)\Lambda_{\text{SM}}^2 - \frac{3\lambda}{8\pi^2}\Lambda_{\text{SM}}^2 \tag{2.59}$$

where each individual term corresponds to the top quark loop, EWK boson loops, and Higgs loop, respectively. The most important of these terms is the one from the top quark, due to its large Yukawa coupling $y_t^2 \approx 1$ that is proportional to the top quark mass. Since this term has a large positive contribution if Λ_{SM} is sufficiently large, in order to produce the relatively small value of the Higgs mass with the true theory, the BSM term must provide a cancellation of roughly equal magnitude and opposite sign as the SM term. This can only be achieved if the fundamental parameters of the theory are fine-tuned to produce a cancellation

Δ that can be bounded below as:

$$\Delta \geq \frac{\delta_{\mathrm{SM}} m_h^2}{m_h^2} = \frac{3y_t^2}{4\pi^2}\left(\frac{\Lambda_{\mathrm{SM}}}{m_h}\right)^2 \approx \left(\frac{\Lambda_{\mathrm{SM}}}{450\ \mathrm{GeV}}\right)^2 \tag{2.60}$$

As an example, if the energy scale to discover new physics turns out to be the scale of a grand unified theory (GUT), $\Lambda_{\mathrm{SM}} = M_{\mathrm{GUT}} \approx 10^{15}$ GeV, where all fundamental forces of nature are described by a single force, the cancellation would be of the order $\Delta \geq 10^{24}$. This is a glaring issue when the true theory parameters that describe the SM and BSM terms, which are completely unrelated, have to produce a 24 digit cancellation in order to explain the Higgs mass.

2.3 Vector-Like Quark Theory Overview

Several BSM theories have been proposed to solve the Hierarchy Problem presented in the previous section, such as Supersymmetry (SUSY) and Composite Higgs (CH). The latter will be the focus of this section, as a key prediction of Higgs compositeness is the existence of new fermionic resonances that are referred to as vector-like quarks (VLQs).

2.3.1 Composite Higgs Models

The idea behind the CH model is that the Higgs is not an elementary particle but instead a bound state of some new particles that interact through a new force. This new force would then give the Higgs boson a finite geometric size l_h, similar to how the quark constituents of the proton are bound within its radius by the strong force. The binding energy of the Higgs is then given by $m_* = 1/l_h$, which can be taken as the cutoff scale in Equation 2.58.

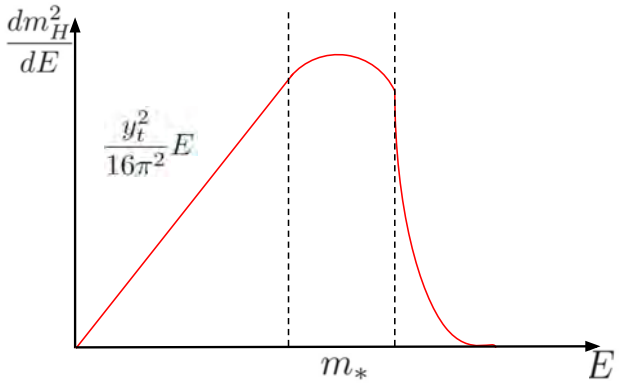


Figure 2.10: Representation of the Higgs mass integrand under the Composite Higgs model. This figure is adapted from [3].

Thus, for energies below m_*, the Higgs boson behaves like an elementary point-like particle since any interaction with the Higgs will not have enough energy to resolve its substructure, just like a photon with a wavelength larger than the radius of the proton cannot resolve the individual quark constituents. In this energy range, the integrand in Equation 2.58 behaves linearly, with the largest quadratic contribution coming from the top quark. As the energy approaches the scale m_*, the finite size of the Higgs becomes evident, which results in the integrand reaching a maximum value. The integrand then sharply decreases at higher energies once the compositeness of the Higgs becomes apparent, as shown in Figure 2.10. The decrease in the integrand for energies above m_* results from the fact that there are no particles in this energy regime that would provide radiative contributions to the Higgs mass. Effectively, the CH model solves the Hierarchy Problem by providing a mechanism that stabilizes the Higgs mass as a result of the interactions of new particles under a new force, which form a bound state corresponding to the Higgs boson.

In order to accommodate the CH model into the SM two different structures will need to be defined: a composite sector (CS), which will contain the new particles that form the Higgs boson bound state along with their interactions, and an elementary sector (ES), which

contains all the SM particles that are known to be elementary. In addition to these two sectors, a new set of Elementary-Composite interactions $\mathcal{L}_{\text{EC}}$ has to be included in order to generate the masses of the SM gauge bosons and fermions since the Higgs is no longer present in the ES.

A potential issue that the CS might have is that if the Higgs boson is a bound state of particles in the CS, then it is expected that m_h should be close to the binding energy scale m_*. This is motivated by observing that the masses of hadrons, which are bound states of QCD interactions, are close to the color confinement scale $\Lambda_{\text{QCD}} \approx 300$ MeV. Specifically, the issue is that if $m_* \approx m_h$, then other bound states of the CS different than the Higgs would have been observed by now. This absence of additional bound states close to the EWK energy scale motivates placing m_* at least at the TeV scale as a minimum.

The problematic lightness of the Higgs mass can be explained by the Higgs boson being a pseudo Nambu-Goldstone Boson (pNGB)[1] of a symmetry group G. The group G is required to contain the SM symmetry groups as a subgroup in order to be compatible with the description of the ES. Since the symmetry to be broken is unrelated to the ES this corresponds to reducing G into an unbroken subgroup $H \leq G$. By Goldstone's Theorem, for each broken symmetry generator that is not an element of H a massless NGB arises. If the Higgs boson is to have mass, then it cannot be fully generated in the CS. This requires that a mass generating mechanism is present in $\mathcal{L}_{\text{EC}}$.

In reality, the CH model is not just a single model but a family of models, which are mostly defined by the nature of the group G and its unbroken subgroup H. The remainder of the discussion will be kept as general as possible regarding the choice of CH model, since

[1] A pNGB arises when an approximate symmetry is spontaneously broken instead of an exact symmetry, therefore giving them mass.

the theory behind these models is outside the scope of this thesis. When additional model-specific details are required, only the Minimal Composite Higgs Model (MCHM) will be discussed. The MCHM is based on the choice of $G = SO(5)$ rotational symmetry in the CS, which is reduced to the $H = SO(4)$ rotational symmetry in order to spawn the Higgs boson as a pNGB. Other extended CH models are based on larger groups G and H, which contain embeddings of their MCHM counterparts.

2.3.2 Gauge Boson Masses

In the MCHM, the Higgs is now represented as a 5-dimensional vector with real entries that can be parametrized as

$$\Phi = f \begin{pmatrix} \sin\frac{\Pi}{f}\frac{\vec{\Pi}}{\Pi} \\ \cos\frac{\Pi}{f} \end{pmatrix} \tag{2.61}$$

where $\vec{\Pi}$ is a 4-dimensional vector with norm Π, of which its entries will be the Goldstone bosons of the MCHM. The parameter f is the Higgs decay constant that represents the energy scale of the spontaneous symmetry breaking of $SO(5)$ into $SO(4)$, which is analogous to the pion decay constant f_π in QCD interactions where pions are pNGB. In order to be consistent with the SM $SU(2)$ representation of the Higgs as given in Equation 2.36, we require that

$$\vec{\Pi} = \begin{pmatrix} \Pi_1 \\ \Pi_2 \\ \Pi_3 \\ \Pi_4 \end{pmatrix} = \begin{pmatrix} \phi_2 \\ \phi_1 \\ \phi_4 \\ \phi_3 \end{pmatrix} \tag{2.62}$$

which can be interpreted as an isomorphism between the unbroken subgroup $SO(4)$ and the group $SU(2)_L \times SU(2)_R$, where $SU(2)_L$ is the same as the one present in the SM, while

$SU(2)_R$ generalizes the SM hypercharge $U(1)_Y$. To spontaneously break the CS symmetry, we follow a similar procedure as done with the EWK spontaneous symmetry breaking, starting with the Lagrangian

$$\mathcal{L} = \frac{1}{2}(\partial_\mu \bar{\Phi})(\partial^\mu \Phi) + V(\Phi) \tag{2.63}$$

and letting the covariant derivative transform as

$$\partial_\mu \to \partial_\mu + ig_2 W_\mu^a T_L^a + ig_1 B_\mu T_R^3 \tag{2.64}$$

where W_μ^a, $a = 1, 2, 3$, and B_μ are the SM gauge bosons and T_L^a, T_R^3 are the generators of $SO(4)$. Next, we expand the composite Higgs around the minimum of an interaction potential $V(\Phi)$, which will be left unspecified as its form depends on the CH model being studied, with the following choice of values for the fields Π_i that corresponds to the unitary gauge

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \vec{\Pi} = \vec{0} \\ \\ V + h(x) \end{pmatrix} \tag{2.65}$$

where V is the VEV of the composite Higgs under the interaction potential. Inserting this expansion into the Lagrangian results in the following expression

$$\mathcal{L} = \frac{1}{2}\partial_\mu h \partial^\mu h + \frac{g_2^2 f^2}{4} \sin^2\left(\frac{V+h}{f}\right)\left(W_\mu^+ W^{-\mu} + \frac{Z_\mu Z^\mu}{2\cos^2\theta_W}\right) \tag{2.66}$$

In order to maintain compatibility with the SM, the following relation must hold true for the gauge boson masses

$$m_W = m_Z \cos\theta_W = \frac{g_2 f}{2} \sin\left(\frac{V}{f}\right) \tag{2.67}$$

which links the EWK symmetry breaking scale v with the Higgs decay constant f. The angle $\theta = V/f$ measures how misaligned the VEV of the Higgs is relative to the direction in 5-dimensional space where the Higgs has vanishing VEV. Thus, the parameter $\xi = v^2/f^2 = \sin^2(V/f)$ measures the relative size of the EWK symmetry breaking scale to the $SO(5)$ spontaneous symmetry breaking scale. If we expand Equation 2.66 in a Taylor series with respect to the Higgs field $h(x)$, the following infinite set of interactions with the gauge bosons are obtained

$$\mathcal{L} = \frac{g_2^2 v^2}{4}\left(W_\mu^+ W^{-\mu} + \frac{Z^\mu Z_\mu}{2\cos^2\theta_W}\right)\left[\sqrt{1-\xi}\frac{2h}{v} + (1-2\xi)\frac{h^2}{v^2} - \xi\sqrt{1-\xi}\frac{4h^3}{3v^3} + \cdots\right] \tag{2.68}$$

From this expansion, we can see that single and double interactions between the gauge and Higgs bosons arise similar to the SM scenario, but with modified coupling strengths

$$k_V = \frac{g^{\text{CH}}_{hVV}}{g^{\text{SM}}_{hVV}} = \sqrt{1-\xi}, \quad k_{Vh} = \frac{g^{\text{CH}}_{hhVV}}{g^{\text{SM}}_{hhVV}} = 1 - 2\xi \tag{2.69}$$

Thus, being able to measure these coupling strengths to high precision or any of the additional Higgs-gauge boson interactions that are absent from the SM could experimentally validate the CH model. It should be noted that as $\xi \to 0$ when v is held fixed and $f \to \infty$, then the modified couplings reduce to their SM values and any interaction beyond the double interaction vanish due to being proportional to ξ. In this limit, the composite nature of the Higgs reduces to its elementary SM behavior, and we recover the SM description of the EWK symmetry breaking.

2.3.3 Fermion Masses

To incorporate the SM fermion masses, the following set of terms must be included in the interaction Lagrangian

$$\mathcal{L}_{\text{int}}^{\text{fermion}} = \lambda_L \bar{\Psi}_L^i \mathcal{O}_i + \lambda_R \bar{\Psi}_R^i \mathcal{O}_i \tag{2.70}$$

The CS operators $\mathcal{O}$ couple to the $SO(5)$ embeddings of the SM $SU(2) \times U(1)$ left-handed fermion doublet, Ψ_L, with coupling strength λ_L, and the right-handed fermion singlet, Ψ_R, with coupling strength λ_R. The embeddings are represented as

$$\Psi_L = \frac{1}{\sqrt{2}} \begin{pmatrix} -i\psi_L \\ -\psi_L \\ -i\psi_L \\ \psi_L \\ 0 \end{pmatrix}, \quad \Psi_R = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \psi_R \end{pmatrix} \tag{2.71}$$

Additionally, the following interaction term between the SM fermions and the CS Higgs must be included, which will determine their coupling strength

$$\tilde{\mathcal{L}}_{\text{int}}^{\text{fermion}} = -\frac{\sqrt{2} m_\psi}{\sqrt{(\xi(1-\xi))}} \Phi_i \bar{\Psi}_L^i \Psi_R \tag{2.72}$$

Only the terms pertaining to the third generation of quarks will be considered as they are the most relevant when discussing the Hierarchy Problem. Following a similar procedure as with the SM gauge bosons, the interaction Lagrangian in Equation 2.72 can be Taylor expanded near the VEV of the Higgs field Φ into the following expression that only involves

the top quark terms:

$$\tilde{\mathcal{L}}_{\text{int}}^{\text{fermion}} = -m_t \bar{t}t - \frac{1-2\xi}{\sqrt{1-\xi}} \frac{m_t}{v} h\bar{t}t + 2\xi \frac{m_t}{v^2} h^2 \bar{t}t + \cdots \tag{2.73}$$

From this expression, it is observed that the top quark coupling to the Higgs is modified as:

$$k_t = \frac{g_{h\bar{t}t}^{\text{CH}}}{g_{h\bar{t}t}^{\text{SM}}} = \frac{1-2\xi}{\sqrt{1-\xi}} \tag{2.74}$$

A similar expansion is obtained for the bottom quark, which yields modified couplings between the Higgs and the bottom quark. However, since each term in the expansion is proportional to m_b and $m_b < m_t$, they are thus irrelevant to the radiative corrections to the Higgs mass. To summarize, the basic structure of the MCHM is described by Equation 2.66, Equation 2.70, and Equation 2.72. Three model-dependent parameters are introduced: λ_L, λ_R, and ξ, of which this last parameter modifies the couplings between the Higgs and the remaining SM particles. It should be noted that a fine tuning requirement similar to the one derived in Equation 2.60 can be constructed for the parameter ξ, which behaves as

$$\Delta \geq \frac{1}{2\xi} \tag{2.75}$$

Thus, as the measurements of the coupling strengths become more stringent, the MCHM becomes more unnatural due to the degree of fine tuning it requires.

2.3.4 Vector-Like Quarks

The exact form of the CS operator $\mathcal{O}$ has remained unspecified until now. By inspecting Equation 2.70, we can determine that the particles associated with this operator must

Vector-Like Quark	Electric Charge
X	+5/3
T	+2/3
B	-1/3
Y	-4/3

Multiplet	Decays	Hypercharge
	$SU(2)$ singlets	
(T)	$T \to Ht/Zt/W^+b$	+2/3
(B)	$T \to Hb/Zb/W^-t$	-1/3
	$SU(2)$ doublets	
(T,B)	$T \to Ht/Zt$, $B \to W^-t$	+1/6
(X,T)	$T \to Ht/Zt$, $X \to W^+t$	+7/6
(B,Y)	$B \to Hb/Zb$, $Y \to W^-b$	-5/3

Table 2.1: Overview of VLQs and their multiplets that are predicted by the MCHM.

have spin 1/2 in order to couple with the SM fermions and for the interaction Lagrangian to remain Lorentz invariant. Secondly, their left-handed and right-handed components must transform similarly under the weak isospin $SU(2)$ gauge group, meaning that they are "vector-like" fermions, which differs from the SM chiral fermions. Finally, these new operators must transform as an $SU(3)$ triplet in order to be compatible with the SM Lagrangian; thus, the particles associated to the operators must have color charge, just like quarks. For the stated reasons, these new particles are known as vector-like quarks (VLQs). An overview of the VLQs as well as their possible multiplets that are predicted by the MCHM is summarized in Table 2.1.

At the LHC, VLQs are expected to be produced in pairs through QCD interactions or singly through weak interactions, as shown in Figure 2.11. The pair production mechanism is analogous to the pair production of SM quarks, taking the form $q\bar{q} \to Q\bar{Q}$ or $gg \to Q\bar{Q}$, where q and g are SM quarks and gluons, respectively, and Q is a VLQ. The cross section

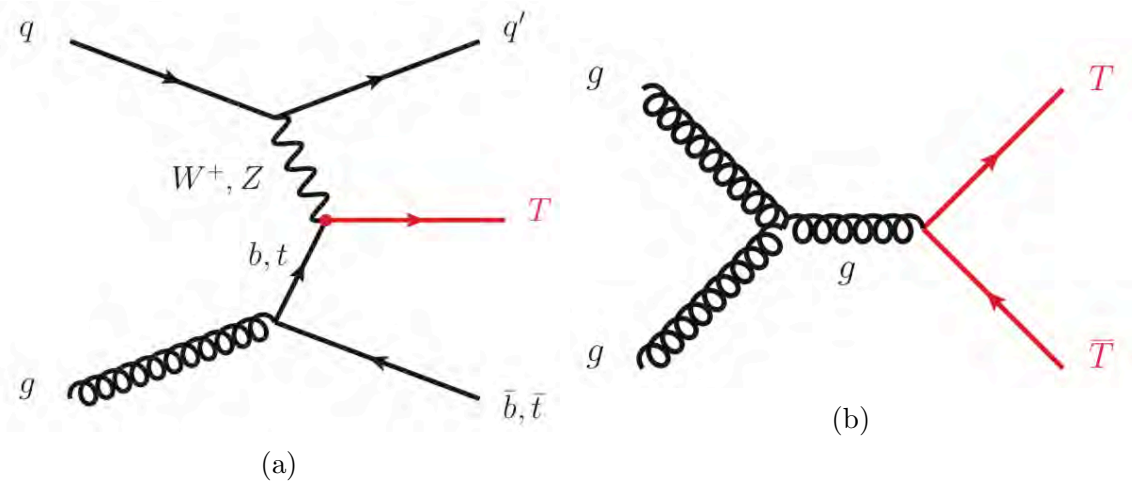


Figure 2.11: Feynman diagrams of single production (a) and pair production (b) of a vector-like top.

for the pair production process only depends on the mass of the VLQ and the center of mass energy of the interaction, as shown in Figure 2.12. On the other hand, the single production cross section is more model dependent since this mechanism couples the VLQs with the third generation SM quarks, the weak gauge bosons, and the Higgs boson. As a consequence of this, the single production cross section overtakes the pair production cross section at higher VLQ masses since the phase space is less suppressed for the production of a single heavy VLQ due to its dependence on the coupling parameters. The branching ratios of the vector-like top and bottom as a function of their mass is shown in Figure 2.13. The vector-like top in the singlet representation is expected to decay into Wb half of the time, with the other half being distributed almost equitably between the Ht and Zt decays as the mass of the VLQ increases. A similar behavior is expected for the vector-like bottom in the singlet representation by exchanging the top and bottom quarks. For the doublet representations (T,B) and (X,T), the decays into Wb are ruled out due to charge conservation. Similarly, the Wt decay is ruled out in the doublet representation (B,Y).

Mixing terms between the SM quarks and their VLQ partners are also generated, of

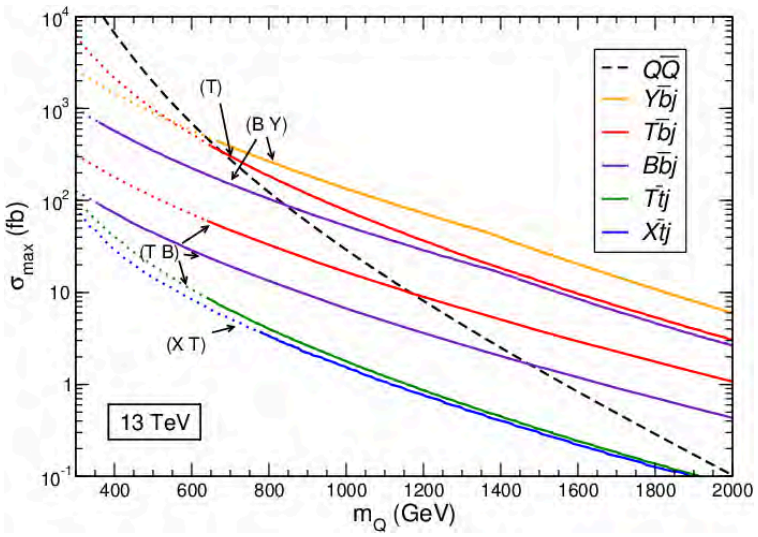
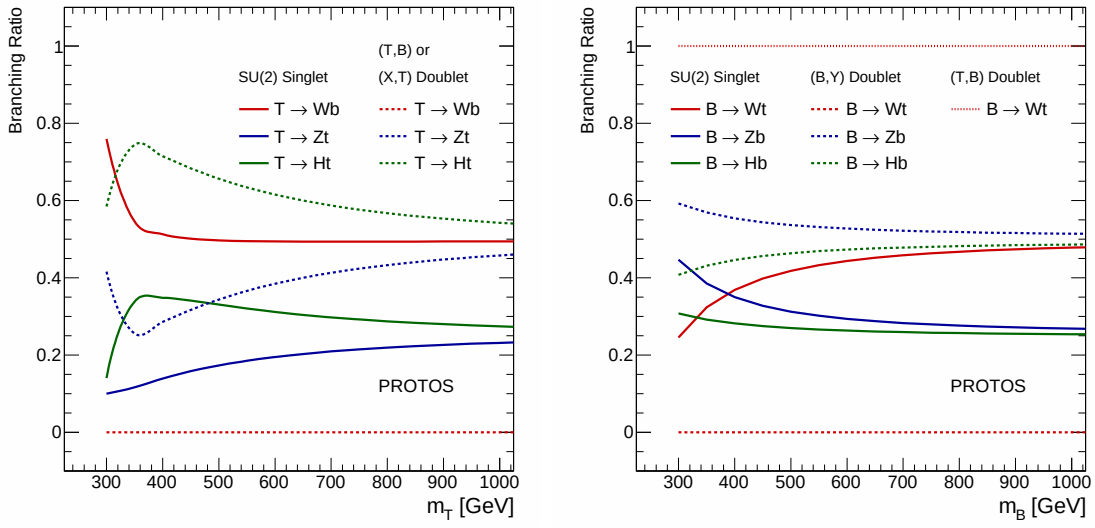


Figure 2.12: Cross section of the different production mechanisms for VLQs as a function of the VLQ mass in GeV at a center-of-mass energy of 13 TeV. The dashed black line represents the cross section of pair production of VLQs while the colored lines represent the cross section of the single production of VLQs for different $SU(2)$ doublet configurations. For single production the maximum cross section at each mass point is obtained by setting the mixing terms between the VLQ and the SM quarks to the maximum value allowed by theoretical constraints. The dotted portion of the colored lines indicate outdated exclusion limits on the cross section. This figure is taken from [12].



(a) (b)

Figure 2.13: Theoretical predictions of the branching ratios of a vector-like top (a) and a vector-like bottom (b) as a function of their mass in GeV for different $SU(2)$ multiplet configurations. This figure is taken from [13].

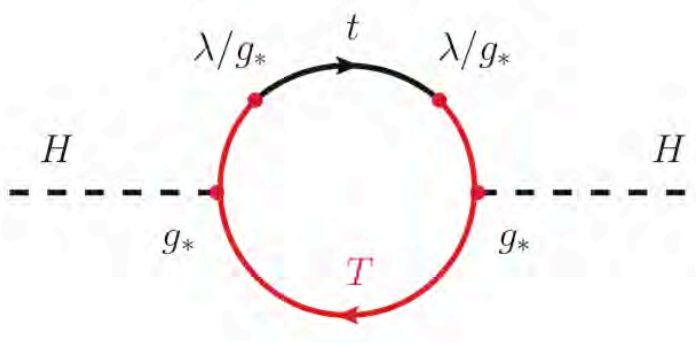


Figure 2.14: Representation of the vector-like top contribution to the Higgs boson mass.

which the most important one is the top and vector-like top mixing term

$$\mathcal{L}_{\text{mix}} = \frac{\lambda_L}{m_*} g_* \bar{T} t_L + \frac{\lambda_R}{g_*} m_* \bar{\tilde{T}} t_R \tag{2.76}$$

where $g_* = m_*/f$, t is the SM top quark and T is vector-like top component of an $SU(2)$ doublet and $\tilde{T}$ the corresponding singlet. The mixing term can be used to generate loop diagrams, as shown in Figure 2.14, that involve the exchange of a virtual top quark and vector-like top resulting in the following expression of the Higgs mass

$$m_h^2 \approx a_L \frac{\lambda_L^2}{16\pi^2} M_T^2 + a_R \frac{\lambda_R^2}{16\pi^2} M_T^2 \tag{2.77}$$

where the coefficients a_L and a_R depend on the actual CH model being studied. If the masses of the VLQs are large, then in order to obtain the observed value of the Higgs mass, the coefficients a_L and a_R must provide a cancellation that has a fine tunning bounded below by

$$\Delta \geq \frac{3y_t^2}{4\pi^2} \left(\frac{M_T}{m_H} \right)^2 \tag{2.78}$$

under the assumption that $\lambda_L = \lambda_R = \sqrt{y_t g_*}$, which minimizes this lower bound. Thus,

this fine tuning requirement provides an experimental handle to constrain the CH models through limits on the mass of VLQs.

Chapter 3

The LHC and the ATLAS Detector

All of the ordinary matter in the universe is mostly composed of up and down quarks, which are bound together into protons and neutrons that form the nuclei of atoms, and electrons, which orbit around the nucleus and dictate the chemistry that allows the formation of complex structures like molecules. The remainder of the SM particles and most of the hypothetical particles that are predicted by BSM theories, such as VLQs, are short-lived due to their large masses and subsequently decay into the lighter particles of the SM. As a consequence of this, these exotic particles can only be produced in highly energetic interactions, such as relativistic collisions between lighter particles. To properly study these exotic particles, an experimental apparatus that can produce them in a controlled experimental environment is required. This is achieved with man-made particle accelerators that accelerate beams of particles to relativistic speeds and focus the beam into desired collision points with the help of strong electric and magnetic fields.

At the time of writing this dissertation, the Large Hadron Collider (LHC) [14] is the most powerful man-made accelerator used to produce the energetic collisions that allow us to make precision measurements of parameters of interest in the SM and perform searches for BSM physics. In order to probe the results of these energetic collisions, the LHC has four main collider detector experiments that are designed for different purposes. The ATLAS (A Toroidal LHC ApparatuS) [15] and CMS (Compact Muon Solenoid) [16] experiments are both general

purpose detectors that are used for SM precision measurements and BSM searches. These two experiments are designed to be independent from each other by having different detector designs and collaborations, which is essential when potential discoveries made at the LHC need to be validated. The LHCb (Large Hadron Collider beauty) [17] experiment focuses on precision measurements of processes that exhibit charge-parity (CP) violations and b-hadron physics. Finally, the ALICE (A Large Ion Collider Experiment) [18] detector focuses on heavy-ion collisions to study QCD interactions and the physics of quark-gluon plasma. In addition to the four main detector experiments, there are also three smaller experiments. The TOTEM (TOTal Elastic and diffractive cross section Measurement) [19] experiment is dedicated to the measurement of the cross-section of proton-proton (pp) collisions and is located at the CMS interaction point. The LHCf (LHC forward) [20] experiment is dedicated to the study of particles that are emitted in the forward regions of LHC collisions and provides calibrations for the hadron interaction models in extremely high-energy cosmic rays. This experiment is located at the ATLAS interaction point. Finally, the MoEDAL (Monopole and Exotics Detector at the LHC) [21] experiment, which is located near the LHCb interaction point, is designed for the search of magnetic monopoles and massive pseudo-stable charged particles that are predicted by BSM theories.

In this chapter, a description of the LHC is given, focusing on how particles are accelerated to achieve the energies required to produce and study exotic particles. Next, a description of the ATLAS detector will be presented, which is the experimental apparatus from which the data and simulations used in the analyses presented in this dissertation were obtained. Finally, the process of reconstructing and calibrating the different physics objects that are used in ATLAS analyses from the inputs of the detector is described.

3.1 The Large Hadron Collider

The LHC is a circular synchrotron accelerator with a circumference of 26.7 km that is located underground at a depth that varies between 50 m and 170 m and crosses the Franco-Swiss border near Geneva, as depicted in the schematic in Figure 3.1. It started its operations

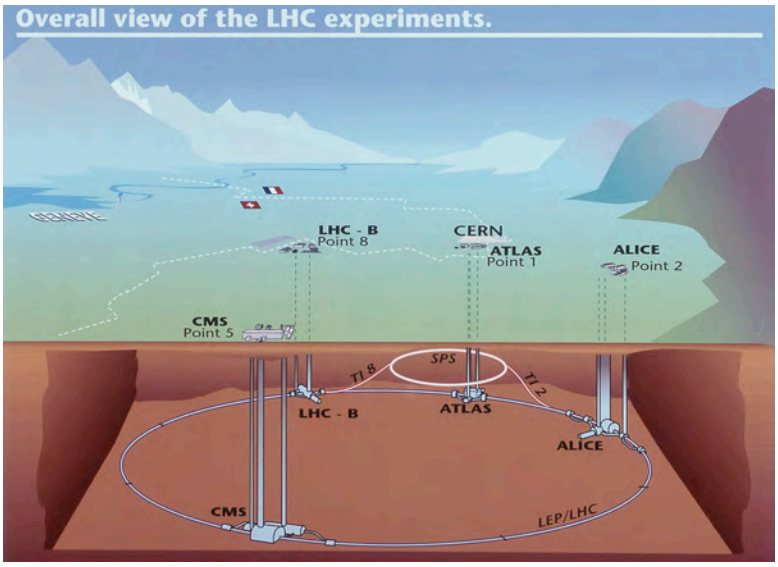


Figure 3.1: Schematic representation of the LHC and its four main experiments [22].

in the year 2009, collecting data from collisions at a center of mass energy of 7 TeV until 2011, at which point the energy was increased to 8 TeV, providing additional collision data until 2013. This data collection period is known as Run-1, after which the LHC was shut down to perform upgrades to the accelerator complex and its detectors. These upgrades allowed the LHC to resume its data collection operations in 2015 using collisions with a center of mass energy of 13 TeV. This data collection period, known as Run-2, lasted until 2018, at which point the LHC entered into another scheduled shutdown for upgrades. At the time of writing this dissertation, the LHC resumed its operations in July 2022, starting

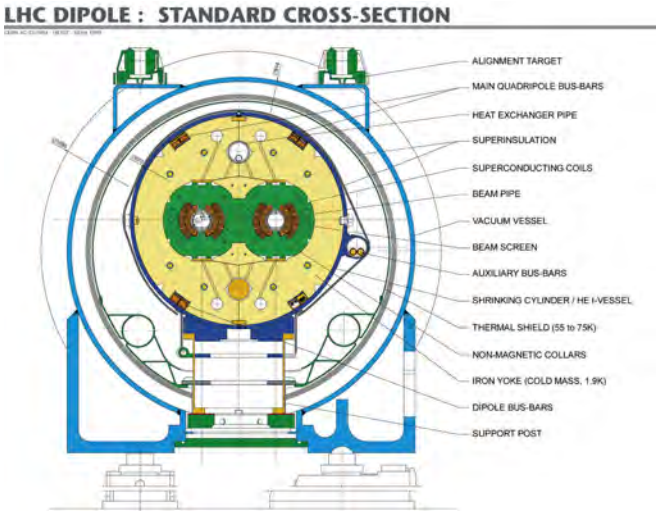


Figure 3.2: Cross section of the LHC beam pipe [23].

the Run-3 data collection period with a center of mass energy of 14 TeV, which will last for approximately four years until the next scheduled shutdown. The data used in the analyses presented in this dissertation are from the Run-2 period.

The main ring of the LHC consists of two separate beam pipes in which a beam of particles to be collided is split to travel in opposite directions along the circumference of the ring, guided by superconducting magnets that surround the pipes, as shown in Figure 3.2. Along the LHC ring, there are four sections known as cryomodules that contain radio frequency (RF) cavities that produce an oscillating electric field tuned to a frequency of 400 MHz designed to accelerate particles up to an energy of 6.5 TeV [1]. The particles in the beam that arrive to the RF cavities in phase with the electric field are accelerated, while those that arrive out of phase are decelerated, which allows to sort the particle beam into particle bunches. Once the bunches reach the desired beam energy, they are collided at the interaction points where the different detector experiments are located. In order for this to occur, the

[1] Since the start of Run-3 particles are now accelerated to 6.8 TeV.

particle beam must consist of stable and charged particles so that they do not decay before reaching the interaction points and can be accelerated and guided through electromagnetic interactions. This limits the choice of beam constituents to electrons, protons[2], or ions. Since the LHC is a circular accelerator, a beam of electrons loses more energy per revolution across the LHC compared to heavier particles due to synchrotron radiation, which can be quantified as:

$$P = \frac{\Delta E}{2\pi R} = \frac{4\pi q^2 \beta^2 E^4}{3Rm^4} \tag{3.1}$$

where q and m are the charge and mass of the particles in the beam, E is the beam energy, R is the orbit radius and $\beta = v/c$ is the ratio of the speed at which the particles in the beam are traveling at to the speed of light, which is approximately equal to 1 for the purposes of the LHC operations. Thus, an electron beam that is accelerated circularly to the same energy as a proton beam will lose more energy by a factor of $(m_p/m_e)^4 \approx 10^{13}$, making the use of an electron beam energetically inefficient to maintain at the LHC. For this reason, the LHC was designed primarily to collide proton-proton (pp) beams to study the fundamental particles they produce. Lead ions have also been used in lead-lead and lead-proton beam collisions. These collisions are used to study a state of matter known as quark-gluon plasma. From this point on, only pp collisions will be discussed, as they are the main interaction of interest for this thesis.

The process of producing the proton beams is depicted in Figure 3.3. This process starts by ionizing hydrogen gas with the use of electric fields so that the protons are separated from the electrons. The protons are then transported through the 33 m long linear accelerator (Linac2), which accelerates them to an energy of 50 MeV. The next steps consist of sequen-

[2]Or their corresponding antiparticles.

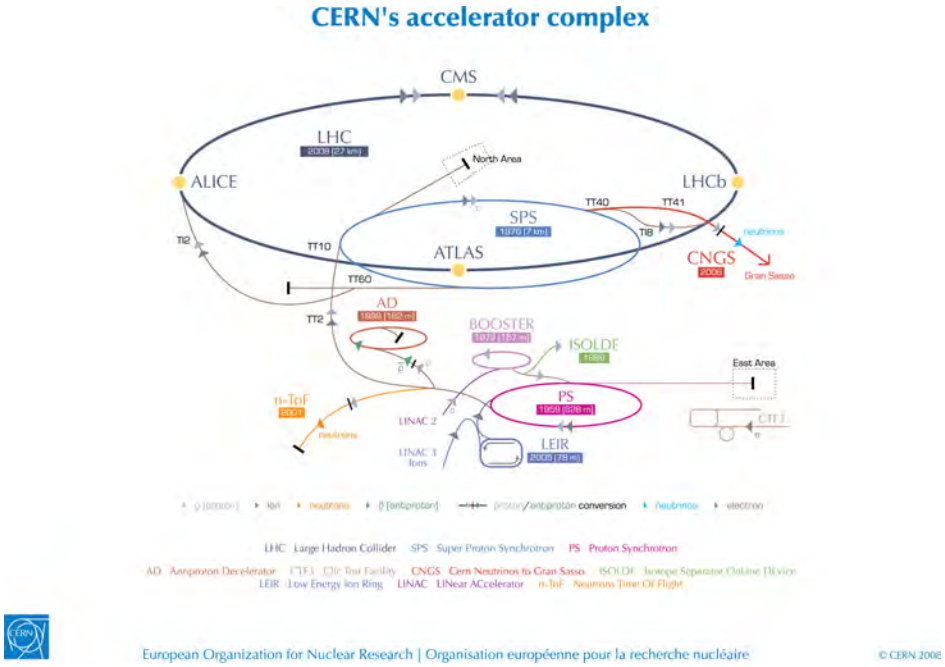


Figure 3.3: Schematic representation of the process of injecting the proton beam into the LHC. The light gray arrows indicate the direction in which the beam of protons travel during the injection process [24].

tially accelerating the beam that is formed in Linac2 through the use of smaller circular accelerators, which also help in sorting the beam into proton bunches. First, the incoming beam from Linac2 is injected into the Proton Synchrotron Booster (PSB) that increases the beam energy to 1.4 GeV. The beam is then transported to the Proton Synchrotron (PS), which increases the beam energy to 25 GeV. The beam is then transported to the last of the small circular accelerators, the Super Proton Synchrotron (SPS), where the beam is further accelerated to an energy of 450 GeV. Finally, the proton beam is split into two beams and injected into the LHC beam pipes, where each beam will travel in opposite directions while being accelerated to its final energy of 6.5 TeV. Once this energy is reached, the opposing beams are ready to collide at the interaction points, resulting in a collision with a center of mass energy of 13 TeV. Under nominal *pp* collision operations, there are 2808 proton

bunches circulating around the LHC ring, with each bunch equidistantly spaced by 25 ns and containing approximately 1.1×10^{11} protons.

In an accelerator experiment, the number of events of a particular process that are produced can be expressed as

$$N_{\text{events}} = \sigma \int L dt \tag{3.2}$$

where σ is the cross section of the process of interest and L is the luminosity of the accelerator. In particle physics it is standard practice to measure the cross section of an event in units of barns (b), where $1\text{b} = 10^{-28}\text{m}^2$. It should be noted that Equation 3.2 is only valid if the detector completely encapsulates the collision target, which is the case for the different detectors at the LHC. The value of the cross section is determined by nature, so the only experimental handle for tuning the event rate of a process comes from the accelerator luminosity. The luminosity of the LHC can be expressed as

$$L = \frac{n_b n_1 n_2 f_r}{2\pi \Sigma_x \Sigma_y} \tag{3.3}$$

where n_b is the number of proton bunches crossing the interaction points, n_1 and n_2 are the number of protons in the colliding bunches, f_r is the collider revolution frequency, and Σ_x and Σ_y are the horizontal and vertical geometric widths of the proton beams, respectively [25]. The cross section of a process is inversely proportional to the energy scale associated to the process. In the case of particles that are predicted by BSM theories with masses above the TeV scale, it is necessary to increase the accelerator luminosity in order to have sufficient statistics to make a claim of discovery of these potentially new particles. This can be achieved by focusing the proton beams as they approach the interaction points, thereby reducing the geometric widths in the denominator of Equation 3.3 and increasing the probability of

collision between opposing proton bunches.

The total integrated luminosity delivered by the LHC and recorded by the ATLAS detector during Run-2 is shown in Figure 3.4. The effect of increasing the interaction rate of

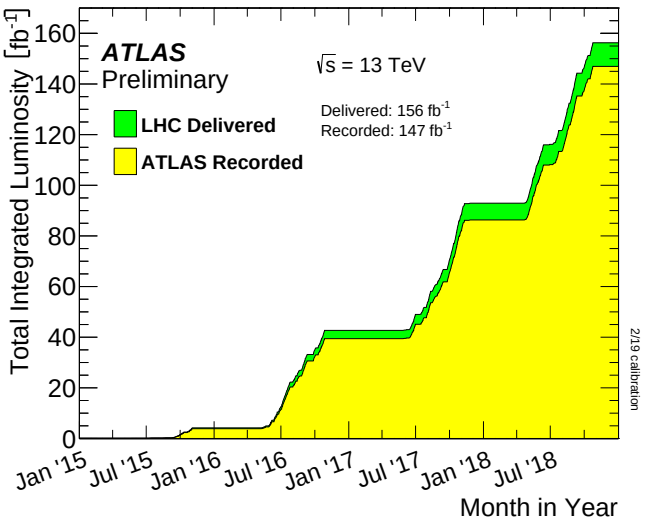


Figure 3.4: The integrated luminosity, in units of inverse femptobarn, over the time period of Run-2 delivered by the LHC (green) and recorded by ATLAS (yellow) during stable beams for pp collisions at a center of mass energy of 13 TeV. This figure is taken from [26].

protons in a single bunch crossing is known as in-time pile-up. The number of interactions in a single bunch crossing follows a Poisson distribution, with the mean number of interactions given by:

$$\mu = \frac{L\sigma_{pp\,\mathrm{inel}}}{f} \tag{3.4}$$

where L is the instantaneous luminosity in Equation 3.3, $\sigma_{pp\,\mathrm{inel}}$ is the cross section of inelastic pp collisions and f is the collision frequency of the LHC. At nominal operations for pp collisions, the peak luminosity of the LHC is $L \approx 10^{34}$ cm^{-2}s^{-1} $= 10$ nb^{-1}s^{-1} with a collision frequency of 40 MHz. The cross section for inelastic pp collisions can be approximated from data as $\sigma_{pp\,\mathrm{inel}} \approx 80$ mb for a center of mass energy of 13 TeV, as shown in Figure 3.5. Using these values, the average number of interactions in a single bunch

crossing can be estimated at approximately 20 interactions. The distribution of the average number of interactions in a single bunch crossing weighted to the luminosity for the Run-2 period is shown in Figure 3.6. Since each proton bunch is equidistantly spaced by 25 ns,

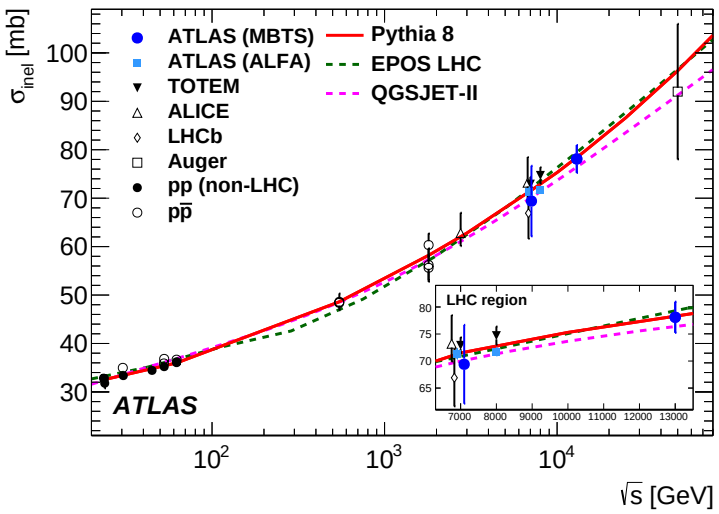


Figure 3.5: The inelastic proton-proton collision cross section measurements as a function of the center of mass energy for different experiments and overlayed with theory predictions. This figure is taken from [27].

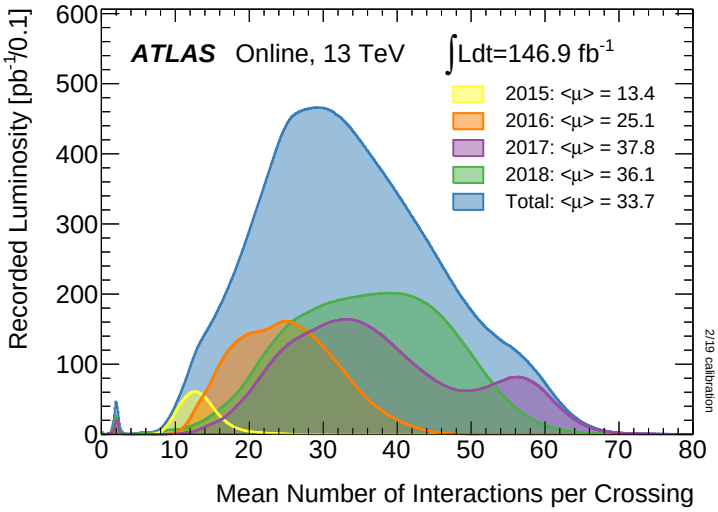


Figure 3.6: The average number of pp interactions in a single bunch crossing for the different years during Run-2, weighted by luminosity. This figure is taken from [26].

emissions from previous bunch crossing interactions may appear as part of the current bunch

crossing interaction due to technical limitations on the detector readout time. This effect is known as out-of-time pile-up. As the LHC increases its luminosity after each upgrade, the effects of pile-up will become a significant source of systematic uncertainty that analyses will need to consider.

3.2 The ATLAS Detector

The LHC provides the energetic collisions that allow us to probe the physics of the SM and potentially make new discoveries that will extend it. However, the collision of particles is only part of the job, as one needs to detect what is produced from the collisions and have the ability to recognize, select, and correctly reconstruct the events of interest. This is achieved with the different detectors that are placed along the LHC ring. The ATLAS detector, shown in Figure 3.7, is a general purpose detector that is made up of many specialized components that measure a wide range of signals from high energy particle interactions. These signals are used in the reconstruction process of pp collisions.

3.2.1 Particle Interactions with Matter

Converting the particles produced in pp collisions into physical signatures requires that the different specialized components of the ATLAS detector are made up of different materials. These materials must elecit specific types of interactions with the particles as they travel throughout the detector. An important factor to consider in the choice of these materials is their ability to contain the particles and their subsequent decays within a certain length of the detector. This ensures that the particles deposit all their energy into the detector for measurement while also shielding exterior detector components from interactions they are

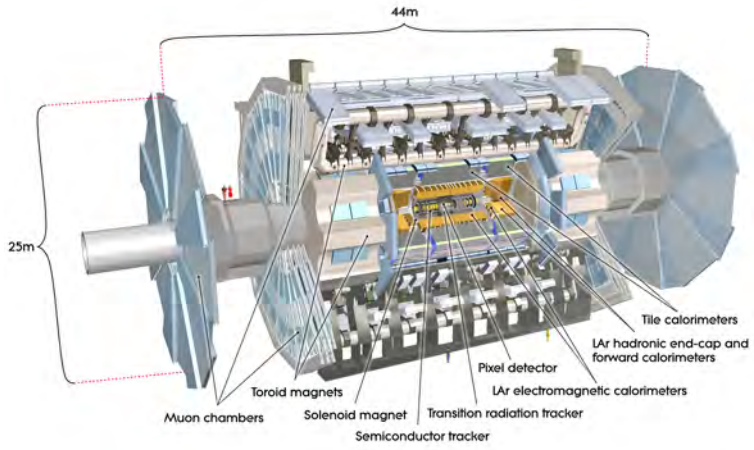


Figure 3.7: Schematic representation of the ATLAS detector showing its dimensions and different components. This figure is taken from [28].

not designed to withstand.

In a pp collision, particles that have electric or color charge can be produced, so the ATLAS detector employs materials that interact with these particles through the electromagnetic or strong force. The electromagnetic interactions usually take the form of bremsstrahlung radiation, which occurs when a charged particle is slowed down by the nuclei of the atoms in the detector material, or through ionizing radiation of the detector material. Electrically charged particles will thus lose energy due to the deceleration caused by these electromagnetic interactions. Photons can either directly ionize the detector material or decay into an electron-positron pair, which are subject to the deceleration process just described. The strong interactions take the form of inelastic nuclear collisions, which start a process known as hadronization, in which a hadron decays into more hadronic particles. Due to color confinement, it is energetically favorable for particles with color charge to stay in color-neutral bound states. The strong interactions that initialize the hadronization of a particle transform its kinetic energy into the creation of quark-antiquark pairs in order to

conform to color confinement. The hadronization process is eventually halted as the total kinetic energy is depleted. This process results in shower-like patterns of particle decay that are eventually reconstructed as objects known as jets (see 3.3.4).

An overview of the different particle interactions as they travel throughout a longitudinal cross-section of the ATLAS detector is shown in Figure 3.8. The innermost layer of the

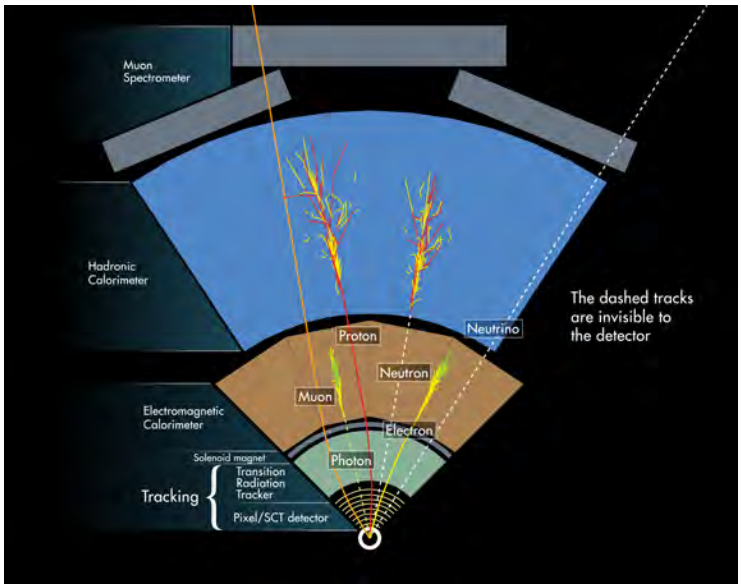


Figure 3.8: Schematic representation of the different detector components of ATLAS in a longitudinal cross-section. Example trajectories and interactions of particles produced in pp collisions with the different detector components are also shown. This figure is taken from [29].

ATLAS detector contains the tracking chamber, which provides information on the position of charged particles as they start to leave the interaction point. It is designed so that electromagnetic interactions are minimal in order for charged particles not to lose most of their energy. Immediately above the tracking chamber is the central solenoid, which exerts an axial magnetic field that curves the trajectory of charged particles. This allows us to identify the charge of a particle and its momentum based on the curvature that the trajectory takes (see 3.3.1). Above the central solenoid is the electromagnetic calorimeter, which maximizes the electromagnetic interactions so that particles like electrons or photons transfer all their

energy into the detector. The hadronic calorimeter is positioned a layer above the electromagnetic calorimeter and is designed to fully stop electrically neutral hadrons. Additionally, it also stops charged hadrons that do not deposit all their energy in the electromagnetic calorimeter. In both cases, the absorption of the energy of the particles by the detector is converted into electric signals, which are then used to reconstruct the underlying event. Finally, particles such as muons and neutrinos barely interact with the detector. Muons, on average, are produced with sufficient energy that makes them minimum-ionizing particles (MIPS), so they traverse the inner detector and calorimeter without depositing the majority of their energy in these detector components. Energy measurements of muons are delegated to the muon spectrometer, which forms the outermost layer of the ATLAS detector. Embedded within the muon spectrometer is a toroidal magnet system that further curves the trajectory of muons. This allows us to measure the energy of muons by measuring the curvature of tracks formed as muons ionize the material of the spectrometer. Neutrinos, on the other hand, are practically invisible to the detector, so they are inferred as missing energy during the event reconstruction process.

3.2.2 Detector Coordinate System

A coordinate system must be established in order to properly describe the kinematics of the particle interactions that are detected by ATLAS. This coordinate system must take into account the fact that the ATLAS detector has a cylindrical symmetry and that the emissions from particle interactions are spherically symmetric. The origin of the system is placed at the nominal interaction point (IP) of the detector with three perpendicular cartesian axes spanning from this point. The x-axis points towards the center of the LHC ring, the y-axis points upwards towards the sky, and the z-axis points along the accelerator beam line in the

direction that makes the coordinate system a right-handed system, as shown in Figure 3.9. The x-y plane is known as the transverse plane of the detector and is characterized by the

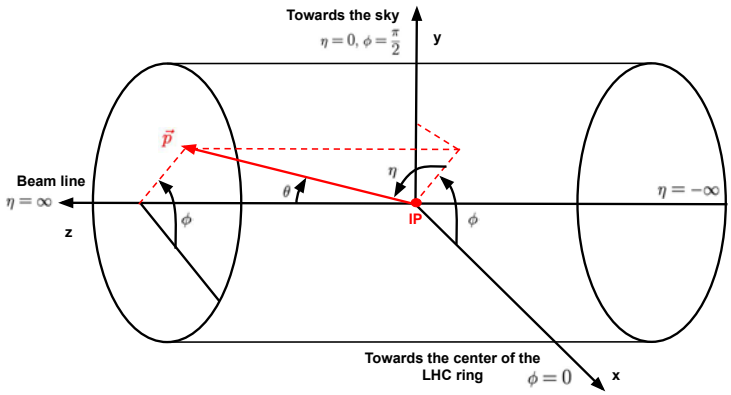


Figure 3.9: Schematic representation of the coordinate system used for the ATLAS detector. The relationships between angles and the cartesian axes are shown. The red solid line represents the three-momentum of a particle that is emitted from the IP.

azimuthal angle ϕ. This is the angle between the x-axis and a two-dimensional vector that starts from the IP and lies completely in the transverse plane. The azimuthal angle ranges between $[0, 2\pi)$, being 0 when the vector is parallel to the x-axis and $\pi/2$ when parallel to the y-axis. The polar angle θ is defined as the angle between the z-axis and any vector starting from the IP. This angle ranges between $[0, \pi]$, attaining the lower and upper bounds of the interval when the vector is parallel and anti-parallel to the z-axis, respectively.

Since the description of particle collisions takes place at the center of mass reference frame, which is boosted along the z-axis, a quantity known as the rapidity of a particle is often used instead of the polar angle. This choice is motivated by the fact that rapidity is invariant under boost transformations along the z-axis. Additionally, the production of particles is uniform with respect to rapidity. The rapidity is defined as

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right) \tag{3.5}$$

where E and p_z are the energy and the momentum component along the z-axis of a particle, respectively. In actual applications, an approximation of the rapidity known as pseudorapidity is used instead. The pseudorapidity is obtained in the limit of massless particles in Equation 3.5 and is applicable to the particles that are detected by ATLAS whose masses are negligible compared to their kinetic energy. The pseudorapidity is defined as

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right) \tag{3.6}$$

where θ is the polar angle. The pseudorapidity has values that range from $(-\infty, \infty)$, approaching $\pm\infty$ along the $\pm z$-axes and being 0 in the transverse plane. The momentum components of a particle can be expressed in terms of η and ϕ as

$$p_x = p_{\mathrm{T}}\cos(\phi),\ p_y = p_{\mathrm{T}}\sin(\phi),\ p_z = p_{\mathrm{T}}\sinh(\eta) \tag{3.7}$$

where p_{T} is the transverse momentum of the particle

$$p_{\mathrm{T}} = |\mathbf{p}|\sin(\theta) = \sqrt{p_x^2 + p_y^2} \tag{3.8}$$

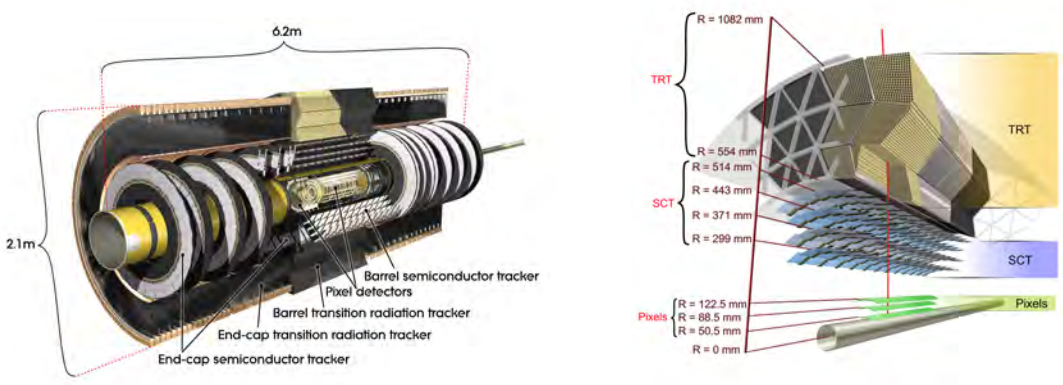
and $\mathbf{p}$ is the three-momentum of the particle. Other kinematic variables can be derived through standard relations involving the four-momentum of the particle. The angular distance between two particles in the detector can be expressed in terms of η and ϕ as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{3.9}$$

where $\Delta\eta$ and $\Delta\phi$ are the separations in pseudorapidity and azimuthal angle, respectively.

3.2.3 Inner Detector

The inner detector (ID) [30] is the innermost part of the ATLAS detector and encapsulates the beam line. It is designed to primarily measure the momentum of charged particles with the aid of a 2 T axial solenoidal field, in which all its subcomponents are immersed. The ID is also used to identify primary and secondary vertices (see 3.3.1) that are used to provide preliminary information for particle identification. The ID has three main subcomponents: the pixel detector, the semiconductor tracker (SCT) system, and the transition radiation tracker (TRT) system. Mechanically, the ID and its three subcomponents span a radius of 1.1 m from the beam line and a length of 6.2 m parallel to the beam line. The ID is split into three regions: a cylindrical barrel region, which provides a pseudorapidity coverage of $|\eta| < 1.2$, and two end-cap regions, which provide a coverage of $1.2 < |\eta| < 2.5$. A schematic representation of the ID and its subcomponents is shown in Figure 3.10.



(a) (b)

Figure 3.10: Schematic representation of the ID including the barrel and end-cap regions (a). A longitudinal cross-section of the different layers along the barrel region with their radial distances measured from the beam line is also shown (b). These figures are taken from [31].

3.2.3.1 Pixel Detector

The pixel detector is the innermost subsystem of the ID, consisting of silicon pixel detector modules. Being the closest component to the IP, it is designed to endure the intense radiation environment it is subject to for a lifetime of ten years. It is mainly used to provide tracking information and pattern recognition of short-lived particles, such as b quarks and τ. The pixel detector was originally designed with three barrel layers and three end-cap disk layers on each side. During the upgrade period between Run-1 and Run-2, a fourth layer directly encompassing the beam line, known as the insertible b-layer (IBL) [32], was introduced to compensate for the radiation damage sustained by the other layers and to reduce readout inefficiencies due to pile-up at higher luminosities. The IBL has the largest granularity in the barrel region of the pixel detector, with approximately 6×10^6 silicon pixels with a $R\phi$-z resolution of 50×250 μm^2. The three layers above the IBL have a smaller granularity with a pixel size of 50×300 μm^2 in the $R\phi$-z plane and a total of approximately 67.2×10^6 pixels. The three end-cap disk layers are positioned at a distance $|z|$ of 495 mm, 580 mm, and 650 mm on each side from the IP, with each disk containing approximately 2.2×10^6 pixels. Overall, the ID contains approximately 86.4×10^6 silicon pixels. The charged particles that are produced at the IP interact with the pixel detector by ionizing the silicon, producing electron-hole pairs, which are then read as an electric current by a sensor.

3.2.3.2 Semiconductor Tracker

The SCT system envelopes the pixel detector and is designed to continue tracking charged particles by measuring their momentum and vertex position and providing pattern recognition of particles. The barrel component of the SCT consists of four layers of silicon microstrip detectors with a size of 6.36×6.40 cm^2 and a resolution of 16×580 μm^2 in the $R\phi$-z plane

per silicon detector. Each layer is positioned at a radii of 300 mm, 373 mm, 447 mm, and 520 mm. The end-cap component of the SCT consists of nine disks with similar silicon modules as the barrel component. Charged particles interact with the SCT in the same way they do with the pixel detector.

3.2.3.3 Transition Radiation Tracker

The TRT is the outermost subsystem of the ID and consists of straw tube detectors with a diameter of 4mm. Each tube encapsulates an anode wire and is filled with a gas mixture composed of 70% Xe, 27% CO_2, and 3% O_2. Charged particles that pass through the straw tubes ionize the gas, and the resulting electrons drift towards the anode, which is then recorded as a signal. The TRT continues to provide tracking information of charged particles, but its main function is to assist in the pattern recognition of particles. This is achieved by including a radiator material between individual straw tubes, which acts as a medium boundary that forces particles to emit transition-radiation photons. These photons then ionize the gas inside the tubes, with the emission rate of transition-radiation photons being characteristic of a particle at a given momentum. Mechanically, the TRT consists of 50000 straw tubes in the barrel region that are placed parallel to the beam line and 320000 straw tubes that are placed at the end-caps in a radial configuration and distributed across 18 wheel structures on each side.

3.2.4 Calorimetry

The calorimetry system of the ATLAS detector encapsulates the ID and solenoid magnet and is located a layer below the muon spectrometer. It is composed of two main subsystems: the liquid argon (LAr) calorimeter and the tile hadronic calorimeter (TileCal). The primary

function of these two subsystems is to stop incoming charged particles and hadrons that are exiting the ID in order to measure their total energy. This is achieved by having different subcomponents in each subsystem with varying material compositions and lengths that are tailored to maximize the interactions between the detector and particles. Together, both subsystems provide a coverage of $|\eta| < 4.9$. In addition to measuring the energy of particles, the calorimeter system also ensures good measurement of the missing transverse energy (E_T^{miss})[3] as a consequence of its wide η coverage and subcomponent material thicknesses. This is important in indetifying many physics signatures, such as the production of neutrinos. A schematic representation of the ATLAS calorimetry system with all its subcomponents to be discussed in the following subsections is shown in Figure 3.11.

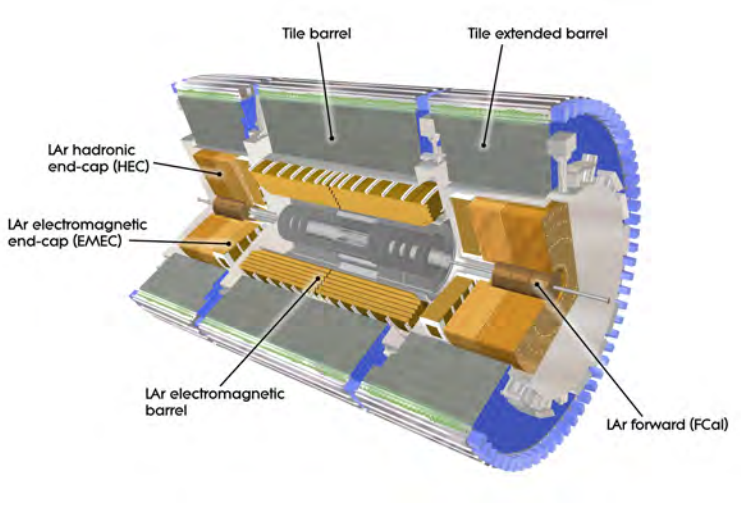


Figure 3.11: Schematic representation of the calorimetry system of the ATLAS detector with its different subcomponents. This figure is taken from [33].

[3]This value is a scalar related to the momentum imbalance in the transverse plane (see 3.3.5).

3.2.4.1 Liquid Argon Calorimeter

The LAr calorimeter provides both electromagnetic and hadronic energy measurement capabilities through the use of four subcomponents, which are distributed in the barrel and end-cap regions. The barrel component consists only of the LAr electromagnetic (EM) calorimeter, which provides a pseudorapidity coverage of $|\eta| < 1.475$. The end-cap region contains the three remaining components, which are distributed in two coaxial wheels and two longitudinal wheels. The outermost coaxial wheels contain the LAr electromagnetic end-cap (EMEC) calorimeter, which is the innermost longitudinal wheel and provides a covering range of $1.375 < |\eta| < 2.5$, and the LAr hadronic end-cap (HEC) calorimeter, which is positioned longitudinally after the EMEC and provides a covering range of $1.5 < |\eta| < 3.2$. The innermost coaxial wheel contains the LAr forward calorimeter (FCal), which covers the high pseudorapidity region of $3.1 < |\eta| < 4.9$.

The EM and EMEC components consist of alternating layers of lead absorption plates and LAr. Charged particles or photons interact with the lead plates either by bremsstrahlung or ionizing radiation, which produces electromagnetic showers that ionize the LAr medium providing electrical signals for measurements. To fully contain the electromagnetic showers, the EM and EMEC are designed to have thicknesses of at least $X_0 > 22$ g.cm^{-2} and $X_0 > 24$ g.cm^{-2} radiation lengths (X_0), respectively. The radiation length is an inherent property of the material that is defined as the average distance at which an electron loses its energy by a factor of $1/e$ while traversing the material.

The HEC component consists of two independent wheels per end-cap that have a similar absorber-LAr structure as the LAr electromagnetic components but use copper plates instead of lead. The choice of copper over lead as the absorption material takes into account the fact

that hadronic showers are longer than electromagnetic showers due to the nuclear interaction length (λ_I) of hadrons being usually larger than the radiation length X_0 by an order of magnitude.

The FCal provides both electromagnetic and hadronic calorimetry with an approximate thickness of 10 λ_I. It is split into three modules: the innermost module uses copper as its absorption material and is optimized to provide electromagnetic measurements, while the two remaining modules use tungsten as their absorption material for its higher density and are optimized to provide hadronic energy measurements.

3.2.4.2 Tile Hadronic Calorimeter

The TileCal is designed to simultaneously measure the energy of hadronic interactions and halt incoming hadrons from leaving the detector. It consists of a barrel component that encapsulates the LAr barrel calorimeter. The TileCal is sectioned into a central long barrel of length 5.8 m that covers a range of $|\eta| < 1.0$, and two extended barrels of length 2.6 m that cover the range $0.8 < |\eta| < 1.7$. The TileCal extends radially from an inner radius of 2.28 m up to an outer radius of 4.25 m and is segmented into three layers that are approximately 1.5, 4.1, and 1.8 λ_I thick in the central barrel region, and 1.5, 2.6, and 3.3 λ_I thick in the extended barrel region. It uses steel as the absorption material and plastic scintillating tiles as the active medium. The particles that are produced in hadronic showers interact with the scintillators producing photons which are read out using photomultiplier tubes.

3.2.5 Muon Spectrometer

The muon spectrometer (MS), shown in Figure 3.12, is the outermost layer of the ATLAS detector. It is designed to provide tracking and momentum measurements of muons, which

hardly interact with the ATLAS calorimeter due to muons being MIPS. It is composed of two subsystems: the muon precision chambers, which contain the monitored drift tubes (MDT) and the cathode strip chambers (CSC) subcomponents; and the muon trigger chambers, which contain the resistive plate chambers (RPC) and the thin gap chambers (TGC) subcomponents. The subcomponents provide the aforementioned muon measurements with the assistance of three superconducting air-core toroidal magnets: one positioned in the central barrel region of $|\eta| < 1.4$, and one at each end-cap covering the regions $1.6 < |\eta| < 2.7$. The central barrel toroidal magnet generates a magnetic field of 0.5 T, while the end-cap toroidal magnets generate a 1 T magnetic field. The magnetic fields are oriented along the azimuthal direction, which curves the trajectories of muons towards the different subcomponents of the MS.

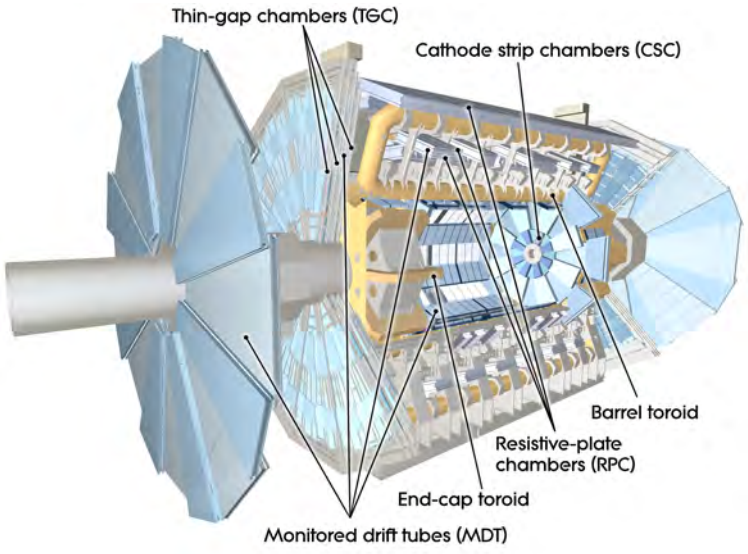


Figure 3.12: Schematic representation of the muon spectrometer of the ATLAS detector with its different subcomponents. This figure is taken from [34].

3.2.5.1 Muon Precision Chambers

The muon precision chambers are designed to provide precise tracking information and momentum measurements of muons at the cost of a higher processing time. The MDT are aluminum tubes that contain a central tungsten-rhenium wire. Each tube is filled with a gas mixture that is mostly composed of argon. The MDT are arranged into chambers, providing a coverage of $|\eta| < 2.0$ in the central region and up to $|\eta| < 2.7$ in the end-cap regions. The CSC is composed of multiwired drift tubes, which contrasts to the monowire design of the MDT, in order to cope with the demanding particle flux in the high pseudorapidity region $2.0 < |\eta| < 2.7$ that it was designed to cover. Muons interact with the MDT and CSC by ionizing the gas inside, which produces electrons that are read out by the central anode wires.

3.2.5.2 Muon Trigger Chambers

The muon trigger chambers are designed to primarily provide well-defined p_{T} thresholds for muons that are used by the ATLAS trigger system. However, it also provides information on bunch-crossing identification and complements the muon tracking measurements performed by the muon precision chambers, as they are orthogonal in direction. The RPC component of the trigger chamber consists of parallel electrode plates that are separated by a gas gap to be ionized by muons passing through. The RPC is located in the barrel region of the detector and provides a covering of $|\eta| < 1.05$. The TGC is composed of multiwired drift tubes and is positioned at the end-cap regions, providing a coverage of $1.05 < |\eta| < 2.7$ that is used for tracking measurements, while the triggering decision information is restricted to $1.05 < |\eta| < 2.4$.

3.2.6 Magnet System

The ATLAS detector magnet system [35] is split into three subsystem: a central solenoid, a barrel toroid, and the end-cap toroids. As previously discussed, the magnetic fields generated by these magnets aid in measuring the momentum of charged particles. A schematic representation of the magnetic field lines produced by the magnet system is shown in Figure 3.13.

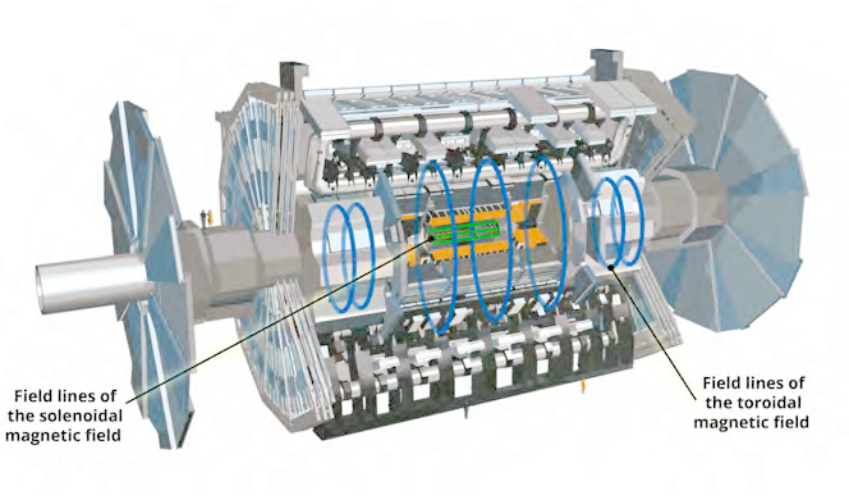


Figure 3.13: Schematic representation of the magnetic field lines produced by the ATLAS detector magnet system. This figure is taken from [36].

The central solenoid is located between the ID and the electromagnetic calorimeter. It has a longitudinal length of 5.3 m, an inner diameter of 2.44 m, and an outer diameter of 2.63 m. The solenoid consists of a single-layer coil of superconducting wire made of an aluminum-stabilized niobium-titanium-copper alloy. The superconducting wire is wound 1173 times around the coil in a supporting cylinder and is designed to operate with a 7.6 kA current.

This allows the central solenoid to provide a nominal 2 T axial magnetic field, although it can produce a field with a peak strength of 2.6 T.

The toroidal magnets are embedded within the MS and are mechanically split into a barrel toroid and two end-cap toroids. The barrel toroid has a longitudinal length of 25.3 m, an inner diameter of 9.4 m, and an outer diameter of 20.1 m. A single end-cap toroid has a longitudinal length of 5 m, an inner diameter of 1.65 m, and an outer diameter of 10.7 m. All three toroids consist of 8 individual air-cored coils that are positioned radially around the detector. Each individual coil contains superconducting wires made from the same material as the superconducting wire in the central solenoid, although the ratio of the different elements varies between the barrel and end-cap toroids. The wires are wound 120 times around each individual coil of the barrel toroid, and 116 times around each individual coil of the end-cap toroids. The operating current is 20.5 kA in the case of the barrel toroid, and 20 kA in the case of the end-cap toroids. This allows the toroids to produce peak magnetic field strengths of 3.9 T and 4.1 T that are directed azimuthally for the barrel toroid and end-cap toroids, respectively.

3.2.7 Trigger and Data Acquisition System

At a nominal collision rate of 40 MHz, it is unfeasible to store the data of all events that are detected by ATLAS. The trigger and data acquisition (TDAQ) system of the ATLAS detector [37] is designed to select events based on certain triggering requirements that are deemed interesting for the different analyses carried out by the ATLAS experiment. The TDAQ system is split into three subsystems. The first subsystem is the Level-1 (L1) trigger, which reduces the initial event rate of 40 MHz down to approximately 100 kHz using direct information from the detector hardware. The second subsystem is the Level-2 (L2) trigger,

which is seeded by the information provided from the L1 trigger to further reduce the event rate down to approximately 3.5 kHz using full granularity and precision from the detector hardware. The third subsystem is the event filter, which takes the information from the L2 trigger and processes it using offline reconstruction algorithms to further reduce the event rate down to approximately 200 Hz. Events that pass the event filter are then written to disk and stored for offline reconstruction.

The L1 trigger searches for events that have patterns of potential candidate high-p_{T} objects, such as electrons, muons, hadronically decaying τ, photons and jets. The L1 trigger also searches for events that have large E_T^{miss} and total transverse energy. The L1 trigger is subdivided into three components: the L1 muon trigger, which identifies potential muon candidates from the hardware information obtained from the MS; the L1 calorimeter trigger (L1Calo), which identifies the remaining aforementioned objects using information from the ATLAS calorimetry system; and the central trigger processor (CTP), which combines the information obtained from the L1 muon and L1Calo triggers to decide whether an event should be selected for further processing by the L2 trigger or not. Additionally, the L1 trigger defines regions of interest (RoI), which include data such as the spatial information in the η-ϕ plane of interesting patterns it has identified from a single detector component, the type of pattern identified, and the passing criteria imposed by the trigger.

The L2 trigger further analyzes the information stored in the RoI using full granularity and precision from the data of all the detector components in the RoI. A trigger menu system that contains individual items from the L1 trigger is used by the L2 trigger to accept or reject events. Events that are accepted by the L2 trigger are built into a single data structure that is sent to the event filter.

The event filter uses the event data structure constructed by the L2 trigger and processes

it with reconstruction algorithms to make the final decision on whether to keep the event or not. Events that pass the event filter are then classified as events to be used for physics analysis or performance measurements, which are then saved in separate data streams.

3.3 Object Reconstruction and Calibrations

In the preceding subsections of this chapter, the preparation of pp collisions at the LHC and how the ATLAS detector is designed to detect processes that arise from these collisions were discussed. Interactions between particles and the different subcomponents of the ATLAS detector are converted into electrical signals. These signals are then processed individually on a subcomponent basis to determine the region in the detector where the interaction took place and the amount of energy deposited by the particles. Events that pass the selection criteria imposed by the ATLAS trigger system are then stored and used in the different analyses and calibrations performed by the ATLAS collaboration. The last step to get events ready for these tasks is to reconstruct and calibrate the different physics objects corresponding to particle detections.

As can be observed in Figure 3.8, the identification of a particle cannot be based solely on a single component of the detector. As an example, both electrons and photons deposit most of their energy in the EM calorimeter as they produce electromagnetic showers. Thus, at the calorimeter level, these two particles are practically indistinguishable. However, when the information of the tracking system is considered, one can use the fact that electrons do interact with the ID while photons do not. Combining these two detector signatures allows us to distinguish between electrons and photons. Thus, the identification of particles produced in events from reconstructed objects is an algorithmic, multi-step procedure that

combines information from all relevant detector components.

Finally, the reconstruction of objects must take into account detector effects. For example, a detector component could have mechanical gaps where there is no material in which particles can deposit their energy. Additionally, some detector components can accumulate damage from the radiation produced by particle interactions. These effects can result in inaccuracies between the energy that is deposited by a particle and what is measured by the detector component. For this reason, reconstructed objects are calibrated in order to address detector effects.

3.3.1 Tracks

Charged particles that pass through the ID interact with its different layers creating hits, which are then reconstructed into tracks that are associated with the trajectories of particles. Since the ID is immersed in a solenoidal magnetic field, the tracks will follow a helicoidal trajectory. The charge of the particles can be determined from the direction of the curvature of the tracks. Additionally, the momentum of the particle can be measured using the curvature since these two quantities are inversely proportional.

The track reconstruction at the ID consists of three subprocesses. The first subprocess takes the data from the pixel and SCT detectors and clusters it into spatial coordinates. The second subprocess consists of applying track-finding algorithms [38, 30] that form track seeds using the spatial coordinates from the pixel detector and the first layer of SCT detector as inputs. These track seeds are then extended through the remainder of the SCT, forming track candidates that are fitted to the track clusters in the SCT. Track candidates that are found to be outliers are removed, while those that are deemed valid are then extended through the TRT to resolve track direction ambiguities. Finally, the extended track candidates are fitted

using all the information from the ID to evaluate the fit quality. A complementary track-finding algorithm, known as back-tracking, is also applied as part of the second subprocess. This algorithm starts from the outermost layer of the ID and works its way through the ID. The purpose of this algorithm is to find leftover tracks that can be extended down into the SCT and pixel detector in order to identify particles that undergo conversion or decay while traversing the ID. The third subprocess consists of applying vertex finding algorithms using tracks as inputs. Vertices are defined as tracks that have a common spatial origin and are close together. Vertices that have their spatial origin near the beam line are denoted as primary vertices. Secondary vertices correspond to particle conversion and decay processes that take place far from the beam line.

3.3.2 Electrons and Photons

The majority of electrons and photons are fully stopped by the electromagnetic calorimeters in the detector. The representation of the energy of these particles is constructed using the energy measurements obtained from this calorimetry subsystem. This process is performed by applying a sliding-window algorithm [39] that scans the different layers of the electromagnetic calorimeter in η-ϕ space using a fixed-size window. The transverse energy that is deposited at a given window interval is added to form candidate energy clusters. The candidate clusters are rejected if their total transverse energy is below the noise threshold of the calorimeter. Clusters that pass the noise threshold criteria are then compared with the reconstructed tracks from the ID in order to determine if these signatures should be reconstructed as an electron or photon. If the energy cluster spatially matches a reconstructed track belonging to a primary vertex, then both objects are reconstructed as an electron. If the energy cluster matches a reconstructed track belonging to a secondary vertex, then both

objects are reconstructed as a photon that decayed into an electron-positron pair. Finally, if the energy cluster does not match any tracks, then it is reconstructed as a photon. In the cases where the energy cluster is matched to a track, the momentum of the reconstructed object is recalculated using the cluster energy and the momentum measurement obtained from the ID tracks.

At the detector level, other particles can produce detector signatures that mimic the signatures of the particle of interest one wants to reconstruct, which can result in the mis-reconstruction of an object. Photons have a low mimic rate due to their unique detector signature. Charged particles will often produce a complete set of tracks in the ID, which makes their signature inconsistent with photons. Particles that carry no charge and interact with the detector, such as neutral hadrons, deposit most of their energy in the hadronic calorimeter, which also makes their signatures inconsistent with those of a photon. Electrons, on the other hand, can be mimicked by several particles. For example, charged pions that decay while traversing the electromagnetic calorimeter can mimic the signature of an electron if the pion decays deposit all of their energy in this detector component, and the tracks associated with the pion are consistent with a primary vertex. Although mimicking detector signatures from electrons is a rare process, the mimic rates can start to become relevant at higher detector luminosities. For this reason, the reconstructed electron candidates are classified as loose, medium, and tight based on selection criteria that lower the mimic rate but also lower the acceptance rate moving from loose to tight.

3.3.3 Muons

The process of reconstructing muons is straightforward compared to other particles due to muons being MIPS and lacking color charge. At the detector level, muons are identified as

tracks in the ID and MS, with the possibility of small energy deposits in the electromagnetic calorimeter. There are four muon reconstruction algorithms [40] that are used based on the availability of the detector signatures. The combined (CB) muon algorithm is used when both the ID and MS tracks are available and have a good spatial match. The segmented-tagged (ST) muon algorithm is used when the ID tracks are fully reconstructed but only partial track segments are available from the MS. In this scenario a muon has low p_{T} or passes through a region of the MS where a single layer is hit. The calorimeter-tagged (CT) muon algorithm is used when there is no information available from the MS but an ID track is matched to an energy cluster in the electromagnetic calorimeter that is consistent with a MIP. Finally, the extrapolated (ME) muon algorithm is used when a full track from the MS is available but there is no ID track. This algorithm is used in the region $2.5 < |\eta| < 2.7$ where there is coverage from the MS but not from the ID.

Although very few particles besides muons and neutrinos do reach the MS, there are certain particles that can mimic muon signatures. An example of such particles are charged pions, which will leave tracks in the ID, traverse the calorimeters if they are produced with a lot of energy, and most likely decay into a muon and neutrino once inside the MS.

3.3.4 Jets

As mentioned in subsection 3.2.1, the particles that have color charge initiate a process known as hadronization due to color confinement. This process starts as soon as quark-antiquark pairs are created and start to separate due to their large momentum. The result of this is the creation of additional quark-antiquark pairs as soon as it becomes energetically favorable in order to conform to color confinement. The inelastic nuclear collisions that occur in the hadronic calorimeter further elicit this process from hadrons. The hadronization process

culminates when all the kinetic energy of the individual quarks is depleted, resulting in the creation of shower-like structures of energy deposits in the calorimeter. The degree of collimation of these showers along the direction of the original particle that initiated the hadronization process is dependent on the p_{T} of the original particle. In addition to quarks, it is also possible to produce other particles during hadronization, such as gluons, photons, electrons, and muons.

The energy deposits that result from the hadronization process are reconstructed as objects known as jets. There are different types of algorithms that are used to reconstruct jets, each yielding jets that have varying properties that are suitable for different applications. An important application of jets in analyses is to identify short-lived particles, such as top quarks, W/Z bosons, and Higgs bosons, that initialize the hadronization process through their decays. This is done with the use of dedicated algorithms, known as jet taggers, that employ different identification techniques based on the target particle to be identified. The concept of jet taggers will be discussed in more detail in Chapter 5.

The jet reconstruction process starts by reconstructing the energy deposits from the hadronization showers into objects known as topological energy clusters [39] (topoclusters). This is done with an algorithm that clusters neighboring calorimeter cells based on whether the total energy of the cluster exceeds a threshold defined on the expected noise of the cells. Unlike the reconstruction of electrons and photons, which uses only the energy deposits in the electromagnetic calorimeter, the reconstruction of topoclusters uses both the electromagnetic and hadronic calorimeters in order to take into account the production of other particles, such as photons and electrons, during the hadronization process.

