A thesis submitted in partial fulfillment of the requirements for the degree *Master of Science*

The Impact of Topo-Cluster Splitting on Boosted Object Identification in ATLAS

Nicolai Weitkemper
born in Soest

September 15, 2025

Technische Universität Dortmund	Fakultät Physik
Università di Bologna	Dipartimento di Fisica e Astronomia
Université Clermont Auvergne	École Universitaire de Physique et d'Ingénierie

This thesis summarizes the scientific work carried out at the Junior Research Group Delitzsch of TU Dortmund University.

Supervisor Dr. Chris Malena Delitzsch
First reviewer Dr. Chris Malena Delitzsch
Second reviewer PD Dr. Dominik Elsässer
Date of submission 2025-09-15

Abstract

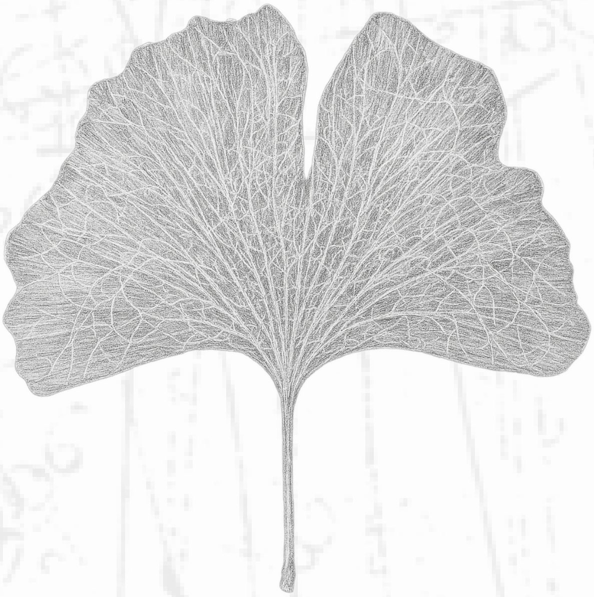
The identification of boosted objects in high-energy particle physics relies on precise jet substructure reconstruction. In ATLAS, a splitting algorithm divides overly large topological clusters (topo-clusters) at local energy maxima to preserve substructure information, but its impact on physics performance has not been systematically studied. This thesis presents a comprehensive investigation of topo-cluster splitting using Monte-Carlo simulations of top and W/Z jets alongside QCD dijet backgrounds. Specialized methodologies are developed, including an approach for comparing clusters before and after splitting as well as an algorithm for matching truth constituents to reconstructed clusters.

The analysis demonstrates that splitting is essential for jet substructure reconstruction. Without it, the discriminating power of variables like τ_{21} and τ_{32} is severely degraded. Comparing the extreme configurations – disabling splitting entirely versus the default settings – reveals that only a small fraction of jets is reconstructed better without splitting. A parameter grid scan indicates that while the choice of splitting parameters has measurable effects, more substantial changes to the algorithm are required for meaningful performance improvements.

Kurzfassung

Die Identifikation von geboosteten Objekten in der Hochenergie-Teilchenphysik beruht auf einer präzisen Rekonstruktion der Jet-Substruktur. In ATLAS werden übergroße topologische Cluster (Topo-Cluster) an lokalen Energiemaxima durch einen Splitting-Algorithmus geteilt, damit Informationen über die Substruktur erhalten bleiben; der Einfluss dieses Verfahrens auf die Rekonstruktionsqualität wurde bislang jedoch nicht systematisch untersucht. In dieser Arbeit wird eine umfassende Untersuchung des Topo-Cluster-Splittings präsentiert, die auf Monte-Carlo-Simulationen von top- und W/Z-Jets sowie QCD-Dijet-Hintergründen basiert. Spezialisierte Methoden werden entwickelt, darunter ein Ansatz zum Vergleich von Clustern vor und nach dem Splitting sowie ein Algorithmus zur Zuordnung von Truth-Konstituenten zu rekonstruierten Clustern.

Es wird gezeigt, dass Splitting für die Rekonstruktion der Jet-Substruktur essenziell ist. Ohne Splitting wird die Trennleistung von Variablen wie τ_{21} und τ_{32} stark vermindert. Der Vergleich extremer Konfigurationen – vollständige Deaktivierung des Splittings gegenüber den Standardeinstellungen – zeigt, dass nur ein kleiner Bruchteil der Jets ohne Splitting besser rekonstruiert wird. Eine Rastersuche im Parameterraum ergibt, dass die Wahl der Splitting-Parameter zwar messbare Effekte hat, jedoch weitergehende Änderungen am Algorithmus erforderlich sind, um eine substanzielle Leistungsverbesserung zu erzielen.



Ist es Ein lebendig Wesen,
Das sich in sich selbst getrennt?
Sind es zwei, die sich erlesen,
Daß man sie als Eines kennt?

— Johann Wolfgang von Goethe

Acknowledgements

I would like to thank all the people and institutions that made this thesis possible. In particular, I extend my gratitude to:

Dr. Chris Malena Delitzsch for the supervision and for giving me the opportunity to stay at CERN.

PD Dr. Dominik Elsässer for taking the time to grade my thesis.

My officemates for their companionship and all the discussions, no matter how (un)related to our work they were.

Unibo / EU for the scholarship, without which the stay at CERN would have been even more costly.

CERN for being an excellent place to get to know so many talented and like-minded people. In the words of T. J. Berners-Lee: “CERN is a wonderful organisation.” [1]

TU Dortmund’s E4 chair for the infrastructure to run my varyingly efficient code on.

My significant other, friends and family for their unconditional support.

Contents

1 Introduction

The search for physics beyond the Standard Model at the Large Hadron Collider [2] depends critically on the precise reconstruction and identification of high-energy objects produced in proton-proton collisions. Among these objects, jets – collimated sprays of particles originating from energetic quarks and gluons – play a central role in nearly every physics analysis. When massive particles such as W and Z bosons or top quarks are produced at high transverse momentum, their decay products are boosted and collimated, forming what is known as a boosted jet. Its internal structure contains information that is crucial for distinguishing signal processes from background, and purpose-built substructure variables are therefore essential tools for analyses [3–5].

Jet reconstruction in the ATLAS collaboration [6] typically includes the formation of topological clusters (topo-clusters) from calorimeter cell signals [7]. These topo-clusters serve as the fundamental input objects for jet reconstruction algorithms. However, the standard topo-clustering algorithm can create overly large clusters that obscure the fine-grained substructure information needed for effective particle identification. To mitigate this issue, a splitting algorithm is applied to identify local energy maxima within large clusters and divide them into smaller, more targeted clusters.

Despite the importance of the splitting algorithm for jet reconstruction quality, the impact of its specific implementation and parameter choices on physics performance has not yet been systematically studied. The algorithm's current parameters, including energy thresholds and neighbor requirements, appear to have been chosen empirically rather than through explicit optimization for particular physics objectives. Furthermore, the splitting procedure operates on rigid thresholds for energies and calorimeter regions, which may not be optimal across the full range of physics scenarios.

This thesis presents a comprehensive study of topo-cluster splitting and its impact on boosted object identification in ATLAS. Using Monte-Carlo simulations of top and W/Z jets from hypothetical W' and Z' boson decays together with dijet background samples, the effect of topo-cluster splitting on jet reconstruction is investigated at multiple levels: from individual calorimeter cells through clusters to complete jets. Novel methods are developed for matching truth-level particle information to reconstructed clusters and for comparing split and non-split cluster configurations. Through systematic analysis of jet substructure variables and their discriminating power, the performance implications of different splitting strategies are quantified.

The thesis is structured as follows. First, the theoretical context is presented, introducing the Standard Model of particle physics, and the physics of jet formation. The Large Hadron Collider and the ATLAS detector are then described, with particular emphasis placed on the calorimeter systems central to this study. The jet reconstruction process is examined in detail, including the topo-clustering and splitting algorithms that form the core of the investigation. The analysis methodology and Monte-Carlo datasets are presented, followed by detailed studies of splitting effects at the jet, cluster, and cell levels. The thesis concludes with a grid search for improved splitting parameters. In an outlook, opportunities for future improvements in jet reconstruction and boosted object identification are discussed.

1.1 The Standard Model

The *Standard Model of particle physics* (SM) [8–11] (Figure 1) is considered the most accurate theory in the entire field of physics. It provides a unified description of the electromagnetic, weak, and strong interactions between fundamental particles. The SM is consistent with numerous experiments and observations, including the discovery of the Higgs boson in 2012 at CERN [12,13].

Mathematically, the SM is a quantum field theory based on the gauge group $\mathrm{SU}(3)_\mathrm{C} \times \mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)_\mathrm{Y}$, where $\mathrm{SU}(3)_\mathrm{C}$ is the gauge group of Quantum Chromodynamics (QCD), describing the strong interaction, and $\mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)_\mathrm{Y}$ describing the electroweak interaction.

Despite its successes, the SM is known to be incomplete, as it does not incorporate gravity[1] [14] or neutrino masses [15], and cannot explain observed phenomena such as the cosmic matter–antimatter asymmetry [16], the nature of dark matter [17], and the accelerated expansion of the Universe (“dark energy”).

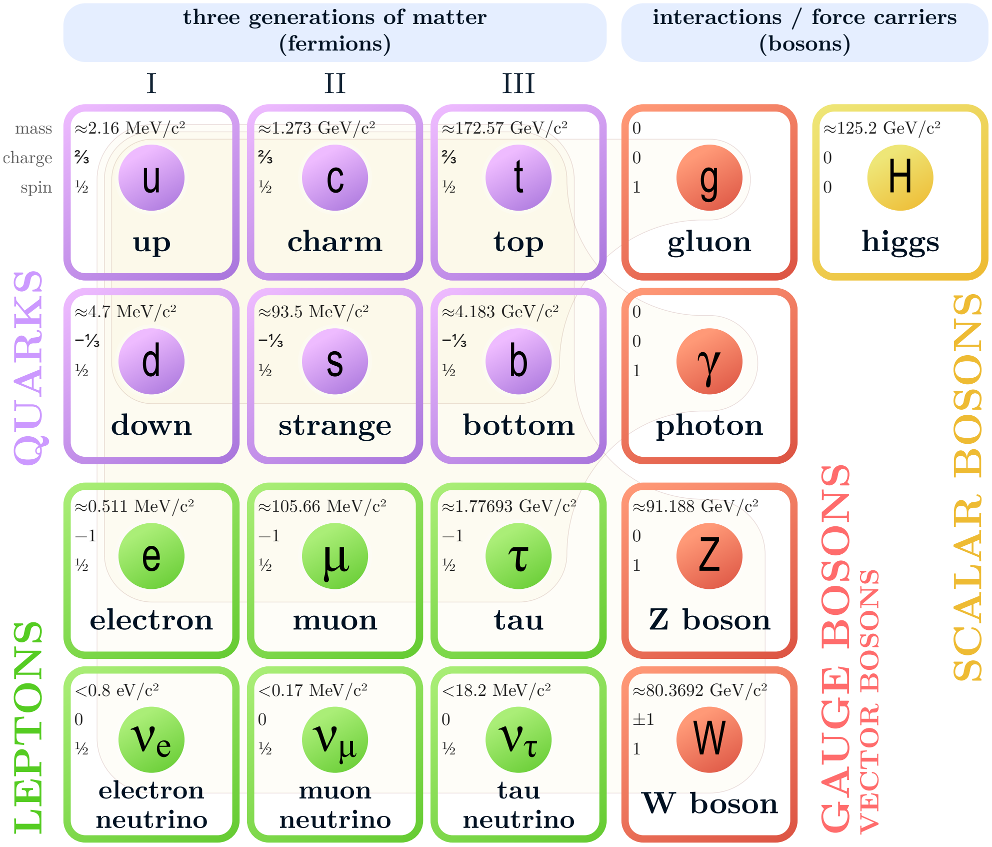

three generations of matter (fermions)
interactions / force carriers (bosons)
I
II
III
mass
charge
spin
≈2.16 MeV/c^2 ⅔ ½ u up
≈1.273 GeV/c^2 ⅔ ½ c charm
≈172.57 GeV/c^2 ⅔ ½ t top
0 0 1 g gluon
≈125.2 GeV/c^2 0 0 H higgs
≈4.7 MeV/c^2 −⅓ ½ d down
≈93.5 MeV/c^2 −⅓ ½ s strange
≈4.183 GeV/c^2 −⅓ ½ b bottom
0 0 1 γ photon
≈0.511 MeV/c^2 −1 ½ e electron
≈105.66 MeV/c^2 −1 ½ μ muon
≈1.77693 GeV/c^2 −1 ½ τ tau
≈91.188 GeV/c^2 0 1 Z Z boson
<0.8 eV/c^2 0 ½ ν_e electron neutrino
<0.17 MeV/c^2 0 ½ ν_μ muon neutrino
<18.2 MeV/c^2 0 ½ ν_τ tau neutrino
≈80.3692 GeV/c^2 ±1 1 W W boson
QUARKS
LEPTONS
GAUGE BOSONS
VECTOR BOSONS
SCALAR BOSONS

Figure 1: Overview of the particles in the Standard Model. Adapted from [18].

[1] The effects of gravity are negligible at the mass/energy scale of collider experiments for the foreseeable future.

Figure 1 shows the fundamental particles and their properties as described by the Standard Model. Fermions (left in the figure; half-integer spin) are the building blocks of matter. They are divided into three generations (flavors) of increasing mass, each consisting of two types of quarks (up/down-type) and two types of leptons (an electrically charged lepton and a neutrino). While ordinary matter consists of first-generation fermions (up/down quarks and electrons), heavier fermions can be produced in high-energy collisions, but quickly decay into lighter particles. Furthermore, each fermion has a corresponding antiparticle with the same mass and spin but opposite charges.

Bosons (right in the figure; integer spin) are force carriers that mediate the interactions between fermions. The photon (γ) mediates the electromagnetic force, the W^+, W^-, and Z^0 bosons mediate the weak force, and the eight gluons (g) mediate the strong force.

As the only scalar boson, the Higgs boson (H) is a manifestation of the Higgs field, which gives mass to elementary particles through the Higgs mechanism [19–24]. The Higgs field has a non-zero vacuum expectation value $v \approx 246\,\mathrm{GeV}$, spontaneously breaking the electroweak symmetry $\mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)_\mathrm{Y} \longrightarrow \mathrm{U}(1)_\mathrm{EM}$. As a result, quarks, charged leptons, and the W/Z bosons acquire mass, while photons, gluons and (in the minimal SM) neutrinos remain massless.

1.1.1 QCD Coupling, Asymptotic Freedom, and Confinement

The strong interaction, being based on the non-Abelian $\mathrm{SU}(3)_\mathrm{C}$ gauge group, has special properties that distinguish it from the electromagnetic and weak interactions. First, unlike photons in Quantum Electrodynamics (QED), gluons carry color charge themselves and can therefore interact with each other. Second, the strength of the strong interaction, characterized by the strong coupling constant α_s, depends on the momentum-transfer scale Q of the interaction. As this scale increases, the strong coupling becomes weaker, a phenomenon known as *asymptotic freedom*. On the other hand, at low momentum scales and large distances, the strong coupling becomes large enough to create quark–antiquark pairs from the vacuum, leading to the formation of bound states called *hadrons* instead of free quarks or gluons. This property is called *confinement*: Isolated color charges are not observed in nature; they are always confined within color-neutral hadrons, such as protons and neutrons.

The interplay of asymptotic freedom and confinement determines the phenomenology of high-energy collisions at *pp* colliders like the LHC. When a high-energy quark or gluon is produced in a collision, it undergoes a cascade of emissions dominated by soft and collinear radiation, known as a *parton shower*. As the shower evolves to lower momentum scales, the increasing strong coupling leads to *hadronization*, where the colored partons are bound into color-neutral hadrons. Because radiation is predominantly collinear at high Q and color must be confined at low Q, the hadrons from one energetic parton form a collimated spray, which is observed in the detector as a *jet*. The overall direction and energy of the jet are directly related to those of the initiating parton, while the internal structure of the jet reflects whether the origin was a generic QCD shower or the decay of a heavy object. In the latter case, the jet has several distinct regions of high energy density, called *prongs*.

1.1.2 Physics Beyond the Standard Model

It is because of the known incompleteness of the Standard Model that physicists are actively searching for New Physics, also known as physics beyond the Standard Model (BSM). BSM physics can manifest itself in various ways, such as the discovery of entirely new particles or interactions, but also through small deviations in precision measurements of known processes.

One possible manifestation of BSM physics is the existence of heavy gauge bosons, commonly referred to as W' and Z'. These hypothetical particles are similar to the Standard Model W and Z bosons, but are predicted to have much higher masses and can arise in various extensions of the Standard Model. A commonly used benchmark is the Sequential Standard Model (SSM) [25], in which W'/Z' have the same fermionic couplings as W/Z and serve as reference signals for reconstruction and tagging studies.

1.2 CERN and the LHC

Since its founding in 1954, the Conseil Européen pour la Recherche Nucléaire[2] (CERN) [26] has been at the forefront of particle physics research. It was at this international institute that W- and Z-bosons were discovered using the Super Proton Synchrotron (SPS) accelerator in 1983 [27,28], and where the Higgs boson was discovered in 2012 [12,13].

Today, it is home to the Large Hadron Collider (LHC) [2], the world's largest and most powerful particle accelerator. The LHC is a 27 km long ring located approximately 100 m underground, where *bunches* of protons or heavy ions are accelerated to near the speed of light and made to collide at four *interaction points* hosting the four major experiments: ATLAS [6], ALICE [29], CMS [30], and LHCb [31].

The LHC has been in operation since 2008, consisting of several *runs* [32], interrupted by *long shutdowns* used for upgrades of the machine and the detectors. Center-of-mass energies $\sqrt{s}$ increased from 7 TeV in Run 1 (2010–2012) to 13 TeV in Run 2 (2015–2018) and 13.6 TeV in Run 3 (2022 – est. 2026). At the same time, the luminosity (explained below) delivered to ATLAS increased drastically, due to increases in $\langle\mu\rangle$ (the average number of simultaneous collisions per bunch crossing) and other improvements to the accelerator complex. In Run 2, a peak instantaneous luminosity of $1.9 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ [33] was reached, exceeding its design value of $1 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ [6] by 90 %, along with a peak $\langle\mu\rangle$ of about 55 [33]. The integrated *good for physics* [33] luminosity at ATLAS for that run was $(140.1 \pm 1.2)\,\mathrm{fb}^{-1}$ [33]. ATLAS selects Luminosity Blocks (LBs), time intervals of typically 1 min during which detector conditions are stable and data quality meets the required standards, as good for physics. Only this subset is used for standard physics analyses.

(Integrated) luminosity L is the measure of the number of potential collisions per area integrated over a given time period. It is related to the number of events N expected for a certain process with cross-section σ via

$$\mu N = L\sigma, \tag{1}$$

where μ is the average number of interactions per bunch crossing.

After another long shutdown with major upgrades to both LHC [34] and ATLAS [35], the *High-Luminosity LHC* (HL-LHC) is scheduled to begin operation in 2029. While the collision energy will increase only slightly to 14 TeV, compared to 13.6 TeV in Run 3, it is planned to increase LHC's integrated luminosity by a factor of 10 and the instantaneous luminosity by a factor of 5 beyond the original design value [34]. The increased luminosity, and thus the larger number of observed interactions, will reduce statistical uncertainties and improve sensitivity to rare processes. This comes at the cost of increased pile-up (explained in Section 1.3.3), posing significant challenges for detector performance and data analysis. Pile-up mitigation techniques will thus become even more crucial in the HL-LHC era.

[2] European Organization for Nuclear Research

1.3 The ATLAS Detector

The ATLAS (A Toroidal LHC ApparatuS) detector [6] is one of the two general-purpose detectors at the LHC, the other being CMS (Compact Muon Solenoid) [30]. Backed by the largest collaboration at CERN, ATLAS is also the largest detector, measuring approximately 44 m in length and 25 m in height, with a total weight of about 7 000 t. As a general-purpose detector, ATLAS provides the means for broad searches for new phenomena and high-precision measurements of Standard Model processes. Accordingly, it is designed to cover nearly the entire solid angle around the collision point, using a variety of subdetectors to identify and measure different types of particles produced in the collisions: an inner detector for tracking charged particles, calorimeters for measuring particle energies, and muon spectrometers for identifying and measuring muons. Strong magnets curve the trajectory of charged particles, allowing for reconstruction of their charge and momenta: A thin 2 T solenoid magnet surrounds the inner detector, while an array of toroidal magnets in between the barrel muon chambers provides a magnetic field in the barrel and endcap regions of the muon spectrometer. The aforementioned components are shown in Figure 2. In Figure 3, the typical signatures of different particles throughout the subdetectors of ATLAS are presented. Charged particles leave tracks in the inner detector, bent by the solenoid magnetic field. Electrons and photons then deposit their energy via *showers* in the electromagnetic calorimeter, while hadrons tend to penetrate deeper into the hadronic calorimeter and shower there. Apart from neutrinos, which escape the detector without interaction, muons are the only charged particles that typically reach the muon spectrometer, where they leave additional tracks bent by the toroidal magnetic field.

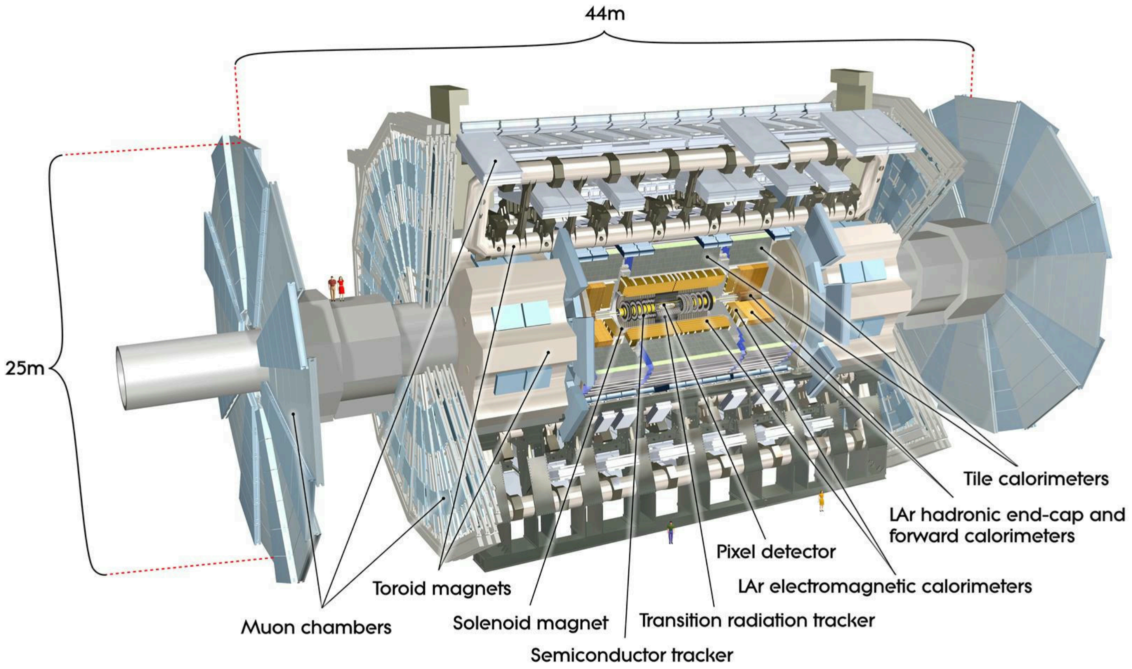

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Figure 2: Subdetectors of the ATLAS experiment in its Run 2 configuration. [6]

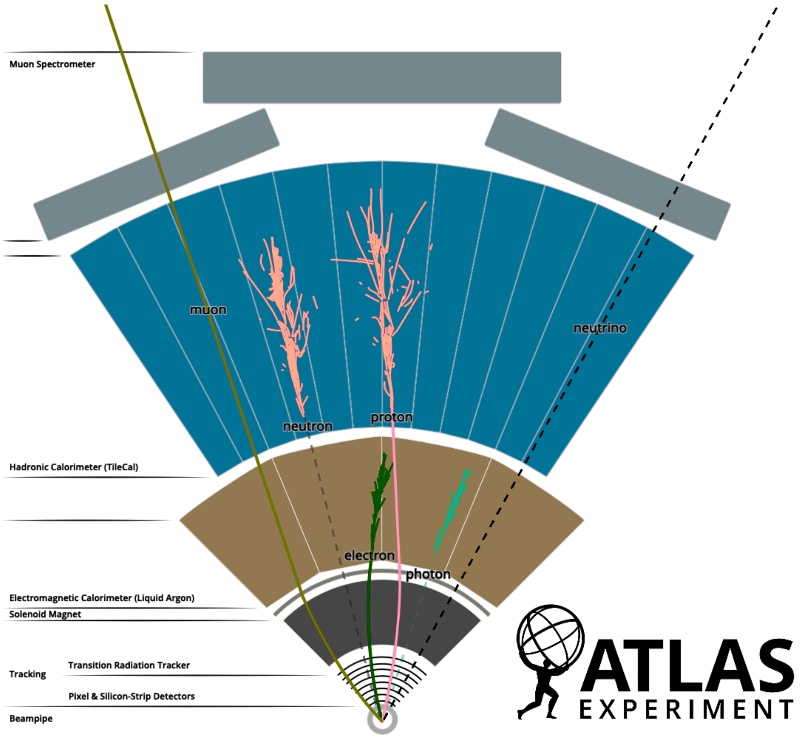

Muon Spectrometer
muon
neutrino
neutron
proton
Hadronic Calorimeter (TileCal)
electron
photon
Electromagnetic Calorimeter (Liquid Argon)
Solenoid Magnet
Tracking
Transition Radiation Tracker
Pixel & Silicon-Strip Detectors
Beampipe
ATLAS
EXPERIMENT

Figure 3: Different particle signatures throughout the subdetectors of the ATLAS experiment. Adapted from [36].

1.3.1 Coordinate System

ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point, the x-axis pointing toward the center of the LHC ring, the y-axis pointing upward, and the z-axis tangential to the beam line [6]. To better capture the rotational symmetry of the detector as well as boosts along the beam axis, positions of particles and calorimeter cells are typically given in pseudorapidity–azimuthal-angle space: The azimuthal angle ϕ is measured around the beam axis and thus lies in the xy plane. The pseudorapidity η is defined as

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

where θ is the polar angle of the particle with respect to the beam axis. For massive particles, the rapidity y is used instead, defined as

$$y = \frac{1}{2}\ln\left(\frac{E+p_z}{E-p_z}\right), \tag{3}$$

where E is the energy and p_z is the longitudinal momentum of the particle (parallel to the beam axis). Angular distances are then measured as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \tag{4}$$

1.3.2 Calorimeter

The ATLAS calorimeter system is designed to precisely measure the energy of both electromagnetic and hadronic showers. It is housed in three cryostats, one barrel and two end-caps, and consists of multiple subsystems using different technologies.

Although specifics vary between the subsystems, most calorimeters in ATLAS are sampling calorimeters using either liquid argon (LAr) or scintillating tiles as active material. Sampling calorimeters, in contrast to homogeneous calorimeters, alternate layers of active and passive material to contain showers in a more compact and cost-effective manner at the expense of energy resolution.

Figure 4 visualizes how the ATLAS calorimeters are arranged and which ones are traversed by particles with different pseudorapidities (η). Longitudinally, a clear separation between the barrel and both endcaps is visible. For $|\eta| \lesssim 1$, particles traverse the barrel calorimeters, while for $1 \lesssim |\eta| \lesssim 3.2$, they pass through the endcap calorimeters. Beyond that, particles enter the forward calorimeters which extend the coverage to $|\eta| \lesssim 4.9$.

A given cell's position in a calorimeter is defined by its location in η, ϕ, and the sampling layer it belongs to. Besides the deposited energy, the information available per cell includes the time of the energy deposit and the significance as the ratio of energy to noise, as it is used in topo-cluster formation (see Section 2.1.2).

Section C in the appendix lists the names and properties of the individual sampling layers in the ATLAS calorimeter system, as they were in Run 2 (2015–2018).

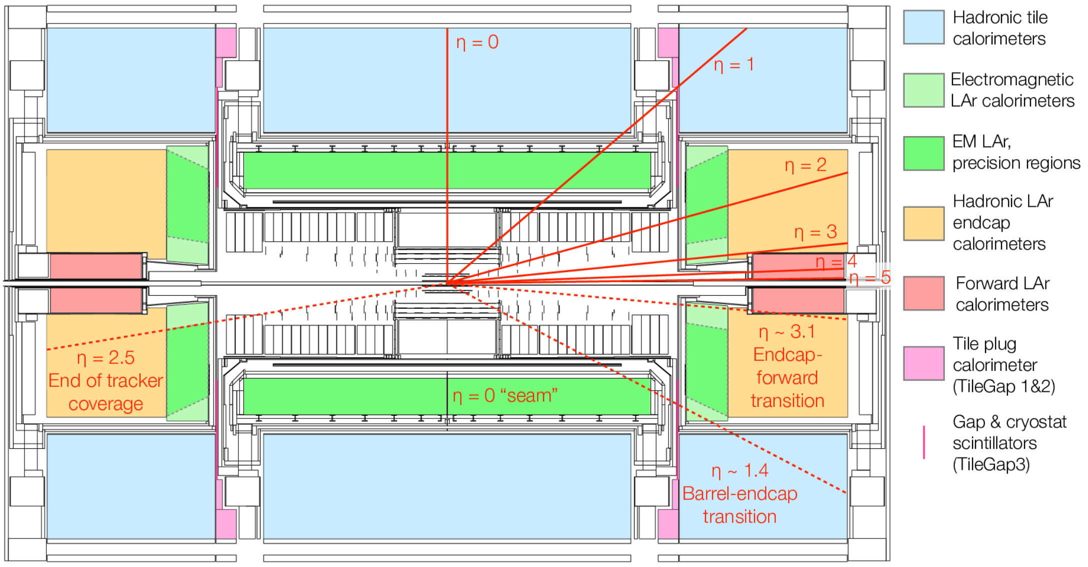


Figure 4: Layout of the ATLAS calorimeters with pseudorapitidy (η) values marked for reference. [37]

1.3.3 Pile-up

Beyond electronic noise, the dominant source of noise in the calorimeters is pile-up, which refers to additional proton-proton interactions.

A distinguishment is made between two types of pile-up: **In-time pile-up** originates from other collisions in the same bunch crossing (the number of which is given by μ), whereas **out-of-time pile-up** stems from collisions in other bunch crossings, both before and after the bunch crossing of interest. For physics analyses, this needs to be accounted for. Mitigations include pulse shaping (Figure 5), which ensures that the energy integrates to zero, averaging out the effect of out-of-time pile-up, and the use of topo-clusters (Section 2.1.2) instead of individual calorimeter cells as input to jet algorithms (Section 2.1.5).

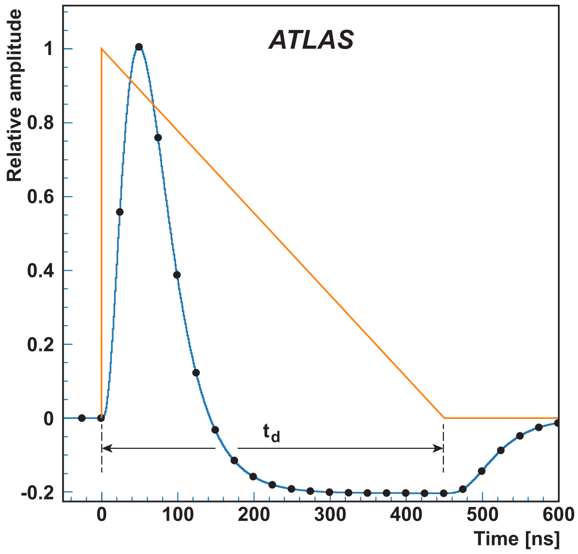


Figure 5: The shaped calorimeter response (🟦) to a triangular pulse (🟧). Points indicate the sampling at 25 ns intervals. Adapted from [6].

2 Jets

Having discussed the QCD foundations that underlie jet formation in Section 1.1.1, this chapter focuses on how jets are defined and reconstructed in ATLAS. It introduces the topo-clustering and splitting algorithms subject to investigation in this thesis.

Jets are not synonymous with physical objects, but rather a helpful concept representing a collection of particles that are likely to have originated from a single high-energy interaction. Importantly, current detector technology cannot provide enough information to reconstruct jets and assign them to particles unambiguously; instead, the original particle's energy and direction are approximated as well as possible while rejecting noise from sources like pile-up and electronics. For this reason, different reconstruction algorithms and configurations (described in Section 2.1.6) are chosen for different purposes.

Figure 6 illustrates how decay products of highly boosted particles can be collimated, so that the otherwise independent jets become *subjets* of a single large-radius jet. Upon reconstruction, they make up a jet's *substructure*, which can be parametrized by various observables (see Section 2.2). Distinct regions of high energy density are also referred to as *prongs*.

Substructure variables are observables designed to study a jet's internal structure to differentiate between jets originating from different processes (e.g. QCD vs. hadronic W/Z decay). They probe features such as the number of hard prongs, the distribution of energy within the jet, and the angular separation of constituents. By quantifying these aspects, substructure variables enable the identification of jets from boosted heavy particle decays and help suppress backgrounds from ordinary QCD jets. They have become essential tools in modern collider analyses, especially in searches for new physics involving highly energetic, collimated decays.

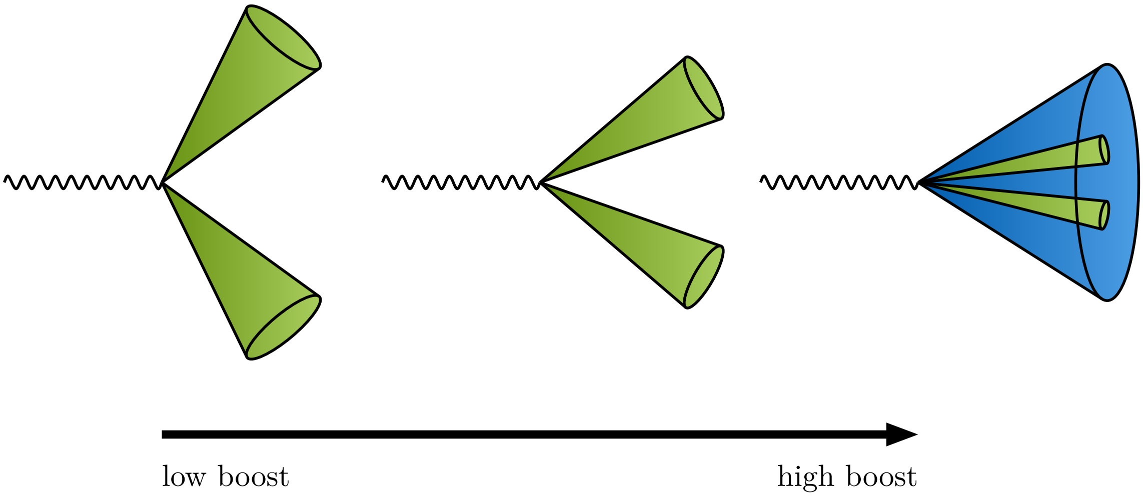


Figure 6: Schematic depiction of how boosted quarks from the decay of a heavy particle are collimated into a single large-radius jet. Adapted from [38].

2.1 Jet Reconstruction

Jet reconstruction consists of multiple steps, including the formation of input objects, the clustering of these objects into jets using a jet algorithm (using e.g. sequential recombination algorithms), and the calibration of the resulting jets to correct for detector effects.

2.1.1 IRC Safety

Sequential recombination algorithms as well as substructure variables must satisfy two important properties: Infrared (IR) safety and collinear safety, often abbreviated as IRC safety. These properties ensure that the algorithm's result does not change when adding soft (IR) emissions that deviate significantly from the jet axis or collinear particles along the jet axis, respectively, which are common occurrences in QCD processes (see Section 1.1.1). Figure 7 illustrates these properties.

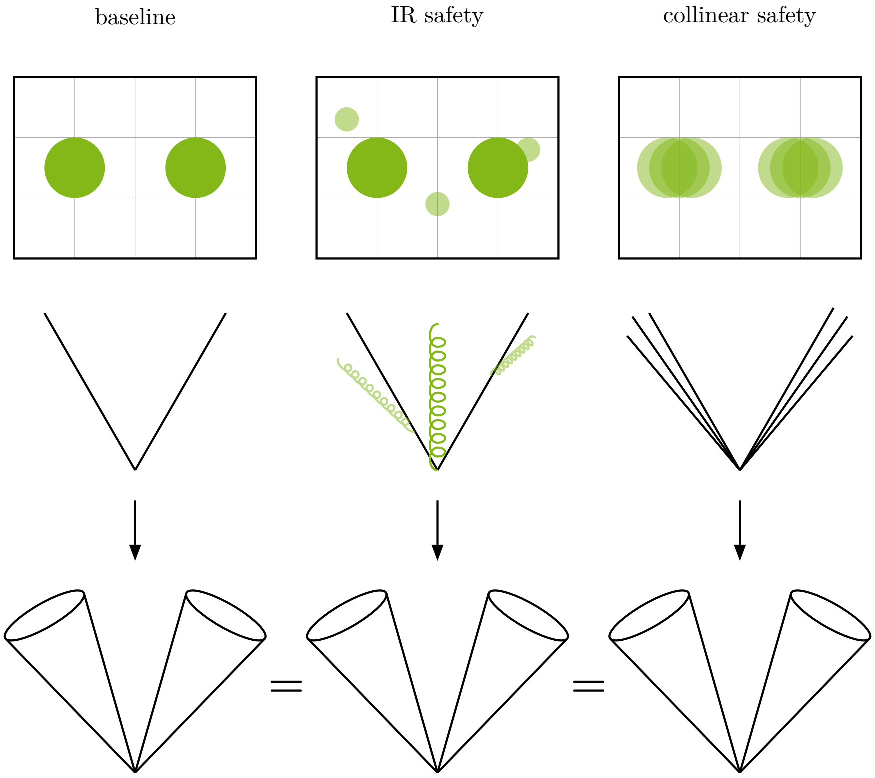


Figure 7: Illustration of infrared and collinear safety. From top to bottom: Energy deposits in the calorimeter, Feynman-like diagrams of the particles making up the jets, and the reconstructed jets. The reconstructed jets are not affected by the addition of soft particles (center) or the "splitting" into collinear particles (right).

2.1.2 Topological Clustering

The topological clustering (topo-clustering) algorithm [7] is used by the ATLAS collaboration to group calorimeter cells into *topo-clusters*. Jet reconstruction algorithms then use these topo-clusters as input objects to form jets. This approach is historically superior to using individual calorimeter cells, as it reduces the impact of electronic noise and pile-up (see Section 1.3.3).

The algorithm utilizes the cells' *significance*, which is defined as the ratio of the cell signal to the average (expected) noise σ in this cell:

$$\varsigma = \frac{\left|E_{\text{cell}}^{\text{EM}}\right|}{\sigma_{\text{noise}}} = \frac{\left|E_{\text{cell}}^{\text{EM}}\right|}{\sqrt{\left(\sigma_{\text{electronic}}\right)^2 + \left(\sigma_{\text{pileup}}\right)^2}}. \tag{5}$$

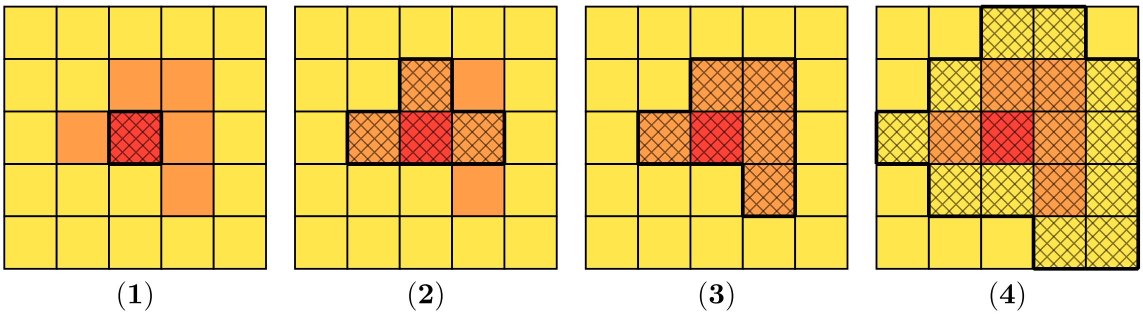


Figure 8: Illustration of the topo-clustering algorithm steps. Cells colored in red/orange/yellow are seed/growth/boundary cells, respectively. The hatching pattern indicates cells belonging to the proto-cluster. For this example, diagonal neighbors are not considered.

Figure 8 demonstrates the steps of the algorithm. In the first step (**1**), the algorithm selects *seed* cells with $\varsigma > S$ that are located in certain allowed calorimeter regions to each "seed" a *proto-cluster*. Secondly (**2** and **3**), these intermediate clusters are grown by repeatedly adding neighboring cells with $\varsigma > N$, where N is a threshold for growth control. Adjacent proto-clusters are merged. Finally (**4**), all cells neighboring the cluster ("boundary cells") that satisfy $\varsigma > P$ are added to the cluster.

In practice, the thresholds $S > N \geq P$ are set as follows [7]:

$$\begin{aligned} \text{seed} \quad S &= 4, \\ \text{neighbor} \quad N &= 2, \\ \text{principal} \quad P &= 0. \end{aligned} \tag{6}$$

The choice of $P = 0$ means that in the last step, all neighbor cells are added to the cluster, regardless of their signal significance. This allows for the inclusion of cells with energies similar to noise levels, improving energy resolution while still suppressing pure noise fluctuations. Cells that have negative energies (due to pulse shaping and electronic noise) require special treatment. While the clustering algorithm considers the absolute value of cell energies, so that clusters can be seeded by and incorporate both positive- and negative-energy cells, jets are built using clusters with positive overall energy only. Still, negative-energy clusters are useful for pile-up suppression and for estimating the overall noise in a given event.

Topo-clusters and jets thus have a many-to-many relationship: A single jet can contribute to multiple topo-clusters, but multiple jets can also overlap and have contributions to the same topo-cluster.

2.1.3 Topo-Cluster Splitting

In some cases, particularly with extremely boosted jets, the proto-clusters formed by the topo-clustering algorithm can grow too large, obscuring the substructure from jet reconstruction algorithms, as those operate on topo-clusters as indivisible units. An example of this is shown in Figure 9, where a single large topo-cluster is formed without splitting, while splitting results in 15 smaller topo-clusters.

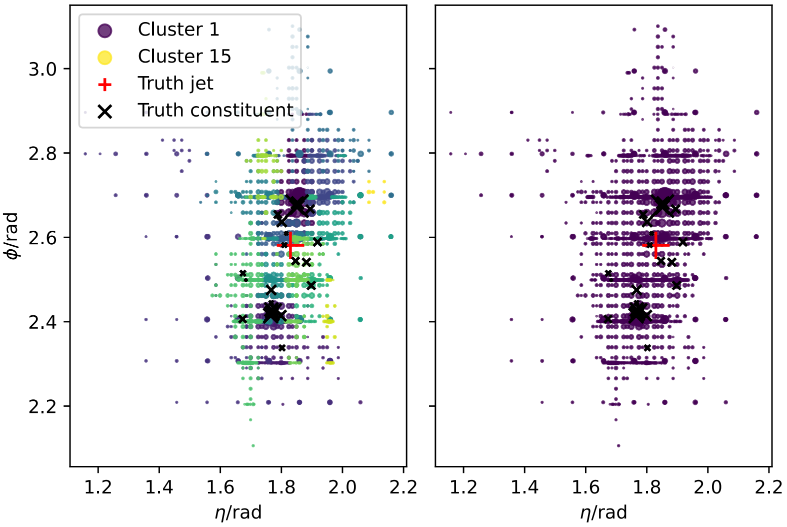


Figure 9: One jet with and without splitting. Without splitting (right), only a single large topo-cluster is formed, while splitting (left) results in 15 smaller topo-clusters.

To counteract this, a splitting algorithm is applied to the proto-clusters after their aforementioned formation. The algorithm identifies local signal maxima within a cluster and splits the cluster between them, by iteratively growing sub-clusters[3] around these maxima. Although the principle of growing clusters from seed cells is similar to the topo-clustering algorithm, there are major differences: The splitting algorithm considers the actual cell energies, which need to exceed 500 MeV to be considered local maxima (seeds) [7]. Furthermore, in a first step, only maxima from the sampling layers EMB2, EMB3, EME2, EME3, and FCAL0 are considered, as these provide more granular information than the coarser hadronic layers. No ς or E thresholds impede the growth of sub-clusters, only the boundary of the original cluster is respected.

If in one growth step two or more sub-clusters meet, boundary cells are assigned to the two highest-energy sub-clusters with weights that depend on the distance r to the sub-cluster centres of gravity d_1, d_2 as well as their energies:

[3] This term is introduced here for clarity; it does not appear in the references.

$$\begin{aligned} w^{\text{geo}}_{\text{cell},1} &= \frac{E^{\text{EM}}_{\text{cluster},1}}{E^{\text{EM}}_{\text{cluster},1} + E^{\text{EM}}_{\text{cluster},2}}, \\ w^{\text{geo}}_{\text{cell},2} &= 1 - w^{\text{geo}}_{\text{cell},1}, \\ r &= \exp(d_1 - d_2). \end{aligned} \tag{7}$$

2.1.4 Local Hadronic Cell Weighting

Several effects cause the energy that a jet deposits in the calorimeters to differ from its true energy at particle level. To correct for these effects, topo-clusters are calibrated using a series of steps, outlined in Figure 10, which together are referred to as Local Hadronic Cell Weighting (LCW).

Due to the non-compensating nature of the ATLAS calorimeters, their response to hadrons is lower than their response to electromagnetic particles of the same energy. Physical reasons for this include invisible energy losses to nuclear breakup and the production of neutrinos. A correcting factor must thus be introduced that is different for electromagnetic and hadronic showers:

$$w^{\text{had-cal}}_{\text{cell}} \neq w^{\text{em-cal}}_{\text{cell}} = 1. \tag{8}$$

Therefore, a classification step is included that estimates the probability $\mathcal{P}^{\text{EM}}_{\text{clus}}$ of a topo-cluster being electromagnetic-like (or hadronic-like). This is done based on the cluster's shape and energy density. The appropriate calibration factors are then applied based on this classification, and the result is weighted by the classifier's confidence $\mathcal{P}^{\text{EM}}_{\text{clus}}$.[4]

$$w^{\text{cal}}_{\text{cell}} = \mathcal{P}^{\text{EM}}_{\text{clus}} \cdot w^{\text{cal-em}}_{\text{cell}} + \left(1 - \mathcal{P}^{\text{EM}}_{\text{clus}}\right) \cdot w^{\text{cal-had}}_{\text{cell}}, \tag{9}$$

where $w^{\text{cal-em}}_{\text{cell}}$ and $w^{\text{cal-had}}_{\text{cell}}$ are composed of several factors outlined in Figure 10.

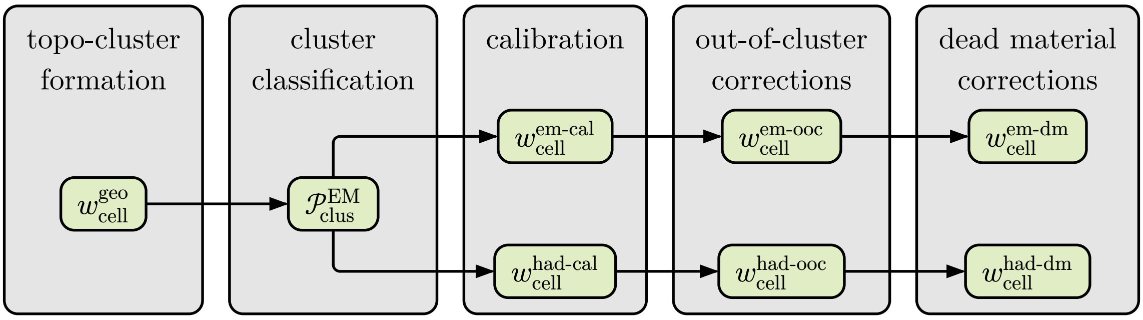

topo-cluster formation
$w^{\text{geo}}_{\text{cell}}$
cluster classification
$\mathcal{P}^{\text{EM}}_{\text{clus}}$
calibration
$w^{\text{em-cal}}_{\text{cell}}$
$w^{\text{had-cal}}_{\text{cell}}$
out-of-cluster corrections
$w^{\text{em-ooc}}_{\text{cell}}$
$w^{\text{had-ooc}}_{\text{cell}}$
dead material corrections
$w^{\text{em-dm}}_{\text{cell}}$
$w^{\text{had-dm}}_{\text{cell}}$

Figure 10: Overview of calibration steps and corresponding weights. Based on [7].

When using variables subject to calibration, their calibration state must be considered. For example, the energy of a cluster can be referred to as $E^{\text{EM}}_{\text{clus}}$ before calibration or $E^{\text{LCW}}_{\text{clus}}$ after LCW calibration. Hereafter, the electromagnetic scale is assumed and the EM superscript is omitted for brevity, with exceptions explicitly noted.

[4] This is preferred over a hard classification to avoid inconsistencies due to misclassifications.

2.1.5 Overview of Jet Algorithms

A variety of algorithms exists to reconstruct jets, drawing on different detector components and providing best results in different kinematic regimes.

As the present work focuses on the topo-clustering aspect of jet reconstruction, LCTopo jets are chosen as the baseline jet definition. While the other jet definitions mentioned below also employ topo-clustering, their additional inputs and reconstruction steps would introduce unnecessary complexity to the study of topo-cluster splitting.

LCTopo jets[5] [7] are reconstructed from calorimeter information only, using the topo-clustering algorithm to group energy deposits in the calorimeter into topological clusters. These topological clusters are then combined into jets using sequential recombination algorithms, such as the anti-k_t algorithm. Because they do not rely on tracking information, LCTopo jets are the only option beyond the inner detector acceptance ($|\eta| > 2.5$; see Section 1.3.2) and remain a robust baseline throughout the detector. The sole reliance on the coarser calorimeter granularity comes at the cost of foregoing track-based angular precision and pile-up mitigation.

PFlow (Particle Flow) jets [39] are reconstructed from a combination of calorimeter and tracking information. The same topo-clustering algorithm is used in the calorimeter, but the resulting topological clusters are combined with tracks to form Particle Flow Objects (PFOs). In practice, PFlow uses track momenta at low–moderate p_T, where they are most precise, and defers to calorimeter energies as p_T increases, yielding near-optimal resolution across the jet's kinematic range. Energy deposits in the calorimeter are then matched to tracks in the inner detector, and the calorimeter energy associated with each matched track is removed from the event, to avoid double-counting of charged particles. PFlow is used primarily for small-R jets at low to moderate p_T, within the tracker acceptance.

TCCs (Track-CaloClusters) [40] use inner-detector track directions together with calorimeter energy to form an input type for jets that resolves dense, collimated structure while retaining calorimeter energy for charged particles (unlike PFOs, which replace it using tracks). They perform best at high p_T, where momentum measurements from the inner detector become less reliable, while the resolution of collimated jets in the calorimeter can still be improved using tracking information. At low p_T, however, TCCs are outperformed by both LCTopo and PFlow jets in terms of resolution and pile-up stability.

Unified Flow Objects (UFOs) [41] are the state-of-the-art in large-R jet reconstruction at ATLAS [42], combining the strengths of both PFlow jets and TCCs at different p_T ranges. By using PFlow-like objects at lower p_T and TCC-like objects at higher p_T, UFO large-R jets provide the best performance across the entire p_T spectrum.

[5] The name **LCTopo** stands for jets built from **topo**logical clusters calibrated with **LC**W (Section 2.1.4) at the cluster level.

2.1.6 Sequential Recombination Algorithms

Though not the only option, the ATLAS collaboration primarily employs sequential recombination algorithms for jet reconstruction. These algorithms iteratively merge pairs of input objects (e.g. calorimeter topo-clusters or particle-flow objects) to form jets, based on a distance measure that typically incorporates both their spatial separation and transverse momenta. There are many variants of sequential recombination algorithms, differing mainly in the specific distance measure used and the order in which objects are merged, but the most widely used sequential recombination algorithms in ATLAS analyses are those in the k_t family.

The family of k_t algorithms operates on two distance measures, namely the distance between pairs of objects i and j,

$$d_{ij} = \min\left(p_{T,i}^{2p}, p_{T,j}^{2p}\right) \frac{\Delta R_{ij}^2}{R^2}, \tag{10}$$

and the distance between each object i and the beam,

$$d_{iB} = p_{T,i}^{2p} \ , \tag{11}$$

where ΔR_{ij} is the distance between particles i and j in the η-ϕ plane (see Equation 4), R is the *jet radius*, controlling the typical size of the jets, $p_{T,i}$ is the transverse momentum of particle i with respect to the beam axis, and p is a parameter (explained below) that impacts the distance measure and consequently the shape of the resulting jet.