Once the topoclusters have been built, they are clustered into jets using a clustering algorithm that combines the spatial and energy information of the topoclusters. The clustering

algorithm is summarized in the following steps:

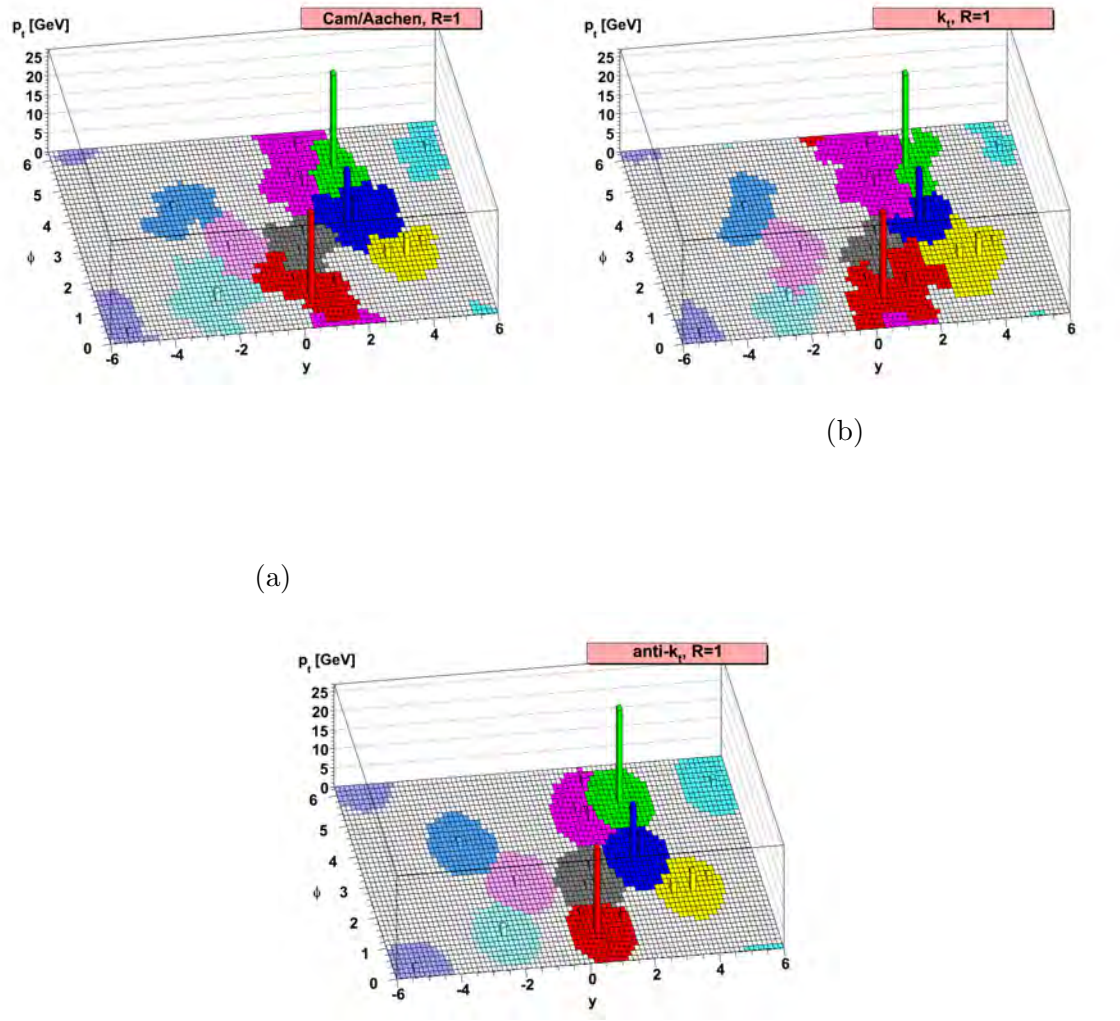
1. Given two topocluster labeled as i and j, their relative momentum-weighted distance $d_{i,j} = \min\{p_{\mathrm{T},i}^{2n}, p_{\mathrm{T},j}^{2n}\}\Delta R_{i,j}^2/R^2$ is calculated. The parameter n determines the momentum dependence of the clustering algorithm, while the parameter R defines the catchment area of the jet.

2. Calculate $d_{i,j}$ for all possible topocluster pairs.

3. Determine $d = \min\{d_{i,j}, d_k\}$, where $d_k = p_{\mathrm{T},k}^{2n}$ is the momentum weight of an individual topocluster k. If $d = d_{i,j}$ for a pair of topoclusters (i,j), then both topoclusters are combined. If $d = d_k$ for a single topocluster k, then the topocluster is considered to be a jet and is removed from further consideration in the clustering algorithm.

4. Repeat the algorithm until all topoclusters are combined into jets.

The most common choices of the parameter n are 0, 1, and -1, which are known as the Cambridge-Aachen (CA) [41], the k_T [42], and the anti-k_T [43] algorithms, respectively. The CA algorithm ignores the topocluster momentum, making it a purely geometric clustering algorithm. The k_T algorithm has the effect of clustering first the low p_{T} topoclusters, while the anti-k_T algorithm prioritizes high p_{T} topoclusters. Figure 3.14 shows the outcome of these three algorithms for an example set of energy deposition. As can be observed from this figure, the anti-k_T algorithm tends to produce jets with circular shapes that have a radius approximately equal to the parameter R. For this reason, this parameter is mostly referred to as the jet radius, a convention that will be adopted throughout the remainder of this thesis. The circular shape of anti-k_T jets is attributed to clustering the most energetic topoclusters first, which defines a stable centroid of the jet. The remaining topoclusters that



(a) (b) (c)

Figure 3.14: Result of applying the Cambrige-Aachen (a), the k_T (b), and the anti-k_T (c) clustering algorithms in the y-ϕ plane as a function of the topocluster p_{T} with parameter $R = 1.0$. Each uniquely colored cluster represents a single jet. This figure is taken from [43].

are added to the jet during the clustering process accumulate around the centroid. A good approximation to determine if a jet captures all the subsequent decays of the particle that initiated the hadronization process within its radius is to define the jet radius as

$$R = \frac{2m_{\text{particle}}}{p_{\text{T particle}}} \tag{3.10}$$

Jets that are used in ATLAS analyses are usually reconstructed using a radius parameter of $R = 0.4$, known as small-R jets, or $R = 1.0$, known as large-R jets. Small-R jets are designed to capture the hadronization of a single non-massive quark and the radiation of gluons. Small-R jets are often used as inputs to flavor tagging algorithms [44, 45], which identify jets that originate from the hadronization of b and c quarks. On the other hand, large-R jets are designed to capture the hadronization of heavier particles such as the top quark and hadronically decaying W/Z and Higgs bosons. For this reason, large-R jets are used as inputs to tagging algorithms dedicated to identify these heavier particles.

Another type of jet reconstruction used in ATLAS analyses consists of using small-R jets as inputs to a reclustering algorithm that combines them into a larger jet known as a reclustered (RC) jet. Unlike large-R jets, which require calibrations to their mass and energy (see 3.3.7), RC jets do not require additional calibrations since they are built from calibrated small-R jets. The RC jets are usually constructed using the anti-k_T algorithm and can have a fixed or variable radius parameter. For fixed-radius RC jets, the radius parameter is set to $R = 1.0$, while for variable-radius RC jets, the radius parameter is set to $R = \rho/p_{\text{T}}$, where ρ is an input parameter that controls the evolution of the effective size of the RC jet [46]. This shape flexibility that variable-radius RC jets offer allows them to capture the decays of boosted particles in a wide p_{T} range. For this reason, the variable-radius RC jets are used

as the inputs to the tagging algorithms used in the VLQ searches presented in Chapter 6.

3.3.5 Missing Transverse Energy

The sources of missing transverse energy (E_T^{miss}) [47] can be attributed to particles that exit the detector without interacting, such as neutrinos, and object misreconstruction, such as reconstructing a jet with mismeasured energy. Since protons that are traveling along the beam line have net zero momentum in the transverse plane prior to colliding[4], the total transverse momentum from all particles produced in the collision must be zero by conservation of momentum. If the total transverse momentum after the collision is not zero, then the corresponding event has a source of E_T^{miss}. The amount of E_T^{miss} is quantified as

$$E_T^{\text{miss}} = \sqrt{\left(p_x^{\text{miss}}\right)^2 + \left(p_y^{\text{miss}}\right)^2} \tag{3.11}$$

where p_x^{miss} and p_y^{miss} are the components of a vector in the transverse plane that quantify the momentum imbalance. This vector also has an associated azimuthal angle that indicates the direction of the net momentum imbalance and is given by

$$\phi^{\text{miss}} = \arctan\left(\frac{p_y^{\text{miss}}}{p_x^{\text{miss}}}\right) \tag{3.12}$$

3.3.6 Tau Leptons

Tau leptons present one of the hardest experimental challenges when trying to reconstruct them from detector signatures due to their short lifetime and possible decay modes, which

[4]This is not quite true since the net momentum of the quarks inside protons can be non-zero. The momentum components of quarks follow a distribution, albeit narrowly centered around zero.

are shown in Figure 3.15. In all decay scenarios the tau always produces a neutrino, which

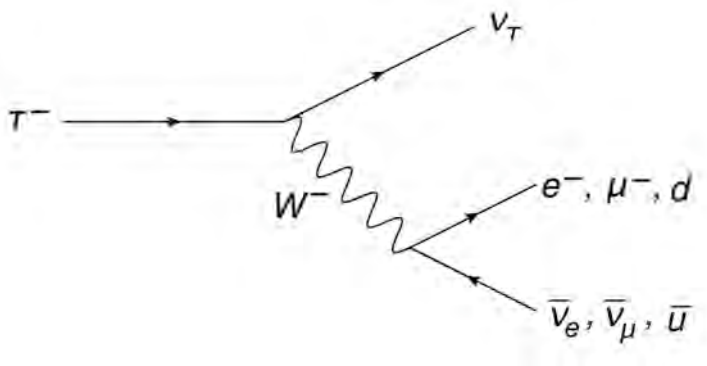


Figure 3.15: Feynman diagram depicting the different decay modes of a tau.

implies that the full energy of a tau cannot be fully reconstructed. In the case of a leptonically decaying tau it is very difficult to distinguish this process from the detector signatures of individual electrons and muons. For a hadronically decaying tau, it is possible to reconstruct and identify the tau from a combination of narrow calorimeter energy clusters and track segments that are consistent with a secondary vertex.

3.3.7 Calibrations

As previously mentioned, the object reconstruction procedure must take into account detector effects so that objects accurately represent the result of the interactions between the detector and particles produced in *pp* collisions. Examples of sources of these effects include, but are not limited to: differences in the detector response due to material variability within a detector subsystem; transitions between different detector technologies, such as different component resolutions; and detector damage due to prolonged radiation exposure. Addition-

ally, the large instantaneous luminosity and pile-up conditions in which the ATLAS detector operates can also have effects on the object reconstruction process by introducing excess energy from other events, which can result in the mismeasurement of an object. In order to compensate for these effects, calibrations are derived for both Monte-Carlo (MC) simulation and data independently. These calibrations are designed to reduce the inaccuracy between the energy measured by the detector and the actual energy that was deposited by particles. The calibrations are accurate by design but can have large variations between MC and data. These variations originate from the differences between the actual detector response and the imperfections in the simulation of the interactions between particles and the detector components. Thus, the differences between the calibrations that are applied to MC and data give rise to sources of systematic uncertainties that need to be considered in analyses.

The calibrations that are applied to jets are some of the most important compared to other physics objects. This is in part due to the complexity of jet reconstruction and the inherent randomness of the hadronization process. Since hadronization is random, jets can have large variations in their particle composition, which results in large variations in detector responses between jets. Additionally, jet calibrations rely on other well-calibrated objects such as electrons and photons. For the stated reasons, the systematic uncertainties that are associated with jet calibrations are a very important source of uncertainty in analyses. The four main types of calibrations applied to jets are on the jet energy scale (JES) and resolution (JER), and the jet mass scale (JMS) and resolution (JMR). A more detailed description of these calibrations, how they are derived, and their associated uncertainties can be found in [48, 49, 50].

These calibrations are first applied to jets from MC. The events from MC simulation contain the information of all stable particles that are produced in a given event, which

is known as the truth information. The truth information includes the four-momentum of the particles and their decay chain, which can be used to reconstruct the full decay tree of the event. Jets can be reconstructed in two ways using the truth information. The first type of jets, known as truth jets, are obtained by applying the clustering algorithm described in 3.3.4 to the four-momentum of the stable particles in the MC record. The second type of jets, simply known as reconstructed jets, are obtained by reconstructing them from simulated detector responses from the truth information particles. Since these simulations are imperfect, the reconstructed jets are calibrated so that their energy and mass match those of truth jets.

The JES and JMS calibrations are designed to correct the energy and mass measurements of jets from sources that can affect these measurements. First, a calibration is applied to correct the jet origin so that the direction of the jet matches that of its primary vertex. Next, a calibration is applied in order to remove effects from pile-up contamination, which can result in the mismeasurement of the jet properties. The next calibration consists of correcting the JES and JMS from the differences between the calorimeter response and the truth jet scales. The final calibration applied to MC jets consists of reducing the dependence of the jet constituent flavor composition on the detector response. An in-situ calibration is applied to jets from data to correct any remaining inaccuracies between data and MC. The JER and JMR calibrations are designed to correct the variance of the jet energy and mass measurements in MC so that they match that of data. These calibrations take into consideration effects such as pile-up contamination, variability in the detector material, and energy deposition in passive detector components.

Chapter 4

Processes of Interest and Data Selection

In this Chapter, the relevant physics processes that are studied in Chapter 5 and Chapter 6 are described. These processes are classified as signal processes, which contain physical signatures of interest, and background processes, which can mimic signal processes in different ways. Both studies use data recorded by the ATLAS detector and MC simulations, which are obtained from the theoretical predictions of the signal and background processes. The details of the MC samples used in these studies are decribed in Appendix A. Events from data and MC go through the object reconstruction and calibration procedures described in section 3.3. The event selection criteria that are used in each study are also described in this Chapter. These are kinematic requirements that are imposed on the reconstructed physics objects from data and MC events and are designed to simultaneously maximize the acceptance of signal-like processes and the rejection of background-like processes.

4.1 Jet Tagging Studies

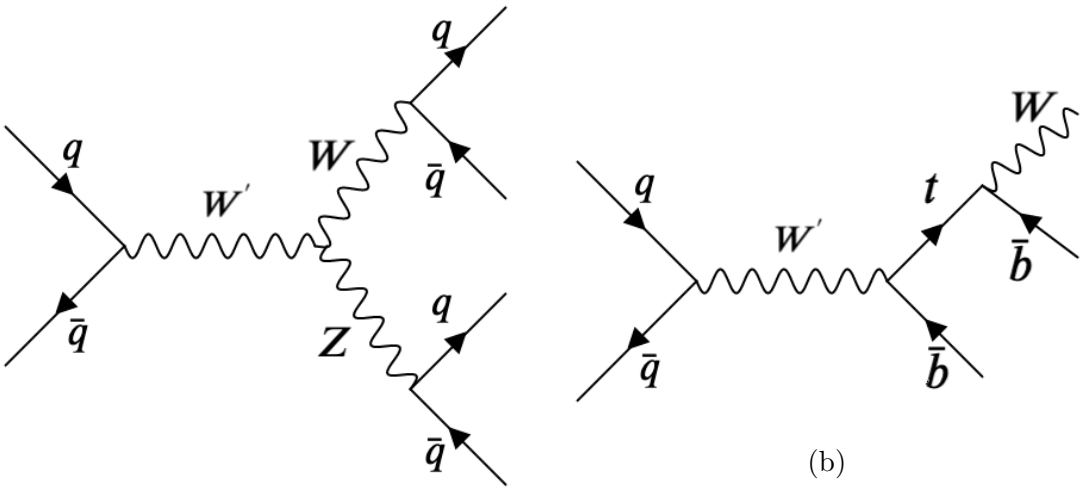
The studies presented in Chapter 5 are dedicated to the optimization and calibration of jet tagging algorithms, which are designed to identify a jet to the particle that originated it. The signal processes in these studies are events that contain a jet that was produced by a

particle of interest. These jets will be referred to as signal jets. Background processes, on the other hand, are events that lack the particle of interest but contain jets that mimic signal jets, which will be referred to as background jets. However, as previously discussed, since hadronization is a stochastic process and jet reconstruction is not a fully efficient process, it is possible that signal jets get mistaken for background jets. This can happen if the kinematic features of the jet are mismeasured or misreconstructed. Both signal and background jets are used as inputs to the tagging algorithms in order to optimize and calibrate the jet taggers, as will be discussed in Chapter 5. The data sample used in the jet tagging studies comes from data recorded from pp collisions at a center of mass energy of 13 TeV in the period 2015-2017 and corresponds to an integrated luminosity of 80.5 fb^{-1}.

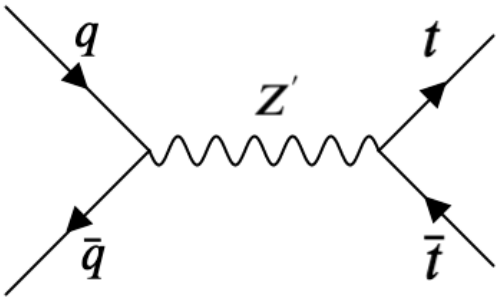
4.1.1 Signal Processes

The jet taggers studied in Chapter 5 are designed to identify jets that originate from the decays of boosted top quarks and W bosons. The taggers are optimized using MC samples from the BSM Heavy Vector Triplet (HVT) model [51] as signal processes. This model predicts the existence of two heavy gauge bosons: the W' and the Z'. The W' production processes considered are the $W' \to WZ \to q\bar{q}q\bar{q}$ decays, which serve as a source of signal W jets, and the $W' \to tb$ decays, which serve as a source of jets arising from top quark decays. The Z' production process considered is the $Z' \to t\bar{t}$ decay, which also provides a source of jets arising from top quark decays. The samples used are required to have the heavy gauge bosons produced with a resonant mass of at least 2 TeV. This ensures that the jets produced in these events are highly boosted. Example Feynman diagrams of these HVT processes are shown in Figure 4.1.

MC simulations of SM processes are used for calibrating the tagger performance to match



(a) (b)



(c)

Figure 4.1: Feynman diagrams depicting the production of the heavy W' and Z' vector bosons. The W' decay into SM vector bosons is shown in (a). The $W' \to tb$ decay channel is shown in (b). The $Z' \to t\bar{t}$ decay is shown in (c).

that of data. Simulations of $t\bar{t}$ and single-top production are used as signal processes for the calibration of the jet taggers. For reasons that will be discussed in subsection 4.1.3, the presence of a single electron or muon that is associated with the production processes is required. Both of these processes serve as potential sources of boosted top and W jets. The jet reconstruction process is simulated in these samples. If a top quark from these events is highly boosted, then it can be reconstructed into a single large-R jet since the decays of the top will be collimated. On the other hand, if the top quark is not sufficiently boosted, then its decays will be more spatially separated, which allows the possibility to individually identify the jets produced from its decays. This scenario is referred to as a resolved top decay. If the W boson that originates from a resolved top decay is sufficiently boosted, then it can be reconstructed into a single large-R jet. Example Feynman diagrams of $t\bar{t}$ and single-top production are shown in Figures 4.2 and 4.3, respectively.

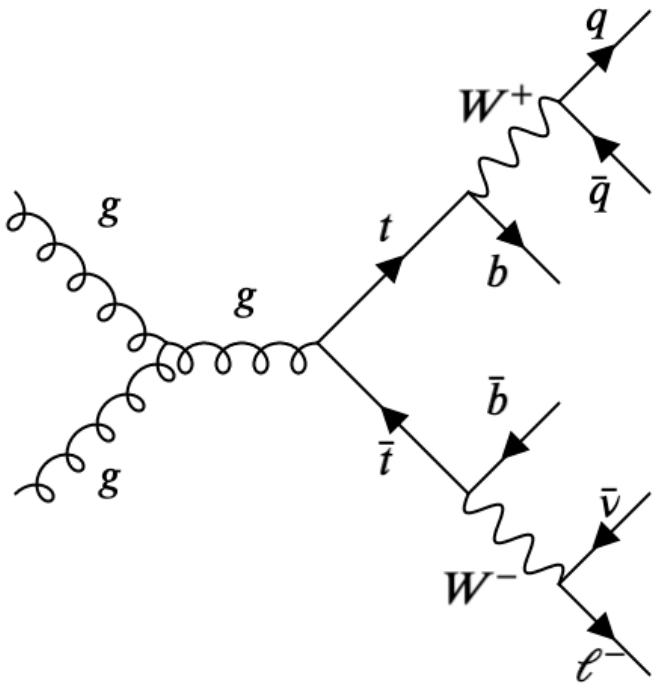
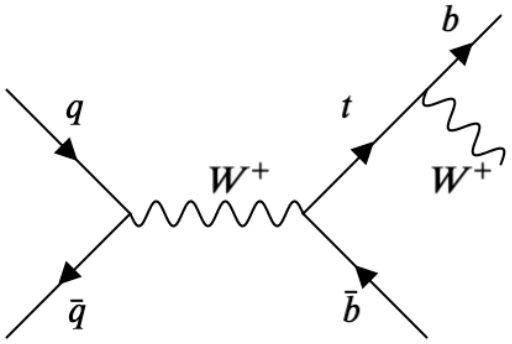
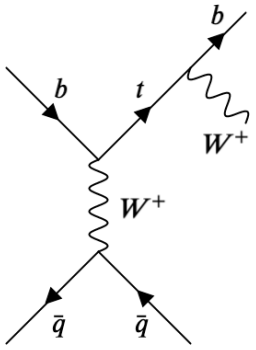


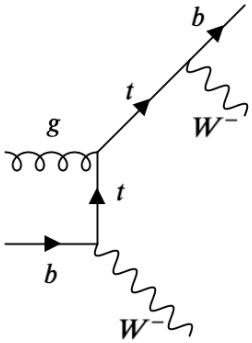
Figure 4.2: Feynman diagram depicting the pair production of top quarks through the strong force where one of the top quarks decays hadronically and the other decays leptonically.



(a)



(b)



(c)

Figure 4.3: Feynman diagrams depicting the production of a single top quark through the t-channel (a), the s-channel (b), and the Wt-channel (c).

4.1.2 Background Processes

The boosted object taggers are optimized to identify signal top and W jets against jets originating from QCD multijet production. The jets produced from this process originate from gluon radiation and the hadronization of non-top quarks. QCD multijet production is one of the most common processes that is initiated from hadron collisions. This ensures that the boosted object taggers are properly optimized against a large selection of background jets that span a wide kinematic regime.

As will be discussed in subsection 4.1.3, the QCD multijet production process is effectively suppressed by the event selection used for the tagger calibration studies. Thus, other sources of background processes that can mimic the production of signal top and W jets must be considered for the tagger calibration. These processes are the production of vector bosons associated with additional jets (V+jets) and the pair production of vector bosons (dibosons). Other potential background processes are significantly suppressed by the event selection criteria that will be discussed in the next subsection. Example Feynman diagrams of V+jets and diboson production processes are shown in Figures 4.4 and 4.5.

4.1.3 Event Selection

The event selection for the boosted object tagging studies requires:

- Exactly one electron or muon[1] with $p_{\mathrm{T}} > 30$ GeV.
- $E_{\mathrm{T}}^{\mathrm{miss}} > 20$ GeV and $E_{\mathrm{T}}^{\mathrm{miss}} + m_{\mathrm{T}}^{W} > 60$ GeV [2].

[1] Electrons and muons are often jointly referred to as leptons in analyses, a convention that will be adopted from this point on.

[2] $m_{\mathrm{T}}^{W} = \sqrt{2p_{\mathrm{T}}^{\ell} E_{\mathrm{T}}^{\mathrm{miss}}(1 - \cos\Delta\phi)}$ is the transverse mass of the lepton and the missing transverse energy system, where p_{T}^{ℓ} is the transverse momentum of the lepton and $\Delta\phi$ is the azimuthal angle separation between the lepton and the direction of the missing transverse energy.

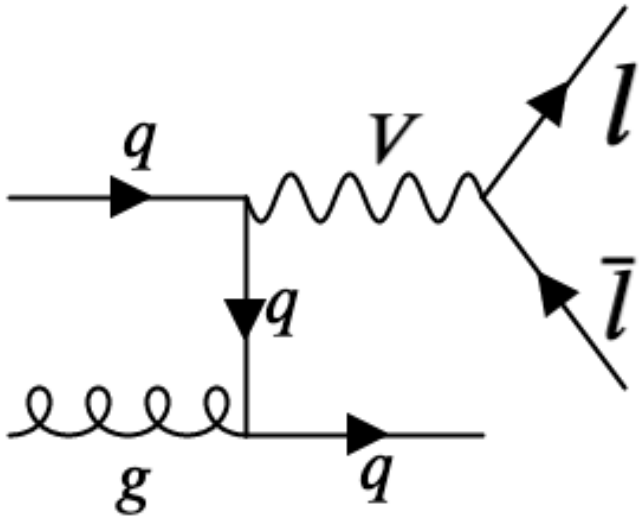


Figure 4.4: Feynman diagram depicting the production of a vector boson V in association with jets that originate from the subsequent hadronization of quarks.

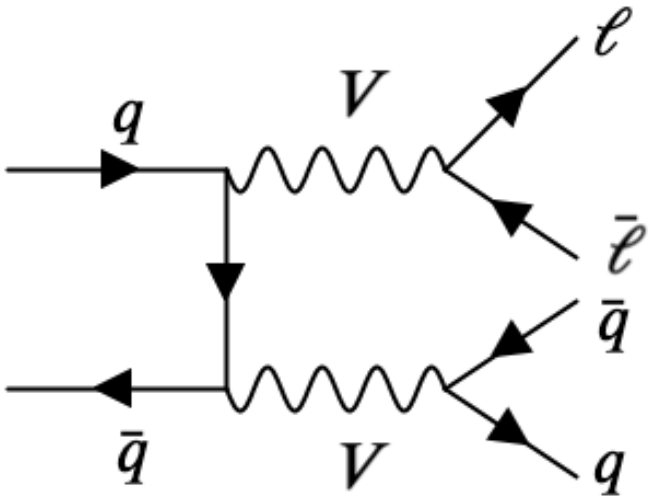


Figure 4.5: Feynman diagram depicting the pair production of vector bosons.

- At least one small-R jet with $p_{\mathrm{T}} > 25$ GeV and $\Delta R(\mathrm{jet}, \mathrm{lepton}) < 1.5$.
- At least one large-R jet with $p_{\mathrm{T}} > 200$ GeV.
- At least one b-tagged variable radius jet with $R_{\mathrm{min}} = 0.02$, $R_{\mathrm{max}} = 0.4$ and $\rho = 30$ GeV. The b-tagging algorithm used is the DL1 [44] algorithm, which uses a deep neural network to determine the probability of a jet originating from a b, c, or other light quarks.

The presence of a single lepton ensures that the signal $t\bar{t}$ and single-top production processes provide a relatively pure sample of signal top and W jets by allowing these jets to be properly reconstructed. This is especially important in $t\bar{t}$ processes, where one top quark must have a leptonic decay while the other must have a hadronic decay in order to satisfy the event selection criteria. This reduces the potential contamination and interference effects that two candidate signal jets might have in the jet reconstruction process. The presence of a single lepton also has the effect of significantly reducing the Z+jets background. Events from this process that pass the lepton selection criteria are likely to come from dileptonic decays of the Z where one of the leptons is misreconstructed as another object. This artificially generates additional missing transverse energy in the event. However, these events are suppressed with the $E_{\mathrm{T}}^{\mathrm{miss}}$ related selection requirements. In QCD multijet processes, the presence of a single lepton originates from a misreconstructed jet, resulting in the artificial generation of $E_{\mathrm{T}}^{\mathrm{miss}}$. These events are also suppressed with the $E_{\mathrm{T}}^{\mathrm{miss}}$ related selection requirements. The large-R jet p_{T} requirement ensures that the jet captures the majority of the decay products of the boosted particles within its radius. The presence of at least one b-tagged jet provides an experimental handle for identifying signal processes in data where no truth information is available. Additionally, as will be discussed in subsection 5.1.4, the b-tagged jet is required

in order to provide a containment-based candidacy criteria in order to distinguish between signal top and W jets.

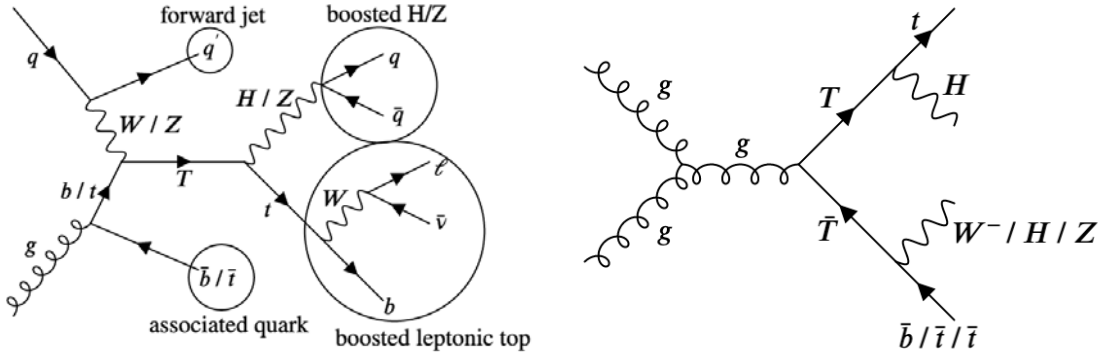
4.2 VLQ Searches

For the VLQ searches presented in Chapter 6, the signal processes are events in which these yet undiscovered particles are produced. The VLQ production processes studied are the single production of a vector-like top (T) and the pair production of vector-like tops ($T\bar{T}$). Separate MC samples are used to model these two processes, including the reconstruction of all the associated physics objects produced in the events of interest. The background processes in these searches are SM processes with events that mimic the kinematic signatures of events in which a T is produced. Processes that have a small fraction of events that pass the event selection criteria are known as reducible backgrounds. On the other hand, background events that have a significant number of events passing the event selection criteria are known as irreducible backgrounds.

As will be discussed in Chapter 6, analysis search regions are defined that elicit kinematic features of signal processes. A statistical analysis is performed in each T production search to determine if there is a significant excess of data events over the SM background prediction in the analysis search regions. If such an excess is found, then a discovery claim can be made on the production of T. The data sample used in these searches comes from data recorded from pp collisions at a center of mass energy of 13 TeV in the period 2015-2018 and corresponds to an integrated luminosity of 139 fb^{-1}.

4.2.1 Signal Processes

The signal processes of interest that are studied in the vector-like T searches are the single production of a T through the electroweak force and the pair production $T\bar{T}$ through the strong force. Only the single production process associated with a single electron or muon, referred to as the 1-lepton channel, is studied. Pair production processes are studied in the 0-lepton and the 1-lepton channels. Example Feynman diagrams that highlight characteristic features of these processes are shown in Figure 4.6.



(a) (b)

Figure 4.6: Feynman diagrams depicting the single production of a vector-like T (a) and the pair production $T\bar{T}$ (b).

The single production of a T is characterized by the simultaneous presence of several unique objects. These objects can be used to experimentally identify single-T production processes against the SM background. The initial quark q that recoils off the off-shell vector boson can often emerge as a jet with high pseudorapidity. These jets, denoted as forward jets (fj), are produced in the forward region of the detector. Additionally, the initial state gluon g can split into $b\bar{b}$ or $t\bar{t}$ with one of the quarks coupling with the vector boson. Since the mass of the top quark is larger than that of the bottom quark, the top-associated production mode is kinematically disfavored. The T decay channels that are studied in the single production

process are $T \to Ht$ and $T \to Zt$. The decay products of the T are expected to be boosted as a result of its large mass. The associated lepton is likely to be produced from the leptonic decay of the boosted top quark. The Higgs and Z bosons are expected to decay hadronically. In the case of the $T \to Ht$ decay channel, the dominant $b\bar{b}$ decay of the Higgs characterizes the signal process with a large presence of b-initiated jets.

The pair production process $T\bar{T}$ lacks experimental handles such as the presence of forward jets and an associated quark with the production. However, the number of boosted objects increases due to the additional T that is produced. This characterizes the $T\bar{T}$ production process with interesting combinatorial decay topologies. In the 0-lepton channel, the main decay topologies of interest are $T\bar{T} \to HtHt$, $HtZt$, and $ZtZt$[3]. An interesting feature of the $ZtZt$ decay topology are events in which at least one Z boson decays as $Z \to \nu\bar{\nu}$. Since the T decays are boosted, this results in a large $E_{\mathrm{T}}^{\mathrm{miss}}$. In the 1-lepton channel the main decay topologies of interest are $T\bar{T} \to HtHt$, $HtZt$ and $HtWb$. The lepton is likely to be produced from a leptonically decaying top quark. Both the 0-lepton and 1-lepton channels are characterized by a large number of b-initiated jets that originate from the predominant $H \to b\bar{b}$ decay and the top quark decays.

4.2.2 Background Processes

The main irreducible background process that can mimic the single and pair production of vector-like T is $t\bar{t}$ production in association with additional jets ($t\bar{t}$+jets). As can be observed by comparing Figures 4.2 and 4.6, the decay topologies are very similar between $t\bar{t}$ production and the different production mechanisms of the T. The presence of a boosted

[3]Throughout the remainder of this thesis, the notation $HtHt$ is used to denote both $HtH\bar{t}$ and its charge conjugate $H\bar{t}Ht$. A similar notation is used for all other $T\bar{T}$ decay topologies.

Higgs boson in T production processes can aid in the discrimination between signal and background events by correctly tagging a jet to the Higgs. However, top-initiated jets can mimic Higgs jets if the full decay of the top is not contained in the jet. This can result in the top jet having kinematic features similar to those of a Higgs jet, such as its mass, which can potentially lead to mistagging the top jet as a Higgs jet.

Single-top and V+jets production are subdominant background processes that can mimic the signal in events that have few b-tagged jets. The reducible background processes are diboson production, $t\bar{t}$ production in association with a vector or Higgs boson ($t\bar{t}V/H$), QCD multijet production, and the production of four top quarks ($t\bar{t}t\bar{t}$). These processes are reduced significantly with the event selection criteria and further kinematic requirements from the analysis search regions, which will be discussed in Chapter 6.

4.2.3 Event Selection

The event selection for the search of a singly produced T is based on the following criteria:

- Exactly one lepton with $p_{\mathrm{T}} > 30$ GeV.
- At least 3 small-R jets.
- At least one DL1 77% working point b-tagged small-R jets.
- $E_{\mathrm{T}}^{\mathrm{miss}} > 20$ GeV and $E_{\mathrm{T}}^{\mathrm{miss}} + m_{\mathrm{T}}^{W} > 60$ GeV.

The event selection criteria for the search of $T\bar{T}$ production in the 1-lepton channel is defined similarly, but instead at least 5 small-R jets and 2 b-tagged jets are required. Additionally, the b-tagging algorithm is switched from DL1 to DL1r [45], which is an optimized version of the DL1 algorithm. The $E_{\mathrm{T}}^{\mathrm{miss}}$ requirements are used to suppress the QCD multijet

production background in the 1-lepton channel. The event selection criteria for the 0-lepton channel of the pair production search is summarized as follows:

- Exactly zero leptons.
- At least 6 small-R jets.
- At least 2 DL1r 77% working point b-tagged small-R jets.
- $E_{\mathrm{T}}^{\mathrm{miss}} > 200$ GeV and $\Delta\phi_{\mathrm{min}}^{4j} > 0.4$[4]

In the 0-lepton channel, the source of missing transverse energy in QCD multijet events is likely to be produced from jet energy mismeasurements. This implies that the most energetic jets of these events are expected to be collinear with the missing transverse energy. Thus, requiring a large azimuthal separation between the leading jets and the associated direction of $E_{\mathrm{T}}^{\mathrm{miss}}$ can significantly reduce the QCD background.

[4] $\Delta\phi_{\mathrm{min}}^{4j} > 0.4$ is the minimum azimuthal angle separation between the four leading in p_{T} jets in the event and the direction of the missing transverse energy.

Chapter 5

Tagging Top Quarks

As previously discussed in Chapter 2, one of the issues stemming from the Hierarchy Problem is the low mass of the Higgs boson. Several BSM theories have been formulated with the goal of solving the Hierarchy Problem by introducing new particles that would provide the quantum loop corrections necessary to explain the value of the Higgs mass. For example, in Composite Higgs models, the VLQs provide the mechanism for the Higgs boson to acquire its mass. Since the masses of these potentially new particles are above the TeV scale, any direct observation with the ATLAS detector is impossible due to their short lifetimes. However, in most scenarios, these new particles are expected to decay into SM bosons and quarks, which can be reconstructed as jets. If the jets are correctly identified to their source particle, then one could use these jets as inputs to reconstruct the BSM particles that initiated the decay chain. The process of identifying a jet to a source particle, known as jet tagging, plays an important role in many BSM analyses searching for new particles. This will become evident in the discussion of Chapter 6, where jet tagging is embedded in several aspects of the analysis strategy, such as the reconstruction of candidate VLQs and the definition of the analysis search regions.

This Chapter is divided into two sections. The first section serves as an introduction to the concept of jet tagging. This is done through the discussion of the optimization and calibration studies of boosted object taggers that are designed to tag jets as top jets and W

jets using information from jet substructure variables. The concept of jet substructure has recently become very important in the design of jet tagging algorithms. A short overview of the jet substructure variables that are relevant to the studies presented in this Chapter is also included in the first section. The second section discusses the optimization studies of two tagging algorithms that are designed to identify top jets using information from topological data analysis (TDA), which has not been used in the context of jet tagging before. An overview of the TDA tools used is given in this section. This is followed by the optimization studies of the two tagging algorithms, which are compared in performance with one of the top taggers discussed in the first section.

5.1 Jet Tagging with Jet Substructure

In this section, jet tagging is introduced through the studies performed on the optimization and calibration of jet taggers designed to identify jets originating from boosted hadronically decaying top quarks and W bosons. The physics processes of interest and the event selection used in these studies were described in section 4.1. The author of this thesis contributed to the calibration effort as part of his ATLAS authorship qualification task. The work culminated with the derivation of data to Monte Carlo (MC) scale factors that are used to calibrate the signal efficiency of the taggers using the data collected in 2015-2017. The development and results of the calibration effort were published in an ATLAS internal note [52]. The taggers studied in this section use information from jet substructure variables that quantify how energy is distributed within the internal structure of jets. The jet tagging and jet substructure concepts that are introduced in this section are also used in the following section of this Chapter. Although the scope of the discussion is mostly limited to top quark and W

boson jet tagging, these concepts are applicable to the tagging of jets to other particles.

5.1.1 Jet Substructure Variables

As discussed in subsection 3.3.4, jets are complex structures that are constructed from the detector response of hadrons that originate from the decay chains initiated by the hadronization process. Different types of particles can produce jets with certain characteristic radiation patterns and structures inside the jet, which define the jet substructure. For example, a jet originating from the hadronic decay of a top quark can be characterized by the presence of three smaller subjets that originate from the b quark and the two quarks from the hadronic decay of the W boson. Depending on the initial energy of the top quark, these substructures can have varying degrees of collimation that can have an effect on the overall radiation pattern and structure of the top-initiated jet. Jet substructure variables quantify how the energy of the jet is distributed across its internal structure. This information allows us to discriminate between different types of jets based on the substructure that is present in them. To achieve this, different types of substructure variables analyze the energy distribution at different scales within the jet, as shown in Figure 5.1. For example, some variables quantify the energy distribution using the topocluster constituents of jets as inputs. Other variables take as input more complex substructures, such as subjets that are obtained by forming smaller jets from constituents that are within a localized region of the jet. In the following, a brief description of some of these substructure variables is given.

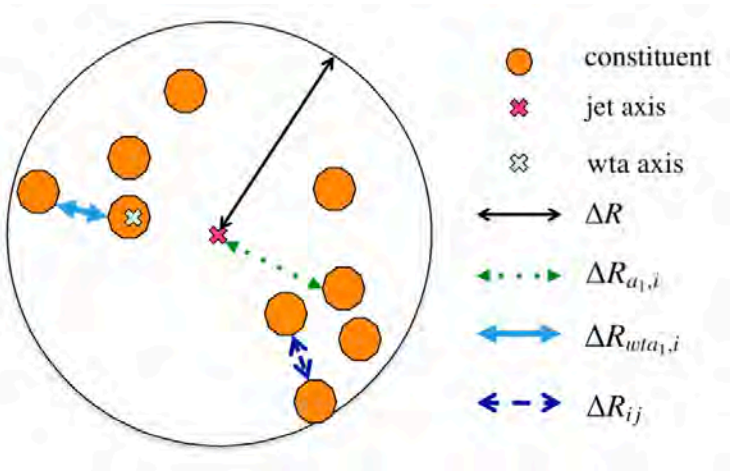


Figure 5.1: Depiction of a jet with radius parameter ΔR and its internal structure. The individual jet constituents are represented by the orange decagons. Different distance scales are shown between the jet constituents that are used to define jet substructure variables. The distance between an individual constituent i and the jet axis, which is defined as the direction of the net momentum of the jet, is denoted as $\Delta R_{a_1,i}$. The "winner-takes-all" (wta) axis is defined as the direction of the constituent with the largest p_{T} in the jet. The distance between a jet constituent i and the wta axis is denoted as $\Delta R_{wta_1,i}$. Finally, the distance between two jet constituents i and j is denoted as ΔR_{ij}. This figure is taken from [53].

n-Subjettiness

The n-subjettiness variables τ_n [54, 55] are designed to measure how well the jet system is represented with a substructure of n subjets. The subjets are obtained by applying the k_T clustering algorithm on the jet constituents. This process is stopped once n subjets are defined. The n-subjettiness variables calculate the sum of the p_{T}-weighted distances between all jet constituents and the closest subjet:

$$\tau_n = \frac{1}{p_{\mathrm{T}}^{\mathrm{ref}}} \sum_i p_{\mathrm{T}\,i} \min\left\{\Delta R_{i\,1}, \cdots, \Delta R_{i\,n}\right\} \tag{5.1}$$

where the summation index i runs over the jet constituents. The distance in the η-ϕ plane between the i^{th} constituent and the j^{th} subjet is denoted by $\Delta R_{i\,j}$. Specifically, these are

the distances between the constituents and the axis of the subjet, which is defined as the direction of the net momentum of the subjet. These variables are weighted by a reference p_{T} scale that is defined as:

$$p_{\mathrm{T}}^{\mathrm{ref}} = \sum_i p_{\mathrm{T}_i} R_0 \tag{5.2}$$

where R_0 is the radius parameter of the jet. A special sub-family of these variables uses the "winner-takes-all" (τ_n^{wta}) configuration [56], where the distances $\Delta R_{i\,j}$ are taken between i^{th} constituent of the jet and the constituent with the largest p_{T} within the j^{th} subjet.

As an example, boosted jets originating from QCD multijet processes can have radiation patterns that consists of large angle soft splittings. This results in jet constituents that are spread apart with relatively low p_{T}. Thus, for small values of n, the values of τ_n will be larger in these jets due to the majority of constituents being farther away from the subjet axes. In contrast, boosted jets that arise from the hadronic decays of W bosons and top quarks are highly collimated. This results in well-defined pronged-like structures that coincide with the direction of the decays of these particles, as shown in Figure 5.2. In the case of W jets, 1- and 2-pronged substructures are usually formed depending on the level of collimation. On the other hand, for top jets, 2- and 3-pronged substructures are usually formed. Since the jet constituents are close to the axes of these substructures, this is reflected in smaller values of τ_n for the corresponding n-prong structure.

The ratios of consecutive n-subjettiness variables, $\tau_{ab} = \tau_a/\tau_b$, are used to discriminate jets based on the relative likelihood of being represented by an n-pronged structure. The comparisons of the distribution of τ_{32} between top and QCD jets and the distribution of τ_{21} between W and QCD jets are shown in Figure 5.3. As observed in the τ_{32} distribution, top jets are better represented by a 3-pronged structure relative to a 2-pronged structure when

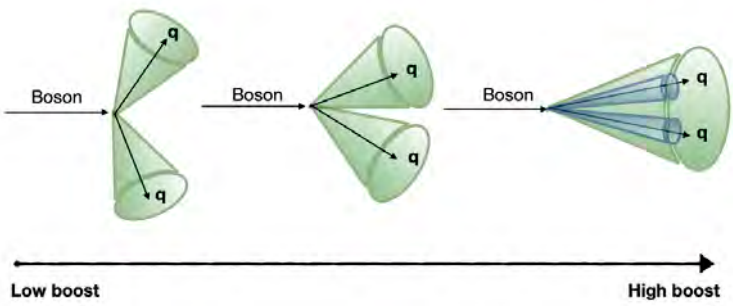
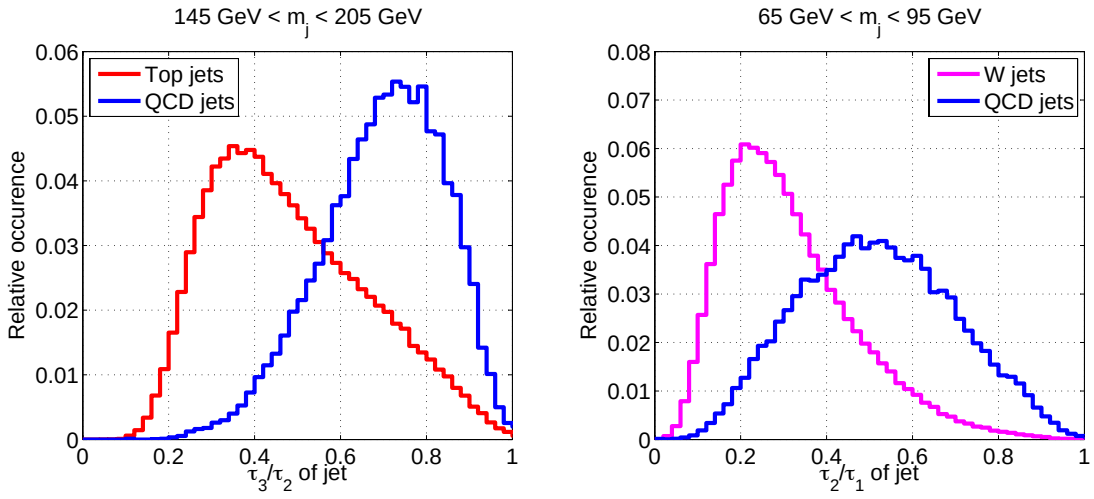


Figure 5.2: Depiction of pronged-like structures that are formed from a jet initiated by a boson. The degree of collimation of these structures is dependent on the momentum of the boson. At lower momentum, the individual jets that arise from the hadronic decays of the boson are resolved. As the momentum of the boson increases, the jets from the decays of the boson start to become more collimated to the point where they can be considered subjets of a large-R jet. The pronged-like structures coincide with the direction of the subjets in the large-R jet. This figure is taken from [57].

compared to QCD jets. Similarly, as observed in the τ_{21} distribution, W jets are better represented by a 2-pronged structure instead of a single prong structure when compared to QCD jets.

k_T Splitting Scales

The next set of substructure variables are the k_T clustering algorithm splitting scales [58]. These scales, which are denoted as $\sqrt{d_{n\,n+1}}$, are defined as the smallest k_T distance between two subjets before they get merged during the clustering step from $n+1$ to n subjets. The process of obtaining these variables can be thought of as a declustering of the jet that probes the original constituent structure of the jet. As an example, for W jets, the scale $\sqrt{d_{12}}$ should be approximately equal to half the mass of the W boson. This is because the jet energy is distributed almost equitably between the two subjets that correspond to the two



(a) (b)

Figure 5.3: The τ_{32} (a) distribution for top and QCD jets with $R_0 = 0.8$ and the τ_{21} (b) distribution for W and QCD jets with $R_0 = 0.6$. The selection criteria for these jets require $p_{\mathrm{T}} > 300$ GeV and $|\eta| < 1.3$. Additionally, a window cut on the jet mass that selects jets that have a mass close to the top quark and W boson mass is applied. These figures are taken from [54].

quark decays of the W boson.

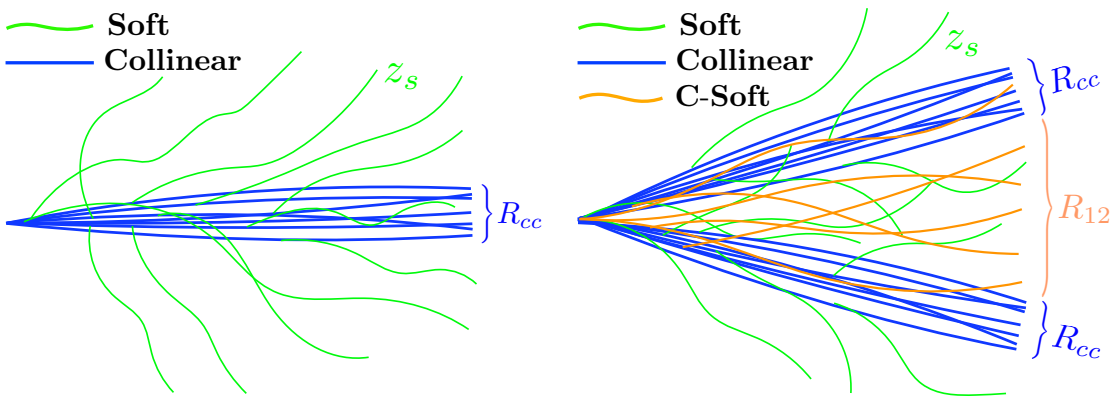
n-Point Energy Correlation Functions

Similar to the n-subjettiness variables, the set of n-point energy correlation functions e_n [59, 60] are designed to probe for n-pronged structures in jets. However, this contrasts with n-subjettiness in that e_n quantifies the jet energy distribution relative to the jet constituents instead of subjets. Mathematically, the n-point energy correlation functions for $n = 2$ and $n = 3$ are defined as:

$$\begin{aligned} e_2 &= \frac{1}{(p_{\mathrm{T}}^{\mathrm{ref}})^2} \sum_{1 \leq i < j \leq N_{\mathrm{const}}} p_{\mathrm{T}\,i} p_{\mathrm{T}\,j} \Delta R_{i\,j} \\ e_3 &= \frac{1}{(p_{\mathrm{T}}^{\mathrm{ref}})^3} \sum_{1 \leq i < j < k \leq N_{\mathrm{const}}} p_{\mathrm{T}\,i} p_{\mathrm{T}\,j} p_{\mathrm{T}\,k} \Delta R_{i\,j} \Delta R_{i\,k} \Delta R_{j\,k} \end{aligned} \tag{5.3}$$

where the summation runs over the constituents in the jet, N_{const} is the number of constituents in the jet, and $p_{\mathrm{T}}^{\mathrm{ref}}$ is defined similarly as in Equation 5.2. The only relevant energy correlation functions for the discussion in this Chapter are e_2 and e_3. The definition of higher order energy correlation functions follows a natural generalization from Equation 5.3.

The phase space of e_2 and e_3 separates jets with single-pronged substructures and two-pronged substructures into different regions. As an example, we consider the case of QCD jets with a single-pronged substructure and Z jets with a two-pronged substructure, as shown in Figure 5.4. A single-pronged QCD jet is usually characterized by collinear radiation that is localized within a small angular region of the jet ($R_{cc} \ll 1$). Additionally, the presence of soft radiation in QCD jets from gluon and quark emissions is characterized by having a low fraction of the jet p_{T} ($z_s = p_{\mathrm{T\,soft}}/p_{\mathrm{T\,jet}} \ll 1$). With these limits in consideration for QCD jets, the n-point energy correlation functions in Equation 5.3 can be shown to behave



(a) (b)

Figure 5.4: Depiction of the radiation patterns of a single-pronged jet (a) and a two-pronged jet (b). The radiation pattern of the single-pronged jet is characterized by a collinear emission core that carries most of the p_{T} of the jet and is localized within a small angular region R_{cc} of the jet. Additionally, soft radiation may be present from gluon and quark emissions, which carries a small fraction z_s of the jet p_{T}. The radiation pattern of the two-pronged jet is characterized by two collinear emission cores that are localized within a small angular region R_{cc} and separated by an angular distance R_{12}. In addition to global soft radiation, there also may be collinear soft-radiation present that is localized within an angular region of size R_{12}. These figures are taken from [60].

approximately as:

$$\begin{aligned} e_2 &\approx R_{cc} + z_s \\ e_3 &\approx R_{cc}^3 + z_s^2 + R_{cc} z_s \end{aligned} \tag{5.4}$$

If the soft radiation in QCD jets has the largest contribution in Equation 5.4, then $e_2 \approx z_s$ and $e_3 \approx e_2^2$. On the other hand, if the collinear radiation from the single prong has the largest contribution, then $e_2 \approx R_{cc}$ and $e_3 \approx e_2^3$. Thus, under these limits, QCD jets with a single-prong substructure populate the region of phase space where $e_2^3 \leq e_3 \leq e_2^2$. Jets with a two-pronged substructure, such as a Z-initiated jet, are usually characterized by the angular size of their collinear emissions satisfying $R_{cc} \ll R_{12} \ll 1$, where R_{12} is the angular separation between the two prong structures. In addition to soft radiation, there may also be collinear soft radiation that is localized in an angular region of size R_{12} with p_{T} fraction z_{cs}. With these limits in consideration for a two-pronged jet, the n-point energy correlation

functions in Equation 5.3 can be shown to behave approximately as:

$$\begin{aligned} e_2 &\approx R_{12} \\ e_3 &\approx R_{12} z_s + R_{12}^2 R_{cc} + R_{12}^3 z_{cs} \end{aligned} \tag{5.5}$$

If the jet has negligible non-collinear soft radiation, then most of its energy is carried by the two prong substructures and $z_{cs} \ll 1$. Under these limits, the jet populates the region of phase space where $e_2^3 \approx R_{12}^3 \gg R_{12}^3 z_{cs} \approx e_3$. The ratios of n-point energy correlation functions $C_2 = e_3/e_2^2$ and $D_2 = e_3/e_2^3$ provide a relative measure of how well a jet is characterized with two prong substructures compared to a single prong substructure. The distribution of these two ratios compared between Z jets and QCD jets is shown in Figure 5.5. As can be observed, Z jets populate the regions $C_2 < 1$ and $D_2 < 1$, which is characteristic of a two-pronged jet, while QCD jets populate the regions $C_2 < 1$ and $D_2 > 1$, which is characteristic of a single-pronged jet.

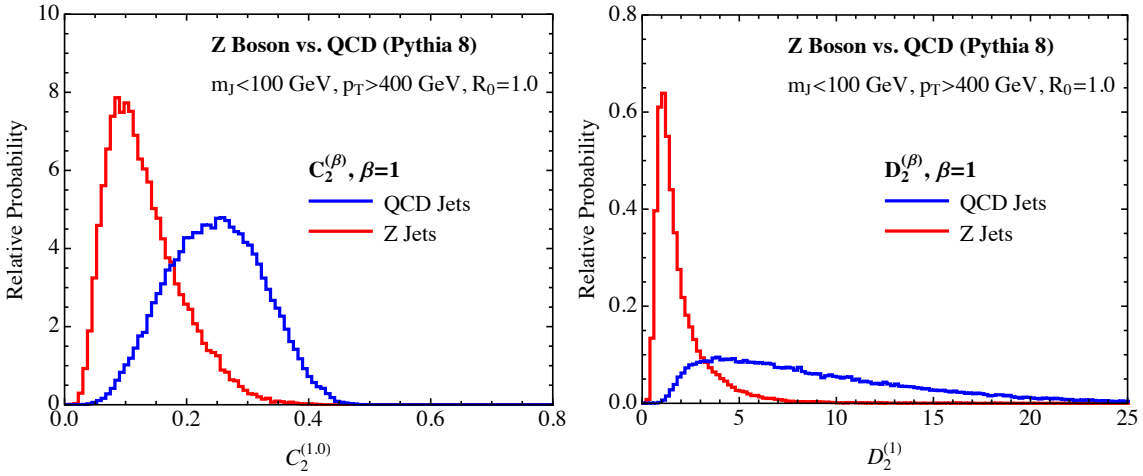


Figure 5.5: Comparison of the energy correlation function ratios C_2 (a) and D_2 (b) between Z large-R jets and QCD large-R jets. The selection criteria for the large-R jets requires the jet mass to be less than 100 GeV and the jet $p_{\mathrm{T}} > 400$ GeV. These figures are taken from [60].

Reconstructed Substructure Mass

The mass of boosted jets originating from particles such as W bosons and top quarks holds discriminatory power against background jets that originate from other particles. This is due to the mass of the jet being close in value to the mass of the particle that originated it. However, if jets are sufficiently boosted or misreconstructed, they can have mass values that differ from the mass of their source particle, which can lead to the misidentification of both signal and background jets. The last set of substructure variables to be considered are the invariant masses of reconstructed substructures within the jet [61]. Candidate substructures can be reconstructed by combining subjets that satisfy a hypothesis based on the radiation pattern of the desired substructure and achieve an invariant mass that is close to the real mass value of the substructure. For example, in jets that arise from the decay products of the top quark, the W boson can be reconstructed by pairing two subjets that are relatively close and result in an invariant mass close to the mass of the W boson. If this procedure is applied to jets originating from particles that do not contain a W boson in their decay products, then the resulting invariant mass can differ significantly from the W mass as a result of reconstructing a W boson from inconsistent radiation patterns.

5.1.2 Jet Substructure Taggers

From the previous discussion, it is clear that the different types of jet substructure variables hold significant discriminatory power to differentiate between signal and background jets. However, in certain kinematic regimes, these variables by themselves may not provide sufficient information in order to classify jets. For example, one could consider a highly boosted top quark jet in which the decays of the W boson are extremely collimated. In this scenario,

the top jet is better represented with a 2-prong structure, which can result in similar values of τ_{32} to those of a W jet with a resolved structure. In order to fully exploit the information from these substructure variables and their correlations, three taggers were optimized for tagging jets to top quarks and W bosons [62]. The optimization of these taggers was performed using the MC samples from the HVT model and QCD multijet production processes described in section 4.1 as sources of signal and background jets, respectively.

Two substructure-based deep neural network (DNN) taggers were optimized for tagging jets to top quarks. One tagger, known as the contained DNN tagger, is designed to tag jets that contain the full decay products of the top quark. The other tagger, known as the inclusive DNN tagger, is designed to tag jets regardless of the full containment of the top decays in the jet. Both DNN taggers use the same set of input variables, which are: the jet mass, the jet p_{T}, the n-subjettiness variables τ_1^{wta}, τ_2^{wta}, τ_3^{wta}, τ_{21}^{wta}, and τ_{32}^{wta}; the 3-point energy correlation function e_3 and the ratios C_2 and D_2; the k_{T} splitting scales $\sqrt{d_{12}}$ and $\sqrt{d_{23}}$; and the invariant mass of the reconstructed W boson that originates from the decay of the top quark. These variables are then combined into a single output variable that quantifies the probability of the jet originating from a top quark. A selection cut can be defined on the output variable that divides it into two regions: a top jet region and a background jet region, which allows for a binary classification of jets.

For W jet tagging, a three-variable tagger was optimized for this task. This tagger takes as input the jet mass, the energy correlation function ratio D_2, and the number of tracks, n_{track}, that are associated with the jet. The tagger defines selection cuts, which are parametrized by the jet p_{T}, on the input variables. These cuts split the phase space into a W jet region and a background jet region. Jets that pass the selection criteria of the W jet region are then tagged as a W jet.

5.1.3 Jet Truth Labeling

A labeling criteria for jets in events from MC signal processes must be provided in order to properly define and optimize the signal efficiency of the taggers. This labeling criteria will allow us to define which jets are signal-like and background-like for each tagger. As it was discussed in subsection 3.3.7, two types of jets can be obtained from MC events: truth jets (J_{truth}) and reconstructed jets (J_{reco}). The reconstructed jets that are used in the tagging studies presented in this Chapter are obtained by applying the anti-k_T clustering algorithm with a radius parameter $R = 1.0$ to the topoclusters obtained from the simulated detector responses from the truth information particles. The taggers are calibrated using J_{reco}, so the labeling criteria is applied to these jets. The labeling procedure starts by spatially matching J_{truth} to a hadronically decaying truth top quark or W boson by requiring that $\Delta R(J_{\text{truth}}, \text{truth particle}) < 0.75$. Next, a J_{reco} is spatially matched to a J_{truth} that has been matched to a truth particle using the same distance criteria. In the case of the DNN top taggers, J_{reco} is labeled as an inclusive top jet if it is matched to a J_{truth} that is matched to a truth top. If, in addition to being matched to a truth top, the mass of J_{truth} satisfies $m_{J_{\text{truth}}} > 140$ GeV and J_{truth} has at least one associated b-hadron [63], then J_{reco} is labeled as a contained top jet. In the case of the W tagger, J_{reco} is labeled as a W jet if it is matched to a J_{truth} that is matched to a truth W with a mass that satisfies $50 < m_{J_{\text{truth}}} < 100$ GeV and J_{truth} has zero associated b-hadrons. If J_{reco} fails the signal label criteria for a given tagger, then it is labeled as a background jet for that tagger.

5.1.4 Tagger Signal Efficiency Optimization

The signal efficiency of a boosted object tagger in MC events is defined as the fraction of signal jets that are correctly tagged:

$$\epsilon_{\mathrm{MC}}(p_{\mathrm{T}}) = \frac{N_{\mathrm{signal}}^{\mathrm{tagged}}(p_{\mathrm{T}})}{N_{\mathrm{signal}}^{\mathrm{tagged}}(p_{\mathrm{T}}) + N_{\mathrm{signal}}^{\mathrm{not\ tagged}}(p_{\mathrm{T}})} \tag{5.6}$$

The efficiency is evaluated in bins of the jet p_{T} that are designed to contain jets that are kinematically similar. The binning used for each tagger will be listed in subsection 5.1.5. Two fixed signal efficiency working points were optimized for the taggers studied. The loose working point requires that the signal efficiency be 80% in all p_{T} bins considered by the tagger. This working point is designed to maximize the acceptance of signal jets, but at the cost of a higher mistag rate for background jets. The tight working point requires that the signal efficiency be 50% in all p_{T} bins. This working point is designed to reject a larger fraction of background jets, but at the cost of lower signal jet acceptance. For the DNN top taggers the working points are obtained by defining cuts on the DNN output variable that achieve the desired efficiency in each p_{T} bin. This procedure is extended for the W tagger by defining simultaneous cuts on the jet mass, D_2, and n_{track} that achieve the desired efficiency. These cuts are then parametrized by performing a functional fit as a function of the jet p_{T} [62]. The results of these fits for the mass and D_2 are shown in Figure 5.6. The jet mass fit defines a window cut, while the D_2 fit defines an upper bound cut. In the case of n_{track}, the fits were found to be consistent with a constant upper bound cut of 26 for the 50% woking point and 34 for the 80% working point.

The performance of the taggers is assessed using the background rejection metric. For a given fixed signal efficiency working point, the background rejection is defined as the number

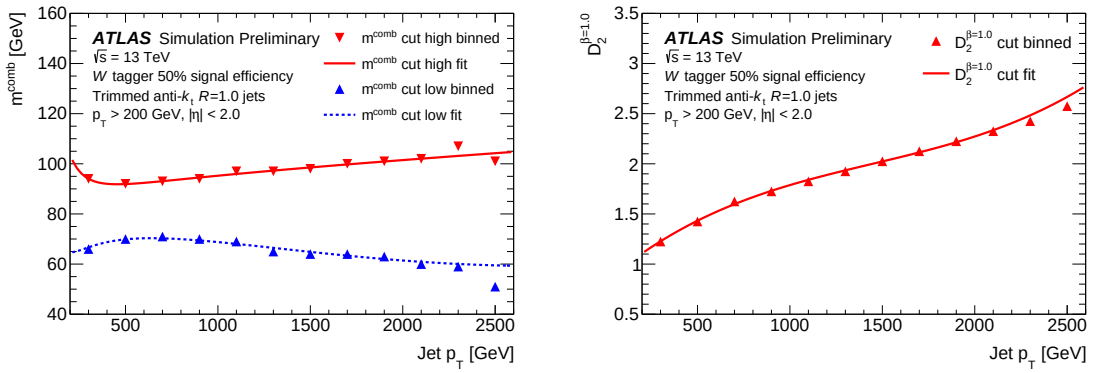
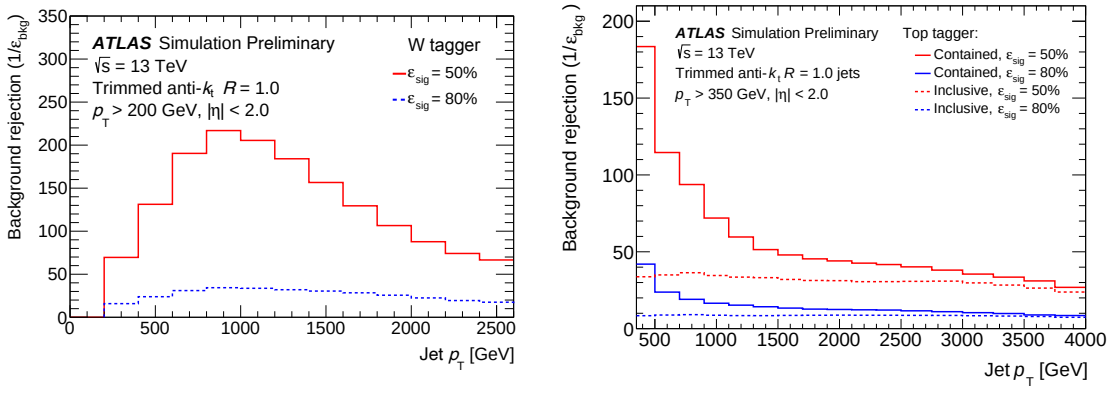


Figure 5.6: The jet mass window cuts and the D_2 upper bound cuts that define the tagging criteria as a function of the jet p_{T} for the 50% fixed signal efficiency working point W tagger (a) - (b) and the 80% fixed signal efficiency working point W tagger (c) - (d). The solid and dotted lines in each plot indicate the parametrized tagging criteria as a function of the jet p_{T}. These figures are taken from [52].

of background jets that are not tagged to the particle of interest. Figure 5.7 shows the QCD multijet background rejection as a function of the jet p_{T} for the DNN top taggers and the W tagger. For the W tagger, it can be observed that the rejection is maximal in the 800-1000 GeV p_{T} interval, which coincides with a narrow jet mass window requirement, as shown in Figure 5.6. Additionally, the D_2 requirement in this p_{T} interval is low, which is indicative of a jet with a 2-pronged structure. In the case of top tagging, the additional requirements imposed on candidate top jets for the contained tagger result in a higher background rejection when compared to the inclusive tagger.



(a) (b)

Figure 5.7: The QCD multijet background rejections for the W tagger (a) and the contained and inclusive top DNN taggers (b) overlayed for the 50% and 80% fixed signal efficiency working points as a function of the jet p_{T}. The jets are required to have a $p_{\mathrm{T}} > 200$ GeV in (a) and $p_{\mathrm{T}} > 350$ GeV in (b) to be in the validity range of the W and top taggers, respectively. These figures are taken from [52].

5.1.5 Tagger Signal Efficiency Calibration

In order to calibrate the signal efficiency of the boosted objects taggers in MC events to that of data, the signal efficiency needs to be extracted from MC simulations of SM processes. As described in subsection 4.1.1, the $t\bar{t}$ and single-top production processes are used as a source of candidate signal jets for the boosted object taggers. The events used for the tagger

calibrations are required to have at least one J_{reco} with $p_{\text{T}} > 200$ GeV and at least one b-tagged small-R jet (j_b). The J_{reco} with the largest p_{T} is selected as a candidate top jet if it satisfies the containment criteria $\Delta R(J_{\text{reco}}, j_b) < 1.0$, which corresponds to the topology of a jet initiated by a top quark. On the other hand, if the jet does not satisfy the containment criteria, then it is selected as a candidate W jet. Candidate MC jets that satisfy the signal labeling criteria described in 5.1.3 are then considered signal jets that are used to extract the signal efficiency of the corresponding tagger. The same procedure outlined for selecting candidate top and W jets in MC is also applied to jets in data.

The MC modeling of data is first assessed in the input variables of the taggers prior to evaluating the tagger efficiencies with data. Several sources of systematic uncertainties are included in the evaluation of the MC modeling. These uncertainties originate from the theory assumptions made in the MC predictions of the samples used and from the reconstruction and calibrations of relevant physics objects. The uncertainties are grouped as follows:

- $t\bar{t}$ modeling uncertainties: These are the uncertainties in the modeling of the signal $t\bar{t}$ process. These uncertainties are associated with the choice of the MC generator used to model the $t\bar{t}$ process, the modeling of the hadronization process, and the modeling of the initial state radiation (ISR) and final state radiation (FSR). These uncertainties are estimated by comparing the nominal MC sample that is used to model the $t\bar{t}$ process with alternative MC samples obtained by varying the MC generator algorithms and modeling parameters, as described in Appendix A.

- Theory uncertainties: These are the uncertainties on the cross-section of the $t\bar{t}$, single-top, and W+jets production processes.

- Large-R jet uncertainties: These are the uncertainties in the calibration of the jet

energy scale and resolution and the jet mass scale and resolution.

- Flavor tagging uncertainties: These are the uncertainties in the efficiency of tagging jets to b-, c-, and light-quarks. Additionally, uncertainties on the data to MC scale factors that are used to calibrate the efficiency of flavor tagging are also taken into account.

- Other experimental uncertainties: These are the uncertainties in the luminosity measurement for the 2015-2017 dataset and the detector response to leptons and missing transverse energy.

The $t\bar{t}$ modeling uncertainties are expected to be the largest source of uncertainty in the calibration of the taggers. Since these uncertainties vary the hadronization, the ISR, and the FSR models of the signal $t\bar{t}$ process, the simulated detector response will also vary, thereby varying the jet reconstruction process. Variations in the reconstructed jets can then impact the outcome of the jet labeling procedure and the different tagger input variables, resulting in variations in the signal efficiency measurement.

Figure 5.8 shows the distributions of the DNN scores and the jet mass for candidate top jets compared between data and the total MC prediction. As can be observed, the MC simulation models the data well. All differences between data and MC in the regions dominated by signal jets are within the $t\bar{t}$ modeling uncertainties. Figure 5.9 shows the distributions of D_2, n_{track}, and jet mass for candidate W jets. The D_2 and n_{track} distributions are shown in W-enhanced regions by requiring that the jet mass be in the $[65, 95]$ GeV interval. In addition to the jet mass cut, a requirement of $D_2 < 1.2$ is included in the n_{track} distribution. As can be observed, the mass distribution shows good agreement between data and MC. However, the D_2 and n_{track} distributions show large differences between data and MC.

These differences, however, are within the total uncertainty considered.

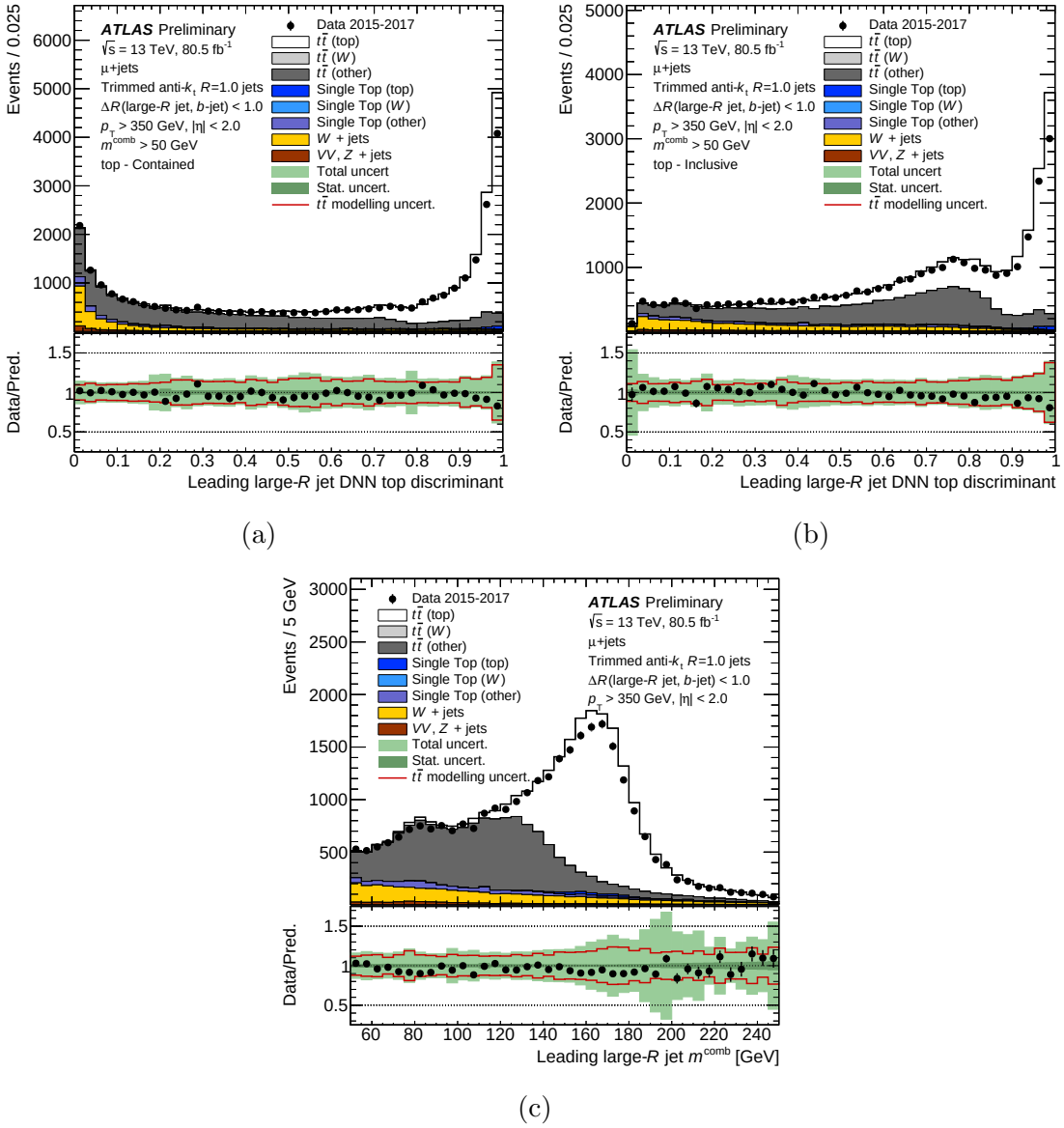


Figure 5.8: Comparisons between data and MC of the contained DNN score (a), the inclusive DNN score (b), and the jet mass (c) distributions for candidate top jets. The candidate top jets from MC signal processes that pass the signal top jet labeling criteria are indicated as $t\bar{t}$ (top) and Single Top (top). The contained top labeling criteria is used in (a), while the inclusive top labeling criteria is used in (b) - (c), as described in subsection 5.1.3. The candidate top jets from MC signal processes that fail the corresponding top labeling criteria are indicated as $t\bar{t}$ (other) and Single Top (other). The bottom panel in each plot shows the ratio of data to the total MC prediction for each bin of the distributions. The dark green band represents the statistical uncertainty, the red line the total $t\bar{t}$ modeling systematic uncertainty, and the light green band the total uncertainty for each bin.

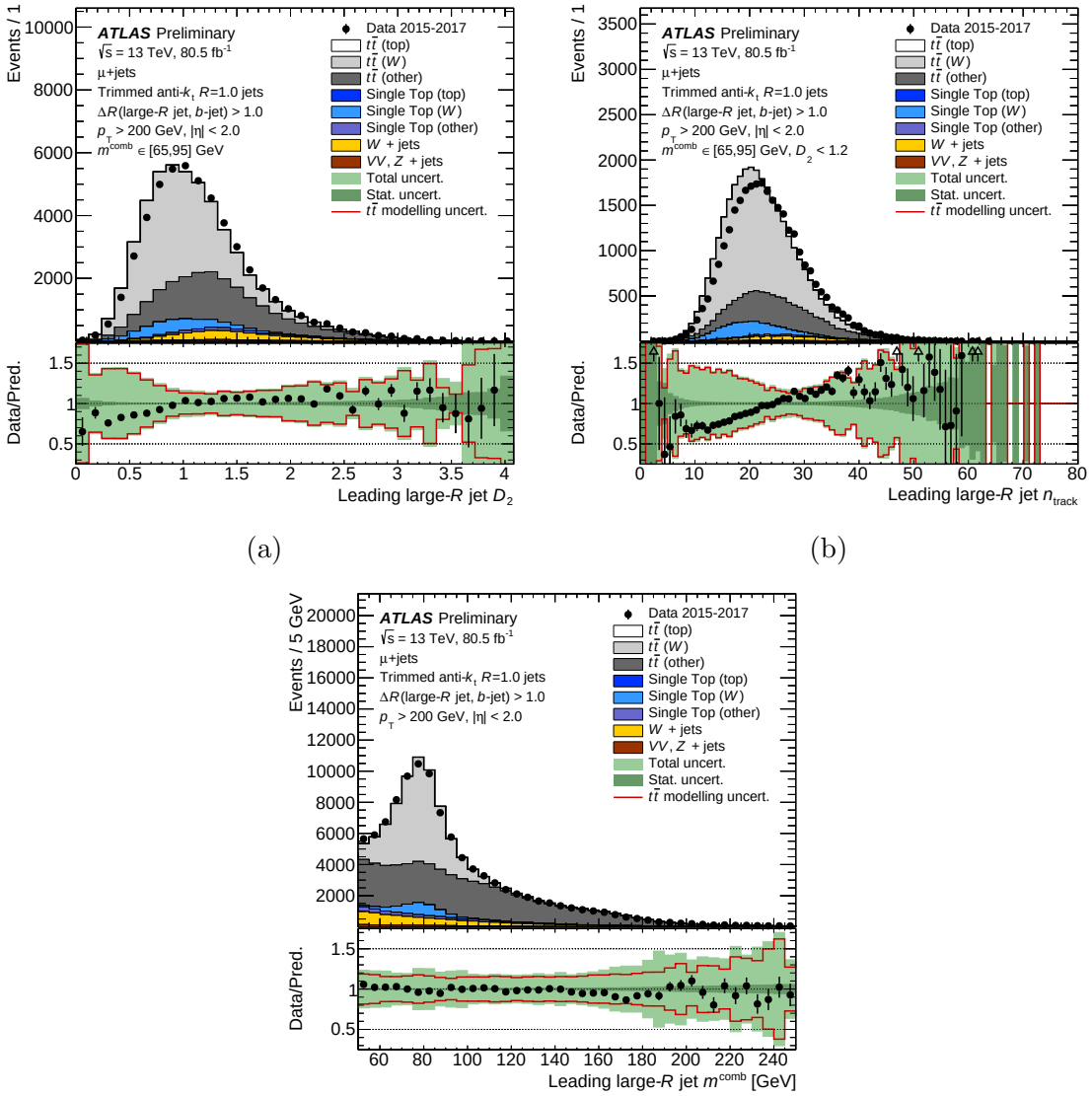


Figure 5.9: Comparisons between data and MC of the D_2 (a), the n_{track} (b), and the mass (c) distributions for candidate W jets. The candidate W jets from MC $t\bar{t}$ (W) and Single Top (W) signal processes are required to pass the W labeling criteria, as described in subsection 5.1.3. The candidate W jets from MC signal processes that fail the W labeling criteria are indicated as $t\bar{t}$ (other) and Single Top (other). A mass window selection of $[65, 95]$ GeV is included in both the D_2 and n_{track} distributions, with an additional selection cut of $D_2 < 1.2$ applied to the n_{track} distribution. These selections are included in order to highlight the observed differences between data and MC in a region that is close to the W tagger acceptance region. The bottom panel in each plot shows the ratio of data to the total MC prediction for each bin of the distributions. The dark green band represents the statistical uncertainty, the red line the total $t\bar{t}$ modeling systematic uncertainty, and the light green band the total uncertainty for each bin.

Unlike MC simulation, there is no truth information in data that can be used to select jets that are signal-like jets. Instead, the number of signal jets in data is determined by performing a χ^2 fit of the number of candidate signal jets in MC to the number of candidate jets in data. The fit is performed both for jets that are tagged and those that are not tagged in order to determine the normalization factors $N^{\text{tagged}}_{\text{fitted signal}}$ and $N^{\text{not tagged}}_{\text{fitted signal}}$. The jet mass distribution is used as the template on which the fits are performed. Additionally, the fits are done in different jet p_{T} bins using an independent χ^2 fit in each bin. For the top taggers, the p_{T} bins in units of GeV are $[350, 400]$, $[400, 450]$, $[450, 500]$, $[500, 600]$, and $[600, 1000]$. For the W tagger, the bins are $[200, 250]$, $[250, 300]$, $[300, 350]$, and $[350, 600]$. Figure 5.10 shows examples of the jet mass distributions for the W and contained top taggers after performing the fit.