The algorithm starts with a list of input objects such as topo-clusters, each representing a proto-jet. It then computes all d_{ij} as well as d_{iB}. If $d_{iB} < d_{ij} \forall j$, object i is considered a final jet and not changed in further iterations.

The choice of p determines the order in which particles are clustered:
For $p = -1$, the **anti-k_t algorithm** [43] is obtained, which clusters hard particles first, leading to more regular, conical jet shapes that are relatively robust against pile-up. This algorithm is the default choice for jet reconstruction in ATLAS analyses.
For $p = 0$, the **Cambridge-Aachen algorithm** [44,45] is obtained, which clusters particles based solely on their angular separation, neglecting their transverse momenta. Cambridge-Aachen is mainly used for *declustering*, a technique that is relevant for jet *grooming* (see Section 2.1.7).
For $p = 1$, the **original k_t algorithm** [46] is obtained, which clusters soft particles first. It is less commonly used for final jet reconstruction but provides the basis for the *exclusive-k_t algorithm* (see below).

The ATLAS collaboration distinguishes between small-R ($R = 0.4$) and large-R ($R = 1$) jets [47]. While smaller radii improve pile-up rejection, larger radii are better suited to reconstruct boosted objects like the W'/Z' decay products discussed in this thesis.

The exclusive-k_t algorithm [48] is a variant of the k_t algorithm (with $p = 1$) that stops the clustering process according to some criterion, usually when a predefined number of proto-jets N or a distance threshold d_{cut} is reached. This is useful for jet substructure studies (see Section 2.2.1), where the same jet is split into different fixed numbers of subjets to probe which configuration best describes the jet's internal structure.

2.1.7 Jet Grooming

Grooming removes soft, wide-angle radiation and pile-up contributions from jets to stabilize jet mass and substructure observables. In ATLAS, LCTopo jets have historically been groomed with *trimming* [49], wherein jet constituents are reclustered into subjets of radius $R_{\mathrm{sub}} \ll R$ and any subjet carrying a p_T fraction below f_{cut} (relative to the ungroomed jet's p_T) is discarded; typical settings are $R_{\mathrm{sub}} = 0.2$ and $f_{\mathrm{cut}} = 0.05$. This also applies to the large-R jets used in this thesis.

More recently, the ATLAS collaboration has also started using *soft drop* [50], which reclusters the jet using the Cambridge-Aachen algorithm and then iteratively *declusters* it, (i.e. reverses the recombination sequence) removing soft wide-angle radiation until a certain condition is met.

Other widely used grooming/tagging variants include *pruning* [51], which discards soft, large-angle recombinations during clustering, and *filtering / (modified) mass-drop* [52,53], which identifies a heavy-particle-like splitting and then refines the jet by reclustering and keeping only the hardest small-radius subjets.

2.1.8 Jet Energy Scale Calibration

While LCW corrects for non-compensation and local response variations at the cluster level, additional effects remain at the jet level. In this work, the jets reconstructed from LCW-calibrated topo-clusters are further corrected by the Jet Energy Scale (JES) calibration [54,55], applied after anti-k_t jet finding and grooming.

It consists of several steps: an *offset correction* removes extra energy from pile-up interactions; a MC-based *response calibration* corrects the mean jet response as a function of p_T and η, so reconstructed jets match particle-level jets in simulation; an *origin (direction) correction* repoints the jet to the primary vertex, updating its four-vector accordingly; and small *in-situ residuals* align data with truth after the MC step.

A more detailed description of the JES calibration procedure can be found in [54] and [55].

2.2 Jet Substructure Variables

Substructure variables are observables that characterise the internal structure of jets. They are useful for distinguishing jets originating from different types of particles, as their decay products lead to different radiation patterns and energy distributions within the jet.

2.2.1 N-subjettiness τ_N

The N-subjettiness τ_N can be thought of as a distance measure, namely between the “energy deposits” in the jet and N *subjet axes*. It was introduced in [3] in analogy to the event-wide N-jettiness [56], where it was used to veto additional jet emissions. Low τ_N indicate that the jet has a clear N-prong substructure, whereas less “clusteredness” or a different number of prongs leads to higher τ_N. The relative value of τ_N for different N can thus be used to estimate the number of subjets in a jet.

The N-subjettiness of a jet is defined as

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \min\{\Delta R_{1,k}, \Delta R_{2,k} ..., \Delta R_{N,k}\}, \tag{12}$$

where k is an index over all particles in the jet, $\Delta R_{J,k}$ is the angular distance between particle k and (implicit) subjet axis J, and d_0 is a normalization factor defined as

$$d_0 = \sum_k p_{T,k} R_0, \tag{13}$$

with the jet radius R_0 that was used in the jet’s construction.

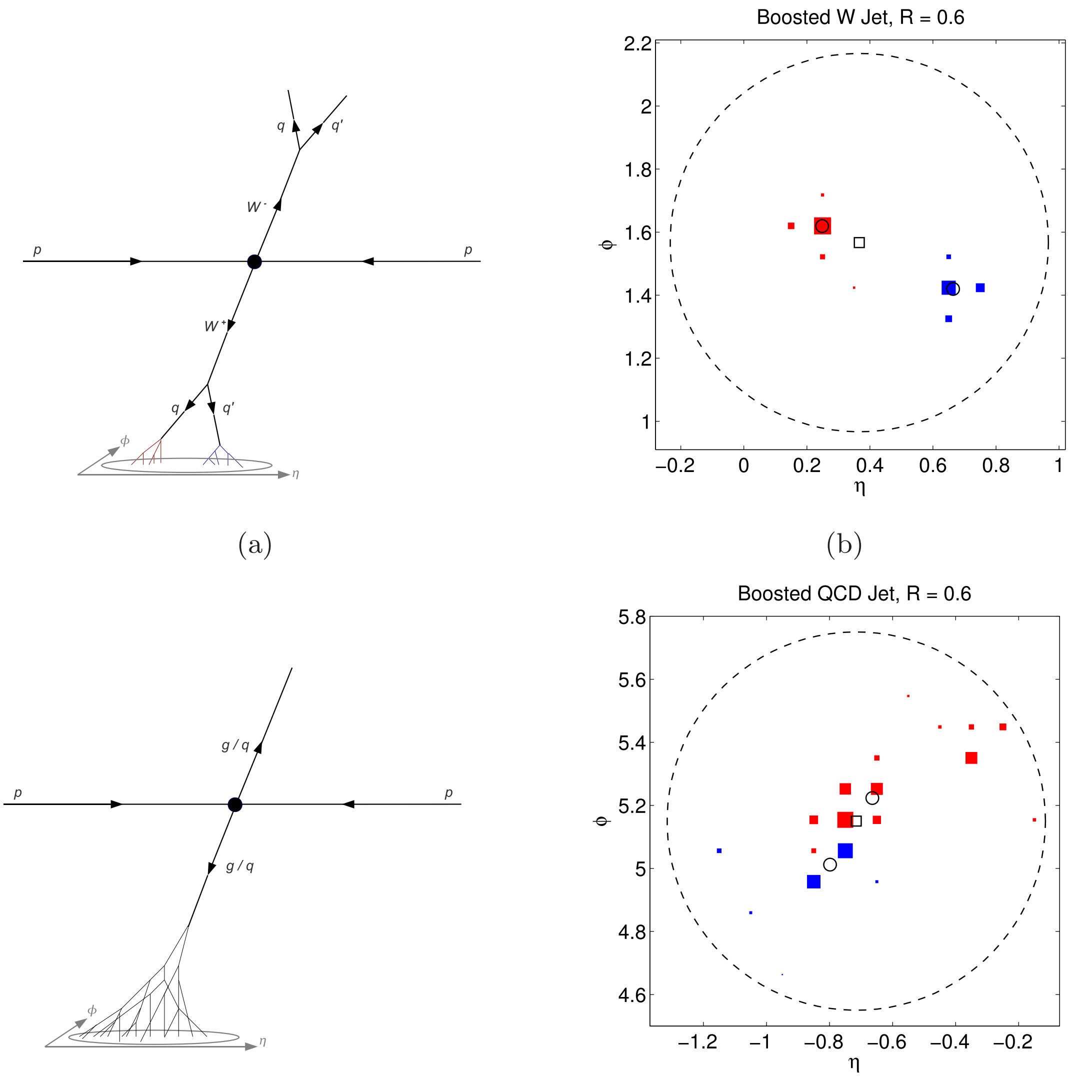
N-subjettiness itself is IRC safe due to its linearity in the momenta of the constituent particles and the smooth angular dependence [3]. Since it depends on subjet axes, these must also be determined in an IRC safe way, e.g. via exclusive-k_T clustering (explained in Section 2.1.6).

In practice, ratios like

$$\tau_{21} = \frac{\tau_2}{\tau_1} \text{ (etc.)} \tag{14}$$

are used due to their superior discriminating power [3].

Figure 11 illustrates how τ_1 and τ_2 can be used to distinguish between jets originating from hadronic W decays and those from QCD processes.



(a)

(b)

(c)

(d)

Figure 11: Schematics of fully hadronic decays in (a) W^+W^- and (c) dijet QCD events alongside typical event displays in (b) and (d). τ_1 resp. τ_2 quantify the alignment of the energy deposits with the outlined square (□) resp. the outlined circles (○). Filled rectangles (■) in (b) and (d) indicate the energy deposits in the calorimeter cells by their size, while the color indicates how the exclusive-k_T algorithm divides them into two clusters. Adapted from [3].

2.2.2 Energy correlation functions e_n and their ratio D_2

Energy correlation functions (ECFs) [4] are another family of jet substructure observables that can be used to identify the prong-like structure of jets. Unlike τ_N and other axis-based substructure observables,[6] ECFs are not only IRC-safe, but also *recoil-insensitive*, meaning that their value does not acquire an extra change when soft, wide-angle radiation merely tilts the jet axis, as they depend only on inter-particle angles and momenta. Soft emissions still contribute directly, but cannot bias the observable through axis displacement.

As an example, in a jet with a soft wide-angle secondary lobe, τ_{21} can falsely indicate 2-pronginess, because exclusive-k_T partitions the jet into two regions and measures distances to those subjet axes. ECF ratios like D_2 instead increase in that configuration, correctly flagging the second prong is soft/wide-angle rather than a genuine hard 2-prong.

The ECFs are defined[7] as

$$\begin{aligned} e_2^{(\beta)} &\equiv \frac{1}{\left(p_{T\,J}\right)^2} \sum_{1\leq i<j\leq n_J} \quad p_{T\,i} p_{T\,j} \qquad R_{ij}^{\beta}, \\ e_3^{(\beta)} &\equiv \frac{1}{\left(p_{T\,J}\right)^3} \sum_{1\leq i<j<k\leq n_J} \quad p_{T\,i} p_{T\,j} p_{T\,k} R_{ij}^{\beta} R_{ik}^{\beta} R_{jk}^{\beta}, \end{aligned} \tag{16}$$

where $\left(p_T\right)_J$ is the transverse momentum of the jet with respect to the beam, $\left(p_T\right)_i$ is the transverse momentum of the i-th particle in the jet, n_J is the number of particles in the jet, and β is the angular exponent. β may be tuned to adjust the sensitivity to collinear splittings. As an example, $\beta \approx 0.2$ is ideal for quark/gluon discrimination [4]. For any $\beta > 0$, ECFs are IRC safe.

As with τ_{21}, a ratio is used to identify two-prong substructure [5]:

$$D_2 = \frac{e_3^{(\beta)}}{\left(e_2^{(\beta)}\right)^3} \tag{17}$$

It has been chosen because it best follows the empirical boundaries between 1-prong QCD and 2-prong jets in the (e_2, e_3) plane and was the best-performing observable for W boson tagging at ATLAS in Run 2 [57].

[6] This applies if the axes are determined using methods like exclusive-k_T, as done here.

[7] The normed ECFs e_n proposed in [5] are given instead of the originally proposed ECF [4]; their relation is

$$e_n^{(\beta)} = \frac{\mathrm{ECF}(n,\beta)}{(\mathrm{ECF}(1,\beta))^n}. \tag{15}$$

3 Analysis

This chapter introduces the Monte-Carlo samples and methods used in this study and presents observations on jet-level, cluster-level and cell-level features. Matching of Monte-Carlo truth constituents to clusters and split to non-split clusters is discussed. Furthermore, the results of a gridsearch for improved parameters for the cluster splitting algorithm are presented.

3.1 Methods

To start, a brief overview of the tools employed in this thesis to compare distributions and evaluate the cutting power of variables is given.

3.1.1 Earth-Mover's Distance (EMD)

In order to compare different distributions of features, the (standardized) Earth-Mover's Distance (EMD) [58], also known as Wasserstein Distance, is used. It is a means of comparing two distributions in a shared metric space by measuring the minimum amount of "work" W (amount × distance as with the physical equivalent) required to transform one distribution into the other. Contrary to comparisons of the mean or select statistical moments, the EMD takes into account the full distribution of the data.

After computing the EMD W[8] using `scipy.stats.wasserstein_distance` [59], it is normalized based on the standard deviation of the combined distributions:

$$W_{\text{std}} = \frac{W}{\sigma_{\text{combined}}}. \tag{18}$$

As exemplified in Figure 12, this makes the EMD invariant to the scale of the distributions, allowing for better comparability of EMDs between different features.

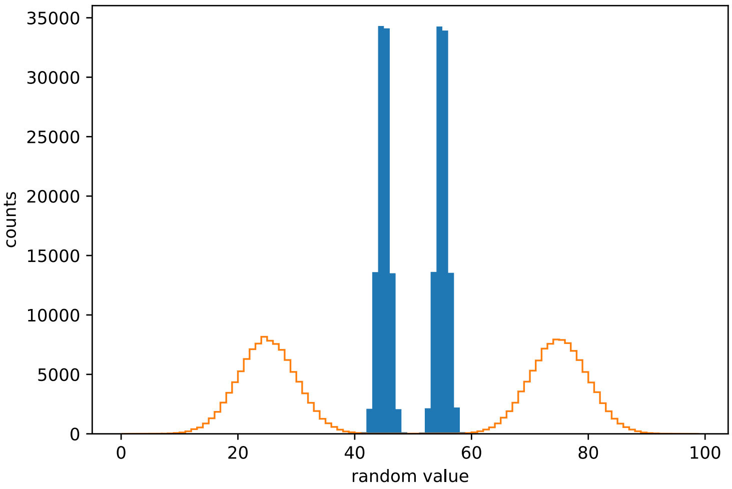


Figure 12: After standard-scaling, the EMD between the two blue (filled) Gaussian peaks is equal to the EMD between the two orange (outlined) Gaussian peaks.

[8] Standard scaling can equivalently be applied to the data *before* computing the EMD.

3.1.2 ROC Curves and AUC Metric

Radio Operating Characteristic (ROC) curves [60] and Area Under the Curve (AUC) are commonly-used metrics in machine learning to evaluate the performance of binary classifiers. They incorporate the trade-off between true positive rate (TPR) and false positive rate (FPR) across different classification thresholds, so that the AUC summarizes the overall performance of a classifier into a single value, namely the area under the ROC curve.

TPR and FPR are defined as follows:

$$\begin{aligned} \mathrm{TPR} &= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \\ \mathrm{FPR} &= \frac{\text{false positives}}{\text{false positives} + \text{true negatives}}, \end{aligned} \tag{19}$$

where positive and negative refer to the two classes of the binary classification problem, and true/false indicates whether the classification was correct or not.

In this thesis, ROC curves are used to evaluate the performance of simple one-sided cuts, where a variable is compared to a threshold to classify events as signal or background. The ROC curve is generated by varying the threshold and plotting the TPR against the FPR at each threshold. In practice, ROC and AUC are computed using `sklearn.metrics.roc_curve` and `sklearn.metrics.roc_auc_score`, respectively [61].

In a balanced dataset, a model with an AUC of 0.5 indicates no discrimination ability (random guessing), whereas an AUC of 1 indicates perfect discrimination. While values below 0.5 correspond to a model that is “worse” than random guessing, it is evident that they still provide some information about the data [62]. Since ROC curves in this thesis only serve as a proxy for the performance of W/Z-boson and top-quark tagging algorithms (which are usually more sophisticated than a simple threshold on a single variable), the target is inverted so that the AUC is always ≥ 0.5.

The shape of the ROC curve depends on the distributions of the signal and background events for the variable being used for classification. In the case of two Gaussian distributions with different means, the ROC curve will have a characteristic convex shape, as shown in Figure 13 and Figure 14. The more the distributions overlap, the closer the ROC curve will be to the identity line, indicating that the classifier is less effective at distinguishing between signal and background. For a Gaussian signal atop a uniform background, however, the ROC curve takes on a sigmoidal shape, as illustrated in Figure 15. Considering a right-sided threshold (i.e. classifying events as signal if the variable exceeds the threshold) that is gradually lowered from 2 to -2, the ROC curve starts off below the identity line (for low FPR values), as the constant background dominates in the $[1, 2]$ region, leading to a TPR that is lower than the FPR. Passing the mean of the Gaussian at 0, where signal and background are equal, the opposite occurs, resulting in the characteristic sigmoidal shape of the ROC curve.

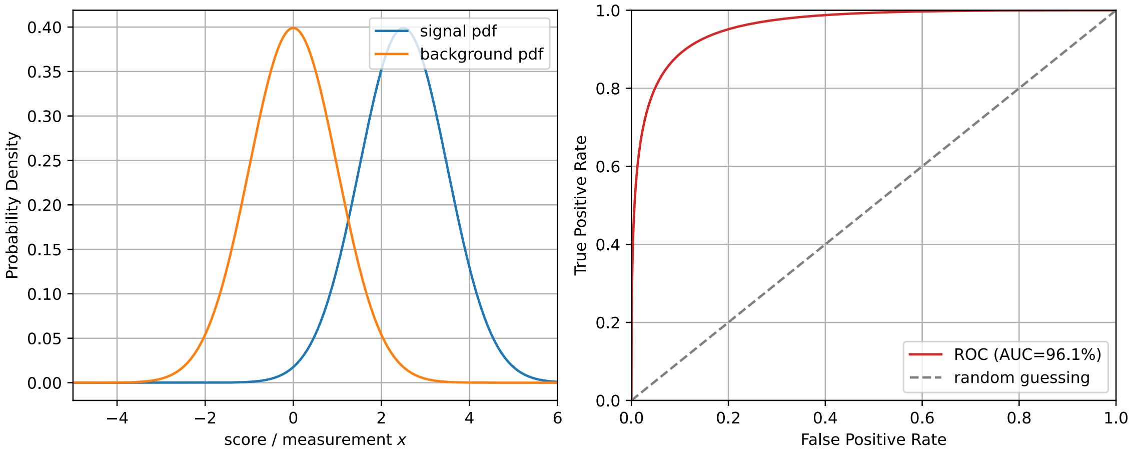


Figure 13: Two Gaussian distributions with different means (left) and the corresponding ROC curve (right).

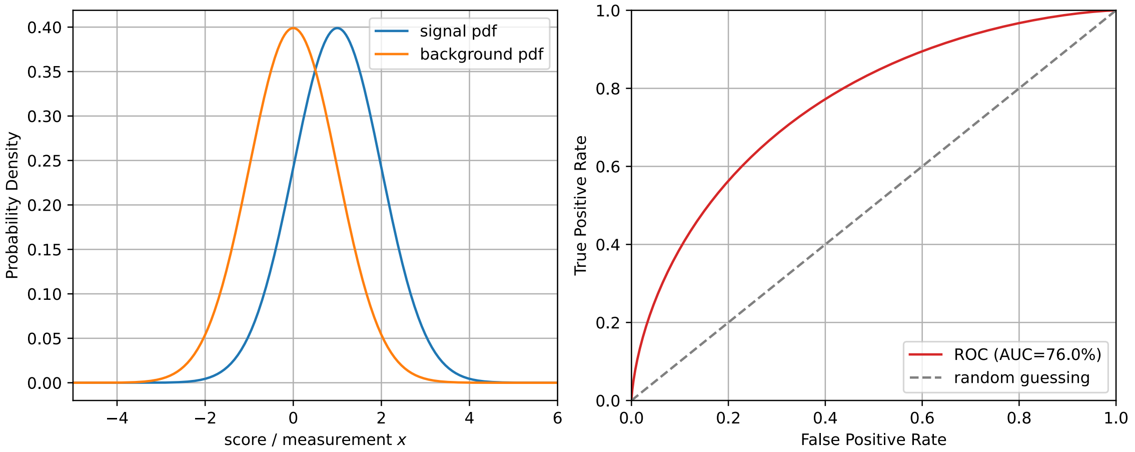


Figure 14: Two Gaussian distributions with more overlap (left) result in a less ideal ROC curve (right).

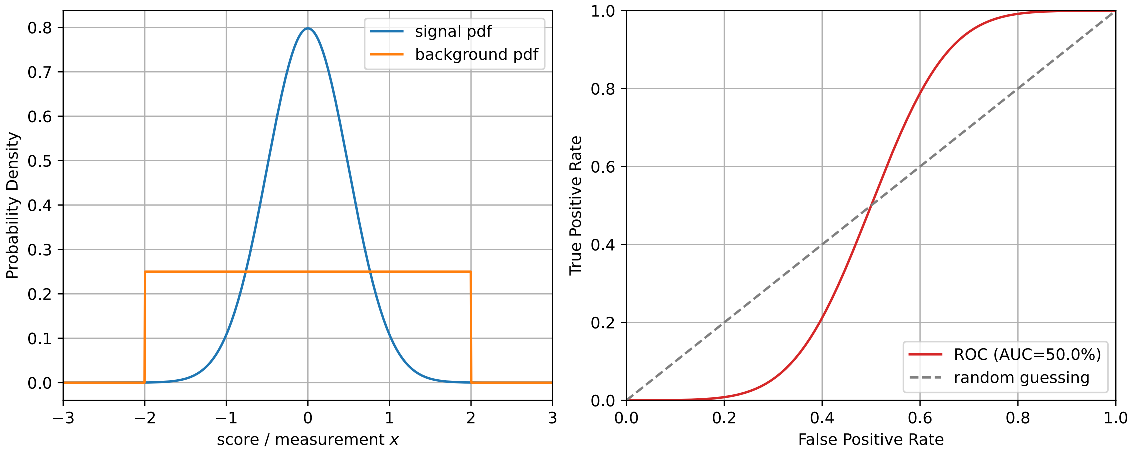


Figure 15: A Gaussian signal distribution atop a uniform background (left) leads to a sigmoidal ROC curve (right).

3.1.3 Processing

Most processing is performed using Awkward Array [63], a Python library that is part of SciKit-HEP. It allows operating on the ragged data structures inherent to the low-level data at hand (different numbers of clusters containing different numbers of cells each) with NumPy-like syntax and performance, eliminating the need for lookup tables, explicit loops, or other cumbersome methods.

Plots are created with Matplotlib [64] and Seaborn [65].

3.2 Monte-Carlo Samples

The Monte-Carlo samples used in this analysis include simulated decay events of exotic W' and Z' bosons (as described in Section 1.1.2) as well as dijet background events. Including both signal and background events allows to study not only the effect of topo-clustering on different kinds of jets, but also the discriminating power of different jet features. The samples are simulated to resemble the conditions (pile-up, $\sqrt{s}$, detector configuration, etc.) during Run 2 of the LHC or more specifically the 2018 data-taking period.

The W' first decays into a W and a Z boson, which then each decay into quark pairs, resulting in a final state with two 2-prong jets (Figure 16). The full simulation procedure is described in [66]. The Z' decays into a top-antitop pair, each decaying fully hadronically into Wb and subsequently into two 3-prong jets (Figure 17). Its simulation procedure is described in [47].

Both the W' and the Z' are defined with a resonance mass of 2 TeV. The corresponding inelastic scattering events are simulated with Pythia 8.235 [67] at leading order (LO), using the NNPDF2.3lo [68] set of parton distribution functions (PDFs) as well as the ATLAS A14 [69] set of tuned parameters for the parton shower and multi-parton interactions. GEANT4 [70] then simulates the response of the ATLAS detector. The cross-sections are reweighted to produce a distribution of jets that is uniformly distributed in p_T.

The dijet background samples (simulation described in [66]) represent the QCD background to the W' and Z' signals. They cover two adjacent p_T regions:
$(1\,300 - 1\,800)$ GeV and $(1\,800 - 2\,500)$ GeV. It is based on leading-order matrix elements simulated with Pythia 8.230 [67] and the aforementioned NNPDF2.3lo [68] and ATLAS A14 [69].

The effect of both in-time and out-of-time pile-up (Section 1.3.3) is modeled by overlaying the simulated hard-scattering event with inelastic minimum-bias proton-proton collision events, generated with Pythia 8.186 [67] using NNPDF2.3lo [68] and ATLAS A3 [71], a set of tuned parameters specifically for modeling minimum-bias events [71]. The number of pile-up interactions to be overlaid is sampled from the measured distribution obtained from Run 2 data [72].

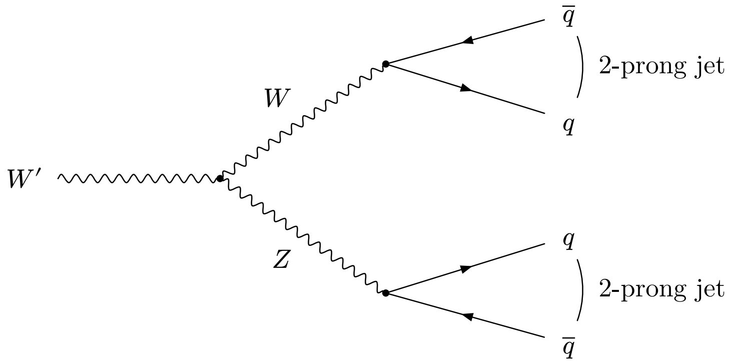


Figure 16: Feynman diagram for the simulated W' decay.

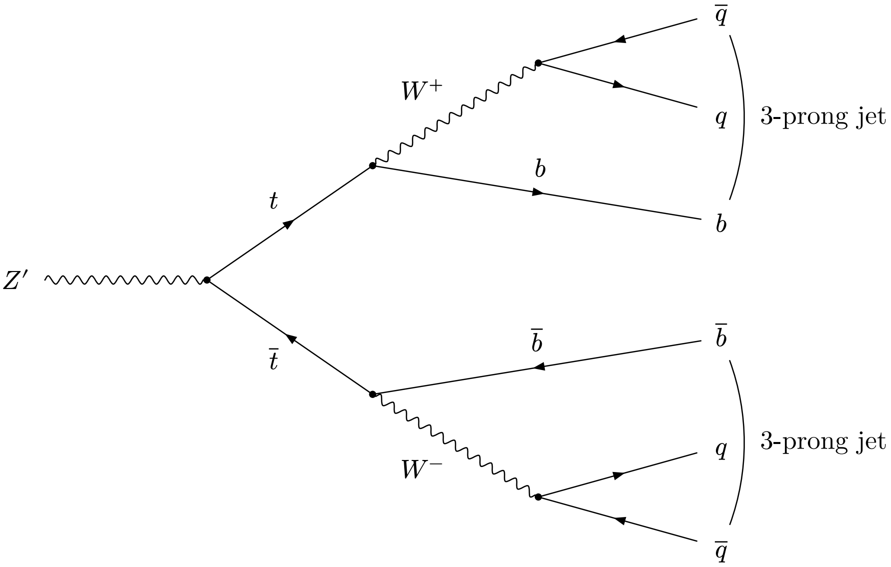


Figure 17: Feynman diagram for the simulated Z' decay.

3.2.1 Data Format and Variations

The samples considered are in the Event Summary Data (ESD) format [73] instead of the more commonly used Analysis Object Data (AOD) format, so that information about individual clusters and their respective cells is available. The samples are provided in ROOT files, each containing a TTree with entries corresponding to individual jets.

Each entry contains a set of features (variables) outlined in Figure 18, describing the jet as a whole (jet-level) as well as its constituents (cluster-level) and cells thereof (cell-level). Two runs of the reconstruction chain are included per event, one with and one without the topo-cluster splitting algorithm enabled; all affected features are duplicated with different names accordingly.[9] In addition, Monte-Carlo truth information about the simulated hard-scattering event is available for each jet. This includes the "true" identity of the jet (signal or background), its "true" kinematic properties (including p_T and mass) and *truth constituents* (i.e. the simulated particles that formed the jet).

Different variations of the Monte-Carlo samples are used to study the impact of pile-up and different hyperparameters of the topo-cluster splitting algorithm. Additional

[9] Not all features are available in all variations; see Section A.

information about the samples is provided in Section B; a complete list of available features is given in Section A.

Except where otherwise noted, a dataset of 200 000 jets is considered, comprised of signal, background, or equal proportions thereof. This suffices to reduce statistical fluctuations to a negligible level for the purposes of this analysis while keeping computation times reasonable. It will be indicated whether W', Z' and/or background events are considered.

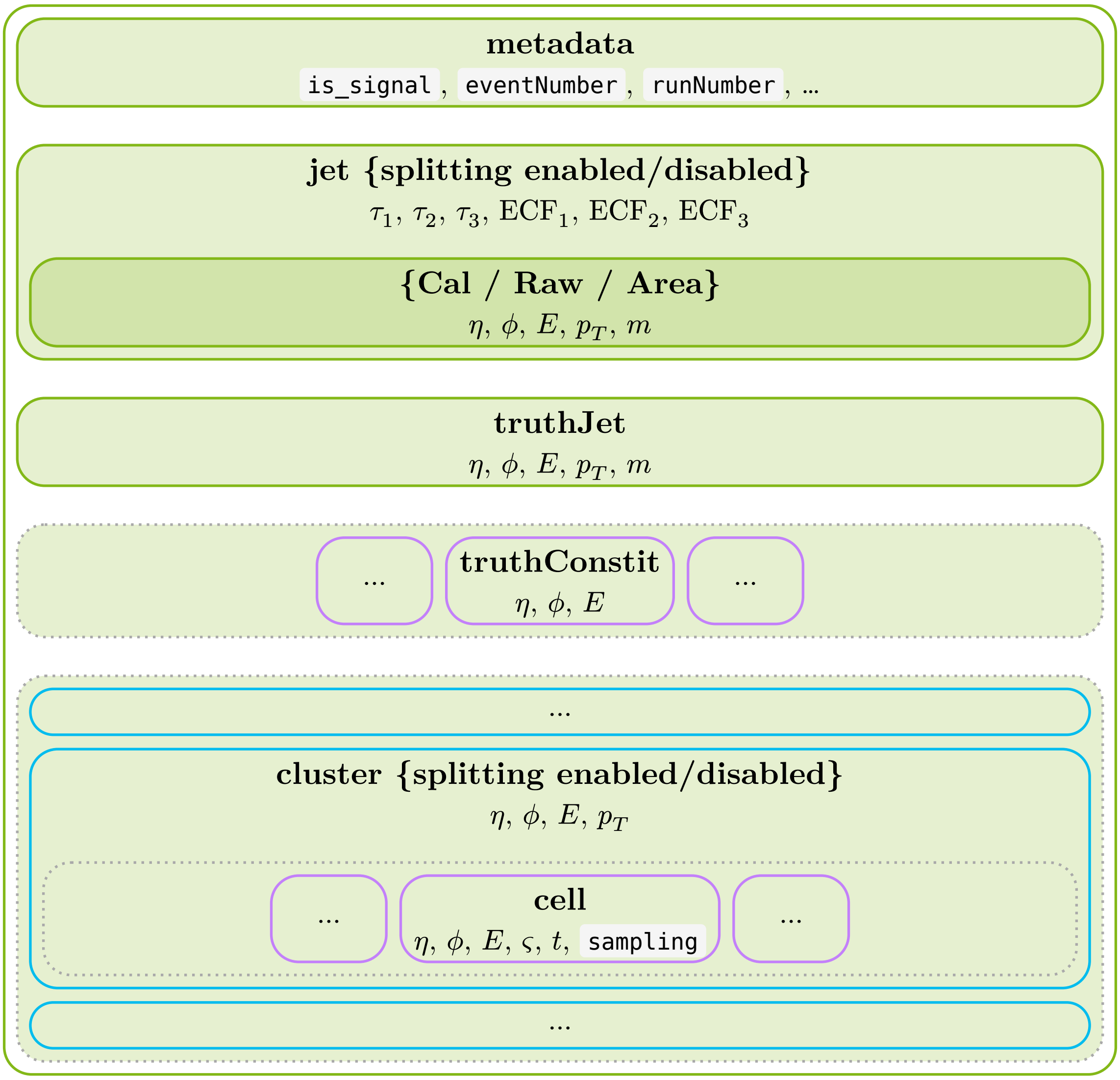

metadata
is_signal, eventNumber, runNumber, …
jet {splitting enabled/disabled}
τ_1, τ_2, τ_3, ECF_1, ECF_2, ECF_3
{Cal / Raw / Area}
η, ϕ, E, p_T, m
truthJet
η, ϕ, E, p_T, m
…
truthConstit
η, ϕ, E
…
…
cluster {splitting enabled/disabled}
η, ϕ, E, p_T
…
cell
η, ϕ, E, ς, t, sampling
…
…

Figure 18: Schematic representation of the data model for the Monte-Carlo samples considered in this analysis. Variations that contain the same data types are indicated with curly braces (e.g. `{splitting enabled/disabled}`).

3.3 Jet-Level Features and Substructure

To put the following studies on topo-cluster splitting into context and to obtain a baseline for comparison, other influences on jet-level features are examined first, including the jet transverse momentum (p_T) and pile-up conditions. Then, the effect of completely disabling topo-cluster splitting on jet features is investigated.

3.3.1 Baseline Reconstruction Performance

Figure 19 shows the N-subjettiness ratios τ_{21}^{truth} and τ_{32}^{truth} (defined in Section 2.2.1) for W' and Z' signal events with the background overlaid. As expected, W' events tend to have lower τ_{21}^{truth} values, coinciding with their 2-prong decay topology. Similarly, Z' events tend to have lower τ_{32}^{truth} values, as they originate from 3-prong decays. In the aforementioned cases, the background distribution peaks at higher values of the respective variable, indicating that these variables are effective at discriminating W' and Z' events from the background.

Figure 20 shows distributions of the same variables reconstructed from detector-level information. The distributions are less distinguishable from each other and from the background than with truth-level information, especially those of τ_{32}^{reco}. In contrast to Figure 19, peaks with values close to 0 are observed, corresponding to jets with few constituents, as can be seen in Figure 21.

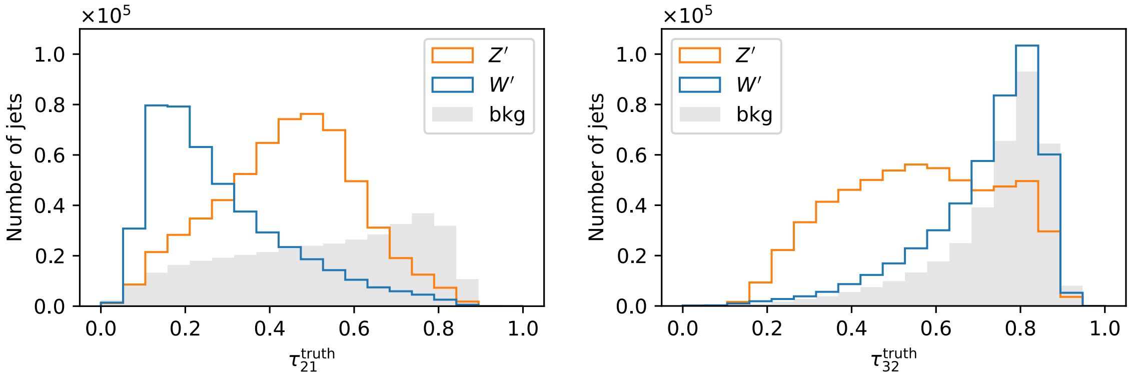


Figure 19: Histograms of τ_{21}^{truth} and τ_{32}^{truth} for W' and Z' signal / background events.

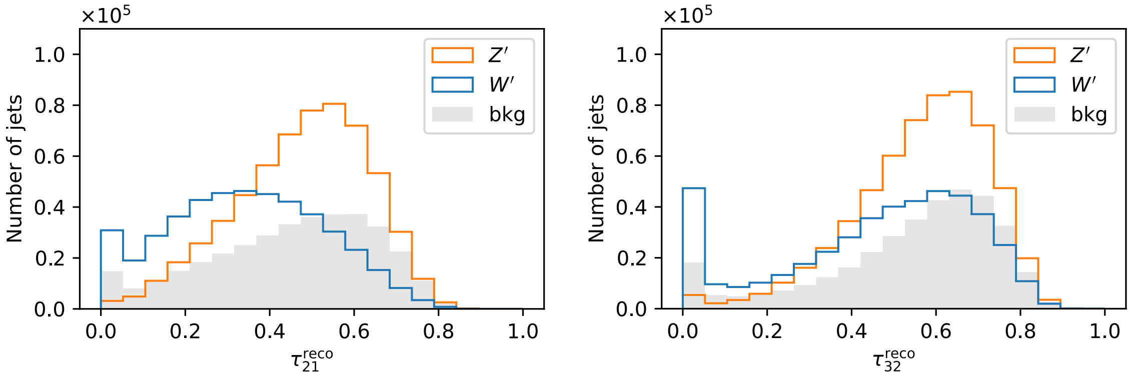


Figure 20: Histograms of τ_{21}^{reco} and τ_{32}^{reco} for W' and Z' signal / background events.

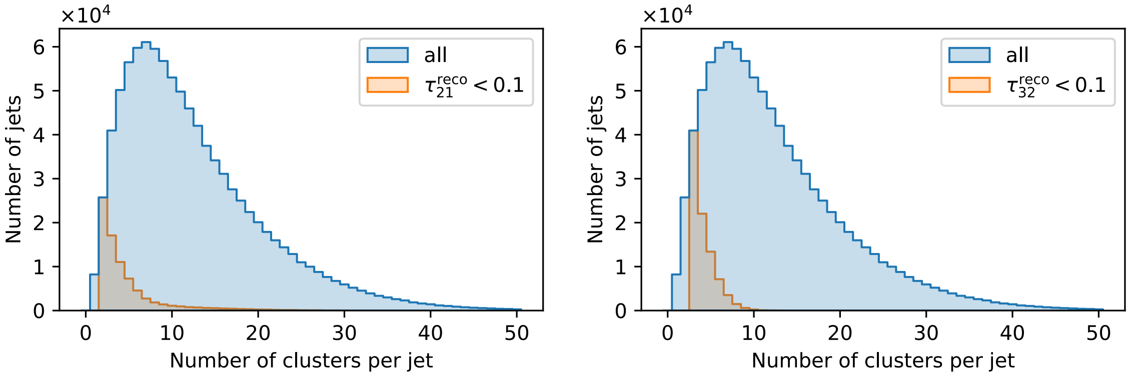


Figure 21: Histograms of the number of topo-clusters for jets with small τ_{21}^{reco} resp. τ_{32}^{reco} versus all jets. (W' + bkg)

Like τ_{21}, D_2 is sensitive to the two-prong substructure of jets originating from W' decays. Figure 22 shows that it is not bound to the $[0, 1]$ range and has similar overlap between signal and background as τ_{21} at both truth and reconstruction level. It is to be noted that a small fraction of infinite values is excluded from all histograms involving D_2.

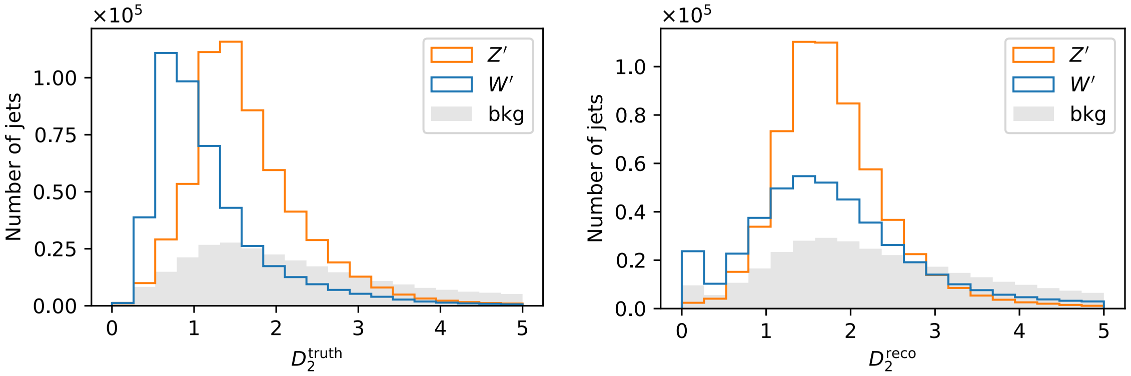


Figure 22: Histograms of D_2^{truth} and D_2^{reco} for W' and Z' signal / background events.

In Figure 23, three examples of jets with high and low τ_{21}^{truth} are given, sampled at different quantiles of the τ_{21} distribution. They are equally scaled to allow for visual comparison. The left and center examples show two high-density regions in close proximity, whereas the right one has a single high-density region.

Figure 24 shows similar examples for D_2^{truth}. The left one shows a jet with two well-separated subjets, the middle one shows two subjets in close proximity, and the right one has three subjets. This exemplifies the fact that high values of D_2^{truth} occur both for jets with a different number of prongs than three and for jets with no discernible substructure.

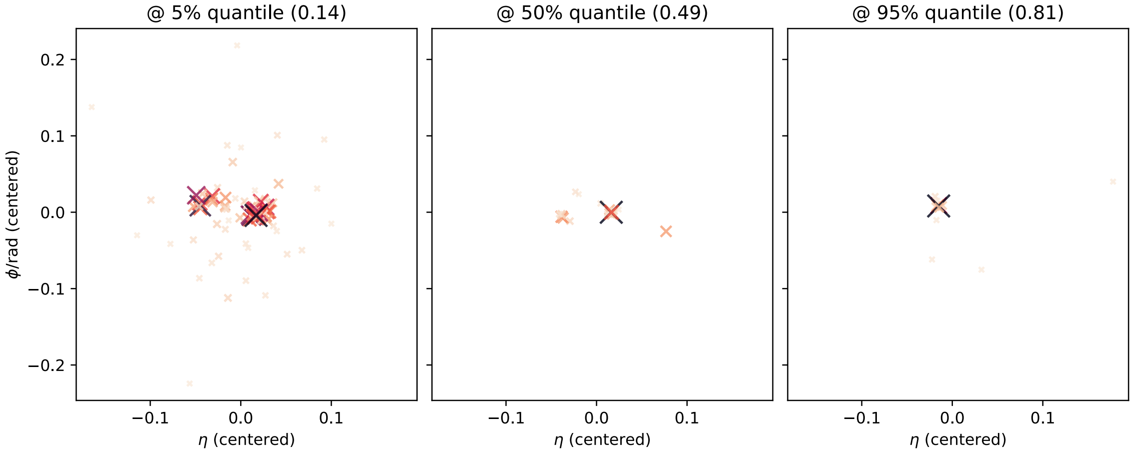


Figure 23: Scatterplots of the truth constituents of three jets at different quantiles of the τ_{21}^{truth} distribution. The × markers are scaled and colored according to their respective energy (larger and darker means higher energy). (W' + bkg)

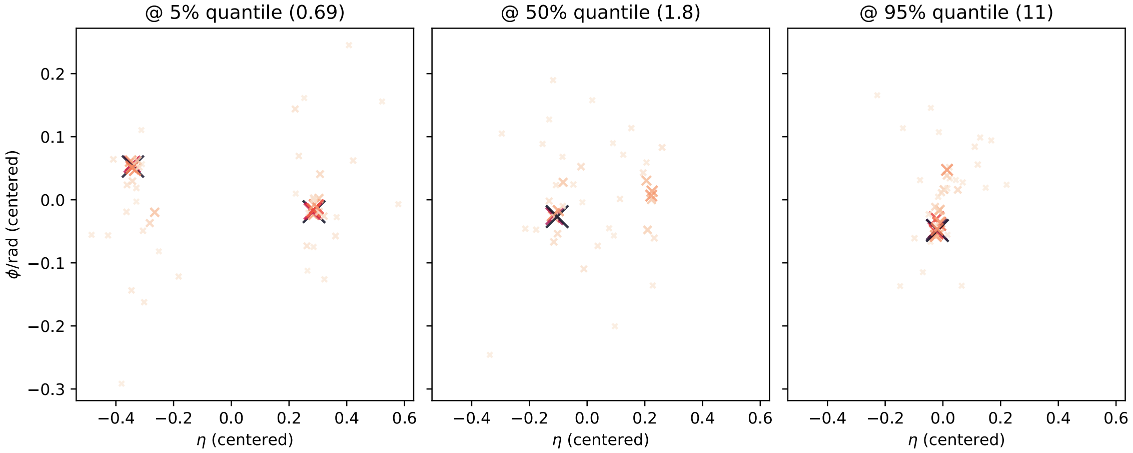


Figure 24: Scatterplots of the truth constituents of three jets at different quantiles of the D_2^{truth} distribution. (W' + bkg)

An overview of the discrimination power of the previously discussed variables is given in Figure 25, which shows ROC curves for the W'/background classification task using cuts on τ_{21}, τ_{32}, and D_2 on either truth or reconstruction level. While the performance of τ_{21} and D_2 is similar, τ_{32} is significantly less effective, due to being designed for three-prong rather than two-prong substructure. On Z'/background samples, shown in Figure 26, τ_{32}^{truth} outperforms all other variables by a large margin. At reconstruction level, however, its performance is degraded enough to be worse than that of τ_{21}^{reco} and D_2^{reco}.

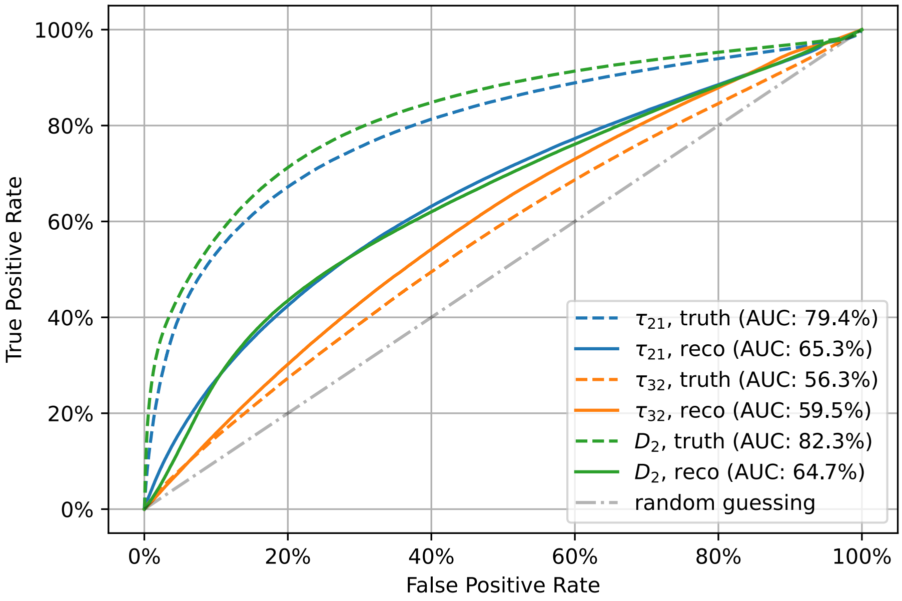


Figure 25: ROC curves for different substructure variables on truth and reconstruction level. (W' + bkg)

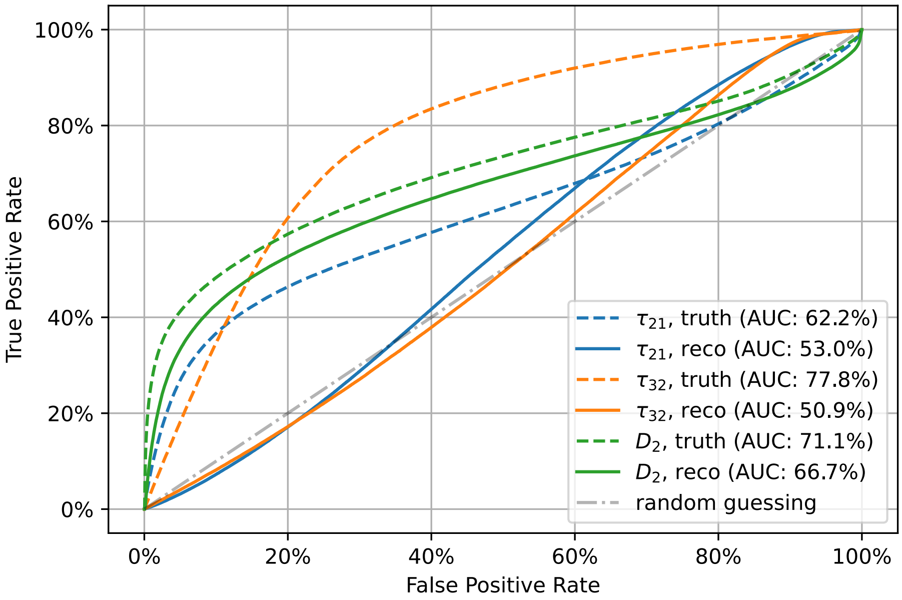


Figure 26: ROC curves for different substructure variables on truth and reconstruction level. (Z' + bkg)

3.3.2 Dependence on p_T

As outlined in Section 2, jets with larger p_T become more collimated, making it harder to resolve their substructure. This effect is studied here by comparing the distributions of τ_{21} and τ_{32} at at low and high p_T^{truth}, as well as the corresponding ROC curves.

Figure 27 shows that there is little effect of p_T^{truth} on the truth-level τ_{21}^{truth} distributions. The reconstructed τ_{21}^{reco} distributions in Figure 28 show a more pronounced shift towards lower values for both signal and background at high p_T^{truth}. In particular, the fraction of jets with $\tau_{21}^{\text{reco}} \approx 0$ increases sharply. As the background distribution is equally affected, this constitutes a loss of discriminating power.

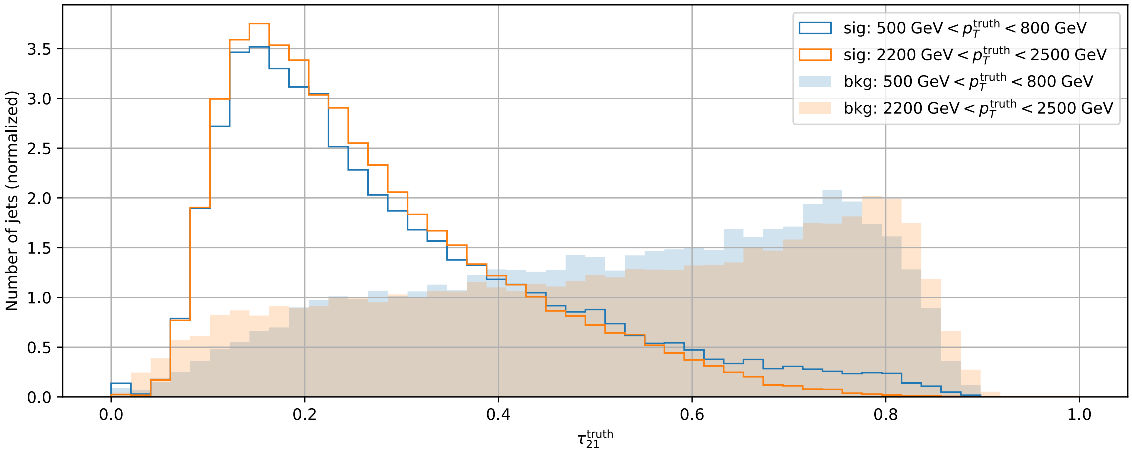


Figure 27: Histogram of τ_{21}^{truth} for high and low p_T^{truth}. Each of the four distributions is normalized to unit area. (W' + bkg)

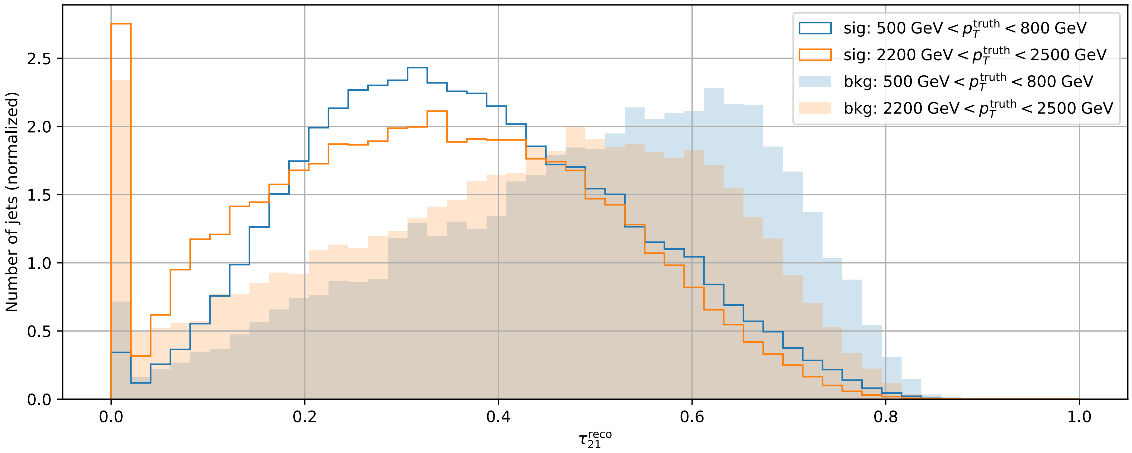


Figure 28: Histogram of τ_{21}^{reco} for high and low p_T^{truth}. (W' + bkg)

In contrast to τ_{21}, D_2^{truth} in Figure 29 is affected by p_T^{truth} differently for signal and background. While the background is slightly shifted towards lower values at high p_T^{truth}, the signal distribution is shifted towards higher values, leading to an increased separation between the two. The reconstructed D_2^{reco} distributions in Figure 30 show a more pronounced shift toward higher values at high p_T^{truth} for signal. As with τ_{21}^{reco}, a peak at $D_2^{\text{reco}} \approx 0$ appears, but it is dominated by background instead of signal.