The tagger signal efficiency in data is defined as:

$$\epsilon_{\text{Data}}(p_{\text{T}}) = \frac{N^{\text{tagged}}_{\text{fitted signal}}(p_{\text{T}})}{N^{\text{tagged}}_{\text{fitted signal}}(p_{\text{T}}) + N^{\text{not tagged}}_{\text{fitted signal}}(p_{\text{T}})} \tag{5.7}$$

Scale factors are calculated to calibrate the MC signal efficiency to that of data. These are defined as:

$$\text{SF}(p_{\text{T}}) = \frac{\epsilon_{\text{Data}}(p_{\text{T}})}{\epsilon_{\text{MC}}(p_{\text{T}})} \tag{5.8}$$

The propagation of the systematic uncertainties to the signal efficiency measurement in MC is obtained by evaluating the tagger signal efficiency with the systematically varied jets. To obtain the systematically varied signal efficiency in data, the χ^2 fits to data are repeated using the systematically varied jet mass distributions. The systematically varied signal efficiencies in MC and data are then used to obtain the systematically varied scale factor. The total uncertainty on the scale factors is obtained by adding in quadrature the individual scale

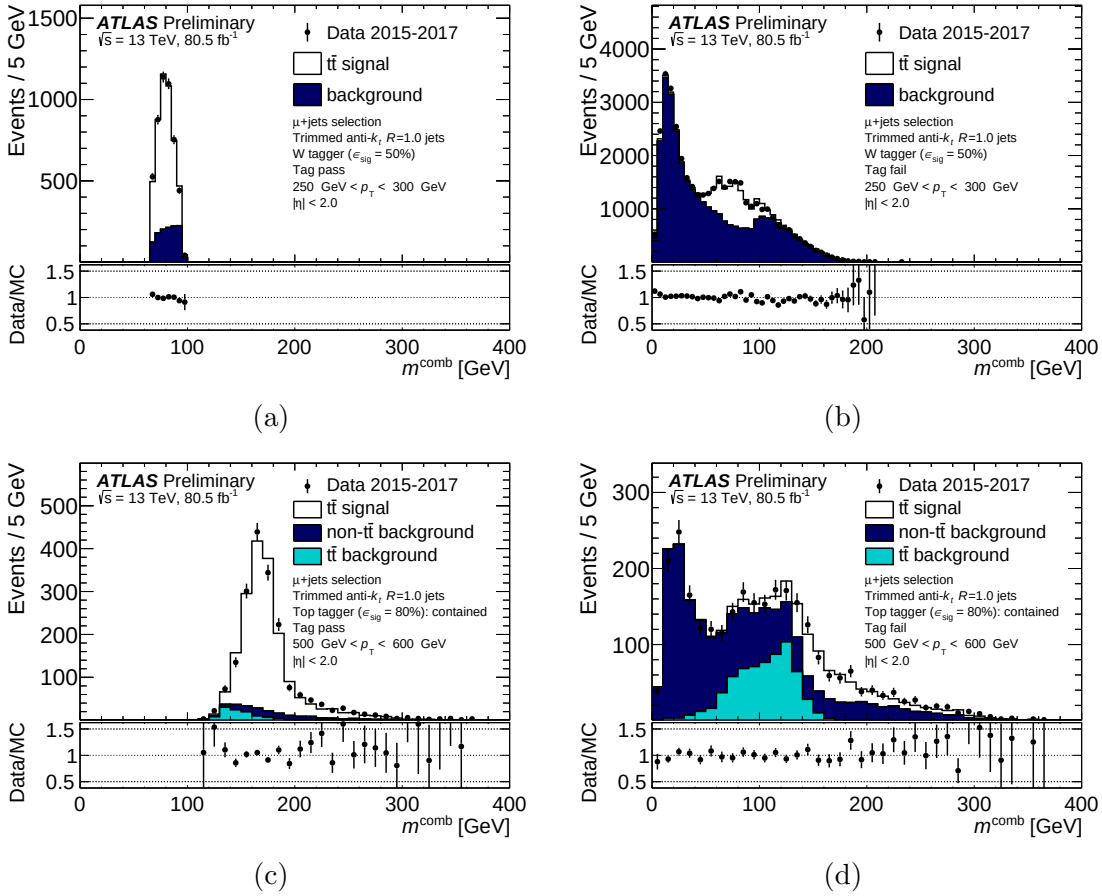
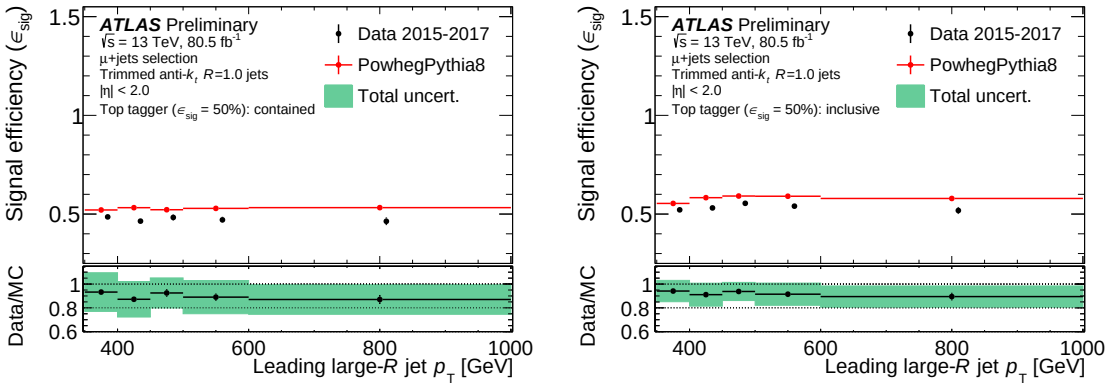


Figure 5.10: Comparison between data and the post-fit MC jet mass distributions for jets that pass and fail the tagging criteria of the 50% fixed signal efficiency W tagger and the 80% fixed signal efficiency contained top DNN tagger. The distributions are shown in the jet p_{T} bin $[250, 300]$ GeV for the W tagger and in the bin $[500, 600]$ GeV for the contained top DNN tagger. The $t\bar{t}$ signal MC template contains candidate jets from signal processes that are labeled as signal jets for the corresponding tagger. The background MC template in the W tagger plots contains candidate jets from signal processes that fail the W labeling criteria and candidate jets from background processes. This template is split in the top tagger plots for visualization purposes. The $t\bar{t}$ background template contains candidate jets from signal processes that fail the contained top labeling criteria. The non-$t\bar{t}$ backgrounnd template contains candidate jets from background processes. In all plots, the $t\bar{t}$ signal template has been scaled with the normalization parameters $N^{\text{tagged}}_{\text{fitted signal}}(p_{\mathrm{T}})$ and $N^{\text{not tagged}}_{\text{fitted signal}}(p_{\mathrm{T}})$ for jets that pass and fail the tagger, respectively. The bottom panel in each plot shows the ratio of data to MC for each bin.

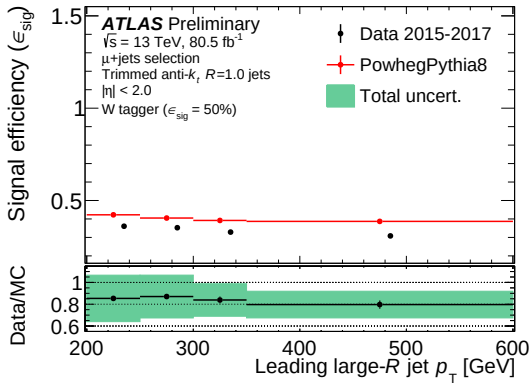
factor variances for each source of systematic uncertainty.

Figures 5.11 and 5.12 show the tagger signal efficiencies in MC and data for the 50% and 80% fixed signal efficiency working points, respectively. The bottom panels in these plots show the corresponding scale factor in the jet p_{T} bin. The total uncertainty on the scale factors is also shown. The efficiency in MC slightly overestimates the efficiency in data, as can be observed from the scale factors ranging between 0.8 and 1. This is more apparent in the W tagger, where the majority of the p_{T} bins have a scale factor below 0.9, which could be a result of the differences observed in the input variables of the W tagger between MC and data. Tables 5.1 - 5.3 show the breakdown of the contribution to the total scale factor uncertainty from the different uncertainty groups considered in the 50% working point taggers. The same information is shown in Tables 5.4 - 5.6 for the 80% working point taggers. Overall, the uncertainty is systematically dominated by the $t\bar{t}$ modeling uncertainties. The last p_{T} bins also show significant contribution from statistical uncertainty due to low statistics in this kinematic region.



(a)

(b)



(c)

Figure 5.11: Comparison between data and MC of the tagger signal efficiencies for the contained top DNN tagger (a), the inclusive top DNN tagger (b), and the W tagger (c) that were optimized to a 50% fixed signal efficiency working point. The bottom panel in each plot shows the ratio of the data signal efficiency to the MC signal efficiency in each jet p_{T} bin, which is equivalent to the tagger scale factor. The green uncertainty band represents the total uncertainty that is propagated to the scale factors.

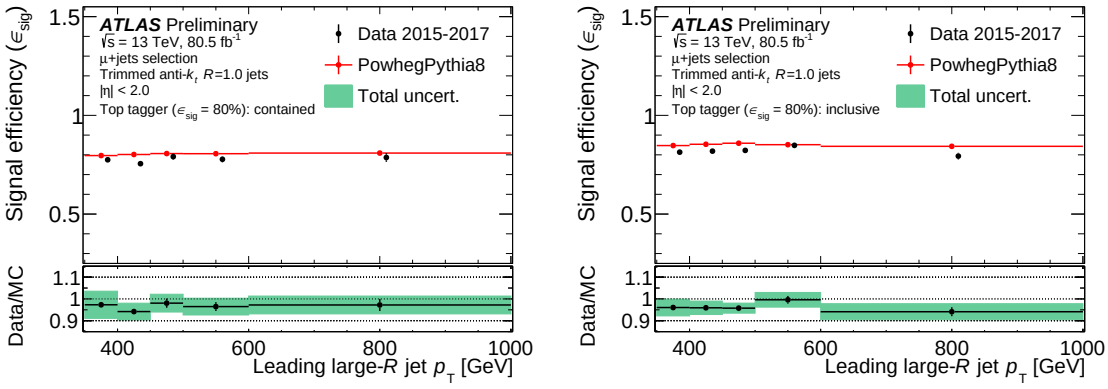


Figure 5.12: Comparison between data and MC of the tagger signal efficiencies for the contained top DNN tagger (a), the inclusive top DNN tagger (b), and the W tagger (c) that were optimized to a 80% fixed signal efficiency working point. The bottom panel in each plot shows the ratio of the data signal efficiency to the MC signal efficiency in each jet p_{T} bin, which is equivalent to the tagger scale factor. The green uncertainty band represents the total uncertainty that is propagated to the scale factors.

Systematic Group	Contained top tagger p_{T} bins [GeV]				
	[350,400]	[400,450]	[450,500]	[500,600]	[600,1000]
Statistical	0.02	0.02	0.03	0.03	0.04
Theory	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
$t\bar{t}$ modeling	0.16	0.15	0.12	0.13	0.11
Large-R jet	0.02	0.01	0.01	0.03	0.03
Other experimental	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
b-tagging	< 0.01	< 0.01	< 0.01	< 0.01	0.02
Total Uncertainty	0.16	0.15	0.13	0.14	0.12

Table 5.1: The uncertainty on the scale factor measurement of the 50% fixed signal efficiency contained top tagger from each individual systematic uncertainty group. Each row shows the uncertainty obtained by adding in quadrature the impact of all uncertainties in the group. The total uncertainty is obtained by adding in quadrature the impact of all uncertainties.

Systematic Group	Inclusive top tagger p_{T} bins [GeV]				
	[350,400]	[400,450]	[450,500]	[500,600]	[600,1000]
Statistical	0.01	0.02	0.02	0.02	0.03
Theory	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
$t\bar{t}$ modeling	0.09	0.09	0.07	0.09	0.08
Large-R jet	0.01	0.01	0.01	0.02	0.02
Other experimental	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
b-tagging	< 0.01	< 0.01	< 0.01	0.01	0.02
Total Uncertainty	0.09	0.10	0.08	0.09	0.09

Table 5.2: The uncertainty on the scale factor measurement of the 50% fixed signal efficiency inclusive top tagger from each individual systematic uncertainty group. Each row shows the uncertainty obtained by adding in quadrature the impact of all uncertainties in the group. The total uncertainty is obtained by adding in quadrature the impact of all uncertainties.

Systematic Group	W tagger p_{T} bins [GeV]			
	[200,250]	[250,300]	[300,350]	[350,600]
Statistical	0.01	0.02	0.03	0.04
Theory	< 0.01	< 0.01	< 0.01	< 0.01
$t\bar{t}$ modeling	0.21	0.20	0.15	0.12
Large-R jet	0.01	0.01	< 0.01	< 0.01
Other experimental	< 0.01	< 0.01	< 0.01	< 0.01
b-tagging	< 0.01	< 0.01	< 0.01	< 0.01
Total Uncertainty	0.21	0.20	0.15	0.12

Table 5.3: The uncertainty on the scale factor measurement of the 50% fixed signal efficiency W tagger from each individual systematic uncertainty group. Each row shows the uncertainty obtained by adding in quadrature the impact of all uncertainties in the group. The total uncertainty is obtained by adding in quadrature the impact of all uncertainties.

Systematic Group	Contained top tagger p_{T} bins [GeV]				
	[350,400]	[400,450]	[450,500]	[500,600]	[600,1000]
Statistical	0.01	0.01	0.02	0.02	0.03
Theory	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
$t\bar{t}$ modeling	0.06	0.03	0.03	0.03	0.02
Large-R jet	0.01	0.01	0.02	0.01	0.02
Other experimental	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
b-tagging	< 0.01	< 0.01	< 0.01	< 0.01	0.02
Total Uncertainty	0.06	0.04	0.04	0.04	0.04

Table 5.4: The uncertainty on the scale factor measurement of the 80% fixed signal efficiency contained top tagger from each individual systematic uncertainty group. Each row shows the uncertainty obtained by adding in quadrature the impact of all uncertainties in the group. The total uncertainty is obtained by adding in quadrature the impact of all uncertainties.

Systematic Group	Inclusive top tagger p_{T} bins [GeV]				
	[350,400]	[400,450]	[450,500]	[500,600]	[600,1000]
Statistical	< 0.01	< 0.01	0.01	0.02	0.02
Theory	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
$t\bar{t}$ modeling	0.04	0.02	0.01	0.02	0.02
Large-R jet	< 0.01	0.01	< 0.01	0.01	< 0.01
Other experimental	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
b-tagging	< 0.01	< 0.01	< 0.01	0.01	0.02
Total Uncertainty	0.04	0.04	0.02	0.03	0.04

Table 5.5: The uncertainty on the scale factor measurement of the 80% fixed signal efficiency inclusive top tagger from each individual systematic uncertainty group. Each row shows the uncertainty obtained by adding in quadrature the impact of all uncertainties in the group. The total uncertainty is obtained by adding in quadrature the impact of all uncertainties.

Systematic Group	W tagger p_{T} bins [GeV]			
	[200,250]	[250,300]	[300,350]	[350,600]
Statistical	< 0.01	0.01	0.02	0.03
Theory	< 0.01	< 0.01	< 0.01	< 0.01
$t\bar{t}$ modeling	0.12	0.10	0.06	0.05
Large-R jet	< 0.01	< 0.01	0.01	< 0.01
Other experimental	< 0.01	< 0.01	< 0.01	< 0.01
b-tagging	< 0.01	< 0.01	< 0.01	< 0.01
Total Uncertainty	0.12	0.10	0.06	0.05

Table 5.6: The uncertainty on the scale factor measurement of the 80% fixed signal efficiency W tagger from each individual systematic uncertainty group. Each row shows the uncertainty obtained by adding in quadrature the impact of all uncertainties in the group. The total uncertainty is obtained by adding in quadrature the impact of all uncertainties.

5.2 Jet Tagging with Topological Data Analysis

This section presents an alternative approach to top jet tagging by using information from topological data analysis (TDA) that has not been used in the context of jet tagging before. TDA is a recent field of statistical analysis that utilizes concepts from algebraic topology to analyze data that has a notion of distance. The main driving hypothesis of TDA is that the data to be analyzed is sampled from an unknown topological manifold. The manifold can be fully characterized by its topological features, or homology, such as connected components and n-dimensional voids. The number of independent features of each homology class, known as Betti numbers, can be used to classify the manifold. The goal of TDA is to infer the Betti numbers from data. This is achieved by reconstructing an approximation of the underlying manifold, known as a simplicial complex, with the data. The simplicial complex consists of a collection of points, edges, triangles, and higher-dimensional polytopes that are formed with the notion of distance between datapoints. Once the simplicial complex is constructed, its simplicial homology is calculated as an approximation of the homology of the underlying manifold. The application of TDA to jet tagging is motivated by the geometric nature that jet topoclusters have. The topoclusters that are associated with a jet can be used as the input dataset into the TDA methodology to be analyzed on a jet-by-jet basis. The Betti numbers and other topological information of the jet can then be used as inputs for a jet tagger.

This section starts with the introduction of the necessary concepts to understand the TDA methodology. Two TDA tools that are used in the workflow to tag jets are presented. The first tool is a persistent homology (PH) analysis [64]. The construction of a simplicial complex is sensitive to the distance scale in the data. Some topological features can appear within a

specific distance scale as new objects that are introduced into the simplicial complex alter its homology. This raises the question of which of these topological features are statistically relevant to classify the underlying manifold. PH is used to determine an optimal distance scale to build the simplicial complex from the input topoclusters. This scale can be optimized to maximize the number of topological features that persist the most while simultaneously minimizing those that emerge within narrow distance scale windows. The second tool used in the workflow is the Mapper algorithm [65]. The Mapper algorithm analyzes the interplay between the simplicial homology of data and functions, also referred to as maps or filters, that are defined on the data. These functions can be used to highlight topological features relative to kinematic features of topoclusters. Finally, the result of applying this workflow to tag top jets from signal $W' \to tb$ processes against jets from QCD multijet background processes in MC is presented. The events used in this study are required to satisfy the event selection described in section 4.1. Additionally, signal jets from $W' \to tb$ processes are required to pass the contained top labeling criteria and signal top jet candidacy criteria that were described in subsection 5.1.3 and subsection 5.1.5, respectively. Two top tagging algorithms were developed that use the information from TDA: a deep neural network (DNN) tagger and a convolutional graph neural network (GNN) tagger. The performances of these two taggers are compared with the contained top DNN tagger that was discussed in the previous section of this Chapter.

5.2.1 Simplicial Complexes and Simplicial Homology

5.2.1.1 Definition of a Simplicial Complex

As previously discussed, the application of TDA techniques involves the construction of a simplicial complex from the input data. Given a finite dataset X, a simplicial complex K of X is a collection of subsets of X that satisfies the following two properties:

1. $\forall v \in X \implies \{v\} \in K$ (Inclusion of points.)

2. If $\sigma \in K$ and $\tau \subset \sigma \implies \tau \in K$ (Closed under the subset operation.)

The elements of K are known as simplices. Simplices are classified by their dimension, with a p-dimensional simplex being a subset of X that has $p+1$ elements. If σ is a p-dimensional simplex and $\tau \subset \sigma$ is a $p-1$-dimensional simplex, then τ is said to be a face of σ. Additionally, the set of all p-dimensional simplices is denoted as K_p. The standard nomenclature for simplices of dimensions 0 through 3 is to denote them as vertices, edges, triangles, and tetrahedra, respectively. The dimension of the simplicial complex K is defined as the maximum dimension of its simplices. An example of a 2-dimensional simplicial complex of a set with four elements is shown in Figure 5.13.

5.2.1.2 Constructing a Simplicial Complex

The process of constructing a simplicial complex is not unique. As a result, different families of complexes exist, each with unique properties and varying degrees of approximation of the underlying manifold. The Čech (Č) and Vietoris-Rips (VR) complexes will be discussed since they are used in the Mapper algorithm and PH studies presented in this Chapter, respectively.

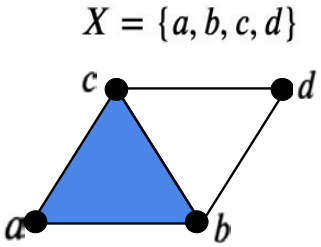


Figure 5.13: Example of a set X with four elements and a simplicial complex K of dimension 2 that is constructed from X. The simplicial complex contains four vertices corresponding to the individual elements of X, five edges corresponding to all possible subsets of X with two elements except for $\{a, d\}$, and one triangle, depicted by the blue shaded area, which corresponds to the subset with three elements $\{a, b, c\}$.

In order to define the Čech complex, the notion of a covering set of a topological space and the nerve of the covering set must be defined first. A covering set $\mathcal{U} = \{U_i\}_{i \in \mathrm{I}}$ of a topological space X is defined as an indexed family of subsets U_i of X with indexing set I, such that for all elements x_j of X there exists at least one cover element U_j that contains x_j. The nerve of a cover is the collection of finite subsets of indices in I corresponding to elements of $\mathcal{U}$ with non-empty intersection. The Čech complex is defined as the nerve of a covering set $\mathcal{U}$ of a topological space. Thus, $\sigma = \{i_0, \cdots, i_p\} \in$ Č is a p-dimensional simplex if $\bigcap_{j=0}^{p} U_{i_j} \neq \emptyset$. In TDA applications, the cover $\mathcal{U}$ is usually taken as the collection of ϵ-spheres that are centered at each datapoint. An ϵ-sphere centered at a point x is defined as the set of all points y within distance ϵ from x in Euclidean n-space. The parameter ϵ is the distance scale that parametrizes the construction of the Čech complex. An example of a topological manifold with a Čech complex that is identical to the example depicted in Figure 5.13 is shown in Figure 5.14.

The VR complex parametrized with a distance scale ϵ is defined as the clique-complex

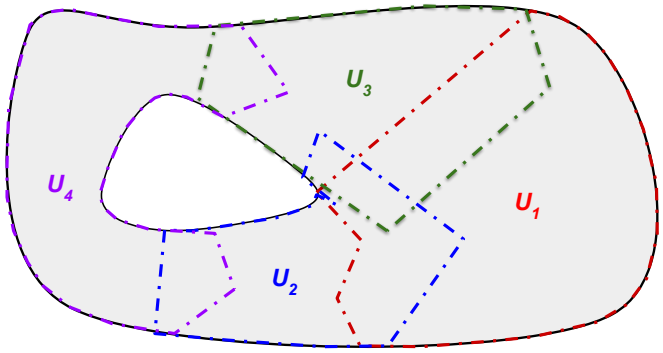


Figure 5.14: Example of a topological manifold with a cover given by $\mathcal{U} = \{U_1, U_2, U_3, U_4\}$. The colored dashed-dotted lines represent the boundaries of each cover element. Note that $U_1 \bigcap U_2 \bigcap U_3 \neq \emptyset$, $U_2 \bigcap U_4 \neq \emptyset$, and $U_3 \bigcap U_4 \neq \emptyset$. By relabeling the cover elements as $U_1 \rightarrow a$, $U_2 \rightarrow b$, $U_3 \rightarrow c$, and $U_4 \rightarrow d$, the simplicial complex shown in Figure 5.13 is obtained. In this example the manifold consists of a single connected component which encompasses a circular void.

of an ϵ-neighborhood graph [66]. The graph can be built from the data with the vertices representing individual datapoints. Edges connect two vertices if the distance between datapoints is less than ϵ. Higher-dimensional simplices are included in the VR complex if all the edges associated with the simplex are in the graph. For example, if a VR complex is constructed from the nerve of the manifold shown in Figure 5.14, then it would also contain the 2-dimensional simplex $\{b, c, d\}$.

From this discussion, it is clear that the Čech and VR complexes satisfy the properties of a simplicial complex that were previously discussed. It should be noted that the approximation of the underlying manifold, and consequently its homology, obtained from the Čech complex improves as the granularity of the cover decreases. On the other hand, the VR complex provides an approximation of the simplicial homology of Čech. Specifically, it can be shown that $\check{\mathrm{C}}_\epsilon \subseteq \mathrm{VR}_\epsilon \subseteq \check{\mathrm{C}}_{\sqrt{2}\epsilon}$. The VR complex is used in the PH studies since it is computationally more efficient to construct compared to the Čech complex.

5.2.1.3 Computing Simplicial Homology

To compute the simplicial homology of a simplicial complex, a relationship between the topological features of the underlying manifold and those of the simplicial complex must be established. This will be achieved with the introduction of Homology groups, which are vector spaces that represent the topological features in each dimension of a manifold as vectors. This will allow us to determine the number of unique, up to path deformation, topological features from the dimensions of these vector spaces.

We first start by introducing the notion of the boundary of a simplex through boundary linear transformations $\partial_p : V(K_p) \rightarrow V(K_{p-1})$, where $V(K_p) = \text{Span}(K_p, \mathbb{F})$ is the vector space spanned by the set of p-dimensional simplices over a field $\mathbb{F}$, which will be left unspecified for the moment. The purpose of the transformation ∂_p is to establish a linear relationship between a p-dimensional simplex σ and its faces τ, such that the result of applying the transformation corresponds to the notion of the region boundary that is encapsulated by σ on the manifold, with its faces forming the boundary. These transformations must preserve the topological features that are bounded by linear combinations of simplices. Additionally, these transformations must satisfy the constraint on their functional composition $\partial_{p-1} \circ \partial_p = 0$, which indicates that the boundary of a boundary is empty. The exact form of these transformations depends on the field $\mathbb{F}$, as different fields can take into account effects such as whether the manifold has a well-defined orientation or not. Throughout the remainder of this Chapter, the field $\mathbb{F}$ is taken as the field with two elements, $\mathbb{Z}_2 = \{0, 1\}$, due to its simplicity in implementation. A full discussion of computing simplicial homology in other fields is outside the scope of this thesis. Under the field $\mathbb{Z}_2$, the boundary transformation

takes the form:

$$\partial_p(\sigma) = \sum_{\tau \subset \sigma,\, \tau \in K_{p-1}} \tau \tag{5.9}$$

In the case where $p = 0$ or is greater than the dimension of the simplicial complex K, then ∂_p is defined as the zero map.

After introducing the concept of the boundary of a simplex, we are now in a position to define the concepts of p-boundaries and p-cycles. Both p-boundaries and p-cycles correspond to path components in the manifold that form closed p-dimensional loops. The elements of the null subspace $\ker(\partial_p)$ are known as p-cycles since all closed paths map to zero in a lower dimension. The elements of the image subspace $\mathrm{Im}(\partial_{p+1})$ are known as p-boundaries since they bound higher-dimensional simplices. From the constraint $\partial_p \circ \partial_{p+1} = 0$, it can be seen that $\mathrm{Im}(\partial_{p+1})$ is fully contained within $\ker(\partial_p)$. All p-cycles that are not p-boundaries correspond to p-dimensional voids in the manifold since there are no higher-dimensional simplices that are encompassed by the p-cycle. These two subspaces are the essential ingredients in the definition of Homology groups. As an example, the subspaces $\ker(\partial_1)$ and $\mathrm{Im}(\partial_2)$ from the simplicial complex in Figure 5.13 are shown in Table 5.7.

$\ker(\partial_1)$	$\mathrm{Im}(\partial_2)$
$\{a,b\} + \{a,c\} + \{b,c\}$	
$\{b,c\} + \{b,d\} + \{c,d\}$	$\{a,b\} + \{a,c\} + \{b,c\}$
$\{a,b\} + \{a,c\} + \{b,d\} + \{c,d\}$	

Table 5.7: The elements of the subspaces $\ker(\partial_1)$ and $\mathrm{Im}(\partial_2)$ of the simplicial complex shown in Figure 5.13. Note that $\ker(\partial_1)$ is a two-dimensional subspace since the first two rows add to the third row with the $\mathbb{Z}_2$ algebra, and $\mathrm{Im}(\partial_2)$ is a one-dimensional subspace. Furthermore, $\{a,b\} + \{a,c\} + \{b,c\}$ is a 1-boundary while $\{b,c\} + \{b,d\} + \{c,d\}$ is a 1-cycle that is not a boundary.

Since the topological features of the simplicial complex are unique up to path deformation, the computation of simplicial homology counts the instances of independent features. This

is achieved by defining the quotient vector space $\mathrm{H}_p = \ker(\partial_p)/\mathrm{Im}(\partial_{p+1})$, known as the p^{th} Homology group. The elements of H_p are the equivalence classes of p-cycles that represent unique p-dimensional topological features up to path deformation. Consequently, the p^{th} Betti number β_p is defined as the dimension of H_p:

$$\beta_p = \dim(\mathrm{H}_p) = \dim(\ker(\partial_p)) - \dim(\mathrm{Im}(\partial_{p+1})) \tag{5.10}$$

The standard procedure to calculate the Betti numbers is to obtain the matrix representations of the linear transformations ∂_p in $\mathbb{Z}_2$, and perform Gaussian elimination to determine the rank and nullity of the matrices. To finalize the example of the simplicial complex in Figure 5.13, the matrix representation of the boundary transformation ∂_1 is given by:

$$\partial_1 = \left(\begin{array}{c|ccccc} & \{a,b\} & \{a,c\} & \{b,c\} & \{b,d\} & \{c,d\} \\ \hline \{a\} & 1 & 1 & 0 & 0 & 0 \\ \{b\} & 1 & 0 & 1 & 1 & 0 \\ \{c\} & 0 & 1 & 1 & 0 & 1 \\ \{d\} & 0 & 0 & 0 & 1 & 1 \end{array}\right) \tag{5.11}$$

This representation is defined by the ordered basis of $V(K_0)$ and $V(K_1)$, which are shown to the left of the vertical line and above the horizontal line in Equation 5.11, respectively.

After performing Gaussian elimination on its columns, the matrix reduces to:

$$\partial_1 = \left(\begin{array}{c|ccccc} & \{a,b\} & \{a,c\} & \{a,b\}+\{a,c\}+\{b,c\} & \{b,d\} & \{b,c\}+\{b,d\}+\{c,d\} \\ \hline \{a\} & 1 & 1 & 0 & 0 & 0 \\ \{b\} & 1 & 0 & 0 & 1 & 0 \\ \{c\} & 0 & 1 & 0 & 0 & 0 \\ \{d\} & 0 & 0 & 0 & 1 & 0 \end{array}\right) \tag{5.12}$$

Similarly, the boundary transformation ∂_2 matrix representation is given by:

$$\partial_2 = \left(\begin{array}{c|c} & \{a,b,c\} \\ \hline \{a,b\} & 1 \\ \{a,c\} & 1 \\ \{b,c\} & 1 \\ \{b,d\} & 0 \\ \{c,d\} & 0 \end{array}\right) \tag{5.13}$$

As it can be observed in Equation 5.12, the rank of the matrix is 3. To determine the value of β_0 we must know $\dim(\ker(\partial_0))$, but since ∂_0 is the zero map, its null space is $V(K_0)$. Thus, $\beta_0 = \dim(V(K_0)) - \dim(\mathrm{Im}(\partial_1)) = 4 - 3 = 1$, which implies that there is a single connected component in the simplicial complex. Similarly, from Equation 5.13, the rank of the matrix is trivially equal to 1. To determine β_1 we use the Rank-Nullity theorem to obtain $\dim(\ker(\partial_1)) = \dim(V(K_1)) - \dim(\mathrm{Im}(\partial_1)) = 5 - 3 = 2$. Thus, we get that $\beta_1 = \dim(\ker(\partial_1)) - \dim(\mathrm{Im}(\partial_2)) = 2 - 1 = 1$, which implies that there is a single circular void in the simplicial complex. Both calculations give the correct number of topological

features of the underlying manifold and the simplicial complex.

5.2.1.4 Filtered Simplicial Complex and Persistent Homology

Up to this point in the discussion, it has been assumed that the simplicial complex is fixed in structure. This limits the simplicial homology analysis of the data to a fixed configuration of the distance scale parameter ϵ. To analyze the simplicial homology as a function of the distance scale, it is necessary to introduce a final construction known as a filtered simplicial complex. This is the central object that drives the PH analysis. A filtered simplicial complex is defined as a finite sequence of nested simplicial complexes, $\{K^i\}_{i \leq N_\epsilon}$, where N_ϵ is the number of filtration steps, which will depend on the distance scale parameter, and $K^i \subset K^j$ if $i < j$. The index i is used to denote the filtration step of the sequence, with larger indices corresponding to larger values of the distance scale parameter ϵ. This allows the definition of inclusion maps, $f_{i \leq j} : \mathrm{H}_p^i \rightarrow \mathrm{H}_p^j$, that give information on how the Betti numbers change between filtration steps. Each new filtration step brings new simplices into consideration. The p^{th} Betti number increases if new path independent p-cycles that are not p-boundaries are formed. Otherwise, the p^{th} Betti number decreases if previous voids are filled in with new simplices. An example of a filtered simplicial complex with four filtration steps is shown in Figure 5.15. In this example, the number of connected components changes along the filtration as $\beta_0 = 4 \rightarrow 1 \rightarrow 1 \rightarrow 1$. The number of circular voids changes as $\beta_1 = 0 \rightarrow 1 \rightarrow 2 \rightarrow 1$. These results are usually interpreted as persistence diagrams, which is shown in Figure 5.16 for the filtered simplicial complex shown in Figure 5.15.

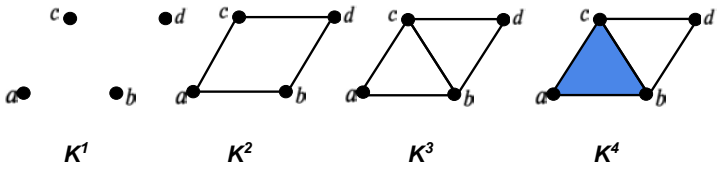


Figure 5.15: Example of a filtered simplicial complex with four filtration steps. The first step consists of four individual points. In the second step, points are pairwise connected so that a single connected component that encompasses a void is formed. In the third step, points b and c are connected, resulting in the creation of a new void. The final filtration step is the same simplicial complex shown in Figure 5.13, which is obtained after filling one of the voids.

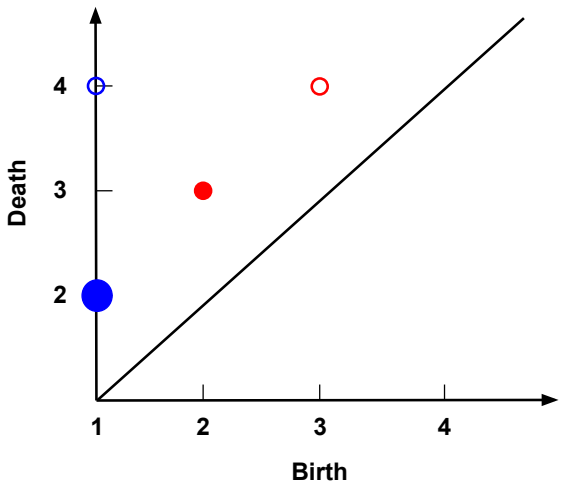


Figure 5.16: Persistence diagram that summarizes the simplicial homology of the filtered simplicial complex shown in 5.15. The horizontal axis, denoted as birth, indicates the filtration step at which a topological feature enters in the simplicial complex. The vertical axis, denoted as death, indicates the filtration step at which a topological feature ceases to exist in the simplicial complex. Topological features are represented as (birth,death) points. The blue points correspond to connected components, while the red points correspond to circular voids. The size of each point is proportional to the Betti number of the topological feature at the corresponding filtration step. Closed points correspond to features that died before the final filtration step. Open points correspond to features that persisted until the final filtration step.

5.2.2 Persistent Homology Studies

In this section, the large-R jets from $W' \to tb$ and QCD multijet events are analyzed with the PH algorithm on a jet-by-jet basis. The jets used in this study are required to have $p_{\mathrm{T}} > 350$ GeV and at least 10 topoclusters. The selection requirement on the number of topoclusters is made since the topoclusters will be used as the inputs to the PH algorithm, thereby removing jets that will not have an interesting topology associated with their topoclusters. The signal jets in this study are large-R jets from the $W' \to tb$ process that pass the contained top jet label requirement, as described in subsection 5.1.3, while all large-R jets from the QCD multijet process are background jets. All topoclusters of a jet are boosted to the center of momentum (CoM) frame of the jet prior to being analyzed with the PH algorithm. This is done so that the PH algorithm processes jets with different levels of collimation on an equal basis. After this preprocessing step, the pseudorapidity and azimuthal angle pairs, (η, ϕ), of each topocluster in the jet are used to build the VR complex by treating each coordinate pair as a vertex of the VR complex. The VR complex is then extended to a filtered VR complex in order to analyze its simplicial homology with the PH algorithm. The processing of the topoclusters of a jet through the PH algorithm is summarized in the following steps:

1. Build the ϵ-neighborhood graph of the jet using the topocluster (η, ϕ) coordinate pairs as the vertices of the graph.

2. Define edges $e_{i,j}$ between all possible topocluster pairs (t_i, t_j) and assign a weight $\omega_{i,j} = \Delta R(t_i, t_j)$ to each edge.

3. Build the VR complex from the ϵ-neighborhood graph by including all simplices up to dimension 2. For each simplex of dimension 2, define its weight as the maximum

weight of all its edges.

4. Construct the filtered VR complex by sorting the weights $\omega_{i,j}$ in ascending order. A filtration step is introduced for each distinct weight value.

5. Calculate the boundary linear transformation matrices in all filtration steps in order to obtain the Betti numbers, as discussed in subsection 5.2.1.

6. Build the persistence diagram of the jet for β_0 and β_1.

Since the topoclusters are represented as two-dimensional coordinate pairs, the only meaningful Betti numbers that can be extracted from the PH analysis are the number of connected components, β_0, and the number of circular voids, β_1. The persistence of the simplicial homology of jets is summarized in the plots shown in Figures 5.17. The β_0 maximum persistence length is the ΔR scale at which all topoclusters in the jet form a single connected component. This scale is analogous to reconstructing the jet from its topoclusters. The β_1 maximum persistence length is defined as the maximal ΔR interval length that a circular void achieves in the filtered VR complex of the jet. Specifically, this is defined as the difference between the ΔR scale at which the void disappears from the filtered VR complex (death scale) and the scale at which the void is introduced in the filtered VR complex (birth scale) that is maximal. As observed in the plots, on average, signal jets become a single connected component at lower distance scales compared to background jets. Both signal and background jets populate the same two regions of the persistence diagram of the most persistent circular void. The majority of jets populate the upper region of the diagram, which corresponds to jets that have their most persistent circular void appearing late in the filtration and disappearing after the topoclusters form a single connected component. The lower region of the diagram contains jets that populate the region close to the death=birth

region for low values of the birth scale, which corresponds to jets with circular voids that appear early in the filtration and are short-lived. The fraction of background jets that populate the lower region of the diagram is larger compared to signal jets. Additionally, on average, the most persistent circular void in signal jets persists longer when compared to background jets.

Based on these observations, the topoclusters in signal jets appear to be clustered along filament-like structures that match in direction with the prong structures that are formed by the top decays in the CoM frame of the jet. Since the topoclusters are clustered along well-defined structures, the jet can be reconstructed as a single connected component at smaller ΔR scales. Additionally, any circular void that is formed in signal jets would be in between the prong structures of the jet. On the other hand, the observations made for background jets could be indicative of the topoclusters being distributed amorphously in the CoM frame of the jet. Since there is no well-defined structure, the jet is reconstructed at larger ΔR scales. This could explain why signal jets have, on average, a smaller β_0 maximum persistence length compared to background jets. Additionally, all short-lived circular voids in background jets could be explained as noise from dispersed topoclusters.

To verify these claims, the three- and two-pronged substructures of signal top jets are taken as hypotheses. These two cases correspond to a resolved top decay and to a collimated decay of the W boson, respectively. To achieve this, the kinematic features of the connected components (CCs) are analyzed separately when there are exactly three and two CCs in the filtered VR complex of the jet. To obtain the kinematic description of a CC, the four-momenta of the topoclusters associated with the CC are added. Thus, the CCs can be interpreted as subjets that originate the prong structures in the large-R jet for each hypothesis. The distributions of ΔR scales at which the filtered VR complexes of jets contain

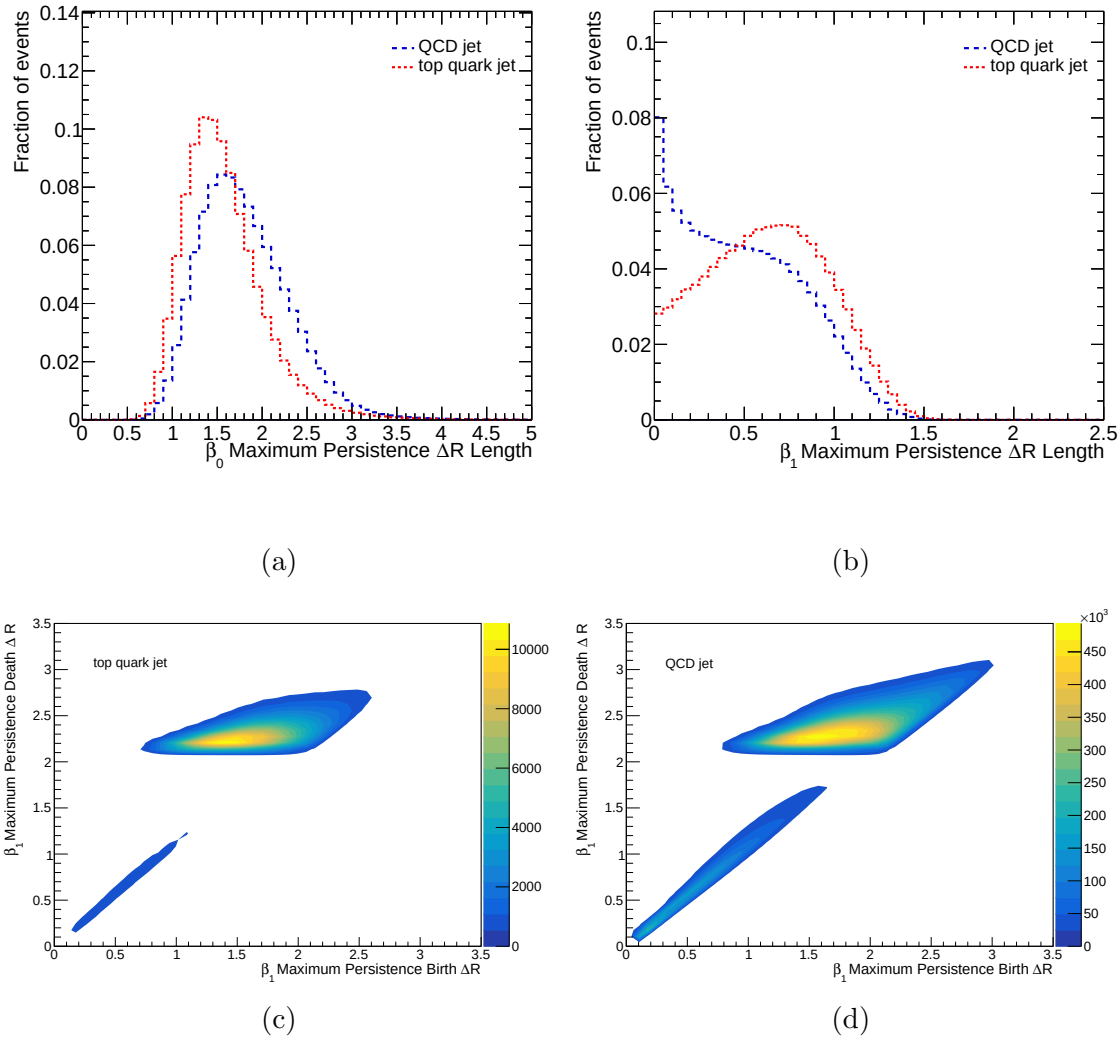
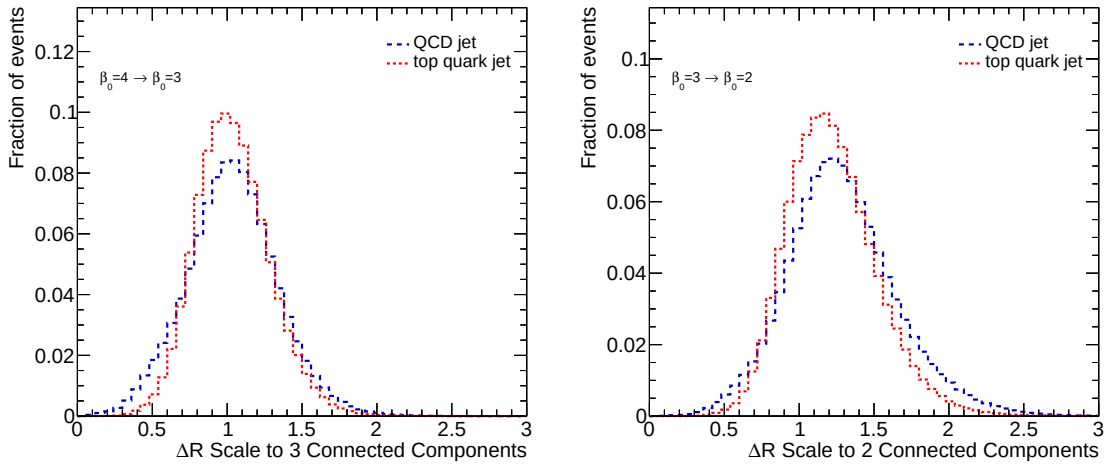


Figure 5.17: The plots in (a) and (b) show the ΔR interval length of the most persistent connected component and circular void of all jets analyzed with the PH algorithm, respectively. The plots in (c) and (d) show the cumulative persistence diagram of the most persistent circular void for signal top jets and background QCD jets, respectively. This corresponds to taking the point of the β_1 persistence diagram of each jet that maximizes the persistence length. The horizontal axis of the persistence diagrams represent the ΔR scale at which the circular void appears in the filtered VR complex of the jet ("birth"), while the vertical axis represents the ΔR scale at which the void disappears from the filtered VR complex ("death").

exactly three and two CCs are shown in Figure 5.18.

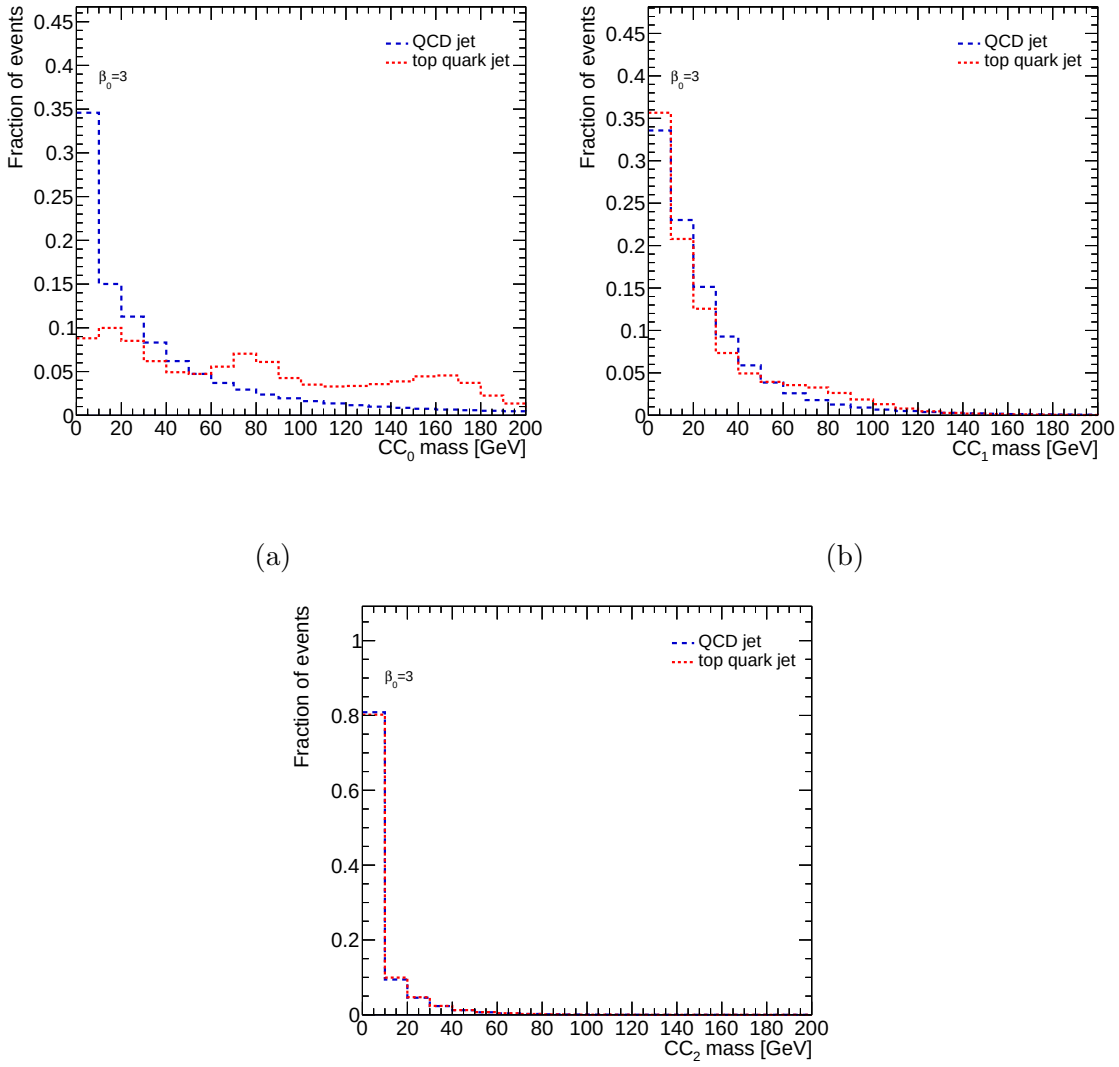


(a) (b)

Figure 5.18: The distributions of the ΔR length scales at which the filtered VR complexes of jets have exactly three connected components (a) and two connected components (b) compared between signal and background jets.

The mass distributions of the CCs after being sorted in descending order by their p_{T} are shown in Figures 5.19 and 5.20 when there are three CCs and two CCs, respectively. The three CCs have reconstructed some of the relevant substructure in signal jets when assuming the three-pronged top jet hypothesis. As can be observed in Figure 5.19, the leading CC shows bumps in the mass distribution near the W and top mass. The subleading CC shows a small bump close to the W mass, while the mass distribution of the third leading CC could correspond to reconstructing the b quark or one of the quarks from the W. These observations indicate that a $\beta_0 = 3$ jet topology has partially resolved some of the relevant substructure in signal top jets. The mass bumps in the leading CC become more prominent when assuming the two-pronged top jet hypothesis. Additionally, the subleading CC mass bump near the W mass becomes slightly more prominent. For background jets, the mass distributions of the CCs peak at lower values and exhibit a long tail at higher values that

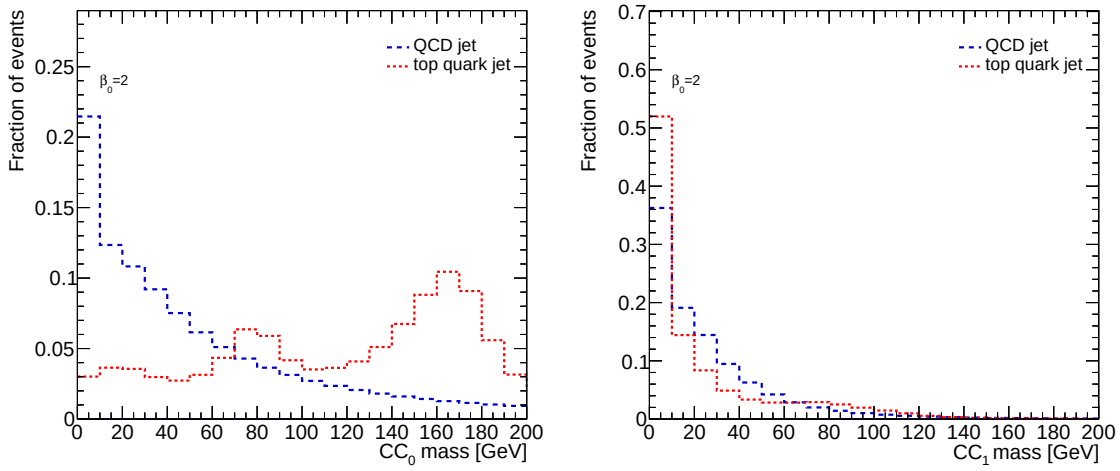
lacks prominent structures like those present in the CCs of signal jets. This implies that the CCs are reconstructing random patterns from the topoclusters of background jets. From these observations, the scale $\Delta R = 1.2$, which is approximately equal to the mean of the ΔR length scale distribution when there are two CCs in signal jets, provides a good qualitative description between signal and background jets. The CCs have reconstructed most of the interesting substructure of signal jets at this scale. The value of this distance scale is used as an input parameter to the Mapper algorithm, as will be discussed in the next section.



(a) (b)

(c)

Figure 5.19: The mass distributions of the leading in p_{T} connected component (CC_0) (a), the second leading in p_{T} connected component (CC_1) (b), and the third leading in p_{T} connected component (CC_2) (c) in the filtration step of the filtered VR complex when there are exactly three CCs. The distributions are compared between signal and background jets.



(a) (b)

Figure 5.20: The mass distributions of the leading in p_{T} connected component (CC_0) (a) and the second leading in p_{T} connected component (CC_1) (b) in the filtration step of the filtered VR complex when there are exactly two CCs. The distributions are compared between signal and background jets.

5.2.3 Mapper Algorithm Studies

The next step in the TDA workflow is to analyze jets with the Mapper algorithm. The mapper algorithm will allow us to analyze the interplay between the simplicial homology of a jet and the kinematic features of the topoclusters in the jet. This will be achieved through the use of continuous filtering functions that map the topoclusters from their underlying manifold to a known image topological space where its simplicial homology can be analyzed. Unlike the PH analysis study presented in the previous section, the Mapper algorithm analyzes the jet at a fixed distance scale, which is known as the resolution scale (ΔR_{res}) of the algorithm. Another parameter that needs to be provided to the Mapper algorithm is a finite covering set for the topological space that the topoclusters are mapped onto. The elements of the covering set will be allowed to overlap so that a topocluster has the possibility of being mapped onto multiple cover elements. As discussed in subsubsection 5.2.1.2, this will allow us to define a non-trivial nerve of the cover from which the Čech complex of the jet can be constructed. The topoclusters will be spatially clustered in each cover element using ΔR_{res} as the clustering distance threshold. The clusters of topoclusters will form the vertices of the Čech complex. The vertices correspond to collections of topoclusters that are spatially near within ΔR_{res} and have a similar response to the filter function since they are mapped to the same cover element. The n-dimensional simplices are obtained from n+1 vertices that share at least one topocluster in common. The higher-dimensional simplices allow us to study how the topocluster response to the filter function transitions along path connected components in the image topological space. Once the Čech complex of the jet is built, its simplicial homology is calculated. Since the filter functions are assumed to be continuous, the simplicial homology obtained from the Čech complex of the jet in the image topological

space is the same as the one from the underlying manifold of the jet.

For the studies presented in this section, the filter function used is the ϕ-projection of the topocluster in the η-ϕ plane. Thus, the topoclusters are mapped to a topological space that corresponds to an arc of a ring. The covering set chosen for this space is the set of overlapping intervals given by:

$$\mathcal{U} = \{[-3.2, -1.2], [-2.0, 0.4], [-0.4, 2.0], [1.2, 3.2]\} \tag{5.14}$$

The topoclusters that are mapped onto each cover element are spatially clustered using a single-linkage clustering algorithm, which defines the distance between two clusters v_n and v_m of topoclusters as:

$$\Delta R(\mathrm{v}_n, \mathrm{v}_m) = \min_{t_i \in \mathrm{v}_n, t_j \in \mathrm{v}_m} \Delta R(t_i, t_j) \tag{5.15}$$

The motivation behind this choice of clustering algorithm is that two clusters of topoclusters are merged at a given clustering step if they achieve the minimal distance between topoclusters that are not in the same cluster, which is similar in behavior to how the topoclusters are aggregated to form CCs by the PH algorithm. The clustering process in a given cover element is stopped after all remaining clusters have a single-linkage distance greater than the resolution scale, which is set to $\Delta R_{\mathrm{res}} = 1.2$ as motivated at the end of the preceding section. A detailed study of the optimization of the Mapper algorithm by using other filter functions and parameter options can be found in Appendix B. The processing of the topoclusters of a jet through the Mapper algorithm is summarized in the following steps:

1. For each topocluster in the jet, evaluate its ϕ-projection and map it to the corresponding cover elements from Equation 5.14.

2. For each cover element, apply the single-linkage clustering algorithm to the topoclusters that are mapped onto the cover element. The clustering process is stopped once all clusters of topoclusters have a single-linkage distance greater than $\Delta R_{\text{res}} = 1.2$. The resulting clusters of topoclusters will form the vertices of the Čech complex of the jet.

3. Construct the Čech complex of the jet from the nerve of the cover by checking which vertices in consecutive cover elements share at least one topocluster. Only simplices up to dimension 2 will be included in the Čech complex.

4. Evaluate the simplicial homology of the Čech complex of the jet.

Similar to the PH analysis study, since the topoclusters are represented as two-dimensional objects, the only meaningful Betti numbers that can be extracted in this study are β_0 and β_1. The CCs that are obtained from the Mapper algorithm correspond to vertices that form path components from sharing topoclusters. These CCs are interpreted as subjets by adding the four-momenta of all distinct topoclusters that are associated with a given CC, similar to how it was done in the PH analysis study. The circular voids correspond to regions in the η-ϕ plane where the path components branch off due to a deficit of topoclusters for a given range of values of η and then merge back to a single branch. An example event display demonstrating a signal top jet being processed through the first three steps of the Mapper algorithm is shown in Figure 5.21.

The distributions of the Betti numbers β_0 and β_1 of jets are shown in Figure 5.22. As can be observed from these plots, the topology of both signal and background jets is characterized by the presence of multiple CCs, with the majority of jets populating the $\beta_0 = 1 - 4$ region, and a lack of circular voids. The absence of circular voids could be a side effect of using the ϕ-projection as the filtering function with a resolution scale of $\Delta R_{\text{res}} = 1.2$ and a covering

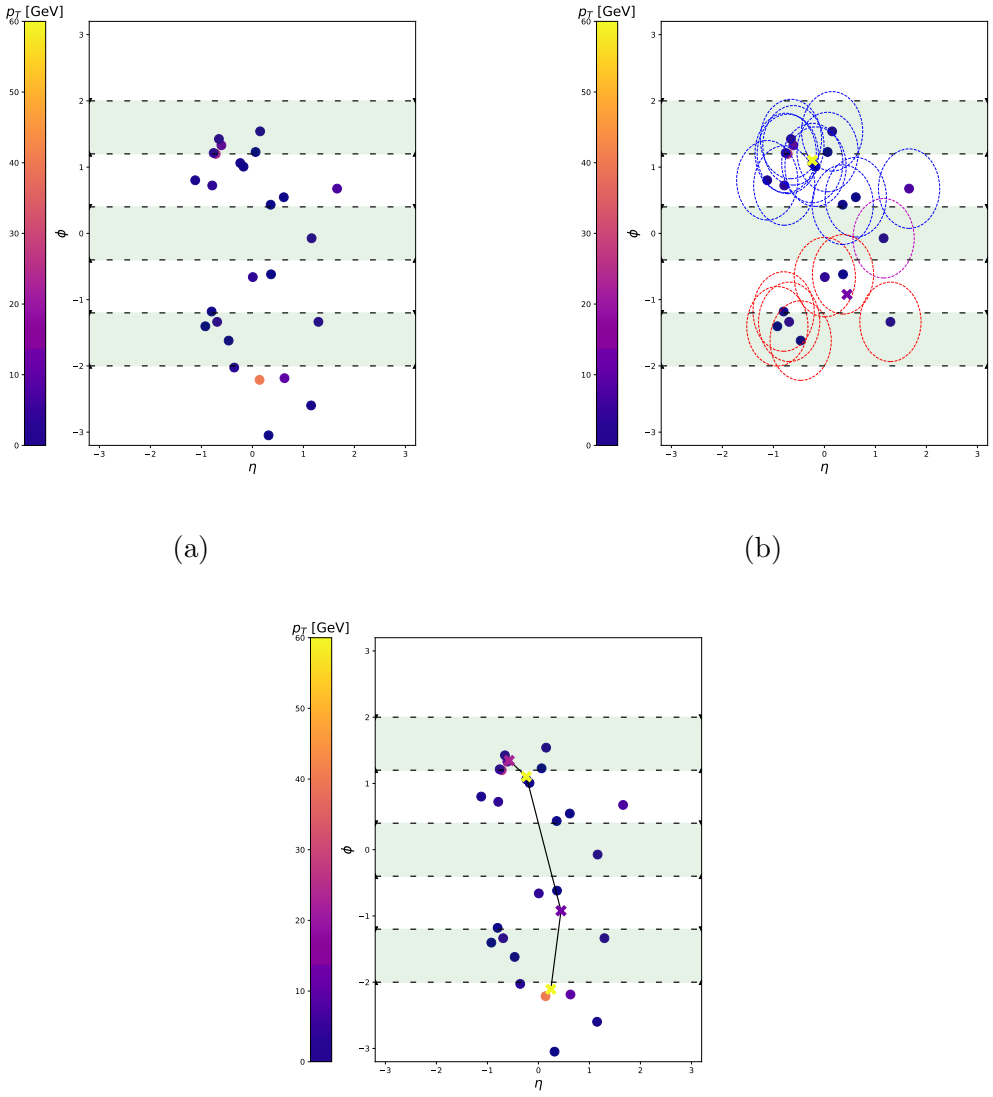
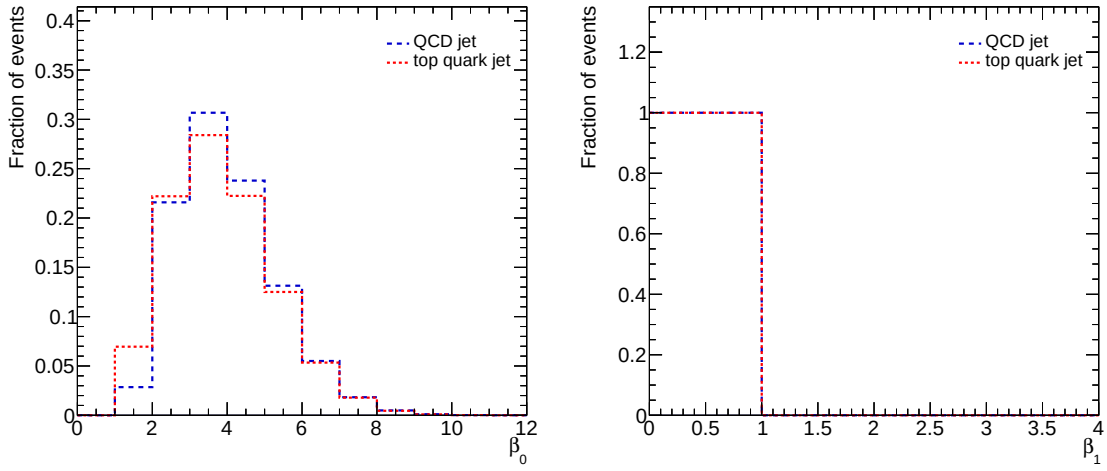


Figure 5.21: The topoclusters of a signal top jet in the CoM frame of the jet are represented as (η,ϕ) coordinate pairs in the η-ϕ plane, as shown in (a). The coordinate pairs are color coded based on the p_{T} of the topoclusters. The topoclusters are mapped onto the intervals of the covering set in Equation 5.14. The light green shaded regions represent the overlap regions of the cover elements. The single-linkage clustering of the topoclusters that are mapped onto the intervals $[-2.0, 0.4]$ and $[-0.4, 2.0]$ is shown in (b). A circle of radius $R = \Delta R_{\mathrm{res}}/2 = 0.6$ is drawn centered around each topocluster. The red and blue circles correspond to the topoclusters that are mapped exclusively onto the intervals $[-2.0, 0.4]$ and $[-0.4, 2.0]$, respectively, while the purple circles correspond to topoclusters that are mapped onto both intervals. All topoclusters that have overlapping circles in a given interval form a vertex of the Čech complex. In this event, a single vertex is formed in each of these two intervals. The "x" marks represent the coordinates of the vertices after adding the four-momenta of the topoclusters associated with the vertex, with their color corresponding to the p_{T} scale. The Čech simplicial complex is constructed after forming edges, represented by the black lines, between vertices from different cover intervals that have at least one topocluster in common, as shown in (c). The vertices obtained from the remaining intervals are shown in this step. The end result is a jet that has a single connected component and no circular voids.

set that is very granular, thereby reducing the ability of the Mapper algorithm to resolve circular features in the jets.



(a) (b)

Figure 5.22: Comparison of the number of connected components (a) and the number of circular voids (b) between signal top jets and background QCD jets after being processed through the Mapper algorithm.

Two metrics are calculated in order to quantify how the topoclusters are distributed in each CC. The first metric is the average ΔR in the CoM frame of the jet between the CC and the topoclusters associated with it ($\Delta R_{\text{avg}}(\text{CC}, t)$). This metric quantifies the effective size of the CC by measuring how displaced the topoclusters are from the axis of the CC. Large values of $\Delta R_{\text{avg}}(\text{CC}, t)$ indicate that the CC has a large fraction of topoclusters far from the CC axis, while small values indicate that the topoclusters are distributed close to the CC axis. The second metric is the average ΔR in the CoM frame of the jet between all possible topocluster pairs that are associated with a given CC ($\Delta R_{\text{avg}}(t_i, t_j \in \text{CC})$). This metric quantifies the eccentricity of the topocluster distribution in the CC. Large values of $\Delta R_{\text{avg}}(t_i, t_j \in \text{CC})$ indicate that the topoclusters in the CC are distributed along large filament-like structures, while small values indicate that the topoclusters are densely distributed in the CC. The

distributions of these two metrics evaluated on the leading (CC_0) and subleading (CC_1) connected components are shown in Figures 5.23 and 5.24, respectively. The distributions are shown separately for jets that have a low number of CCs ($\beta_0 = 2-3$) and a high number of CCs ($\beta_0 \geq 4$) in order to highlight any effects that the value of β_0 may have on these metrics. As can be observed from the $\Delta R_{\text{avg}}(\text{CC}_0, t)$ distribution, the CC_0 in signal top jets tends to be larger in effective size when compared to background QCD jets. Additionally, from the $\Delta R_{\text{avg}}(t_i, t_j \in \text{CC}_0)$ distribution, it is observed that the topoclusters in CC_0 from signal top jets are more eccentrically distributed when compared to background QCD jets. These observations are independent on the value of β_0 of the jet. However, the effective size and eccentricity of CC_0 from jets with $\beta_0 \geq 4$ are slightly smaller when compared to jets with $\beta_0 = 2-3$, which may be due to CC_0 containing a lower fraction of topoclusters in the former case. These observations are indicative that the topoclusters in CC_0 from signal top jets are spatially spread out, forming long path-connected structures. On the other hand, the topoclusters in CC_0 from background QCD jets are densely distributed, forming smaller structures. No significant differences are observed in the distributions for CC_1 between signal and background jets.

The inclusive distributions of mass and p_{T} of CCs from jets with $\beta_0 = 1-3$ are shown in Figure 5.25. The mass distribution exhibits bumps close to the W and top mass, which gives confidence in the top jet substructure reconstruction with the CCs. On the other hand, the mass distribution in background QCD jets peaks at lower values and exhibits a long tail, which is consistent with the CCs reconstructing objects from random patterns of topoclusters. In order to extend the kinematic description of the CCs, variables that are inspired from the jet substructure observables discussed in subsection 5.1.1 are defined using the CCs as subjets. The n-subjettiness variables are obtained by calculating the ΔR distance

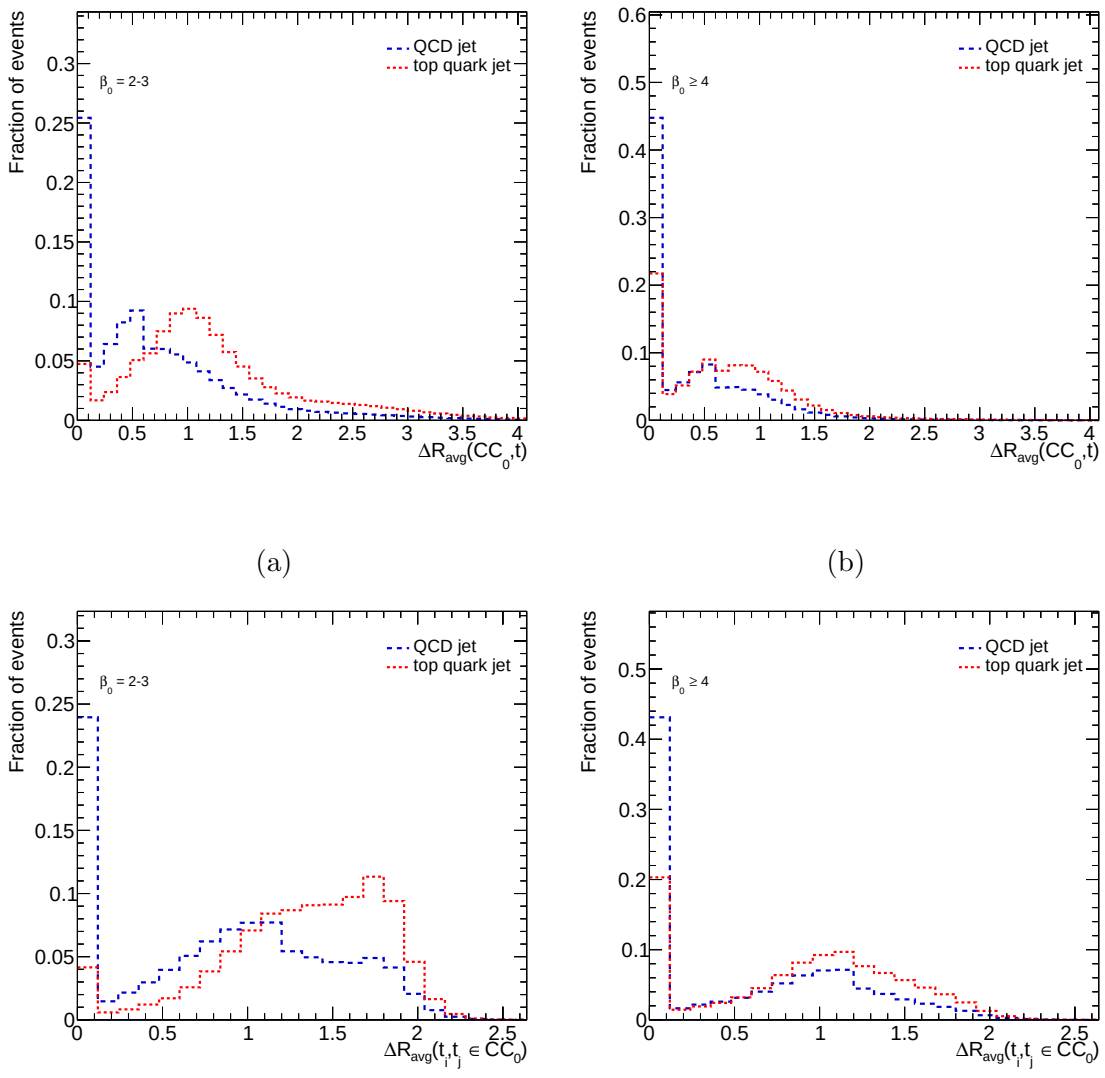


Figure 5.23: Comparison between signal top jets and background QCD jets of the metrics $\Delta R_{\text{avg}}(\text{CC}_0, t)$ and $\Delta R_{\text{avg}}(t_i, t_j \in \text{CC}_0)$. The distribution of $\Delta R_{\text{avg}}(\text{CC}_0, t)$ and $\Delta R_{\text{avg}}(t_i, t_j \in \text{CC}_0)$ are shown for jets with $\beta_0 = 2 - 3$ in (a) - (c) and for jets with $\beta_0 \geq 4$ in (b) - (d).

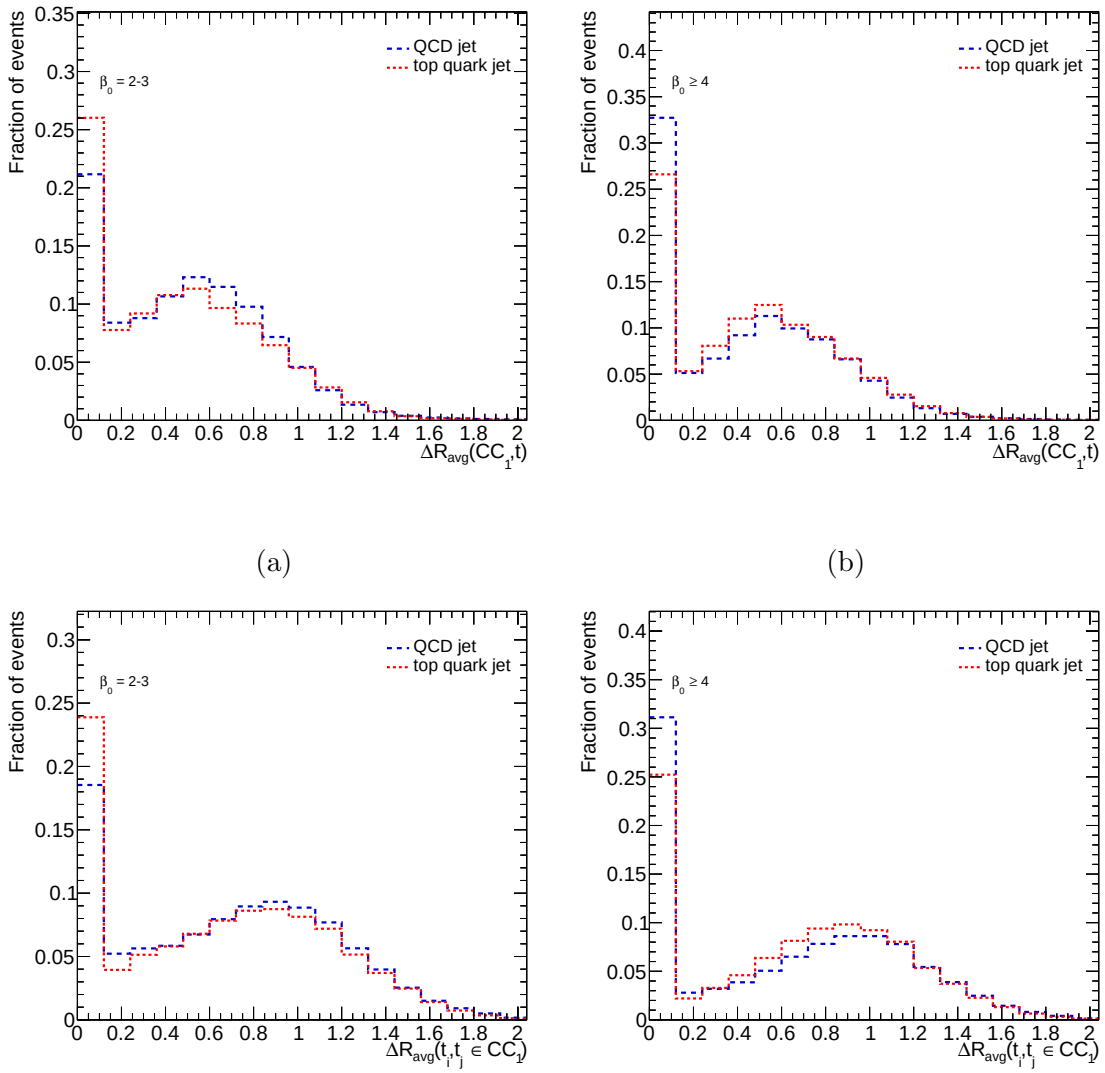
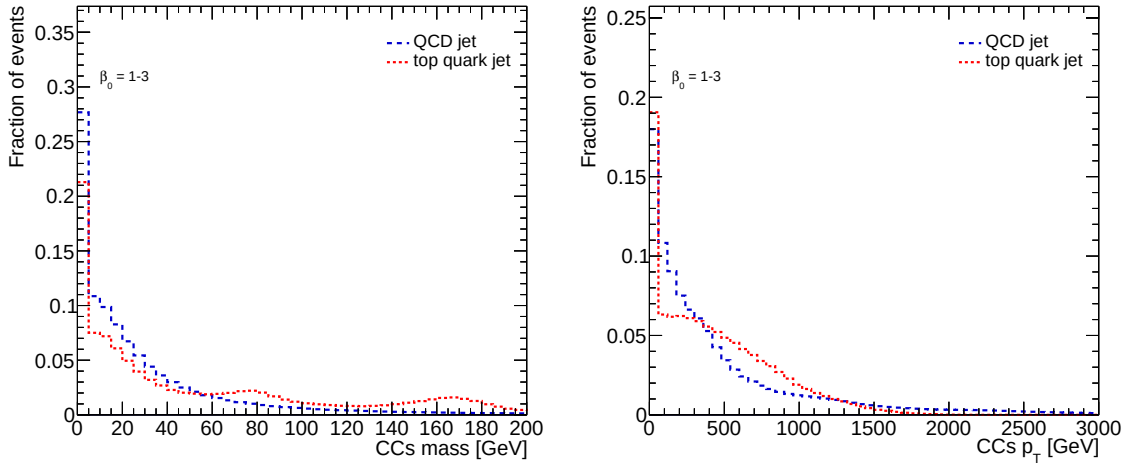


Figure 5.24: Comparison between signal top jets and background QCD jets of the metrics $\Delta R_{\text{avg}}(\text{CC}_1, t)$ and $\Delta R_{\text{avg}}(t_i, t_j \in \text{CC}_1)$. The distribution of $\Delta R_{\text{avg}}(\text{CC}_1, t)$ and $\Delta R_{\text{avg}}(t_i, t_j \in \text{CC}_1)$ are shown for jets with $\beta_0 = 2 - 3$ in (a) - (c) and for jets with $\beta_0 \geq 4$ in (b) - (d).



(a) (b)

Figure 5.25: The inclusive distributions of mass (a) and p_{T} (b) of connected components from jets with $\beta_0 = 1 - 3$.

between topoclusters and the closest CC. If the jet has $\beta_0 > n$, the CCs are reclustered using the Cambridge-Aachen algorithm until there are exactly n CCs in the jet. The Cambridge-Aachen algorithm is used in order to maintain consistency with the spatial clustering that is used by the Mapper algorithm when creating vertices. Furthermore, by the same reasoning, splitting scales that are analogous to the k_T splitting scales are defined as the minimum distance between two CCs before they are merged using the Cambridge-Aachen algorithm. Two variations of the n-point energy correlation functions and their ratios are defined. The first set calculates the energy correlation of the jet by using the CCs as the jet constituents. The second set calculates the energy correlation of a CC by using the topoclusters associated with the CC as its constituents. Figure 5.26 shows example distributions of these variables that are inspired by the jet substructure observables, with additional plots presented in Appendix C. As can be observed from the plots, the interplay between the topological structure of jets and the kinematics of the CCs contains discriminatory power between signal

and background jets. In the next section, the information obtained from the topology of jets and the kinematics of the CCs will be used to train two taggers that are designed to tag signal top jets against background QCD jets.

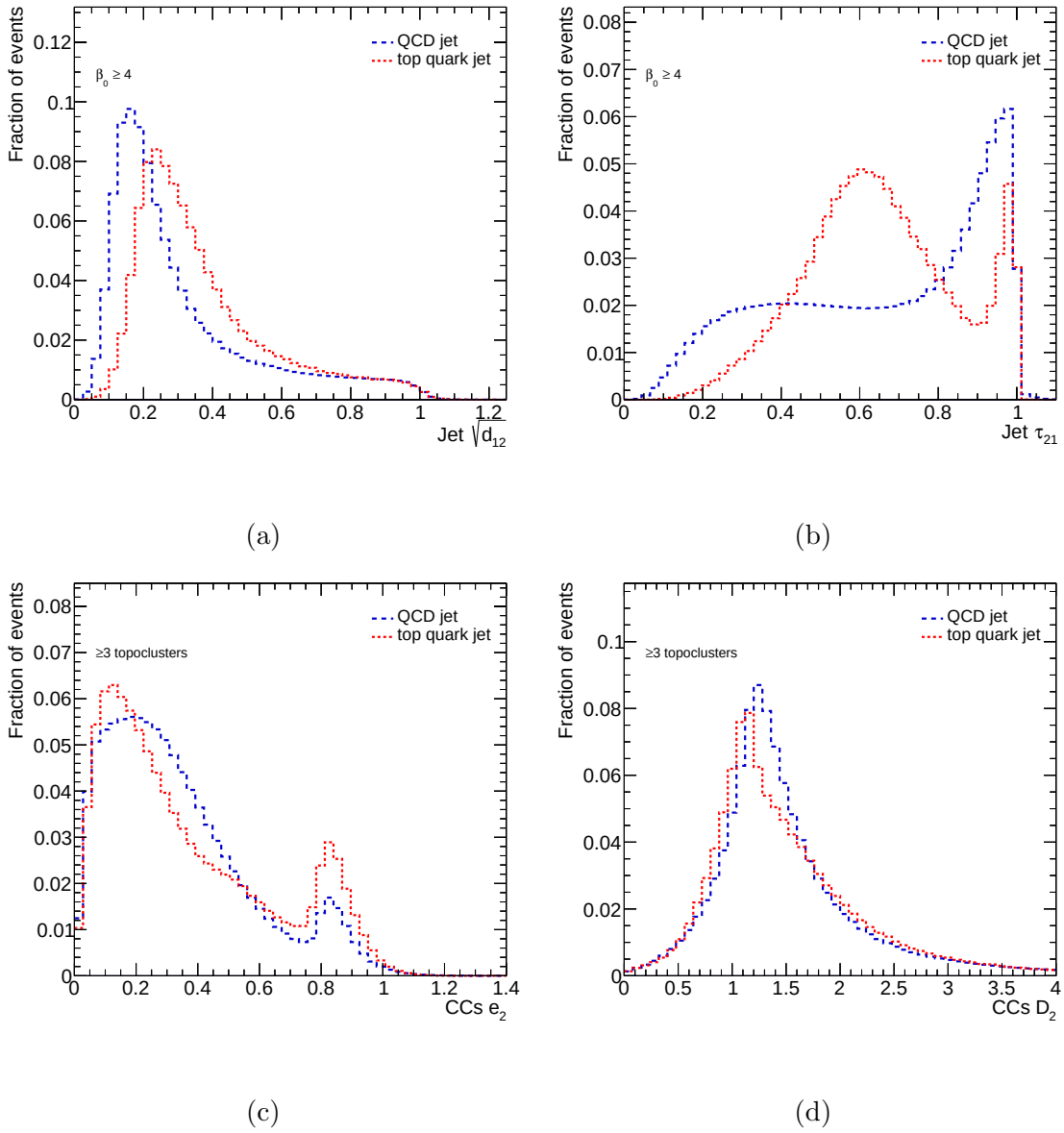


Figure 5.26: Comparisons of jet substructure variable distributions that are evaluated using the CCs of the jet and the topoclusters associated with a given CC between signal and background jets. The Cambridge-Aachen splitting scale of a jet when the last two remaining CCs are merged into a single CC is shown in (a) for jets with $\beta_0 \geq 4$. The n-subjettiness ratio τ_{21} is shown in (b) for jets with $\beta_0 \geq 4$. The observables τ_n are obtained by reclustering the CCs in the jet until there are n CCs using the Cambridge-Aachen algorithm and evaluating the minimum distance between the topoclusters in the jet and the reclustered CCs. The inclusive distributions of the e_2 energy correlation function and the ratio D_2 for CCs that contain at least three topoclusters are shown in (c) and in (d), respectively. These observables are evaluated using the topoclusters that are associated with the CC as the input constituents.

5.2.4 Machine Learning Studies

In this section, the design and optimization of two tagging algorithms for tagging jets from signal $W' \to tb$ and background QCD multijet production processes as either signal top or QCD background jets are presented. Both tagging algorithms are designed to use the topological and kinematic information of jets, vertices, and CCs that was obtained from the Mapper algorithm. The first tagging algorithm is a deep neural network (DNN) tagger that uses variables introduced in the previous section that are inspired by the jet substructure observables as input. The design of the DNN tagger is motivated in order to determine if there is residual information in the jet substructure observables obtained from the TDA that is not utilized by the contained and inclusive top taggers discussed in section 5.1. The second tagging algorithm is a convolutional graph neural network (GNN) that uses graph representations of jets as input. As will be detailed shortly, the graph representation of a jet is built from the Čech complex of the jet that is obtained from the Mapper algorithm. The design of the GNN is motivated in part by the ability of a graph to encode the topological information of jets in a single structure. Additionally, this allows the definition of a simpler tagging algorithm that does not utilize high-level information from the jet substructure variables. Both tagging algorithms were trained using signal top jets that satisfy the contained top jet labeling criteria, as discussed in subsection 5.1.3. The taggers are optimized to a 50% and 80% fixed signal efficiency working points. Finally, the performance of both taggers is compared to the performance of the contained top DNN tagger, which is referred to as the jet substructure (JSS) tagger in this section.

The optimization process of the DNN tagger started with the selection of the input variables from a baseline set of 74 variables. These variables consist of the CCs substructure

information of jets, the kinematic information of vertices, the kinematic information of CCs, and the topocluster substructure information of CCs. As discussed in the previous section, the vertices and CCs obtained from the Č complex are interpreted as subjets by adding the four-momenta of the topoclusters that are associated with these objects. The baseline set of variables was reduced by clustering variables into groups based on their correlation. The correlation distance metric between two variables is defined as:

$$d(x, y) = \sqrt{1 - \rho^2(x, y)} \tag{5.16}$$

where $\rho(x, y)$ is the correlation coefficient between two variables x and y. Variables that are highly correlated or anti-correlated are mapped close to zero with this metric. The distance between two clusters of variables A_i and A_j is determined using the complete-linkage distance:

$$D(\mathrm{A}_i, \mathrm{A}_j) = \max_{x \in \mathrm{A}_i,\, y \in \mathrm{A}_j} d(x, y) \tag{5.17}$$

Two clusters of variables are merged if they achieve the minimal complete-linkage distance between all possible pairs of clusters of variables. This corresponds to grouping together all variables that contain approximately the same amount of information. The clustering process was carried out up to a threshold distance of 0.92, which corresponds to a minimum absolute correlation within the variable cluster of 0.39. The end result of the clustering process yielded 26 clusters of variables, which are summarized in the dendogram shown in Figure 5.27. A single variable was retained from each individual cluster based on the separation power that the variable has between signal and background jets. This intermediate set of variables was reduced to 21 variables by removing those that did not contain sufficient discriminatory power. The final set of variables chosen as the inputs for the DNN tagger

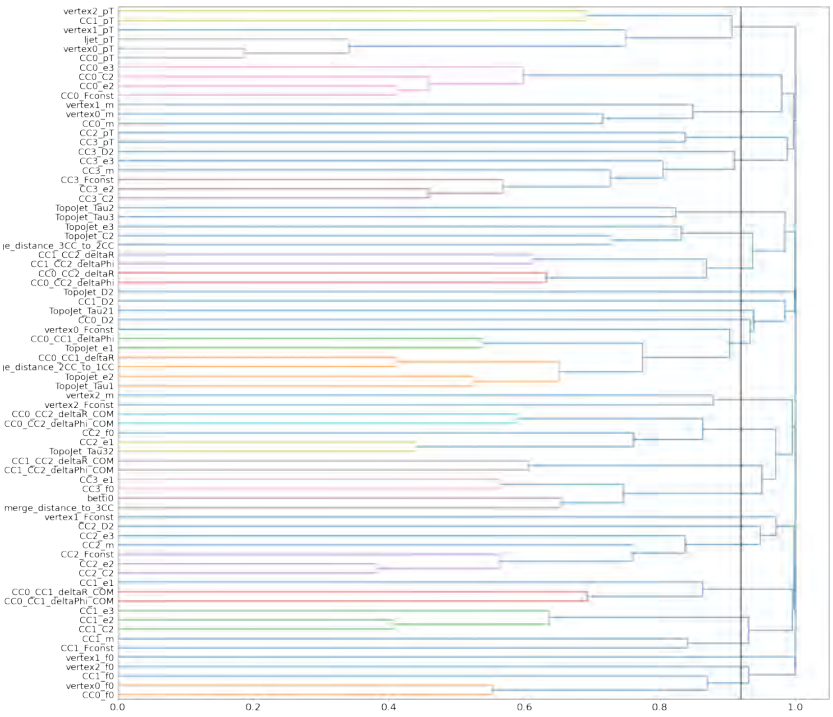


Figure 5.27: The clustering dendrogram of the initial baseline set of variables that shows how the variable clusters are formed based on the correlation metric in Equation 5.16. The black vertical line is the clustering distance threshold at which the intermediate set of variables was chosen.

is summarized in Table 5.8. The Keras [67] software package was used in the design of the DNN tagging algorithm. The architecture and optimized hyperparameters of the DNN tagger are summarized in Table 5.9. The tagger was trained using 200000 contained top jets from the $W' \to tb$ process as signal jets and 4167611 jets from the QCD multijet production process as background jets. Both signal and background jets were split evenly into training and validation datasets.

The optimization process of the GNN tagger started with the design of the graph representation of jets. Each individual jet is assigned a graph whose vertices correspond to

Variable types	Variables
Fraction of contained topoclusters	v_1, CC_1, CC_3
Mass	v_2, CC_0, CC_2
p_T	v_2, CC_0, CC_2
e_2	CC_0, CC_1
D_2	CC_0, CC_1, CC_2
$\Delta R_{CoM}(CC_i, CC_j)$	$(i,j) = (0,1), (0,2), (1,2)$
$\Delta R(CC_i, CC_j)$	$(i,j) = (0,2)$
Cambridge-Aachen splitting scales	$\sqrt{d_{12}}$, $\sqrt{d_{23}}$, $\sqrt{d_{34}}$

Table 5.8: List of variables used as input to the DNN grouped by variable type. The variables that are defined on vertices and connected components obtained from the Mapper algorithm are denoted by v_i and CC_i respectively, where the index i is used to denote the ordering of the objects based on their p_T. The fraction of contained topoclusters is the ratio of the number of topoclusters associated with the object to the total number of topoclusters in the jet. The mass and p_T are obtained by adding the four-momentum vectors of the topoclusters associated with the object. The energy correlation functions are calculated by using the associated topoclusters of the connected components as the constituents.

Hyperparameter	Option used
Layer	Dense
Number of hidden layers	3
Number of nodes per hidden layer	20, 15, 10
Activation function	Scaled exponential linear unit (selu) [68]
L1 regularizer	None
L2 regularizer	None
Weight initializer	lecun normal
Optimizer	Adam with Nesterov momentum (Nadam) [69]
Learning rate	0.00001
Batch size	500
Batch normalization	Yes
Number of epochs	1000
Loss function	Binary crossentropy

Table 5.9: List of hyperparameters optimized for the DNN tagger. The DNN consists of 3 hidden layers with the number of nodes decreasing in each subsequent layer.

the CCs of the Čech complex of the jet. Each vertex of the graph is assigned a set of input features that consist of the CC four-momentum, mass, and p_T, which are evaluated in the CoM frame of the jet. The graph is made fully connected by including edges $e_{i,j}$ between

all possible pairs of CCs. In order to encode the degree of disconnectedness between two CCs, each edge is assigned a set of input features that consist of the angular distances between CCs evaluated in the CoM frame of the jet: $\Delta R_{\mathrm{CoM}}(\mathrm{CC}_i, \mathrm{CC}_j)$, $\Delta\phi_{\mathrm{CoM}}(\mathrm{CC}_i, \mathrm{CC}_j)$ and $\Delta\eta_{\mathrm{CoM}}(\mathrm{CC}_i, \mathrm{CC}_j)$. This graph structure with the corresponding set of input features is used as the input to the GNN tagger in order to classify jets. The Spektral [70] software package was used in the design of the GNN tagging algorithm. The architecture and optimized hyperparameters of the GNN tagger are summarized in Table 5.10. The tagger was trained using 200000 contained top jets from the $W' \to tb$ process as signal jets and 500000 jets from the QCD multijet production process as background jets, which were split evenly into training and validation datasets. The number of background jets used for the training of the GNN had to be reduced compared to the DNN training due to limits on the available memory resources. This is because the GNN requires all graphs from the dataset to be available during the training process, and the amount of memory that each graph takes is large, which can exceed the available resources if a large number of jets are included in the training.

The performance during the training process of the DNN and GNN tagging algorithms are summarized in Figure 5.28. The accuracy, which quantifies the frequency of a given tagger correctly classifying jets as either signal or background jets, and the loss function of both models were evaluated as a function of the training epoch, both with the training and validation datasets. Both the DNN and GNN taggers show no sign of overtraining since the performance between the training and validation datasets agrees well. However, the GNN tagger shows signs of undertraining since the accuracy of the validation dataset exceeds that of the training dataset at later epochs. The tagger score distributions shown in Figure 5.29 indicate that the GNN tagger is not robust enough when classifying signal top

Hyperparameter	Option used
Layer 1	Graph Convolution with skip connection (GCS) [71]
Number of output channels	6
Activation function	Exponential linear unit (elu) [72]
Weight initializer	Glorot uniform [73]
L1 regularizer	None
L2 regularizer	None
Layer 2	Edge-conditioned convolutional layer (ECC) [74]
Number of output channels	6
Activation function	None
Weight initializer	Glorot uniform
MLP number of hidden layers	2
MLP number of nodes per hidden layer	9, 7
L1 regularizer	None
L2 regularizer	0.0001
Layer 3	Global sum pool
Layers $4-6$	Dense
Number of hidden layers	3
Number of nodes per hidden layer	10, 9, 8
Activation function	elu
L1 regularizer	None
L2 regularizer	0.0001
Weight initializer	Glorot uniform
Optimizer	Nadam
Learning rate	0.001
Batch size	350
Batch normalization	Yes
Number of epochs	200
Loss function	Binary crossentropy

Table 5.10: List of hyperparameters optimized for the GNN tagger. The input graphs are first processed through the GCS layer. The output of this convolution layer is used as an input to the ECC layer. The output of the ECC layer is pooled by summing the individual output channel features per node. The pooled features are then used as the input into a DNN with three hidden layers, which performs the jet classification.

jets, as the GNN tagger score peaks at lower values when compared to the DNN tagger score for signal top jets. However, both models show good separation power between signal and background jets. Both 50% and 80% fixed signal efficiency working points were defined for

both taggers using the training dataset. The performance of each tagger at these working points is compared to the corresponding working point of the JSS tagger. The number of signal and background jets that pass or fail a given working point is summarized in Figure 5.30. As can be observed in the plots, both the DNN and GNN taggers are slightly outperformed by the JSS tagger. However, a similar level of background jet rejection is obtained for all taggers considered.

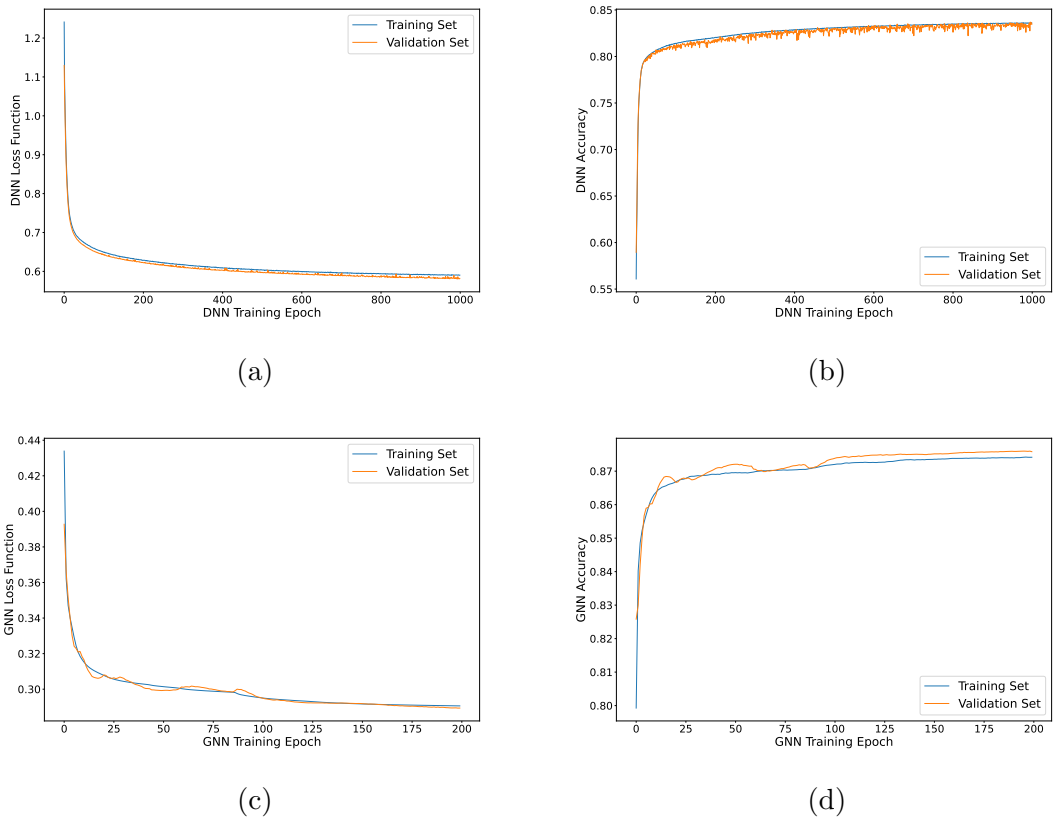
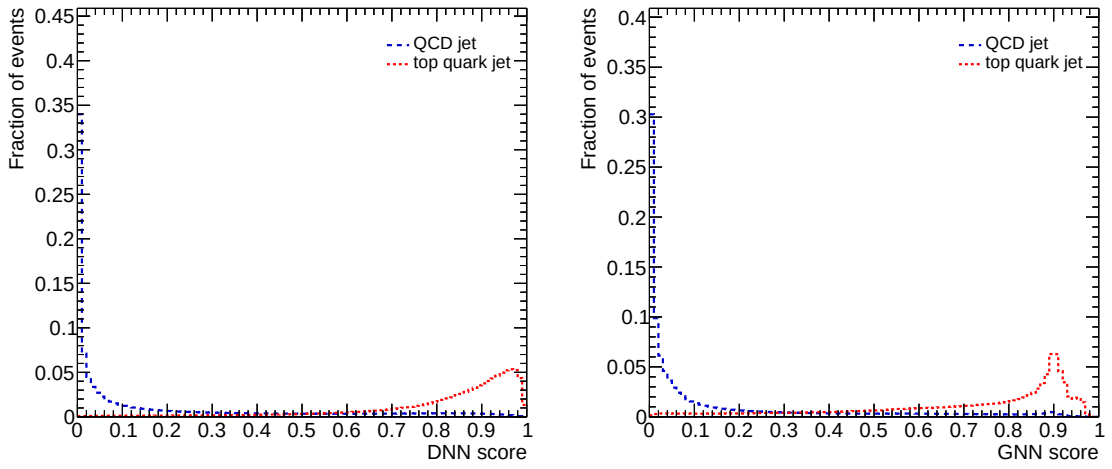


Figure 5.28: The DNN loss function (a) and accuracy (b), and the GNN loss function (c) and accuracy (d). Both metrics are shown for the training and validation datasets as a function of the training epochs of the networks.

In order to determine if there is residual information from the topology of jets that is not being used by the taggers, the distributions of variables obtained from the TDA are compared between signal and background jets in different tagging selection regions. The

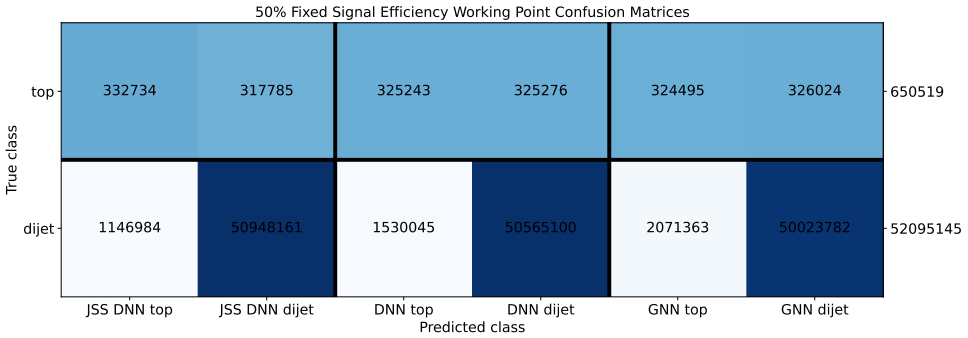


(a) (b)

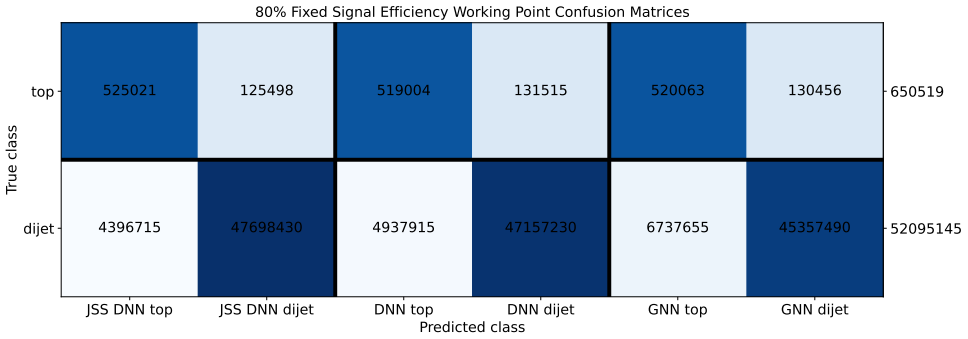
Figure 5.29: The score distributions of the DNN (a) and the GNN (b) models overlayed between signal top jets and background QCD jets.

distributions are compared for jets that pass either the 80% working point tagging criteria of the DNN or GNN taggers independently while simultaneously failing the 50% working point tagging criteria of the JSS tagger. Conversely, the distributions are also compared for jets that fail either the 50% working point of the DNN or GNN taggers independently while simultaneously passing the 80% JSS tagger. The jets that satisfy these tagging selections populate a phase space where the classification of jets by the taggers is ambiguous. For example, signal top jets in these regions contain features that are deemed background-like by the tight signal requirements of a tagger while being loosely considered as signal by another tagger. Thus, if the distributions of variables show differences between signal and background jets in these regions, then this implies that there is residual information from the topology of jets that is not fully used by the taggers and can improve the jet classification.

First, the number of connected components in jets were compared in order to determine if the topology of the jets contained residual information in the tagging phase spaces considered.



(a)



(b)

Figure 5.30: The 50% (a) and 80% (b) fixed signal efficiency working point confusion matrices for the jet substructure contained top DNN tagger (JSS), the DNN trained using the information from the Mapper algorithm, and the GNN using the graph representation of the jets. The rows of the matrices correspond to the number of events for the signal top jet and background QCD multijet classes while the columns correspond to the number of predicted events in each class for a given tagger. The total number of events for each class is shown on the right vertical axis.

As can be observed in Figure 5.31, the number of connected components between signal top jets and background QCD jets does not show any differences. Thus, at a surface level, the information from the homology of jets has been fully used by the taggers. As a next step, the kinematic and substructure observables of the objects that define the homology of jets were analyzed in order to determine if there is residual information in the interplay between the topology and kinematics of jets.

The kinematic distributions of the vertices of the $\check{\mathrm{C}}$ complex were compared in order to determine if the mapping onto the filter function feature space and spatial clustering of the topoclusters contains residual information. Figures 5.32 and 5.33 show the p_{T} distributions of the second leading and third leading vertices in the $\check{\mathrm{C}}$ complex of the jets, respectively. As can be observed from these distributions, signal top jets that fail the 50% working points of the DNN or GNN and pass the 80% working point of the JSS tagger have a narrower p_{T} distributions for their vertices when compared to background jets. Thus, the JSS tagger is able to identify signal top jets as having vertices with well-defined p_{T} values. On the other hand, jets that pass either the 80% working point of the DNN or GNN and fail the 50% working point of the JSS tagger do not show significant differences in the kinematics of vertices. This indicates that the DNN and GNN taggers have used most of the information from vertices.

The Cambridge-Aachen splitting scales were compared in order to assess if there is residual information from how vertices form CCs and how the CCs are distributed within the jet. Figures 5.34 and 5.35 show the Cambridge-Aachen splitting scales from the three-to-two and two-to-one CC mergings, respectively. As can be observed in these plots, signal top jets tend to have larger merging scales compared to QCD jets in all tagging selection regions considered except for jets that pass the 80% working point of the DNN tagger and fail the

50% working point of the JSS tagger. Thus, all taggers except the DNN tagger have not used to their full extent the information that signal top jets tend to have CCs that are more spread out within the jet when compared to background jets.

The n-subjettiness and n-point energy correlation observables in jets were compared in order to determine if there is additional discriminatory information from the substructure and radiation pattern of the CCs in the jets. The distributions of the 2-point energy correlation function e_2 and the n-subjettiness ratios τ_{21} and τ_{32} are shown in Figures 5.36 - 5.38. The 2-point energy correlation function e_2 contains some residual information in all phase spaces considered except for jets that pass the 80% DNN and fail the 50% JSS taggers. Thus, all taggers except the DNN have not used all the information available from how the energy of the jet is distributed across its CCs. The τ_{21} distribution is bimodal for both signal and background jets while the τ_{32} distribution peaks sharply at 1 for both signal and background jets. These observations indicate that the jets that populate these phase spaces are better modeled with either two CCs or a single CC, with the degree of the preferred substructure varying across the different tagging criteria.

Finally, the kinematic distributions of CCs and the substructure observables evaluated using the topoclusters associated with a given CC are compared. This is done in order to determine if there is residual information from the energy and radiation patterns that the CCs reconstruct from the topoclusters. The distributions of the energy correlation function ratio D_2 of the second leading CC and the mass of the first leading CC are shown in Figures 5.39 and 5.40, respectively. The CC_1 D_2 distribution in background QCD jets is narrower compared to signal top jets, which indicates that there is some residual information in the energy distribution of topoclusters within the CCs that is not fully used by the taggers. The mass bumps near the W boson mass and top quark mass that are observed in the CC_0

mass distribution for signal top jets indicate that CC_0 has partially reconstructed some of the substructure of the jet in these phase spaces. Additionally, the CC_0 mass distribution for background QCD jets that pass the 80% GNN tagger and fail the 50% JSS tagger shows a peak below the top quark mass with a long tail that extends to higher values, which is characteristic when trying to reconstruct the substructure of top jets from inconsistent radiation patterns. This could be indicative that the convolutional layers of the GNN have learned to reconstruct the top mass from the graph structure of jets but have not fully used this information for jet classification.

To summarize these observations, the variables obtained from the TDA of jets contain residual information that is not used to its full extent by the taggers studied. This information could be used to improve the separation between signal top jets and background QCD jets in phase spaces where their classification by the taggers is ambiguous. As discussed, both signal and background jets that populate these phase spaces are characterized by a topology that best models the jets with a single CC or two CCs. In the case of signal top jets, the CCs are more spatially spread out compared to the CCs in background QCD jets. Additionally, the leading CC in signal top jets has partially reconstructed some of the relevant substructure of the jet, while in background QCD jets the reconstruction is more consistent with reconstructing substructures from random patterns of topoclusters.

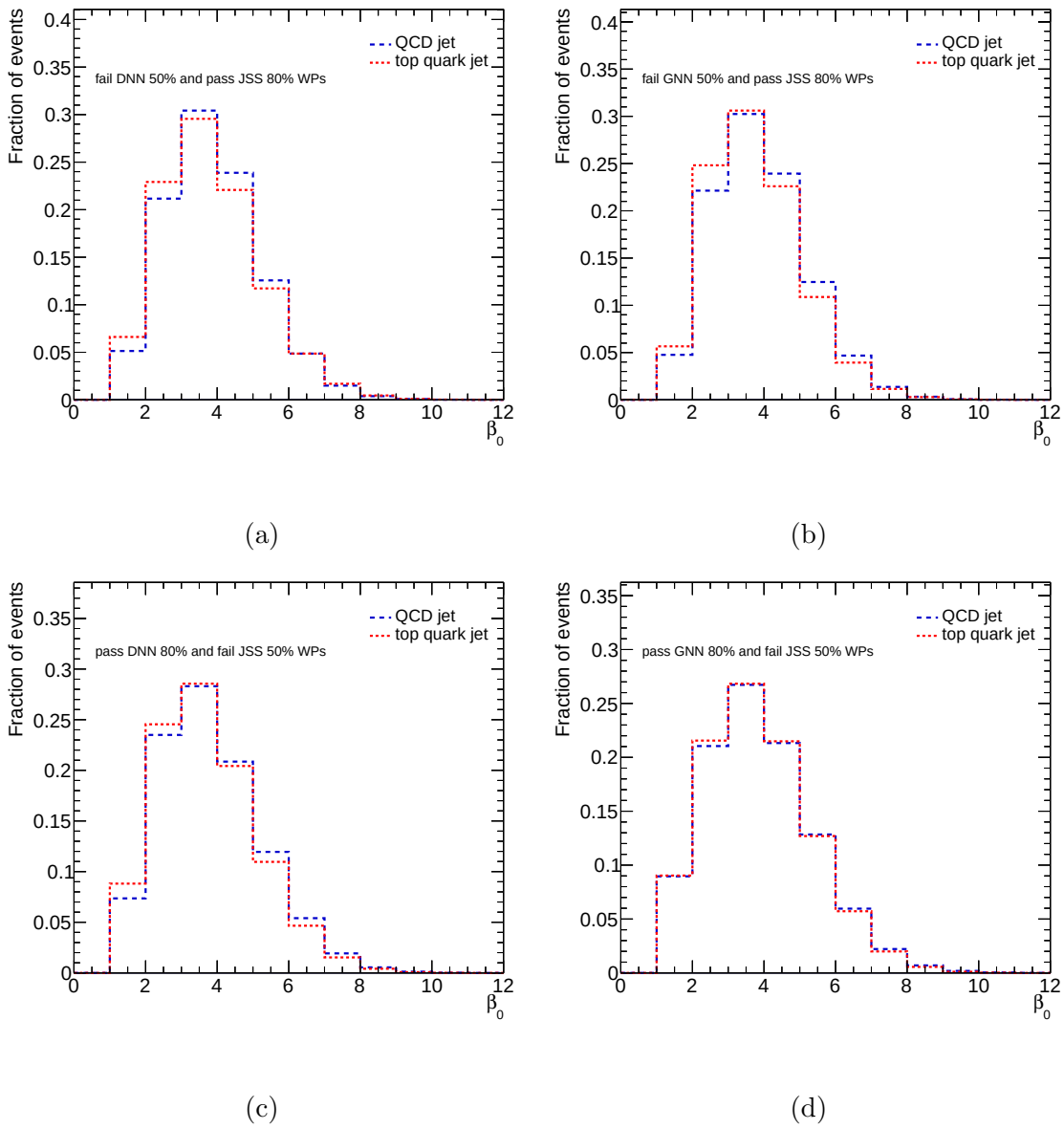


Figure 5.31: The distribution of the number of connected components in jets that pass the 50% working point of the JSS tagger and fail the 80% working point of the DNN tagger (a) and the 80% of the GNN tagger (b) respectively. The same distributions are shown for jets that fail the 50% working point of the JSS tagger and pass the 80% working point of the DNN tagger (c) and the 80% working point of the GNN tagger (d) respectively.

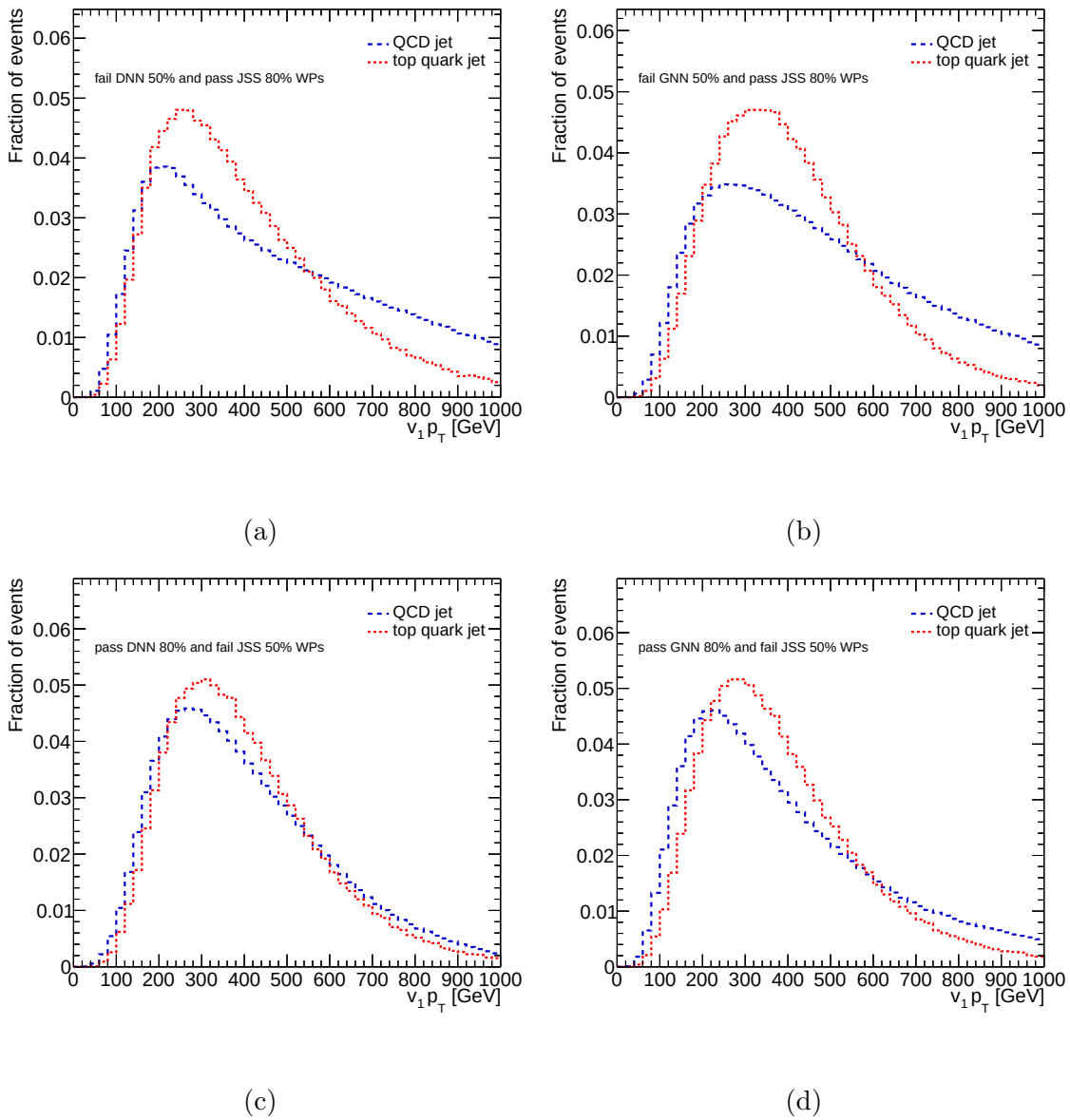


Figure 5.32: The p_{T} distribution of the second leading vertex of the Č complex of jets that pass the 50% working point of the JSS tagger and fail the 80% working point of the DNN tagger (a) and the 80% working point of the GNN tagger (b) respectively. The same distributions are shown for jets that fail the 50% working point of the JSS tagger and pass the 80% working point of the DNN tagger (c) and the 80% working point of the GNN tagger (d) respectively.

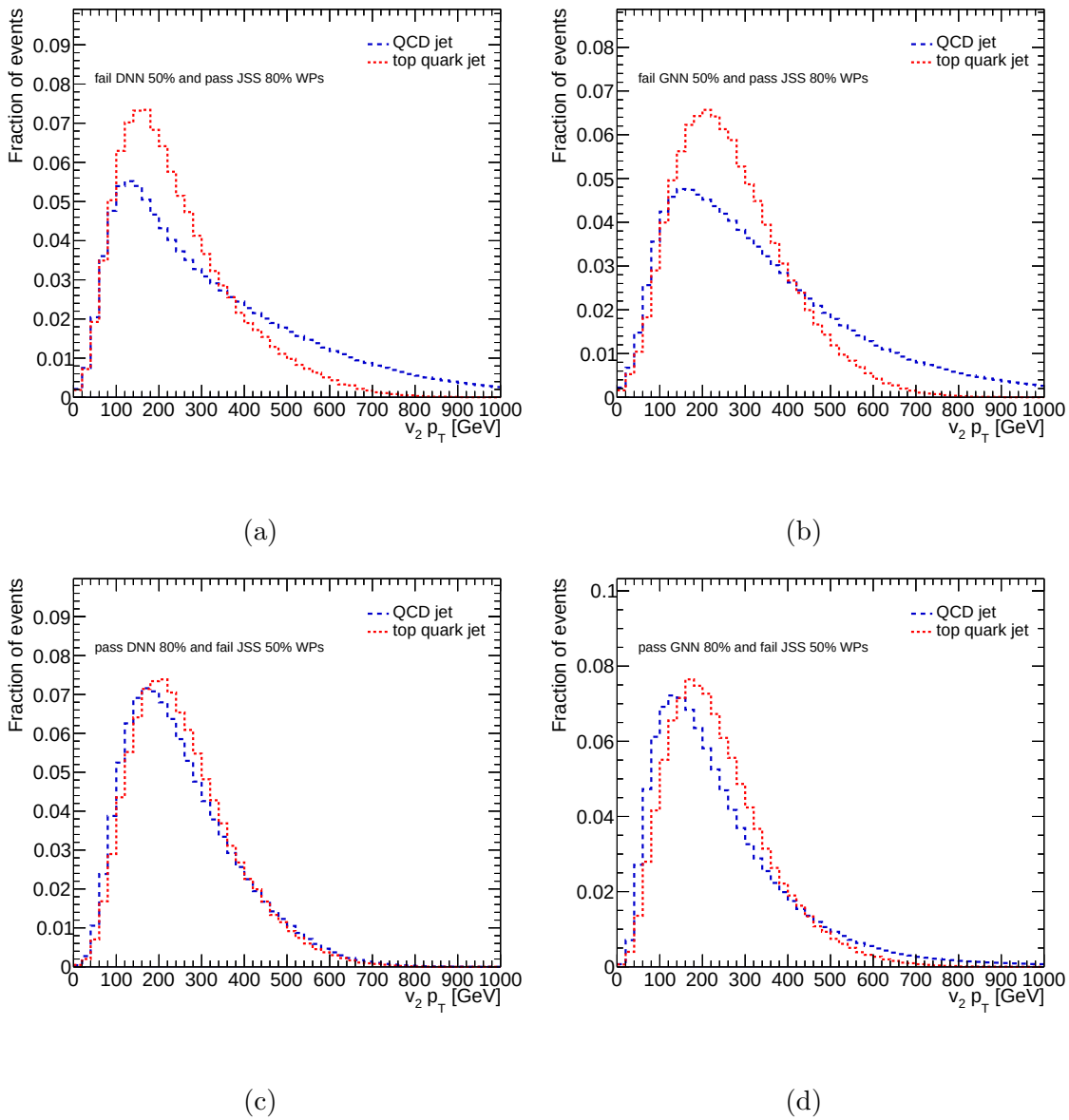


Figure 5.33: The p_{T} distribution of the third leading vertex of the Č complex of jets that pass the 50% working point of the JSS tagger and fail the 80% working point of the DNN tagger (a) and the 80% working point of the GNN tagger (b) respectively. The same distributions are shown for jets that fail the 50% working point of the JSS tagger and pass the 80% working point of the DNN tagger (c) and the 80% working point of the GNN tagger (d) respectively.

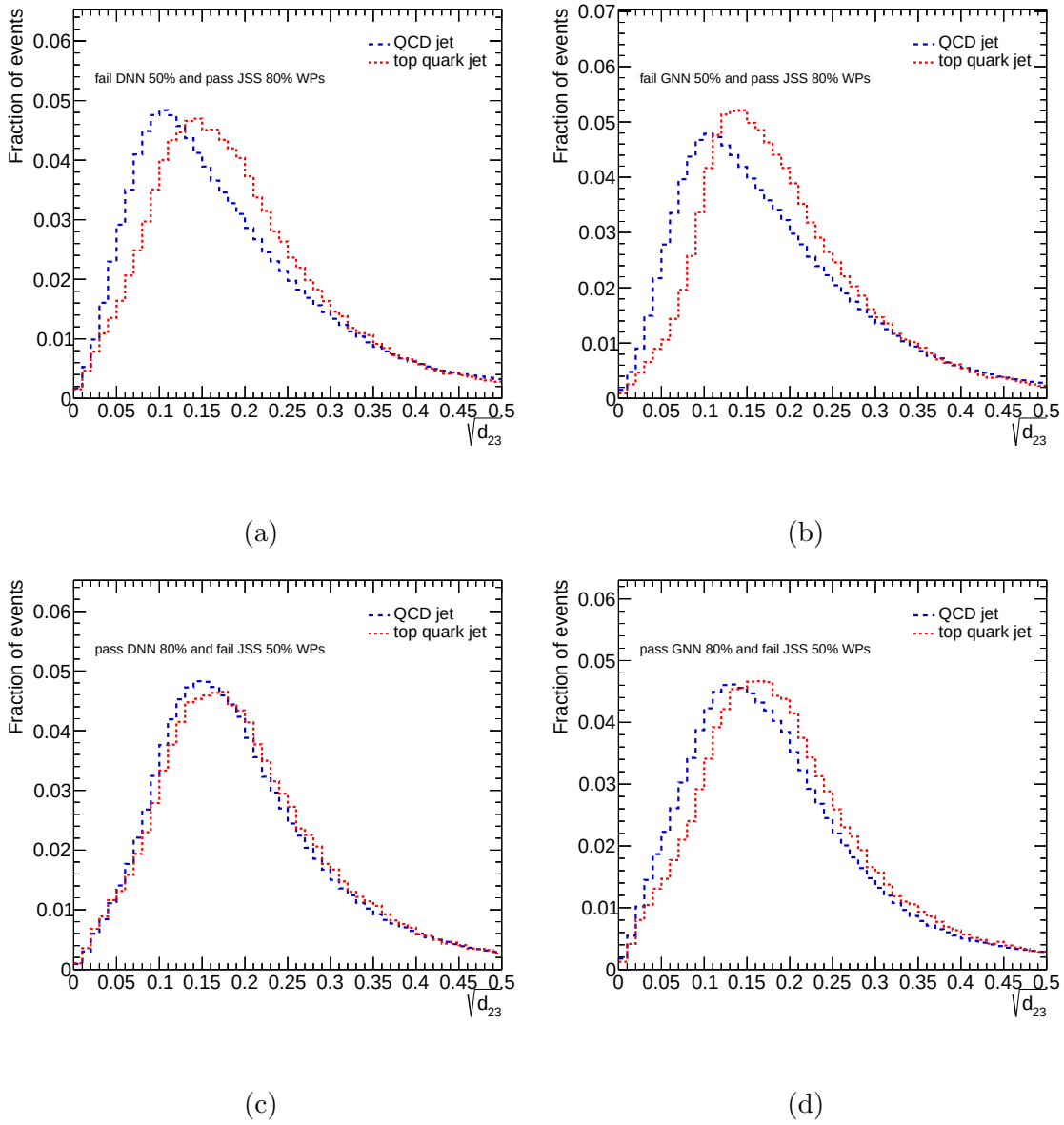


Figure 5.34: The three-to-two connected component Aachen splitting scale of jets that pass the 50% working point of the JSS tagger and fail the 80% working point of the DNN tagger (a) and the 80% working point of the GNN tagger (b) respectively. The same distributions are shown for jets that fail the 50% working point of the JSS tagger and pass the 80% working point of the DNN tagger (c) and the 80% working point of the GNN tagger (d) respectively.

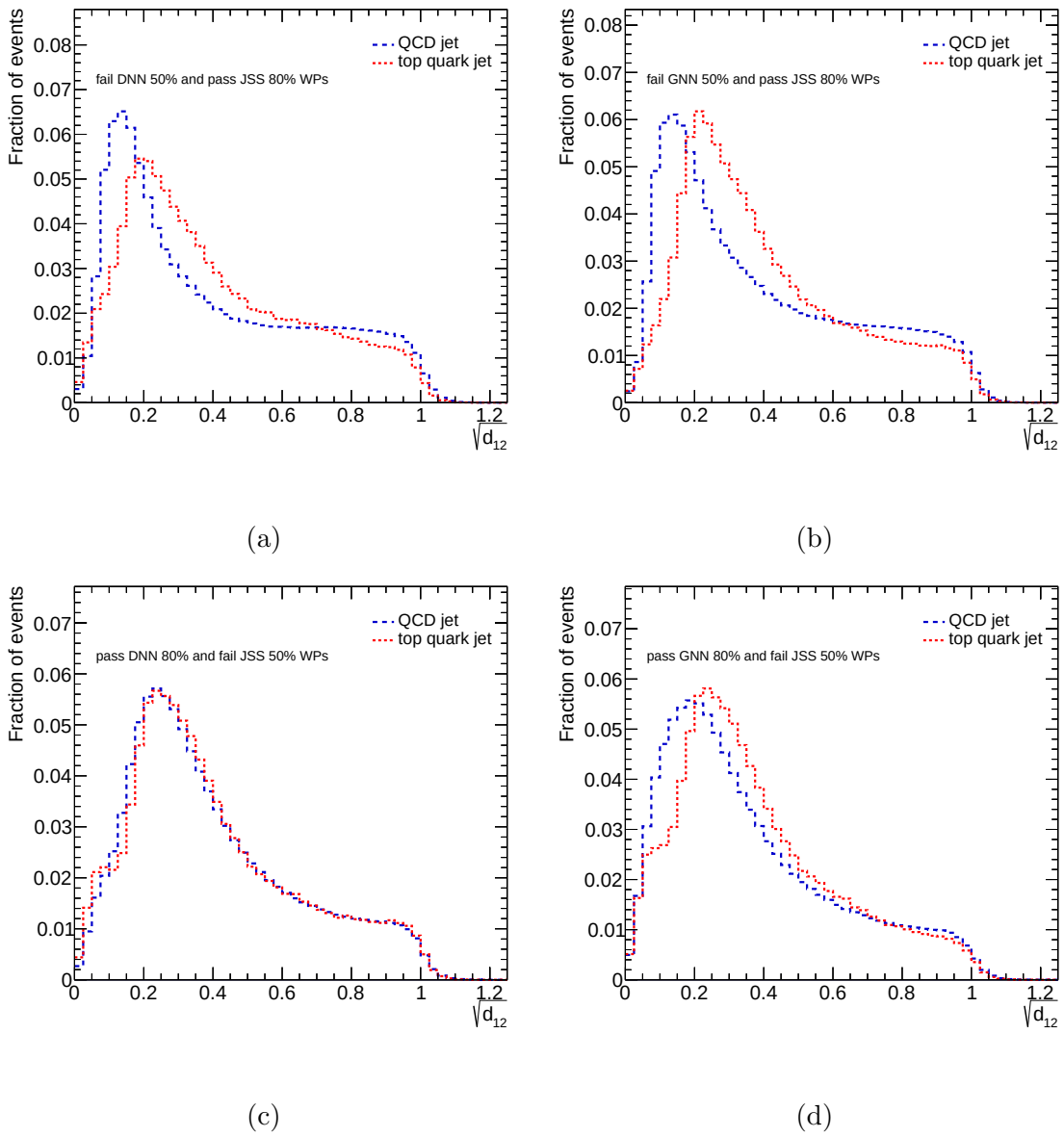


Figure 5.35: The two-to-one connected component Aachen splitting scale of jets that pass the 50% working point of the JSS tagger and fail the 80% working point of the DNN tagger (a) and the 80% working point of the GNN tagger (b) respectively. The same distributions are shown for jets that fail the 50% working point of the JSS tagger and pass the 80% working point of the DNN tagger (c) and the 80% working point of the GNN tagger (d) respectively.

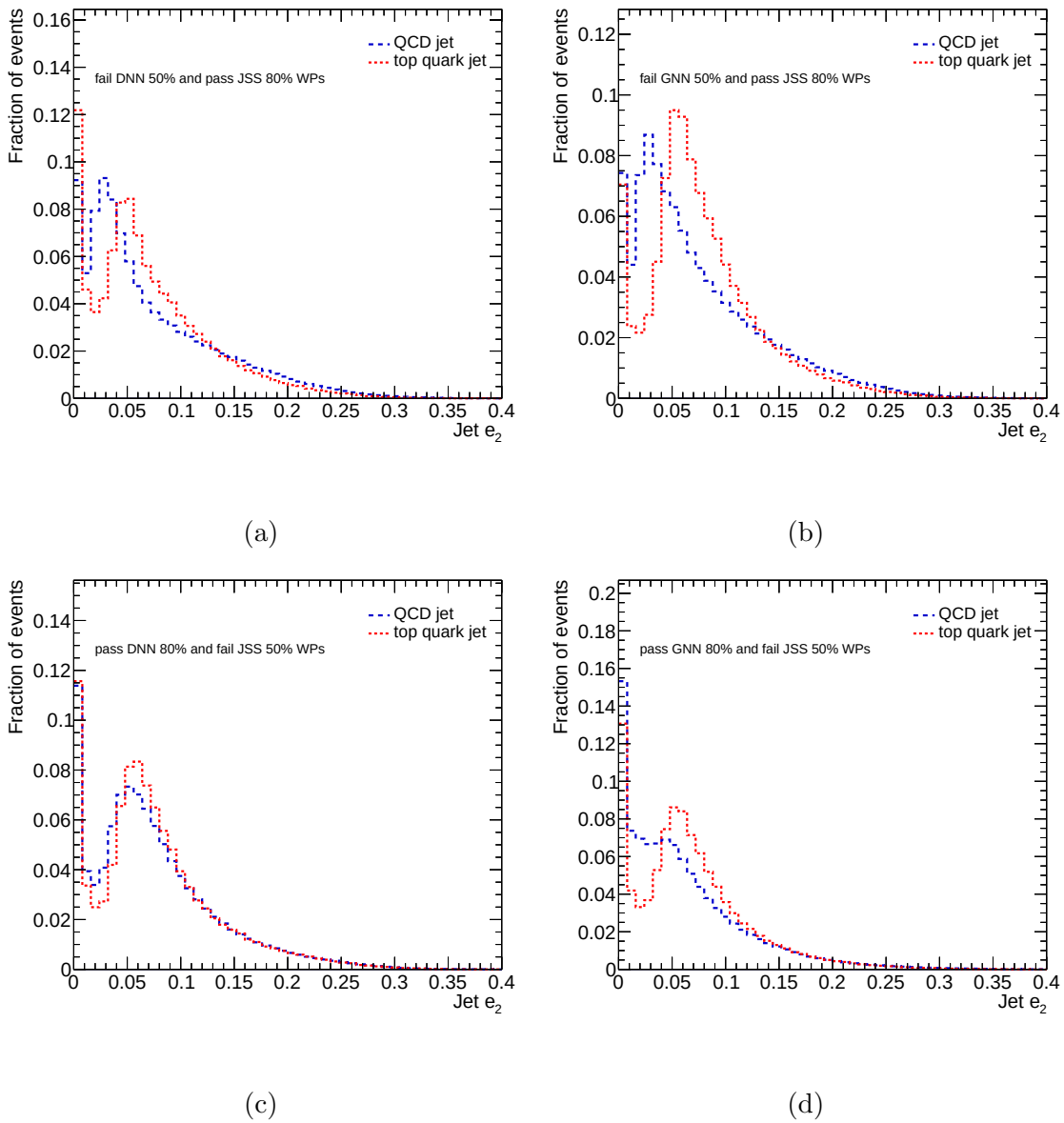


Figure 5.36: The energy correlation function e_2 evaluated with the CCs of jets that pass the 50% working point of the JSS tagger and fail the 80% working point of the DNN tagger (a) and the 80% working point of the GNN tagger (b) respectively. The same distributions are shown for jets that fail the 50% working point of the JSS tagger and pass the 80% working point of the DNN tagger (c) and the 80% working point of the GNN tagger (d) respectively.

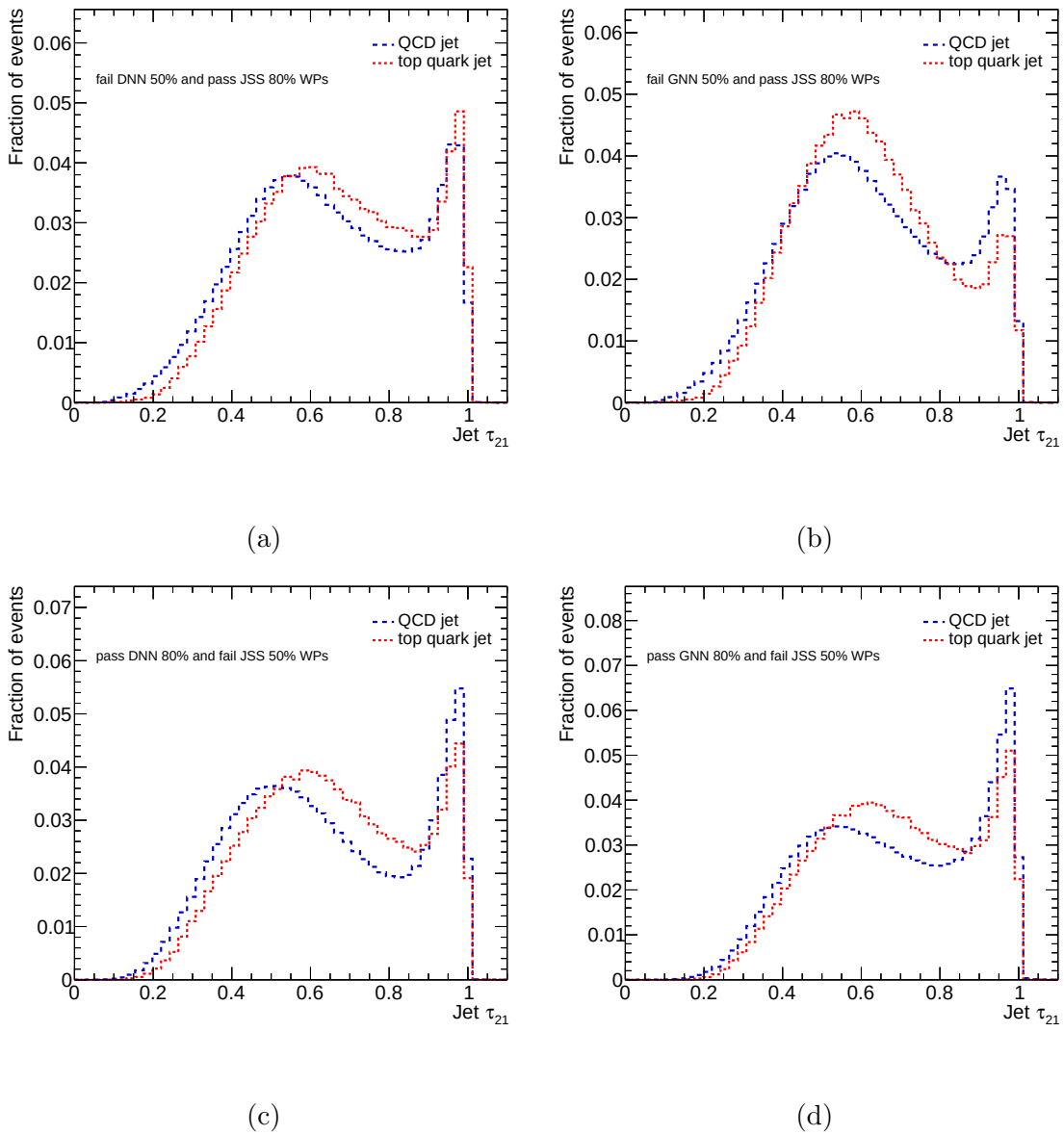


Figure 5.37: The τ_{21} ratio evaluated with the CCs of jets that pass the 50% working point of the JSS tagger and fail the 80% working point of the DNN tagger (a) and the 80% working point of the GNN tagger (b) respectively. The same distributions are shown for jets that fail the 50% working point of the JSS tagger and pass the 80% working point of the DNN tagger (c) and the 80% working point of the GNN tagger (d) respectively.

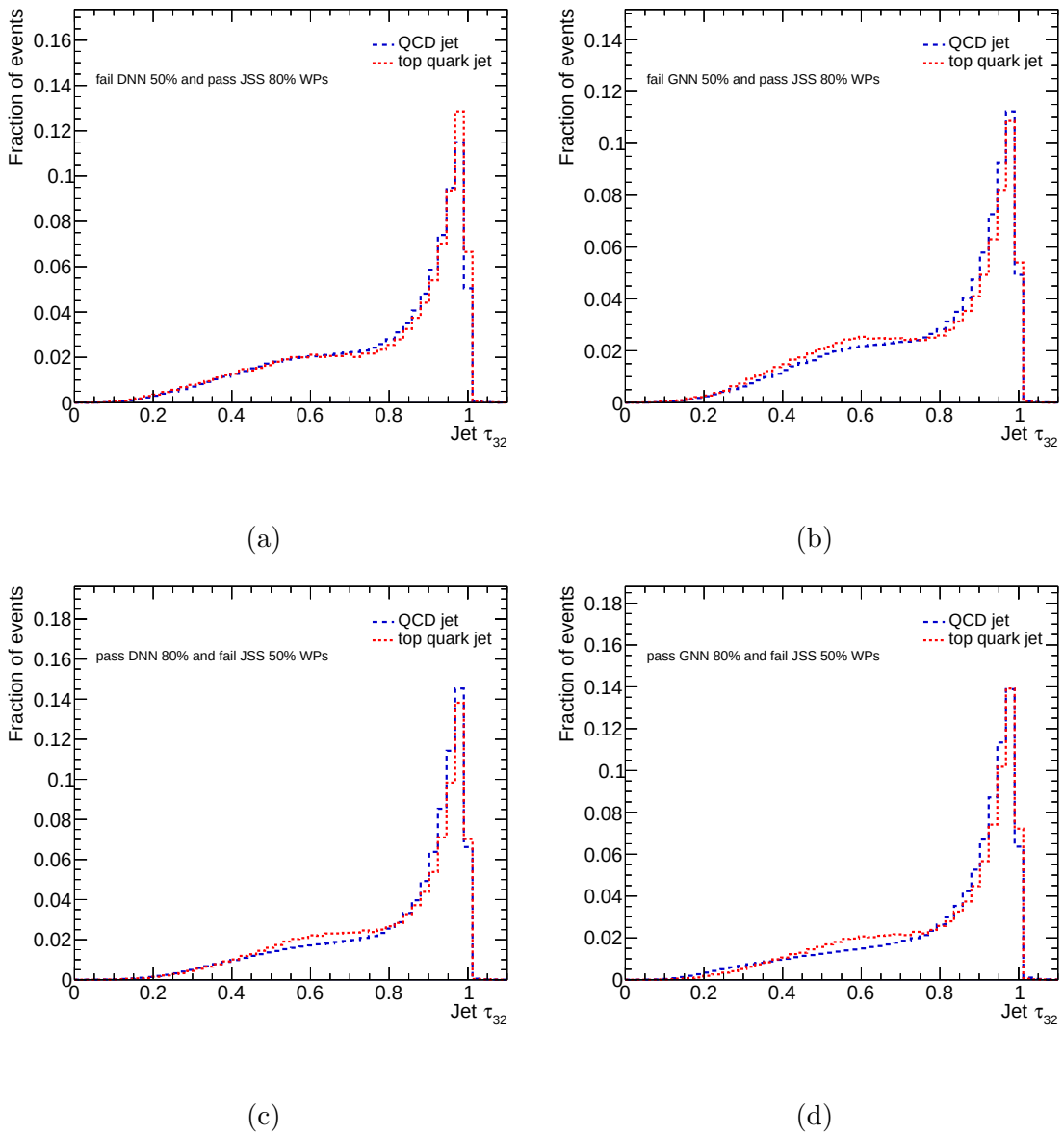


Figure 5.38: The τ_{32} ratio evaluated with the CCs of jets that pass the 50% working point of the JSS tagger and fail the 80% working point of the DNN tagger (a) and the 80% working point of the GNN tagger (b) respectively. The same distributions are shown for jets that fail the 50% working point of the JSS tagger and pass the 80% working point of the DNN tagger (c) and the 80% working point of the GNN tagger (d) respectively.

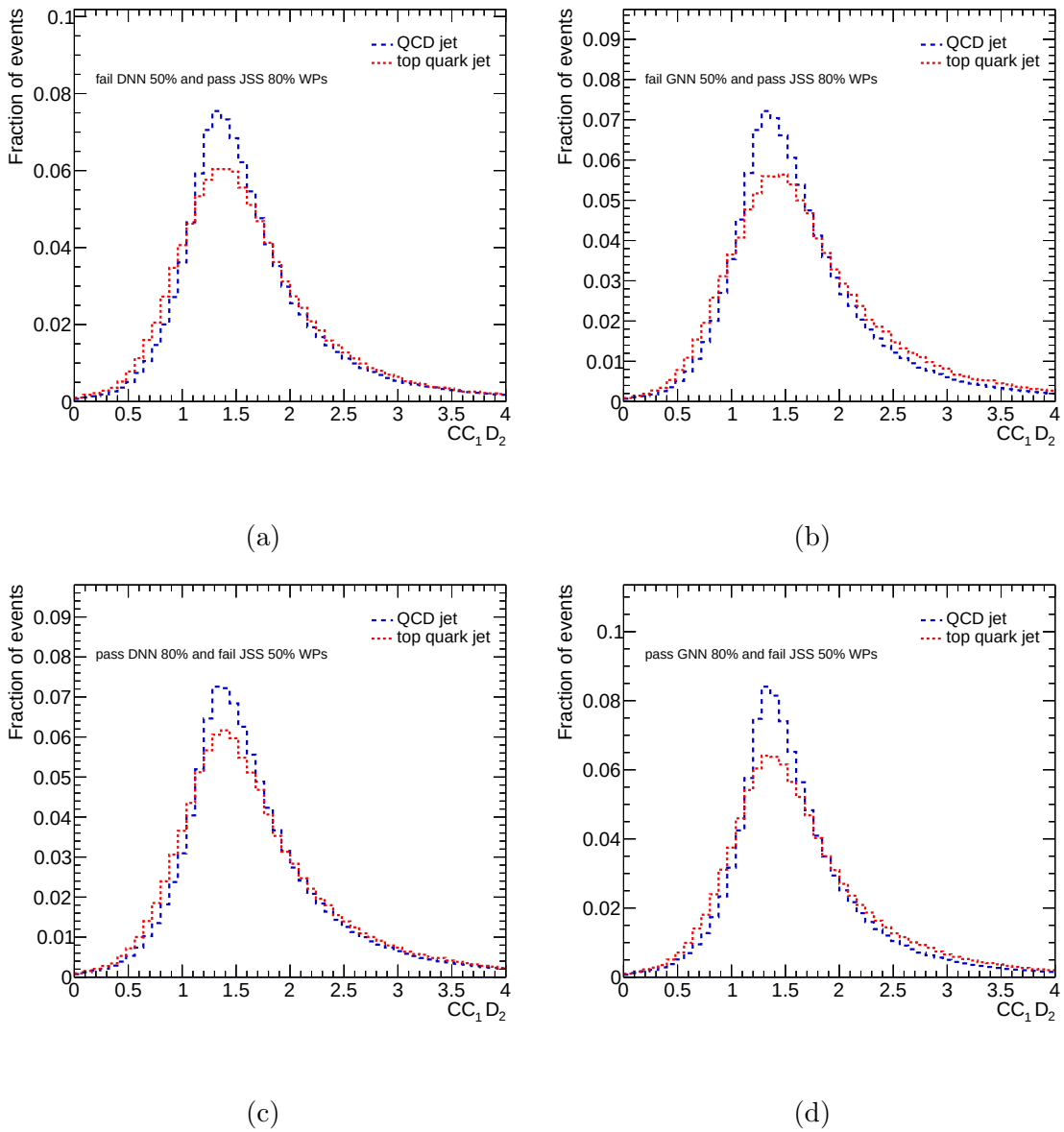


Figure 5.39: The D_2 of the second leading CC of jets that pass the 50% working point of the JSS tagger and fail the 80% working point of the DNN tagger (a) and the 80% working point of the GNN tagger (b) respectively. The same distributions are shown for jets that fail the 50% working point of the JSS tagger and pass the 80% working point of the DNN tagger (c) and the 80% working point of the GNN tagger (d) respectively.

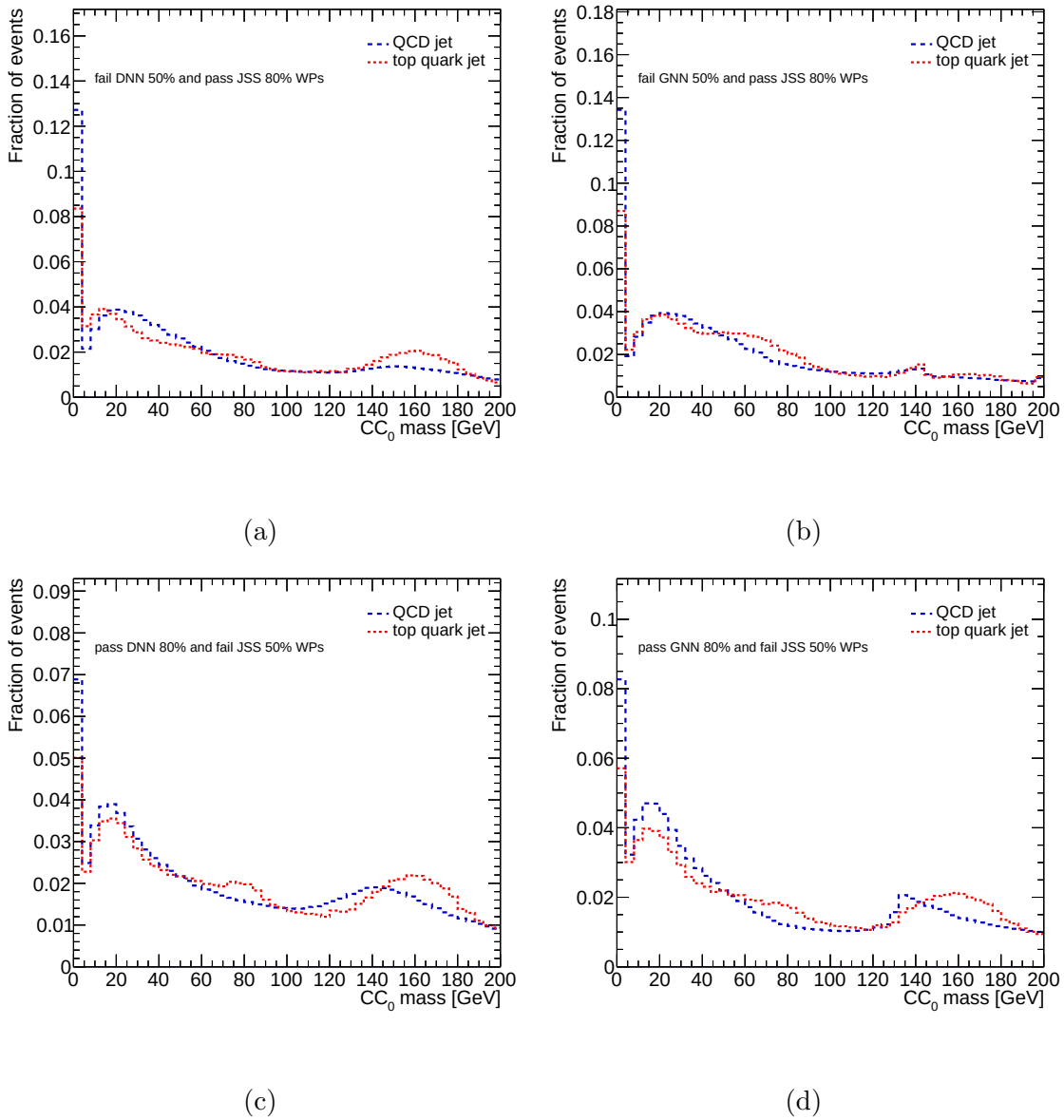


Figure 5.40: The mass of the leading CC of jets that pass the 50% working point of the JSS tagger and fail the 80% working point of the DNN tagger (a) and the 80% working point of the GNN tagger (b) respectively. The same distributions are shown for jets that fail the 50% working point of the JSS tagger and pass the 80% working point of the DNN tagger (c) and the 80% working point of the GNN tagger (d) respectively.

Chapter 6

Searches for Vector-Like Quarks

In this Chapter, the search for Vector-Like Quarks (VLQs) is presented. Two analyses were performed for the search of Vector-Like top quarks (T) where the T decays into Ht or Zt. The first analysis is dedicated to the search for a singly produced T in association with an electron or muon, referred to as the 1-lepton channel [1]. The second analysis is dedicated to the search for pair-produced Ts both in the 0-lepton and 1-lepton channels. Both analyses use 139 fb^{-1} of data recorded corresponding to the year period 2015-2018. Additionally, both analyses follow similar background modeling and event selection criteria as discussed in Chapter 4, as well as a similar search strategy that will be discussed in this Chapter.

The first part of this Chapter is devoted to the single production analysis and will discuss its strategy and results. The results are interpreted using two signal benchmarks: the $SU(2)$ singlet ($T^{2/3}$) gauge representation and the $SU(2)$ doublet ($T^{2/3}\,B^{-1/3}$) gauge representation. The involvement of the author of this thesis in this analysis was mostly limited to the derivation of correction factors that were designed to improve the modeling of the background processes. This has an important role in the overall analysis since having a well-modeled background is essential in the design of the analysis strategy and interpretation of results. More emphasis will be given to the overall search strategy of the analysis in this first part of the Chapter, which will serve as an introduction and motivation for the pair

[1] Throughout the remainder of this Chapter, the word lepton will refer specifically to either an electron or muon, unless otherwise stated.

production analysis. This analysis is now concluded, and its results have been published [75].

The second part of this Chapter is devoted to the currently ongoing pair production analysis. At the time of writing this dissertation, the 1-lepton channel of the analysis is far more developed than the 0-lepton channel. Only the 1-lepton channel will be discussed in the second part of this Chapter. However, the 0-lepton channel will follow a similar analysis strategy as the one that will be discussed for the 1-lepton channel. Since several aspects of the search strategy of this analysis are shared with the single production analysis, only the strategy components that are different will be discussed. The results of this analysis are interpreted using four signal benchmarks: the $SU(2)$ singlet gauge representation, the $SU(2)$ doublet gauge representation, assuming the branching ratio $\mathrm{BR}(T \to Ht) = 1$, and assuming the branching ratio $\mathrm{BR}(T \to Zt) = 1$.

6.1 Single Production of Vector-Like Quarks

6.1.1 Analysis Strategy

The single production analysis is optimized to search for the production of a T that decays to a top quark and either a Higgs or Z boson in the 1-lepton channel. The lepton is mainly expected to be produced from the leptonic decay of a top quark. However, other less frequent sources for the lepton include the dileptonic decay of the Z, in which one of the leptons is misreconstructed, for example. As discussed in subsection 4.2.1, the single production of a T is initiated through an electroweak interaction that results in the production of an associated top or bottom quark, referred to as the t-associated and b-associated production modes respectively. Thus, the signal processes in this analysis can be categorized based on the T decays and the associated production modes as follows:

1. $T(\rightarrow Ht)qb$ for the b-associated production of a T decaying into Ht

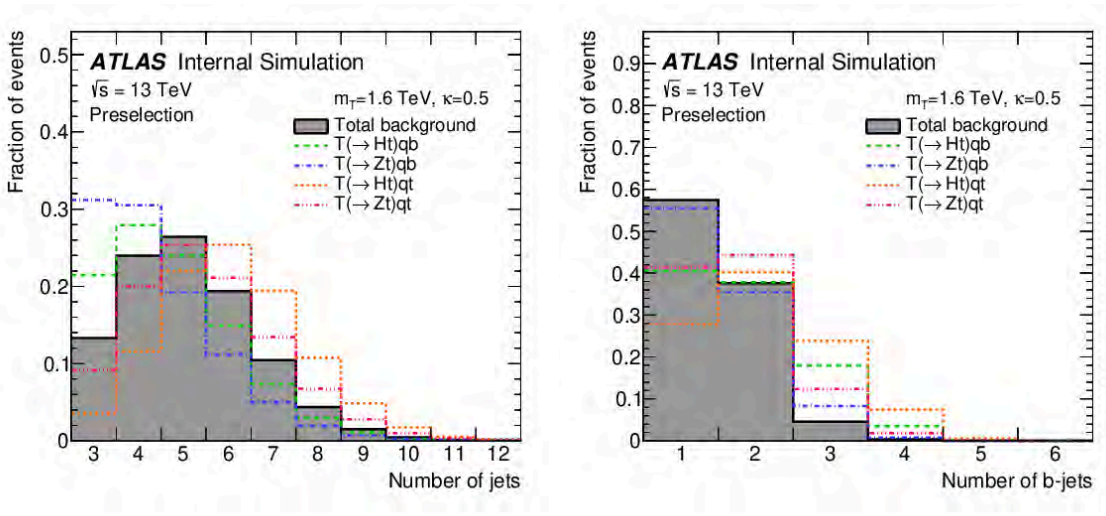
2. $T(\rightarrow Zt)qb$ for the b-associated production of a T decaying into Zt

3. $T(\rightarrow Ht)qt$ for the t-associated production of a T decaying into Ht

4. $T(\rightarrow Zt)qt$ for the t-associated production of a T decaying into Zt

Although the analysis strategy is optimized for these production modes and decay channels of the T, the search is aimed at the $SU(2)$ singlet ($T^{2/3}$) and doublet ($T^{2/3}\,B^{-1/3}$) gauge representations of VLQs, which are the signal benchmarks of the analysis. It should be noted that the coupling of the T to the W or Z boson is dependent on the $SU(2)$ gauge representation. In the case of the doublet representation, the coupling to the W boson vanishes due to charge conservation considerations, thereby making the t-associated production mode the only allowed mode for this representation. As previously discussed in subsection 4.2.1, even though the t-associated production mode is kinematically suppressed due to the mass of the top quark, studying both production modes is well motivated by the theory and extends the interpretability of the parameter space in the analysis results.

The distributions of the number of jets and b-tagged jets overlayed between the different signal processes just described and the total SM background are shown in Figure 6.1. The distributions are shown at the event preselection level that was discussed in subsection 4.2.3. For the b-associated production mode, the number of jets that originate directly from the main decay topology, which includes the associated b quark and the decay products of the T, is on average expected to be 4. However, if the Higgs or Z boson that originate from the T are highly boosted, then the two-pronged decay of these particles might not be identified and instead be reconstructed as a single jet, thereby reducing the number of jets in the event. Additionally, the associated b quark that originates from the initial gluon split can



(a) (b)

Figure 6.1: The distributions of the multiplicities of jets (a) and b-tagged jets (b) at preselection level overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

potentially decay in the high-pseudorapidity region of the detector due to its low mass, which will not be reconstructed as a central jet. As observed in Figure 6.1a, the bulk of the b-associated production mode populates the $3-5$ jet region, which is denoted as the low-jet (LJ) multiplicity region. On the other hand, the t-associated production mode mostly populates the ≥ 6 jets region, which is denoted as the high-jet (HJ) multiplicity region. Since the analysis is performed in the 1-lepton channel, at least one of the top quarks in the t-associated production mode is expected to decay hadronically. Thus, the total number of jets that are expected from the top quark decays alone can range from $2-5$, depending on the degree of collimation of the top decay products. In addition to the jets that originate directly from the main decay topology, the final state radiation of the signal processes can lead to the production of additional jets. Although, these additional jets will not be as energetic as the ones that arise directly from the main decay topology of signal processes.

Baseline selections on jet and b-tag multiplicities			
Jet multiplicity	b-tag multiplicity	Channel name	Targeted signal
3–5	1–2	LJ, 1–2b	$T(\rightarrow Zt)qb$
3–5	$\geq$3	LJ, $\geq$3b	$T(\rightarrow Ht)qb$
$\geq$6	1–2	HJ, 1–2b	$T(\rightarrow Zt)qt$
$\geq$6	$\geq$3	HJ, $\geq$3b	$T(\rightarrow Ht)qt$

Table 6.1: Definition of the four baseline analysis search regions based on jet and b-tagged jet multiplicity and the signal process which they are designed to target.

Another feature that distinguishes signal events from background events is the number of b-tagged jets in the event. From all the T production modes considered, the ones with the $T \rightarrow Ht$ decay channel are expected to have the largest number of b-tagged jets. This is due to the $H \rightarrow b\bar{b}$ decay channel, which has the largest branching ratio for the Higgs boson. Thus, for the $T \rightarrow Ht$ decay channel the number of jets that originate from the main decay topology and can potentially be b-tagged is 4. However, as previously discussed, the b-associated production mode could have fewer b-tagged jets due to the possibility of the associated b quark decaying in the high-pseudorapidity region, which lies outside of the validty range of the b-tagger used. On the other hand, for the $T \rightarrow Zt$ decay channel the expected number of b-tagged jets from the main decay topology ranges between $1 - 2$.

Taking these observations into account, four baseline analysis search regions are defined solely based on the multiplicity of jets and b-tagged jets that individually target each signal process. These baseline regions are summarized in Table 6.1.

6.1.2 Signal Discrimination

From the previous discussion it is clear that signal events can be isolated from background events by placing selection cuts on the multiplicity of jets and b-tagged jets. However,

these requirements are dependent on the signal decay channel and production modes by definition. Instead, a clever observation is to note that due to the large mass of the T, its decay products are expected to be highly boosted regardless of the signal process considered. Thus, the production of a large number of jets, of which a significant fraction is expected to be boosted, a potentially boosted lepton, and a significant amount of $E_{\mathrm{T}}^{\mathrm{miss}}$ from the leptonic decay of a boosted top quark motivates the definition of the effective mass (m_{eff}) variable:

$$m_{\mathrm{eff}} = \sum_{\text{central jets}} p_{\mathrm{T}}^{j} + \sum_{\text{leptons}} p_{\mathrm{T}}^{\ell} + E_{\mathrm{T}}^{\mathrm{miss}} \tag{6.1}$$

which is the scalar sum of the p_{T} of the jets, the lepton, and the $E_{\mathrm{T}}^{\mathrm{miss}}$ that are produced in the event. This variable allows us to discriminate between signal and background processes in a way that is agnostic to the signal decay channels and production modes. The distribution of m_{eff} overlayed between the different signal processes and the total SM background is shown in Figure 6.2 at the event preselection level. All signal processes that are shown are for a T with a mass $m_T = 1.6$ TeV. As can be observed in the plot, the distribution peaks close to m_T for signal processes, while for the SM background processes the distribution decays rapidly at higher values of m_{eff} due to these processes lacking the sufficient energy to produce highly boosted final states. As the mass of the T gets larger, the separation power between signal and background processes improves since the signal processes will populate the high m_{eff} region. Based on these observations, the m_{eff} variable is chosen as the variable on which the fit is performed in the statistical analysis (see subsection 6.1.7).

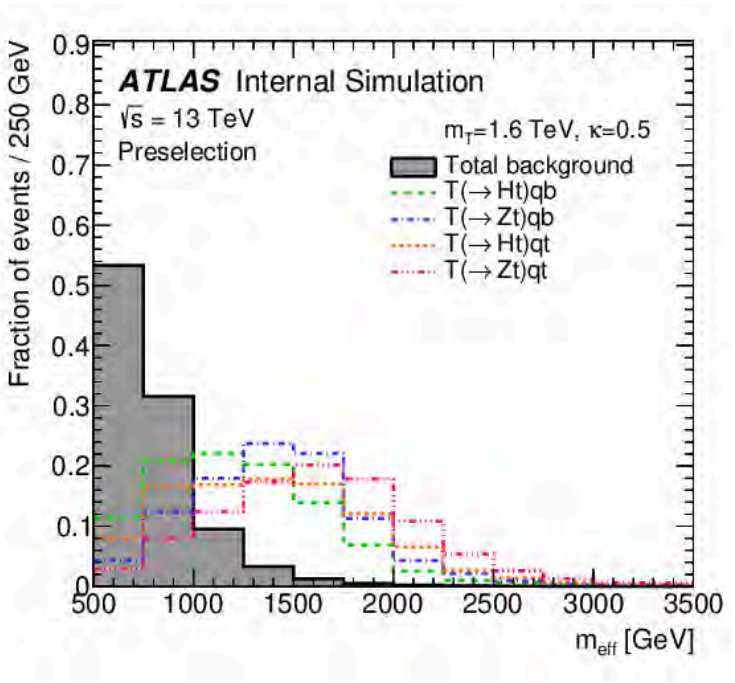
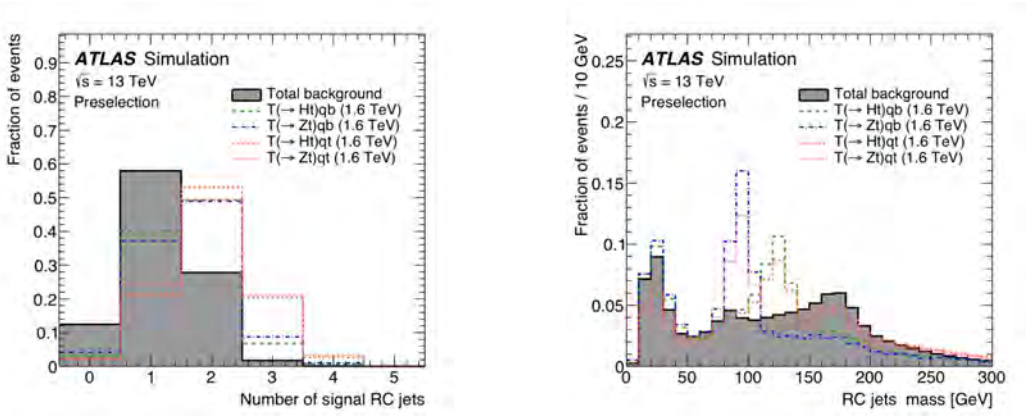


Figure 6.2: Distribution of m_{eff} at preselection level overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

6.1.3 Boosted Object Tagging and Reconstruction

As discussed in the previous section, due to the large mass of the T, a large number of boosted jets can be produced from the hadronic decays of the top quark, the Higgs boson, and the Z boson that are produced in the main decay topology of signal processes. Depending on the degree of collimation of the decay products of these particles, the jets that are produced can be reclustered into a single large-R jet. These reclustered jets can be used to identify the particle that originated them with the use of a tagging algorithm. Thus, this allows us to potentially reconstruct the direct decay products of the T by correctly tagging the reclustered large-R jets to their source particle. As discussed in subsection 3.3.4, variable radius RC jets are used as the inputs to the tagging algorithm due to their flexibility in capturing the decay products of boosted objects over a wide p_{T} regime. The distributions of the number of RC

jets in an event and their masses are shown in Figure 6.3. As can be observed, the number of RC jets is, on average, larger in signal events compared to events from the SM background. Additionally, the RC jets in signal processes exhibit prominent mass peaks that correspond to the direct decay products of the T. For the b-associated production modes, the mass peak near the top quark mass is less prominent when compared to the t-associated production modes. This is because the top quark in the b-associated production mode is likely to decay leptonically; thus, jets cannot be used to identify the leptonically decaying top.



(a) (b)

Figure 6.3: The distributions of the multiplicities of reclustered large-R jets (a) and their mass (b) at preselection level overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

The tagging algorithm that is implemented to identify the RC jets to their source particles is a simple kinematic variable cut-based tagger. The tagger takes as input the p_{T}, mass, and the number of subjet constituents (N_{const}) of the RC jets. The tagger is designed to

identify jets that are produced from hadronically decaying top quarks, Higgs bosons, and vector bosons inclusively. The tagger does not distinguish between W and Z bosons due to the similarity of the jets that are produced by these particles in the input variables of the tagger. However, this ambiguity does not impact the analysis significantly since the production of W bosons is not a central focus of the analysis. The kinematic requirements to tag a jet to a given particle are summarized in Table 6.2. The requirements on the RC

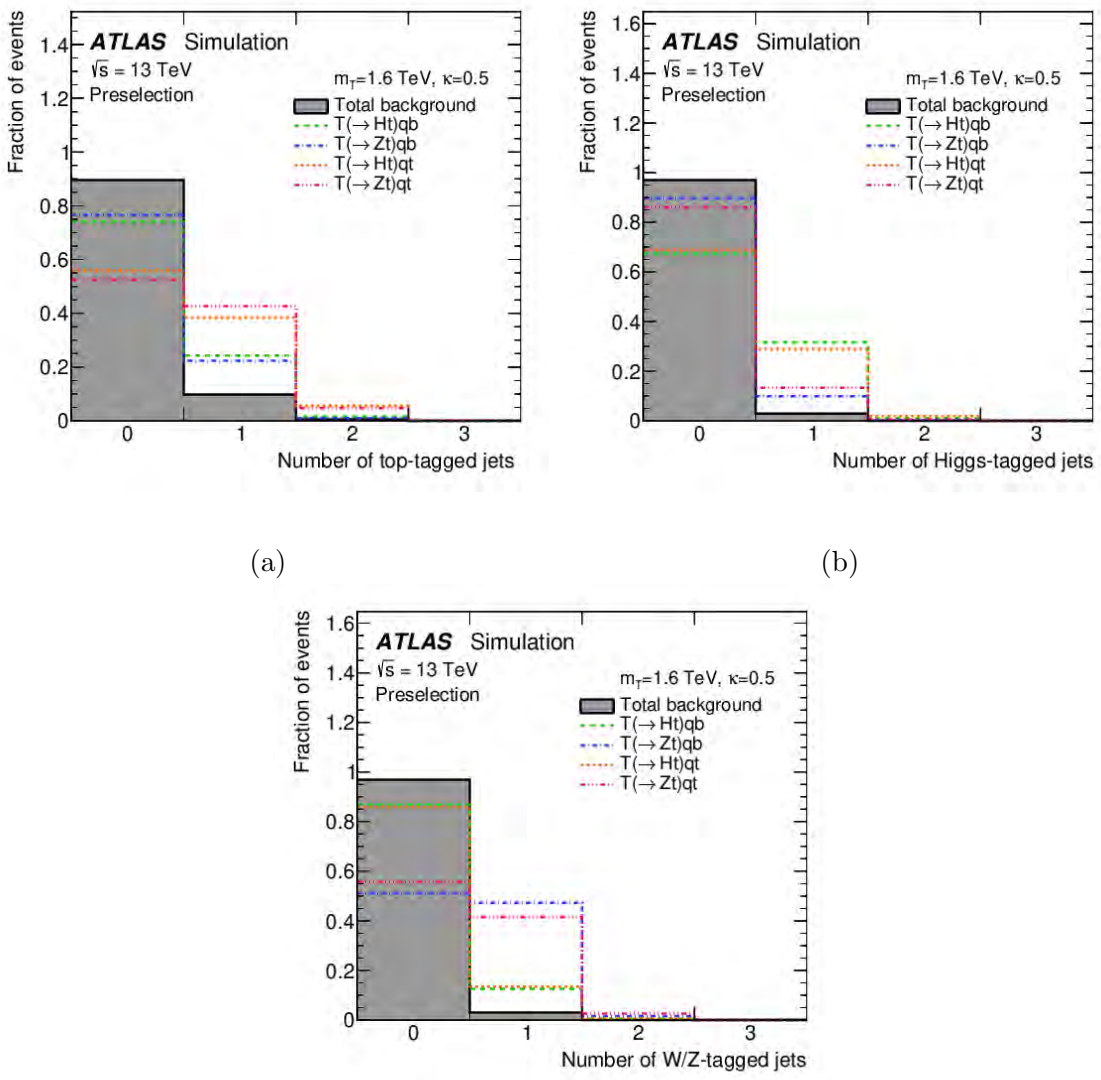
Kinematic Observable	t-tagged	H-tagged	V-tagged
p_{T} [GeV]	> 400	> 350	> 350
Mass [GeV]	> 140	$[105, 140]$	$[70, 105]$
N_{const}	≥ 2 if $p_{\mathrm{T}} < 700$ ≥ 1 if $p_{\mathrm{T}} > 700$	$= 2$ if $p_{\mathrm{T}} < 600$ ≤ 2 if $p_{\mathrm{T}} > 600$	$= 2$ if $p_{\mathrm{T}} < 450$ ≤ 2 if $p_{\mathrm{T}} > 450$

Table 6.2: Kinematic requirements on RC jets to be tagged to a top quark, a Higgs boson, or a vector boson.

jet p_{T} ensure that the majority of the decay products of the particles are captured within the jet. The mass requirements for each type of particle are designed to be orthogonal in order to have well-defined particle classes. Additionally, they allow the tagger to be flexible on jets that are highly boosted or do not capture all the subsequent decays of the source particle. Finally, the requirement on N_{const} is designed to capture the jet substructure by introducing a p_{T} dependence that adjusts to the desired jet topology. Jets that are highly boosted tend to have collimated decay products; thus, the requirements on N_{const} at high p_{T} allow for the merging of subjets. On the other hand, jets that have a lower p_{T} tend to have a resolved decay topology; thus, the requirements on N_{const} are higher or more exclusive compared to their corresponding high p_{T} requirement. The distributions of the number of jets that are tagged to a top quark, Higgs boson, and vector boson are shown in Figure 6.4 at the event preselection level. As can be observed from the plots, the signal processes tend to have a larger fraction of events with at least one jet tagged to a hadronically decaying

boosted object when compared to the SM background. The signal t-associated production modes have a larger fraction of events with at least one top-tagged jet when compared to the b-associated production modes, which is expected due to the presence of an additional top quark in the t-associated production mode. Similarly, the signal processes with the $T \to Ht$ decay channel have a larger fraction of events with a Higgs-tagged jet, while the processes with the $T \to Zt$ decay channel have a larger fraction of V-tagged jets.