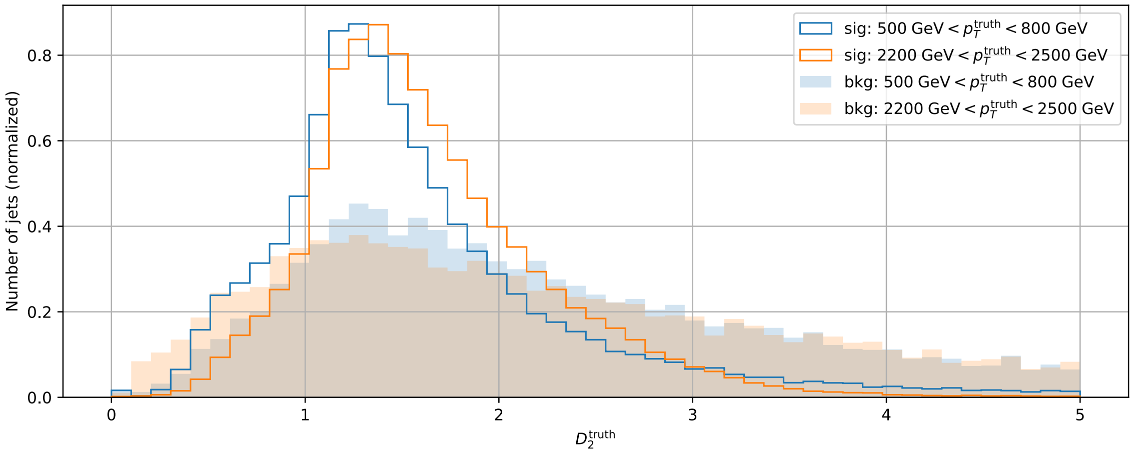


Figure 29: Histogram of D_2^{truth} for three p_T^{truth} ranges. (Z' + bkg)

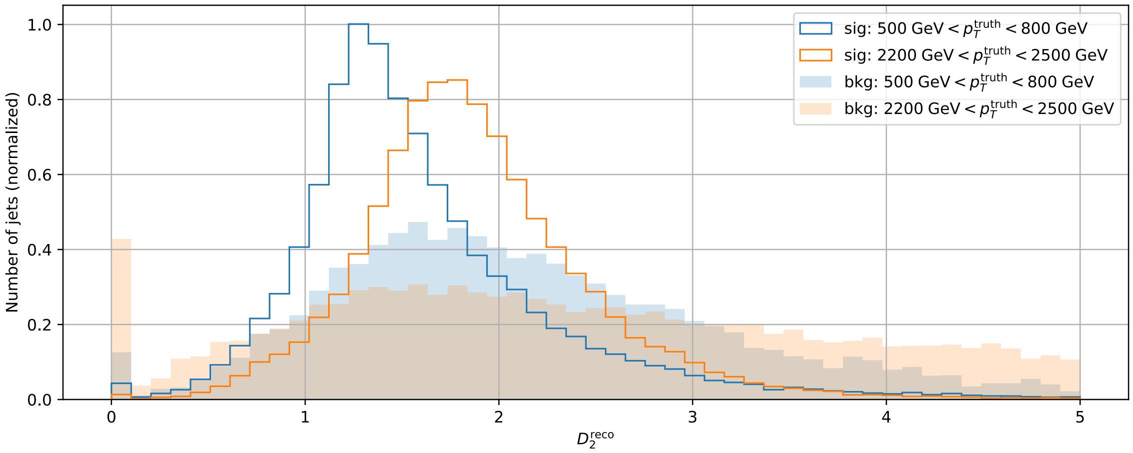


Figure 30: Histogram of D_2^{reco} for three p_T^{truth} ranges. (Z' + bkg)

In Figure 31, the changes to truth-level τ_{32} distributions with p_T^{truth} are shown. While the background distribution again remains largely unchanged, the mode of the signal distribution shifts from ≈ 0.8 to ≈ 0.5, possibly due to increased collimation.

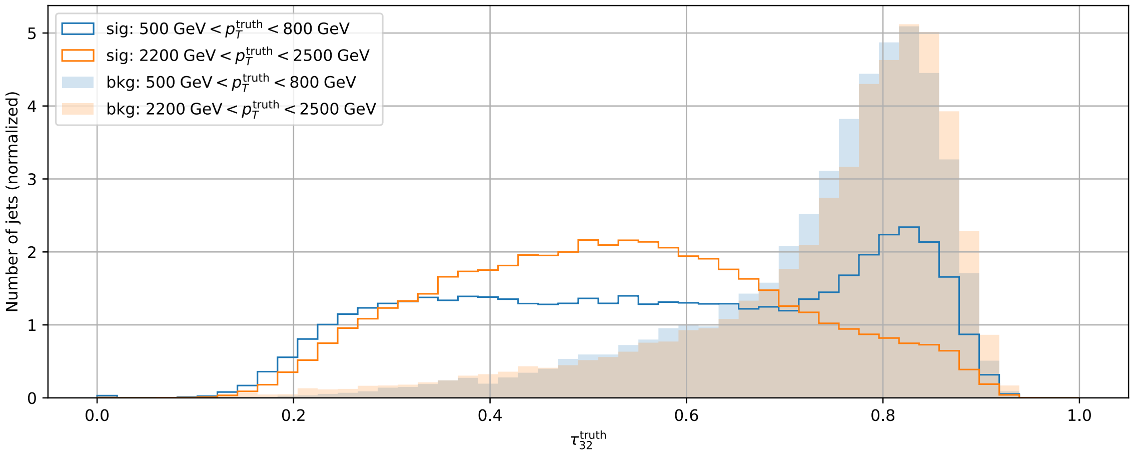


Figure 31: Histogram of τ_{32}^{truth} for three p_T^{truth} ranges. (Z' + bkg)

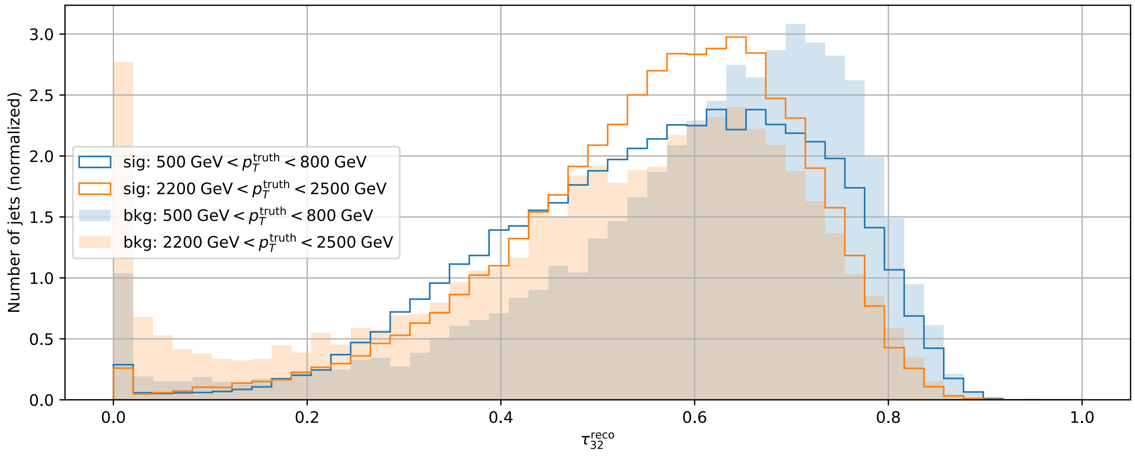


Figure 32: Histogram of τ_{32}^{reco} for three p_T^{truth} ranges. (Z' + bkg)

As reasoned above, the ROC curves in Figure 33 show a decrease in performance at high p_T^{truth} for both τ_{21} and τ_{32}. It is intuitively clear that this is due to the increased collimation of jets at high p_T, making it harder to resolve their substructure. D_2 shows a slight increase in ROC-AUC at high p_T^{truth}, due to changes in the signal distribution outlined above.

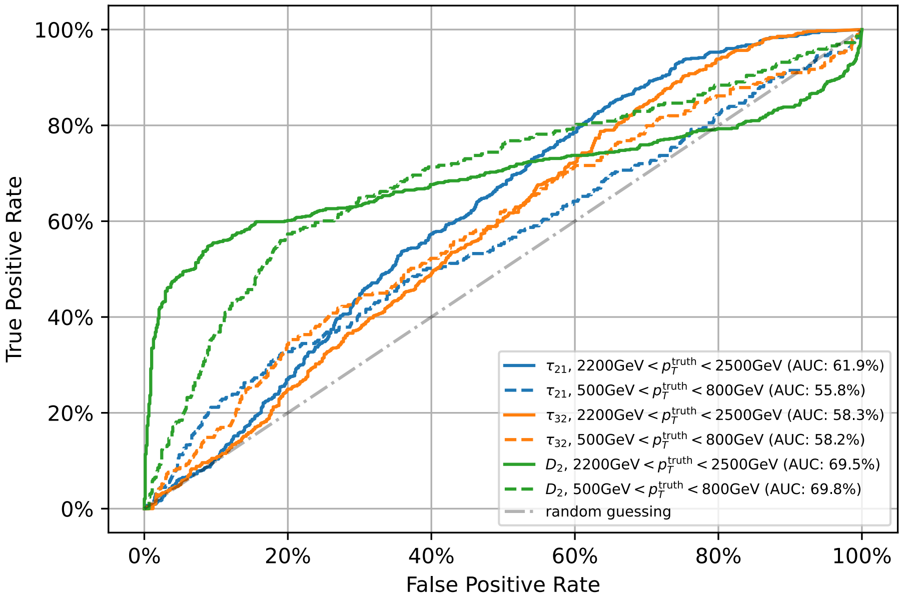


Figure 33: ROC curves for different substructure variables on truth and reconstruction level. (Z' + bkg)

3.3.3 Dependence on Pile-Up

The previous comparisons were performed on samples with realistic pile-up conditions. To isolate the effect of pile-up on jet substructure variables, these baseline samples are compared to equivalent samples without pile-up.

Qualitatively, the removal of pile-up is expected to improve the resolution of jet substructure variables. Indeed, for all three considered variables, a shift towards lower values is observed that is larger for signal than for background, as shown in Figure 34. Truth-level variables are unaffected by pile-up by design and not presented here.

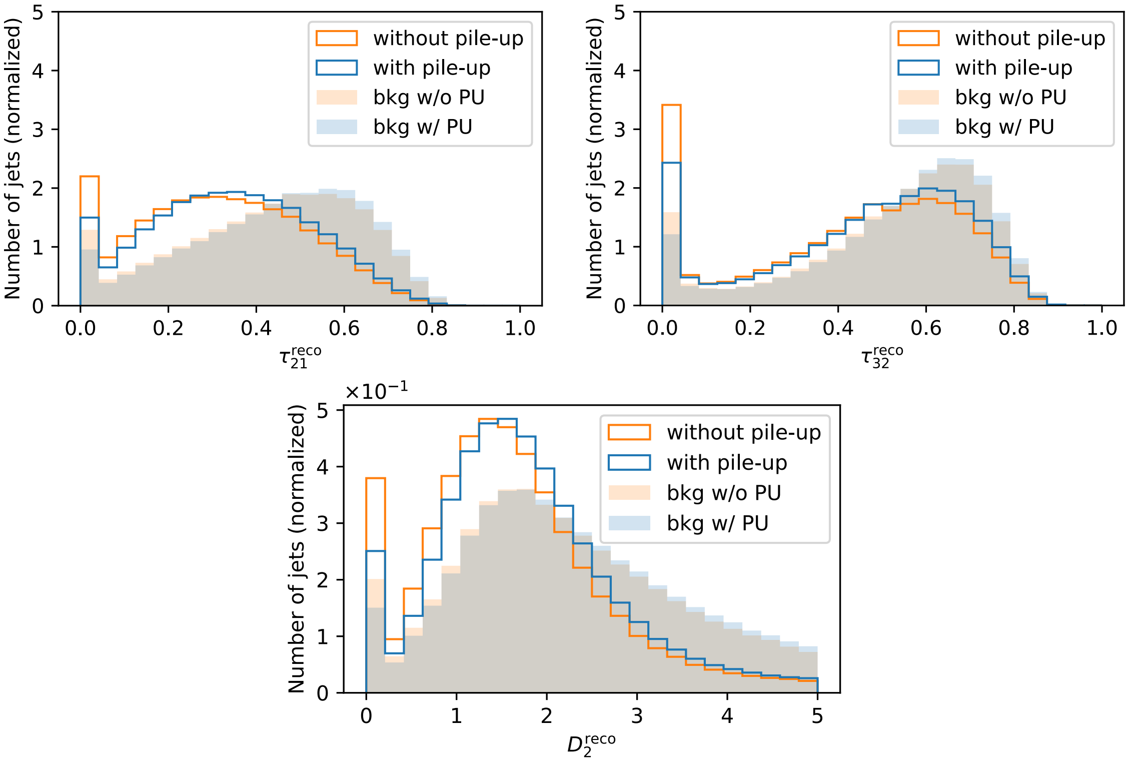


Figure 34: Histograms of τ_{21}^{reco}, τ_{32}^{reco}, and D_2^{reco} with and without pile-up. (W' + bkg)

Figure 35 compares the ROC curves of the substructure variables with and without pile-up for the W' + bkg samples. As was shown in Figure 25, τ_{21}^{reco} and D_2^{reco} have similar performance, while τ_{32}^{reco} is not suitable for 2-prong jets. Pile-up degrades the performance of all three variables, but only by a small amount. In Figure 36, τ_{32} is again outperformed by D_2, but D_2 *gains* discriminating power with pile-up. This might be due to it being used “the wrong way”, i.e. to tag 3-prong instead of 2-prong jets, while still outperforming the other variables, as the classification task at hand is only between ≥1-prong signal and 1-prong background.

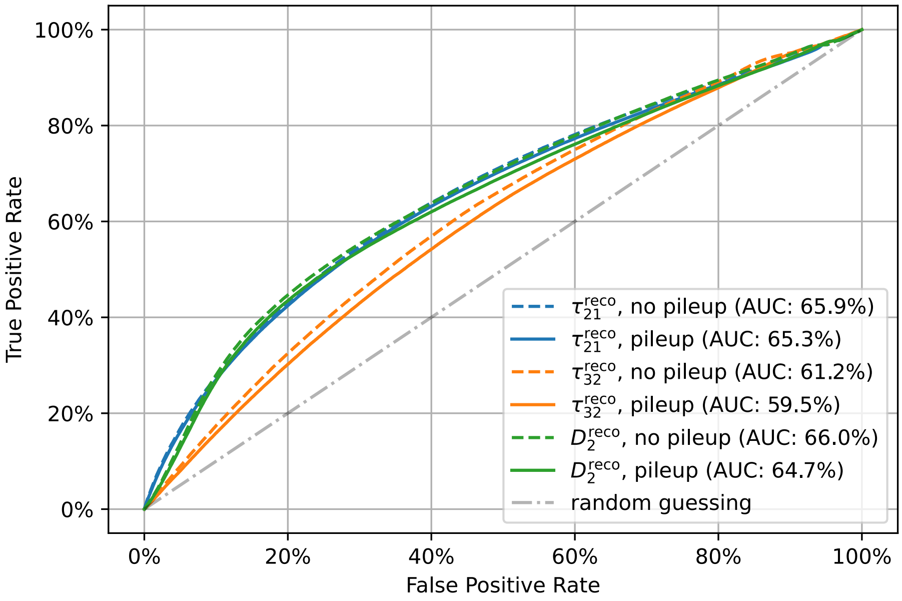


Figure 35: ROC curves for various substructure variables, with and without pile-up. (W' + bkg)

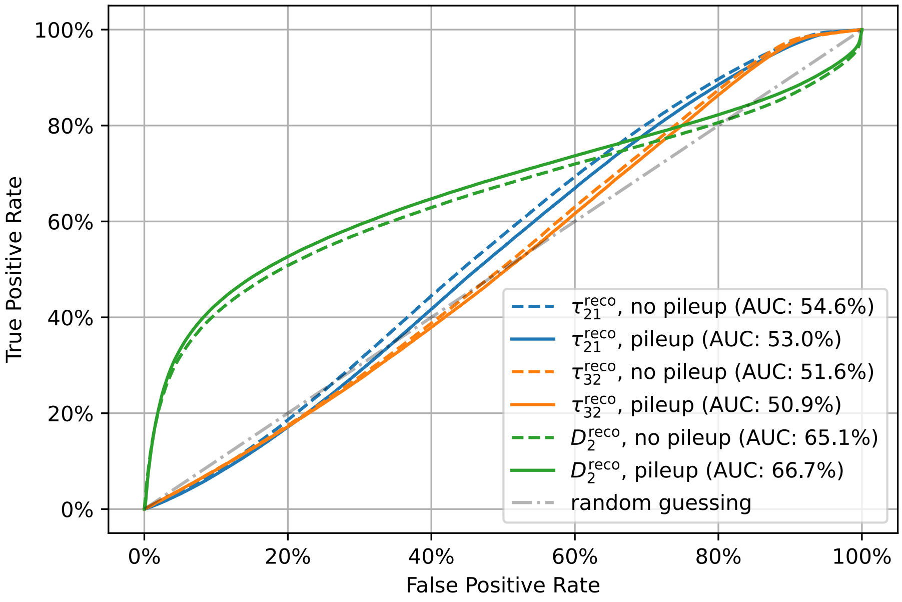


Figure 36: ROC curves for various substructure variables, with and without pile-up. (Z' + bkg)

3.3.4 Dependence on Topo-Cluster Splitting

The most evident impact of topo-cluster splitting (described in Section 2.1.3) is in the number of clusters per jet, histogrammized in Figure 37. With splitting, the mean number of clusters per jet increases from 2.169 to 12.468. A per-jet comparison in Figure 38 shows that this increase is not simply proportional to the original number of clusters. While almost always matching or exceeding the original number of clusters,[10] the number of clusters with splitting enabled is largely independent of the original number of clusters. Without splitting, many jets contain only a single cluster, which is detrimental to the discriminative power of substructure variables, as they require at least two constituents to be meaningful. This can be seen in Figure 39: Both τ_{21} and τ_{32} collapse towards 0 without splitting, while with splitting, the distributions are much closer to the truth-level ones. Accordingly, the ROC curves in Figure 40 are close to random guessing without splitting. The reconstruction error of the jet axis, as measured by the ΔR between truth and reconstructed jets, is not affected by the splitting, while the jet energy tends to be underestimated without splitting, as shown in Figure 41.

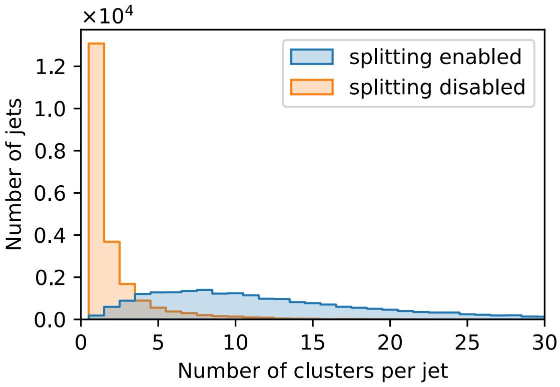


Figure 37: Distribution of the number of clusters per jet with splitting enabled/ disabled. (W' + bkg)

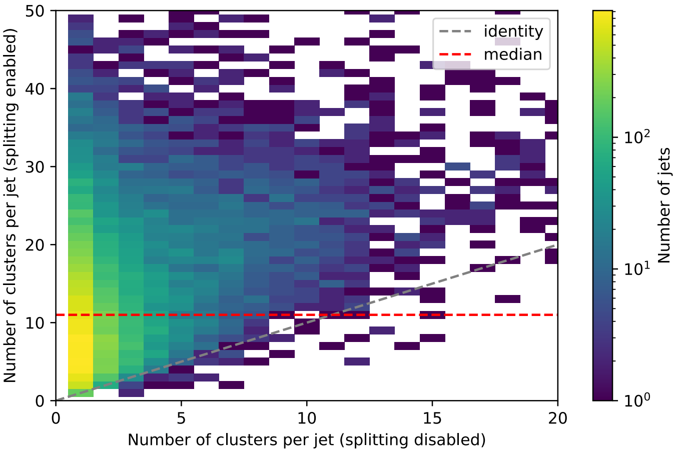


Figure 38: 2D Distribution of the number of clusters with splitting enabled/disabled. (W' + bkg)

[10] The few exceptions, located below the identity line in the figure, are presumably the result of trimming.

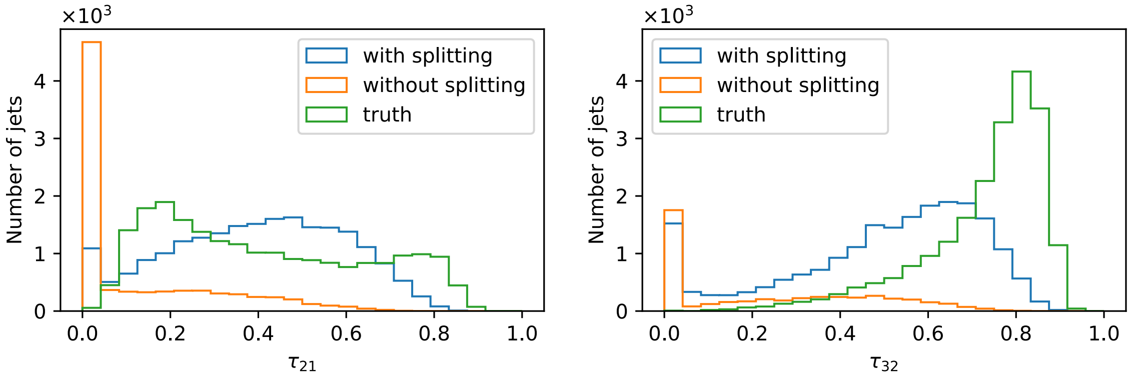


Figure 39: Histograms of τ_{21} and τ_{32} of signal events with topo-cluster splitting enabled / disabled, compared to the truth-level distributions. (W')

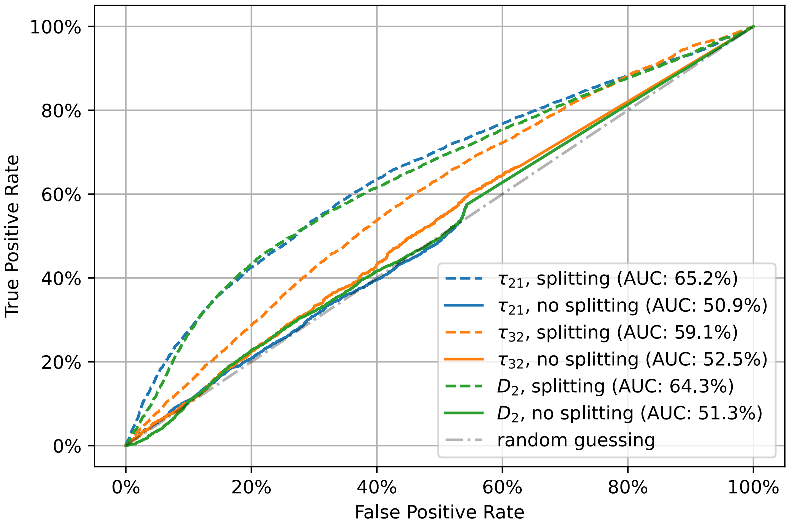


Figure 40: Comparison of ROC curves for substructure variables with and without topo-cluster splitting. (W' + bkg)

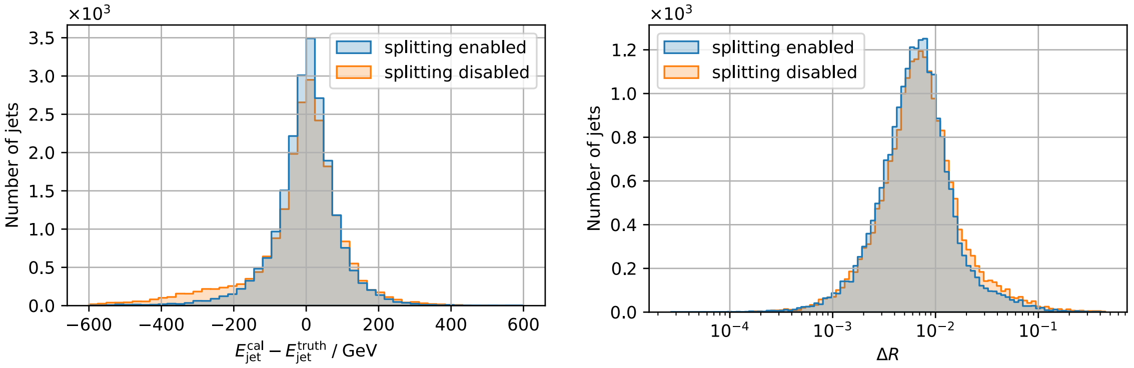


Figure 41: Comparison of jet energy and angular resolution with splitting enabled/ disabled. (W' + bkg)

3.3.5 Upper Limit on Performance with Binary Splitting

The given data only contains jet-level variables either with topo-cluster splitting enabled or with topo-cluster splitting disabled.[11] Without modifications to the clustering algorithm itself, which are outside the scope of this thesis, only these two extreme cases can be studied here.

A hypothetical algorithm operating on the given data could decide for each jet whether to return the jets/variables that were obtained with splitting enabled or disabled. An upper limit on the performance gain achievable by such an algorithm can be obtained by always choosing the better option for each jet using Monte-Carlo truth information.