In order to identify potential leptonically decaying top quarks that are produced in events, jets cannot be used due to the presence of the lepton and the $E_{\mathrm{T}}^{\mathrm{miss}}$ that originate from this decay. Instead, a dedicated algorithm is implemented to reconstruct a candidate leptonic top system from simple kinematic considerations. A schematic representation of this algorithm is shown in Figure 6.5. First, a candidate leptonically decaying W boson is reconstructed under the assumption that all the $E_{\mathrm{T}}^{\mathrm{miss}}$ from the event and its azimuthal angle are the same as those of the p_{T} of the neutrino that is produced from the leptonically decaying W boson. The longitudinal momentum component of the neutrino is determined by performing algebraic manipulations on the four-momenta of the neutrino and lepton in the event under the constraint that the invariant mass of the lepton-neutrino system is consistent with the mass of the W boson. The candidate leptonically decaying W boson is then reconstructed by adding the four-momenta of the lepton and reconstructed neutrino. Next, the candidate leptonically decaying W boson is spatially matched with the closest b-tagged jet within a distance of $\Delta R < 1.5$. Additionally, the b-tagged jet must not be a constituent of any tagged RC jet in the event. This is done in order to avoid potential double counting and to ensure that the leptonically decaying W is matched with the appropriate b-tagged jet that originates from the same top quark decay. If no such b-tagged jet exists, then the leptonic top is not reconstructed in the event. On the other hand, if the candidate leptonically decaying W



(a) (b)

(c)

Figure 6.4: The distributions of the multiplicities of reclustered large-R jets that are tagged to a hadronically decaying top quark (a), Higgs boson (b), and vector boson (c) at preselection level overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

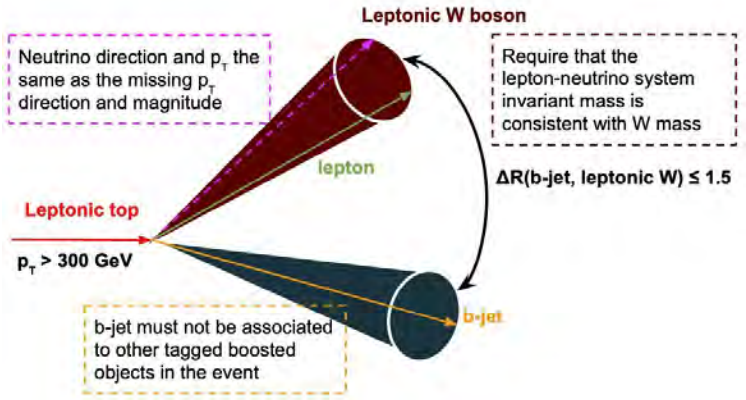


Figure 6.5: Schematic representation of the leptonic top reconstruction algorithm.

boson is matched with a free b-tagged jet and the p_{T} of the reconstructed W boson and b-tagged jet system satisfies $p_{\mathrm{T}} > 300$ GeV, then the resulting system is considered to be a reconstructed leptonic top. The distribution of the number of reconstructed candidate leptonic tops at the event preselection level is shown in Figure 6.6.

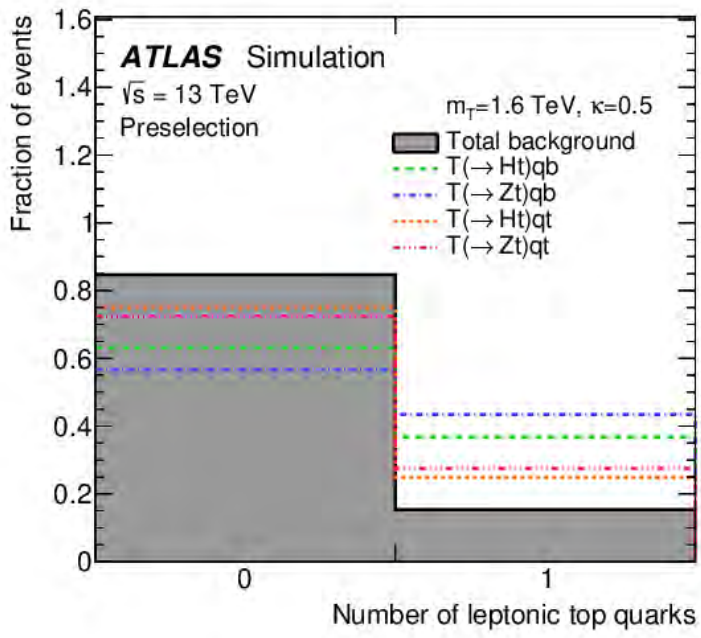


Figure 6.6: The number of reconstructed candidate leptonic tops at preselection level overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

6.1.4 Analysis Search Regions

In order to carry out the statistical analysis, dedicated search regions that are pure enough in the different signal processes considered in this search must be defined. Starting from the baseline search regions that were defined in subsection 6.1.1, the signal purity for the different signal processes can be enhanced by making additional requirements on the number of boosted objects in the event, which are tailored to a particular signal process. These requirements are also motivated by the fact that a larger number of boosted objects in events drives the m_{eff} distribution towards higher values in signal processes, thereby increasing the overall separation power of m_{eff} between signal and background.

For example, the baseline regions that require $\geq 3b$ are designed to be sensitive to the $T \to Ht$ decay channel; therefore, requiring the presence of at least one Higgs-tagged jet would increase the purity of this decay channel. Similarly, in the 1–2 b-tagged jets regions the purity of the $T \to Zt$ decay channel can be increased by requiring at least one V-tagged jet. Additionally, the presence of a V-tagged jet can also improve the sensitivity of signal events where a semi-boosted hadronically decaying top quark is produced. This could happen in events where no jets are tagged to the top quark but instead to the W boson originating from the top decay. The presence of top-tagged jets can be used to increase the sensitivity of t-associated production modes where one of the top quarks must decay hadronically. Regions with at least one top-tagged jet can also improve the sensitivity of rare processes such as $T \to Ht$ decays where $H \to WW/\tau\tau$ and the lepton is produced from a leptonically decaying W or τ. Similarly, signal events with $T \to Zt$ decays where $Z \to \ell\ell$ and one of the leptons is misreconstructed can also gain sensitivity if a jet is tagged to a top quark. Finally, as discussed in subsection 4.2.1, at least one forward jet in signal processes is expected to be

produced from the initial quark that recoils off from the off-shell W/Z boson. Background events, on the other hand, are usually not energetic enough to produce jets in the forward region of the detector, as shown in Figure 6.7. Thus, the overall signal purity can be further improved by requiring the presence of at least one forward jet in the analysis search regions.

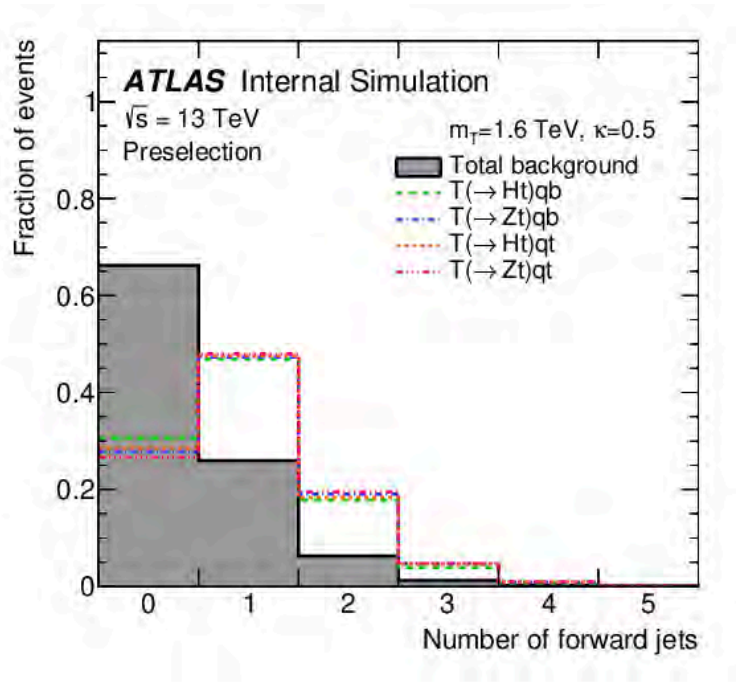


Figure 6.7: The distribution of the number of forward jets at preselection level overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

To summarize this discussion, the analysis search regions, which will be referred to as analysis fit regions, are obtained from the baseline search regions by making additional requirements on the number of forward jets (fj), Higgs-tagged jets (H), top-tagged jets (t_h), vector boson-tagged jets (V), and reconstructed leptonic tops (t_l). The combined use of all these regions in a likelihood fit allows for the analysis search to retain sensitivity to all of the signal processes that can occur in a given signal benchmark model. In addition to the fit regions that target the different signal processes, two background control regions

are included in the set of fit regions. The purpose of these regions is to calibrate and constrain the normalization of the $t\bar{t}$ production in association with at least one b-tagged jet ($t\bar{t}+ \geq 1b$) when performing the likelihood fit. These control regions are defined by requiring the presence of a leptonic top, at least 4 b-tagged jets, and a veto of forward jets and hadronically decaying boosted objects. In total there are 24 analysis fit region, which are summarized in Table 6.3.

In order to ensure that the background is well-modeled in the analysis fit regions, a set of 20 validation regions that are kinematically similar to the fit regions and are signal-depleted are defined. This is achieved by either requiring a veto on forward jets or inverting the most relevant boosted object multiplicity requirement of a given fit region. The validation regions are summarized in Table 6.4.

Fit regions with 3–5 jets			
b-tag mult.	Boosted-object mult.	Region name	Targeted signal / bkg
1	0(t_h+t_l), 0H, ≥1V	LJ, 1b, ≥1fj, 0(t_h+t_l), 0H, ≥1V	$T(\to Zt)qb$
1	0t_h, ≥1t_l, 0H, ≥1V	LJ, 1b, ≥1fj, 0t_h, ≥1t_l, 0H, ≥1V	$T(\to Zt)qb$
2	0(t_h+t_l), 0H, ≥1V	LJ, 2b, ≥1fj, 0(t_h+t_l), 0H, ≥1V	$T(\to Zt)qb$
2	0t_h, ≥1t_l, 0H, ≥1V	LJ, 2b, ≥1fj, 0t_h, ≥1t_l, 0H, ≥1V	$T(\to Zt)qb$
3	0(t_h+t_l), ≥1H, 0V	LJ, 3b, ≥1fj, 0(t_h+t_l), ≥1H, 0V	$T(\to Ht)qb$
3	0t_h, ≥1t_l, ≥1H, 0V	LJ, 3b, ≥1fj, 0t_h, ≥1t_l, ≥1H, 0V	$T(\to Ht)qb$
3	≥1t_h, 0t_l, ≥1H, 0V	LJ, 3b, ≥1fj, ≥1t_h, 0t_l, ≥1H, 0V	$T(\to Ht)qb$
≥4	0(t_h+t_l), ≥1H, 0V	LJ, ≥4b, ≥1fj, 0(t_h+t_l), ≥1H, 0V	$T(\to Ht)qb$
≥4	0t_h, ≥1t_l, ≥1H, 0V	LJ, ≥4b, ≥1fj, 0t_h, ≥1t_l, ≥1H, 0V	$T(\to Ht)qb$
≥4	≥1t_h, 0t_l, ≥1H, 0V	LJ, ≥4b, ≥1fj, ≥1t_h, 0t_l, ≥1H, 0V	$T(\to Ht)qb$
≥4	≥1t_l, 0H, 0(V+t_h)	LJ, ≥4b, 0fj, ≥1t_l, 0H, 0(V+t_h)	$t\bar{t}+ \geq 1b$
Fit regions with ≥6 jets			
b-tag mult.	Boosted-object mult.	Region name	Targeted signal / bkg
1	0t_h, 1t_l, 0H, ≥1V	HJ, 1b, ≥1fj, 0t_h, 1t_l, 0H, ≥1V	$T(\to Zt)qt$
1	1t_h, 0t_l, 0H, ≥1V	HJ, 1b, ≥1fj, 1t_h, 0t_l, 0H, ≥1V	$T(\to Zt)qt$
1	≥2(t_h+t_l), 0H, ≥1V	HJ, 1b, ≥1fj, ≥2(t_h+t_l), 0H, ≥1V	$T(\to Zt)qt$
2	0t_h, 1t_l, 0H, ≥1V	HJ, 2b, ≥1fj, 0t_h, 1t_l, 0H, ≥1V	$T(\to Zt)qt$
2	1t_h, 0t_l, 0H, ≥1V	HJ, 2b, ≥1fj, 1t_h, 0t_l, 0H, ≥1V	$T(\to Zt)qt$
2	≥2(t_h+t_l), 0H, ≥1V	HJ, 2b, ≥1fj, ≥2(t_h+t_l), 0H, ≥1V	$T(\to Zt)qt$
3	1t_l, ≥1H, 0(V+t_h)	HJ, 3b, ≥1fj, 1t_l, ≥1H, 0(V+t_h)	$T(\to Ht)qt$
3	0t_l, ≥1H, 1(V+t_h)	HJ, 3b, ≥1fj, 0t_l, ≥1H, 1(V+t_h)	$T(\to Ht)qt$
3	≥1H, ≥2(V+t_l+t_h)	HJ, 3b, ≥1fj, ≥1H, ≥2(V+t_l+t_h)	$T(\to Ht)qt$
≥4	1t_l, ≥1H, 0(V+t_h)	HJ, ≥4b, ≥1fj, 1t_l, ≥1H, 0(V+t_h)	$T(\to Ht)qt$
≥4	0t_l, ≥1H, 1(V+t_h)	HJ, ≥4b, ≥1fj, 0t_l, ≥1H, 1(V+t_h)	$T(\to Ht)qt$
≥4	≥1H, ≥2(V+t_l+t_h)	HJ, ≥4b, ≥1fj, ≥1H, ≥2(V+t_l+t_h)	$T(\to Ht)qt$
≥4	≥1t_l, 0H, 0(V+t_h)	HJ, ≥4b, 0fj, ≥1t_l, 0H, 0(V+t_h)	$t\bar{t}+ \geq 1b$

Table 6.3: Definition of the 24 analysis search regions (referred to as "fit regions"). The events are categorized based on the multiplicity of central jets (j), b-tagged jets (b), forward jets (fj), V-tagged jets (V), Higgs-tagged jets (H), hadronic top tagged jets (t_h), and reconstructed leptonic tops (t_l).

Validation regions with 3–5 jets			
b-tag mult.	Fwd-jet mult.	Boosted-object mult.	Region name
1	0	$0t_h$, $0t_l$, 0H, $\geq$1V	LJ, 1b, 0fj, $0t_h$, $0t_l$, 0H, $\geq$1V
1	0	$0t_h$, $\geq1t_l$, 0H, $\geq$1V	LJ, 1b, 0fj, $0t_h$, $\geq1t_l$, 0H, $\geq$1V
1	$\geq$1	$\geq1(t_h+t_l)$, 0H, 0V	LJ, 1b, $\geq$1fj, $\geq1(t_h+t_l)$, 0H, 0V
1	$\geq$1	$\geq1t_h$, $0t_l$, 0H, $\geq$1V	LJ, 1b, $\geq$1fj, $\geq1t_h$, $0t_l$, 0H, $\geq$1V
2	0	$0t_h$, $0t_l$, 0H, $\geq$1V	LJ, 2b, 0fj, $0t_h$, $0t_l$, 0H, $\geq$1V
2	0	$0t_h$, $\geq1t_l$, 0H, $\geq$1V	LJ, 2b, 0fj, $0t_h$, $\geq1t_l$, 0H, $\geq$1V
2	$\geq$1	$\geq1(t_h+t_l)$, 0H, 0V	LJ, 2b, $\geq$1fj, $\geq1(t_h+t_l)$, 0H, 0V
2	$\geq$1	$\geq1t_h$, $0t_l$, 0H, $\geq$1V	LJ, 2b, $\geq$1fj, $\geq1t_h$, $0t_l$, 0H, $\geq$1V
$\geq$3	0	$0(t_h+t_l)$, $\geq$1H, 0V	LJ, $\geq$3b, 0fj, $0(t_h+t_l)$, $\geq$1H, 0V
$\geq$3	$\geq$1	0H, $\geq1(V+t_l+t_h)$	LJ, $\geq$3b, $\geq$1fj, 0H, $\geq1(V+t_l+t_h)$
Validation regions with $\geq$6 jets			
b-tag mult.	Fwd-jet mult.	Boosted-object mult.	Region name
1	0	$1(t_h+t_l)$, 0H, $\geq$1V	HJ, 1b, 0fj, $1(t_h+t_l)$, 0H, $\geq$1V
1	0	$\geq2(t_h+t_l)$, 0H, $\geq$1V	HJ, 1b, 0fj, $\geq2(t_h+t_l)$, 0H, $\geq$1V
1	$\geq$1	$0t_h$, $0t_l$, $\geq$1H, $\geq$1V	HJ, 1b, $\geq$1fj, $0t_h$, $0t_l$, $\geq$1H, $\geq$1V
1	$\geq$1	$\geq2(t_h+t_l)$, $\geq$1H, 0V	HJ, 1b, $\geq$1fj, $\geq2(t_h+t_l)$, $\geq$1H, 0V
2	0	$1(t_h+t_l)$, 0H, $\geq$1V	HJ, 2b, 0fj, $1(t_h+t_l)$, 0H, $\geq$1V
2	0	$\geq2(t_h+t_l)$, 0H, $\geq$1V	HJ, 2b, 0fj, $\geq2(t_h+t_l)$, 0H, $\geq$1V
2	$\geq$1	$0t_h$, $0t_l$, $\geq$1H, $\geq$1V	HJ, 2b, $\geq$1fj, $0t_h$, $0t_l$, $\geq$1H, $\geq$1V
2	$\geq$1	$\geq2(t_h+t_l)$, $\geq$1H, 0V	HJ, 2b, $\geq$1fj, $\geq2(t_h+t_l)$, $\geq$1H, 0V
$\geq$3	0	$\geq$1H, $\geq1(V+t_l+t_h)$	HJ, $\geq$3b, 0fj, $\geq$1H, $\geq1(V+t_l+t_h)$
$\geq$3	$\geq$1	0H, $\geq1(V+t_l+t_h)$	HJ, $\geq$3b, $\geq$1fj, 0H, $\geq1(V+t_l+t_h)$

Table 6.4: Definition of the 20 analysis validation regions that are designed to validate the kinematic modeling of the background processes in the analysis fit regions. The validation regions are obtained by either vetoing forward jets in the events or inverting the most relevant boosted object multiplicity requirements in the fit regions that are defined in Table 6.3.

6.1.5 Kinematic Reweighting of Background

As discussed in subsection 6.1.2, the m_{eff} variable has a good separation power between signal and background, which stems from its dependence on the p_{T} and multiplicities of final state objects in an event. However, recent measurements have demonstrated that the MC simulations of the dominant $t\bar{t}$ background process and subdominant V+jets background processes mismodel the p_{T} and multiplicity of jets that are produced in events from these processes. In the case of $t\bar{t}$ processes, it is observed that the MC simulation overestimates the cross section of this process at large values of the jet p_{T} spectrum [76] and underestimates it at high jet multiplicities [77]. A similar issue is also present in the modeling of V+jets processes in the high jet multiplicity region and $H_{\text{T}}^{\text{had}2}$ [78]. These mismodelings on the MC simulation enter as an additional source of mismodeling on m_{eff} due to how it is defined. This strongly impacts the high-tail of the m_{eff} distribution where most of the signal is expected to reside, as shown in Figure 6.8.

In order to fix the mismodeling introduced by these background processes, data-driven correction factors are derived in kinematic regions that are enriched in the background process that is to be corrected and are depleted in signal events. These regions are referred to as reweighting source regions (RSRs) and are summarized in Table 6.5. The potential contamination from signal processes in the RSRs was quantified and found to be negligible. The reweighting of $t\bar{t}$+jets is done jointly with the single-top Wt-channel background, denoted as $t\bar{t} + Wt$, due to both processes sharing the same final state and therefore being subject to interference. Additionally, both processes are generated using the same MC generator, thus sharing similar mismodeling. The $t\bar{t} + Wt$ RSR is defined in a way that is close to the preselection level of the analysis, which is dominated by the $t\bar{t}$+jets process. For the V+jets

[2] $H_{\text{T}}^{\text{had}}$ is defined as the scalar sum of the p_{T} of all central jets in the event.

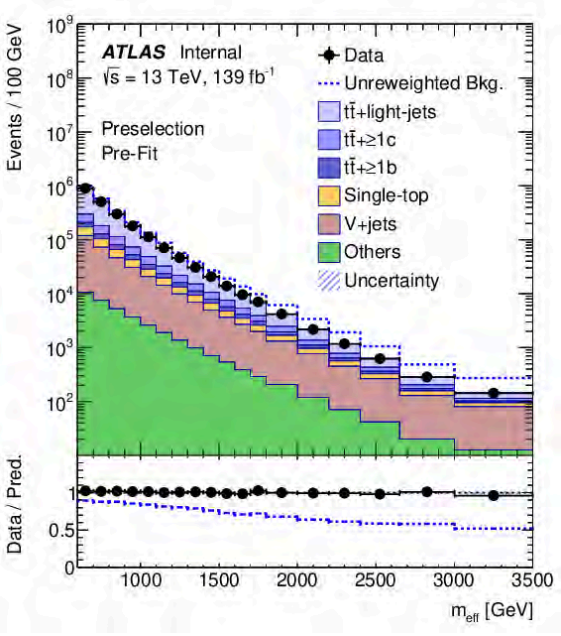


Figure 6.8: Comparison of the m_{eff} distribution between the data, the mismodeled background prediction (blue dashed line), and the reweighted background prediction at preselection level before performing the likelihood fit. The "Others" background includes the $t\bar{t}V/H$, VH, tZ, $t\bar{t}t\bar{t}$, diboson, and multijet production background processes. The bottom panel shows the ratios of data to the total mismodeled background prediction and the total reweighted background prediction.

background, there is no region in the analysis that is pure enough in the W+jets process so that it can be isolated to derive a correction factor for it. Instead, the correction factor is derived for Z+jets in a RSR that requires exactly two same-flavored leptons in order to isolate this process from other backgrounds. The Z+jets RSR requires at least 3 jets with exactly one b-tagged jet in order to maintain kinematic consistency with the preselection region. Additionally, in order to increase the purity of Z+jets in its RSR, the invariant mass of the dilepton system ($m_{\ell\ell}$) is required to be consistent with the mass of the Z boson. Finally, to further reduce the contamination from $t\bar{t}$+jets, a $E_{\text{T}}^{\text{miss}} < 100$ GeV cut is applied. The correction factor that is derived for Z+jets is assumed to be valid also for W+jets.

The correction factors are defined in the same way for all background processes that are to be reweighted. For a background process a and kinematic variable x, the correction factor

Reweighting source regions				
Lepton multiplicity	Jet multiplicity	b-tag multiplicity	Additional cuts	Targeted background
1	$\geq$3	2	–	$t\bar{t}+Wt$
2	$\geq$3	1	$\lvert m_{\ell\ell}-M_Z\rvert \leq 10$ GeV, $E_{\mathrm{T}}^{\mathrm{miss}} < 100$ GeV	Z+jets

Table 6.5: Reweighting source regions from which the reweighting functions for $t\bar{t}$ and Wt production and W/Z+jets production are derived.

is calculated as:

$$R_a(x) = \frac{\mathrm{Data}(x) - \mathrm{MC}_{\mathrm{non}-a}(x)}{\mathrm{MC}_a(x)} \tag{6.2}$$

where $\mathrm{Data}(x)$, $\mathrm{MC}_{\mathrm{non}-a}(x)$, and $\mathrm{MC}_a(x)$ denote the distributions of the kinematic variable x in data, the total background MC simulation excluding the process to be reweighted, and the MC simulation of the process to be reweighted respectively. Since Equation 6.2 is defined through binned distributions of a kinematic variable x, the correction factor $R_a(x)$ is also a binned distribution that is a function of x. The reweighting procedure for the background processes to be corrected can be summarized in the following steps:

1. Derive a bin-by-bin jet multiplicity correction factor ($R_a(N_{\mathrm{jets}})$) in the RSR for process a.

2. For each event of the process a, apply the binned correction factor that corresponds to the value of N_{jets} in the event as an event weight.

3. Derive a bin-by-bin correction factor for the reduced effective mass variable ($m_{\mathrm{eff}}^{\mathrm{red}}$)[3] in the RSR for process a after applying $R_a(N_{\mathrm{jets}})$.

4. Perform a functional fit on $R_a(m_{\mathrm{eff}}^{\mathrm{red}})$ in order to mitigate statistical effects and then

[3]The reduced effective mass variable is defined as $m_{\mathrm{eff}}^{\mathrm{red}} = m_{\mathrm{eff}} - (N_{\mathrm{jets}} - 3) \times 50$ GeV.

apply it as an event weight in conjuction with $R_a(N_{\text{jets}})$ to the unreweighted process a.

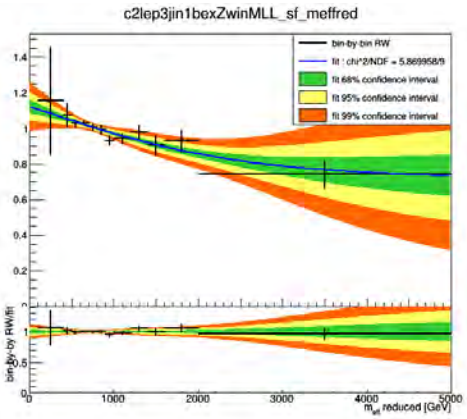
This procedure is first applied to the V+jets background and then subsequently applied to $t\bar{t} + Wt$ using the corrected V+jets background. The motivation of using $m_{\text{eff}}^{\text{red}}$ as the kinematic variable to derive the correction factor that addresses the m_{eff} mismodeling is twofold. First, the constant factor of 50 GeV in its definition approximately corresponds to the average p_{T} of each additional jet that is produced at preselection level in $t\bar{t}$ events. Second, the reweighting of $m_{\text{eff}}^{\text{red}}$ for $t\bar{t} + Wt$ is performed in exclusive N_{jets} bin regions for $N_{\text{jets}} = 3, 4, 5$, and 6, and inclusively for $N_{\text{jets}} \geq 7$. This is implemented in this way because the additional jets in $t\bar{t}$ processes can arise directly from the main decay topology of this process in the exclusive N_{jets} bin regions and therefore can strongly influence the shape of the correction factor in a N_{jets}-dependent way. Thus, including the $N_{\text{jets}} - 3$ shift in the definition of $m_{\text{eff}}^{\text{red}}$ reduces the N_{jets} dependence on the correction factors. In the case of $N_{\text{jets}} \geq 7$, the additional jets start to arise from outside the main decay topology of the $t\bar{t}$ system and thus lack sufficient energy to influence the shape of the correction factor. Thus, the use of $m_{\text{eff}}^{\text{red}}$ allows for an inclusive correction factor in high jet multiplicities instead of requiring an individual factor for each N_{jets} bin. Unlike the derivation of the $m_{\text{eff}}^{\text{red}}$ correction factor for $t\bar{t} + Wt$, the V+jets correction factor is derived inclusively in jet multiplicity due to the low event statistics available in the Z+jets RSR. Furthermore, the additional jets in V+jets processes come mostly from final state radiation and thus lack the energy to strongly influence the shape of the correction factor.

The final step in the reweighting procedure is to perform a functional fit to $R_a(m_{\text{eff}}^{\text{red}})$ in order to reduce the effects from statistical limitations in the extreme regions of $m_{\text{eff}}^{\text{red}}$. All the different $R_a(m_{\text{eff}}^{\text{red}})$ use the same functional form template, which is given by the following

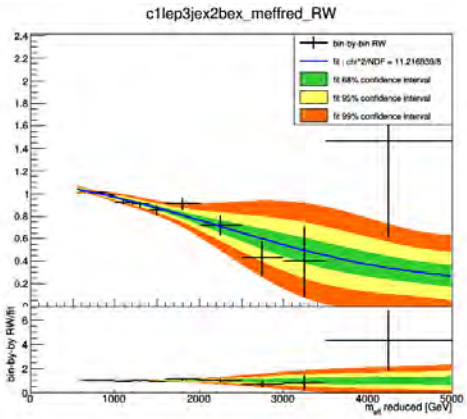
sigmoid function:

$$f(x) = p_1 - \frac{p_2}{1 + \exp\left(p_3(x - q)\right)} \tag{6.3}$$

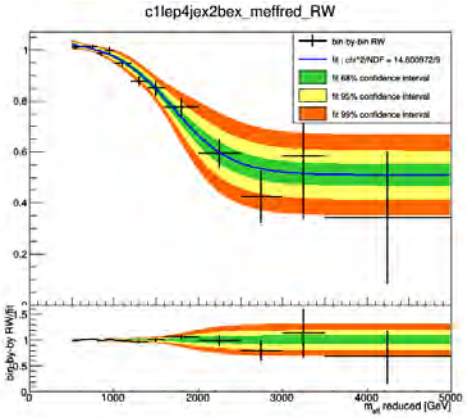
The values of the p_i parameters are determined from the fit, while the parameter q is a fixed parameter that varies for each different fit. An uncertainty is assigned to the reweighting procedure by applying a $\pm 2\sigma$ variation on the fit function to take into account the statistical uncertainty on the bin-by-bin reweighting and the potential shape differences on the fit template. The uncertainties for each fit are propagated as nuisance parameters on the binned likelihood fit which will be discussed in the following sections. The correction factors and their fits with the corresponding 1, 2, and 3σ bands are shown in Figure 6.9. As shown in Figure 6.8 the background reweighting improves significantly the MC modeling at preselection level. Additional plots are included in Appendix D that demonstrate how well the background reweighting procedure extends to other kinematic observables and selection regions in the analysis.



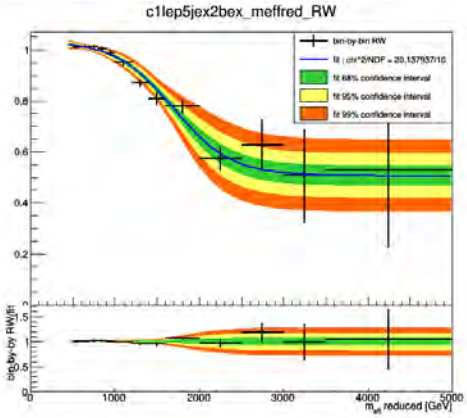
(a)



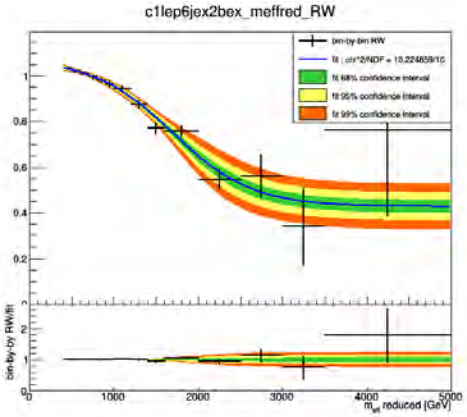
(b)



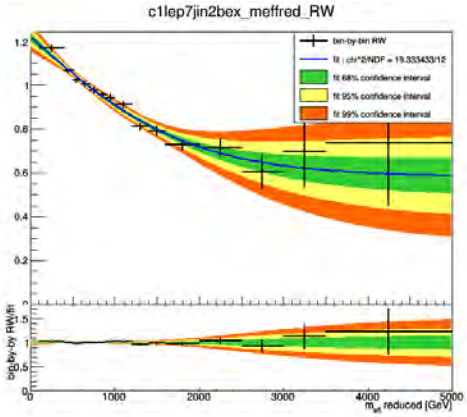
(c)



(d)



(e)



(f)

Figure 6.9: The V+jets reweighting correction factor in the Z+jets RSR and the $t\bar{t}$+Wt reweighting correction factors in their corresponding RSRs. The black markers correspond to the bin-by-bin correction factors with their associated statistical uncertainty. The solid blue line corresponds to the best fit obtained using the sigmoid template. The green, yellow, and orange bands correspond to the ± 1, 2, and 3σ confidence intervals of the fit, respectively. The bottom panel shows the ratio of the bin-by-bin correction factor to the fit.

6.1.6 Systematic Uncertainties

Several sources of systematic uncertainty are considered that can affect either the normalization or both the normalization and the shape of the m_{eff} distribution. Each systematic uncertainty is considered to be correlated across processes, analysis regions, and bins of m_{eff}, unless explicitly stated otherwise in the following description. Uncertainties from different sources are considered to be uncorrelated from each other. The sources of systematic uncertainty can be classified as either experimental uncertainties or modeling uncertainties.

6.1.6.1 Experimental Uncertainties

The experimental uncertainties are associated with the data taking and object reconstruction procedures by the ATLAS detector. The leading sources of experimental uncertainties in this search arise from the jet flavor tagging efficiencies and the jet mass resolution.

Luminosity The uncertainty in the combined 2015-2018 integrated luminosity is 1.7%, which affects the overall normalization of all simulated processes. It is obtained using the LUCID-2 detector [79] for the primary luminosity measurements.

Lepton Uncertainties These uncertainties are associated with the lepton triggering, selection, reconstruction, and identification processes. Additionally, data to MC scale factors are derived to calibrate the efficiencies of these processes in MC to data. The uncertainties associated with these scale factors are also considered. The overall effect of these uncertainties results in a normalization uncertainty in signal and background of approximately 1%.

Jet and $E_{\mathrm{T}}^{\mathrm{miss}}$ Uncertainties The uncertainties associated with jets arise from the jet energy scale (JES) and resolution (JER), the jet mass scale (JMS) and resolution (JMR), and the efficiency of the jet vertex tagger (JVT) [80] requirements that are imposed to reject jets from pile-up. The JES and JER uncertainties are estimated from a combination of collision data, test-beam data and simulation. The JES and JER uncertainties are split into 30 and 8 uncorrelated components, respectively, that correspond to different physical sources. The JMS uncertainty is estimated by comparing each nominal sample to two corresponding alternative event samples in which the mass of each jet is shifted up and down by 10%, respectively. A similar procedure is applied to estimate the JMR uncertainty by comparing each nominal sample to an alternative event sample in which the mass of each jet is smeared by a Gaussian function whose width is shifted by 20% relative to the nominal JMR.

The $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction is affected by uncertainties associated with the energy scales and resolutions of leptons and jets that are propagated to $E_{\mathrm{T}}^{\mathrm{miss}}$. Additional small uncertainties associated with the impact on the p_{T} scale and resolution of the unclustered energy from the underlying event are also taken into account as part of the $E_{\mathrm{T}}^{\mathrm{miss}}$ uncertainties.

Flavor Tagging Uncertainties The uncertainties associated with the efficiency of tagging jets to b-, c-, and light-quarks and the data to MC scale factors used to calibrate the b-tagging algorithm efficiency. These uncertainties are broken down into a set of 9 independent uncertainty sources for b-jets, 5 independent uncertainty sources for c-jets, and 6 independent uncertainty sources for light-jets. Additionally, an extrapolation uncertainty component is considered for high p_{T} jets that are outside the kinematic regime of the data sample that is used to calibrate the b-tagger. This component is taken to be correlated amongst the different jet flavors.

Table 6.6 summarizes the sources of experimental uncertainties, including whether they affect only the normalization (N), or both the normalization and the shape (SN) of the m_{eff} distribution, as well as the number of uncorrelated components.

6.1.6.2 Modeling Uncertainties

The modeling uncertainties are associated to the modeling of the MC simulations of SM background processes. The sources of uncertainties considered are normalization uncertainties related to the cross section of the different processes and the uncertainties related to the modeling parameters used to obtain the simulation samples. For small background processes in the analysis, only the cross section uncertainties are considered. For the dominant background source in the analysis, modeling uncertainties are considered in addition to the cross section uncertainties. The modeling parameter uncertainties are estimated by comparing the nominal simulation sample with specialized alternative samples that are obtained by systematically varying these modeling parameters. These alternative samples are described in Appendix A. Finally, the uncertainties associated with the background reweighting procedure described in subsection 6.1.5 are also included as modeling uncertainties. The leading sources of modeling uncertainties in this search arise from the modeling of $t\bar{t}$ and single-top Wt-channel backgrounds.

Cross Section Uncertainties An uncertainty of $+5.5/-6.1\%$ is assigned to the inclusive $t\bar{t}$ production cross section [81], which includes contributions from varying the factorization and renormalization scales, the PDF, the α_{S} parameter, and the value of the top quark mass. Additionally, a normalization uncertainty of 50% is assigned to the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ processes individually. These uncertainties are motivated by the level of agreement between

data and MC simulation in dedicated measurements of the cross section of the $t\bar{t}+\geq 1b$ process [82].

For single-top processes a $\pm 5\%$ uncertainty on the total cross section is estimated as a weighted average of the theoretical uncertainties on the t-, s-, and Wt-channel productions [83, 84, 85].

For the V+jets backgrounds a $\pm 30\%$ normalization uncertainty is applied and kept correlated between W+jets and Z+jets processes but uncorrelated between different b-tag multiplicities (1,2,3,$\geq$4). This is based on variations of the factorization and renormalization scales and the SHERPA [86] MC generator matching parameters [87] that are used to generate samples of these processes.

For the $t\bar{t}W/Z$ and $t\bar{t}H$ processes, since their contribution in the analysis fit regions is small, only cross section uncertainties of these processes are considered. The assigned uncertainty is $\pm 15\%$ for $t\bar{t}W/Z$, which is decorrelated between the LJ and HJ regions, and $+9/-12\%$ for $t\bar{t}H$, which is kept correlated.

For diboson processes a 5% uncertainty on the inclusive cross section calculated at NLO [88] is included. An additional uncorrelated 24% uncertainty on the production cross section is considered for each additional jet in the event that is based on a comparison amongst different algorithms for merging LO matrix elements and parton showers [89]. This uncertainty is computed based on the average jet multiplicity in each fit region which is approximately 3 in the LJ regions and 6 in the HJ regions. A $\pm 30\%$ normalization uncertainty on the production of additional heavy flavor jets is also considered and only applied in the fit regions with $\geq 3b$-jets. All of these uncertainties are added in quadrature and decorrelated between the LJ and HJ regions, as well as between low b-tag and high b-tag multiplicity regions. The total magnitude of the normalization uncertainty on diboson processes in the

LJ region is 24% and 38% in the low b-tag and high b-tag multiplicity regions respectively, whereas the HJ regions it is 48% and 56% in the low b-tag and high b-tag regions respectively.

Sample Modeling Uncertainties A number of sources of systematic uncertainties that affect the modeling of $t\bar{t}$+jets and single-top production processes are considered. The uncertainties associated to the choice of the NLO generator, the modeling of the parton showering and hadronization processess, and the modeling of the initial-state radiation (ISR) and final-state radiation (FSR) are estimated by comparing the nominal simulation samples to alternative samples as outlined in Appendix A. These uncertainties are all treated as uncorrelated between the $t\bar{t}$+light-jets, $t\bar{t}+\geq 1c$, $t\bar{t}+\geq 1b$, and single-top samples but correlated across the the single-top s-, t-, and Wt-channels. Furthermore, these uncertainties are treated as uncorrelated amongst the LJ and HJ analysis regions and regions with 0, 1, or $\geq$2 tagged boosted objects. An additional systematic uncertainty on the Wt-channel production concerning the separation between $t\bar{t}$ and Wt at NLO is assessed by comparing the nominal sample that utilizes the diagram-subtraction scheme to an alternative sample using the diagram-removal scheme.

An additional set of modeling uncertainties on the V+jets background is also considered. These uncertainties are estimated from variations in the internal renormalization and factorization scale parameters in the Sherpa MC generator.

Background Reweighting Uncertainties As discussed in subsection 6.1.5, an uncertainty is associated to the kinematic reweighting of the $t\bar{t}$+Wt and V+jets processes by varying the functional fit of the bin-by-bin reweighting by $\pm 2\sigma$ in order to consider the statistical limitations in the derivation procedure and the choice of the functional template. Additionally, the reweighting procedure is applied to the alternative $t\bar{t}$ and single-top Wt-

channel samples. This is required since the alternative samples also share similar mismodelings as the ones affecting the nominal samples and also in order to maintain kinematic consistency between the nominal and alternative samples when estimating the modeling uncertainties. The reweighting correction factors that are applied to the alternative samples are derived from the nominal correction factor $R_{a\ \mathrm{nom}}(x)$ sample as:

$$R_{a\ \mathrm{alt}}(x) = \frac{\mathrm{MC}_{a\ \mathrm{nom}}(x)}{\mathrm{MC}_{a\ \mathrm{alt}}(x)} R_{a\ \mathrm{nom}}(x) \tag{6.4}$$

where $\mathrm{MC}_{a\ \mathrm{nom}}(x)$ and $\mathrm{MC}_{a\ \mathrm{alt}}(x)$ are the distributions of the kinematic variable x in the nominal MC simulation and alternative MC simulation of the background process to be reweighted, respectively.

Table 6.7 summarizes the sources of modeling uncertainties, including whether they affect only the normalization (N), or both the normalization and the shape (SN) of the m_{eff} distribution, as well as the number of uncorrelated components.

Experimental uncertainty	Type	Components
Luminosity	N	1
Electron trigger+reco+ID+isolation	SN	5
Electron energy scale+resolution	SN	2
Muon trigger+reco+ID+isolation	SN	12
Muon momentum scale+resolution	SN	5
Jet vertex tagger	SN	1
Jet energy scale	SN	30
Jet energy resolution	SN	8
Jet mass scale	SN	1
Jet mass resolution	SN	1
$E_{\mathrm{T}}^{\mathrm{miss}}$ scale and resolution	SN	3
$E_{\mathrm{T}}^{\mathrm{miss}}$ trigger efficiency	N	1
b-tagging efficiency	SN	9
c-tagging efficiency	SN	5
Light-jet tagging efficiency	SN	6
b-tagging extrapolation	SN	2

Table 6.6: List of experimental systematic uncertainties considered. An "N"("S") means that the uncertainty is taken as normalization-only (shape-only) for all processes and channels affected, whereas "SN" means that the uncertainty is taken on both shape and normalization. Some of the systematic uncertainties are split into several components for a more accurate treatment.

Modeling uncertainty	Type	Components
$t\bar{t}$ cross section	N	1
$t\bar{t}+\geq 1b$, $t\bar{t}+\geq 1c$ normalizations	N	2
$t\bar{t}$+light parton shower+hadronization	SN	5
$t\bar{t}$+light NLO generator	SN	5
$t\bar{t}$+light radiation	SN	20
$t\bar{t}+\geq 1c$ parton shower+hadronization	SN	5
$t\bar{t}+\geq 1c$ NLO generator	SN	5
$t\bar{t}+\geq 1c$ radiation	SN	20
$t\bar{t}+\geq 1b$ parton shower+hadronization	SN	5
$t\bar{t}+\geq 1b$ NLO generator	SN	5
$t\bar{t}+\geq 1b$ radiation	SN	20
Single-top cross section	N	1
Single-top parton shower+hadronization	SN	5
Single-top NLO generator	SN	5
Single-top radiation	SN	20
Single-top DR/DS	SN	1
V+jets normalization	N	4
W+jets modeling	S	1
Z+jets modeling	S	1
Diboson normalization	N	8
$t\bar{t}V$ normalization	N	2
$t\bar{t}H$ cross section	N	1
V+jets reweighting	SN	1
$t\bar{t}+Wt$ reweighting	SN	5

Table 6.7: List of modeling systematic uncertainties considered. An "N"("S") means that the uncertainty is taken as normalization-only (shape-only) for all processes and channels affected, whereas "SN" means that the uncertainty is taken on both shape and normalization. Some of the systematic uncertainties are split into several components for a more accurate treatment.

6.1.7 Statistical Analysis

6.1.7.1 Maximum Likelihood Function

For each signal benchmark scenario and mass hypotheses considered in this analysis, the m_{eff} distributions across all the analysis search regions are jointly analyzed to test for the presence of the predicted signal. To perform the statistical analysis a likelihood function $\mathcal{L}(\mu, \theta)$ is constructed as the product of Poisson probability terms over all the m_{eff} bins considered in the analysis. This likelihood function depends on the signal-strength parameter μ, which enters as a multiplicative factor of the predicted production cross section for signal, and θ, which is a set of nuisance parameters that encode the effect of systematic uncertainties in the signal and background expectations. Therefore, the expected total number of events in a given bin depends on μ and θ. The likelihood function can be expressed mathematically as:

$$\mathcal{L}(\mu,\theta) = \prod_{i}^{\text{bins}} \prod_{j}^{\text{regions}} \frac{\left(\mu S_{ij}(\theta) + B_{ij}(\theta)\right)^{n_{ij}}}{n_{ij}!} e^{-\left(\mu S_{ij}(\theta) + B_{ij}(\theta)\right)} \times \prod_{\theta_k}^{\theta} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{\theta_k - \theta_k^0}{\sigma_k}\right)^2} \tag{6.5}$$

The terms in the first two products are the Poisson terms that quantify the probability of observing n_{ij} data events in the m_{eff} bin i of a given analysis fit region j subject to $\mu S_{ij}(\theta) + B_{ij}(\theta)$ expected events, where $S_{ij}(\theta)$ and $B_{ij}(\theta)$ are the predicted number of signal and background events in that bin, respectively. The signal is normalized to the signal production cross section times the decay branching ratio of a given signal benchmark times the integrated luminosity. The terms in the third product are Gaussian prior terms that parametrize the effect of a given systematic uncertainty with a corresponding nuisance parameter θ_k, where k enumerates the systematic uncertainties that are considered in the

analysis. The values that θ_k takes represent variations from the nominal value of the nuisance parameter θ_k^0, which is usually taken as 0, measured in units of one standard deviation σ_k. The value of θ_k will be determined by obtaining the maximum likelihood estimator (MLE) of the nuisance parameter. As a result of this, the Gaussian prior acts as a penalty function by reducing the likelihood if the fitted value of θ_k is moved away, or pulled, from its nominal value θ_k^0.

For a given value of μ, the variations in the nuisance parameters θ allow the expectations for signal and background to change according to the corresponding systematic uncertainties. The fitted values of θ that are obtained correspond to deviations from the nominal expectations that globally provide the best fit to the data. This procedure reduces the impact of the systematic uncertainties on the search sensitivity and improves the background prediction by taking advantage of the highly populated background-dominated regions that are included in the likelihood fit.

6.1.7.2 Hypothesis Testing

The statistical analysis is designed to test the agreement of the observed data with two hypotheses: the background-only hypothesis, $\mu = 0$, where only the physics of the Standard Model is assumed, and the signal-plus-background hypothesis, $\mu > 0$, where the new physics from beyond the Standard Model is also assumed. This is achieved by defining a test statistic, which will depend on the likelihood function, that quantifies the agreement of the observed data with a given hypothesis. In order to define the test statistic, the likelihood function will be fit to the data. The fits are done in two ways: a conditional fit where the value of μ is specified, and an unconditional fit where μ is not specified but obtained from its MLE. A desirable property of the test statistic is that it is a function that maps observed data to a

numerical value that monotonically orders the observations based on how extreme they are under the assumption of a given hypothesis. This will allow us to calculate probabilities of making an observation that is at least as incompatible with the observed data and set limits on the signal-strength parameter based on whether the signal-plus-background hypothesis is rejected for a given value of μ.

6.1.7.3 Profile Likelihood Ratio Test Statistic

The test statistic that is employed for testing the signal-plus-background hypothesis for a given value of $\mu > 0$ is the profile likelihood ratio q_μ:

$$q_\mu = -2 \ln \left(\mathcal{L}(\mu, \hat{\hat{\theta}}_\mu) / \mathcal{L}(\hat{\mu}, \hat{\theta}) \right) \tag{6.6}$$

where $\hat{\mu}$ and $\hat{\theta}$ are the values of the parameters μ and θ that simultaneously maximize the likelihood function $\mathcal{L}(\mu, \theta)$ subject to the constraint $0 \leq \hat{\mu} \leq \mu$, which are obtained through the unconditional fit. The values $\hat{\hat{\theta}}_\mu$ are the values of the nuisance parameters that maximize the likelihood function for a given value of μ, which are obtained through the conditional fit. Effectively, this statistic chooses the value of the signal-strength parameter that best matches the observed data, $\hat{\mu}$, and compares it with the value set by the analyzer. If the data agrees well with the specified value of μ then $\mathcal{L}(\mu, \hat{\hat{\theta}}_\mu)/\mathcal{L}(\hat{\mu}, \hat{\theta})$ tends to 1, and consequently, q_μ tends to 0. On the other hand, if the data strongly disagrees with the specified value of μ, then $\mathcal{L}(\mu, \hat{\hat{\theta}}_\mu)/\mathcal{L}(\hat{\mu}, \hat{\theta})$ tends to 0, and consequently, q_μ tends to larger positive values. To test for a discovery, which amounts to testing the compatibility of the observed data with the background-only hypothesis, a similar test statistic is used as in Equation 6.6 by setting

$\mu = 0$ and leaving $\hat{\mu}$ unconstrained when performing the unconditional fit:

$$q_0 = -2 \ln \left(\mathcal{L}(0, \hat{\hat{\theta}}_0) / \mathcal{L}(\hat{\mu}, \hat{\theta}) \right) \tag{6.7}$$

Both test statistics behave similarly in that they attain values near zero when the observed data agrees well with the hypothesized value of μ and tend to large positive values when in disagreement. The p-value of the hypothesis test, which is the probability of making an observation that is at least as incompatible with the observed data, is given by:

$$p_{\mathrm{H}} = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu \mid \mathrm{H}) dq_\mu \tag{6.8}$$

where $q_{\mu,\mathrm{obs}}$ is the value of the test statistic that is obtained from the observed data and $f(q_\mu \mid \mathrm{H})$ is the probability density function of the test statistic q_μ under the assumption of a hypothesis H. Specifically, the p-value of the background-only hypothesis, p_b, is obtained by integrating the probability density function of q_0 that one would obtain by assuming the background-only hypothesis. On the other hand, the p-value of the signal-plus-background hypothesis for a specific value of $\mu > 0$, p_{s+b}, is obtained by integrating the probability density function of q_μ that one would obtain assuming the signal-plus-background hypothesis, where the number of signal events is scaled by μ.

The probability distributions of the test statistics q_0 and q_μ are often unknown and thus require estimation. This can be done through MC pseudo-experiments in which toy datasets are sampled from the Poisson distributions in Equation 6.5. The sampling is done individually for the background-only hypothesis and the signal-plus-background hypothesis under a specified value of μ. These toy datasets are then used to evaluate the test statistic of

the assumed hypothesis from which they were sampled. A distribution of values of the test statistic is obtained from multiple toy datasets, which is then used to calculate the p-values. However, an asymptotic approximation of the expected distributions of the test statistic probability density functions [90, 91] can also be used to do these computations since the sampling of a large number of toy datasets is often needed in order to obtain reliable test statistic distributions for each value of μ, which can be a time-consuming process. Under this approximation, the values of the test statistics follow Gaussian distributions with different probability density functions, where q_μ tends to lower values when in agreement with the signal-plus-background hypothesis and q_0 tends to higher values when in agreement with the background-only hypothesis. As a consequence of this, the logic of the p-value p_b is flipped, so that the probability of making an observation that is at least as incompatible as $q_{0,\text{obs}}$ is obtained by integrating the probability distribution below this value.

6.1.7.4 The CL_s Method

In traditional statistics, when performing a hypothesis test, a hypothesis is rejected if its p-value is below a certain threshold. In the field of particle physics, it is standard practice to set the rejection threshold to 0.05 for the signal-plus-background hypothesis in an analysis that searches for potential new particles. However, instead of applying this threshold to p_{s+b}, it is applied to the quantity CL_s [92, 93], which is defined as

$$\text{CL}_\text{s} = \frac{p_{s+b}}{1 - p_b} \tag{6.9}$$

where p_b is the p-value of the background-only hypothesis. In the majority of searches for new particles, the predicted number of signal events is very low compared to the expected number

of background events ($S_{ij} << B_{ij}$), to the point where the physics effects of the signal-plus-background hypothesis can be approximated by the background-only hypothesis. For some values of μ, the expected number of events from the signal-plus-background hypothesis Poisson distributions in Equation 6.5 can effectively behave as the background-only hypothesis distributions since $\mu S_{ij} + B_{ij} \approx B_{ij}$. This can lead to probability density functions of the test statistics that can significantly overlap when estimated from toy datasets, as shown in Figure 6.10. A problematic situation that can arise when excluding the signal-plus-background hypothesis based on p_{s+b} alone is when there is an observed deficit of data compared to the expected number of events predicted from the background-only hypothesis. The test statistic tends towards negative values when in agreement with the signal-plus-background hypothesis and towards positive values when in agreement with the background-only hypothesis. Thus, an observed deficit will result in a positive test statistic that is distributed towards the right-hand tail of $f(q \mid s+b)$, which will result in a very small value of p_{s+b}. However, even if q_{obs} agrees more with the background-only hypothesis, a deficit of observed events also has poor compatibility with the background-only hypothesis. If the rejection threshold were to be applied to p_{s+b}, then this could lead to the rejection of the hypothesis that favors new physics based on a statistical test that has negligible sensitivity to the signal-plus-background model. The main motivation for using CL_s to reject the signal-plus-background hypothesis is that by dividing by $1 - p_b$, the rejection threshold is increased when there is an observed deficit of data compared to the background-only model, thereby providing a more conservative rejection criteria for the signal-plus-background hypothesis.

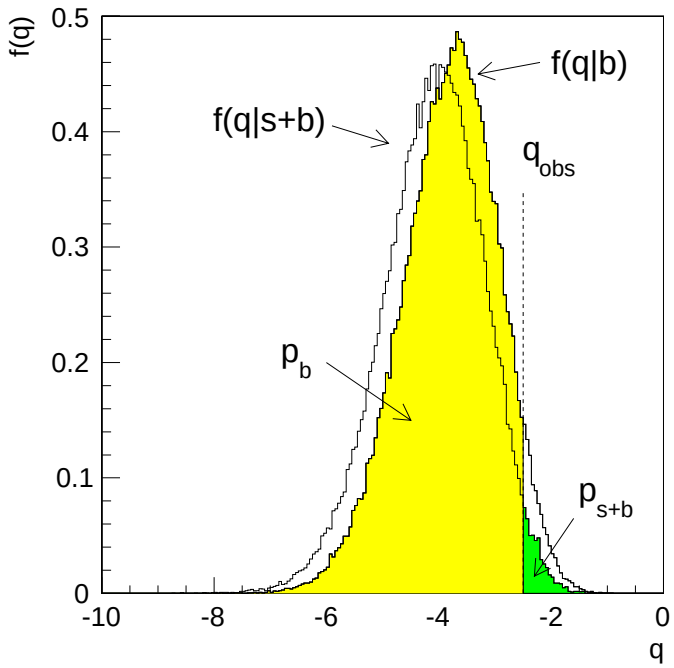


Figure 6.10: Example of the probability distribution of a hypothetical test statistic q that is used for testing a signal-plus-background hypothesis against a background-only hypothesis. The p-value for rejecting the background-only hypothesis, p_b, is obtained by integrating the probability distribution of the test statistic under the background-only hypothesis, $f(q \mid b)$, below the observed value of the statistic, q_{obs}, which is depicted as the yellow region. The p-value for rejecting the signal-plus-background hypothesis is obtained by integrating the probability distribution under the signal-plus-background hypothesis, $f(q \mid s+b)$, above the observed value, which is depicted as the green region. This figure is taken from [94].

6.1.7.5 Limit Calculation

Upper limits on the signal production cross section are computed using q_μ in the CL_s method, where the production cross section is parametrized by μ. Two sets of upper limits are computed: the expected limit, and the observed limit. Since $f(q_0 \mid b)$ is a Gaussian distribution probability density function, the expected limit is defined when $p_b = 0.5$, which in turn defines a value for q_μ, denoted as $q_{\mu,\text{exp}}$. Next, p_{s+b} is determined by setting the lower bound of the integral to $q_{\mu,\text{exp}}$, which will allow us to determine the value of CL_s. If CL_s is different than 0.05, then μ is varied until CL_s is equal to the threshold value of 0.05. The value of μ that achieves $\text{CL}_\text{s} = 0.05$ is the expected upper limit. The observed limit is obtained in the

same way but using the observed value of the test statistic obtained from the data instead of the expected value. This process is repeated for all mass points for a given signal benchmark scenario, with the end result being an upper limit band that depends on the mass of the Vector-Like top quark. The $\pm 1\sigma$ and $\pm 2\sigma$ contour bands of the expected limit are obtained by repeating these steps using the values of $q_{\mu,\text{exp}}$ that correspond to the given standard deviation of $f(q_0 \mid b)$.

6.1.8 Results

6.1.8.1 Maximum Likelihood Fits to Data

A binned likelihood fit as described in subsection 6.1.7 is performed under the background-only hypothesis on the m_{eff} distributions across all analysis fit regions. A comparison between the overall observed and expected yields in each fit region before and after the fit to data is shown in Figure 6.11. As can be observed in the bottom panels of the plots, the combined impact of the systematic uncertainties considered in the analysis has been constrained as a result of the fit, using information from the large number of events in the signal-depleted regions with different background contributions. Consequently, an improved background prediction is obtained with reduced uncertainty across all regions, including those with a significant fraction of expected signal events. This is summarized in Tables 6.8 and 6.9, which contain the number of observed data events, and the pre-fit and post-fit background yields in the four most sensitive analysis fit regions respectively. Furthermore, the pre-fit and post-fit distributions of m_{eff} in these four regions are shown in Figures 6.12 and 6.13 to highlight the overall good post-fit agreement between data and the MC background prediction.

The improved background prediction is verified by checking the agreement between data

and the post-fit background in the analysis validation regions, which are not included in the fit and are designed to be orthogonal from the analysis search regions that are used in the fit. The pre-fit and post-fit comparison of the observed and expected yields in all validation regions is shown in Figure 6.14. Overall the post-fit results in a reduced impact of the systematic uncertainties and an improved background prediction that agrees with data within uncertainties. Furthermore, the pre-fit and post-fit m_{eff} distributions in the corresponding validation regions of the four most sensitive analysis fit regions are shown in Figure 6.15 and Figure 6.16. The general post-fit improvement in the estimated background in the analysis validation regions gives confidence in the background estimation procedure.

	LJ, 2b, ≥1fj, $0t_h$, $\geq 1t_l$, 0H, ≥1V	LJ, ≥4b, ≥1fj, $0t_h$, $\geq 1t_l$, ≥1H, 0V	HJ, 2b, ≥1fj, $\geq 2(t_h+t_l)$, 0H, ≥1V	HJ, ≥4b, ≥1fj, ≥1H, $\geq 2(V+t_l+t_h)$
T singlet ($m_T = 1.6$ TeV, $\kappa = 0.5$)	31.8 ± 4.9	7.2 ± 3.5	1.3 ± 0.4	1.0 ± 0.5
T doublet ($m_T = 1.6$ TeV, $\kappa = 0.5$)	21.8 ± 2.4	8.5 ± 5.6	7.3 ± 2.1	7.1 ± 4.5
$t\bar{t}$+light-jets	1170 ± 210	1.6 ± 2	39.1 ± 9.5	0.49 ± 0.29
$t\bar{t}+\geq 1c$	143 ± 80	1.5 ± 1.3	15.3 ± 9.9	0.86 ± 0.58
$t\bar{t}+\geq 1b$	57 ± 32	4.8 ± 3.9	6.1 ± 4.3	2.6 ± 2
Single-top	250 ± 50	0.66 ± 0.87	7.3 ± 7.5	<0.001
$t\bar{t}W/Z$	13.2 ± 3.1	0.33 ± 0.19	2.5 ± 1.1	0.22 ± 0.82
$t\bar{t}H$	1.5 ± 0.2	0.51 ± 0.15	0.34 ± 0.14	0.42 ± 0.12
W+jets	25.7 ± 9.4	0.70 ± 1.3	1.2 ± 1.1	0.24 ± 0.15
Z+jets	4.4 ± 1.7	<0.001	0.25 ± 0.10	0.007 ± 0.007
Dibosons	3.8 ± 1.4	0.02 ± 0.03	0.21 ± 0.15	<0.001
Multijet	12.9 ± 7.3	0.025 ± 0.017	0.61 ± 0.46	0.16 ± 0.14
Rare backgrounds	2.0 ± 0.3	0.03 ± 0.04	0.25 ± 0.14	0.33 ± 0.06
Total background	1690 ± 280	10.2 ± 4.8	73 ± 20	5.4 ± 2.5
Data	1519	10	64	7

Table 6.8: Predicted and observed yields in four of the most sensitive search regions (depending on the signal scenario) considered. The "rare backgrounds" category includes the VH, tZ and $t\bar{t}t\bar{t}$ backgrounds. The background prediction is shown before the fit to data. Also shown are the signal predictions for different benchmark scenarios considered. The individual systematic uncertainties for the different background processes can be correlated, and do not necessarily add in quadrature to equal the systematic uncertainty in the total background yield. The quoted uncertainties are the sum in quadrature of statistical and systematic uncertainties in the yields.

	LJ, 2b, ≥1fj, $0t_h$, $\geq 1t_l$, 0H, ≥1V	LJ, ≥4b, ≥1fj, $0t_h$, $\geq 1t_l$, ≥1H, 0V	HJ, 2b, ≥1fj, $\geq 2(t_h+t_l)$, 0H, ≥1V	HJ, ≥4b, ≥1fj, ≥1H, $\geq 2(V+t_l+t_h)$
$t\bar{t}$+light-jets	1033 ± 72	0.6 ± 0.8	33.6 ± 4.5	0.57 ± 0.24
$t\bar{t}+\geq 1c$	144 ± 54	1.5 ± 1.0	15.6 ± 5.5	0.82 ± 0.32
$t\bar{t}+\geq 1b$	75 ± 22	8 ± 3	8.2 ± 2.3	3.8 ± 1.1
Single-top	223 ± 55	0.09 ± 0.55	2.3 ± 4.5	<0.001
$t\bar{t}W/Z$	12.1 ± 2.3	0.36 ± 0.18	2.3 ± 0.8	0.62 ± 0.76
$t\bar{t}H$	1.46 ± 0.21	0.51 ± 0.11	0.29 ± 0.08	0.40 ± 0.09
W+jets	26.6 ± 7.1	0.6 ± 1.0	0.8 ± 0.5	0.22 ± 0.13
Z+jets	4.5 ± 1.2	<0.001	0.27 ± 0.08	0.005 ± 0.006
Dibosons	3.4 ± 1.2	0.017 ± 0.029	0.17 ± 0.13	<0.001
Multijet	9.5 ± 5.7	0.018 ± 0.015	0.45 ± 0.41	0.12 ± 0.12
Rare backgrounds	2.0 ± 0.2	0.03 ± 0.03	0.22 ± 0.08	0.31 ± 0.05
Total background	1534 ± 56	12.1 ± 3.5	64 ± 8	6.8 ± 1.5
Data	1519	10	64	7

Table 6.9: Predicted and observed yields in four of the most sensitive search regions (depending on the signal scenario) considered. The "rare backgrounds" category includes the VH, tZ and $t\bar{t}t\bar{t}$ backgrounds. The background prediction is shown after the fit to data under the background-only hypothesis. The individual systematic uncertainties for the different background processes can be correlated, and do not necessarily add in quadrature to equal the systematic uncertainty in the total background yield. The quoted uncertainties are computed after taking into account correlations among nuisance parameters and among processes. The statistical uncertainty is added in quadrature to the systematic uncertainties.

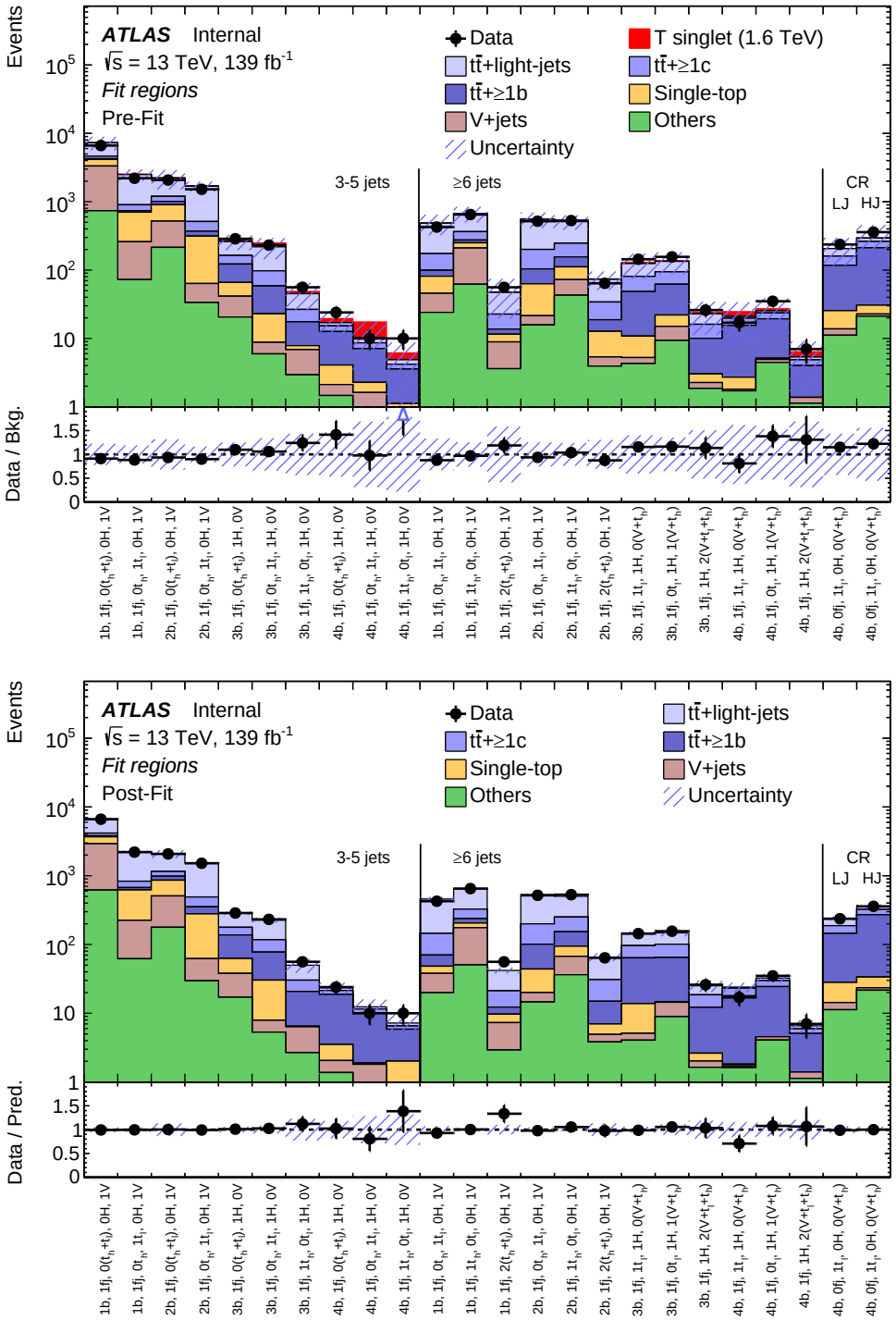
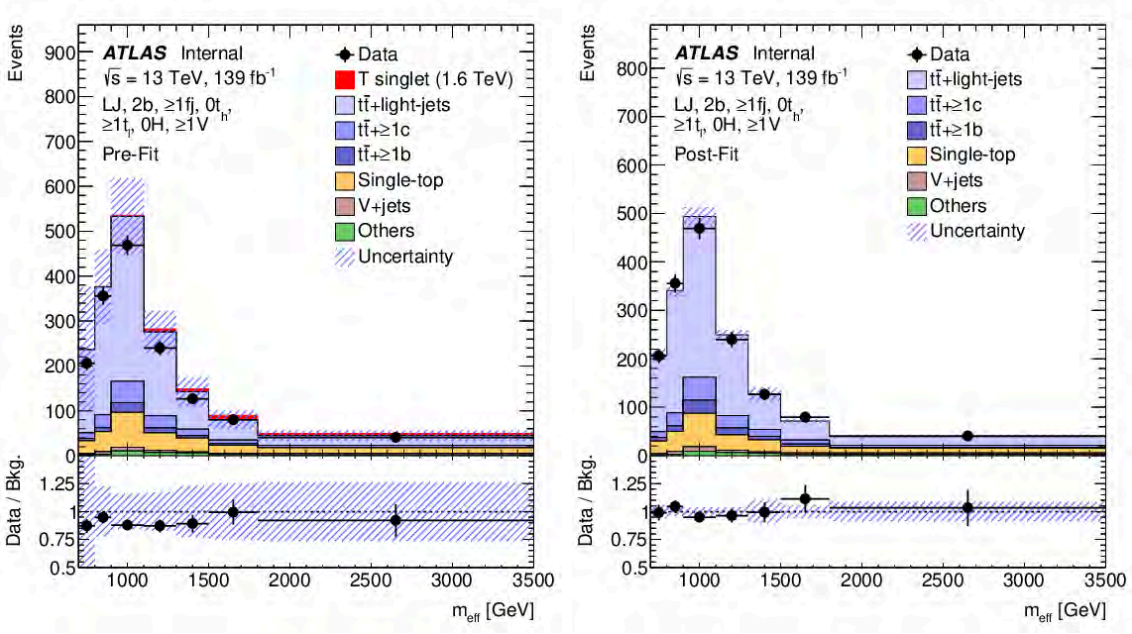


Figure 6.11: Comparison between the data and background prediction for the yields in each of the fit regions considered (top) pre-fit and (bottom) post-fit, performed under the background-only hypothesis. The two right-most regions shown in the plot are the 3–5j (LJ) and ≥6j (HJ) control regions, respectively. The "others" background includes the $t\bar{t}$ V/H, VH, tZ, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The expected T singlet signal (solid red) for $m_T = 1.6$ TeV and $\kappa = 0.5$ is included in the pre-fit figure. The bottom panels display the ratios of data to the total background prediction. The hashed area represents the total uncertainty on the background.



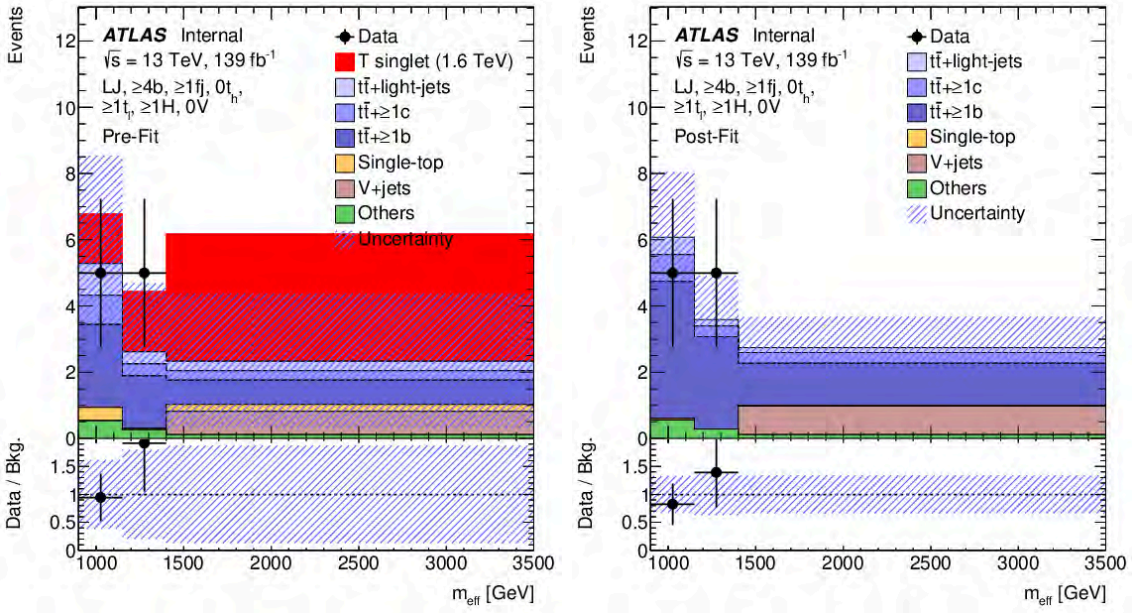


Figure 6.12: Comparison between the data and prediction for the m_{eff} distribution under the background-only hypothesis, in the (LJ, 2b, ≥1fj, 0t$_h$, ≥1t$_l$, 0H, ≥1V) region (a) pre-fit and (b) post-fit, and the (LJ, ≥4b, ≥1fj, 0t$_h$, ≥1t$_l$, ≥1H, 0V) region (c) pre-fit and (d) post-fit. The expected T singlet signal (solid red) for $m_T = 1.6$ TeV and $\kappa = 0.5$ is included in the pre-fit figures. The "others" background includes the $t\bar{t}$ V/H, VH, tZ, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The bottom panels display the ratios of data to the total background prediction. The hashed area represents the total uncertainty on the background. The last bin in each distribution contains the overflow.

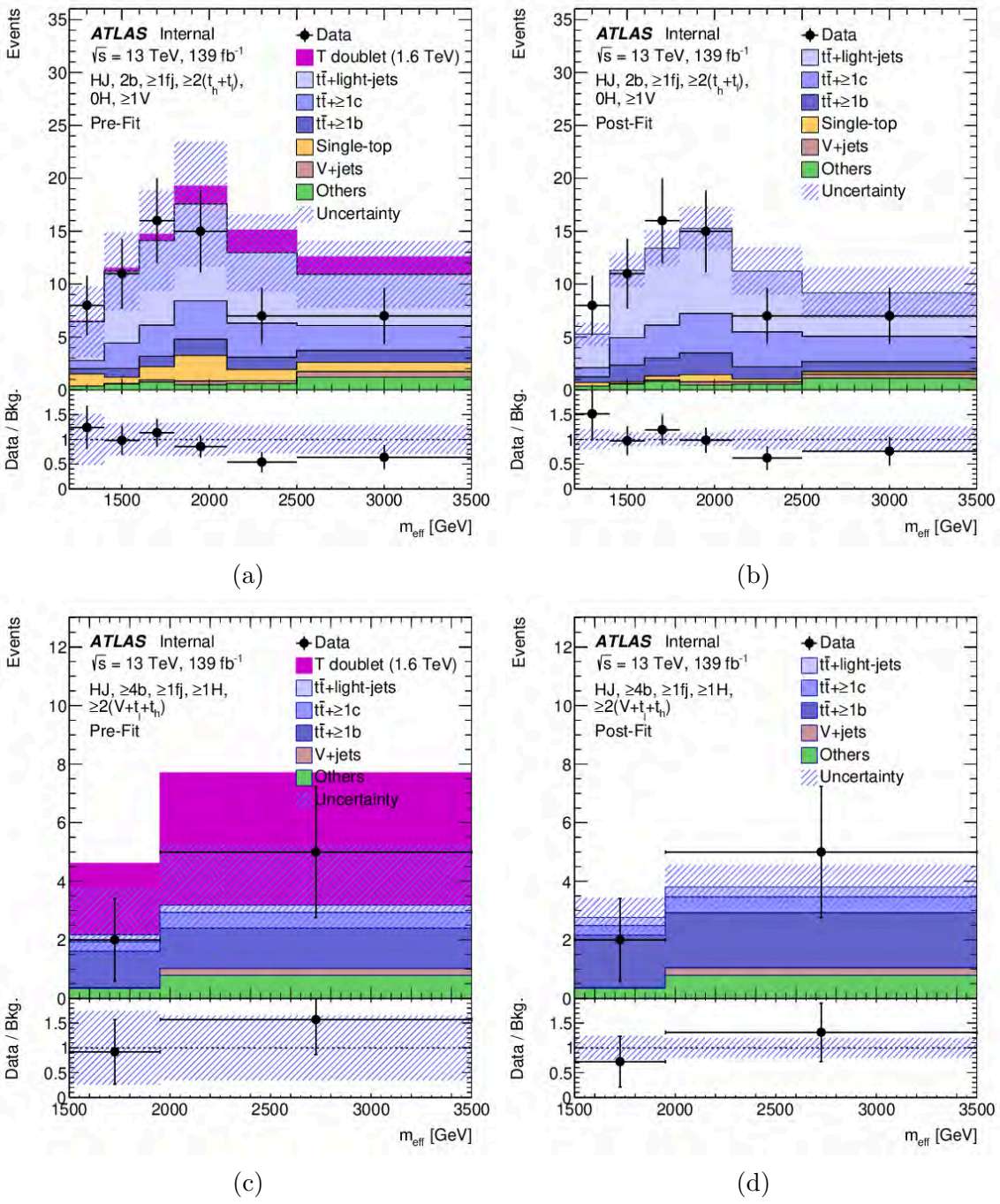
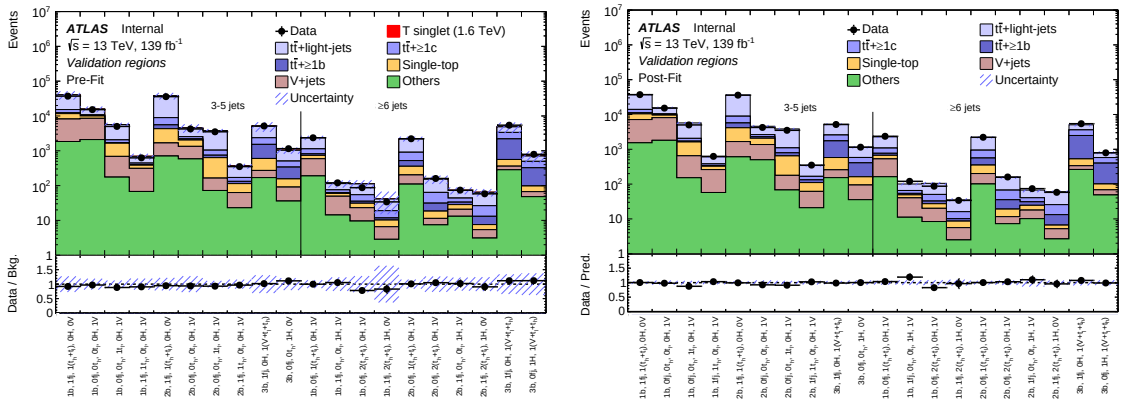
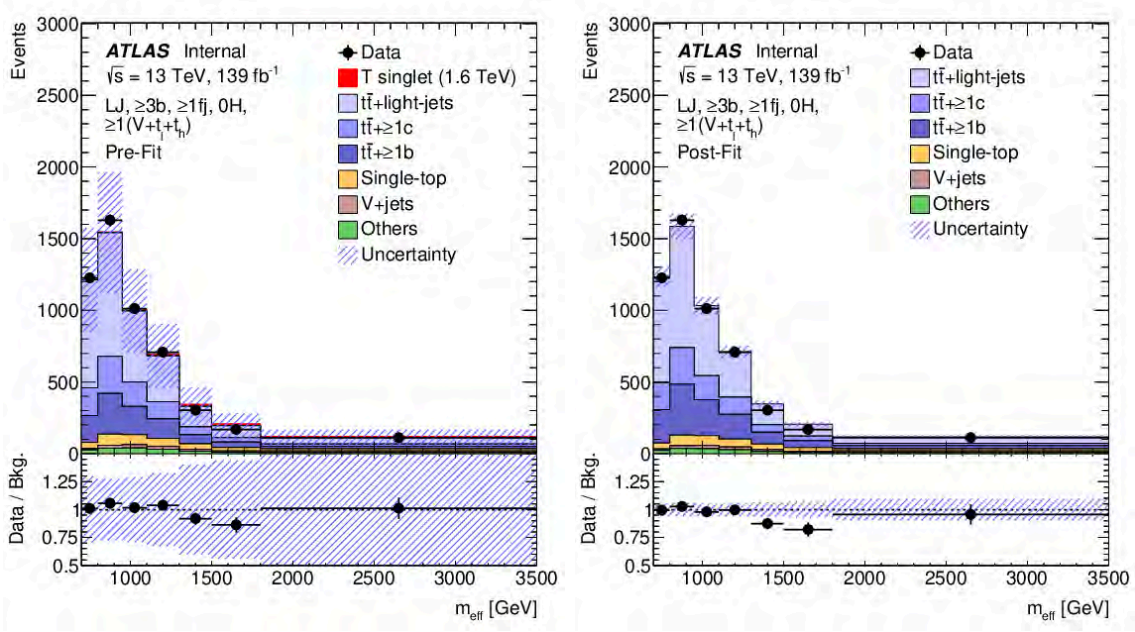


Figure 6.13: Comparison between the data and prediction for the m_{eff} distribution under the background-only hypothesis, in the (HJ, 2b, ≥1fj, ≥2(t_h+t_l), 0H, ≥1V) region (a) pre-fit and (b) post-fit, and the (HJ, ≥4b, ≥1fj, ≥1H, ≥2(V+t_h+t_l)) region (c) pre-fit and (d) post-fit. The expected T doublet signal (solid purple) for $m_T = 1.6$ TeV and $\kappa = 0.5$ is included in the pre-fit figures. The "others" background includes the $t\bar{t}$ V/H, VH, tZ, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The bottom panels display the ratios of data to the total background prediction. The hashed area represents the total uncertainty on the background. The last bin in each distribution contains the overflow.



(a) (b)

Figure 6.14: Comparison between the data and background prediction for the yields in each of the VRs considered (top) pre-fit and (bottom) post-fit, performed under the background-only hypothesis considering only the fit regions. The "others" background includes the $t\bar{t}\,V/H$, VH, tZ, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The expected T singlet signal (solid red) for $m_T = 1.6$ TeV and $\kappa = 0.5$ is included in the pre-fit figure. The bottom panels display the ratios of data to the total background prediction. The hashed area represents the total uncertainty on the background.



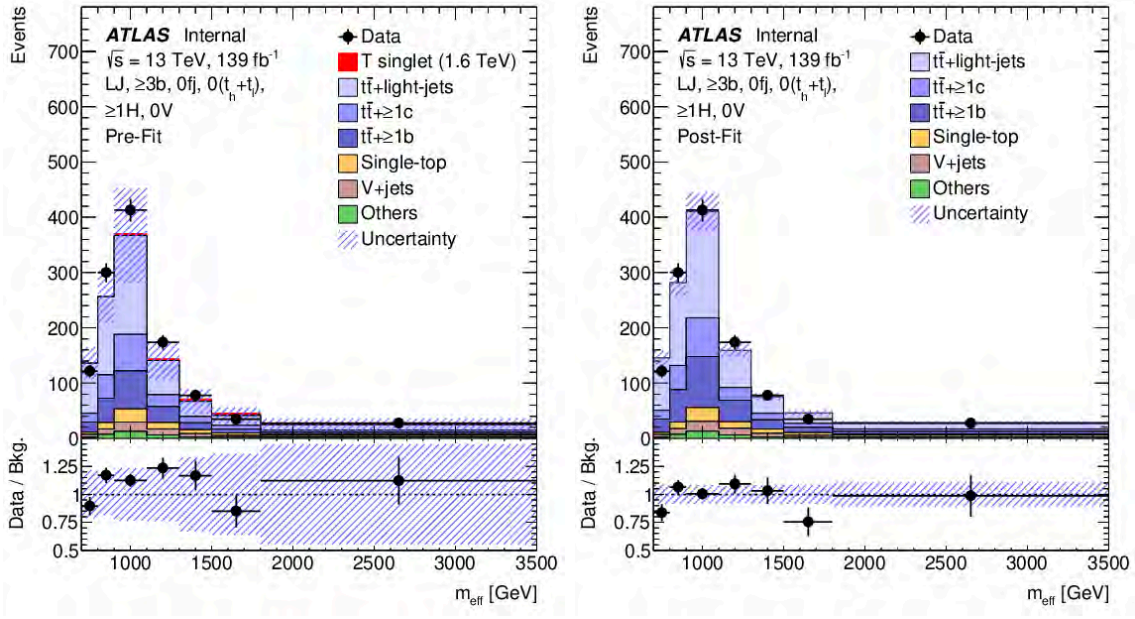
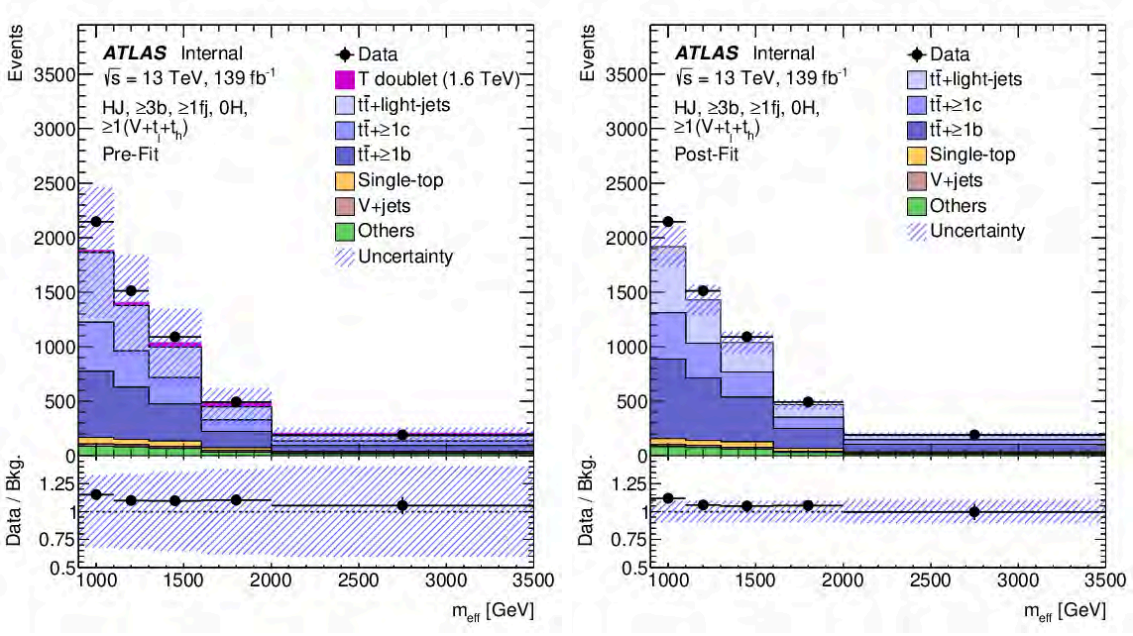


Figure 6.15: Comparison between the data and prediction for the m_{eff} distribution under the background-only hypothesis, in the (LJ, ≥3b, ≥1fj, 0H, ≥1(V+t_l+t_h)) validation region (a) pre-fit and (b) post-fit, and the (LJ, ≥3b, 0fj, 0(t_h+t_l), ≥1H, 0V) validation region (c) pre-fit and (d) post-fit. The expected T singlet signal (solid red) for $m_T = 1.6$ TeV and $\kappa = 0.5$ is included in the pre-fit figures. The "others" background includes the $t\bar{t}\,V/H$, VH, tZ, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The bottom panels display the ratios of data to the total background prediction. The hashed area represents the total uncertainty on the background. The last bin in each distribution contains the overflow.



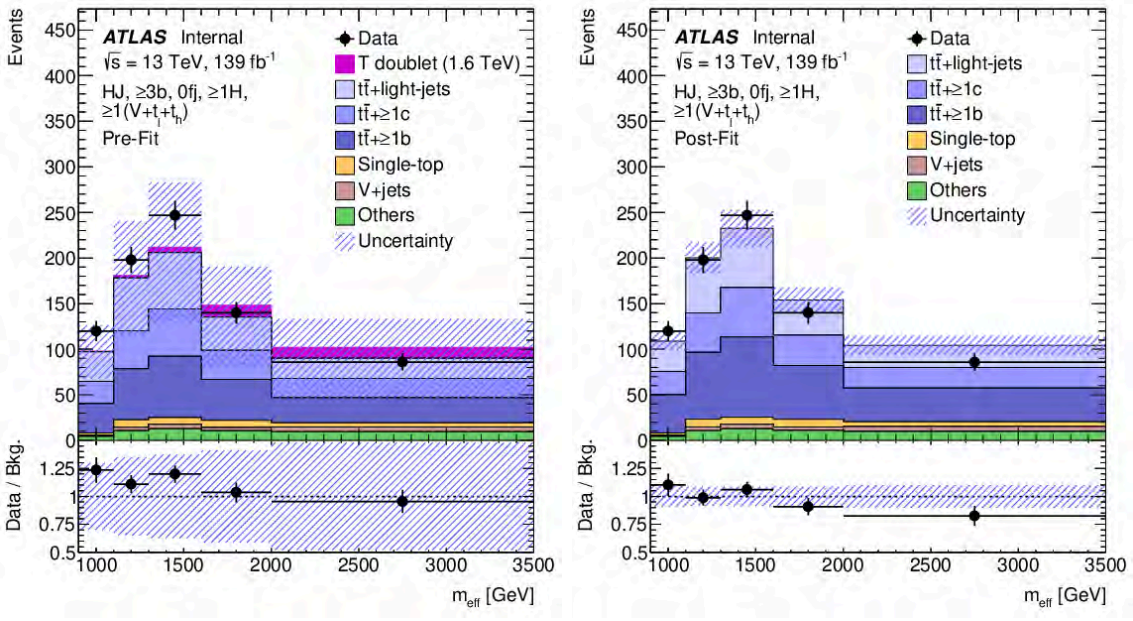


Figure 6.16: Comparison between the data and prediction for the m_{eff} distribution under the background-only hypothesis, in the (HJ, ≥3b, ≥1fj, 0H, ≥1(V+t_l+t_h)) validation region (a) pre-fit and (b) post-fit, and the (HJ, ≥3b, 0fj, ≥1H, ≥1(V+t_l+t_h)) validation region (c) pre-fit and (d) post-fit. The expected T doublet signal (solid purple) for $m_T = 1.6$ TeV and $\kappa = 0.5$ is included in the pre-fit figures. The "others" background includes the $t\bar{t}\,V/H$, VH, tZ, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The bottom panels display the ratios of data to the total background prediction. The hashed area represents the total uncertainty on the background. The last bin in each distribution contains the overflow.

6.1.8.2 Limits on Single Vector-Like Quark Production

As discussed in the previous subsection, no significant excess above the SM prediction is found in any of the considered regions in the background only fit. Furthermore, the unconditional fits with a floating signal-strength parameter μ were also consistent with the background-only hypothesis. Upper limits at the 95% CL on the T production cross section are derived in both the singlet (T) and doublet (TB) scenarios. The observed cross section limits are compared to the NLO theoretical prediction to set exclusion limits on model parameters. The reliability range of the theory cross section calculations with finite width effects and non-resonant contributions is up to a relative T decay width (Γ/M) of approximately 50% [95], thus results are shown only in this restricted regime.

The obtained limits corresponding to the singlet and doublet scenarios are shown in Figure 6.17 and Figure 6.18 respectively for a set of three values of the common coupling parameter κ that are chosen to approximately span the sensitivity range of the search in each scenario. The upper limits are also interpreted as a function of the T mass and coupling strength, which are shown in Figure 6.19 and Figure 6.20 for the singlet and doublet scenarios respectively. All T masses below 2.1 TeV (expected 1.9 TeV) are excluded for the singlet scenario at couplings $\kappa \geq 0.6$. At a mass of 1.6 TeV the coupling strength values above 0.3 (expected 0.41) are all excluded. For the doublet scenario the limits on the considered mass range extend down to coupling values of $\kappa = 0.55$ corresponding to a T quark mass limit of 1.0 TeV. At a coupling strength of $\kappa = 0.75$, masses up to 1.68 TeV are excluded at the threshold of 50% relative T decay width.

The expected limits on the production cross section get progressively stronger at larger T masses in both scenarios, as the decay products of the T become more boosted, and the

fraction of signal in the highest m_{eff} bins increases. The limits deteriorate at larger values of κ since this regime corresponds to large resonance width and a larger fraction of the signal resides in the low mass regime away from the peak of the resonance. The observed limits exceed the expected limits in both signal benchmarks in a few cases, with deviations reaching almost 2σ at high masses for the singlet scenario. This can be ascribed to the downward statistical fluctuations in a few of the most signal sensitive bins such as the last m_{eff} bin of the LJ, $\geq$4b, $\geq$1fj, 0t$_h$, $\geq$1t$_l$, $\geq$1H, 0V region (Figure 6.12d) which has no data events. The origin of these discrepancies has been investigated and no evidence of any systematic bias was found. Furthermore, as previously discussed, the pre-fit and post-fit m_{eff} distributions in the corresponding validation regions exhibit good agreement between data and expectations.

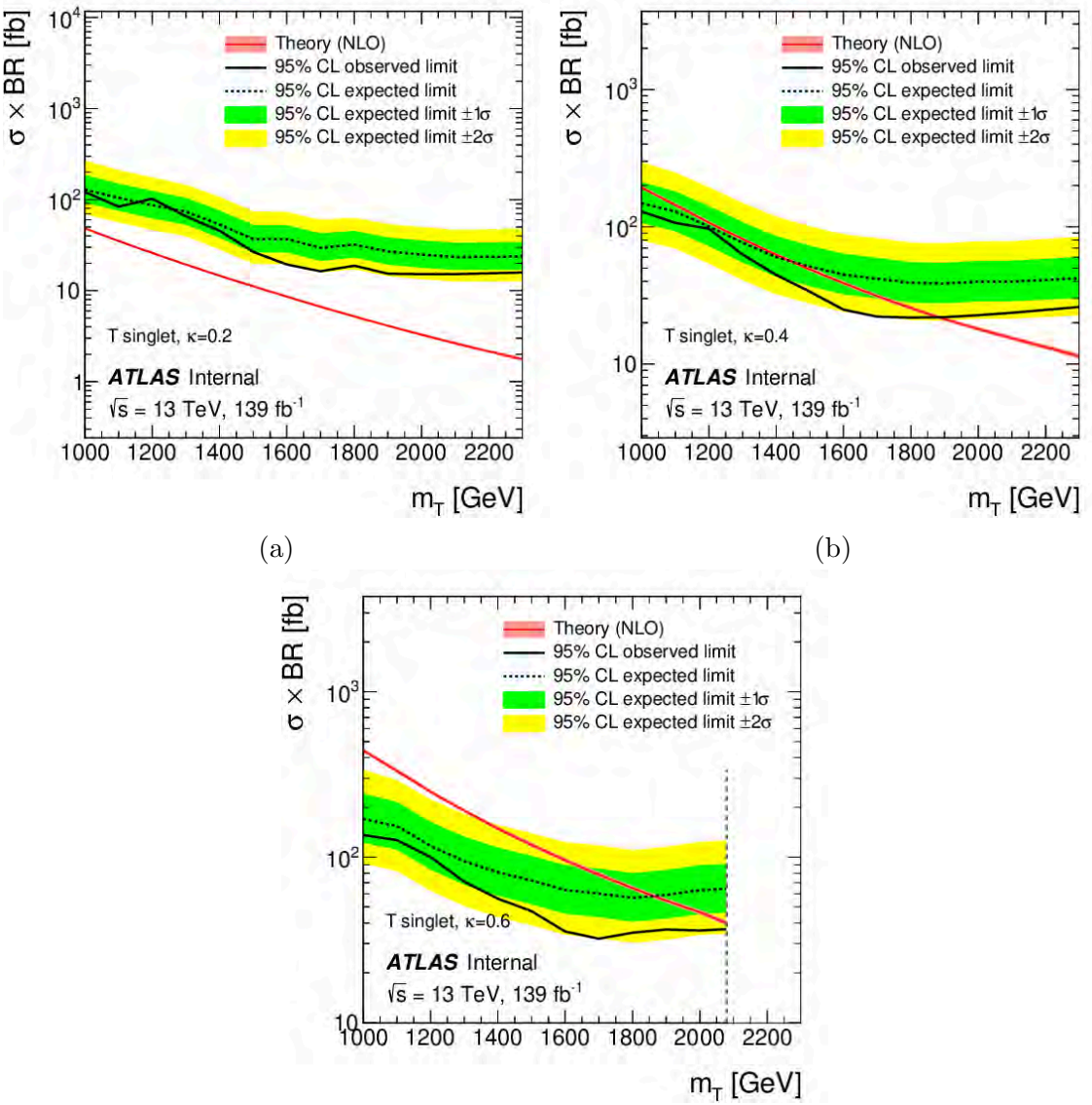


Figure 6.17: Observed (solid line) and expected (dashed line) 95% CL upper limits on the single T production cross-section as a function of the T quark mass in the singlet scenario with the common coupling parameter (a) $\kappa = 0.2$, (b) $\kappa = 0.4$, and (c) $\kappa = 0.6$. The surrounding shaded bands correspond to ± 1 and ± 2 standard deviations around the expected limit. The red line shows the NLO theoretical cross-section prediction, with the surrounding shaded band representing the corresponding uncertainty. Limits are only presented in the regime $\Gamma/M \leq 50\%$, where the theory calculations are known to be valid, as indicated by the vertical gray dashed line.

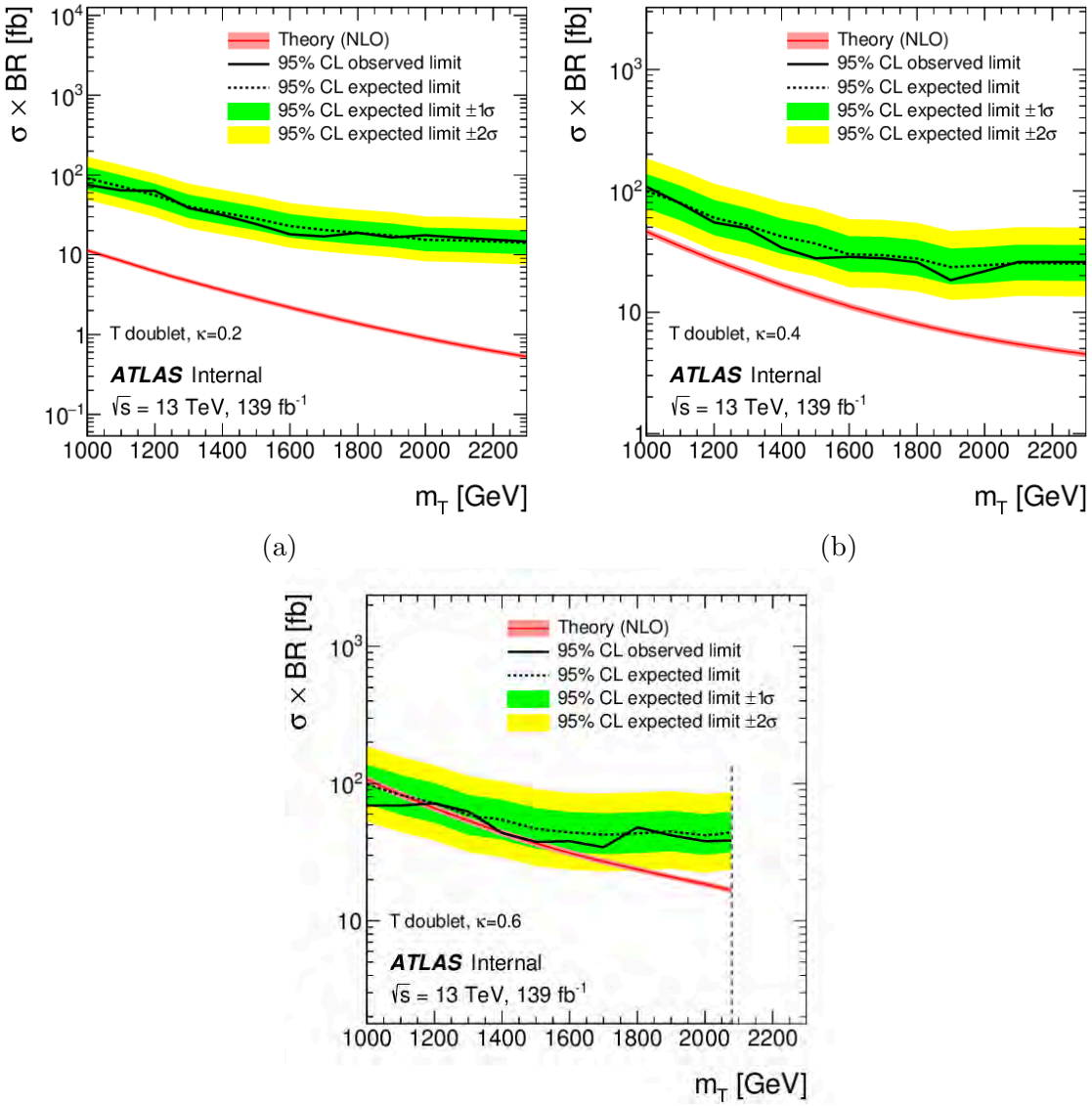
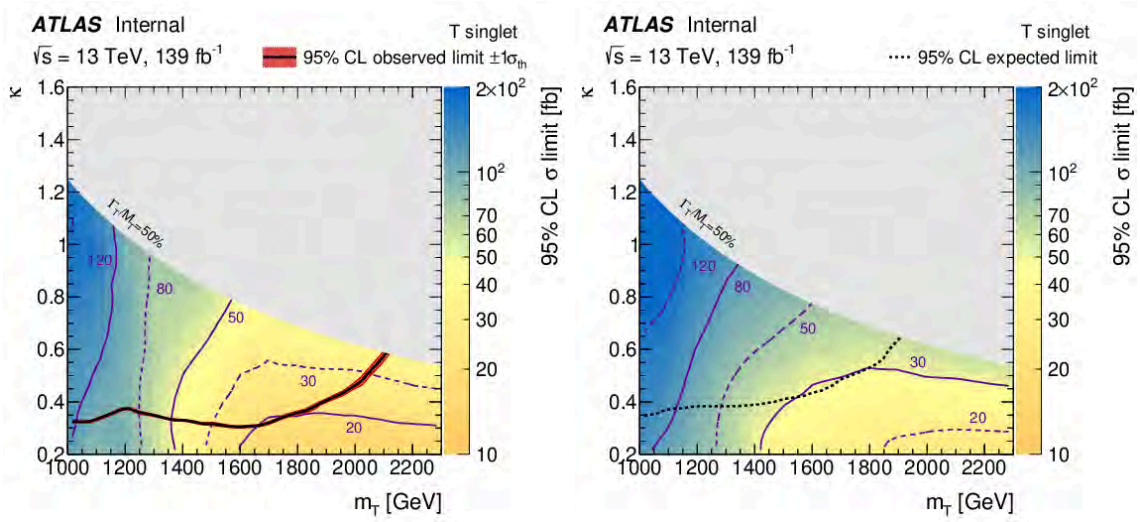
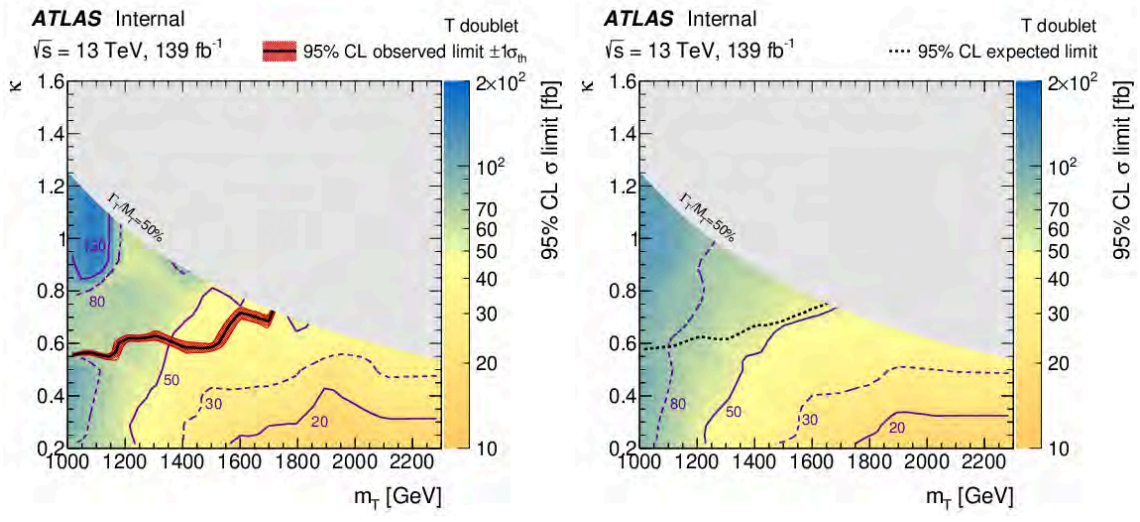


Figure 6.18: Observed (solid line) and expected (dashed line) 95% CL upper limits on the single T production cross-section as a function of the T quark mass in the doublet scenario with the common coupling parameter (a) $\kappa = 0.2$, (b) $\kappa = 0.4$, and (c) $\kappa = 0.6$. The surrounding shaded bands correspond to ± 1 and ± 2 standard deviations around the expected limit. The red line shows the NLO theoretical cross-section prediction, with the surrounding shaded band representing the corresponding uncertainty. Limits are only presented in the regime $\Gamma/M \leq 50\%$, where the theory calculations are known to be valid, as indicated by the vertical gray dashed line.



(a) (b)

Figure 6.19: (a) Observed and (b) expected 95% CL exclusion limits on the cross section times branching ratio of single T quark production as a function of the universal coupling constant κ and the T quark mass in the the SU(2) singlet scenario. The red hashed area around the observed limit corresponds to the theoretical uncertainty on the NLO theoretical cross-section prediction. All values of κ above the black contour line are excluded at each mass point. The purple contour lines denote exclusion limits of equal cross section times branching ratio in units of fb. Limits are only presented in the regime $\Gamma/M \leq 50\%$, where the theory calculations are known to be valid.



(a) (b)

Figure 6.20: (a) Observed and (b) expected 95% CL exclusion limits on the cross section times branching ratio of single T quark production as a function of the universal coupling constant κ and the T quark mass in the the SU(2) doublet scenario. All values of κ above the black contour line are excluded at each mass point. The purple contour lines denote exclusion limits of equal cross section times branching ratio in units of fb. Limits are only presented in the regime $\Gamma/M \leq 50\%$, where the theory calculations are known to be valid.

6.2 Pair Production of Vector-Like Quarks

6.2.1 Analysis Strategy

The pair production analysis is optimized to search for pairly produced Ts where one of the Ts decays to a top quark and either a Higgs or Z boson. The analysis is divided into a 0-lepton and 1-lepton channel, however, at the time of writing this dissertation, the 1-lepton channel is far more developed than the 0-lepton channel, so only the 1-lepton channel will be discussed. Since this analysis closely follows the background model and shares the same signal decay channels as the single production analysis, some elements from the single production analysis strategy carry over to the pair production analysis. In particular, the boosted object tagging and reconstruction and the choice of m_{eff} as the analysis discriminant variable remain the same. The systematic uncertainty model that is used in the pair production analysis also closely follows the model from the single production analysis due to the background model and event kinematics being almost identical. Furthermore, since the simulation of the most relevant background processes is identical between the analyses, the background reweighting procedure designed for the single production analysis is also implemented in this analysis. However, some minor modifications are made to this procedure in order to accommodate the event preselection of the pair production analysis, which will be further elaborated in the following sections. Finally, the results presented for this analysis were obtained using the same statistical analysis methodology as the one used in the single production analysis.

Signal processes in this analysis are categorized based on their combinatorial decay topologies as follows:

1. $HtHt$ for both vector-like tops decaying into Ht

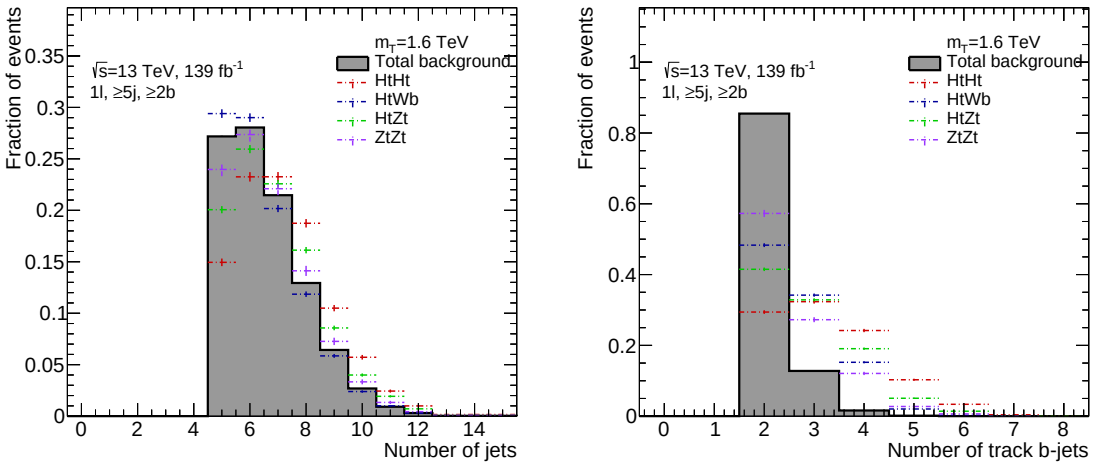
2. $HtWb$ for one vector-like top decaying into Ht and the other into Wb

3. $HtZt$ for one vector-like top decaying into Ht and the other into Zt

4. $ZtZt$ for both vector-like tops decaying into Zt

The 1-lepton channel strategy is optimized for the pair production of Ts with one $T \to Ht$ decay. However, the 1-lepton channel has some sensitivity to $ZtZt$ processes in which a large amount of E_T^{miss} is produced, as will be discussed shortly. The combinatorial nature of the decays of the T provides interesting kinematic features that are taken advantage of in the design of the analysis strategy. The signal processes are characterized by the production of a large number of jets and b-tagged jets, as can be observed in the plots shown in Figure 6.21 at the event preselection level discussed in subsection 4.2.3. The increase in jet multiplicity in signal events is attributed to the dominant decay modes of the decay products of the Ts, such as $H \to b\bar{b}$ decays and hadronically decaying top quarks. Background events that have a large multiplicity of jets are expected to come from the dominant $t\bar{t}$+jets processes. However, these jets mostly originate outside the main $t\bar{t}$ decay topology as final state radiation and are thus not very energetic. The number of b-tagged jets in signal is overall larger compared to the total SM background, with the distribution being shifted towards higher multiplicities for signal processes with a $T \to Ht$ decay. Another kinematic feature that characterizes signal processes is the production of a large amount of E_T^{miss}, as can be observed in the plot shown in Figure 6.22. The source of E_T^{miss} in signal processes is mostly expected to come from $Z \to \nu\nu$ decays or a leptonically decaying top quark that are boosted as a consequence of the large mass of the T.

Due to both Ts being produced with a large mass and low p_{T}, their decay products often emerge as boosted objects that are back-to-back. As a result of this, the number of jets



(a) (b)

Figure 6.21: The distributions of the multiplicities of jets (a) and b-tagged jets (b) at the 1-lepton channel preselection level overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

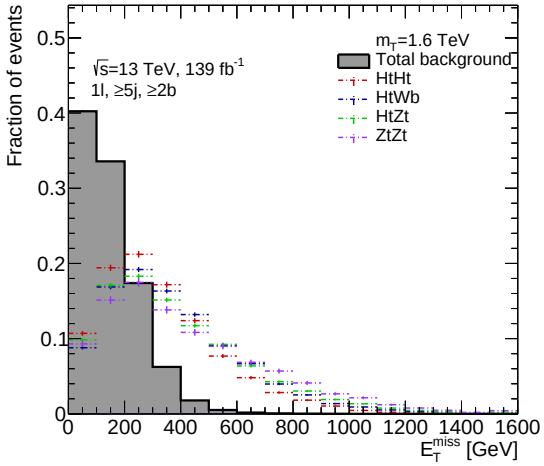


Figure 6.22: The distribution of $E_{\mathrm{T}}^{\mathrm{miss}}$ at the 1-lepton channel preselection level overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

that are tagged to hadronically decaying boosted objects increases due to the additional T that is produced in signal processes, as can be observed in the plot shown in Figure 6.23. The distributions of the number of top-tagged jets, Higgs-tagged jets, W/Z-tagged jets, and reconstructed leptonic tops are shown in Figure 6.24. The tagging of RC jets and the reconstruction of the leptonic top is performed as discussed in subsection 6.1.3. The fraction of signal events in the different bins of the distributions are expected for the signal decay topologies that are considered. Similar to the single production analysis, the lepton that is produced in the 1-lepton channel is expected to originate from a leptonically decaying top quark in signal processes. The combined presence of these boosted objects allows for the reconstruction of candidate Ts, which will be further elaborated in subsection 6.2.4.

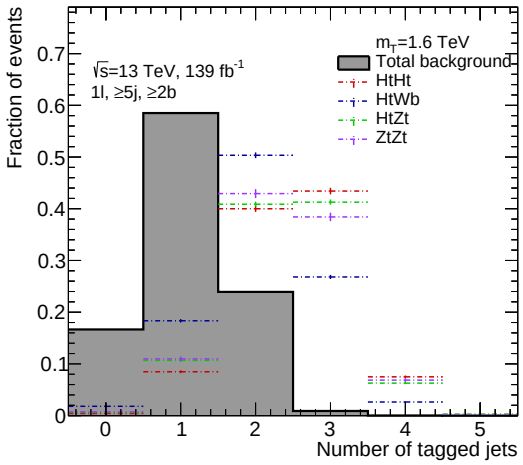
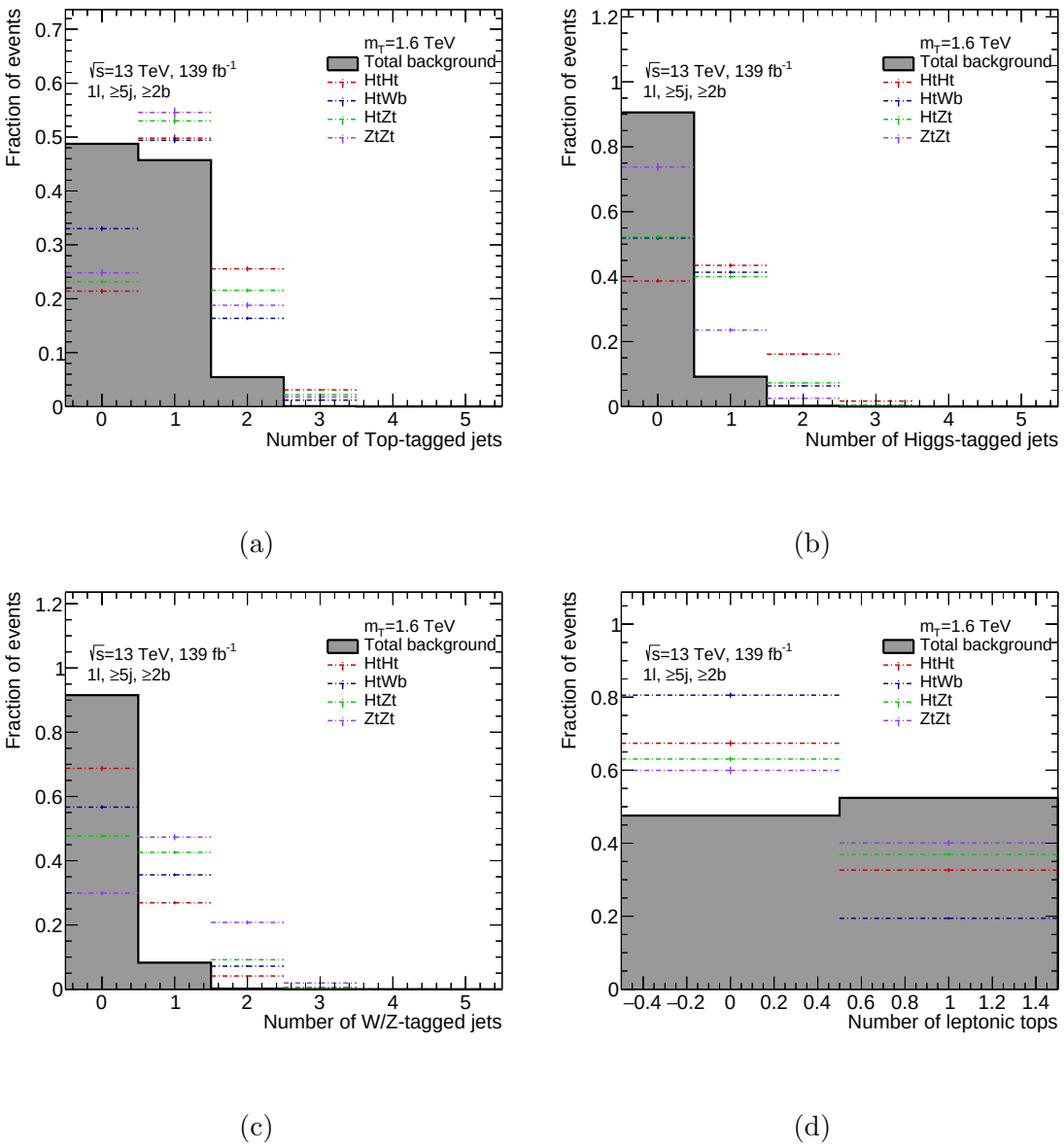


Figure 6.23: The distribution of the number of reclustered large-R jets that are tagged to either a top quark, Higgs boson, or a vector boson at the 1-lepton channel preselection level overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

6.2.2 Signal Discrimination

As previously discussed, the m_{eff} variable is used as the analysis discriminant as it is done in the single production analysis. The use of m_{eff} in the pair production analysis is further

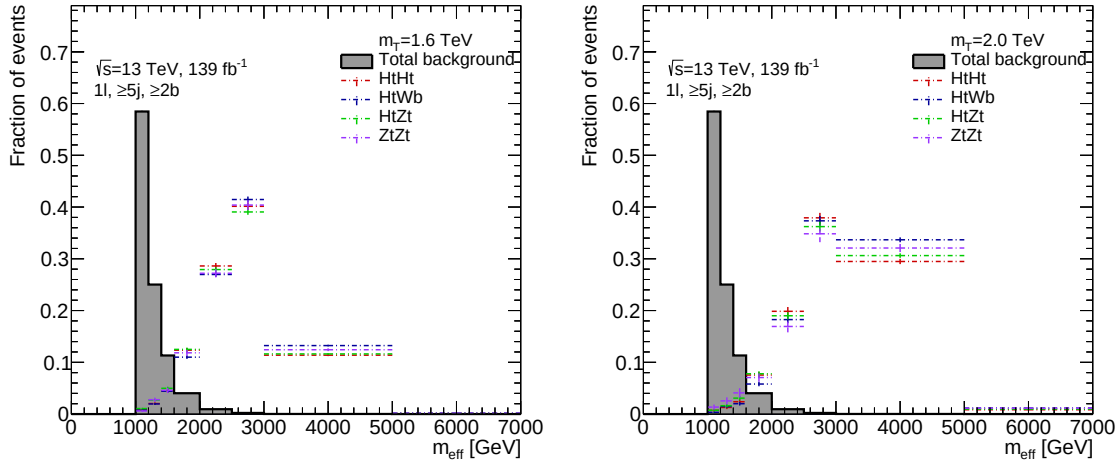


(a) (b)

(c) (d)

Figure 6.24: The distributions of the number of reclustered large-R jets that are tagged to a hadronically decaying top quark (a), Higgs boson (b) and vector boson (c), and the number of reconstructed leptonically decaying top quark (d). The distributions are shown at the 1-lepton channel preselection level and overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

motivated by the additional T that is produced in signal processes. First, the number of final state objects in signal processes increases due to the additional T. Second, both Ts are produced with low p_{T}, but due to their large mass their decay products emerge as boosted objects, which results in energetic final states. This results in an increased separation power between signal and background processes when compared to the single production analysis. As can be observed in Figure 6.25, the m_{eff} distribution peaks close to twice the mass of the T.



(a) (b)

Figure 6.25: The distributions of m_{eff} at the 1-lepton channel preselection level overlayed between the different signal processes for a T mass of 1.6 TeV (a) and of 2.0 TeV (b), and the SM background. An additional cut of $m_{\mathrm{eff}} > 1.0$ TeV is included to highlight the separation between signal and background in a region that is close to the analysis search regions.

6.2.3 Kinematic Reweighting of Background

Both the single production and pair production analyses follow a similar background modeling. The $t\bar{t}$+jets background is the main irreducible background in both analyses, followed by subdominant contributions from single-top and V+jets production processes. The MC

simulation samples that are used to model these background processes in the pair production analysis were generated using the same MC generators as the samples used in the single production analysis; therefore, the same MC mismodeling that was discussed in subsection 6.1.5 is present in the pair production anaylsis. It should be noted that the version of the SHERPA MC generator that is used to model the V+jets background processes was updated from v2.2.1 to v2.2.11. This results in an overall improvement in the modeling of V+jets, however, these backgrounds are still mismodeled in the kinematic regime that the signal is expected to reside in, albeit to a lesser degree when compared to the single production analysis. The distributions of m_{eff} and N_{jets} overlayed between the total SM background simulation prior to applying any correction factors and data are shown in Figure 6.26. Both distributions are shown in the 1l, $\geq$ 5j, 2b region, which is background-dominated. As can be observed, the MC prediction overestimates the m_{eff} distribution at high values and underestimates the N_{jets} distribution at high jet multiplicities, which is where the signal is expected to reside.

A background reweighting procedure is applied in the 1-lepton channel of this analysis in order to improve the MC simulation modeling. The implementation of the reweighting procedure closely follows the strategy outlined in subsection 6.1.5, with a few modifications. In order to accommodate the event preselection of the pair production analysis, the RSRs for both $t\bar{t} + Wt$ and V+jets are defined starting at $\geq$ 5j instead of $\geq$ 3j. Additionally, for the $t\bar{t} + Wt$ reweighting, the 1l, $\geq$ 7j, 2b RSR that was used in the single production analysis is now split into the RSRs 1l, 7j, 2b and 1l, $\geq$ 8j, 2b. This choice is motivated by the presence of sufficient background statistics in the resulting RSRs, thereby allowing the derivation of correction factors that are more reliable in the $\geq$ 7j region. As a result of the improved V+jets modeling from SHERPA v2.2.11, it was determined that only N_{jets}

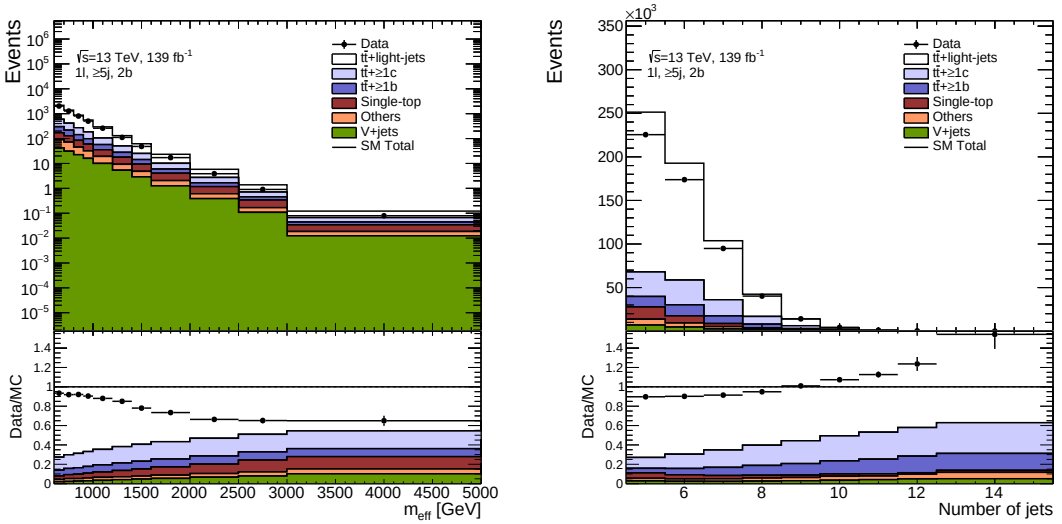


Figure 6.26: The distributions of m_{eff} (a) and N_{jets} (b) overlayed between data and the total SM background simulation prior to applying any correction factors. The "Others" background includes the $t\bar{t}V/H$, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The distributions are shown in the 1l, ≥5j, 2b region. All bins are weighted by the bin width.

correction factors were sufficient to address the remaining mismodeling of these processes. The N_{jets} correction factors for V+jets are applied as a bin-by-bin jet multiplicity correction factor prior to the derivation of the $t\bar{t}+Wt$ correction factors. For the $t\bar{t}+Wt$ reweighting, the $m_{\text{eff}}^{\text{red}}$ is redefined in order to better capture the behavior of the additional jets from $t\bar{t}$ processes. Thus, instead of a $N_{\text{jets}}-3$ shift, the new definition of $m_{\text{eff}}^{\text{red}}$ uses a $N_{\text{jets}}-5$ shift. In order to determine the constant p_{T} scale that multiplies the N_{jets} shift, the average jet p_{T} was calculated as a function of the number of additional jets at the event preselection, as shown in Figure 6.27. Three linear fits were performed to determine the p_{T} scale from the slope of the line. The fits differ in the range of N_{jets} that is considered. As can be observed, the average jet p_{T} is slightly higher in events with at least 10 jets. However, as previously discussed, these jets originate mostly from outside the main $t\bar{t}$ decay topology. The choice of the p_{T} scale is obtained from the linear trend observed for $N_{\text{jets}} \leq 9$. Based from this

discussion, the new definition of $m_{\text{eff}}^{\text{red}}$ is given by:

$$m_{\text{eff}}^{\text{red}} = m_{\text{eff}} - (N_{\text{jets}} - 5) \times 40 \text{ GeV} \tag{6.10}$$

The $m_{\text{eff}}^{\text{red}}$ background correction factors are fitted using the same sigmoid functional template that was used in the single production analysis. The distributions of m_{eff} and N_{jets} overlayed between the total SM background simulation after applying the background correction factors and data are shown in Figure 6.28. As can be observed, the modeling of MC simulation is significantly improved. The background MC simulation is reweighted throughout the remaining discussion of this analysis.

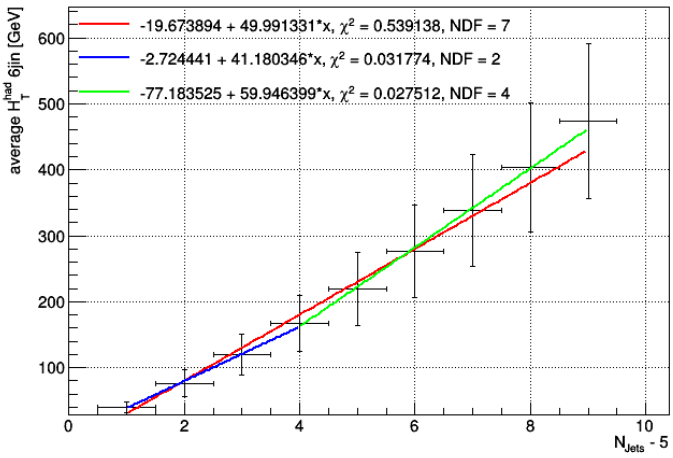
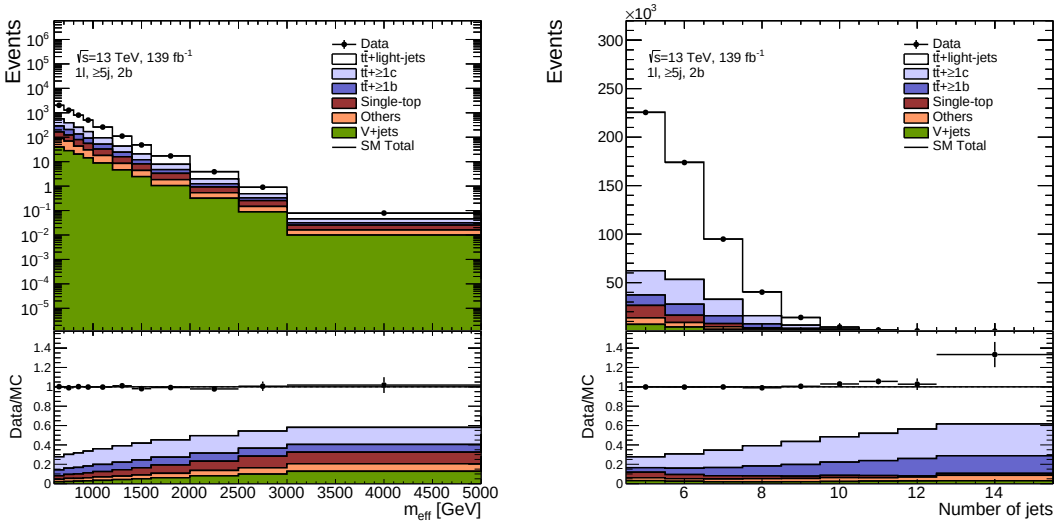


Figure 6.27: Plot of the average p_{T} of additional jets at preselection level (average H_T^{had}) as a function of the number of additional jets at preselection level. The colored lines show the linear fits that were performed to determine constant p_{T} scale in the $m_{\text{eff}}^{\text{red}}$ definition. The red line shows the fit obtained by including all values of the average H_T^{had}, the blue line shows the fit obtained by including the values of the average H_T^{had} up to $N_{\text{jets}} = 9$, and the green line shows the fit obtained by including the values of the average H_T^{had} for $N_{\text{jets}} \leq 9$.



(a) (b)

Figure 6.28: The distributions of m_{eff} (a) and N_{jets} (b) overlayed between data and the total SM background simulation after applying all background correction factors. The "Others" background includes the $t\bar{t}V/H$, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The distributions are shown in the 1l, $\geq$5j, 2b region. All bins are weighted by the bin width.

6.2.4 VLQ Reconstruction

The reconstruction of candidate Ts in events from signal processes is possible due to the combined presence of boosted objects that are identified as the direct decay products of the Ts. As can be observed in Figure 6.23, a large fraction of signal events have at least two RC jets that are tagged to a hadronically decaying boosted object, which allows the possible reconstruction of at least one T. A dedicated algorithm is implemented to reconstruct candidate Ts using the identified boosted objects in the event. The algorithm works under the assumption that the decay topology of each T is resolved due to their low p_{T}, and therefore the two boosted objects that emerge from each T are approximately back-to-back. First, all the identified boosted objects in the event are grouped into pairs based on the possible decays that each T might have had in the event. Next, all pairs of boosted objects are sorted by descending ΔR distance, in accordance with the resolved decay topology assumption that

was made for each T. Finally, the leading and subleading sorted pairs of boosted objects are used to reconstruct the candidate Ts in the event by adding the four-momenta of the two boosted objects in each pair. The distribution of the invariant mass of the leading and subleading pairs are shown in Figure 6.29. As can be observed, the invariant mass in signal processes has a narrow peak close to the corresponding mass value of the Ts, which gives confidence in the reconstruction algorithm. For background processes, the invariant mass distributions peak at lower values and have a large tail at higher mass values, which is expected from trying to reconstruct Ts from random background processes.

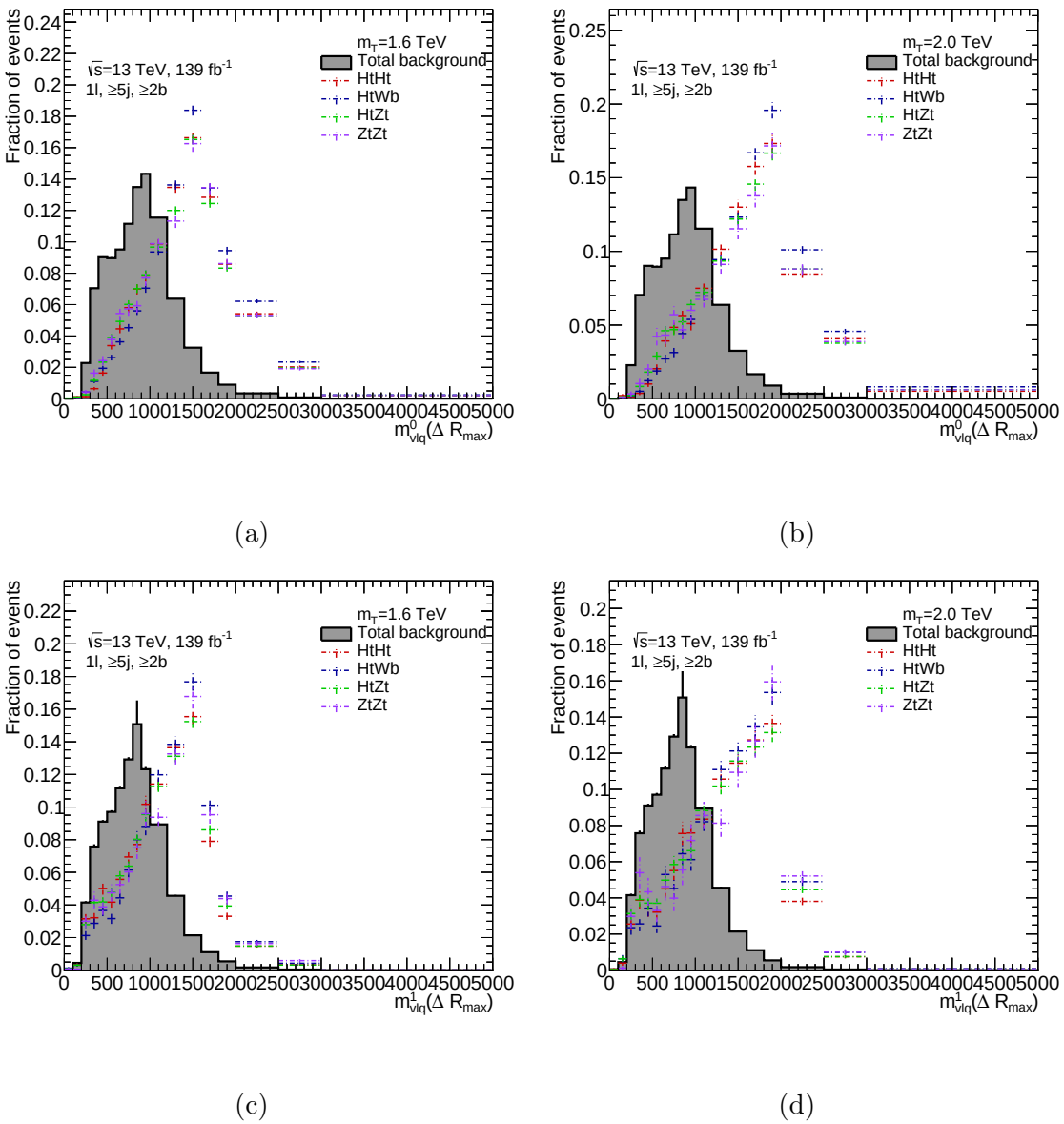


Figure 6.29: The invariant mass distributions of the reconstructed candidate Ts at the 1-lepton channel preselection level overlayed between the different signal processes and the SM background. The distributions are shown for candidate Ts with a mass of $m_{\mathrm{T}} = 1.6$ TeV and $m_{\mathrm{T}} = 2.0$ TeV that are reconstructed from the leading (a)-(b) and subleading (c)-(d) pairs of boosted objects in ΔR distance.

6.2.5 Multivariate Analysis

As previously discussed, kinematic variables such as m_{eff}, the invariant masses of the reconstructed candidate Ts, and the multiplicity of boosted objects in the event contain discriminatory power between signal and background processes. Additionally, other kinematic features that take advantage of the combinatorial nature of the decays of the Ts have been defined. Examples of such variables include the p_{T} of boosted objects in the event, the angular separations of boosted objects in the event, $E_{\text{T}}^{\text{miss}}$ related variables, and the number of subjet constituents within tagged jets. The distributions of some of these variable types are shown in Figure 6.30 in a 1-lepton channel region that requires at least 6 jets, at least 3 b-tagged jets, and at least 3 RC jets, of which at least 2 must be tagged to a boosted hadronically decaying object (1l, $\geq$ 6j, $\geq$ 3b, $\geq$ 2M, $\geq$ 3J). These requirements are made in order to select events from background processes that contain multiplicities of jets and boosted objects that resemble those from events in signal processes at the event preselection level. Additionally, a $m_{\text{eff}} \geq 1$ TeV cut is also applied to this region in order to select background events that are in a kinematic regime where most of the signal is expected to reside.

A multivariate analysis (MVA) was performed in the 1-lepton channel in order to fully exploit the information present in all these variables to classify events as either signal pair production events or SM background events. The MVA consisted in the training of three separate DNN models to perform the event classification task. The DNNs were trained in the region 1l, $\geq$ 6j, $\geq$ 3b, $\geq$ 2M, $\geq$ 3J with the $m_{\text{eff}} \geq 1$ TeV cut applied in order for the DNNs to learn to separate background processes in events that are kinematically similar to signal processes. Furthermore, the DNNs were trained to be agnostic on the decay modes of the Ts and independent of their mass by including all relevant signal decay channels with

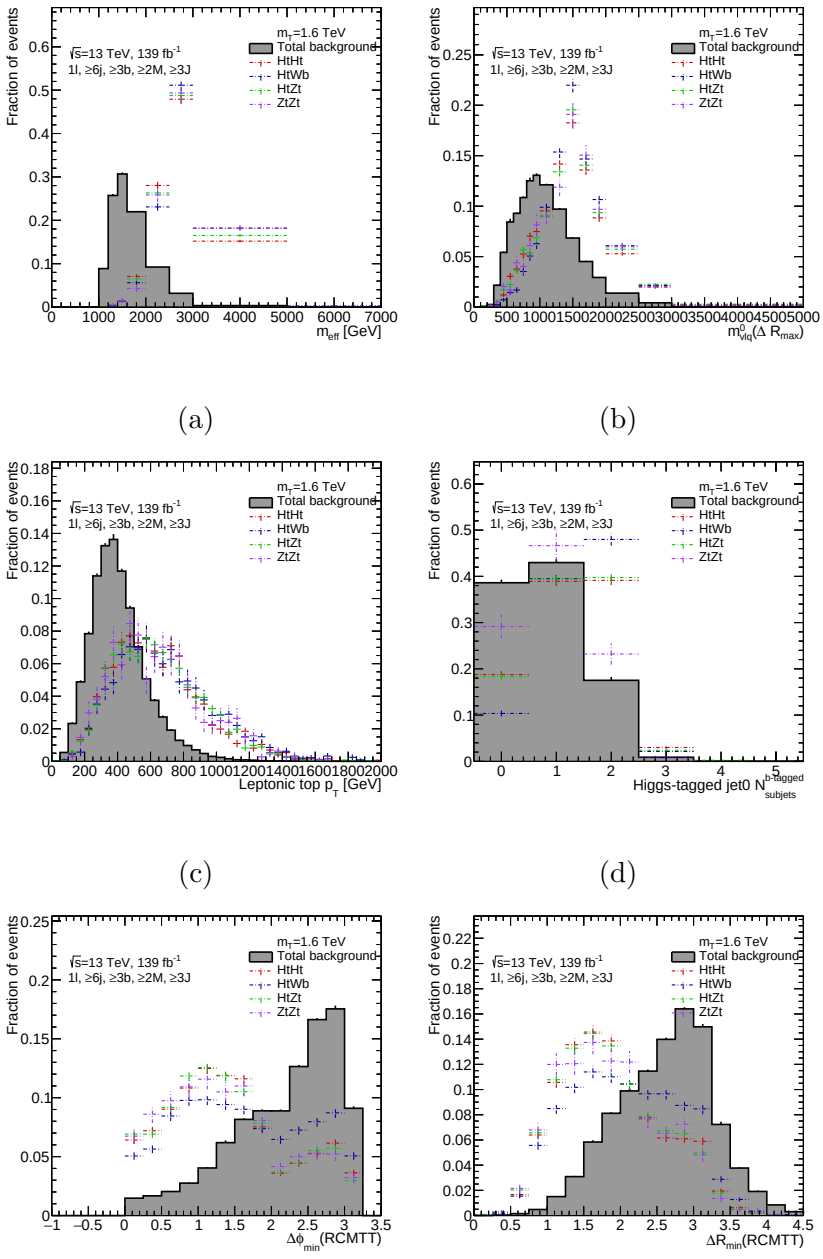


Figure 6.30: The distributions of m_{eff} (a), the reconstructed candidate T invariant mass from the leading ΔR pair of boosted objects (b), the p_{T} of the reconstructed candidate leptonic top (c), the number of b-tagged subjets in the p_{T} leading Higgs-tagged jet (d), the minimum absolute value of $\Delta\phi$ between two tagged jets in the event (e), and the minimum ΔR between two tagged jets in the event (f). The distributions are overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

$m_{\mathrm{T}} = 1.4$ TeV, 1.6 TeV, and 1.8 TeV as part of the training events. Two DNN models were trained with 30 input variables: one with all background processes as part of the training (30 vars., all bkgd.), and the other with $t\bar{t}$ events as the only background process in the training (30 vars., $t\bar{t}$ only). The remaining DNN was trained with 20 input variables and all background processes as part of the training (20 vars., all bgkd.). The list of input variables used by each DNN is summarized in Table 6.10.

Variable	30 vars., all bkgd.	30 vars., $t\bar{t}$	20 vars., all bkgd.
m_{eff}	•	•	•
$m^{b}_{\text{T,min}}$	•	•	•
m^{W}_{T}	•	•	
$E^{\text{miss}}_{\text{T}}$	•	•	•
Residual $E^{\text{miss}}_{\text{T}}$	•	•	
p_{T}(leptonic top)	•	•	•
p_{T}(RCjet2)	•	•	•
p_{T}(RCMHiggs0)	•	•	•
p_{T}(RCMHiggs1)	•	•	
p_{T}(RCMV0)	•		•
p_{T}(RCMTop0)	•	•	
N_{const}(RCMHiggs0)	•	•	•
N_{const}(RCMV0)	•	•	•
N_{bconst}(RCMHiggs0)	•	•	•
N_{bconst}(RCMV0)		•	
$\Delta\phi_{\text{min}}$(RCTTM)	•	•	•
$\Delta\phi_{\text{min}}$(RCjets)	•	•	•
$\Delta\phi_{\text{avg}}$(RCjets)	•	•	•
$\Delta\eta_{\text{min}}$(RCTTM)	•	•	
$\Delta\eta_{\text{min}}$(RCjets)	•	•	
Leading $\Delta\eta$(RCjets)	•	•	
ΔR_{min}(RCMTT)	•	•	•
ΔR_{min}(RCjets)	•	•	•
ΔR_{avg}(RCjets)	•	•	•
Leading ΔR(RCMTT)	•	•	
Leading ΔR(RCjets)	•	•	
m^{0}_{vlq}(RCTTM, ΔR_{max})	•	•	
m^{1}_{vlq}(RCTTM, ΔR_{max})	•	•	•
m^{1}_{vlq}(RCjets, ΔR_{max})	•	•	•
N_{bjets}	•	•	•
N_{RCjets}	•	•	•

Table 6.10: List of input variables of the three DNN models that were trained in the 1-lepton channel MVA. Variables that are ticked indicate that they are used as an input to a given DNN model.

The performance of each DNN is assessed with their receiver operating characteristic (ROC) curve. The ROC curve is a parametric curve that shows the fraction of background events that are correctly identified as background ($1 - \epsilon_{\text{background}}$) as a function of the fraction of signal events that are correctly identified as signal (ϵ_{signal}) at a given DNN score value selection cut. The ROC curve for the three DNN models is shown in Figure 6.31. As can be observed in the plot, the performance between the three DNN models is similar. The DNN that was trained with 30 input variables and with all background processes was chosen as the event classifier due to the robustness in the background model used in its training. Figure 6.32 shows the distribution of the DNN score of this model, which is also referred to as the MVA score. As can be observed in the plot, the DNN achieves good separation power between signal and background processes. Two MVA score working points were optimized based on the background rejection they achieve, which are shown in Figure 6.31. The low working point was set to 0.16, which achieves a background rejection of approximately 75%. The high working point was set to 0.81, which achieves a background rejection of approximately 95%. The signal efficiencies that are achieved were found to be consistent across the different signal processes considered and T mass values. The overall signal efficiency is approximately 95% at the low working point and 77% at the high working point. These working points are used to define the analysis baseline search regions, which will be discussed in the next section.

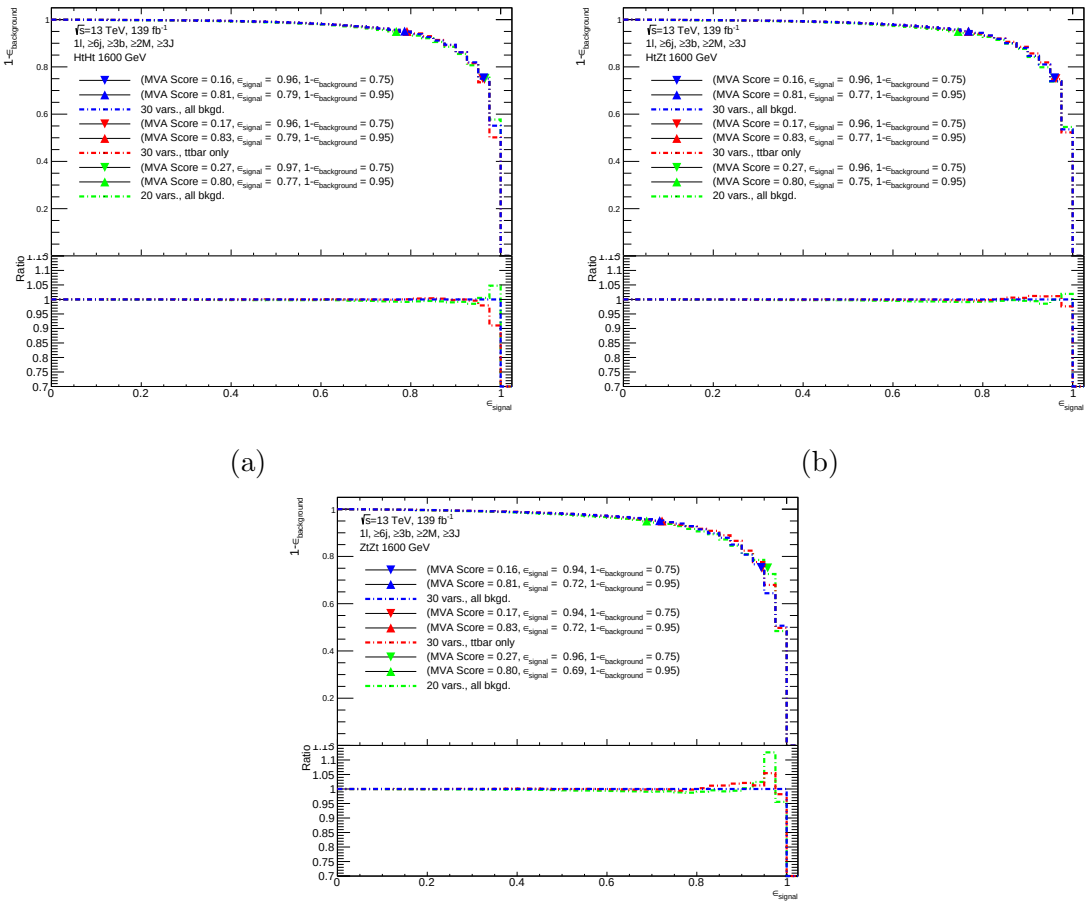


Figure 6.31: The ROC curve of the three DNN models that were trained in the 1-lepton MVA. The horizontal axis shows the signal efficiency evaluated for the signal processes $HtHt$ (a), $HtZt$ (b), and $ZtZt$ (c) for a T mass of 1.6 TeV. The vertical axis shows the background rejection that is achieved at a given value of the signal efficiency. The upwards (downwards) triangle marks indicate points in the ROC curve that achieve a 95% (75%) background rejection. The corresponding signal efficiencies and the MVA score that achieves these values are also displayed. The bottom panel in each plot shows the ratio of the ROC curve value of each model to the 30 vars., all bkgd. model.

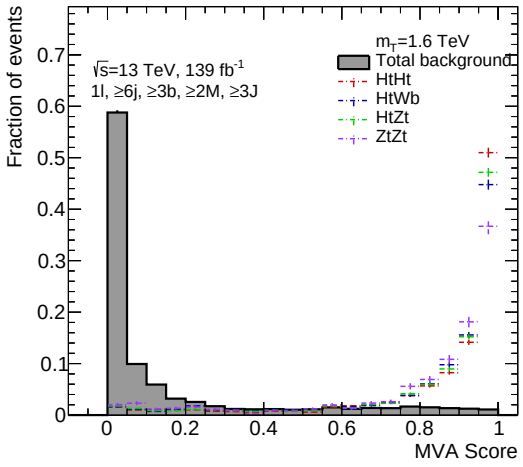


Figure 6.32: The distribution of the MVA score obtained from the DNN that was trained with 30 input variables and all background processes. The distribution is shown for events in the DNN training region 1l, $\geq$ 6j, $\geq$ 3b, $\geq$ 2M, $\geq$ 3J with $m_{\text{eff}} \geq 1$ TeV. The plot is overlayed between the different signal processes for a T mass of 1.6 TeV and the SM background.

1-lepton channel MVA cut	Region name
MVA score ≥ 0.81 (High MVA score)	1l, $\geq$ 5j, $\geq$ 3b, $\geq$ 2M, $\geq$ 3J, HMVA
$0.16 \leq$ MVA score < 0.81 (Mid MVA score)	1l, $\geq$ 5j, $\geq$ 3b, $\geq$ 2M, $\geq$ 3J, MMVA
MVA score < 0.16 (Low MVA score)	1l, $\geq$ 5j, $\geq$ 3b, $\geq$ 2M, $\geq$ 3J, LMVA

Table 6.11: The cuts on the DNN score that are used to define the baseline analysis search regions.

6.2.6 Analysis Search Regions

As discussed in the previous section, the MVA performed in the 1-lepton channel resulted in the development of a DNN that is designed to classify events as either signal T pair production events or SM background events. Two working points on the MVA score were defined based on the background rejection metric. The low working point is defined to achieve a 75% background rejection, which corresponds to an MVA score of 0.16 and an overall signal efficiency of 95%. The high working point is defined to achieve a 95% background rejection, which corresponds to an MVA score of 0.81 and an overall signal efficiency of 77%. These two working points allow for the definition of simpler analysis search regions when compared to the ones defined in the single production analysis. This is due to the training of the DNN being agnostic on the signal processes that are considered in the analysis. Therefore, the purity of signal events is expected to be higher in regions that are defined by requiring events to have an MVA score higher than the high working point. Conversely, regions that are defined by requiring events to have an MVA score lower than the low working point are expected to be background-dominated and can serve as background control regions. These observations motivate the following definitions of the baseline analysis search regions that are listed in Table 6.11. The baseline regions are more inclusive than the DNN training regions by requiring the presence of at least 5 jets instead of 6. This is done in order to increase the signal sensitivity of the analysis. The potential impact that the loosening of this

requirement might have had on the performance of the DNN was assessed and found to be negligible. An intermediate region (1l, $\geq$ 5j, $\geq$ 3b, $\geq$ 2M, $\geq$ 3J, MMVA) is also included as a baseline region in order to retain some sensitivity to signal processes that are not targeted by the 1-lepton channel, such as the $ZtZt$ signal process.

Since the 1-lepton channel of the analysis targets the pair production of Ts with one $T \to Ht$ decay, the purity of the different signal processes in the baseline regions can be further improved by making additional requirements on the multiplicities of b-tagged jets and Higgs-tagged jets. For example, the purity of the $HtHt$ signal process can be increased by requiring at least 4 b-tagged jets and at least one Higgs-tagged jet in the baseline regions. On the other hand, the purity of the $HtZt$ and $HtWb$ can be increased by requiring exactly 3 b-tagged jets or no Higgs-tagged jets. The list of the analysis search regions, also referred to as fit regions, are summarized in Table 6.12. The number of expected events in the $SU(2)$ doublet and singlet signal benchmarks for a T mass of 1.6 TeV and the total background in each fit region are also listed. In addition to the regions that are defined with the MVA score, three regions that are defined only through boosted object multiplicities are also included. These regions serve an identical purpose as the $t\bar{t}$ control regions that were included as part of the single production analysis fit regions. The background composition in each of the fit regions is summarized in the pie charts shown in Figure 6.33. The dominant background in each fit region comes from $t\bar{t}$+jets processes. The regions that have a low b-tagged jet multiplicity requirement are dominated by $t\bar{t}$+light-jets, while regions that have a higher b-tagged jet multiplicity are dominated by $t\bar{t}+ \geq 1b$. The single-top, $t\bar{t}V/H$, and $t\bar{t}t\bar{t}$ processes have subdominant contributions in the HMVA regions; however, the total number of expected background events in these regions is small.

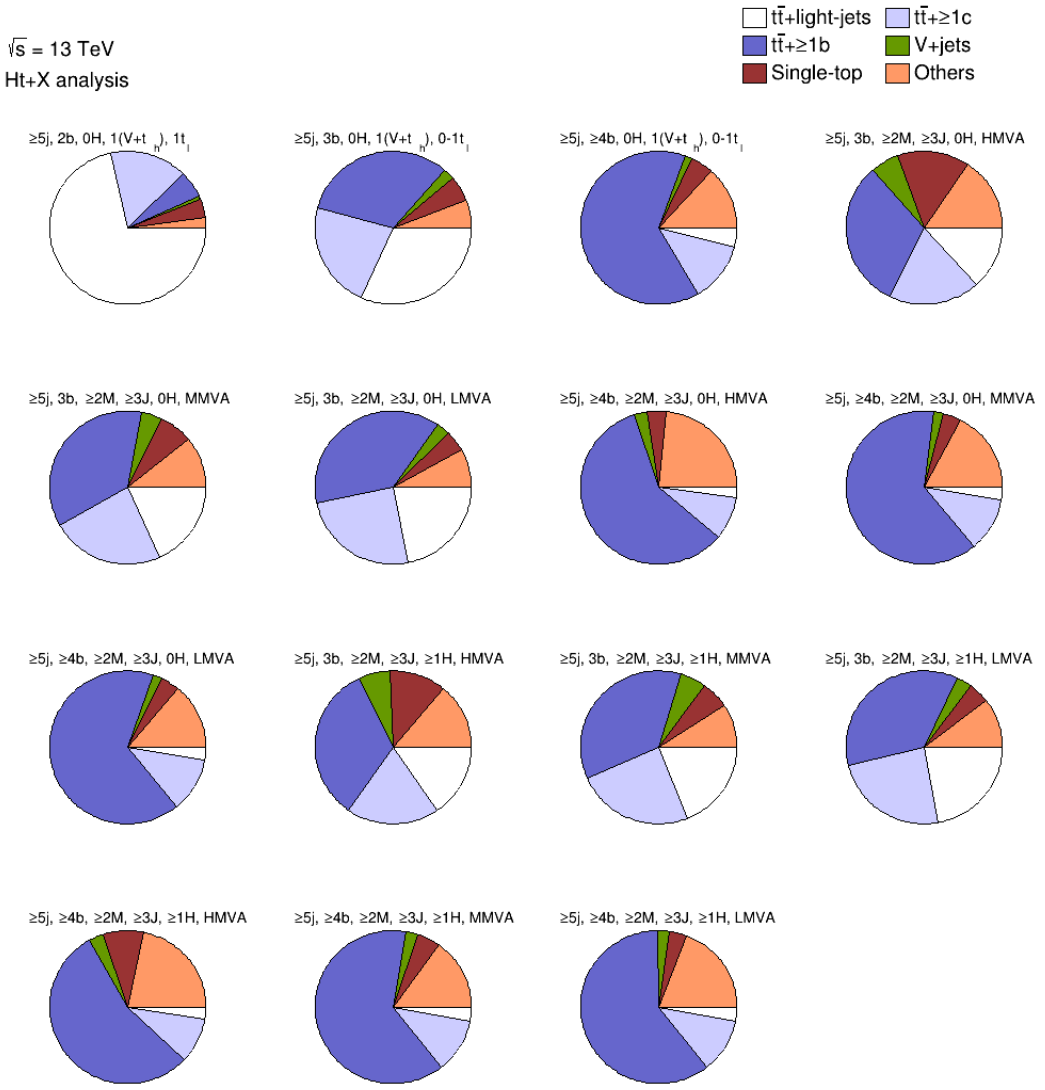


Figure 6.33: The breakdown of the background composition in the 1-lepton channel fit regions.The "Others" background includes the $t\bar{t}V/H$, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds.

Region	$SU(2)$ doublet	$SU(2)$ singlet	Total background
1l, $\geq$ 5j, 3b, $\geq$ 2M, $\geq$ 3J, 0H, HMVA	1.46	1.05	16.3
1l, $\geq$ 5j, 3b, $\geq$ 2M, $\geq$ 3J, 0H, MMVA	0.914	0.633	127
1l, $\geq$ 5j, 3b, $\geq$ 2M, $\geq$ 3J, 0H, LMVA	0.246	0.15	620
1l, $\geq$ 5j, 3b, $\geq$ 2M, $\geq$ 3J, $\geq$ 1H, HMVA	2.67	1.55	20.3
1l, $\geq$ 5j, 3b, $\geq$ 2M, $\geq$ 3J, $\geq$ 1H, MMVA	0.506	0.284	46.3
1l, $\geq$ 5j, 3b, $\geq$ 2M, $\geq$ 3J, $\geq$ 1H, LMVA	0.129	0.0631	142
1l, $\geq$ 5j, $\geq$ 4b, $\geq$ 2M, $\geq$ 3J, 0H, HMVA	1.7	0.846	13.7
1l, $\geq$ 5j, $\geq$ 4b, $\geq$ 2M, $\geq$ 3J, 0H, MMVA	0.402	0.195	36.1
1l, $\geq$ 5j, $\geq$ 4b, $\geq$ 2M, $\geq$ 3J, 0H, LMVA	0.0656	0.031	102
1l, $\geq$ 5j, $\geq$ 4b, $\geq$ 2M, $\geq$ 3J, $\geq$ 1H, HMVA	3.24	1.34	11.1
1l, $\geq$ 5j, $\geq$ 4b, $\geq$ 2M, $\geq$ 3J, $\geq$ 1H, MMVA	0.197	0.0765	12.1
1l, $\geq$ 5j, $\geq$ 4b, $\geq$ 2M, $\geq$ 3J, $\geq$ 1H, LMVA	0.0272	0.0103	18.5
1l, $\geq$ 5j, 2b, 0H, 1(V+t_h), 1t_l	0.554	0.481	2.37e+04
1l, $\geq$ 5j, 3b, 0H, 1(V+t_h), 0–1t_l	0.879	0.828	6.29e+03
1l, $\geq$ 5j, $\geq$ 4b, 0H, 1(V+t_h), 0–1t_l	0.561	0.381	816

Table 6.12: Definition of the 15 analysis search regions (referred to as "fit regions"). The events are categorized based on the multiplicity of central jets (j), b-tagged jets (b), tagged RC jets (M), RC jets (J), Higgs-tagged jets (H), W/Z-tagged jets (V), hadronic top-tagged jets (t_h), reconstructed leptonic tops (t_l), and MVA score. The expected yields of the $SU(2)$ doublet and singlet benchmarks for $m_{\mathrm{T}} = 1.6$ TeV and total background are shown in each fit region.

6.2.7 Systematic Uncertainties

As previously discussed, the pair production analysis follows closely the background model and shares the same signal decay channels with the single production analysis. As a result of this, the sources of systematic uncertainties that are relevant to the pair production analysis are the same as the single production analysis, which were described in subsection 6.1.6. Thus, the systematic uncertainty model that is implemented in the pair production analysis follows closely the model from the single production analysis, with a few modifications which will be detailed in this section.

The only experimental systematic uncertainty that has been modified is the uncertainty in the combined 2015-2018 integrated luminosity. The uncertainty associated with the measurement of luminosity for the ATLAS detector for Run-2 has decreased from 1.7% to 0.83%, which is based on the final measurement made during the Run-2 data taking period [96].

The only modeling uncertainties that have been modified in the pair production analysis are the uncertainties associated with the modeling of the $t\bar{t}$ background process and the uncertainties associated with the background reweighting procedure. The $t\bar{t}$ parton shower and hadronization modeling uncertainty, NLO generator modeling uncertainty, and radiation modeling uncertainty are kept uncorrelated between the $t\bar{t}$+light-jets, $t\bar{t}+ \geq 1c$, and $t\bar{t}+1b$ samples, but treated as correlated amongst all analysis fit regions for each sample. As discussed in subsection 6.2.3, only a bin-by-bin N_{jets} correction factor is applied to the V+jets background processes. The uncertainty assigned to this correction factor is obtained by applying the nominal correction factor shifted by the statistical error of the corresponding N_{jets} bin, both as a positive and negative variation. The $t\bar{t}+Wt$ background reweighting uncertainties are obtained from the $\pm 2\sigma$ variations of the $m_{\text{eff}}^{\text{red}}$ fits that are performed in each

RSRs, similar to how it was done in the single production analysis. An additional nuisance parameter is included for the $t\bar{t} + Wt$ background reweighting procedure that corresponds to the 1l, ≥8j, 2b RSR resulting from the 1l, ≥7j, 2b RSR split.

6.2.8 Results

The results of the search for pair-produced Ts in the 1-lepton channel following the statistical analysis methodology described in subsection 6.1.7 will be presented in this section. As previously mentioned, the analysis is in the early stages of going through the rigorous internal review that all ATLAS analyses must go through in order to establish full confidence in the results obtained. The 0-lepton channel of the analysis is currently at the stage of finalizing validation studies to assess the modeling of the MC prediction to data in order to initialize the 0-lepton channel fit studies. The 1-lepton channel of the analysis has been reviewed up to partially blinded data results in order to assess the fit model behavior in the background-enriched and signal-depleted search regions. This review process has deemed the fit model behavior to be reliable up to this stage, establishing full confidence in the background and uncertainty models of the 1-lepton channel. While the fully unblinded data results have not been internally reviewed, they are not expected to change significantly once they reach that stage. Furthermore, the interpretation of the results given here will be limited to the 1-lepton channel only. These interpretations may change once the results of the 0-lepton channel become available and are combined with the 1-lepton channel results in order to give a broader description in the full phase space of the analysis.

6.2.8.1 Maximum Likelihood Fits to Data

A likelihood fit under the background-only hypothesis is performed on the m_{eff} distributions across the search regions of the 1-lepton channel. A comparison between the overall observed and expected yields in each search region before and after the fit to data is shown in Figure 6.34. As can be observed in the bottom panel of the pre-fit plot, the agreement between data and the predicted background is reasonable. The combined impact of the systematic uncertainties is significantly constrained as a result of the fit by using the information from the background-enriched and signal-depleted regions. This results in an overall improved background prediction with reduced uncertainties in the majority of the search regions. However, the search regions that require $\geq$4b, 0H, and a low and mid MVA score show excesses in the observed data that are not covered by the post-fit uncertainty in these regions. The post-fit event yield breakdown from these two regions is summarized in Table 6.13. As can be observed, there is a 17% and 29% excess of data over the post-fit background prediction in the $\geq$4b, 0H, LMVA and $\geq$4b, 0H, MMVA regions, respectively. For comparison, the post-fit event yield breakdown in the signal-enriched HMVA regions shows no significant excesses, as can be observed in Table 6.14. Furthermore, as can be observed between the pre-fit and post-fit yield comparison in Figure 6.34, the $t\bar{t}$ control regions and the 3b, LMVA regions drive the fit. These observations are indicative that the fit might be missing additional degrees of freedom that are needed to correct the background in the $\geq$4b, 0H, LMVA and $\geq$4b, 0H, MMVA regions. The pre-fit and post-fit m_{eff} distributions in these two regions are shown in Figure 6.35. As can be observed in the plots, the overall post-fit agreement between data and the background prediction is sensible and within the post-fit uncertainty in all bins except the second bin of the MMVA region. Since the $t\bar{t}+$ $\geq$1b background dominates in these

regions, an alternative test fit was performed with the $t\bar{t}+ \geq$1b normalization uncertainty being decorrelated across the high-statistics 2-3b regions and the low-statistics $\geq$4b regions in order to test if the initial fit model configuration has missing degrees of freedom related to this background. The results of this decorrelation test did not deviate significantly from those of the initial fit configuration. These observations will require further investigation into the fit model; however, as it was argued, these excesses are observed in signal-depleted regions and the overall agreement between data and the post-fit background on the m_{eff} distributions in these regions is good. Furthermore, the signal-enriched HMVA regions also show overall good agreement between data and the post-fit background prediction, with only the last bin of the $\geq$4b, 0H, HMVA region showing a downward fluctuation, as can be observed in Figures 6.36 and 6.37. Thus, these observed excesses can be deemed as not significant.