Figure 42 shows the distribution of the τ_{21} and τ_{32} substructure variables with splitting enabled, disabled, and the hypothetical best-of-both. It can be seen that while disabled splitting yields significantly higher errors, the best-of-both approach only improves slightly upon always choosing to enable splitting. In fact, on a per-jet basis, only 8 % of jets would benefit from having splitting disabled instead of enabled in the case of τ_{21}, 4.55 % in the case of D_2, and 3.17 % for τ_{32}. It is to be noted that this is an extreme case, meaning that algorithms that actually modify the splitting behavior beyond a binary decision could achieve much higher performance gains.

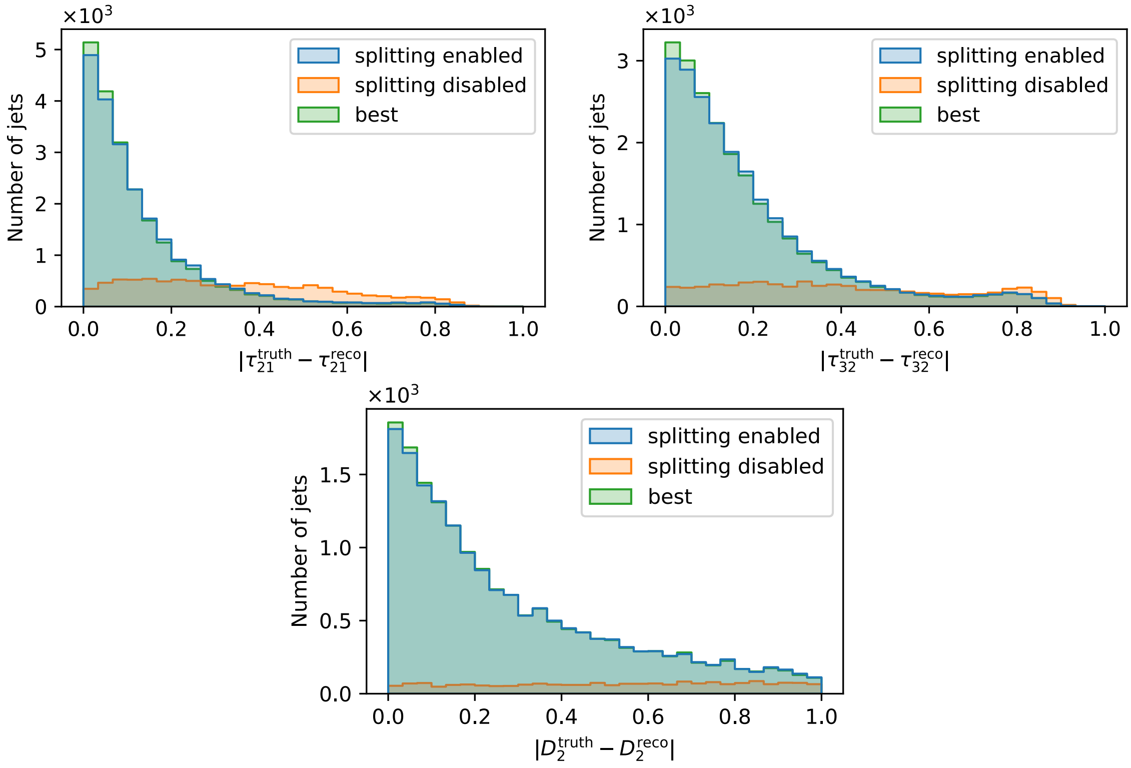


Figure 42: Distributions of the absolute error (compared to Monte-Carlo truth) for splitting enabled, disabled, and the hypothetical best-of-both. (Z' + bkg)

[11] Outputs for variations of the splitting algorithm are used in a grid search for hyperparameter optimization, but each sample corresponding to a hyperparameter set contains distinct events. Hence, no direct comparison is possible.

3.3.6 Number of Constituents

Figure 43 shows a 2D density plot comparing the number of reconstructed constituents (i.e. topo-clusters) to the number of truth constituents. Evidently, the number of truth constituents is substantially higher than the number of reconstructed constituents; almost all events are below the identity line (gray dashed), hinting at the lossy nature of the reconstruction process. Though the distributions (or rather the areas they cover) are similar for signal and background events, the signal events tend to have more truth- and – proportionally – more reconstructed constituents. Figure 44 shows the same for W' signal events; in this case, the difference between signal and background is less pronounced. This difference in the number of constituents can be attributed at least partially to the different decay topologies of W' and Z' bosons (2-prong vs. 3-prong decays).

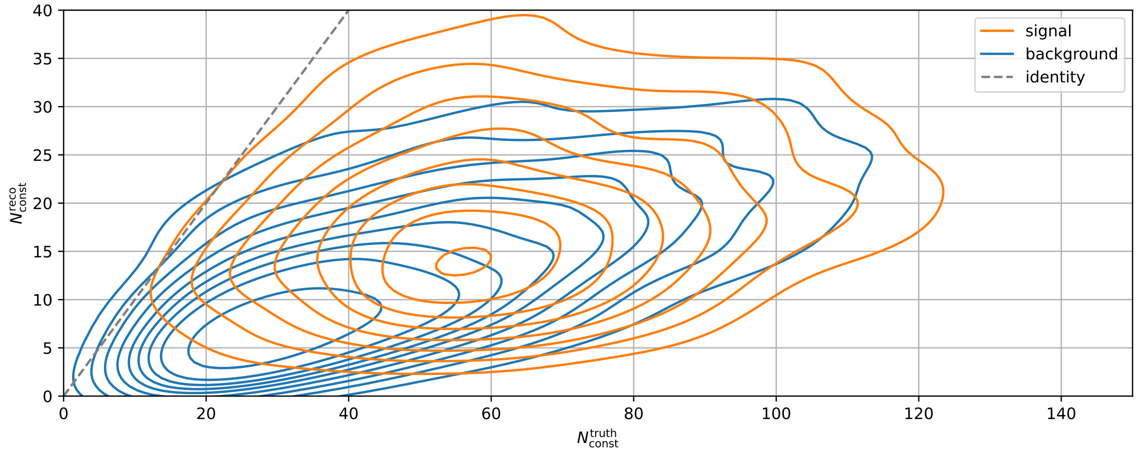


Figure 43: 2D density plot of the number of reconstructed constituents vs. the number of truth constituents. As with topographical maps, lines connect points of equal density; areas surrounded by more lines are denser. The dashed line represents the identity, where the number of reconstructed constituents equals the number of truth constituents. (Z' + bkg)

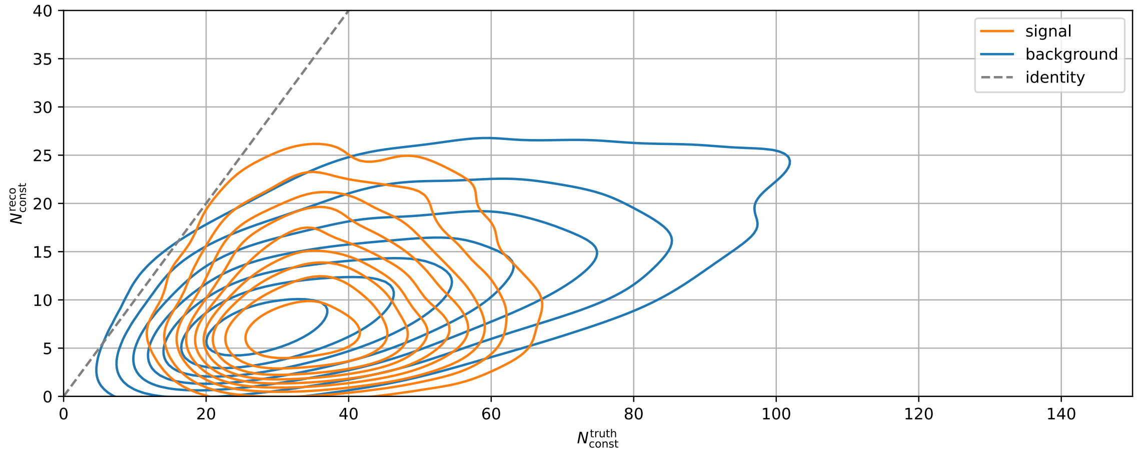


Figure 44: 2D density plot of the number of reconstructed constituents vs. the number of truth constituents. (W' + bkg)

3.4 Cluster-Level Features

Having studied the jet-level substructure variables, the next step is to investigate features describing the individual clusters within the jets. Contrary to the previous studies, no additional samples are available to study the impact of pile-up. Furthermore, while some truth information is available at the cluster level, there is typically no corresponding reconstructed quantity. The focus will thus be on the dependence of cluster-level features on p_T and the number of clusters per jet. Impacts of topo-cluster splitting are examined using more complex methods and will be discussed in Section 3.5.

In order to reason about one-dimensional distributions of cluster-level features, the features need to be aggregated first, so that a cluster-level feature is described by a single scalar *per jet*. This can be done in various ways, e.g. by taking the mean or median, minimum or maximum, sum, etc. Alternatively, the features of all clusters in a jet can be concatenated (flattened), yielding $\sum_{j \in \text{ jets}} n_{\text{clus},j}$ instead of n_{jets} values. As the choice of aggregation method can have a significant impact on the resulting distributions, it will be specified in all following plots.

3.4.1 Dependence on p_T

Because of the large number of cluster-level variables available, an EMD-based comparison (see Section 3.1.1) is performed to find the variables most (and least) affected by changes in p_T. Variables in this subsection are concatenated before EMD calculation.

A distinction needs to be made between the p_T of a whole jet and the p_T of individual clusters within. Results of a comparison based on the former are given in Table 1. They indicate increased significances for the overall cluster (`SIGNIFICANCE`) and for the dominant cell within a cluster (`CELL_SIGNIFICANCE`), as well as increased mass and energy. Even though higher-p_T should penetrate deeper into the calorimeter, the average cartesian distance of a cluster from the nominal vertex (`CENTER_MAG`) decreases with increasing jet p_T, presumably because of the parallel decrease in $|\eta|$ combined with the barrel geometry of the calorimeter. Meanwhile, Table 2 shows that high-p_T clusters tend to contain a significantly larger fraction of the jet's energy (`fracE`) and are more likely to be surrounded by other clusters (lower `ISOLATION` ; defined in [7]).

As features of particular interest for later investigations, `fracE` and the number of cells per cluster are also given as histograms in Figure 45 and Figure 46, respectively.

Section A.2 in the appendix lists all cluster-level variables used in this work along with their definitions and descriptions. Complete tables of EMD results for all variables are available in Section F.

Table 1: Normalized EMDs between cluster-level variables with high/low $p_{T\,\mathrm{jet}}^{\mathrm{truth}}$. ($W'$ + bkg)

variable	unit	EMD	mean	mean
			($p_{T\,\mathrm{jet}}^{\mathrm{truth}} < 800\,\mathrm{GeV}$)	($p_{T\,\mathrm{jet}}^{\mathrm{truth}} > 2\,000\,\mathrm{GeV}$)
CENTER_MAG	mm	0.594	2897.664	2352.156
CELL_SIGNIFICANCE		0.566	54.704	340.264
MASS	GeV	0.558	1198.655	13 324.379
ENG_CALIB_TOT	GeV	0.551	28.808	310.518
SIGNIFICANCE		0.548	27.426	161.036

Table 2: Normalized EMDs between cluster-level variables with high/low $p_{T\,\mathrm{cluster}}^{\mathrm{reco}}$. ($W'$ + bkg)

variable	unit	EMD	mean	mean
			($p_{T\,\mathrm{cluster}}^{\mathrm{reco}} < 1\,\mathrm{GeV}$)	($p_{T\,\mathrm{cluster}}^{\mathrm{reco}} > 40\,\mathrm{GeV}$)
ENG_CALIB_OUT_L	GeV	1.673	0.113	5.796
ENG_CALIB_OUT_T	GeV	1.377	0.252	2.298
fracE		1.224	0.000	0.279
ISOLATION		1.208	0.660	0.326
LATERAL		1.206	0.483	0.859

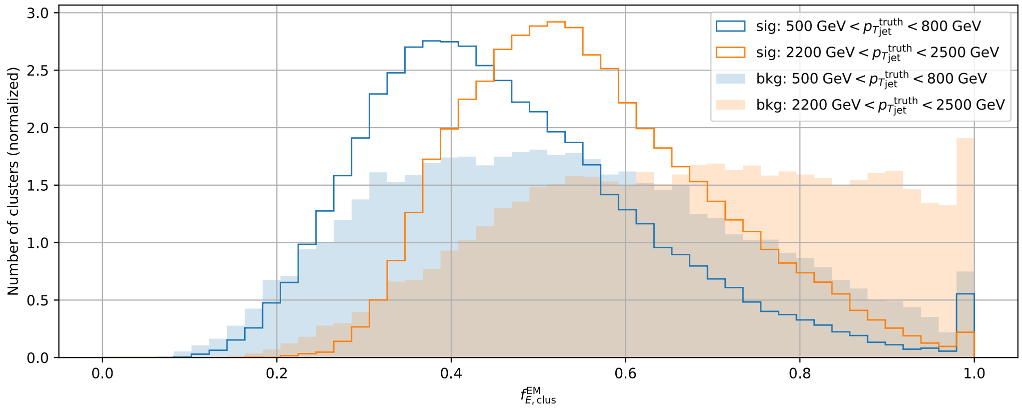


Figure 45: Histogram of the fraction of jet energy contained in a cluster for high and low $p_{T\,\mathrm{jet}}^{\mathrm{truth}}$. Each of the four distributions is normalized to unit area. (W' + bkg)

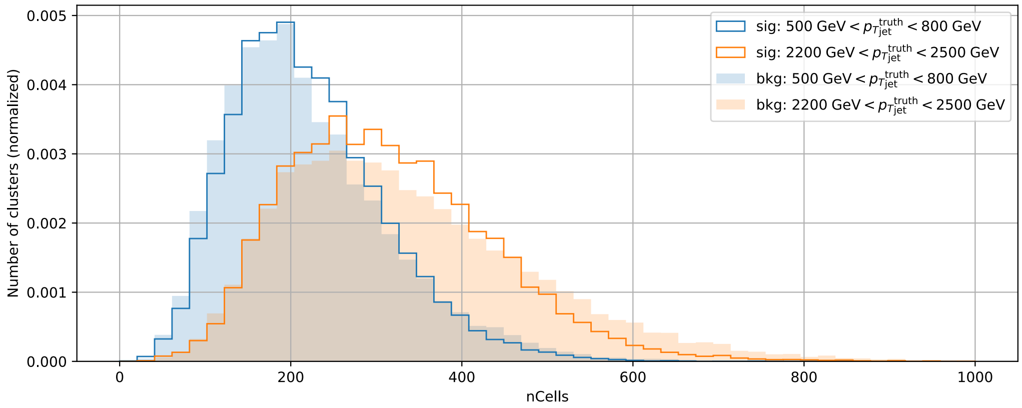


Figure 46: Histogram of the number of cells in a cluster for high and low $p_{T\,\mathrm{jet}}^{\mathrm{truth}}$. ($W'$ + bkg)

3.4.2 Dependence on N_{cluster}

Another dependency to consider is that on the number of clusters per jet, n_{clus}. It is largely independent of the jet's p_T, The distributions of cluster-level features for high and low n_{clus} are compared in Table 3. Unsurprisingly, the fraction of jet energy per cluster (`fracE`) decreases with many clusters. The same can be said about the mass per cluster (`MASS`), decreasing nearly 10-fold. The cluster's distance from the nominal vertex (`CENTER_MAG`) increases; showers that penetrate deeper into the calorimeter tend to form more clusters.

Table 3: Normalized EMDs between cluster-level variables for jets with high/low number of clusters. (Z' + bkg)

variable	unit	EMD	mean	mean
			($n_{\text{clus}} \leq 8$)	($n_{\text{clus}} \geq 40$)
ENG_CALIB_OUT_T	GeV	1.189	1.653	0.825
CENTER_MAG	mm	0.768	2401.060	3161.950
fracE_ref		0.766	0.150	0.016
fracE		0.764	0.179	0.021
MASS	GeV	0.709	14 230.989	1560.402

As can be seen in Figure 47, the distribution of the first η moment of clusters shows pronounced peaks at $\eta \approx \pm 1.4$, for jets with a high number of clusters, coinciding with the transition region between the barrel and endcap calorimeters (see Figure 4). This underscores the influence of detector geometry on cluster formation and gives an example of underlying correlations that may affect the following analyses.

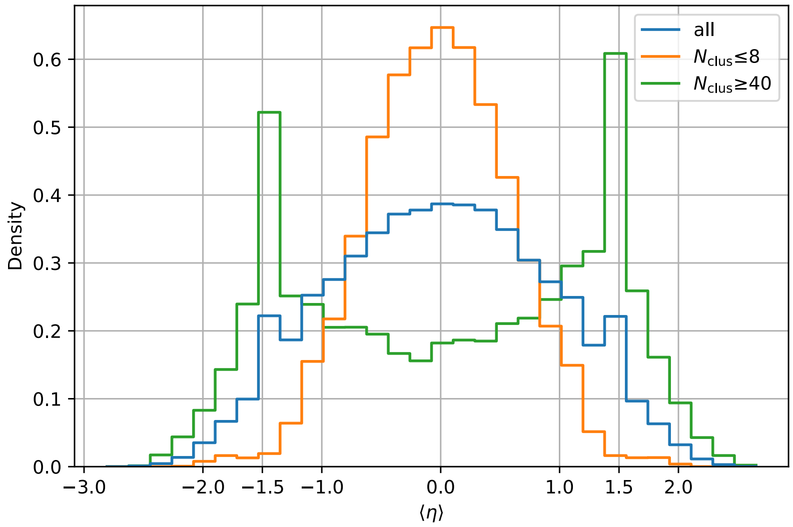


Figure 47: Distribution of the first η moment of clusters, for different numbers of clusters per jet.

3.5 Comparison of Split and Non-Split Clusters

After considering the impact of topo-cluster splitting on jet-level variables in Section 3.3.4, in this section, the effect of topo-cluster splitting on cluster-level variables is studied. For this, two distinct aspects can be considered: On one hand, the distributions of *any* variable with topo-cluster splitting enabled/disabled can be compared. On the other hand, the distributions of *cluster-level* variables can be compared for clusters that have or have not been split by the algorithm. The latter approach gives insights into the "decision" of the algorithm, but requires a matching of split and non-split clusters because information about them is stored separately (see Section 3.2.1). Both approaches are pursued in the following.

3.5.1 Splitting Enabled/Disabled

The most different cluster-level variables with/without topo-cluster splitting are those directly related to the cluster size and shape, as shown in Table 4.

As the average size of clusters decreases with splitting, so do the number of cells with positive / arbitrary energies in a cluster (`nCells` / `nCells_tot`) and fraction of $E_{\text{jet}}^{\text{truth}}$ contained in the cluster (`fracE<Calib>_ref`). The truth-level energy deposited in the calorimeter, but outside of the associated cluster (`ENG_CALIB_OUT_L`)[12], is also expected to scale inversely with the number of clusters, but highlights the negative impact of disabling splitting on the energy resolution.

Table 4: Normalized EMDs between cluster-level variables with and without topo-cluster splitting enabled. See Table A13 for all variables.

variable	unit	EMD	mean (no splitting)	mean (splitting)
nCells_tot		1.480	701.013	137.022
nCells		1.460	539.010	106.664
fracECalib_ref		1.363	0.443	0.077
ENG_CALIB_OUT_L	GeV	1.358	11.535	1.998
fracE_ref		1.356	0.385	0.065

3.5.2 Matching Split Clusters to Non-Split Clusters

As motivated before, the comparison of topo-clusters that were and were not split by the algorithm requires a matching of split clusters to their respective non-split clusters. For clarity, the terms "pre-splitting" and "post-splitting" will be used to refer to clusters obtained with the splitting algorithm disabled / enabled, respectively. This is to distinguish them from "split" and "non-split" clusters, which shall refer to a post-splitting clusters' intrinsic property of having been split or not. A set-theoretic approach to this problem is given below:

For brevity, let n_{pre} / n_{post} be the number of clusters in a jet pre- / post-splitting. Note that post-splitting does not imply that all clusters were actually split, but $n_{\text{pre}} \leq n_{\text{post}}$ should always hold. Let $X_{\text{pre},i}$ / $X_{\text{post},j}$ be the set of (fully identifying) cell coordinates

[12] `_L` stands for "loose" matching, i.e. $\Delta\alpha = 1.0$. See `CaloClusterMoment.h` ° in [74].

$(\varphi, \eta, \text{sampling})$ in pre- / post-splitting clusters i / j. Then, three cases of matching can be distinguished:

- If $X_{\text{pre},i} \cap X_{\text{post},j} = \emptyset$, there is no match.
- If $X_{\text{pre},i} \supseteq X_{\text{post},j}$, there is a (perfect) match.
- Else, if $X_{\text{pre},i} \cap X_{\text{post},j} \neq \emptyset$, there is a partial match. This is a theoretical possibility only; it does not occur in the present samples.

Finally, m_i $(i \in [1, ..., n_{\text{pre}}])$ is the number of post-splitting clusters that were matched to the i-th pre-splitting cluster according to the above definition. It can be 0 (no match), 1 (perfect match) or >1 (split). Based on the idea that topo-cluster splitting should not change the overall content of a jet,

$$\left(\bigcup_{n_{\text{post}}} X_{\text{post}} \right) \subseteq \left(\bigcup_{n_{\text{pre}}} X_{\text{pre}} \right) \tag{20}$$

and $m_i \geq 1 \forall i$ should hold true.

However, both conditions are found to be violated in some cases. Examples of this are shown in Figure 48: New clusters appear to be created that do not match any pre-splitting cluster, while some sections of pre-splitting clusters do not match any post-splitting cluster. Though no reason for this behavior is demonstratately identified, it is assumed to be the result of trimming (see Section 2.1.7 and Figure 48) that was applied independently to the jets with and without splitting enabled.

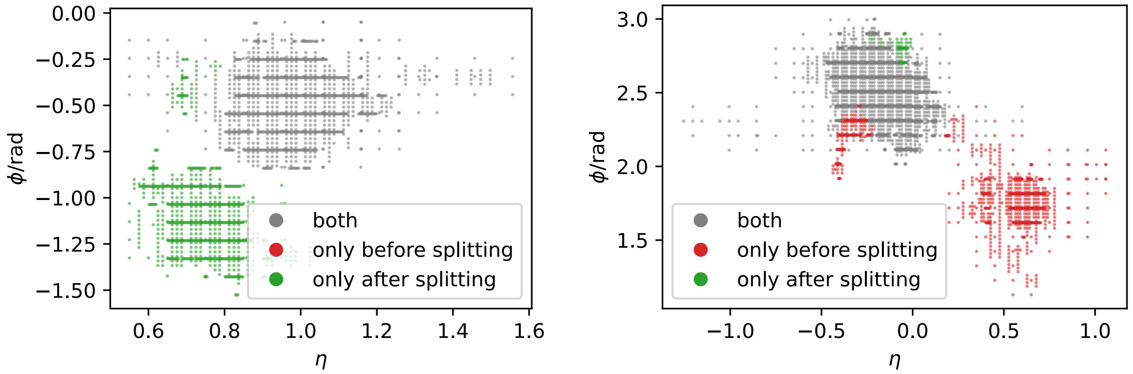


Figure 48: Two examples of jets with a mismatch between split and non-split clusters. Gray cells are present in both pre- and post-splitting clusters, while red / green cells are only present in pre- / post-splitting clusters, respectively.

A histogram of m_i, given in Figure 49, shows that most pre-splitting clusters are matched to exactly one post-splitting cluster, indicating that they were not split by the algorithm. Values of $m_i > 1$, accounting for 44.51 % of all pre-splitting clusters, indicate that the respective pre-splitting cluster was split into m_i clusters. Values of $m_i = 0$, indicate that the respective pre-splitting cluster was not matched to any post-splitting cluster. This applies to about 7.17 %.

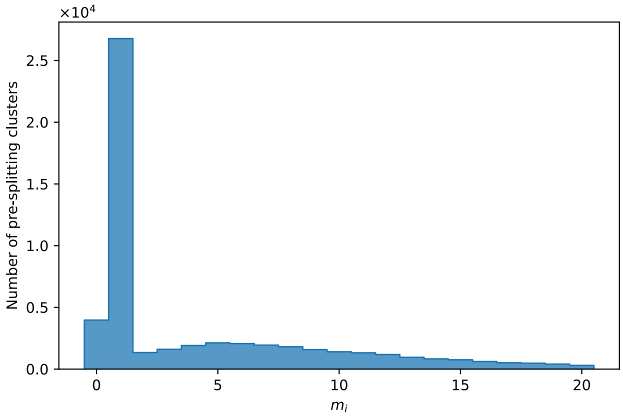


Figure 49: Histogram of the number m_i of matched split clusters per pre-splitting cluster. (Z' + bkg)

3.5.3 Matching-Based Comparisons

To evaluate the correctness of the matching procedure, the number of clusters post-splitting n_{post} is compared to the sum of the number of matched split clusters per pre-splitting cluster $\sum_i m_i$ in Figure 50. If the matching worked perfectly, these two numbers would be equal for each jet, and the two distributions would be identical. However, it is observed that n_{post} tends to be larger than $\sum_i m_i$, again indicating that some split clusters are not matched to any non-split cluster, presumably due to the aforementioned grooming algorithm.

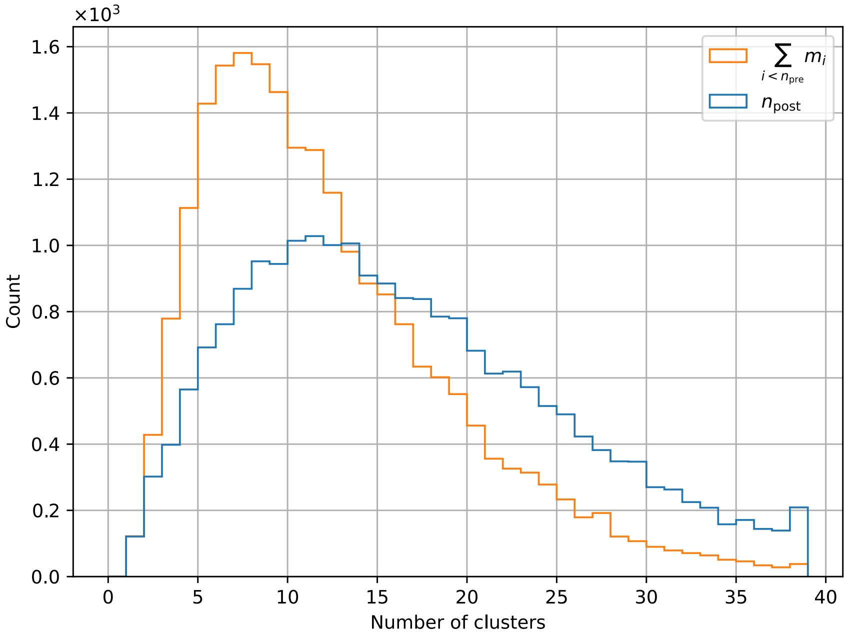


Figure 50: Histograms of the number of clusters per jet with splitting enabled n_{post} and the sum of the number of matched split clusters per pre-splitting cluster $\sum_i m_i$. (Z' + bkg)

With the aforementioned issues in mind, a comparison of cluster-level features is performed between clusters that were or were not split, as indicated by $m_i = 1$ and $m_i \geq 2$, respectively. The largest resulting normalized EMDs are shown in Table 5. Similarly to Section 3.5.1, the largest differences are observed in features that scale with the cluster size, such as `fracE` and `sumCellE`, the sum of positive EM-scale cell energies in the cluster. `SECOND_LAMBDA`[13] is also increased approximately 30-fold, indicating that non-split clusters tend to be more elongated in the direction of their principal axis.