	≥4b, 0H, MMVA	≥4b, 0H, LMVA
$t\bar{t}$+light-jets	1.58 ± 1.2	4.35 ± 2.76
$t\bar{t}$+≥1c	3.97 ± 1.36	10.97 ± 3.076
$t\bar{t}$+≥1b	28.88 ± 3.03	83.71 ± 8.07
Single-top	2.42 ± 2.94	3.12 ± 2.52
W+jets	0.7 ± 0.26	1.7 ± 0.58
Z+jets	0.12 ± 0.045	0.27 ± 0.092
$t\bar{t}$V	1.75 ± 0.81	3.87 ± 1.08
$t\bar{t}$H	2.12 ± 0.3	5.91 ± 0.76
$t\bar{t}t\bar{t}$	2.52 ± 0.75	4.16 ± 1.25
Dibosons	0.14 ± 0.1	0.38 ± 0.23
QCD	0.13 ± 0.13	0.22 ± 0.19
Total	44.34 ± 5.15	118.66 ± 8.14
Data	57	139

Table 6.13: Predicted and observed yields in the two search regions that show an observed excess of data after performing the background-only fit. The individual systematic uncertainties for the different background processes can be correlated, and do not necessarily add in quadrature to equal the systematic uncertainty in the total background yield. The quoted uncertainties are computed after taking into account correlations among nuisance parameters and among processes. The statistical uncertainty is added in quadrature to the systematic uncertainties.

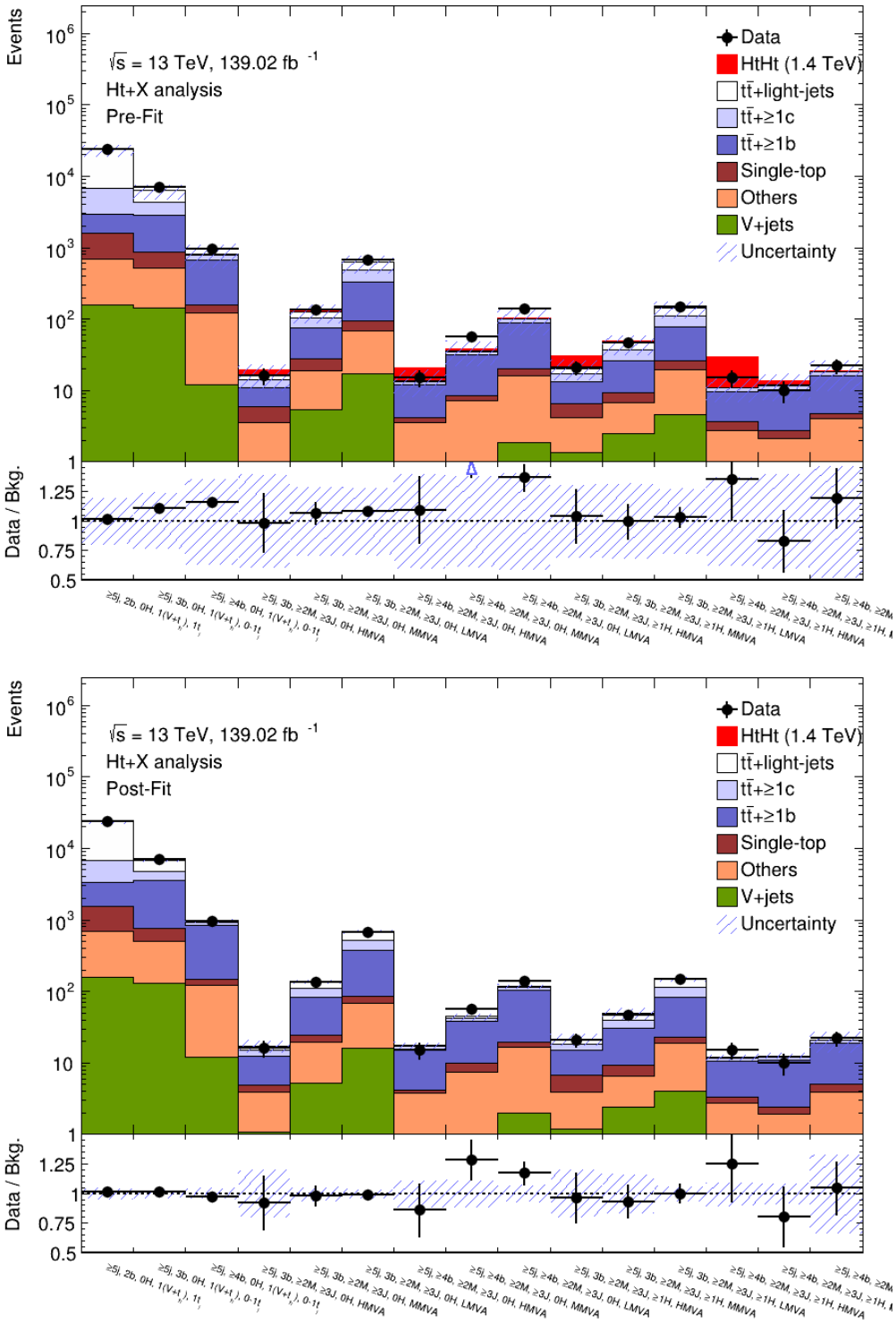
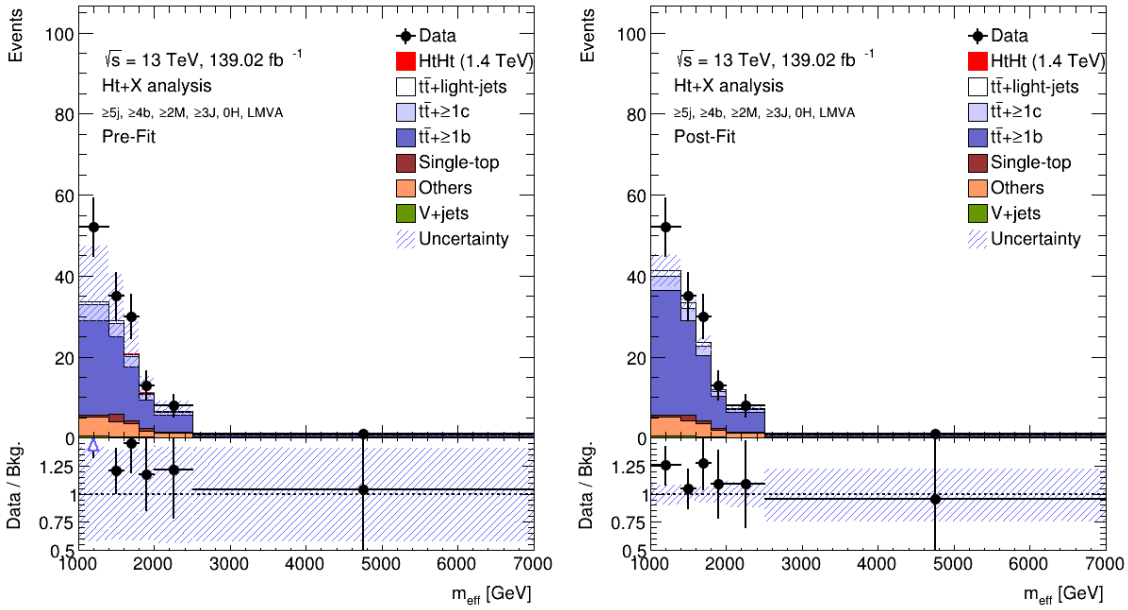


Figure 6.34: Comparison between the data and background prediction yields in each of the fit regions considered (top) pre-fit and (bottom) post-fit, performed under the background-only hypothesis. The "Others" background includes the $t\bar{t}V/H$, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The expected $TT \to HtHt$ signal (solid red) for $m_T = 1.4$ TeV is included in the pre-fit figure. The bottom panels display the ratios of data to the total background prediction. The hashed area represents the total uncertainty on the background.

	3b, 0H, HMVA	≥4b, 0H, HMVA	3b, ≥1H, HMVA	≥4b, ≥1H, HMVA
$t\bar{t}$+light-jets	2.33 ± 0.46	0.54 ± 0.38	3.25 ± 0.57	0.3 ± 0.21
$t\bar{t}$+≥1c	2.62 ± 0.89	1.34 ± 0.48	3.54 ± 0.92	1.05 ± 0.52
$t\bar{t}$+≥1b	7.47 ± 1.17	11.49 ± 1.6	8.22 ± 1.25	7.31 ± 1.019
Single-top	1.1 ± 1.22	0.28 ± 0.43	2.93 ± 1.06	0.59 ± 0.7
W+jets	0.96 ± 0.37	0.4 ± 0.15	1.09 ± 0.38	0.31 ± 0.14
Z+jets	0.11 ± 0.051	0.042 ± 0.016	0.11 ± 0.045	0.035 ± 0.0188
$t\bar{t}$V	1.26 ± 0.35	0.57 ± 0.33	1.04 ± 0.33	0.63 ± 0.41
$t\bar{t}$H	0.46 ± 0.092	0.84 ± 0.18	0.71 ± 0.1	0.91 ± 0.14
$t\bar{t}t\bar{t}$	0.75 ± 0.22	1.53 ± 0.47	0.51 ± 0.16	0.62 ± 0.19
Dibosons	0.16 ± 0.22	0.083 ± 0.055	0.21 ± 0.13	0.1 ± 0.07
QCD	0.15 ± 0.073	0.37 ± 0.46	0.19 ± 0.096	0.16 ± 0.11
Total	17.36 ± 3.4	17.48 ± 2.04	21.8 ± 4.25	12 ± 1.21
Data	16	15	21	15

Table 6.14: Predicted and observed yields in the four of the most sensitive search regions considered after performing the background-only fit. The individual systematic uncertainties for the different background processes can be correlated, and do not necessarily add in quadrature to equal the systematic uncertainty in the total background yield. The quoted uncertainties are computed after taking into account correlations among nuisance parameters and among processes. The statistical uncertainty is added in quadrature to the systematic uncertainties.



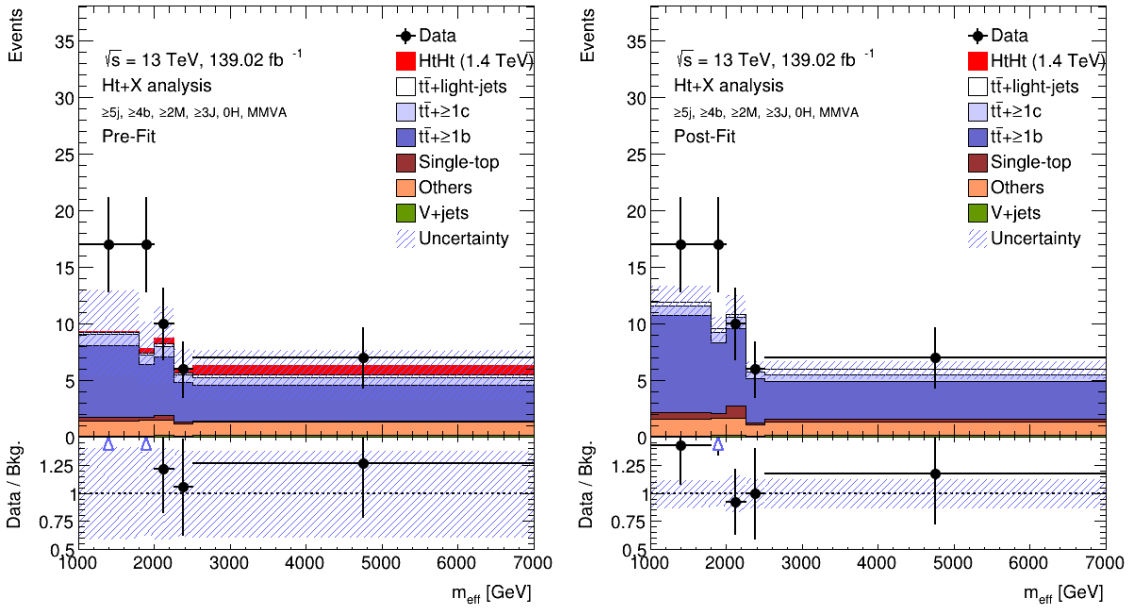
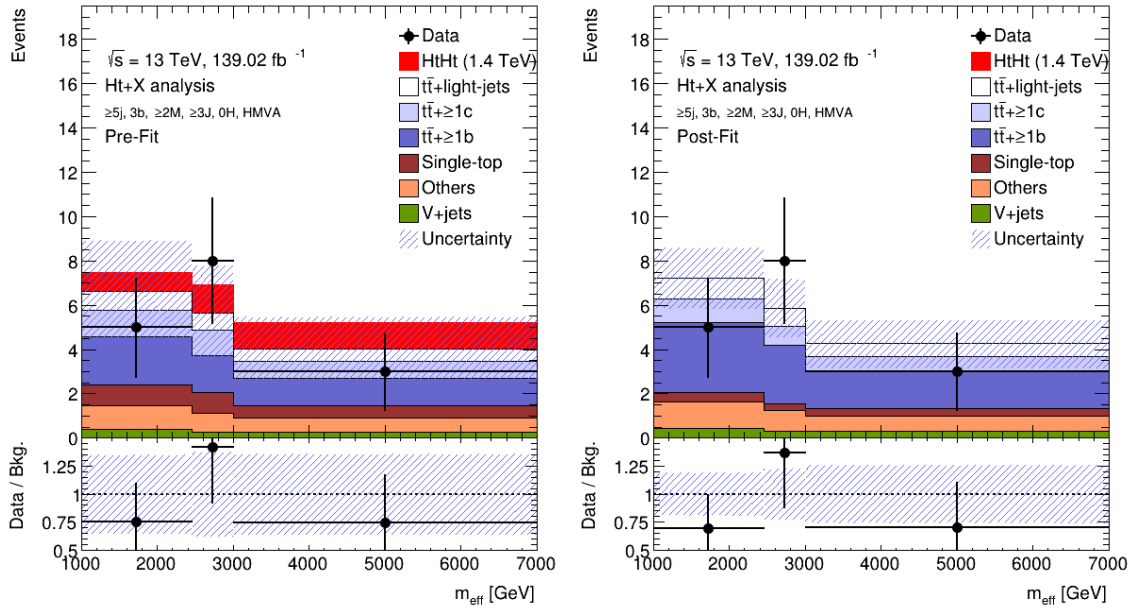
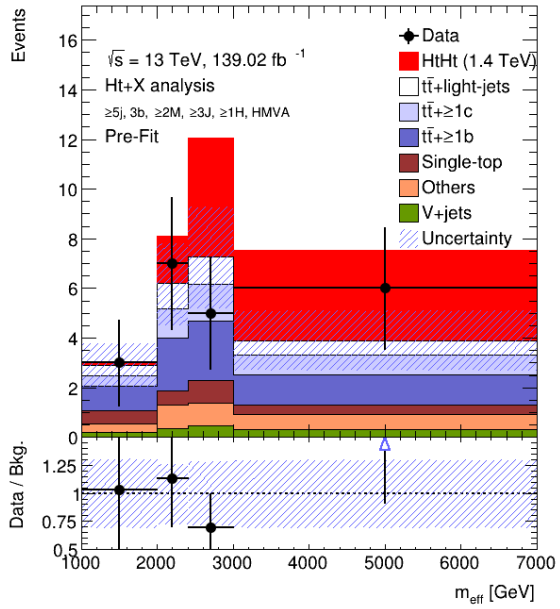


Figure 6.35: Comparison between the data and prediction for the m_{eff} distribution under the background-only hypothesis, in the ($\geq$5j, $\geq$4b, $\geq$2M, $\geq$3J, 0H, LMVA) region (a) pre-fit and (b) post-fit, and the ($\geq$5j, $\geq$4b, $\geq$2M, $\geq$3J, 0H, MMVA) region (c) pre-fit and (d) post-fit. The "Others" background includes the $t\bar{t}V/H$, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The expected $TT \to HtHt$ signal (solid red) for $m_T = 1.4$ TeV is included in the pre-fit figures. The bottom panels display the ratios of data to the total background predictions. The hashed area represents the total uncertainty on the background.

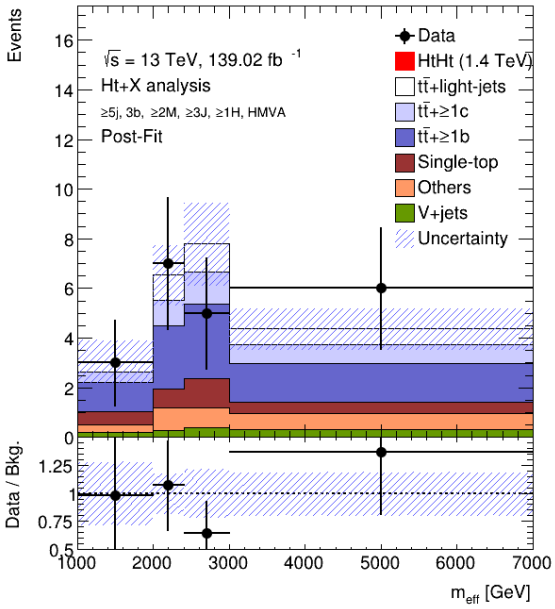


(a)

(b)



(c)



(d)

Figure 6.36: Comparison between the data and prediction for the m_{eff} distribution under the background-only hypothesis, in the ($\geq$5j, 3b, $\geq$2M, $\geq$3J, 0H, HMVA) region (a) pre-fit and (b) post-fit, and the ($\geq$5j, 3b, $\geq$2M, $\geq$3J, $\geq$1H, HMVA) region (c) pre-fit and (d) post-fit. The "Others" background includes the $t\bar{t}V/H$, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The expected $TT \to HtHt$ signal (solid red) for $m_T = 1.4$ TeV is included in the pre-fit figures. The bottom panels display the ratios of data to the total background predictions. The hashed area represents the total uncertainty on the background.

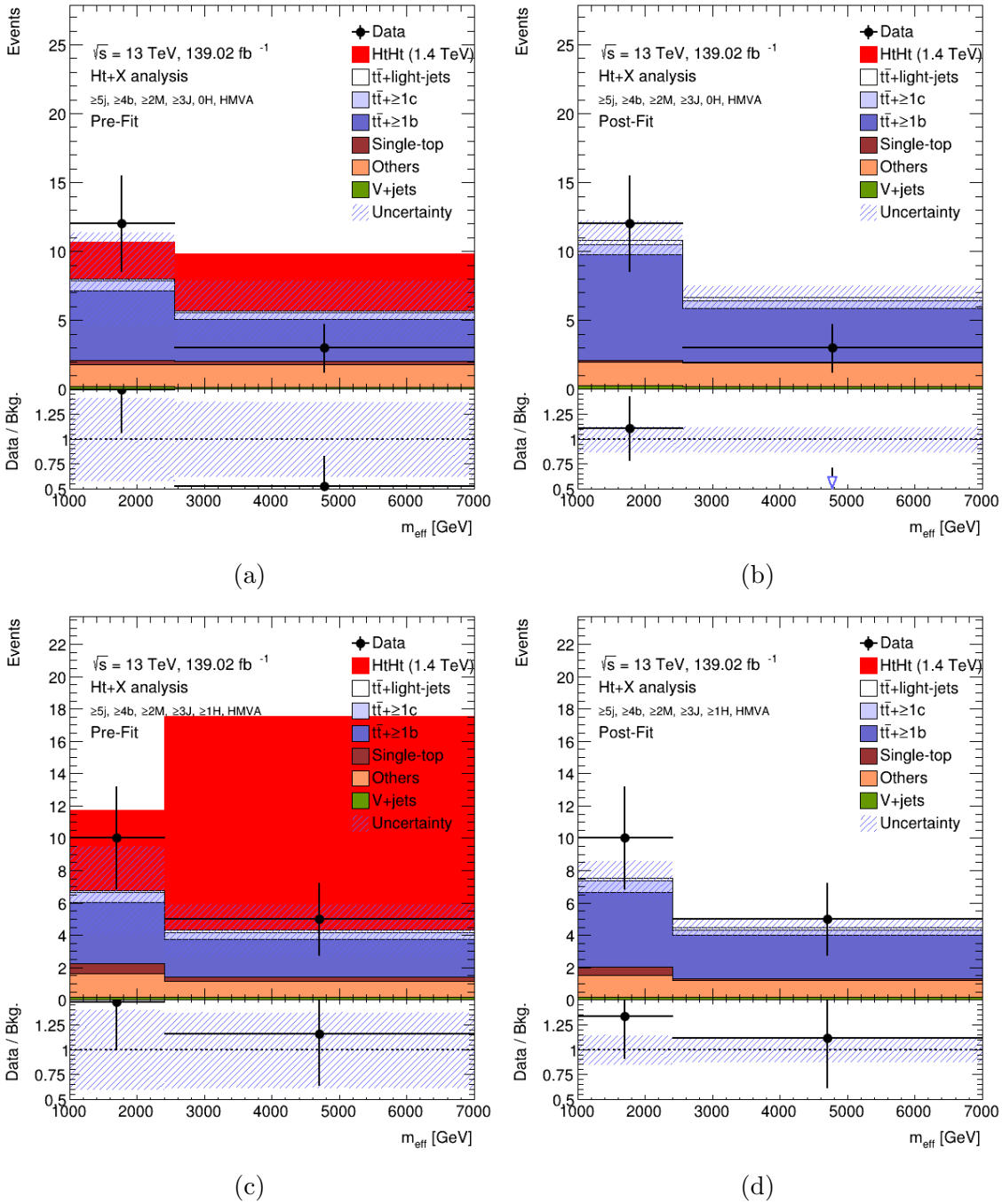


Figure 6.37: Comparison between the data and prediction for the m_{eff} distribution under the background-only hypothesis, in the ($\geq$5j, $\geq$4b, $\geq$2M, $\geq$3J, 0H, HMVA) region (a) pre-fit and (b) post-fit, and the ($\geq$5j, $\geq$4b, $\geq$2M, $\geq$3J, $\geq$1H, HMVA) region (c) pre-fit and (d) post-fit. The "Others" background includes the $t\bar{t}V/H$, $t\bar{t}t\bar{t}$, diboson, and multijet backgrounds. The expected $TT \to HtHt$ signal (solid red) for $m_T = 1.4$ TeV is included in the pre-fit figures. The bottom panels display the ratios of data to the total background predictions. The hashed area represents the total uncertainty on the background.

To further investigate the robustness of the fit model, a likelihood fit under the signal-plus-background hypothesis was performed assuming the different signal benchmark scenarios and mass points that are considered in this analysis. The observed data is fitted with the signal-strength parameter μ treated as a floating parameter of the fit. In all scenarios, the post-fit signal strength is negative, which indicates that the fit model disfavors the signal-plus-background hypothesis. Furthermore, the robustness of the uncertainty model was assessed by performing the signal-plus-background fit four times for each individual nuisance parameter, with the value of the nuisance parameter θ_k being fixed to one of the following values per fit: $\theta_k^{\text{pre-fit}} \pm \Delta\theta_k^{\text{pre-fit}}$, and $\theta_k^{\text{post-fit}} \pm \Delta\theta_k^{\text{post-fit}}$. Here $\Delta\theta_k$ denotes the uncertainty of the nuisance parameter θ_k. These fits allow us to determine the impact that each nuisance parameter has on the signal-strength parameter by calculating the difference between the μ obtained in these fits and the one obtained from the nominal signal-plus-background fit. This information is summarized in Figures 6.38 and 6.39, which show the 20 leading nuisance parameters ranked based on their post-fit impact on μ assuming the two signal benchmarks that the 1-lepton channel is most sensitive to, which are the $HtHt$ and doublet signals, with $m_T = 1.6$ TeV. In addition to the nuisance parameter ranking, these plots also show the deviation, or pull, of each nuisance parameter post-fit value from its nominal value, as well as the constraint on the nuisance parameter uncertainty that results from the fit.

As can be observed from these plots, a large fraction of the top-ranked nuisance parameters are associated with the modeling uncertainties of the $t\bar{t}$ background processes and the jet experimental uncertainties. Overall, the top-ranked nuisance parameters are well-behaved, with only a few exhibiting mild pulls and constraints, the most noticeable of which comes from the uncertainty associated with the normalization of the $t\bar{t}+ \geq$1b background. The pull and constraint of this uncertainty can be ascribed to the high-statistics $t\bar{t}$ control re-

gions with 3b and ≥4b, which the fit uses to correct the normalization of this background and consequently reduce its uncertainty. The nuisance parameters that exhibit the strongest post-fit impact on μ are associated with the $t\bar{t}+Wt$ background reweighting uncertainty in $N_{\text{jets}} \geq 8$, the modeling and normalization uncertainties of the $t\bar{t}+$ ≥1b background, the uncertainty on the jet mass resolution (JMR), and the uncertainty on the extrapolation of the b-jet tagging scale factors for jets that have a p_{T} greater than the validity range of the data sample used for the calibration of the tagger. The impacts from the b-jet tagging extrapolation and the modeling and normalization of $t\bar{t}+$ ≥1b background are expected since the $HtHt$ and doublet signals mostly populate the 3b and ≥4b search regions, which are characterized by a large presence of b-tagged jets by construction and are dominated by the $t\bar{t}+$ ≥1b background. The impact from the JMR uncertainty can be ascribed to the effect that the jet mass smearing associated with this uncertainty has when it is propagated to the RC jets. This can potentially cause event migration between search regions, which can happen as the JMR-varied RC jet is tagged to a different particle compared to the nominal RC jet or as the MVA input variables that depend on RC jets change significantly, thereby resulting in a different MVA score than the nominal event. Finally, the impact from the uncertainty on the $t\bar{t}+Wt$ background reweighting in $N_{\text{jets}} \geq 8$ is also expected as a large number of jets drives the m_{eff} towards higher values where the signal is expected to reside.

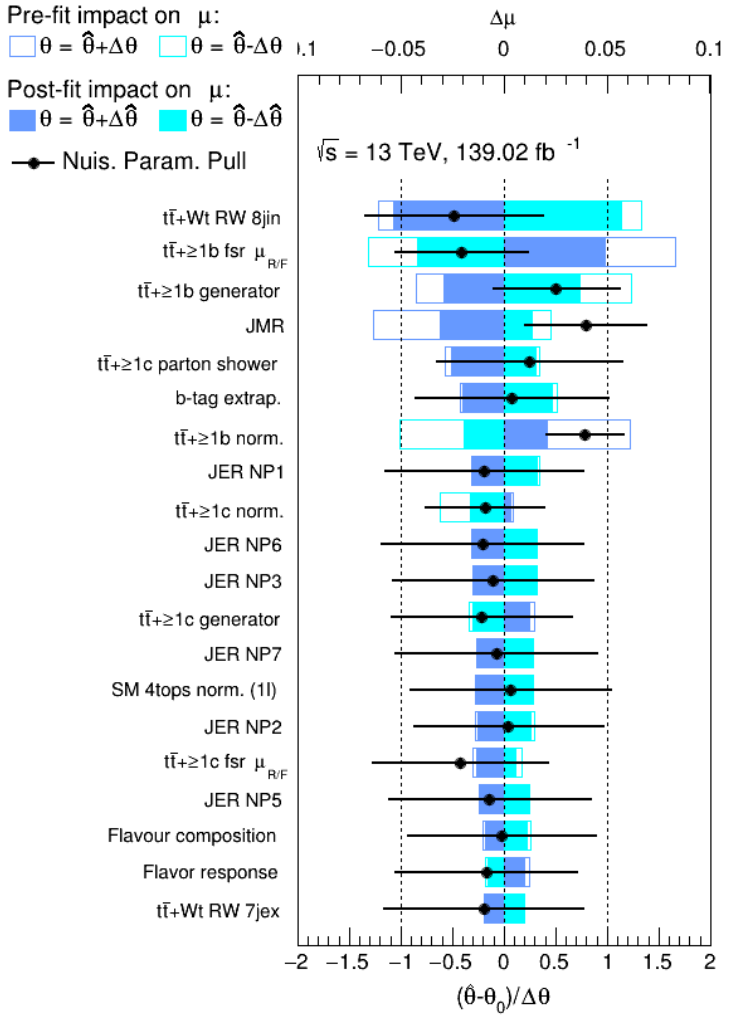


Figure 6.38: The pre-fit and post-fit impacts of the 20 leading nuisance parameters on the signal strength parameter μ under the signal-plus-background hypothesis, assuming a 1.6 TeV $HtHt$ signal. Each nuisance parameter is ranked based on their post-fit impact on μ, which is indicated by the filled colored rectangles. The vertical axis lists the top ranked nuisance parameters in descending order. The pre-fit impact on μ is indicated by the unfilled rectangles. The impact on μ, denoted by $\Delta\mu$, is read from the top horizontal axis. The black markers represent the deviation, or pull, of the corresponding post-fit nuisance parameter from their nominal value, measured in units of the pre-fit standard deviation $\Delta\theta$. The black error bars represent the post-fit uncertainty of the corresponding nuisance parameter. This information is read from the bottom horizontal axis.

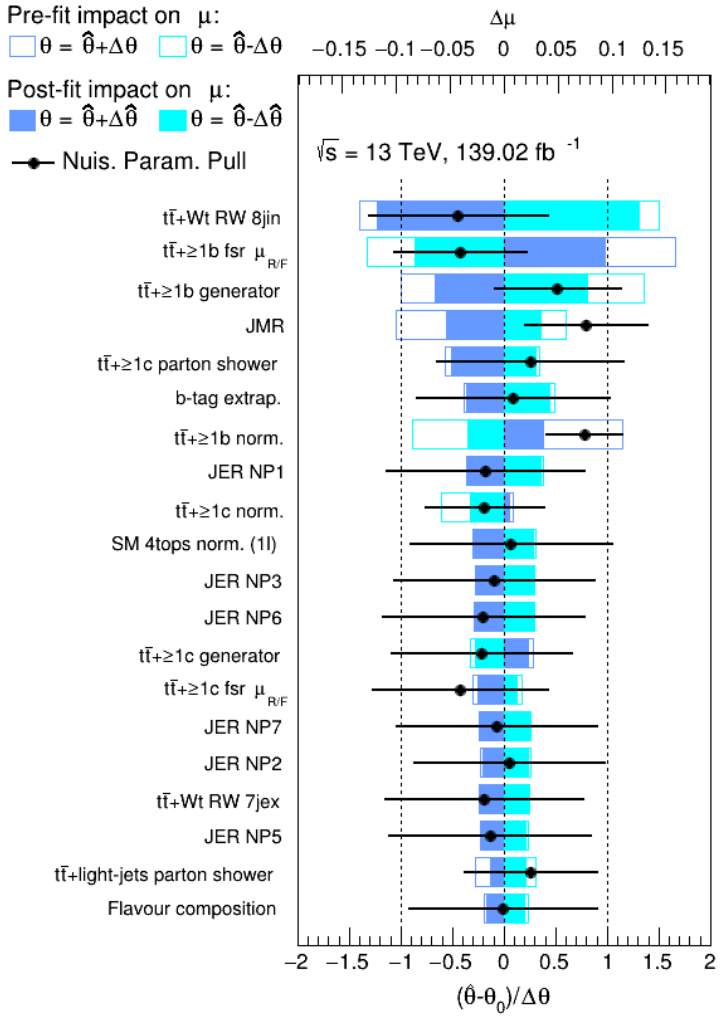


Figure 6.39: The pre-fit and post-fit impacts of the 20 leading nuisance parameters on the signal strength parameter μ under the signal-plus-background hypothesis, assuming a 1.6 TeV doublet signal. Each nuisance parameter is ranked based on their post-fit impact on μ, which is indicated by the filled colored rectangles. The vertical axis lists the top ranked nuisance parameters in descending order. The pre-fit impact on μ is indicated by the unfilled rectangles. The impact on μ, denoted by $\Delta\mu$, is read from the top horizontal axis. The black markers represent the deviation, or pull, of the corresponding post-fit nuisance parameter from their nominal value, measured in units of the pre-fit standard deviation $\Delta\theta$. The black error bars represent the post-fit uncertainty of the corresponding nuisance parameter. This information is read from the bottom horizontal axis.

6.2.8.2 Limits on Pair Vector-Like Quark Production

As argued in the preceding section, no significant excess above the SM prediction is found in the 1-lepton channel regions. Furthermore, the fits performed under the signal-plus-background hypothesis with the signal-strength parameter μ free-floating were consistent with the background-only hypothesis. Upper limits at the 95% CL on the $T\bar{T}$ production cross section are derived in the four signal benchmarks considered in this analysis and compared to the leading order theory prediction. The obtained limits are shown in Figure 6.40. As can be observed from the plots, an excess above the expected limit is observed for $m_T < 1$ TeV, which ranges between 1-2σ for the doublet and $HtHt$ signal benchmarks. For $m_T > 1.2$ TeV, a deficit below the expected limit is observed, being close to -1σ for the singlet and $ZtZt$ signal benchmarks. This is expected as the 1-lepton channel has little sensitivity to the singlet and $ZtZt$ signals. Since no significant excess is observed in any of the signal benchmarks, a claim for the evidence of $T\bar{T}$ production cannot be made. Thus, the $T\bar{T}$ production is excluded for all T masses where the observed limit is below the theory prediction. The lower limits set on the T mass for each signal benchmark considered are shown in Table 6.15. The limits set by the previous iteration of this analysis in the 1-lepton channel [97], which was based on a dataset of recorded collisions in 2015-2016 corresponding to an integrated luminosity of 36 fb^{-1}, are also listed for comparison. As can be observed, a significant improvement on the lower limits on the T mass has been achieved when compared to the limits obtained in the 1-lepton channel from the previous iteration of the analysis.

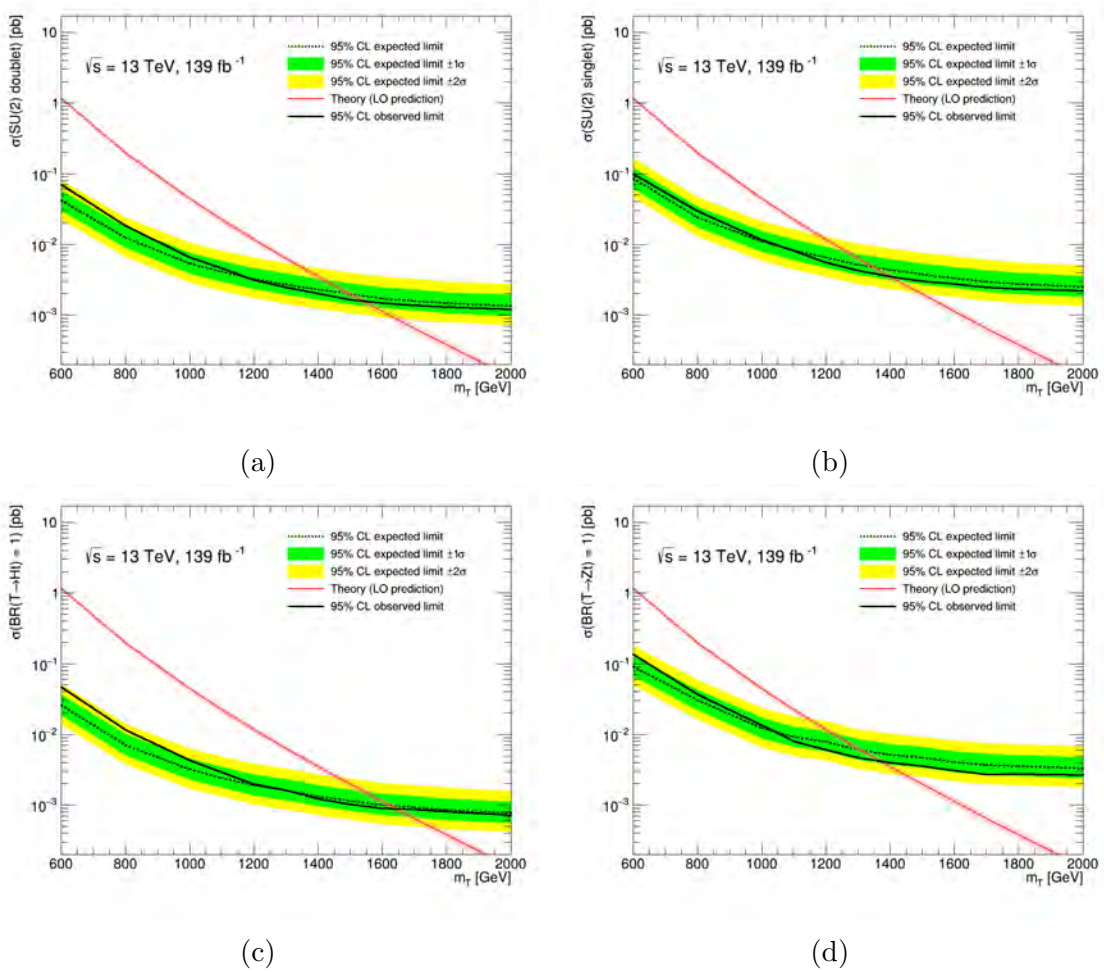


Figure 6.40: Observed (solid line) and expected (dashed line) 95% CL upper limits on the $T\bar{T}$ production cross section as a function of the T quark mass in the doublet (a), singlet (b), $HtHt$ (c), and $ZtZt$ (d) signal scenarios. The surrounding shaded bands correspond to ± 1 and ± 2 standard deviations around the expected limit. The red line shows the LO theoretical cross section prediction.

1-lepton channel 95% CL lower limits on T quark mass [TeV]				
Analysis iteration	BR($T \to Ht$)=1	BR($T \to Zt$)=1	Doublet	Singlet
Current	1.64 (1.62)	1.37 (1.31)	1.54 (1.50)	1.40 (1.35)
Previous	1.47 (1.30)	1.12 (0.91)	1.36 (1.16)	1.23 (1.02)

Table 6.15: Summary of the observed (expected) 95% CL lower limits on the T quark mass for the different signal benchmarks considered that were obtained in the 1-lepton channel of the current and previous iteration of the analysis.

Chapter 7

Conclusion

This dissertation describes two research topics. The first topic describes the tagging of collimated sprays of particle decays that are initiated by the process of hadronization, known as jets, to the particle that instigated the process. This topic covers two jet tagging studies. The first one consists of the development and calibration of three jet substructure-based taggers to tag jets to top quarks and W bosons. The second study consists of performing a topological data analysis (TDA) of jets, which has not been used in the context of jet tagging previously, and applying the information obtained in the design of two top tagging algorithms. The second topic describes two search analyses for a hypothetical vector-like top quark (T) that decays to a Higgs boson and a top quark (Ht), or a Z boson and a top quark (Zt), associated with the presence of a single electron or muon. The first analysis focuses on the single production of a T, which is mediated through the electroweak force. This analysis allows to probe the universal coupling strength κ, which controls the coupling of the T to the W, Z and Higgs bosons and the production cross section. The results of this analysis are interpreted in the $SU(2)$ singlet ($T^{2/3}$) and doublet ($T^{2/3}\,B^{-1/3}$) signal scenarios. The second analysis focuses on the pair production $T\bar{T}$, which is mediated by the strong force. The results of this analysis are interpreted in the $SU(2)$ singlet and doublet scenarios, as well as assuming the branching ratios $\mathrm{BR}(T \to Ht) = 1$ and $\mathrm{BR}(T \to Zt) = 1$. A summary of both topics as well as potential outlooks is given in this Chapter.

7.1 Tagging Top Quark Studies

For the first jet tagging study, a three-variable W tagger and two deep neural network (DNN) top taggers, one designed to tag jets that contain the full decay products of the top quark and the other designed to tag jets regardless of the full containment, were optimized to perform their corresponding tagging tasks. The performance of the taggers in Monte Carlo (MC) simulation was calibrated to the performance of the data. The MC modeling of the data in the input variables of the taggers was assessed, including the effects of various sources of systematic uncertainty that are associated with the modeling of physics processes and the reconstruction and calibration of relevant physics objects. The overall agreement between MC and the data is good in the input variables. Some moderate MC modeling discrepancies are observed in the input variables of the W tagger in a region close to the W-tagged region; however, these differences are within the total uncertainty considered. Data to MC scale factors were derived for the signal jet tagging efficiency measurement. The scale factors were found to range between 0.8 and 1, with the lowest values being attained by the W tagger. The MC overestimate of the W tagging signal efficiency is attributed to the moderate discrepancies observed in the modeling of the tagger input variables. Finally, the effects of the systematic uncertainties were propagated to the scale factor calculation in order to provide an uncertainty of these measurements. The uncertainties associated with the modeling of the $t\bar{t}$ production process are observed to be the primary source of uncertainty. This is expected as these uncertainties can significantly vary the hadronization of the signal $t\bar{t}$ process, which in turns varies the simulated detector response and the jet reconstruction process, thereby having ramifications for the signal efficiency measurement.

For the second jet tagging study, two TDA techniques were applied to analyze the ho-

mology of top jets and QCD jets using their associated topoclusters to build simplicial complexes. The first technique is a Persistent Homology (PH) analysis, which was used to study how the homology of jets varied as a function of a distance scale parameter. A hypothesis was formulated that the number of connected components that are formed by the topoclusters corresponds to the decay topology of a signal top jet. A distance scale parameter of $\Delta R = 1.2$ was determined from the average distance scale at which signal top jets have two connected components, which corresponds to a top quark decay topology where the decays of the W boson are collimated under the assumed hypothesis. The homology of both signal top and background QCD jets at this distance scale was characterized with the presence of a circular void that disappeared after the topoclusters formed a single connected component. This circular void was found to persist longer in signal jets. A kinematic description of the connected components was achieved by adding the four-momenta of the topoclusters associated with a connected component and interpreting it as a subjet. The mass distribution of the connected components in signal top jets indicate that these objects are reconstructing relevant substructures of top jets, as evidenced by the mass bumps near the W and top mass. The corresponding mass distributions from background QCD jets are indicative of reconstructing inconsistent substructures from random patterns of topoclusters. These observations give confidence in the hypothesis formulated for this study.

The second technique used in the TDA of jets is the Mapper algorithm, which analyzes the homology of jets at a fixed distance scale. The distance scale used in the Mapper algorithm studies, which governs the formation of vertices in the Čech (Č) simplicial complex of jets was set to $\Delta R = 1.2$, motivated by the results obtained from the PH analysis. A dedicated study was performed to optimally select the other parameters needed to use the Mapper algorithm. The homology of jets obtained by the Mapper algorithm is characterized by the presence of

multiple connected components and a lack of circular voids, with no significant differences observed between signal and background jets. The absence of voids in the Čech complex of jets is attributed to the combined use of a granular covering set with the distance scale parameter $\Delta R = 1.2$, which hinders the ability of the algorithm to resolve circular features in jets. Similar to the PH studies, a kinematic description of the vertices and connected components in the Čech complex of jets was given. Additionally, jet substructure-inspired observables were defined in order to quantify how the energy of the jet is distributed across its connected components. The connected components were observed to achieve a similar degree of jet substructure reconstruction as the one achieved in the PH study. Furthermore, the jet substructure-inspired observables showed differences in how the energy of the jet is distributed in vertices and connected components between signal and background jets.

Two tagging algorithms were designed to use the information obtained from the TDA of jets to classify jets as either signal top jets or background QCD jets. The first algorithm consists of a DNN tagger that uses the kinematic and substructure information of the vertices and connected components obtained from the Mapper algorithm in order to classify jets. The second algorithm consists of a convolutional graph neural network (GNN) that uses a graph representation of jets that is built from the connected components of the Čech complex of a jet. Both taggers achieved a good separation power between signal and background jets. However, the GNN tagger presented signs of undertraining, evidenced by its moderate ability to confidently tag signal jets when compared to the DNN tagger. The undertraining is attributed to the limited computational memory resources that were available when this tagger was trained. The GNN training required that the graph of the jets used for the training dataset were readily available, which exceeded the memory resources when a large number of jets were included.

The performance of the DNN and GNN taggers was compared to the contained top tagger from the first jet tagging study. Both the DNN and GNN taggers were found to be slightly outperformed by the contained top tagger. The variables obtained from the TDA of jets were compared between signal and background jets in tagging selection regions that corresponded to ambiguous classifications between the DNN or GNN and the contained top tagger. This was done in order to determine if there was any residual information from the TDA of jets that the taggers were not using to their full extent and could further improve the separation between signal and background jets. The connected components from signal top jets in these tagging selection regions were observed to partially retain their ability to reconstruct relevant substructures of top jets. Furthermore, the jet substructure-inspired observables of connected components showed differences in how the energy is distributed in these structures between signal and background jets. These observations indicate that there is residual information from the TDA of jets that the taggers are not fully utilizing.

The TDA of jets has untapped potential that can be harnessed in future endeavors. First, the assumed hypothesis that was made in the PH analysis that the number of connected components should correspond to the decay topology of a top quark may not be optimal. Instead, a more descriptive distance scale of the homology of jets could be obtained from the merging of two connected components that results in a mass close to the W boson mass, which could happen well before the jet has two or three connected components. This opens the possibility of analyzing jets with the Mapper algorithm using a distance scale parameter that varies on a jet-by-jet basis instead of a fixed-value distance parameter that may not properly characterize the homology of all jets. Another aspect that can be improved in the TDA methodology is to apply the Mapper algorithm with the use of a covering set that is composed of finer elements. For the studies presented in this thesis, a set with four granular

elements was used to cover the topocluster ϕ-projection image space. When combined with the distance scale parameter $\Delta R = 1.2$, the vertices that are formed in each cover element will tend to have large fractions of topoclusters, which trivializes the Čech complex of the jet. Thus, a covering set with finer elements can improve the resolution of the Mapper algorithm by increasing the number of vertices that better capture the small-scale structure of jets. This can be further improved by combining the use of the ϕ-projection filter function with a η-projection filter function, which increases the spatial resolution of the Mapper algorithm. Finally, the TDA of jets described in this thesis was limited to a geometric point of view. A prospect of this analysis is to study how the homology of jets is affected with the use of a distance metric that takes into account the energy of the topoclusters.

7.2 Searches for Vector-Like Quarks

The search analyses for a vector-like T quark presented in this dissertation covered the single production mechanism, which is mediated by the electroweak force, and the pair production mechanism, which is mediated by the strong force. Both analyses target the decay topology $T \to Ht$ in final states that include the presence of a single electron or muon, referred to as the 1-lepton channel. The pair production analysis will cover the 0-lepton channel; however, the 0-lepton channel analysis is at the stage of finalizing validation studies that are needed prior to performing the statistical analysis. Thus, the 0-lepton channel results are not covered in this dissertation. Both single and pair production analyses were performed using 139 fb^{-1} of data and shared the same background and systematic uncertainty models. The main irreducible background in these searches is $t\bar{t}$ production in association with additional jets. Subdominant background contributions come from the single-top and W/Z+jets production

processes.

The design of the analysis strategy for the single production search took advantage of the simultaneous presence of several unique objects in signal processes, such as forward jets, an associated top or bottom quark with the T production, and a hadronically decaying boosted Higgs boson produced from the decay of the T. The presence of these objects allowed the definition of search regions that were relatively pure in the different T decay topologies and associated production modes considered. The design of the pair production analysis strategy took advantage of the interesting decay topology combinatorics that became available with the production of an additional T. This allowed the definition of many discriminating variables between signal and background processes. An example of one of these variables is the invariant mass of reconstructed candidate Ts. The distribution of this variable in signal processes peaked sharply at the mass of the T, while for background processes it peaked at lower values and exhibited a long tail, which is characteristic of reconstructing a candidate T from inconsistent kinematics. A multivariate analysis was performed using all the discriminating variables, which resulted in the definition of a DNN that classified events as either signal $T\bar{T}$ production events or SM background events. The DNN allowed for the definition of simpler search regions that were agnostic to the decay topologies of signal processes, which contrasts with the single production analysis search regions that are tailored to the different T decay topologies and associated production modes.

Both analyses use the effective mass (m_{eff}) variable, which is defined as the scalar sum of the p_{T} of the final state jets, leptons, and $E_{\text{T}}^{\text{miss}}$ in an event, as the final discriminant between signal and background processes. The definition of this variable is motivated by the presence of a large number of energetic final state objects in signal processes that arise from the decays of the massive Ts. As a result of this, the m_{eff} allows for a discrimination between

signal and background processes that is agnostic on the decay topologies and associated production modes of the Ts. The MC simulations of the $t\bar{t}$ and W/Z+jets background processes are known to mismodel the upper tail of the jet p_{T} spectrum and the distribution of the number of jets at high multiplicities. This enters as a source of mismodeling in m_{eff} in the region where the signal is expected to reside due to how it is defined. To address this issue, data-driven correction factors were derived to improve the MC modeling of these backgrounds in this kinematic regime. The correction factors were derived in regions that are enriched in the background to be reweighted and signal-depleted in order to ensure that the presence of potential signal events is not removed by the correction factors. The modeling of the background MC to the data in the regions used to derive these correction factors, as well as orthogonal validation regions that are background-enriched and signal-depleted, was compared before and after applying the correction factors. A significant improvement in the modeling of m_{eff}, as well as other variables that are not related to m_{eff} but showed signs of being mismodeled, was observed after applying the correction factors.

In both search analyses, a statistical analysis in the form of a maximum likelihood fit was performed, where the m_{eff} distributions in all search regions of a given analysis were jointly analyzed to test for the presence of potential signal T production events in the data. In the single production analysis, no significant excess above the SM prediction was found in all search regions considered. Upper limits at the 95% CL on the cross section of the single production of a T were derived in both the singlet and doublet signal scenarios. The limits are interpreted as exclusion lower limits of the T mass and universal coupling strength κ. For the singlet scenario, masses below 2.1 TeV are excluded for $\kappa \geq 0.6$, while values of $\kappa \geq 0.3$ are excluded for a T mass of 1.6 TeV. For the doublet scenario, values of $\kappa \geq 0.55$ are excluded for a T mass of 1 TeV.

Finally, for the pair production analysis, an excess of data that is not covered by the post-fit uncertainty was observed in two search regions that are not signal-enriched. The post-fit agreement in the remaining search regions, which includes the signal-enriched search regions, is overall good. The agreement between the data and the post-fit MC background on the m_{eff} distribution in these two search regions is sensible and within the post-fit uncertainty in the majority of the m_{eff} bins. Furthermore, the likelihood fits performed under the signal-plus-background hypothesis were consistent in rejecting the signal-plus-back-ground hypothesis in favor of the background-only hypothesis. These observations indicate that the fit model is missing degrees of freedom that are required to improve the correction of the background MC prediction, which will need to be further investigated. However, based from the observations made, the observed data excesses can be deemed as non-significant. Upper limits at the 95% CL on the cross section of $T\bar{T}$ pair production were derived for the BR($T \to Ht$) = 1, BR($T \to Zt$) = 1, doublet, and singlet signal scenarios in the 1-lepton channel. These limits were interpreted as exclusion lower limits of the T mass for each signal scenario. The 1-lepton channel search excludes T masses below 1.64 TeV, 1.37 TeV, 1.54 TeV and 1.40 TeV for the BR($T \to Ht$) = 1, BR($T \to Zt$) = 1, doublet, and singlet scenarios, respectively. These limits show a significant improvement from the ones that were obtained in the previous iteration of this analysis in the 1-lepton channel. The interpretations of the results obtained may change once the results of the 0-lepton channel become available and are combined with the 1-lepton channel results.

APPENDICES

Appendix A

Monte Carlo Simulations

This appendix describes the MC simulation samples that were used to simulate the different signal and background processes of interest in the studies presented in this thesis. The samples are generated with computational tools known as MC generators that apply the MC sampling method to simulate the events of a given process of interest in order to produce distributions of kinematic variables of the process. The MC generators simulate multiple steps in a given process. First, the collision of two protons is simulated down to the level of the quarks and gluons inside protons, also known as partons. This is done using parton distribution functions (PDFs) that represent the probabilities of two given partons interacting and carrying a given fraction of the total energy of the proton. The second step consists of simulating the final state particles that are produced from the interacting partons for a given process. The third step consists of simulating the hadronization of quarks that are produced from the final state of the process of interest. Additionally, the emission of quarks and gluons from partons prior to the collision, known as initial state radiation, and after the collision, known as final state radiation, are modeled using parton shower MC generators. Finally, the detector response is simulated using the final state leptons and hadronized quarks from the previous step. In the following, the list of MC samples used in the different studies, as well as the MC generators, PDFs, and modeling parameters, is given.

Jet Tagging Study Samples

The samples used in the design and optimization of the jet taggers studied in Chapter 5 are divided into two categories: signal and background. The signal samples are generated with BSM processes that are described in the Heavy Vector Triplets framework [51], which is an extended gauge symmetry model that predicts the existence of heavy W' and Z' gauge bosons. These samples were simulated using the Pythia 8.235 [98] generator with the NNPDF2.3LO [99] PDF set and the A14 set of tuned parameters [100]. The background events used for the tagger optimization are QCD multijet events. These are generated using Pythia 8.230 with the NNPDF2.3LO PDF set and the A14 set of tuned parameters.

The samples used for the signal efficiency calibration of the jet substructure taggers are also divided into signal and background. The $t\bar{t}$ and single top signal samples are used to model events with jets originating from top quarks and W bosons. These samples were simulated with Powheg [101, 102, 103] interfaced with the Pythia 8.230 generator. Alternative $t\bar{t}$ samples were also used for the evaluation of systematic uncertainties of the signal efficiency calibration. The samples used to assess the uncertainty on the matching of the next to leading order (NLO) matrix-elements and parton shower for $t\bar{t}$ samples were generated with MadGraph5_aMC@NLO v2.6.0 [104] interfaced with Pythia 8.230. To assess the uncertainty on the choice of the parton shower and hadronization algorithm, samples were simulated using Powheg interfaced with Herwig 7.04 [105, 106] to model the parton shower and hadronization.

The background samples used for the signal efficiency calibration consists of simulations of W/Z+jets (V+jets) and diboson production processes. The V+jets samples were generated with Sherpa v2.2.1 [86], while the diboson samples were generated with Sherpa v2.1.

Single and Pair Production of Vector-Like Quarks Samples

The single production of T vector-like quarks was simulated with samples produced with the MadGraph5_aMC@NLO v2.3.3 generator interfaced with Pythia 8.212 for the modeling of the parton showering and hadronization. The NNPDF3.0LO PDF set and the A14 set of tuned parameters are used. The VLQs are assumed to couple exclusively to the third generation SM quarks. Separate samples were generated for the $T(\rightarrow Ht)qb$ and $T(\rightarrow Zt)qb$, $T(\rightarrow Ht)qt$ and $T(\rightarrow Zt)qt$ processes in the 1.1-2.3 TeV mass range at fixed values of mass and coupling strength parameter κ.

The pair production of T vector-like quarks was simulated with samples produced with the Protos [107] generator using the NNPDF2.3LO PDF set and the A14 set of tuned parameters. These events were interfaced with Pythia 8.212 to model the parton showering and hadronization. The samples were generated assuming singlet couplings and forced to decay with equal branching ratios to Ht, Zt, and Wb. Additionally, the samples were generated in the T mass range 600-2000 GeV in steps of 100 GeV.

Both the single and pair production analyses have a similar background model; thus, the majority of the MC samples that are used to simulate the background processes were generated with same configurations for both analyses. The following description of the background samples applies to both analyses, unless otherwise stated.

The $t\bar{t}$ and single top production background processes were modeled using the Powheg generator at NLO with the NNPDF3.0LO PDF set. The events were interfaced to Pythia 8.230 to model the parton shower and hadronization. The $t\bar{t}$ samples were generated inclusively, but events are categorized based on the flavor content of additional particle jets that

do not originate from the decay of the $t\bar{t}$ system. These events are labeled as $t\bar{t}+ \geq 1b$, $t\bar{t}+ \geq 1c$, and $t\bar{t}$+light-jets.

The associated production of a single top quark with W bosons has significant contributions in regimes of high transverse momentum. Samples to model the single top Wt-channel were generated using the diagram removal scheme [108] in order to remove interference and overlap with $t\bar{t}$ production. The uncertainty associated with this procedure is estimated by comparing with an alternative Wt sample generated using the diagram subtraction scheme [109] and the same generator setup as the nominal sample. Separate samples were generated to model the s-channel and t-channel of single top production.

Additional alternative $t\bar{t}$ and single top production samples were used to evaluate systematic uncertainties on the modeling of these processes. The impact on the choice of the parton shower and hadronization model is evaluated with samples that were generated with the Powheg generator using the NNPDF3.0NLO PDF set, but interfaced with Herwig 7.04. The uncertainty on the matching of NLO matrix-element and parton shower for the $t\bar{t}$ samples is evaluated by comparing the nominal sample that was generated using Powheg with an alternative sample generated with MadGraph5_aMC@NLO v2.6.0. For single top production, the nominal sample was compared with an alternative sample generated with MadGraph5_aMC@NLO v2.6.2.

The V+jets production background process in the single production analysis was simulated with the Sherpa v2.2.1 using the NNPDF3.0NNLO PDF set. In the pair production analysis this process was simulated with Sherpa v2.2.11, which improves the modeling of this background. Diboson production in the single production analysis was simulated with the Sherpa v2.2.1 or Sherpa v2.2.2 generators depending on the process. In the pair production analysis this process was simulated with Sherpa v2.2.11. The production of $t\bar{t}W$

and $t\bar{t}Z$ ($t\bar{t}V$) were simulated using the MadGraph5_aMC@NLO v2.3.3 generator with the NNPDF3.0NLO PDF set interfaced to Pythia 8.210. The production of $t\bar{t}H$ was simulated using the Powheg generator with the NNPDF3.0NLO PDF set interfaced to Pythia 8.230. The production of four top quarks was simulated with the MG5_aMC v2.2.2 generator with the NNPDF2.3LO PDF set interfaced to Pythia 8.186. Finally, the QCD multijet samples were simulated using Pythia 8.230.

Appendix B

Mapper Algorithm Optimization

This appendix summarizes the optimization studies that were performed to determine the optimal set of parameters to be used with the Mapper algorithm. These parameters are: the filter function that maps the topoclusters to an image topological space; the covering set of the image topological space, from which the Čech complex of the jet is obtained; the clustering algorithm that is applied in each cover element, which provides the vertices of the Čech complex; and the distance resolution scale ΔR_{res}, which determines the clustering threshold distance that governs the formation of vertices. The final choice of the parameters consists of projecting the topoclusters to the ϕ-axis of the η-ϕ plane, which is covered by the set of overlapping intervals $\mathcal{U} = \{[-3.2, -1.2], [-2.0, 0.4], [-0.4, 2.0], [1.2, 3.2]\}$. The topoclusters in each interval are clustered using a single-linkage clustering algorithm with $\Delta R_{\text{res}} = 1.2$. These parameters were chosen due to their interpretability on how the Mapper algorithm works and their effectiveness in reconstructing the relevant substructures in signal top jets.

The optimization was performed with a grid search that varied a single parameter option at a time. The topological and kinematic distributions of the jets, vertices and connected components were analyzed in order to make the final choice of parameter options. The variations in the filter function and clustering algorithm exhibited some of the largest differences in the output of the Mapper algorithm. In the following, the distributions of topological and

kinematic variables of different objects are compared between different options that were considered for the filter function and clustering algorithm during the optimization process, both for signal top jets and background QCD jets.

Filter Function Optimization

A comparison between a subset of the filter functions considered for the Mapper algorithm is presented. The other parameters of the Mapper algorithm are set to their final choice in the results shown here, with the exception of the covering set, which varies depending on the definition of the filter function. Although the final choice of the filter function consists of a single function, the Mapper algorithm can be extended to use an arbitrary number of filter functions. The use of two filter functions was considered during the optimization process. In this case, the covering set consists of square grids that are built from the covering intervals of the individual filter functions. The overlap region, from which edges between the vertices of the Č complex are defined, can maximally consist of four overlapping square grids instead of two overlapping intervals as in the case of a single filter function. A description of the subset of filter functions shown in this comparison study is given below:

η-Projection

This function projects the coordinate pair of topoclusters onto the η-axis of the η-ϕ plane. The covering set that is used with this filter function is the same as the one used with the ϕ-projection filter function.

Log Sigmoid ΔR

This function is defined as the absolute value of the natural logarithm of the product of three sigmoid functions. Each individual sigmoid function is designed to measure the distance response of a topocluster t to a reference topocluster t' in the jet. The reference topoclusters chosen are the three leading in p_{T} topoclusters in the jet: t_0, t_1, and t_2. This filter function can be expressed mathematically as:

$$f(t) = |\ln\left(s(t,t_0)s(t,t_1)s(t,t_2)\right)| \tag{B.1}$$

where the sigmoid function $s(t,t_i)$ is defined as

$$s(t,t_i) = \frac{1}{1+e^{-\Delta R(t,t_i)/\Delta R_{\mathrm{res}}}} \tag{B.2}$$

The distance between topoclusters t and t_i is scaled by the threshold distance that is used in the clustering process of the Mapper algorithm. The motivation of this filter function is to map topoclusters that are spatially close to the three leading topoclusters of the jet onto the same region of the image space of the filter function. This is done in order to construct substructures that are centered around the most energetic topoclusters of the jet. The optimal covering set that is used with this filter function is $\mathcal{U} = \{[0, 0.7], [0.66, 1.1], [0.88, 1.4], [1.31, 2.1]\}$.

Log Sigmoid $E_{\mathbf{T}}$

This function is defined as the absolute value of the natural logarithm of a sigmoid function that measures the energy response between a topocluster t relative to the average transverse

energy of the topoclusters in the jet. This filter function can be expressed mathematically as:

$$f(t) = \left| \ln \left(\frac{1}{1 + e^{-\sqrt{E_{\mathrm{T},t}/E_{\mathrm{T}}^{\mathrm{avg}}}}} \right) \right| \tag{B.3}$$

where $E_{\mathrm{T},t}$ is the transverse energy of the topocluster t and $E_{\mathrm{T}}^{\mathrm{avg}}$ is the average transverse energy of the topoclusters in the jet. The motivation of this filter function is to map topoclusters that are energetically similar onto the same region of the image space of the filter function. The optimal covering set that is used with this filter function is $\mathcal{U} = \{[0, 0.1], [0.04, 0.4], [0.21, 0.54], [0.49, 0.7]\}$.

Momentum Fractions

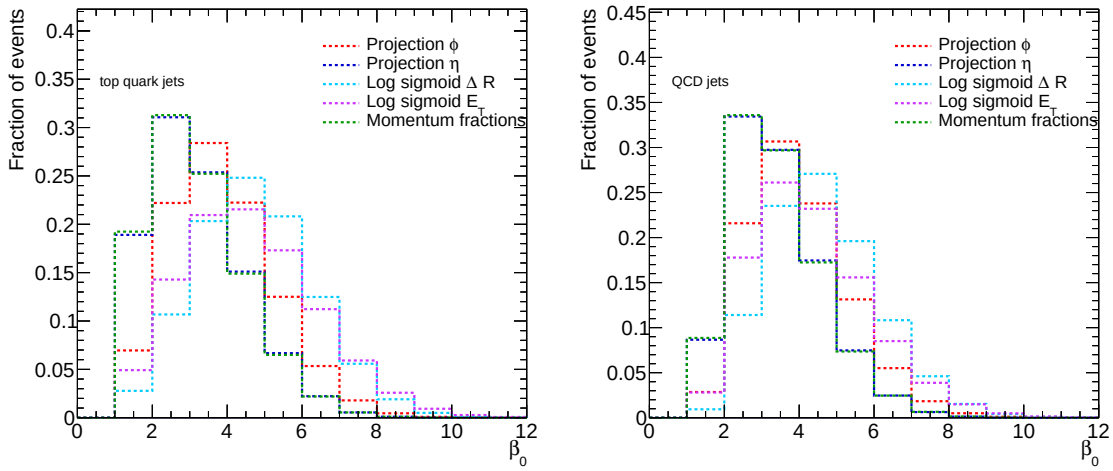
These two filter functions are defined as the ratio of the x and y momentum components of a topocluster t with the scalar sum of the p_{T} of the topoclusters in the jet. These filter functions can be expressed mathematically as:

$$f(t) = \frac{p_{i,t}}{\sum_{t' \in J} p_{\mathrm{T},t'}} \tag{B.4}$$

where i is the x or y momentum component and the sum in the denominator runs along all the topoclusters t' in the jet J. These two functions are used together in the Mapper algorithm. The optimal covering sets for both functions are the same and is given by $\mathcal{U} = \{[-1.1, -0.25], [-0.45, 0.45], [0.25, 1.1]\}$. These intervals are combined in order to form a covering grid.

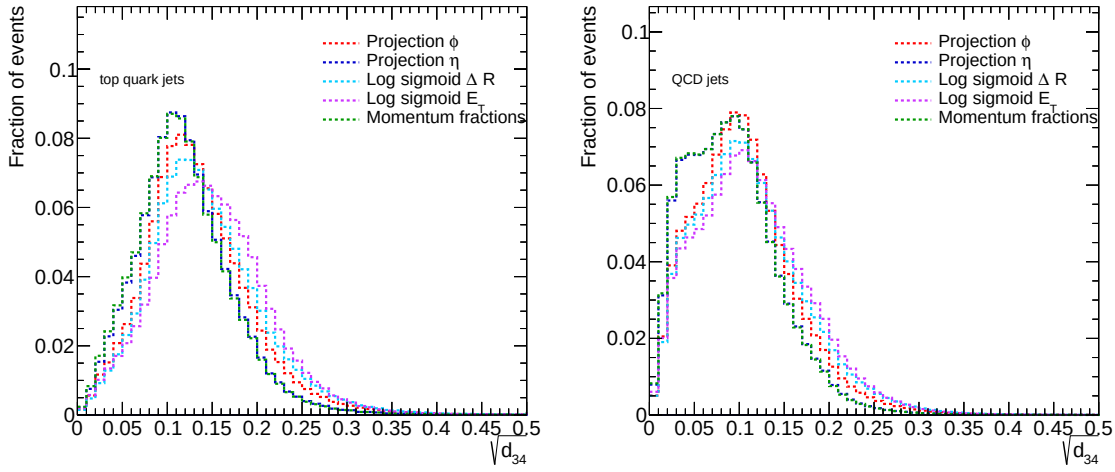
As can be observed in Figure B.1, both for signal and background jets, the η-projection and momentum fraction filter functions produce fewer connected components on average,

the log sigmoid filter functions tend to produce more connected components on average, and the ϕ-projection filter function results in an average number of connected components. From the Cambridge-Aachen splitting scales shown in Figures B.2 - B.4, it is observed that the η-projection and momentum fraction filter functions tend to produce CCs that are, on average, spatially closer at each reclustering step. The n-subjettiness ratio distributions behave similarly for all filter functions as can be observed in Figures B.5 and B.6. For both signal and background jets, the τ_{32} distribution peaks sharply at values close to 1, which indicates that the jets are better modeled with two CCs as subjets instead of three. The τ_{21} distribution for signal jets is bimodal, with low values corresponding to jets that are better modeled with two CCs as subjets and higher values corresponding to jets that are better modeled with a single CC. The mass distributions of the leading vertex and CC are shown in Figures B.7 and B.8, respectively. As can be observed in the distributions for signal jets, both the leading vertex and CC tend to reconstruct larger substructures in the jet when the η-projection and momentum fraction filter functions are used. Additionally, it is observed that the momentum fraction filter functions reconstruct most of the top jet at the leading vertex level, while for the η-projection a similar degree of reconstruction is achieved at the leading CC level. On the other hand, the log sigmoid ΔR tends to reconstruct smaller substructures with these objects in signal jets, while the ϕ-projection and log sigmoid E_{T} act as a compromise between small-scale and large-scale substructure reconstruction. Finally, as can be observed in Figures B.9 and B.10, the different filter functions result in varying behaviors of the energy correlation of the topoclusters that are associated with the CCs.



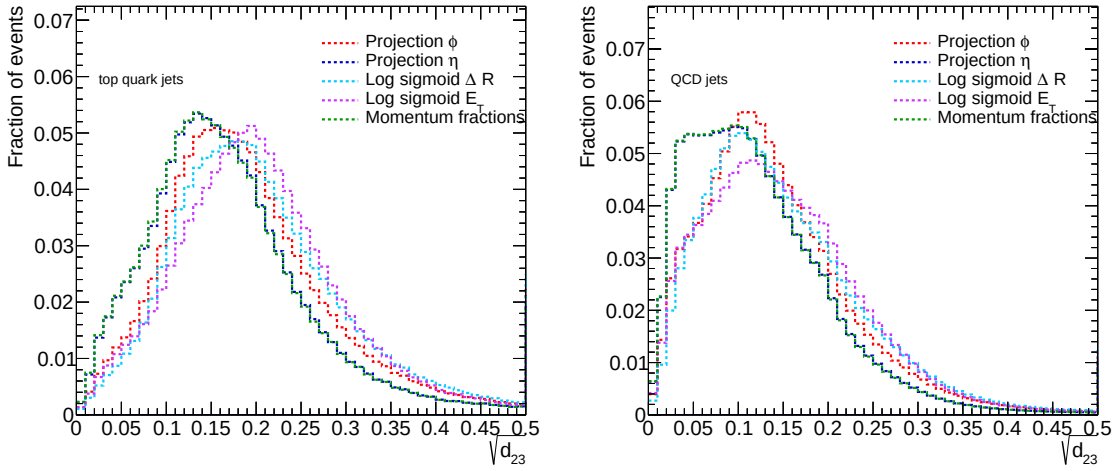
(a) (b)

Figure B.1: The number of connected components in the jet for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of filter functions.



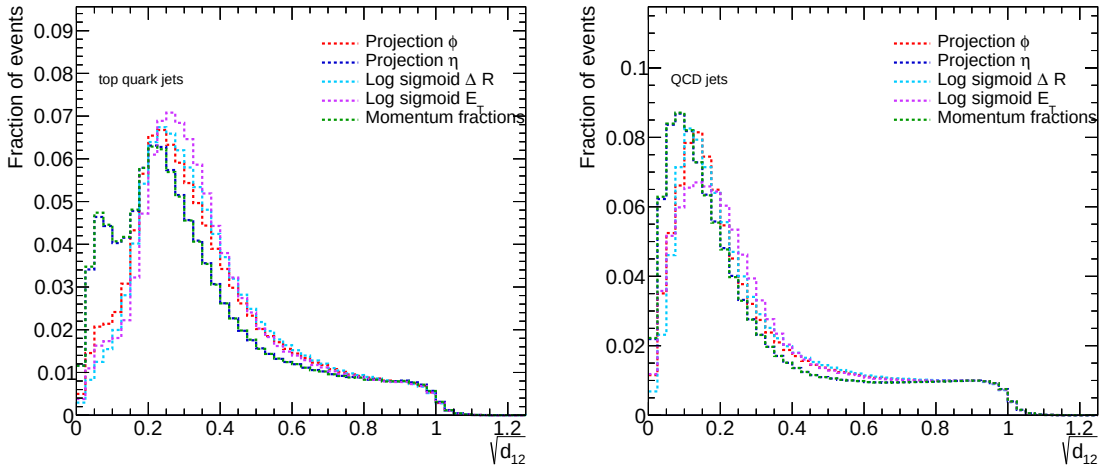
(a) (b)

Figure B.2: The Cambridge-Aachen splitting scale to three connected components for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of filter functions.



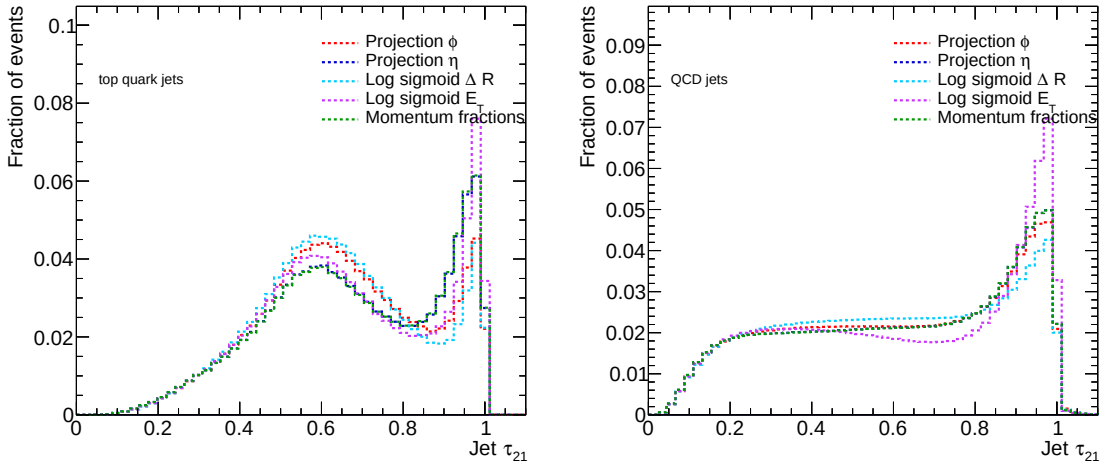
(a) (b)

Figure B.3: The Cambridge-Aachen splitting scale to two connected components for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of filter functions.



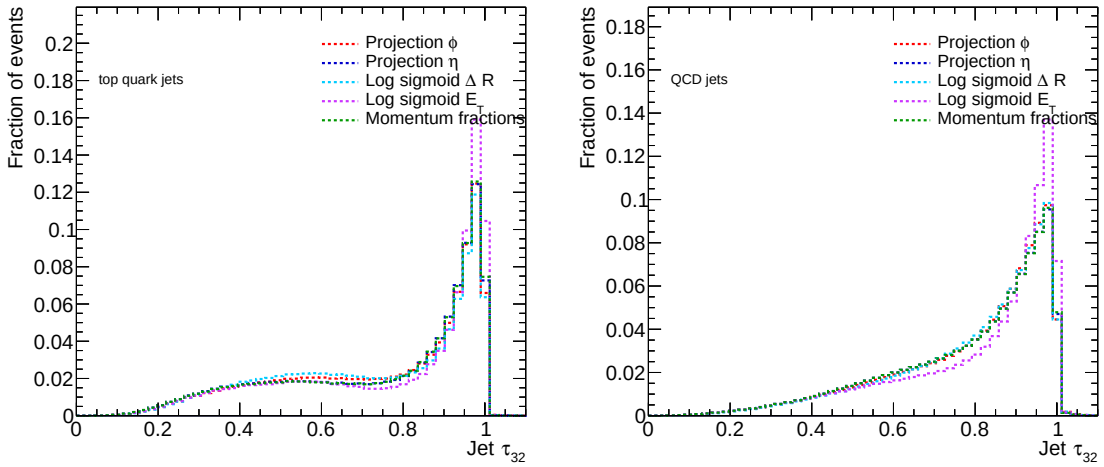
(a) (b)

Figure B.4: The Cambridge-Aachen splitting scale to one connected component for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of filter functions.



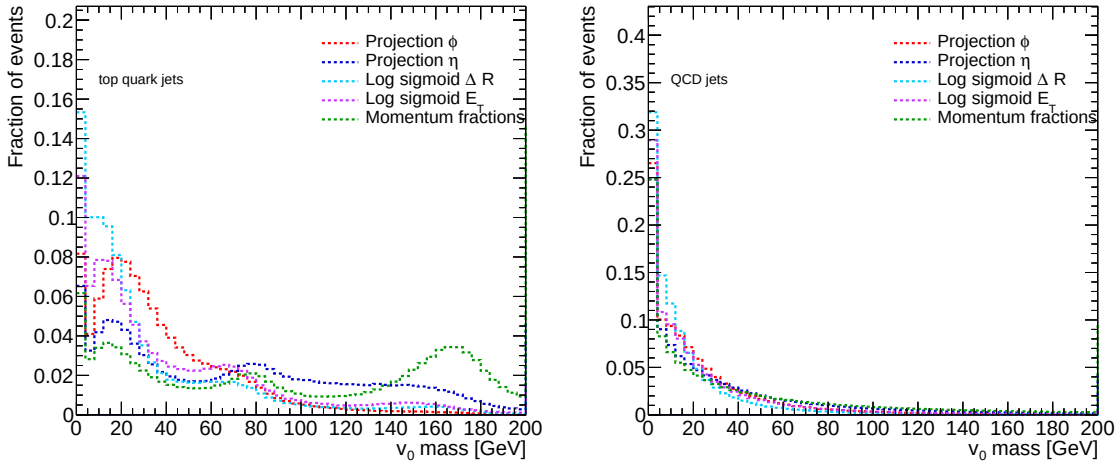
(a) (b)

Figure B.5: The n-subjettiness ratio τ_{21} for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of filter functions. The individual n-subjettiness variables are calculated by interpreting the CCs of the jet as subjets.



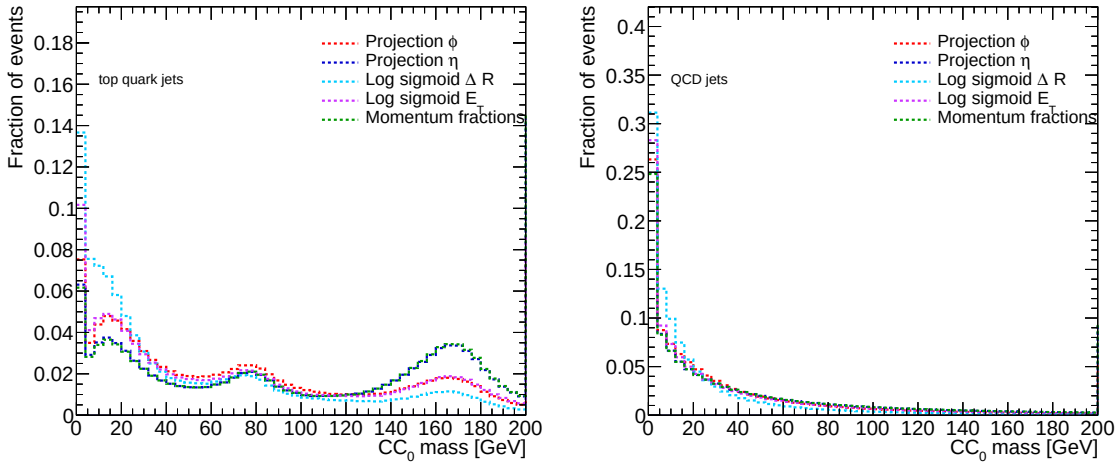
(a) (b)

Figure B.6: The n-subjettiness ratio τ_{32} for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of filter functions. The individual n-subjettiness variables are calculated by interpreting the CCs of the jet as subjets.



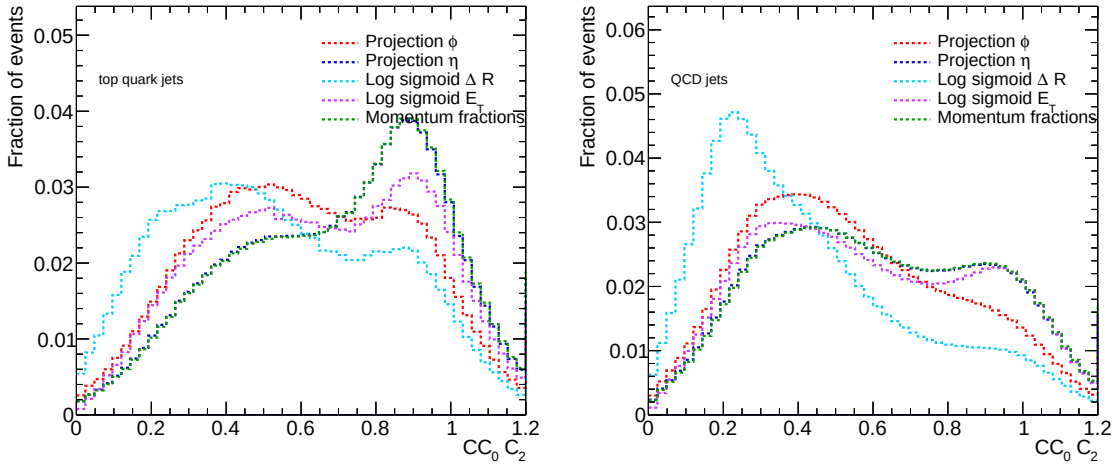
(a) (b)

Figure B.7: The mass of the leading in p_{T} vertex for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of filter functions.



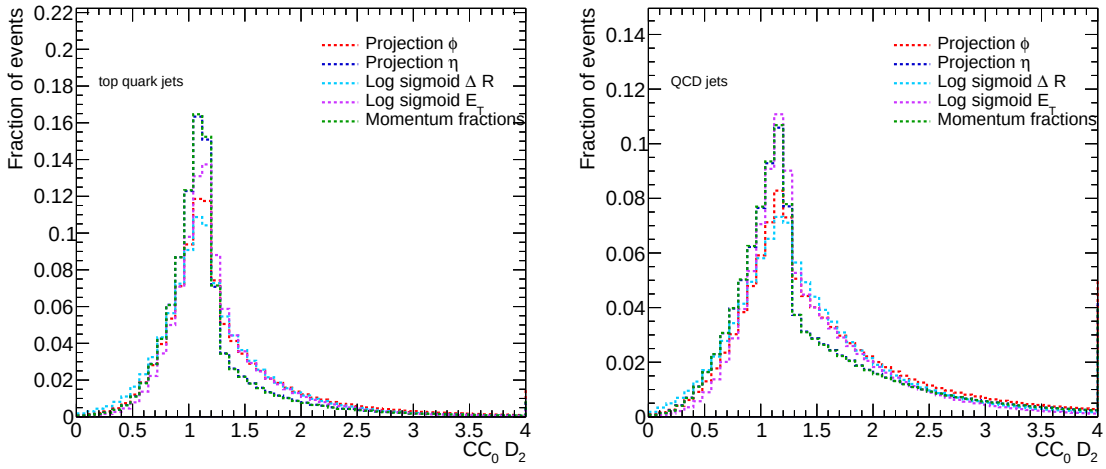
(a) (b)

Figure B.8: The mass of the leading in p_{T} connected component for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of filter functions.



(a) (b)

Figure B.9: The energy correlation function ratio C_2 of the leading in p_{T} connected component for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of filter functions. The energy correlation function ratios are evaluated using the topoclusters associated with the connected component.