Table 5: Normalized EMDs between cluster-level variables for non-split clusters matching 1 or ≥ 2 split clusters.

variable	**unit**	**EMD**	**mean**	**mean**
			$(m_i = 1)$	$(m_i \geq 2)$
fracE		1.885	0.013	0.924
fracE_ref		1.882	0.011	0.779
SECOND_LAMBDA	mm^2	1.552	16 065.308	452 754.603
ENG_POS	GeV	1.526	16 259.747	1 744 003.767
sumCellE	GeV	1.526	16.231	1743.671

As cluster-level variable with the largest EMD, the fraction of jet energy contained in a given cluster `frac_E` is shown as a histogram in Figure 51. Note that only the singular value pre-splitting is considered here. Sharp peaks are visible at 0 % and 100 %, corresponding to clusters that contain none or all of the jet energy, respectively. Although clusters that were not split tend to contain very little of the jet energy, there are also some clusters that nearly contain the entirety of it. The sparse interim region proves again that many jets contain only a single cluster without splitting.

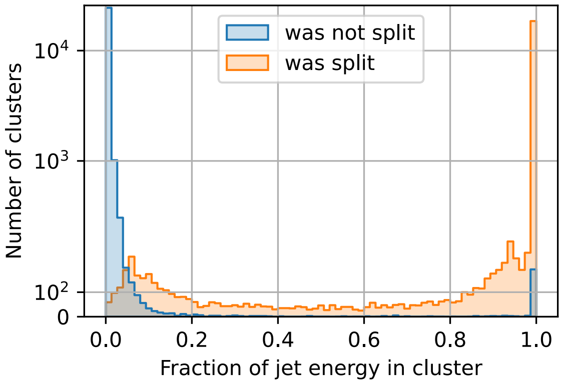


Figure 51: Clusters containing a large fraction of the overall energy tend to be split. A symlog scale is used to make interim values visible. (Z' + bkg)

Besides comparisons of clusters that were or were not split, it is also possible to compare the features of split clusters to those of their non-split counterparts. Because this is a one-to-many relationship, 2D histograms are used to visualize these comparisons. Each value on the x-axis corresponds to a cluster before splitting, while the y-axis shows the values of all matched split clusters.

[13] `SECOND_LAMBDA` is the second moment of λ, i.e. $\langle \lambda^2 \rangle$, where λ are the distances of cells from cluster center along the principal axis. See [7] for a detailed definition.

Figure 52 proves that split clusters always have less total energy than the non-split clusters, as expected. Beyond this trivial observation, the energy after splitting does not increase with the energy before splitting; instead, it is distributed over a larger number of clusters. In Figure 53, the points $(0,0)$ and $(1,1)$ are most prominent, corresponding to clusters that are either fully contained in the electromagnetic calorimeter or have no energy in it at all, both before and after splitting. Additional regions of high density are visible at the top and bottom around $x = 0.6$, meaning that split cluster are significantly more likely to be fully or not at all contained in the EM calorimeter. The diagonal corresponds to clusters that are not split.

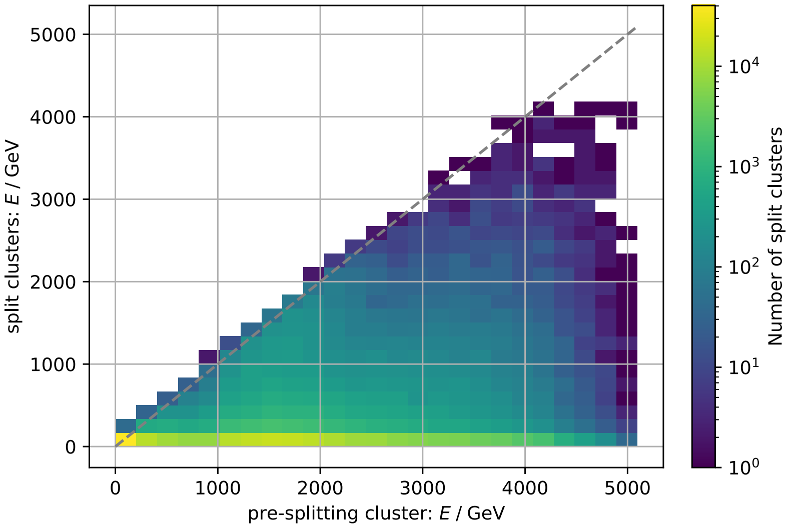


Figure 52: 2D histogram of the total energy of clusters before splitting vs. all matching split clusters. (Z' + bkg)

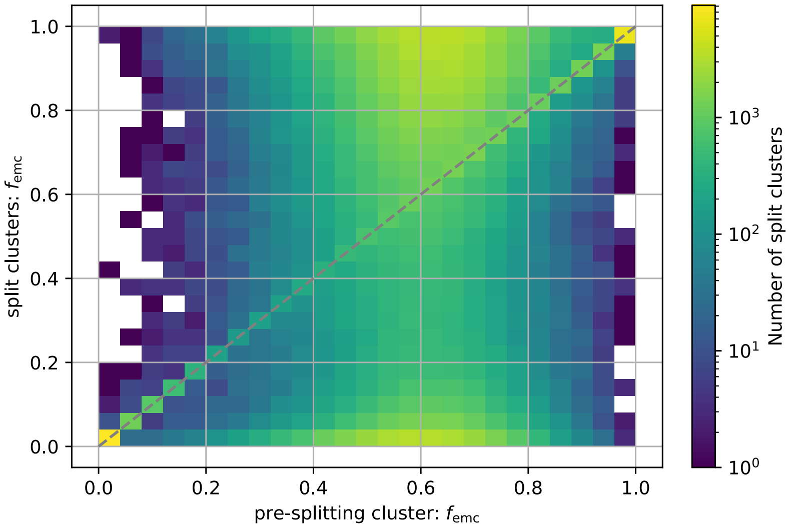


Figure 53: 2D histogram of the fraction of energy contained in the EM calorimeter before splitting vs. all matching split clusters. (Z' + bkg)

3.6 Cell-Level Studies

Although topo-cluster splitting does not directly affect cell-level variables (apart from the weight of border cells, which is not present in the given data), they in turn affect the splitting algorithm, for example due to the 500 MeV threshold that defines local maxima (see Section 2.1.3). As building blocks of clusters, fundamental properties of cells within clusters are studied in the following.

Figure 54 shows that the vast majority of cells has miniscule energy deposits relative to the cluster's maximum, resulting in a peak near zero. A second peak is visible near 100 %, corresponding to the highest-energy cells in each cluster. The left tail of the graph corresponding to negative cell energies extends beyond − 100 %. This region covers clusters that are dominated by cells with negative energy deposits. 0.81 % of all clusters satisfy this condition. Topo-clusters with negative *overall* energy are not part of the sample at hand; consequently, the fraction of clusters with negative-energy cells is lower than in reality.

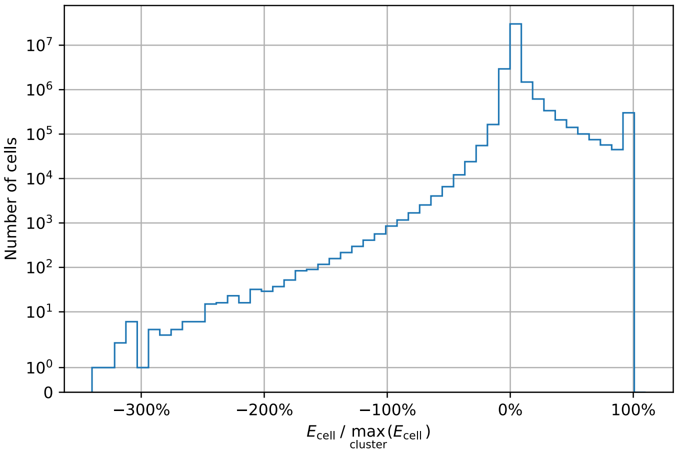


Figure 54: Distribution of cell energy deposit relative to the maximum energy deposit in the cluster. (W' + bkg)

Figure 55 and Figure 56 show the relationship between the second-highest cell energy in a cluster and the number of cells in that cluster, with splitting disabled and enabled, respectively. The second-highest cell energy is chosen because it relates to the 500 MeV threshold for splitting, indicated by a dashed line in both figures. Stripes for low numbers of cells are visible due to the logarithmic binning of the histogram. When splitting is disabled, two regions of high density are visible: One at lower cell energies in the order of hundreds of MeV, and one at higher cell energies in the order of tens of GeV. Due to the lower number of clusters without splitting, less statistics are available, as can be seen by comparing the color scales. With splitting, the overall distribution is more uniform, as maxima from more clusters are considered, while each maximum is less likely to be very high in energy. A cut becomes visible near the 500 MeV threshold, as expected. All cells to the right of this line belong to clusters that were not split, even though they contained at least two cells above the threshold. This is due to the additional conditions for splitting (see Section 2.1.3).

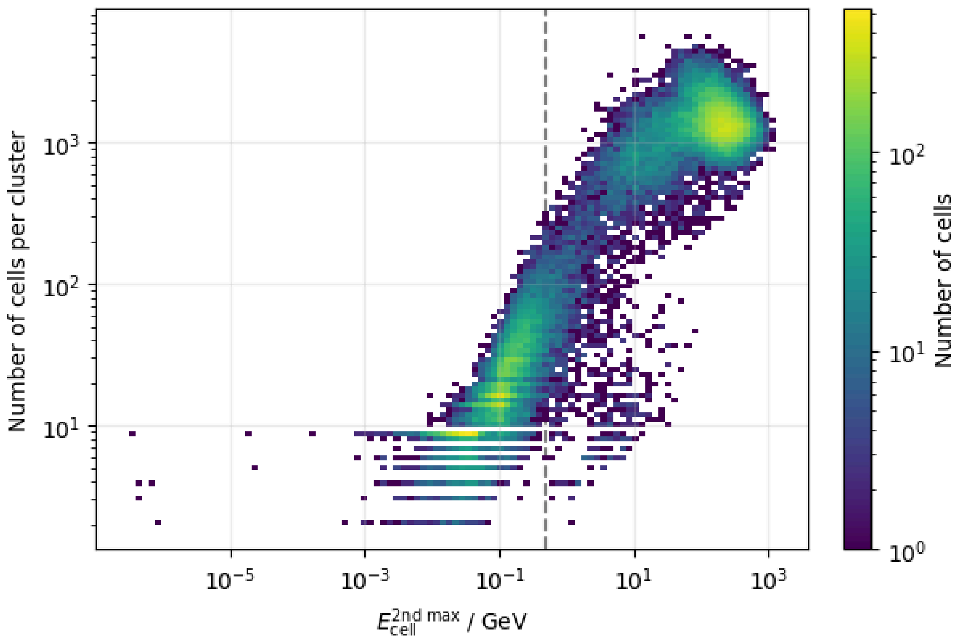


Figure 55: 2D Histogram of the second-highest cell energy in a cluster vs. the number of cells in that cluster, with splitting disabled. A dashed line indicates the 500 MeV threshold for splitting.

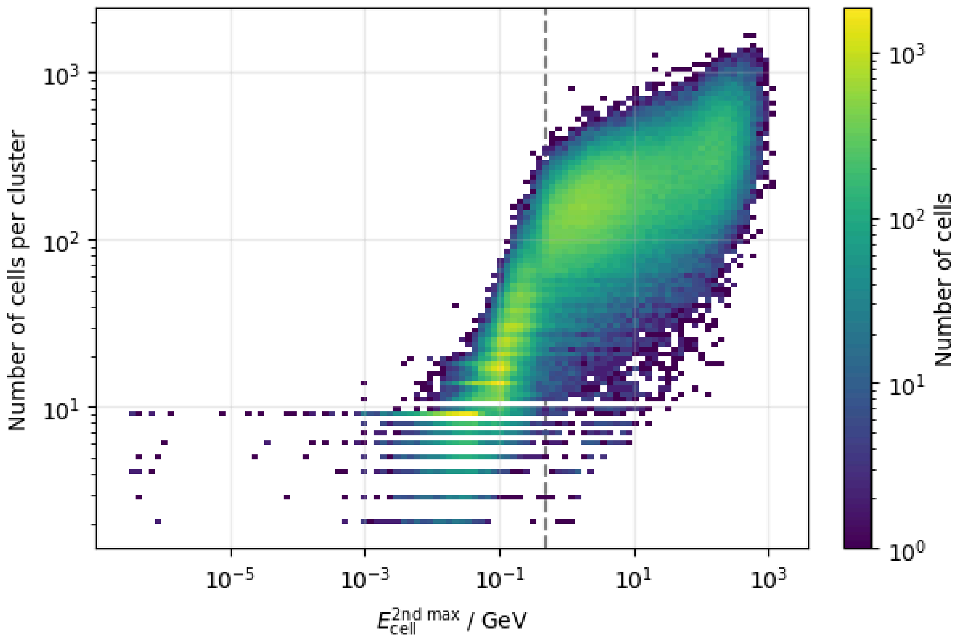


Figure 56: 2D Histogram of the second-highest cell energy in a cluster vs. the number of cells in that cluster, with splitting enabled.

3.7 Agreement of Clusters with Truth Constituents

One aspect to consider when judging the impact of topo-cluster splitting is how the distribution of simulated shower particles (*truth constituents*) relates to the distribution of cell signals, or rather, clusters. If clusters were often split in a way that some of the resulting clusters do not contain any truth constituents, this would be a sign of deficiency of the splitting algorithm.

The underlying question of whether a truth constituent belongs to a cluster is not unambiguous, though, because clusters are observed only via their member cells, each represented as a point in the present samples (see Section A).[14] Since the truth constituent data does not include any information about sampling (/ depth with respect to the beam axis), the problem is reduced to two dimensions, namely the (η, ϕ) plane. Obtaining the actual shape of individual cells based on their center coordinates is not feasible in this analysis, as it would require setting up Athena and modifying the reconstruction chain to store this information.

Most basic solutions would require a grid of cells that is equidistant, which is not the case in the ATLAS calorimeter, especially throughout sampling layers. Therefore, two more complex solutions were considered: A convex hull approach that can adapt to the shape of the cluster, and a “pitch-aware” nearest-neighbor search that takes into account the varying cell density, both of which are described in the following sections.

3.7.1 Option 1: Convex Hulls

A fully cell-layout-independent method is to use a convex hull around the cells of a cluster. Convex hulls are a way to define a boundary around a set of points in space. Convexity means that for any two points within the boundary, the straight line connecting them is also entirely within the boundary. Imposing this constraint gives a well-defined and computationally efficient way to create a boundary, compared to more general shapes that could be concave or have holes.

Generating convex hulls is computationally more expensive than preparing an optimized nearest-neighbor search and non-trivial to parallelize, but still sufficient for offline analysis, as the results for a given dataset can be cached. The implementation `scipy.spatial.ConvexHull` from [59] is used for this purpose.

One drawback of requiring convexity is that it does not always fit the shape of a cluster well. To give an example, a cluster with a half-moon shape would not be well represented by a convex hull, as the hull would include the empty space between the two outer ends of the half-moon. Similarly, outliers may dictate the shape of the hull. As a result, convex hulls would tend to overestimate the area covered by a cluster, thus overestimating the number of truth constituents contained therein. As a countermeasure, the cells used for

[14] It can be argued that this ambiguity is related to the “coastline paradox” in geography, where the length of a coastline depends on the resolution of the measurement. In the present case, the “coastline” is the boundary of a cluster, and its shape depends on the choice of cells used to define it. This is further complicated by the fact that clusters are not necessarily convex, which means that the boundary can take on complex shapes.

construction of the convex hull are restricted to those with sufficient energy deposit. Motivated by the typical energy distribution of cells within a cluster (see Figure 54), the 70 % quantile of the cell energy deposit was found to be a functional lower threshold for this purpose.

Another issue arises for small clusters with very few cells. If all cells of a cluster are collinear (i.e. lie on a straight line), the convex hull does not enclose any area and is thus not defined. Those cases are treated by setting the number of matching truth constituents to `-1`.

Figure 57 shows an example of a convex hull. Notably, there are some cells that are not part of the hull; they are barely visible due to their low energies.

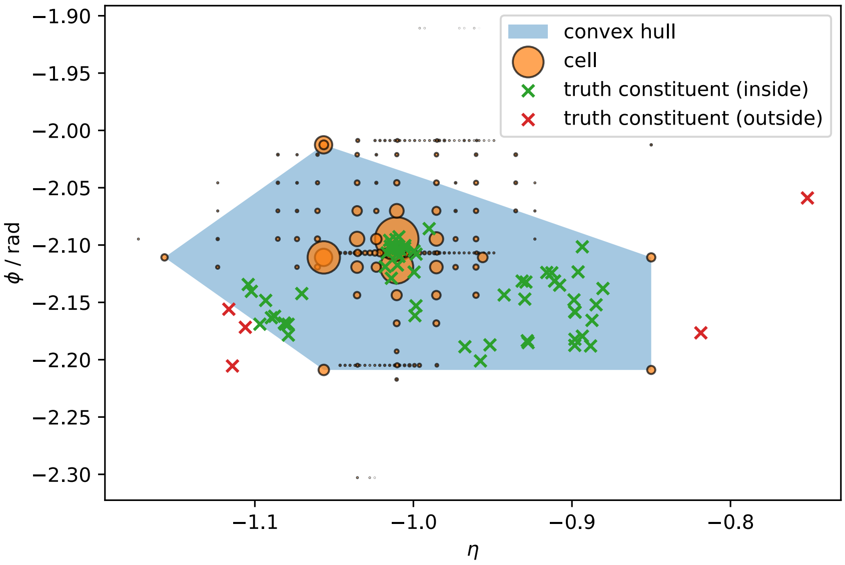


Figure 57: Example of a convex hull. The × markers denoting truth constituents are colored based on whether they are part of the hull or not. The cells of a single cluster are shown as circles, with their size proportional to their energy deposit. The convex hull is represented by a blue surface.

3.7.2 Option 2: "Pitch-Aware" Matching

As discussed before, a basic nearest-neighbor search is not feasible due to the inhomogeneous cell layout (shown in Figure 58). This becomes evident when considering a subsection of calorimeter cells in a (ϕ, η) region, shown in Figure 58: Without knowledge of the sampling layer, it is not possible to determine the typical cell spacing (pitch), as it varies significantly between layers. Even within a sampling layer, pitches are not entirely uniform. They are not only different for η and ϕ (especially in the strip layers), but also vary within layer and direction, as exemplified in Figure 59 for sampling layer 5. Therefore, distance thresholds per sampling layer and coordinate direction are obtained from the 90-percentile of all corresponding individual cell pitches. The complete table of obtained pitches is given in Table A9. They will be referred to as thresholds $t_{\eta,s}$ and $t_{\phi,s}$ in the following.

A truth constituent is considered part of a cluster if it is within $t_{\eta,s}$ in η and $t_{\phi,s}$ in ϕ of any cell of the cluster, where $t_{\eta,s}$ and $t_{\phi,s}$ correspond to the sampling layer s of that cell:[15]

$$|\Delta\eta| < t_{\eta,s} \wedge |\Delta\phi| < t_{\phi,s}. \tag{21}$$

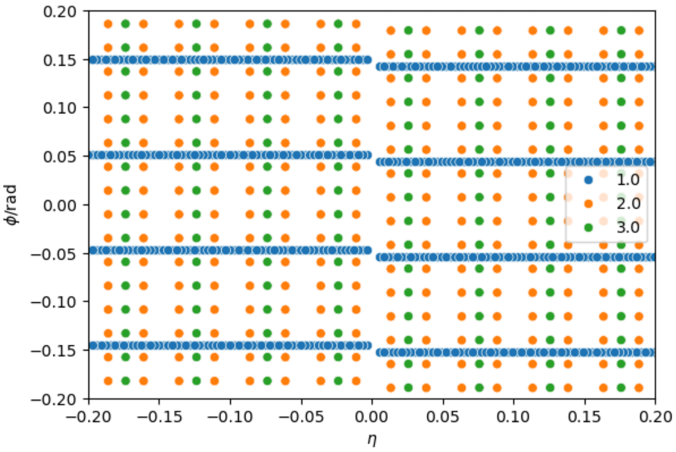


Figure 58: Layout of calorimeter cells in a (ϕ, η) subsection, colored by sampling layer. Differences in pitch between layers are clearly visible. For example, ■ EMB1 (index 1) has a strip geometry and is dense in η, and a "seam" at $\eta = 0$ (see Figure 4).

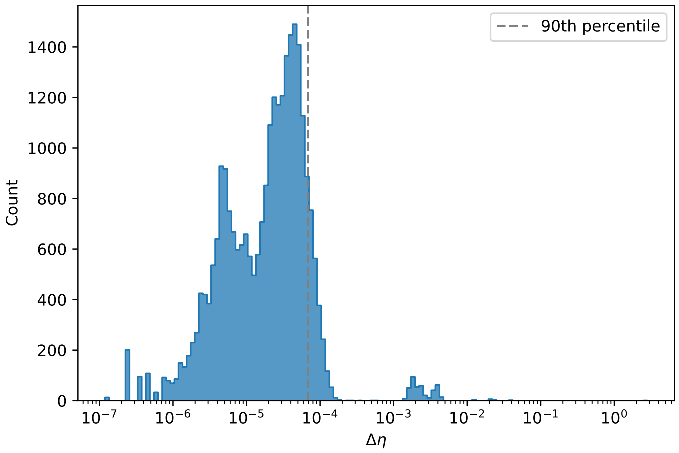


Figure 59: Histogram of pitches $\Delta\eta$ of cells in sampling layer 5.

[15] After scaling by the thresholds, this is the Chebyshev distance / maximum metric.

3.7.3 Results

Figure 60 and Figure 61 show the distributions of the number of truth constituents contained in clusters, as determined by the convex hull and pitch-aware methods, respectively. While the distributions are similar towards higher numbers of truth constituents, the behavior up to about 10 truth constituents is different: The convex hull method shows a significant number of clusters with `-1` truth constituents, emphasizing the drawback of undefined convex hulls for small clusters with very few cells. Both methods show a significant number of clusters with zero truth constituents, presumably due to noise from pile-up. The effect might be amplified by the fact that charged truth constituents' tracks are bent in the magnetic field, which cannot be accounted for with 2D information alone.

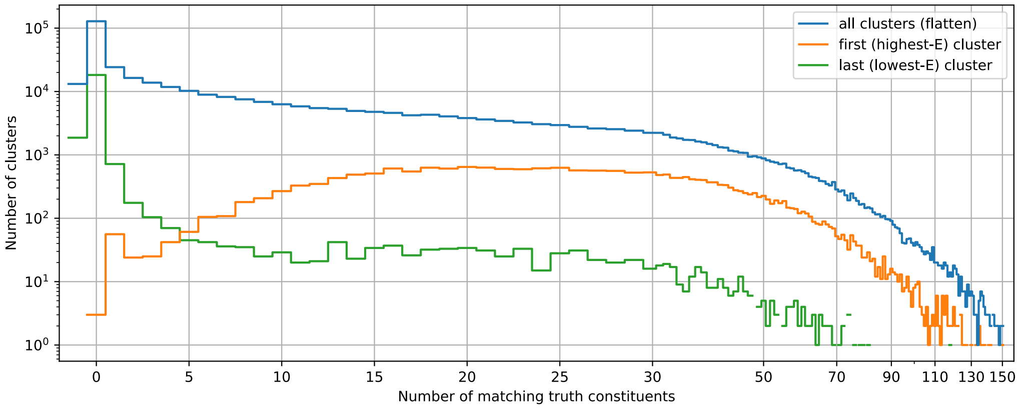


Figure 60: Histogram of the number of truth constituents contained in the convex hull of a cluster. Note: The x axis is symlog-scaled to allow for `-1` to be shown. (Z' + bkg)

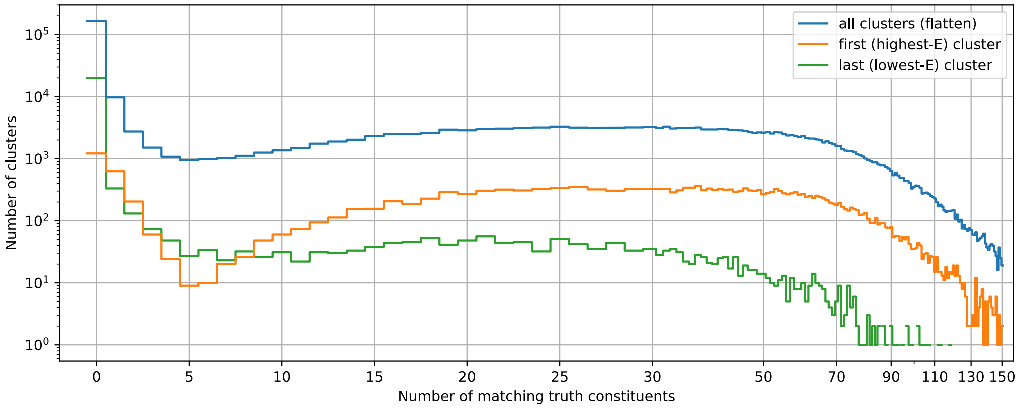


Figure 61: Histogram of the number of truth constituents matched via pitch-aware nearest-neighbor search. (Z' + bkg)

Figure 62 shows the correlation between the fraction of truth constituents that do not match any cluster and the reconstruction error of τ_{21} as an exemplary cluster variable. A weak correlation is observed, indicating that – counterintuitively – jets with many unmatched truth constituents tend to have a lower reconstruction error of τ_{21}. This might be due to untreated correlations, as τ_{21} is calculated from the leading two clusters in a jet, which might still be well reconstructed even if many truth constituents are unmatched.