(a) (b)

Figure B.10: The energy correlation function ratio D_2 of the leading in p_{T} connected component for signal top jets from $W' \rightarrow tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of filter functions. The energy correlation function ratios are evaluated using the topoclusters associated with the connected component.

Clustering Algorithm Optimization

A comparison between the different clustering algorithms that are used by the Mapper algorithm to form the vertices of the Čech complex is presented. The other parameters of the Mapper algorithm are set to their final choice in the results shown here. In addition to the single-linkage clustering algorithm, a centroid-linkage clustering and the anti-k_T clustering algorithms were considered during the optimization process of the Mapper algorithm. These two options are described below:

Centroid-Linkage Clustering

The centroid-linkage clustering algorithm defines the distance between two clusters of topoclusters v_i and v_j as:

$$D(\mathrm{v}_i, \mathrm{v}_j) = \sqrt{(\eta_i^{\mathrm{avg}} - \eta_j^{\mathrm{avg}})^2 + (\phi_i^{\mathrm{avg}} - \phi_j^{\mathrm{avg}})^2} \tag{B.5}$$

where η_k^{avg} and ϕ_k^{avg} are the average pseudorapidity and azimuthal angles of the topoclusters in a given cluster v_k. The two clusters that achieve the minimum centroid-linkage distance are merged together.

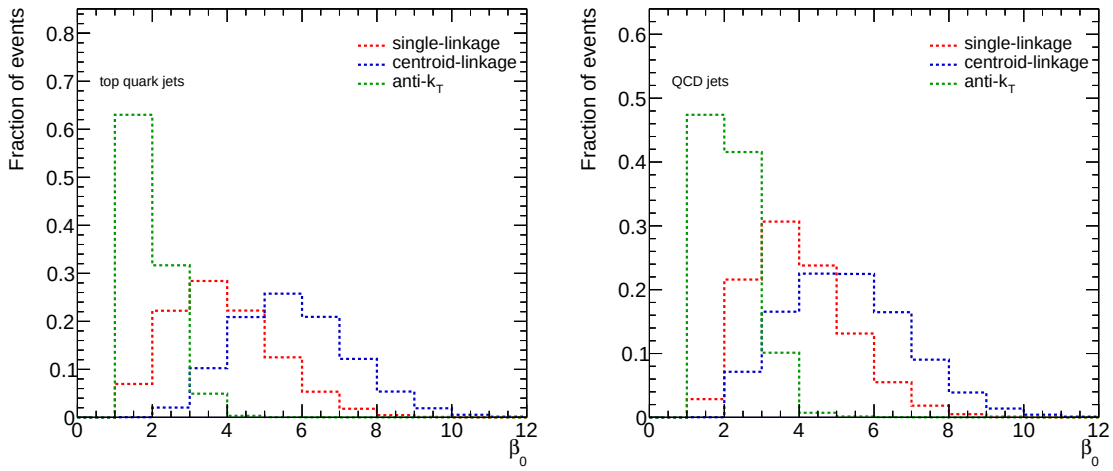
Anti-k_T Clustering

This is the standard anti-k_T clustering algorithm applied to clusters of topoclusters. The distance between clusters is defined as

$$D(\mathrm{v}_i, \mathrm{v}_j) = \min\{p_{\mathrm{T}\,i}^{-2}, p_{\mathrm{T}\,j}^{-2}\}\Delta R(\mathrm{v}_i, \mathrm{v}_j) \tag{B.6}$$

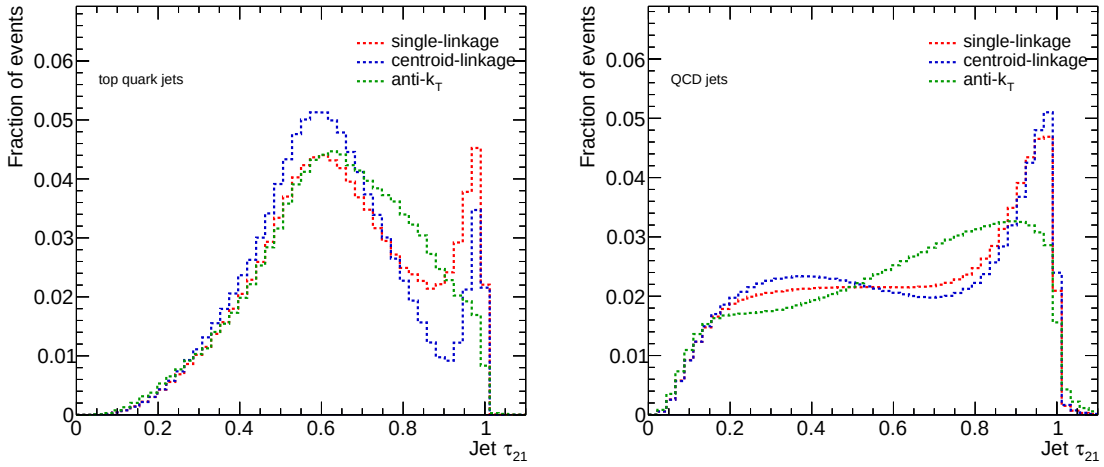
The two clusters that achieve the minimum anti-k_T distance are merged together.

As can be observed in Figure B.11, the anti-k_T clustering tends to produce fewer CCs in jets, while the centroid-linkage clustering results in more CCs. The distributions of the n-subjettiness ratios τ_{21} and τ_{32} are shown in Figures B.12 and B.13, respectively. The τ_{32} distribution indicates that both signal and background jets that have at least three CCs are better modeled with two CCs instead of three. On the other hand, when using the single-linkage and centroid-linkage clustering algorithms, the τ_{21} distribution is bimodal for signal jets, while preferring a single CC substructure for background jets. The anti-k_T clustering tends to produce CCs that better model signal jets with two CCs when compared to the other clustering options. This observation also holds for background jets; however, the relative modeling between two CCs and a single CC is more ambiguous compared to the other n-subjettiness ratio distributions. The mass distributions of the leading vertex and CC are shown in Figures B.14 and B.15, respectively. These two objects tend to reconstruct smaller substructures in signal jets when the centroid-linkage is used. On the other hand, the anti-k_T clustering reconstructs most of the substructure in signal jets with these two objects. This can be observed from the bump in the leading vertex mass distribution near the W boson mass and the prominent peak in the leading CC mass distribution near the top quark mass. No significant differences are observed in the mass distributions of these objects for background jets between the single-linkage and centroid-linkage clustering algorithms. On the other hand, the anti-k_T clustering tends to produce leading vertices and CCs in background jets that have more mass compared to the other clustering options. This could be explained by the behavior of the anti-k_T algorithm, which clusters the most energetic objects first. Finally, the different clustering algorithms tend to produce CCs with varying behaviors in their topocluster substructure, as can be observed from the energy correlation function ratios in Figures B.16 and B.17.



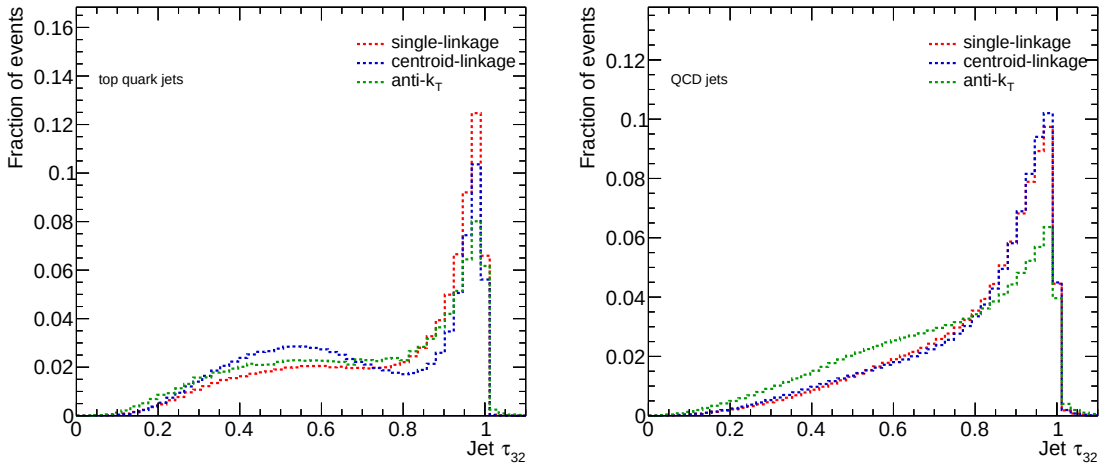
(a) (b)

Figure B.11: The number of connected components in the jet for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of clustering algorithms.



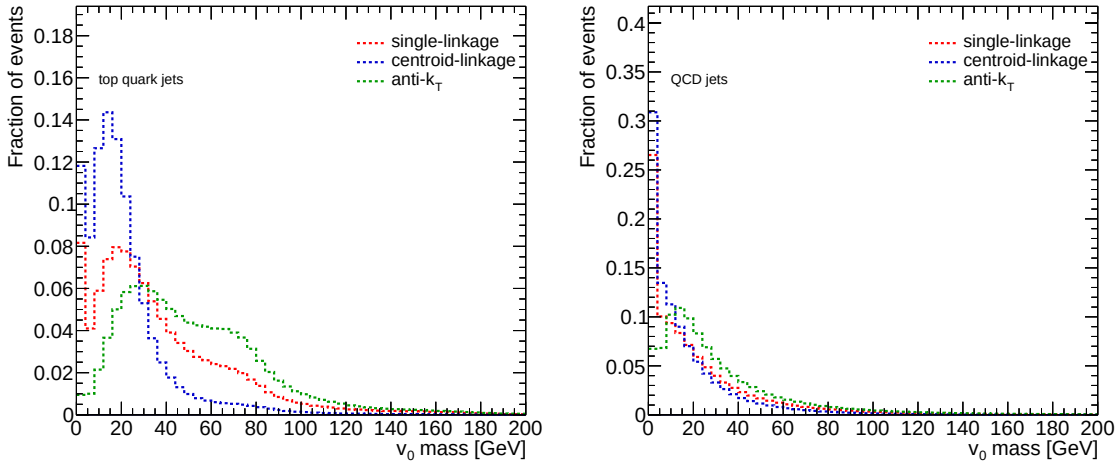
(a) (b)

Figure B.12: The n-subjettiness ratio τ_{21} for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of clustering algorithms. The individual n-subjettiness variables are calculated by interpreting the CCs of the jet as subjets.



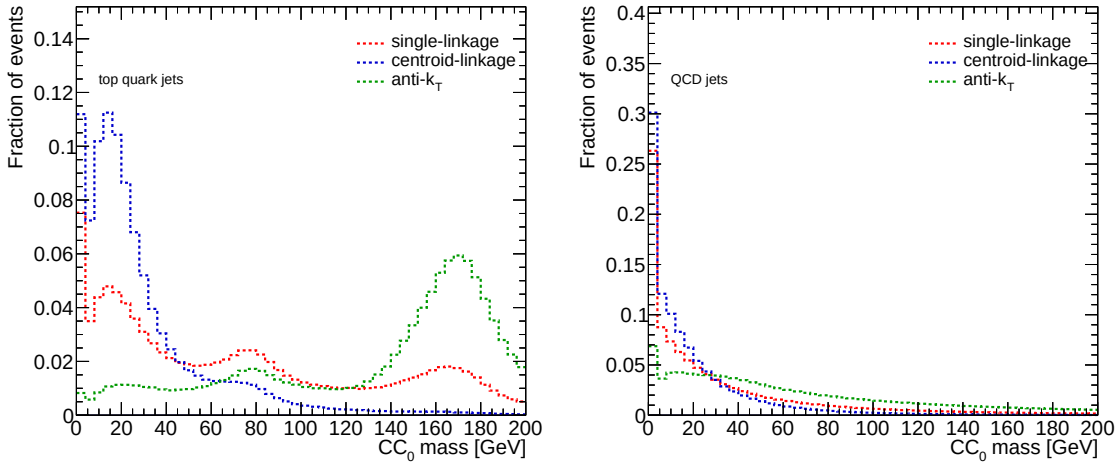
(a) (b)

Figure B.13: The n-subjettiness ratio τ_{32} for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of clustering algorithms. The individual n-subjettiness variables are calculated by interpreting the CCs of the jet as subjets.



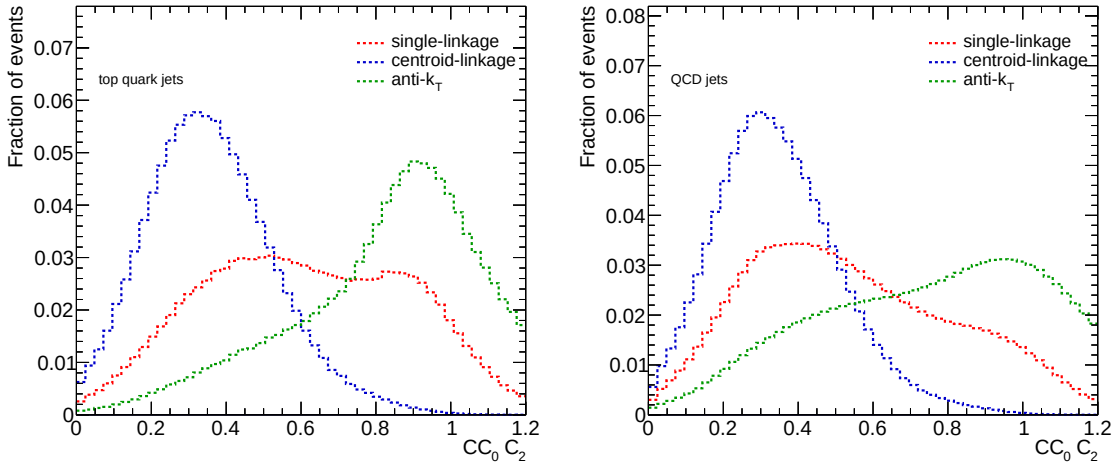
(a) (b)

Figure B.14: The mass of the leading in p_{T} vertex for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of clustering algorithms.



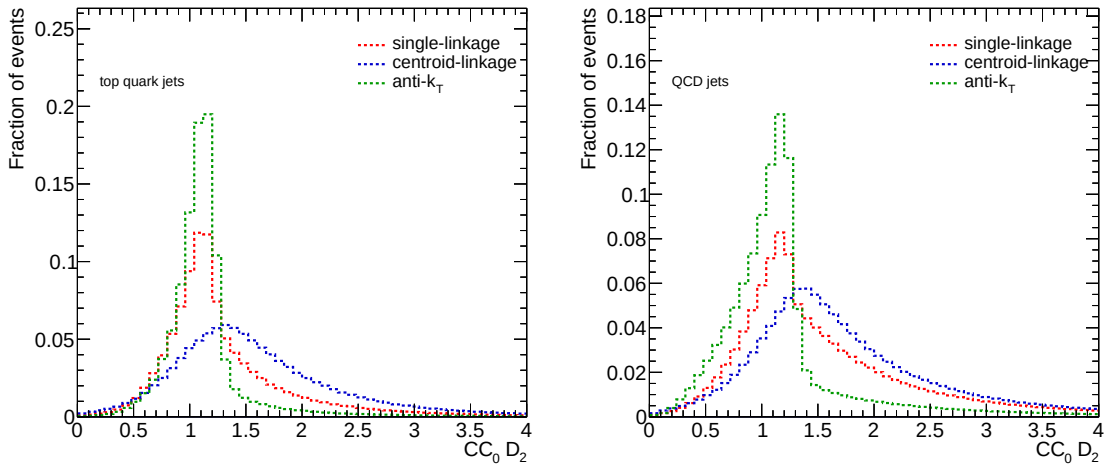
(a) (b)

Figure B.15: The mass of the leading in p_{T} connected component for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of clustering algorithms.



(a) (b)

Figure B.16: The energy correlation function ratio C_2 of the leading in p_{T} connected component for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of clustering algorithms. The energy correlation function ratios are evaluated using the topoclusters associated with the connected component.



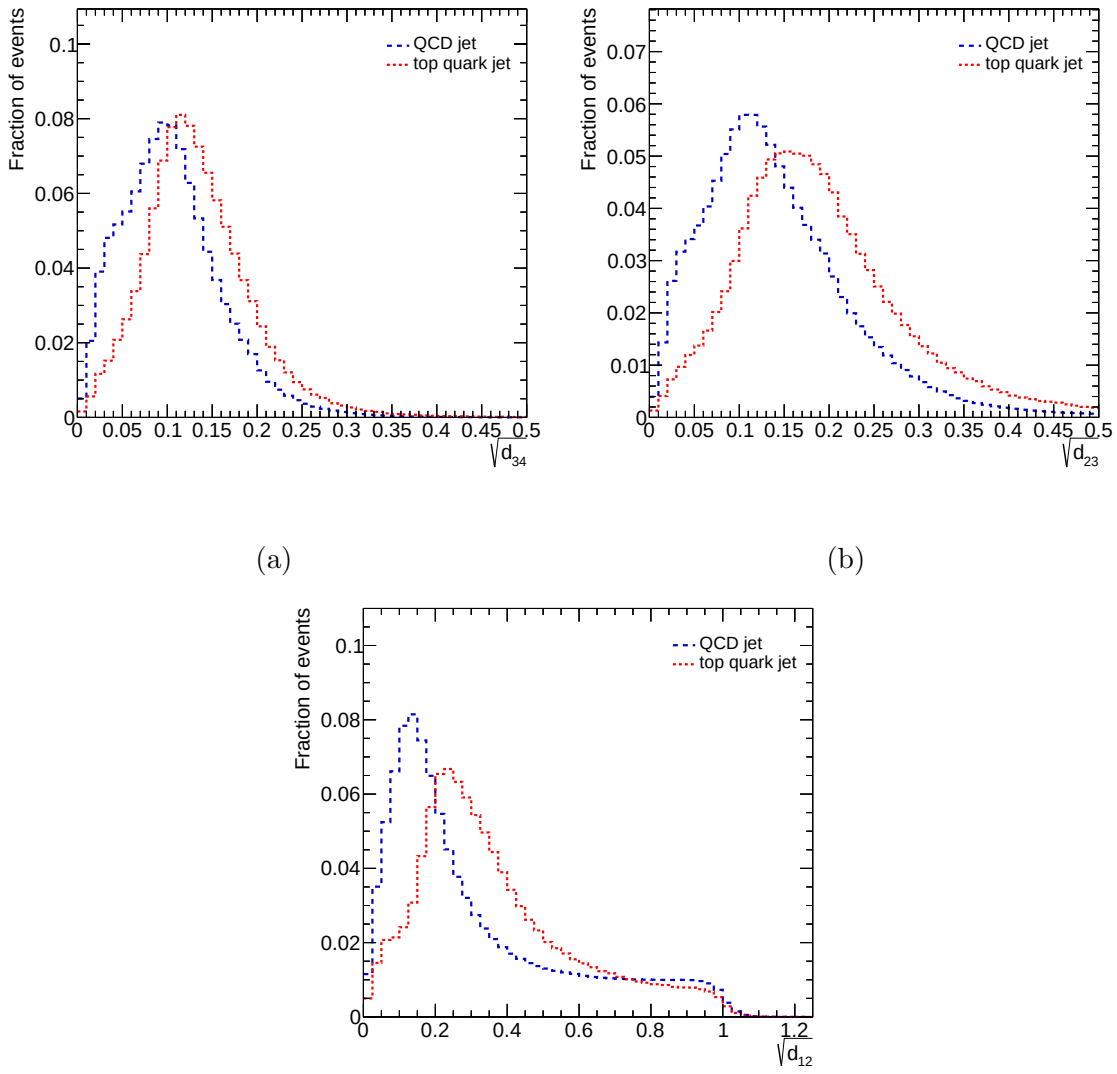
(a) (b)

Figure B.17: The energy correlation function ratio D_2 of the leading in p_{T} connected component for signal top jets from $W' \to tb$ processes (a) and background jets from QCD processes (b) overlayed between the different options of clustering algorithms. The energy correlation function ratios are evaluated using the topoclusters associated with the connected component.

Appendix C

Mapper Algorithm Comparison Plots

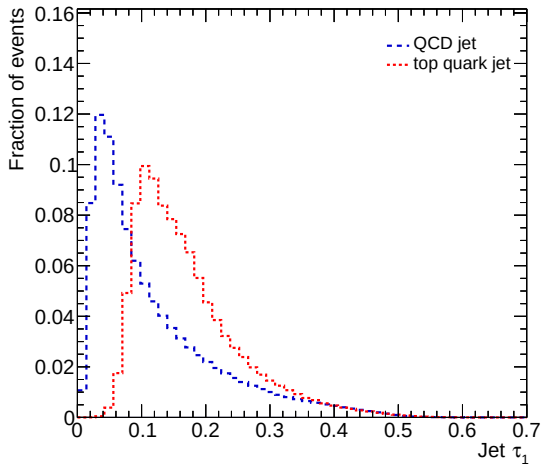
As discussed in subsection 5.2.3, variables that are inspired by the jet substructure observables were defined using the information obtained from the topological data analysis of jets. The vertices and connected components (CCs) of the Čech complex of jets obtained from the Mapper algorithm are interpreted as subjets. This is achieved by adding the four-momenta of the topoclusters that are associated with these objects. This allows us to use vertices and CCs as inputs to the jet substructure observables to quantify how the energy of a jet is distributed across structures formed by these objects. Additionally, some of these substructure observables were also defined for the CCs in jets by using the topoclusters associated with a given CC as inputs when evaluating these variables. This allows us to quantify how the energy is distributed in the substructures that are reconstructed by CCs. As discussed in subsection 5.2.4, some of these variables are used as inputs to the DNN and GNN taggers that were trained to classify jets as either signal top jets or background QCD jets. This appendix contains plots comparing the distributions of these variables between signal and background jets.



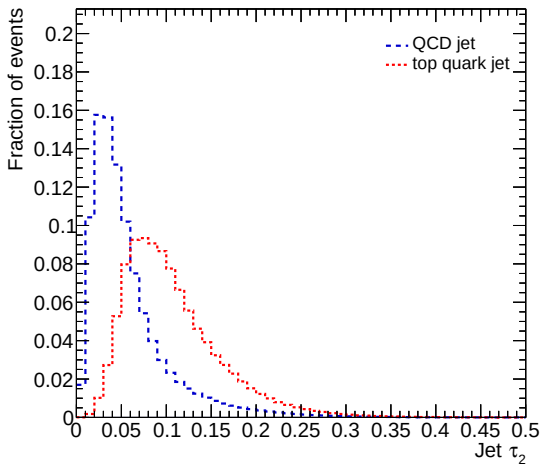
(a) (b)

(c)

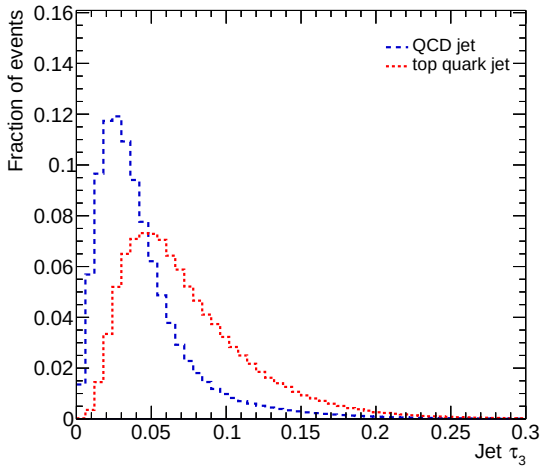
Figure C.1: Cambridge-Aachen splitting scales of jets to three connected components (a), two connected components (b), and one connected component (c).



(a)

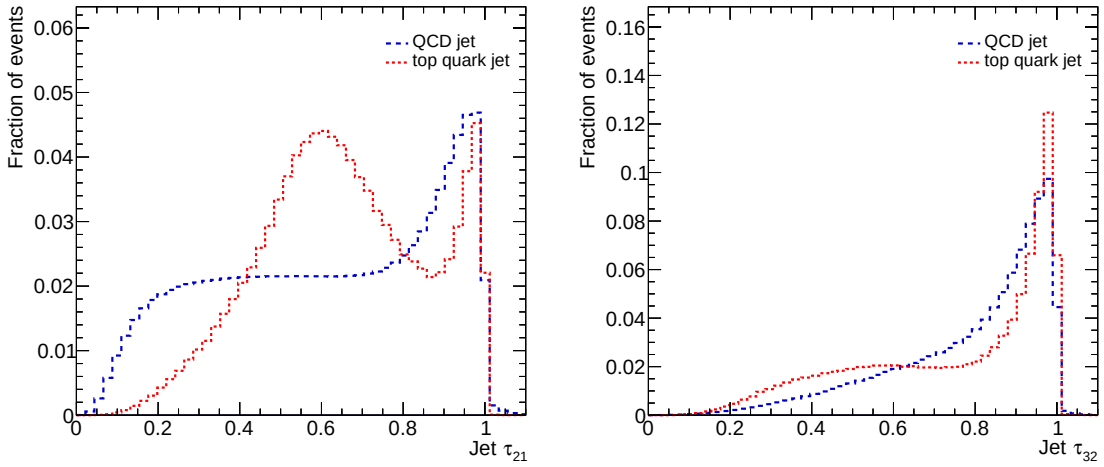


(b)



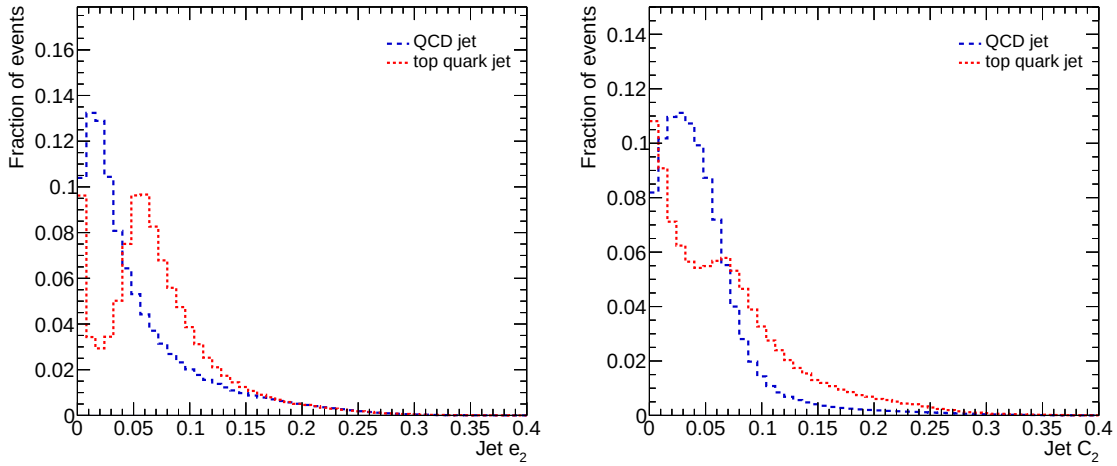
(c)

Figure C.2: n-subjettiness distributions τ_1 (a), τ_2 (b), and τ_3 (c) using the connected components as the subjets of the jet.

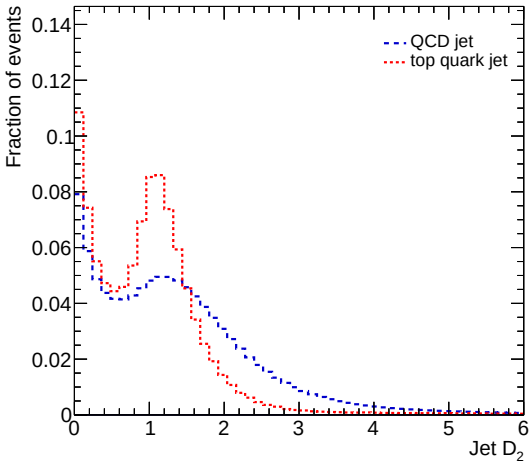


(a) (b)

Figure C.3: n-subjettiness ratios $\tau_{2\,1} = \tau_2/\tau_1$ (a) and $\tau_{3\,2} = \tau_3/\tau_2$ (b) using the connected components as the subjets of the jet.

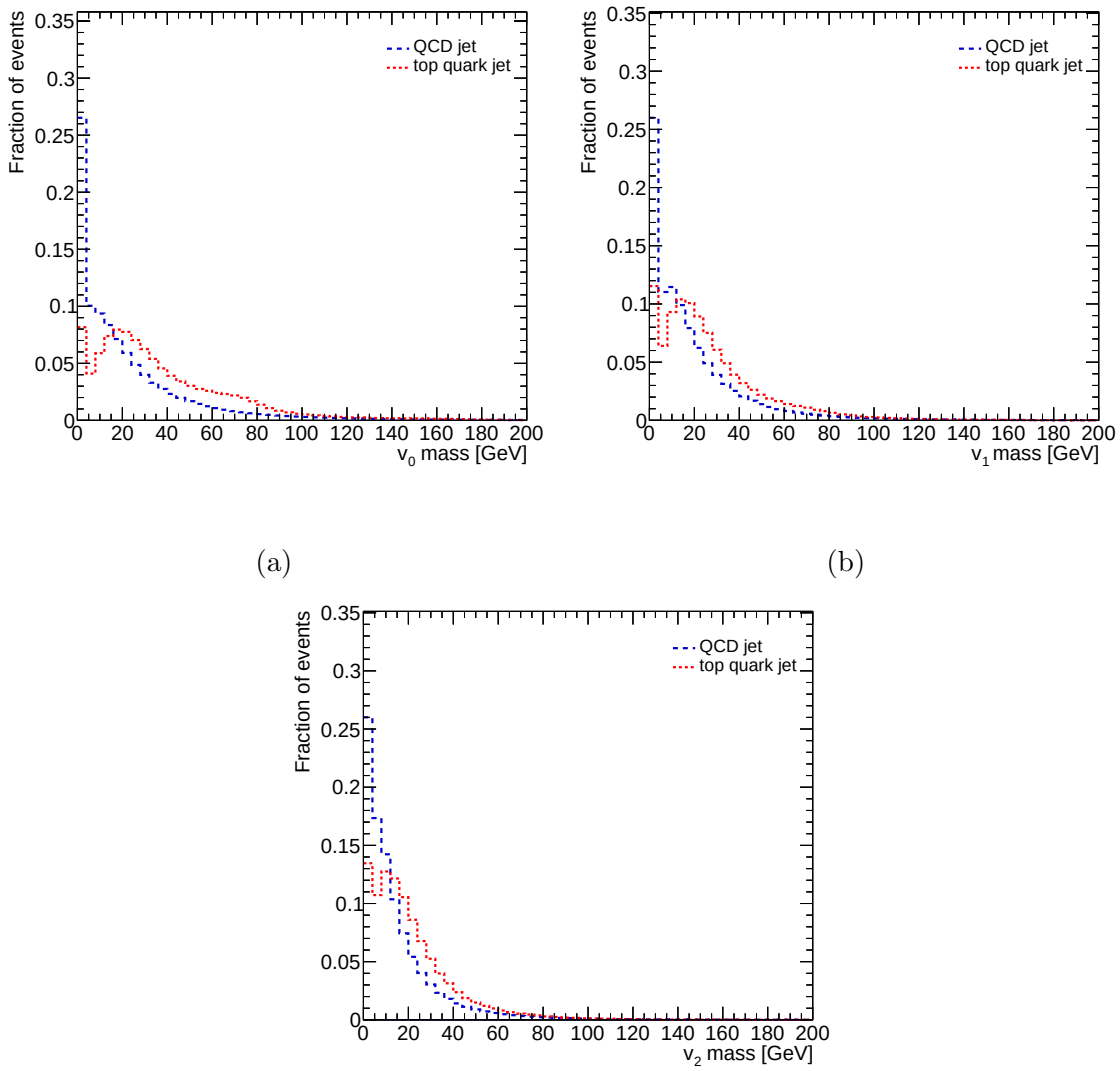


(a) (b)



(c)

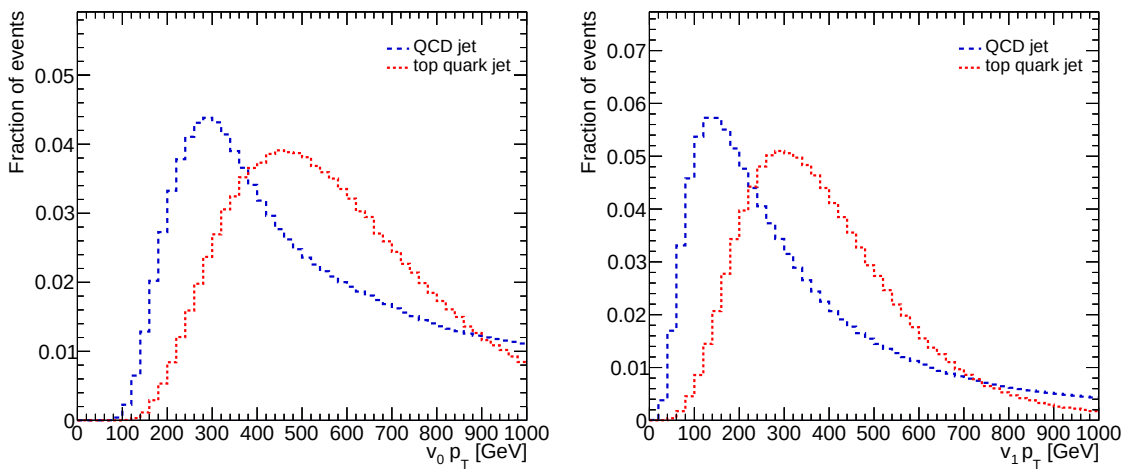
Figure C.4: n-point energy correlation function e_2 (a) and the ratios $C_2 = e_3/e_2^2$ (b) and $D_2 = e_3/e_2^3$ (c) using the connected components as the constituents of the jet.



(a) (b)

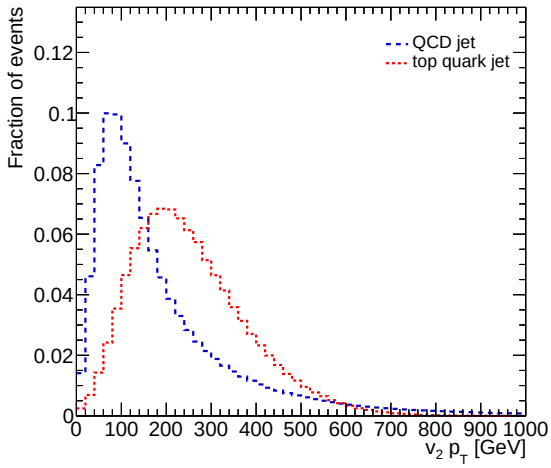
(c)

Figure C.5: Mass distributions of the leading (a), second leading (b), and third leading (c) in p_{T} vertices obtained from the Mapper algorithm by clustering topoclusters in the cover elements of the filter function image space.



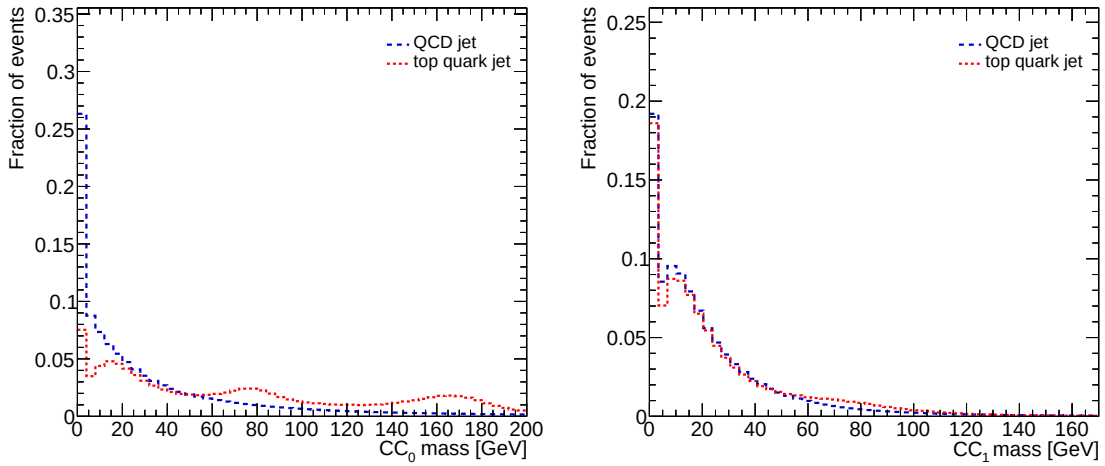
(a)

(b)

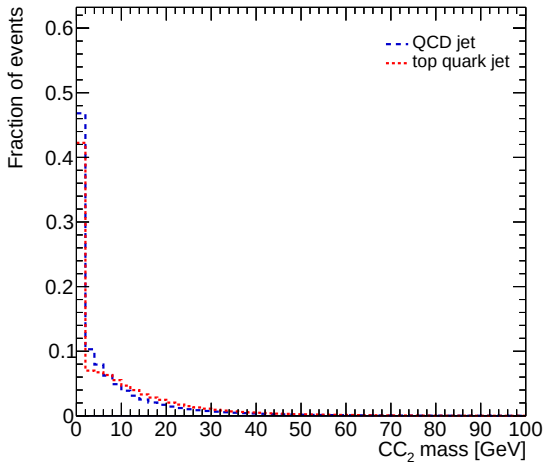


(c)

Figure C.6: p_{T} distributions of the leading (a), second leading (b), and third leading (c) in p_{T} vertices obtained from the Mapper algorithm by clustering topoclusters in the cover elements of the filter function image space.

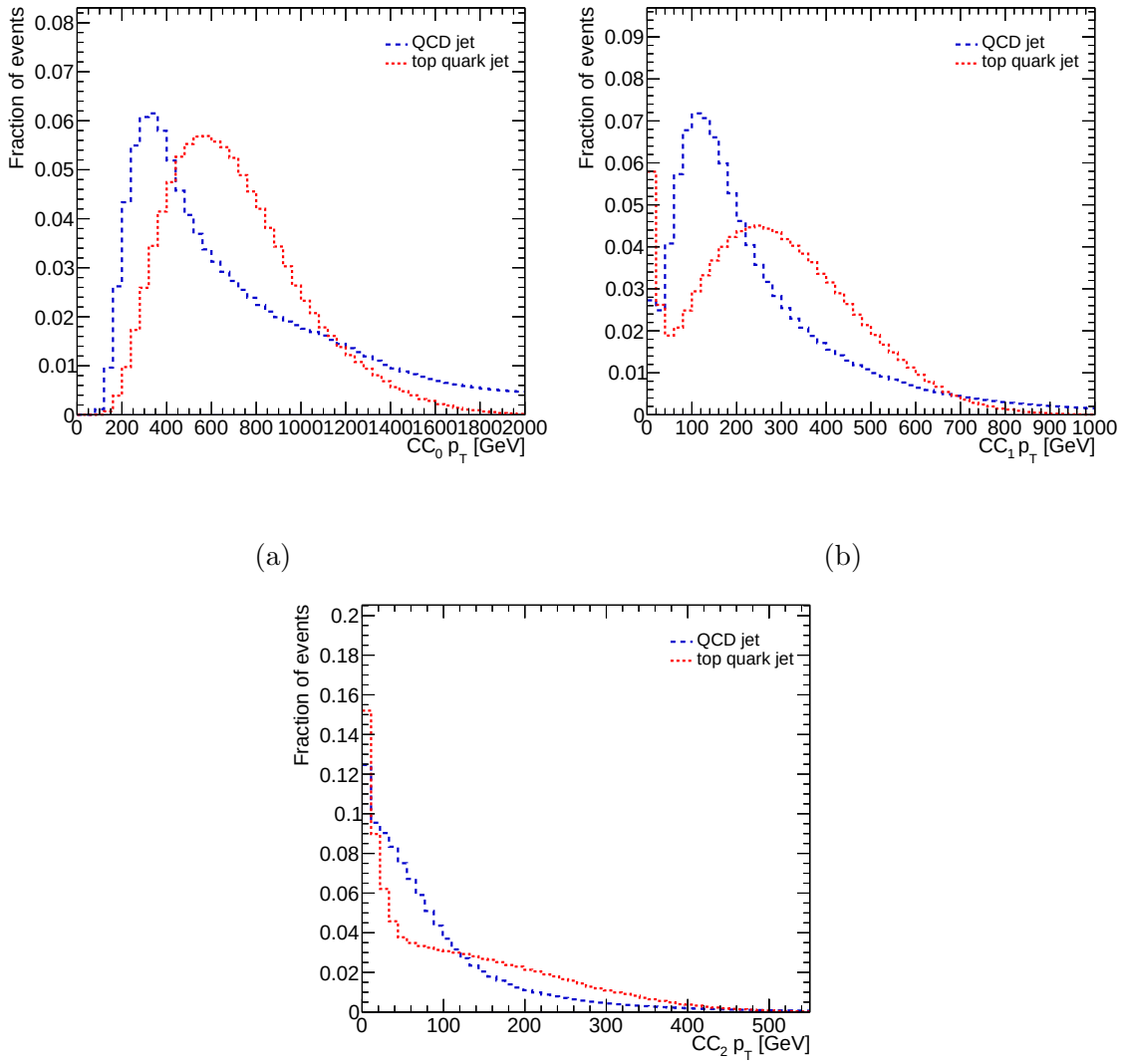


(a) (b)



(c)

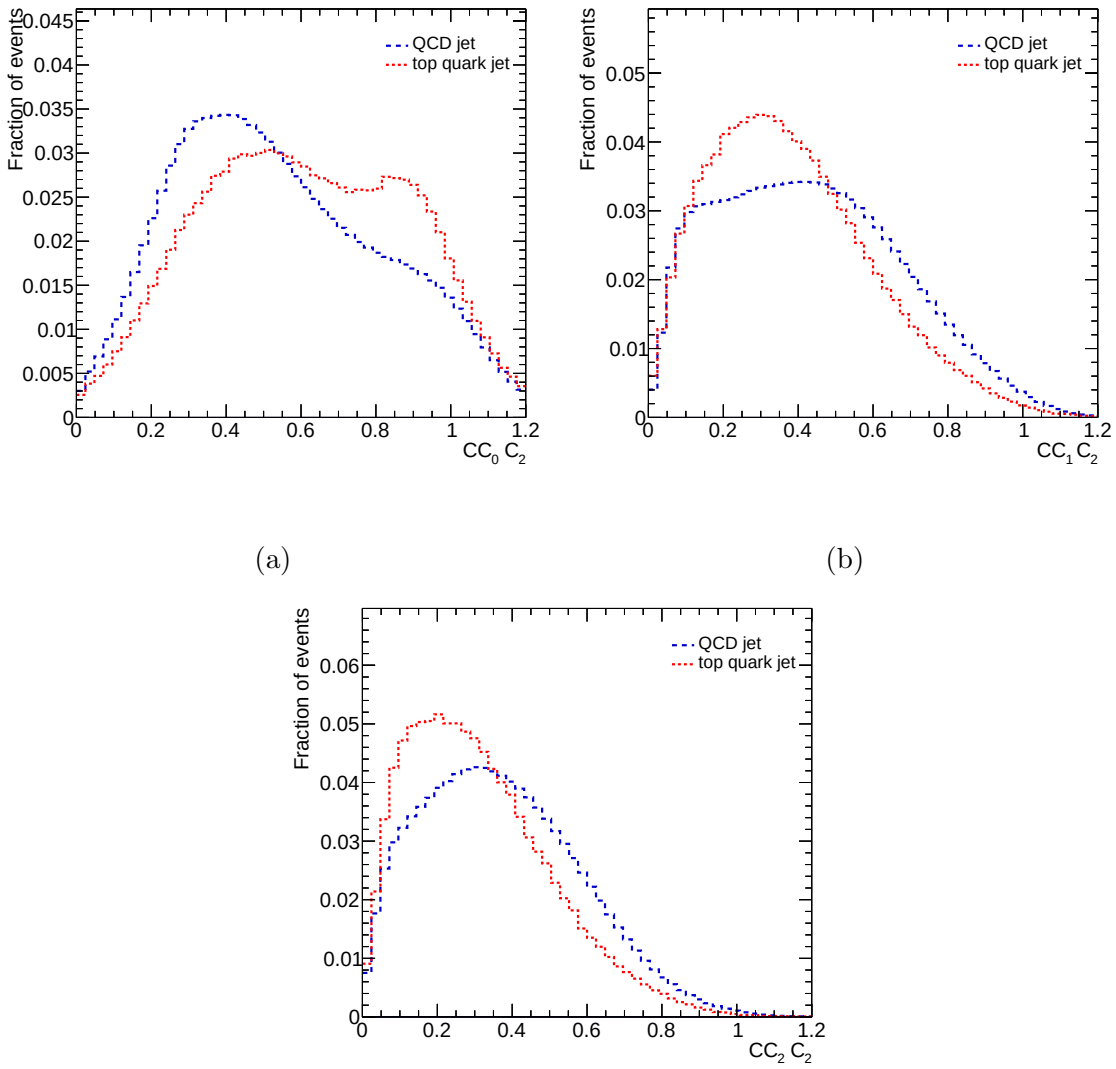
Figure C.7: Mass distributions of the leading (a), second leading (b), and third leading (c) in p_{T} connected components obtained from the Mapper algorithm by adding the four-momenta of the topoclusters associated with the connected component.



(a) (b)

(c)

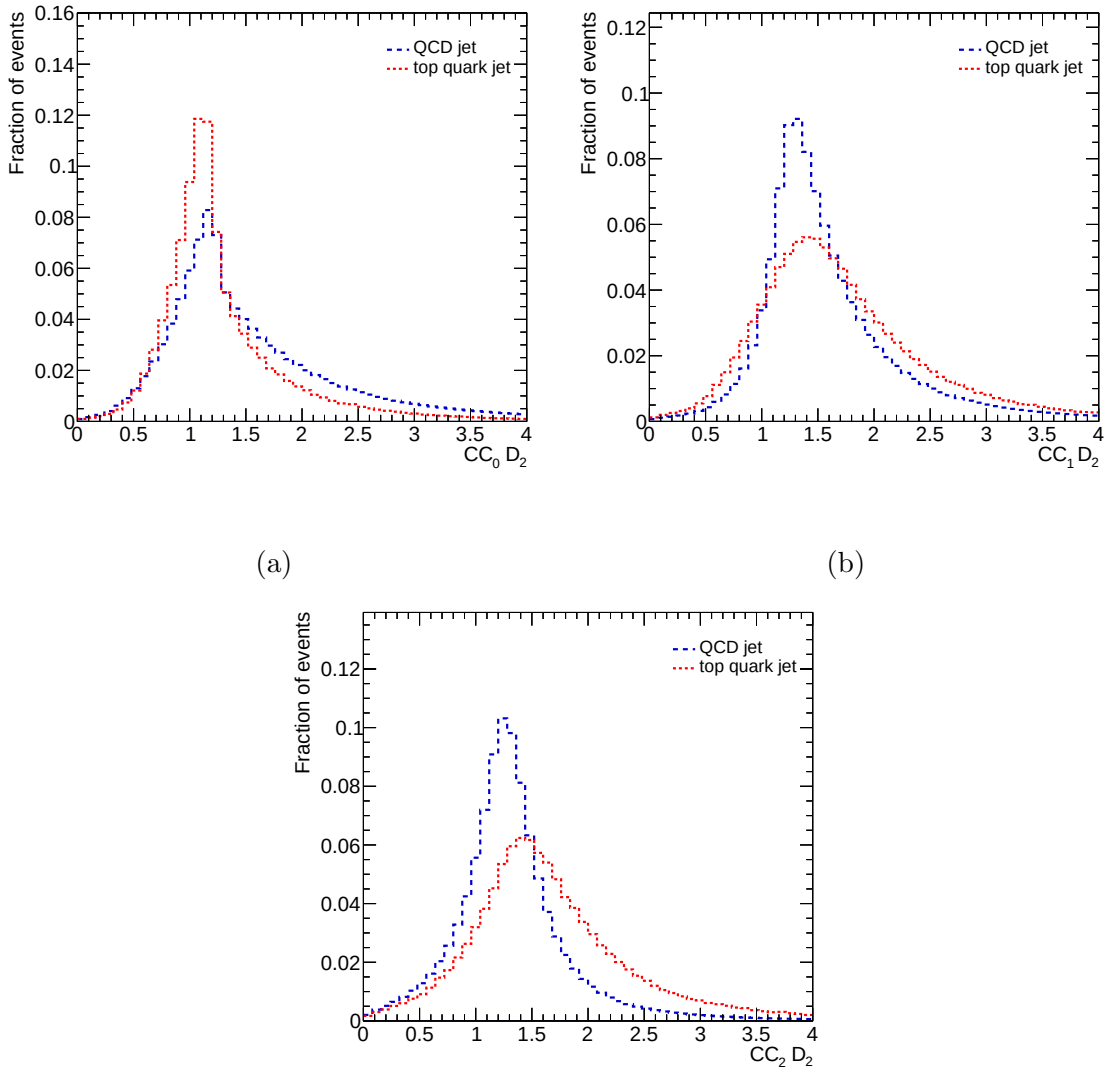
Figure C.8: p_{T} distributions of the leading (a), second leading (b), and third leading (c) in p_{T} connected components obtained from the Mapper algorithm by adding the four-momenta of the topoclusters associated with the connected component.



(a) (b)

(c)

Figure C.9: Energy correlation function ratio $C_2 = e_3/e_2^2$ distributions of the leading (a), second leading (b), and third leading (c) in p_{T} connected components obtained from the Mapper algorithm.



(a) (b)

(c)

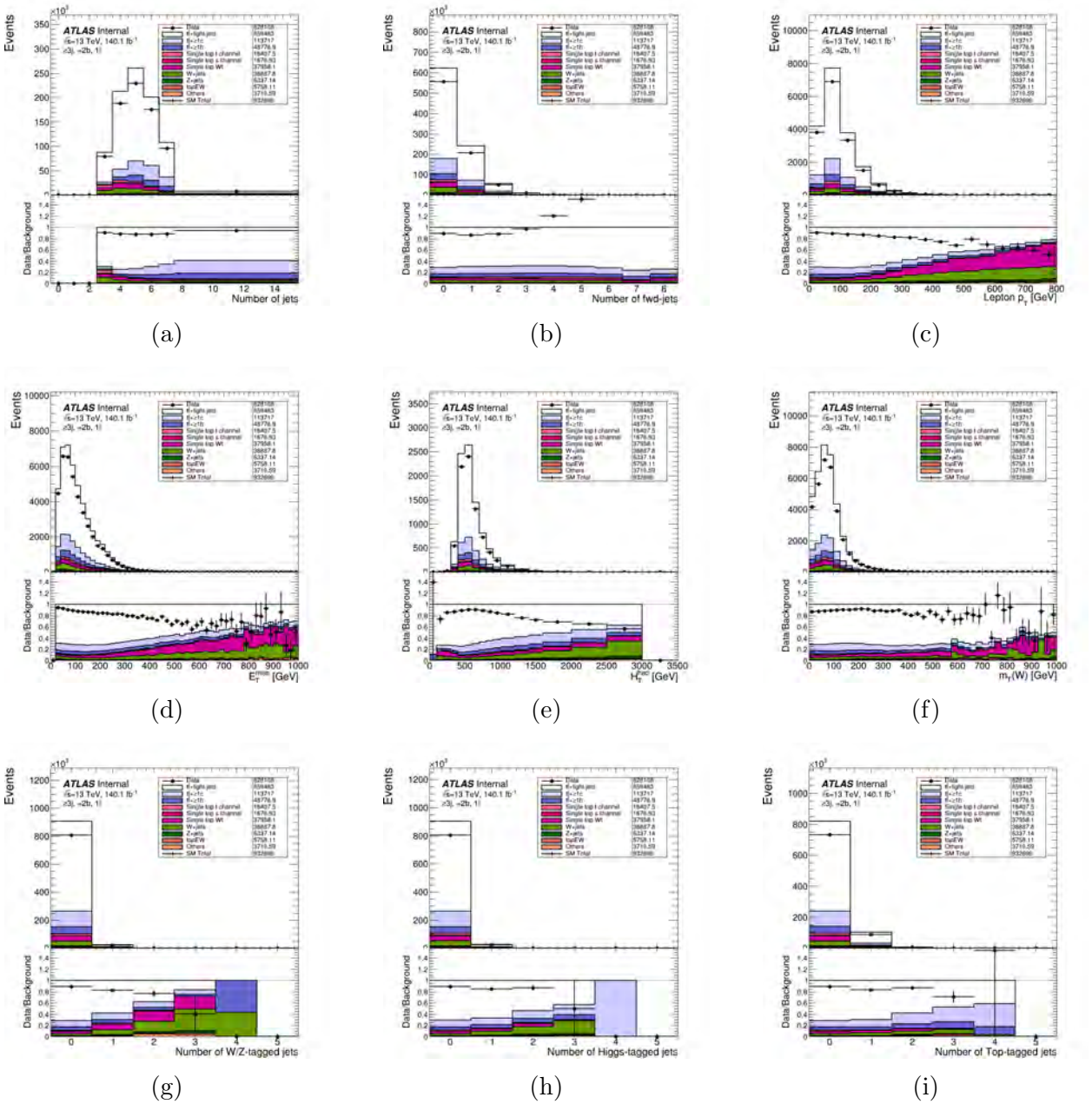
Figure C.10: Energy correlation function ratio $D_2 = e_3/e_2^3$ distributions of the leading (a), second leading (b), and third leading (c) in p_{T} connected components obtained from the Mapper algorithm.

Appendix D

Single VLQ Background Reweighting

Background Reweighting Kinematic Comparisons

This appendix contains plots that compare kinematic distributions before and after applying the background correction factors that were derived in the single production of a vector-like T analysis, as discussed in subsection 6.1.5. The selection of kinematic distributions that are shown is varied and highlights the applicability of the reweighting procedure to other kinematic variables that are not related to the variables that are used in the derivation of the correction factors. The distributions are shown at the $t\bar{t} + Wt$ reweighting source region before applying the correction factors in Figure D.1 and after applying the correction factors in Figure D.2. Additionally, these kinematic distributions are also shown in a region that requires exactly one b-tagged jet. This region is orthogonal to the $t\bar{t} + Wt$ region by definition and is used to validate the full reweighting procedure. Figures D.3 and D.4 show the distributions before and after applying the correction factors, respectively. As can be observed, the modeling of the MC simulation improves significantly in this validation region after applying all correction factors, which gives confidence in the overall background reweighting procedure.

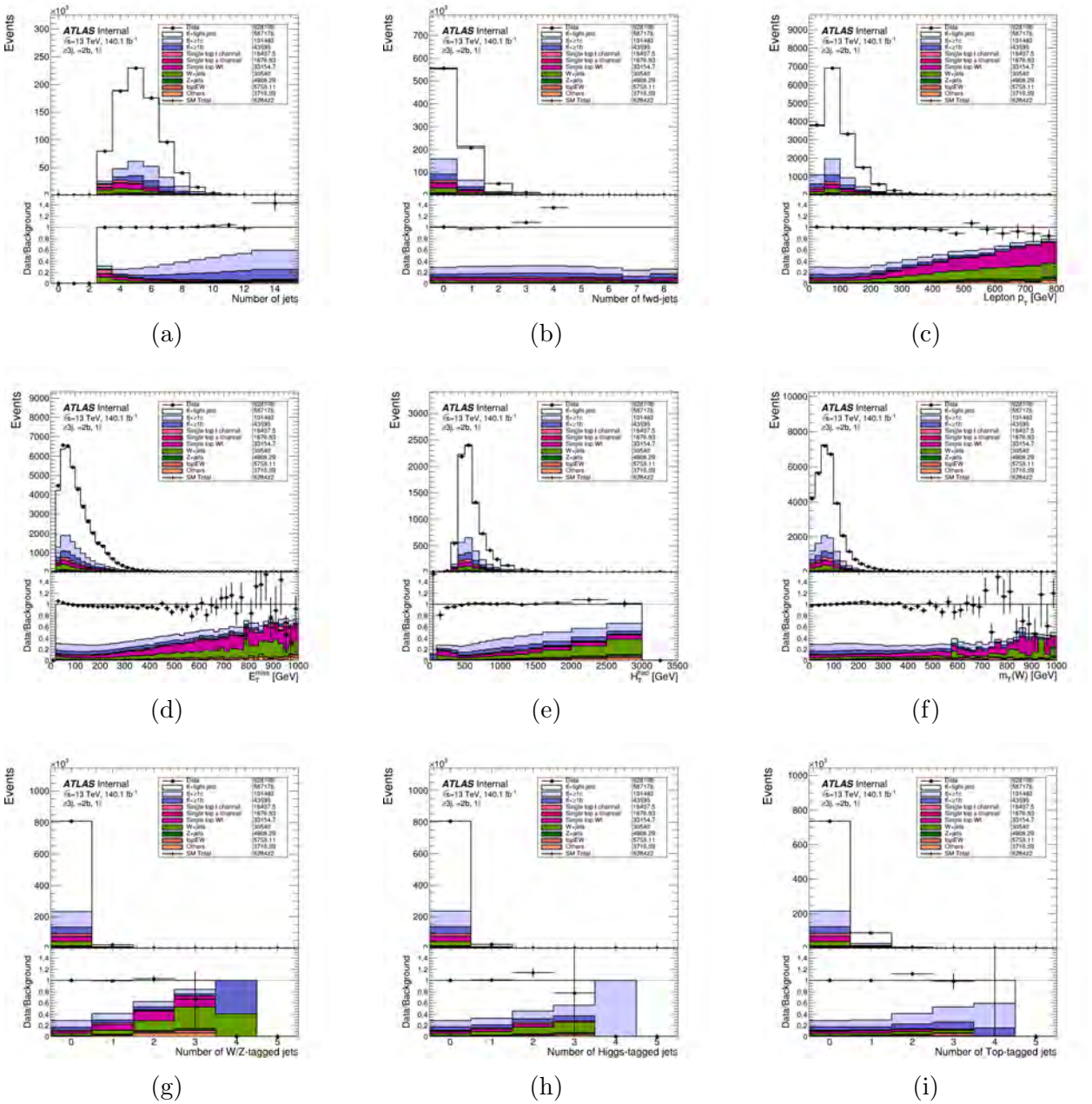


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure D.1: Comparison between data and unreweighted MC prediction in the preselection region with 1 lepton, at least 3 jets, and 2 b-tagged jets. From top to bottom and left to right, the variables displayed are: number of jets, number of forward jets, leading lepton p_{T}, $E_{\mathrm{T}}^{\mathrm{miss}}$, $H_{\mathrm{T}}^{\mathrm{had}}$, m_{T}^{W}, number of V-tagged jets, number of H-tagged jets, number of top-tagged jets.

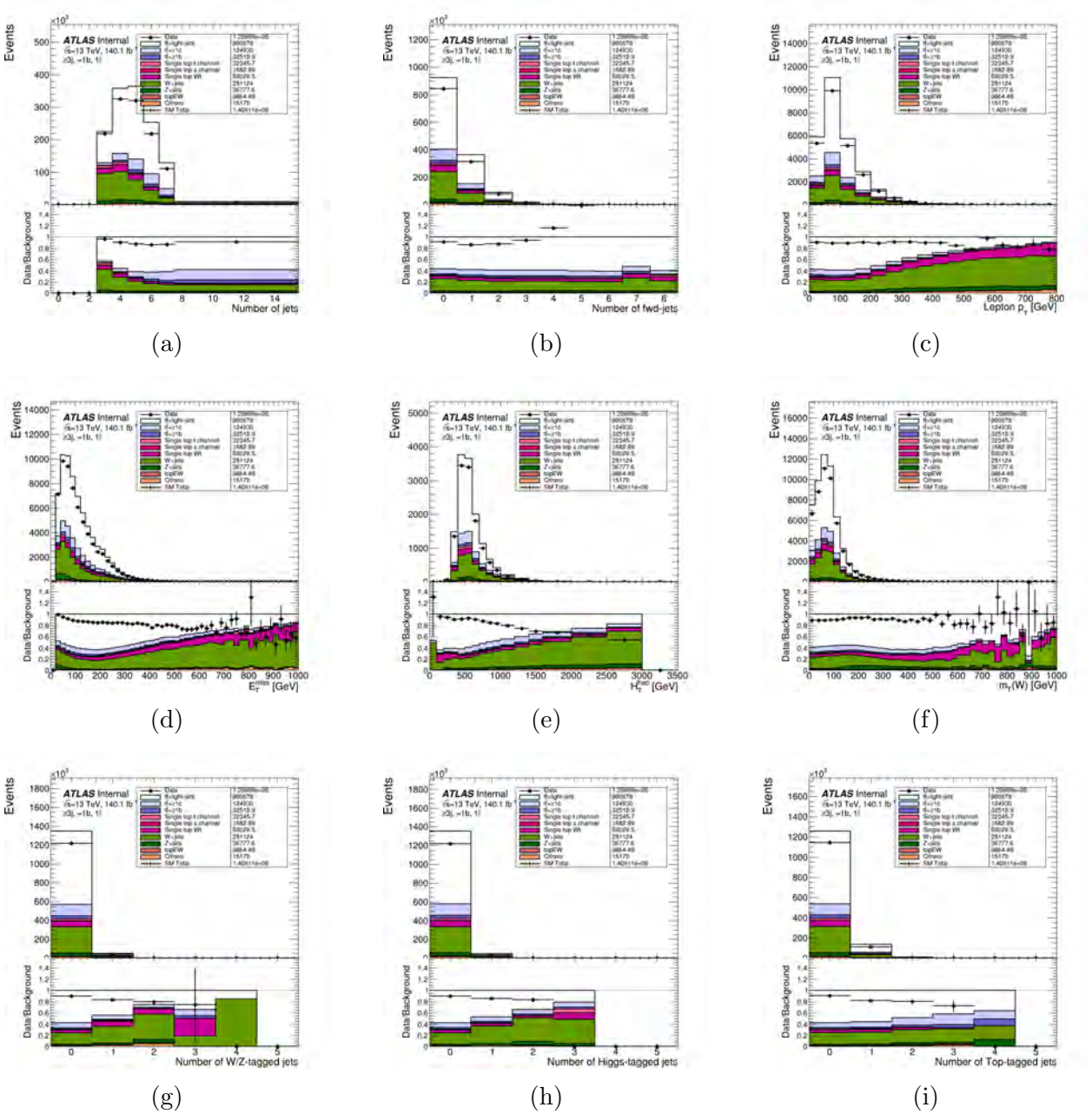


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure D.2: Comparison between data and fully reweighted MC prediction in the preselection region with 1 lepton, at least 3 jets, and 2 b-tagged jets. From top to bottom and left to right, the variables displayed are: number of jets, number of forward jets, leading lepton p_{T}, $E_{\mathrm{T}}^{\mathrm{miss}}$, $H_{\mathrm{T}}^{\mathrm{had}}$, m_{T}^{W}, number of V-tagged jets, number of H-tagged jets, number of top-tagged jets.

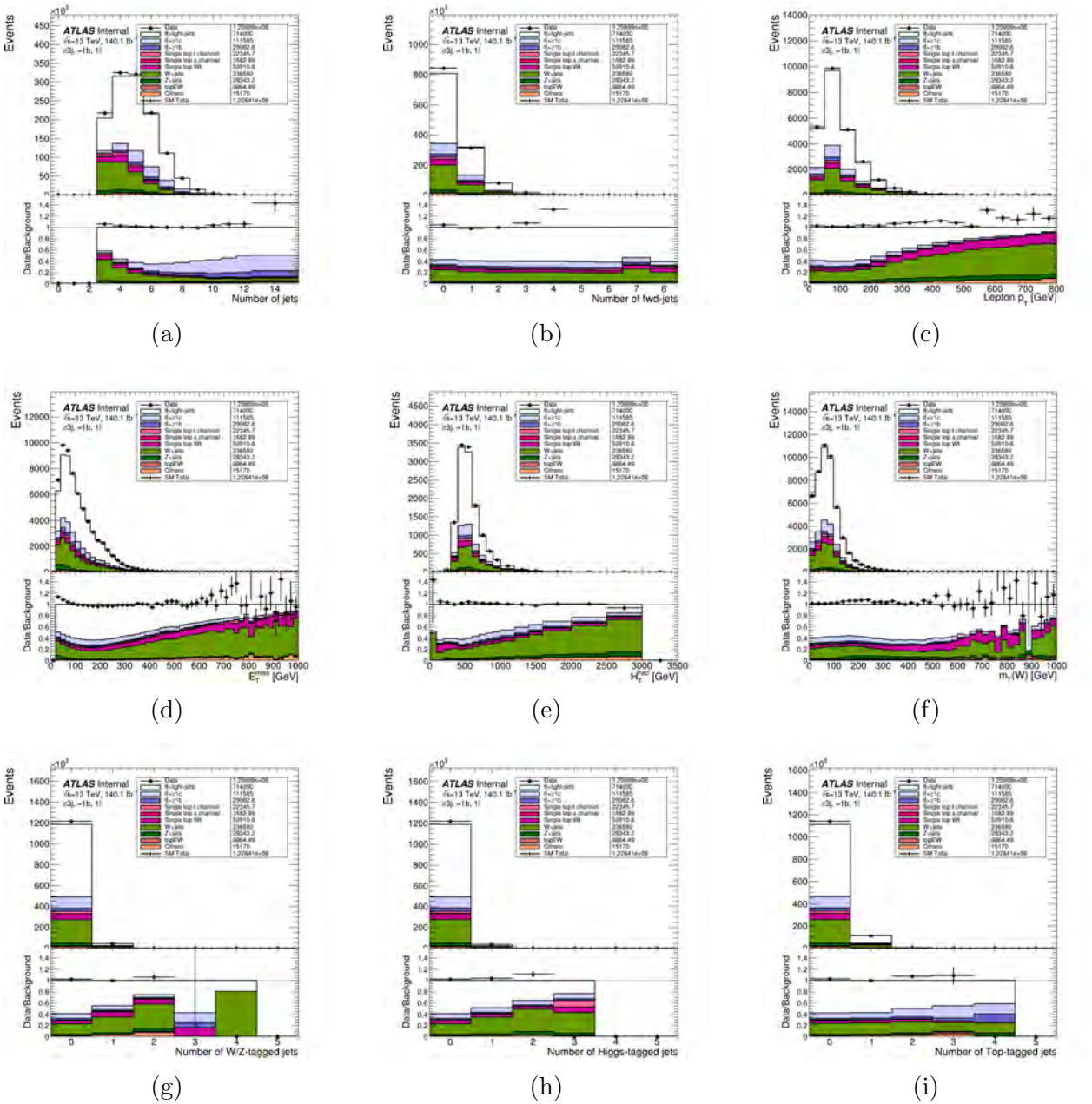


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure D.3: Comparison between data and unreweighted MC prediction in the preselection region with 1 lepton, at least 3 jets, and 1 b-tagged jets. From top to bottom and left to right, the variables displayed are: number of jets, number of forward jets, leading lepton p_{T}, $E_{\mathrm{T}}^{\mathrm{miss}}$, $H_{\mathrm{T}}^{\mathrm{had}}$, m_{T}^{W}, number of V-tagged jets, number of H-tagged jets, number of top-tagged jets.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure D.4: Comparison between data and fully reweighted MC prediction in the preselection region with 1 lepton, at least 3 jets, and 1 b-tagged jets. From top to bottom and left to right, the variables displayed are: number of jets, number of forward jets, leading lepton p_{T}, $E_{\mathrm{T}}^{\mathrm{miss}}$, $H_{\mathrm{T}}^{\mathrm{had}}$, m_{T}^{W}, number of V-tagged jets, number of H-tagged jets, number of top-tagged jets.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] G. Kane. *Modern Elementary Particle Physics: Explaining and Extending the Standard Model*. Cambridge University Press, 2017.

[2] A. Zee. *Group Theory in a Nutshell for Physicists*. In a Nutshell. Princeton University Press, 2016.

[3] Andrea Wulzer. Behind the Standard Model. In *2015 European School of High-Energy Physics*, 1 2019.

[4] Standard Model of Elementary Particles. `https://upload.wikimedia.org/wikipedia/commons/0/00/Standard_Model_of_Elementary_Particles.svg`, 2019.

[5] Maggiore Michele. *A Modern Introduction to Quantum Field Theory.* Number Vol. 12 in Oxford Master Series in Physics. OUP Oxford, 2005.

[6] Alistair Savage. Introduction to Lie Groups. `https://alistairsavage.ca/mat4144/notes/MAT4144-5158-LieGroups.pdf`, 2015.

[7] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. Experimental test of parity conservation in beta decay. *Phys. Rev.*, 105:1413–1415, Feb 1957.

[8] John Ellis. An illustration of the Higgs potential. `https://cds.cern.ch/record/1638469/plots`.

[9] K. G. Begeman, A. H. Broeils, and R. H. Sanders. Extended rotation curves of spiral galaxies: dark haloes and modified dynamics. *Monthly Notices of the Royal Astronomical Society*, 249(3):523–537, 04 1991.

[10] Edvige Corbelli and Paolo Salucci. The extended rotation curve and the dark matter halo of M33. *Monthly Notices of the Royal Astronomical Society*, 311(2):441–447, 01 2000.

[11] M. Markevitch, A. H. Gonzalez, D. Clowe, A. Vikhlinin, W. Forman, C. Jones, S. Murray, and W. Tucker. Direct Constraints on the Dark Matter Self-Interaction Cross Section from the Merging Galaxy Cluster 1E 0657-56. *The Astrophysical Journal*, 606(2):819–824, May 2004.

[12] J. A. Aguilar-Saavedra, R. Benbrik, S. Heinemeyer, and M. Pérez-Victoria. Handbook of vectorlike quarks: Mixing and single production. *Physical Review D*, 88(9), nov 2013.

[13] The ATLAS Collaboration. Search for production of vector-like quark pairs and of four top quarks in the lepton-plus-jets final state in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *Journal of High Energy Physics*, 2015(8), aug 2015.

[14] Lyndon Evans and Philip Bryant. LHC machine. *Journal of Instrumentation*, 3(08):S08001–S08001, aug 2008.

[15] The ATLAS Collaboration. The ATLAS experiment at the CERN large hadron collider. *Journal of Instrumentation*, 3(08):S08003–S08003, aug 2008.

[16] The CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08004–S08004, aug 2008.

[17] The LHCb Collaboration. The LHCb detector at the LHC. *Journal of Instrumentation*, 3(08):S08005–S08005, aug 2008.

[18] The ALICE Collaboration. The ALICE experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08002–S08002, aug 2008.

[19] The TOTEM Collaboration. The TOTEM Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3(08):S08007, aug 2008.

[20] The LHCf Collaboration. The LHCf detector at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3(08):S08006, aug 2008.

[21] James Pinfold. The MoEDAL experiment at the LHC. *EPJ Web Conf.*, 145:12002, 2017.

[22] AC Team. The four main LHC experiments. `http://cds.cern.ch/record/40525`, 1999.

[23] AC Team. Diagram of an LHC dipole magnet. Schema d'un aimant dipole du LHC. `https://cds.cern.ch/record/40524`, 1999.

[24] Christiane Lefevre. The CERN accelerator complex. Complexe des accelerateurs du CERN. `http://cds.cern.ch/record/1260465`, 2008.

[25] The ATLAS Collaboration. Luminosity determination in pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector at the LHC. *The European Physical Journal C*, 71(4), apr 2011.

[26] ATLAS Collaboration. Public ATLAS Luminosity Results for Run-2 of the LHC. `https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2`.

[27] The ATLAS Collaboration. Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV with the atlas detector at the lhc. *Phys. Rev. Lett.*, 117:182002, Oct 2016.

[28] Joao Pequenao. Computer generated image of the whole ATLAS detector. `https://cds.cern.ch/record/1095924`, 2008.

[29] Joao Pequenao and Paul Schaffner. How ATLAS detects particles: diagram of particle paths in the detector. `https://cds.cern.ch/record/1505342`, 2013.

[30] The ATLAS Collaboration. *ATLAS inner detector: Technical Design Report, 1.* Technical design report. ATLAS. CERN, Geneva, 1997.

[31] Joao Pequenao. Computer generated image of the ATLAS inner detector. `https://cds.cern.ch/record/1095926`, 2008.

[32] M Capeans, G Darbo, K Einsweiller, M Elsing, T Flick, M Garcia-Sciveres, C Gemme, H Pernegger, O Rohne, and R Vuillermet. ATLAS Insertable B-Layer Technical Design Report. Technical report, ATLAS Collaboration, 2010.

[33] Joao Pequenao. Computer Generated image of the ATLAS calorimeter. `https://cds.cern.ch/record/1095927`, 2008.

[34] Joao Pequenao. Computer generated image of the ATLAS Muons subsystem. `https://cds.cern.ch/record/1095929`, 2008.

[35] The ATLAS Collaboration. *ATLAS magnet system: Technical Design Report, 1.* Technical design report. ATLAS. CERN, Geneva, 1997.

[36] Ana Maria Rodriguez Vera and Joao Antunes Pequenao. ATLAS Detector Magnet System. `https://cds.cern.ch/record/2770604`, 2021.

[37] The ATLAS Collaboration. Performance of the ATLAS trigger system in 2015. *The European Physical Journal C*, 77(5):317, 2017.

[38] T Cornelissen, M Elsing, S Fleischmann, W Liebig, E Moyse, and A Salzburger. Concepts, Design and Implementation of the ATLAS New Tracking (NEWT). Technical report, CERN, Geneva, 2007.

[39] W Lampl, S Laplace, D Lelas, P Loch, H Ma, S Menke, S Rajagopalan, D Rousseau, S Snyder, and G Unal. Calorimeter Clustering Algorithms: Description and Performance. Technical report, CERN, Geneva, 2008.

[40] The ATLAS Collaboration. Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV. *The European Physical Journal C*, 76, may 2016.

[41] Yu.L. Dokshitzer, G.D. Leder, S. Moretti, and B.R. Webber. Better jet clustering algorithms. *Journal of High Energy Physics*, 1997(08):001, sep 1997.

[42] Stephen D. Ellis and Davison E. Soper. Successive combination jet algorithm for hadron collisions. *Phys. Rev. D*, 48:3160–3166, Oct 1993.

[43] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063, apr 2008.

[44] ATLAS Collaboration. ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV. *The European Physical Journal C*, 79(11), nov 2019.

[45] ATLAS Collaboration. ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset. 11 2022.

[46] David Krohn, Jesse Thaler, and Lian-Tao Wang. Jets with variable R. *Journal of High Energy Physics*, 2009(06):059–059, jun 2009.

[47] The ATLAS Collaboration. Expected performance of missing transverse momentum reconstruction for the ATLAS detector at $\sqrt{s} = 13$ TeV. Technical report, CERN, Geneva, 2015.

[48] Georges Aad et al. Jet energy scale and resolution measured in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 81(8):689, 2021.

[49] ATLAS Collaboration. Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D*, 96:072002, Oct 2017.

[50] The ATLAS Collaboration. Jet reconstruction and performance using particle flow with the ATLAS Detector. *The European Physical Journal C*, 77(7):466, 2017.

[51] Duccio Pappadopulo, Andrea Thamm, Riccardo Torre, and Andrea Wulzer. Heavy vector triplets: bridging theory and data. *Journal of High Energy Physics*, 2014(9), sep 2014.

[52] ATLAS Collaboration. Boosted hadronic vector boson and top quark tagging with ATLAS using Run 2 data. Technical report, CERN, Geneva, 2020. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2020-017.

[53] Georges Aad et al. Identification of boosted, hadronically decaying W bosons and comparisons with atlas data taken at $\sqrt{s} = 8$ TeV. *Eur. Phys. J. C*, 76(3):154, 2016.

[54] Jesse Thaler and Ken Van Tilburg. Identifying boosted objects with n-subjettiness. *Journal of High Energy Physics*, 2011(3), mar 2011.

[55] Jesse Thaler and Ken Van Tilburg. Maximizing boosted top identification by minimizing n-subjettiness. *Journal of High Energy Physics*, 2012(2), feb 2012.

[56] Andrew J. Larkoski, Duff Neill, and Jesse Thaler. Jet shapes with the broadening axis. *Journal of High Energy Physics*, 2014(4), apr 2014.

[57] CMS Collaboration. Displays of candidate events in the search for new heavy resonances decaying to dibosons in the all-jets final state in the CMS detector, 2022. CMS Collection.

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

[59] Andrew J. Larkoski, Gavin P. Salam, and Jesse Thaler. Energy correlation functions for jet substructure. *Journal of High Energy Physics*, 2013(6), jun 2013.

[60] Andrew J. Larkoski, Ian Moult, and Duff Neill. Power counting to better jet observables. *Journal of High Energy Physics*, 2014(12), dec 2014.

[61] Jesse Thaler and Lian-Tao Wang. Strategies to identify boosted tops. *Journal of High Energy Physics*, 2008(07):092–092, jul 2008.

[62] ATLAS Collaboration. Performance of top-quark and W-boson tagging with atlas in run 2 of the lhc. *Eur. Phys. J. C*, 79:375, 2019.

[63] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The catchment area of jets. *Journal of High Energy Physics*, 2008(04):005–005, apr 2008.

[64] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1), aug 2017.

[65] Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors, *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007.

[66] Afra Zomorodian. Fast construction of the Vietoris-Rips complex. *Computers and Graphics*, 34(3):263–271, 2010.

[67] François Chollet et al. Keras. `https://keras.io`, 2015.

[68] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

[69] Timothy Dozat. Incorporating Nesterov Momentum into Adam. In *Proceedings of the 4th International Conference on Learning Representations*, pages 1–4, 2016.

[70] Daniele Grattarola and Cesare Alippi. Graph neural networks in tensorflow and keras with spektral, 2020.

[71] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2016.

[72] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus), 2015.

[73] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, May 2010. PMLR.

[74] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs, 2017.

[75] ATLAS Collaboration. Search for single production of vector-like T quarks decaying into Ht or Zt in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, 2023.

[76] The 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. *The European Physical Journal C*, 79(12), dec 2019.

[77] The ATLAS Collaboration. Measurement of the $t\bar{t}$ production cross-section and lepton differential distributions in $e\mu$ dilepton events from pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *The European Physical Journal C*, 80(6), jun 2020.

[78] ATLAS Collaboration. ATLAS simulation of boson plus jets processes in Run 2. Technical report, CERN, Geneva, 2017.

[79] The ATLAS Collaboration. Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC. *The European Physical Journal C*, 76(12), nov 2016.

[80] ATLAS Collaboration. Tagging and suppression of pileup jets with the ATLAS detector. Technical report, CERN, Geneva, 2014.

[81] Michał Czakon and Alexander Mitov. Top++: A program for the calculation of the top-pair cross-section at hadron colliders. *Computer Physics Communications*, 185(11):2930–2938, nov 2014.

[82] The ATLAS Collaboration. Measurements of inclusive and differential fiducial cross-sections of $t\bar{t}$ production with additional heavy-flavour jets in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Journal of High Energy Physics*, 2019(4), apr 2019.

[83] Nikolaos Kidonakis. Next-to-next-to-leading-order collinear and soft gluon corrections for t-channel single top quark production. *Physical Review D*, 83(9), may 2011.

[84] Nikolaos Kidonakis. Two-loop soft anomalous dimensions for single top quark associated production with a W^- or H^-. *Physical Review D*, 82(5), sep 2010.

[85] Nikolaos Kidonakis. Next-to-next-to-leading logarithm resummation for s-channel single top quark production. *Physical Review D*, 81(5), mar 2010.

[86] Enrico Bothmann et al. Event Generation with Sherpa 2.2. *SciPost Phys.*, 7(3):034, 2019.

[87] The ATLAS Collaboration. Measurement of higgs boson decay into b-quarks in associated production with a top-quark pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Journal of High Energy Physics*, 2022(6), jun 2022.

[88] J. M. Campbell and R. K. Ellis. Update on vector boson pair production at hadron colliders. *Physical Review D*, 60(11), nov 1999.

[89] J. Alwall, S. Hoche, F. Krauss, N. Lavesson, L. Lonnblad, F. Maltoni, M.L. Mangano, M. Moretti, C.G. Papadopoulos, F. Piccinini, S. Schumann, M. Treccani, J. Winter, and M. Worek. Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions. *The European Physical Journal C*, 53(3):473–500, dec 2007.

[90] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2), feb 2011.

[91] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Erratum to: Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 73(7), 2013.

[92] A L Read. Presentation of search results: the CLs technique. *Journal of Physics G: Nuclear and Particle Physics*, 28(10):2693, sep 2002.

[93] Thomas Junk. Confidence level computation for combining searches with small statistics. *Nucl. Instrum. Meth. A*, 434:435–443, 1999.

[94] ATLAS Statistics Forum. The CL_s method: information for conference speakers. `https://www.pp.rhul.ac.uk/~cowan/stat/cls/CLsInfo.pdf`.

[95] Aldo Deandrea, Thomas Flacke, Benjamin Fuks, Luca Panizzi, and Hua-Sheng Shao. Single production of vector-like quarks: the effects of large width, interference and NLO corrections. *JHEP*, 08:107, 2021.

[96] ATLAS Collaboration. Luminosity determination in pp collisions at $\sqrt{s} = 13$ tev using the atlas detector at the lhc, 2022.

[97] The ATLAS Collaboration. Search for pair production of up-type vector-like quarks and for four-top-quark events in final states with multiple b-jets with the ATLAS detector. *Journal of High Energy Physics*, 2018(7):89, 2018.

[98] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An introduction to PYTHIA 8.2. *Computer Physics Communications*, 191:159–177, jun 2015.

[99] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Christopher S. Deans, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Nathan P. Hartland, José I. Latorre, Juan Rojo, and Maria Ubiali. Parton distributions with LHC data. *Nuclear Physics B*, 867(2):244–289, feb 2013.

[100] ATLAS Collaboration. Studies on top-quark Monte Carlo modelling with Sherpa and MG5_aMC@NLO. Technical report, CERN, Geneva, 2017. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2017-007.

[101] Stefano Frixione, Paolo Nason, and Giovanni Ridolfi. A Positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction. *JHEP*, 09:126, 2007.

[102] Paolo Nason. A new method for combining NLO QCD with shower Monte Carlo algorithms. *JHEP*, 11:040, 2004.

[103] Stefano Frixione, Paolo Nason, and Carlo Oleari. Matching NLO QCD computations with parton shower simulations: the POWHEG method. *JHEP*, 11:070, 2007.

[104] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and

next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014.

[105] M. Bahr et al. Herwig++ physics and manual. *Eur. Phys. J. C*, 58:639, 2008.

[106] Johannes Bellm et al. Herwig 7.0/Herwig++ 3.0 release note. *Eur. Phys. J. C*, 76(4):196, 2016.

[107] J. A. Aguilar-Saavedra. Protos - program for top simulations. `http://jaguilar.web.cern.ch/jaguilar/protos/`.

[108] Stefano Frixione, Eric Laenen, Patrick Motylinski, Chris White, and Bryan R Webber. Single-top hadroproduction in association with a w boson. *Journal of High Energy Physics*, 2008(07):029–029, jul 2008.

[109] The ATLAS Collaboration. Studies on top-quark Monte Carlo modelling for Top2016. Technical report, CERN, Geneva, 2016. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2016-020.