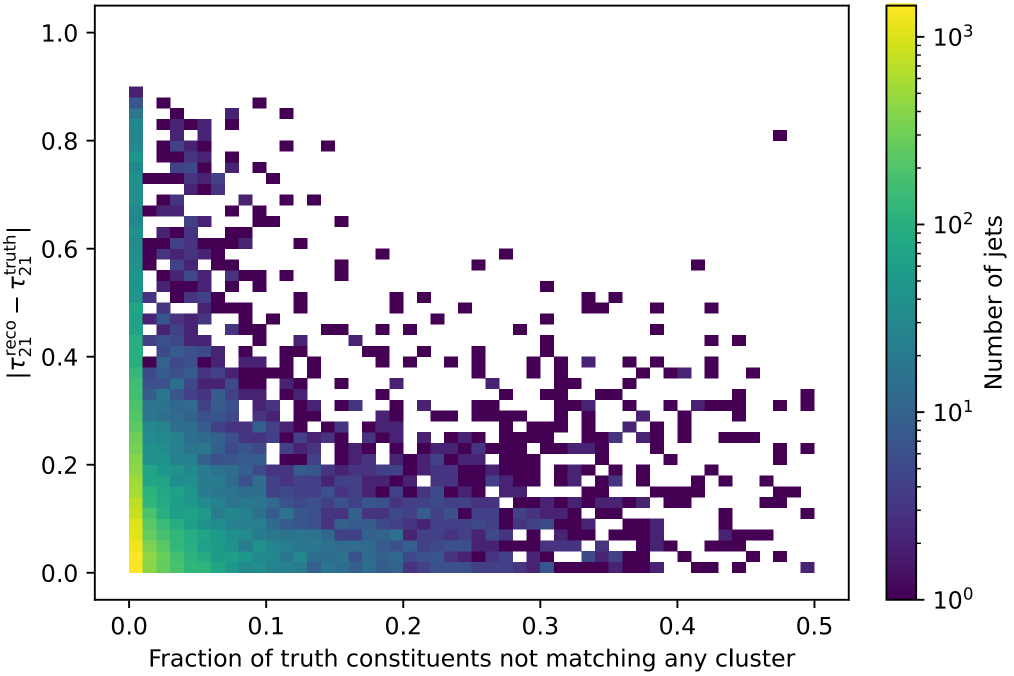


Figure 62: τ_{21} reconstruction error vs. fraction of truth constituents not matched to any cluster. (convex hull method) (Z' + bkg)

Table 6 gives an overview of cluster-level variables that differ the most between clusters with few ($N_{\text{matching}} \leq 5$) and many ($N_{\text{matching}} \geq 20$) truth constituents contained therein, as determined by the convex hull method. The thresholds were chosen to yield populations of similar size. As in previous comparisons, the most prominent features are those that are intuitively correlated with the size of the cluster, such as the fraction of jet energy contained therein, or the maximum cell significance within the cluster. Additionally, a ~40-fold increase in mass and a ~20-fold increase in cluster-level significance are observed for clusters with many truth constituents. Results for the alternative pitch-aware method are shown in Table 7. While overall trends are similar, some additional variables appear among the leading in terms of EMD differences. For example, clusters matching many truth constituents tend to be less isolated (lower `ISOLATION`), and the effective weights accounting for out-of-cluster deposits are lower.

Figure 63 shows the distributions underlying the EMD for the (truth-level) energy deposited in cells outside of the cluster but associated with it, either via the convex hull or pitch-aware method. This variable is of special interest as it is related to the goal of quantifying how well clusters capture the energy deposits of particles. Though the distributions for either method are similar in shape, the pitch-aware method appears to match more truth constituents on average. Both methods show that clusters with many truth constituents tend to have more associated out-of-cluster energy deposits, possibly because of being larger in the first place. This is in line with the observation that clusters with many truth constituents tend to contain a larger fraction of the jet energy, as mentioned above.

Table 6: Normalized EMDs between $N_{\text{matching}} \leq 5$ and $N_{\text{matching}} \geq 20$. (convex hull method)

variable	**unit**	**EMD**	**mean**	**mean**
			($N_{\text{matching}} \leq 5$)	($N_{\text{matching}} \geq 20$)
ENG_CALIB_OUT_T	GeV	2.690	0.604	1.754
ENG_CALIB_OUT_M	GeV	1.438	0.544	3.350
ENG_CALIB_OUT_L	GeV	1.369	0.571	4.568
fracE		1.272	0.006	0.194
fracE_ref		1.247	0.005	0.158
CELL_SIGNIFICANCE		1.226	18.454	515.279
MASS	GeV	1.169	404.816	17 571.809
SIGNIFICANCE		1.166	11.096	235.392
ENG_CALIB_TOT	GeV	1.153	8.491	403.457
ENG_POS	GeV	1.142	8822.642	387 454.358

Table 7: Normalized EMDs between $N_{\text{matching}} \leq 5$ and $N_{\text{matching}} \geq 20$. (pitch-aware method)

variable	**unit**	**EMD**	**mean**	**mean**
			($N_{\text{matching}} \leq 5$)	($N_{\text{matching}} \geq 20$)
ENG_CALIB_OUT_T	GeV	2.360	0.479	1.449
ENG_CALIB_OUT_L	GeV	1.092	0.593	2.926
OOC_WEIGHT		1.079	1.347	1.041
SECOND_LAMBDA	mm^2	1.070	17 920.171	175 924.693
ENG_CALIB_OUT_M	GeV	1.054	0.505	2.275
CELL_SIG_SAMPLING		1.041	1.579	6.936
CENTER_LAMBDA	mm	0.972	211.208	802.358
ISOLATION		0.934	0.552	0.304
LATERAL		0.865	0.617	0.862
AVG_TILE_Q		0.851	0.319	14.113

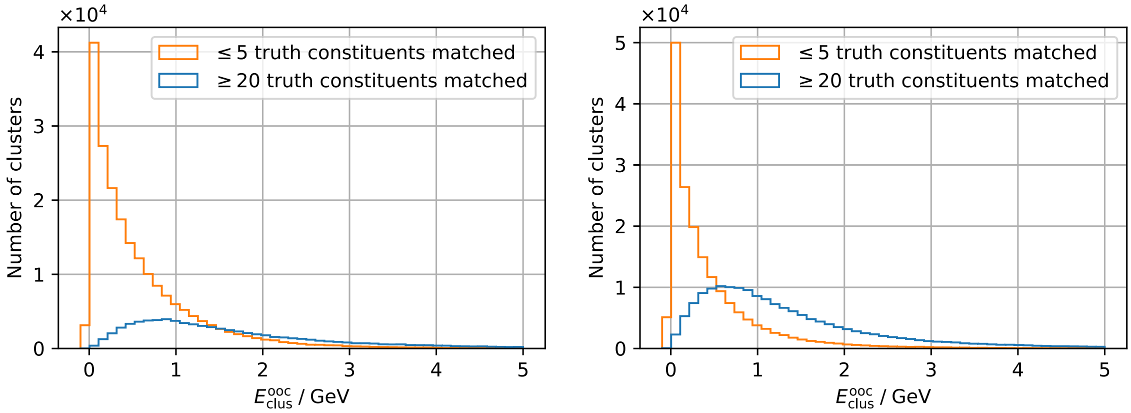


Figure 63: Histograms of truth-level energy deposited in cells outside of the cluster but associated with it for clusters with few ($N_{\text{matching}} \leq 5$) and many ($N_{\text{matching}} \geq 20$) truth constituents contained therein, as determined by the convex hull method (left) and pitch-aware method (right). (Z' + bkg)

3.8 Variations of Splitting Hyperparameters

As discussed in Section 2.1.3, the topo-cluster splitting algorithm depends on a set of hyperparameters, most importantly a minimum number of neighbor cells that are part of the same cluster (default value: 4) as well as an energy threshold for *local maxima* (default value: 500 MeV). The paper introducing the algorithm [7] does not provide any information on how these hyperparameters were chosen or optimized, thereby leaving their performance impact unspecified and motivating an empirical study.

A basic grid search is performed with 8 additional datasets (listed in Section B) to gauge the impact of these hyperparameters on the performance of the splitting algorithm. For each dataset, the performance of substructure variables is evaluated in terms of the area under the ROC curve (AUC). The datasets are of sufficient size to ensure that statistical fluctuations are small compared to the observed differences in AUC and too small to be visible in the ROC curves.

Figure 64 superimposes the ROC curves for all considered substructure variables and different splitting hyperparameters to give an overview of the results. An example of all ROC curves for a single substructure variable (τ_{21}^{reco}) is shown in Figure 65. Differences in the TPR of up to 4 % are observed relative to the default hyperparameters, at medium FPR values.

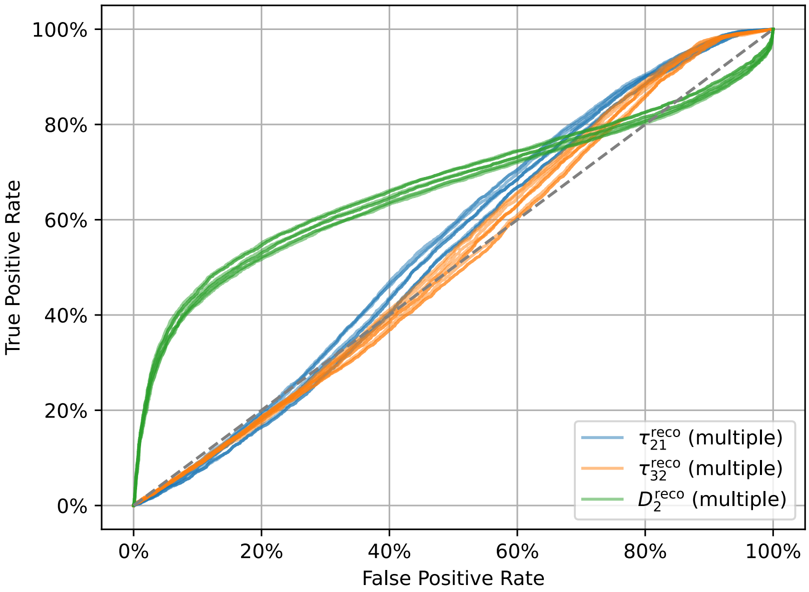


Figure 64: Superimposed ROC curves for all considered substructure variables (color-coded) and different splitting hyperparameters. (Z' + bkg)

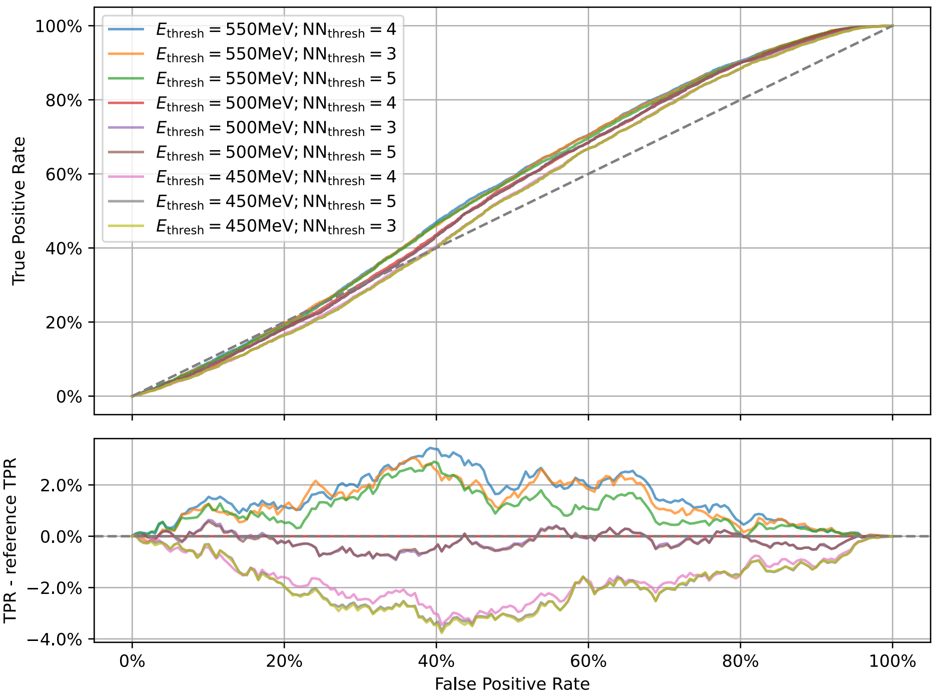


Figure 65: ROC curves for $\tau_{21}^{\mathrm{reco}}$ for different splitting hyperparameters. Residuals are shown relative to the default hyperparameters; positive residuals indicate better performance. (Z' + bkg)

The numerical results for Z'/background discrimination are listed in Table 8. The table is sorted by the mean improvement across all considered substructure variables, reaching improvements of up to 0.5 % in AUC relative to the default hyperparameters in the case of an increased energy threshold (550 MeV). However, no single set of hyperparameters consistently outperforms the others across all substructure variables. Though jet tagging algorithms can combine the information from different substructure variables, the lack of a clear overall winner suggests that either the default hyperparameters are already close to optimal, or more extreme changes to the hyperparameters are required to achieve meaningful improvements. For W', the hierarchy of hyperparameter performance is nearly unchanged, but overall improvements are even smaller. The corresponding results are listed in Table A17 in the appendix.

Table 8: Overview of the performance of different splitting hyperparameters in terms of the AUC of different substructure variables as well as their improvement relative to the default hyperparameters. (Z' + bkg)

E_{thresh}/ MeV	NN$_{\mathrm{thresh}}$	AUC($\tau_{21}^{\mathrm{reco}}$)	AUC($\tau_{32}^{\mathrm{reco}}$)	AUC(D_2^{reco})	ΔAUC($\tau_{21}^{\mathrm{reco}}$)	ΔAUC($\tau_{32}^{\mathrm{reco}}$)	ΔAUC(D_2^{reco})	⟨ΔAUC⟩
550	4	0.557	0.531	0.656	+0.014	+0.012	−0.010	0.005
550	5	0.552	0.528	0.657	+0.009	+0.009	−0.009	0.003
550	3	0.555	0.523	0.654	+0.012	+0.004	−0.012	0.001
500	4	0.542	0.519	0.667	+0.000	+0.000	+0.000	0.000
500	5	0.541	0.517	0.666	−0.001	−0.001	−0.000	−0.001
500	3	0.541	0.517	0.666	−0.001	−0.001	−0.001	−0.001
450	4	0.526	0.507	0.680	−0.016	−0.011	+0.013	−0.004
450	5	0.524	0.504	0.678	−0.018	−0.014	+0.011	−0.007
450	3	0.524	0.504	0.678	−0.018	−0.014	+0.011	−0.007

4 Conclusions

This thesis has presented a comprehensive investigation of topo-cluster splitting in the ATLAS calorimeter and its impact on boosted object identification. Through systematic analysis of Monte-Carlo simulations featuring top and W/Z jets alongside QCD dijet backgrounds, the effects of the splitting algorithm have been examined across multiple levels of reconstruction detail.

The study has demonstrated that topo-cluster splitting plays a fundamental role in jet substructure reconstruction. Without splitting, the discriminating power of substructure variables such as τ_{21}, τ_{32}, and D_2 is severely degraded, confirming that the splitting procedure is essential for effective boosted object identification. The analysis revealed significant differences in cluster-level properties between split and non-split configurations (Section 3.4), with split clusters generally exhibiting more appropriate size distributions and improved energy resolution for substructure analyses.

Studies at cell-level (Section 3.6) provided insights into the distribution of energy deposits within clusters, and how these distributions are altered by the splitting process. In particular, it was demonstrated how the distribution of peak cell energies shifts after splitting.

Novel methodologies have been developed for this study, including a systematic approach for comparing split and non-split cluster configurations (Section 3.5), and two matching algorithms for associating Monte-Carlo truth constituents with reconstructed clusters (Section 3.7). Though subject to inherent limitations, such as the imperfect association of clusters before and after splitting due to grooming or the non-trivial shapes that clusters can take on, correlations have been identified that could inform future algorithm development.

The binary splitting analysis (Section 3.3.5) established an upper performance limit for a subclass of potential future algorithms that would make binary split/no-split decisions on a per-cluster basis. The results indicate that only a small fraction of jets (approximately 8% for τ_{21} and 3% for τ_{32}) would benefit from disabling splitting entirely, suggesting that the current default approach of enabling splitting is generally appropriate.

A grid search optimization of two splitting algorithm hyperparameters (Section 3.8) revealed that while measurable differences exist between parameter choices, the magnitude of performance improvements achievable through hyperparameter tuning alone is limited. The energy threshold for local maxima identification and the minimum number of neighbor cells both show modest effects on substructure variable performance, with optimal values varying slightly depending on the specific observable considered.

The findings of this thesis provide a starting point for future research and algorithm development. The limited performance gains achievable through optimization of energy and neighbor thresholds alone (Section 3.8) suggest that more fundamental algorithmic improvements may be necessary to substantially enhance jet substructure reconstruction. Other hyperparameters of the splitting algorithm, such as the selection of sampling layers that provide local maxima, are also worth investigating, as they have been chosen before the advent of boosted object analyses.

Advanced splitting algorithms that move beyond the binary split/no-split decisions examined in Section 3.3.5 represent a natural next step. Rather than applying uniform splitting criteria, future algorithms could implement adaptive strategies that tailor splitting behavior to the specific characteristics of individual clusters or the physics context of particular events. Machine learning approaches could potentially identify optimal splitting strategies by learning from the relationship between cluster properties and final reconstruction performance.

The choice of optimization targets also warrants further consideration. From a physics point of view, splitting should resolve subjets, meaning that e.g. gluon-initiated 1-prong jets should not be split excessively. The specific number of clusters that should be formed per jet, however, is not well-defined. Different physics scenarios and algorithms may require different levels of granularity for optimal performance. A middle ground must be found between “under-splitting”, which could obscure important substructure details, and “over-splitting”, increasing noise and complicating reconstruction.

One approach explored in this thesis is to optimize splitting algorithms based on their alignment with Monte-Carlo truth constituents (Section 3.7). This strategy provides a benchmark that is not directly tied to specific physics analyses while still reflecting important aspects of jet substructure. However, it is reliant on a clear definition of matching between truth constituents and reconstructed clusters; a matching radius must be chosen, and the 3-dimensional shape of every cell should be considered. Tight integration with the detector simulation could introduce depth information to truth constituents, potentially improving the fidelity of the association. Nonetheless, reliance on Monte-Carlo truth information introduces potential biases, as different event generators may produce varying constituent-level structures.

In order to find cluster-level proxies for splitting performance, the matching of split to non-split clusters (Section 3.5) could be further refined. While the current approach provides a useful starting point, tighter integration with the reconstruction chain is needed to mitigate ambiguities due to jet grooming or similar effects.

While topo-clustering remains a crucial step in ATLAS calorimeter reconstruction, the advent of successor algorithms for LCTopo jets presents new opportunities for improvement. As an example, TCCs (Track-CaloClusters) [40] already add to the default splitting behavior by incorporating tracking information in an additional step. This does not preclude further enhancements to the underlying topo-clustering and splitting algorithms, as TCCs still build on top of the existing clusters. Neutral particles in particular could benefit from improved calorimeter-level reconstruction, given that TCCs primarily enhance charged particle reconstruction through track association.

The systematic framework developed in this thesis provides a foundation for future splitting algorithm research. The multi-level analysis approach, novel matching methodologies, and performance evaluation techniques established here can be applied to the assessment of more sophisticated algorithmic approaches and guide the development of next-generation calorimeter reconstruction software.

Ultimately, while current splitting algorithms provide essential functionality for jet substructure analyses, substantial room for improvement remains. The combination of more intelligent algorithmic approaches, expanded application domains, and refined optimization techniques offers the potential for major advances in boosted object identification capabilities, with corresponding benefits for both Standard Model precision measurements and beyond-Standard-Model searches at ATLAS and future experiments.

Appendix

A Features

A.1 Restructured

Below is the *Awkward Array* data structure for a cell-level sample. It does not directly correspond to the structure in the ROOT files; instead, it has been restructured to make use of nested arrays and records wherever possible, reducing redundancy and improving usability.

```
{
  meta: {
    eventNumber: int64,
    runNumber: int32,
    seqNumber: int64,
    _is_signal: bool
  },
  jet / jetNS: {
    Cal / Raw / Area: {
      E: float32,
      Pt: float32,
      Eta: float32,
      Rap: float32,
      Phi: float32,
      M: float32
    },
    Tau1: float32,
    Tau2: float32,
    Tau3: float32,
    ECF1: float32,
    ECF2: float32,
    ECF3: float32,
    NConst: int32
  },
  truthJet: {
    E: float32,
    Pt: float32,
    Eta: float32,
    Rap: float32,
    Phi: float32,
    M: float32,
    Tau1: float32,
    Tau2: float32,
    Tau3: float32
  },
  truthConstit: var * {
    E: float32,
    Pt: float32,
    Eta: float32,
    Phi: float32
  },
  clusters / clustersNS: var * {
    cells: var * cell[
      eta: float32,
      phi: float32,
      sampling: float32,
      E: float32,
      significance: float32,
      time: float32
    ],
    E: float32,
    Pt: float32,
    Eta: float32,
    Phi: float32,
    sumCellE: float32,
    time: float32,
    fracE: float32,
    fracE_ref: float32,
    ePerSampling: 28 * float32,
    EM_PROBABILITY: float32,
    HAD_WEIGHT: float32,
    OOC_WEIGHT: float32,
    DM_WEIGHT: float32,
    ENG_CALIB_TOT: float32,
    ENG_CALIB_OUT_T: float32,
    ENG_CALIB_OUT_L: float32,
    ENG_CALIB_OUT_M: float32,
    ENG_CALIB_DEAD_TOT: float32,
    ENG_CALIB_FRAC_EM: float32,
    ENG_CALIB_FRAC_HAD: float32,
    ENG_CALIB_FRAC_REST: float32,
    CENTER_MAG: float32,
    FIRST_ENG_DENS: float32,
    FIRST_PHI: float32,
    FIRST_ETA: float32,
    SECOND_R: float32,
    SECOND_LAMBDA: float32,
    DELTA_PHI: float32,
    DELTA_THETA: float32,
    DELTA_ALPHA: float32,
    CENTER_X: float32,
    CENTER_Y: float32,
    CENTER_Z: float32,
    CENTER_LAMBDA: float32,
    LATERAL: float32,
    LONGITUDINAL: float32,
    ENG_FRAC_EM: float32,
    ENG_FRAC_MAX: float32,
    ENG_FRAC_CORE: float32,
    SECOND_ENG_DENS: float32,
    ISOLATION: float32,
    ENG_BAD_CELLS: float32,
    N_BAD_CELLS: float32,
    N_BAD_CELLS_CORR: float32,
    BAD_CELLS_CORR_E: float32,
    BADLARQ_FRAC: float32,
    ENG_POS: float32,
    SIGNIFICANCE: float32,
    CELL_SIGNIFICANCE: float32,
    CELL_SIG_SAMPLING: float32,
    AVG_LAR_Q: float32,
    AVG_TILE_Q: float32,
    ENG_BAD_HV_CELLS: float32,
    N_BAD_HV_CELLS: float32,
    PTD: float32,
    MASS: float32,
    SECOND_TIME: float32
  }
}
```

A.2 Cluster-Level Variables

name	formula	description
`E`	$E^{\text{EM}}_{\text{clus}}$	EM-scale cluster energy.
`Pt`	$p^{\text{EM}}_{T,\text{clus}}$	EM-scale cluster transverse momentum.
`Eta`	η_{clus}	Cluster pseudorapidity (EM scale).
`Phi`	ϕ_{clus}	Cluster azimuth (EM scale).
`sumCellE`	$\sum_{i:\,E^{\text{EM}}_{\text{cell},i}>0} w^{\text{geo}}_{\text{cell},i}\, E^{\text{EM}}_{\text{cell},i}$	Sum of positive EM-scale cell energies in the cluster (with geometric weights).
`fracE`	$f^{\text{EM}}_{\text{E, clus}} = \frac{E^{\text{EM}}_{\text{clus}}}{\sum_{i=1}^{N^{\text{particle}}_{\text{clus}}} E^{\text{EM}}_{\text{clus},i}}$	Cluster's fraction of the parent particle/jet EM energy carried by all its topo-clusters.
`fracE_ref`	$F^{\text{EM}}_{\text{E, clus}} = \frac{E^{\text{EM}}_{\text{clus}}}{E^{\text{truth}}_{\text{particle (jet)}}}$	Cluster EM energy over truth particle/jet energy.
`EM_PROBABILITY`	$P^{\text{EM}}_{\text{clus}} \in [0,1]$	Probability that the cluster is EM-like (used for LCW interpolation).
`HAD_WEIGHT`	$P^{\text{EM}}_{\text{clus}}\, w^{\text{em-cal}}_{\text{clus}} + (1 - P^{\text{EM}}_{\text{clus}})\, w^{\text{had-cal}}_{\text{clus}}$	Effective cluster-level weight combining EM and HAD calibrations.
`ECalib`	$E^{\text{LCW}}_{\text{clus}}$	LCW-calibrated cluster energy.
`PtCalib`	$p^{\text{LCW}}_{T,\text{clus}}$	LCW-calibrated cluster transverse momentum.
`EtaCalib`	$\eta^{\text{LCW}}_{\text{clus}}$	Cluster pseudorapidity after LCW.
`PhiCalib`	$\phi^{\text{LCW}}_{\text{clus}}$	Cluster azimuth after LCW.
`sumCellECalib`	$\sum_{i:\,E^{\text{EM}}_{\text{cell},i}>0} w^{\text{LCW}}_{\text{cell},i}\, E^{\text{EM}}_{\text{cell},i}$	Sum of calibrated (LCW) cell energies in the cluster.
`fracECalib`	$f^{\text{LCW}}_{\text{E, clus}} = \frac{E^{\text{LCW}}_{\text{clus}}}{\sum_{i=1}^{N^{\text{particle}}_{\text{clus}}} E^{\text{LCW}}_{\text{clus},i}}$	LCW energy fraction within the parent particle/jet.
`fracECalib_ref`	$F^{\text{LCW}}_{\text{E, clus}} = \frac{E^{\text{LCW}}_{\text{clus}}}{E^{\text{truth}}_{\text{particle (jet)}}}$	LCW energy over truth particle/jet energy.

`CENTER_X`	X_{clus}	Cluster center of gravity, x (detector frame).
`CENTER_Y`	Y_{clus}	Cluster center of gravity, y (detector frame).
`CENTER_Z`	Z_{clus}	Cluster center of gravity, z (detector frame).
`CENTER_MAG`	r_{clus}	Distance of cluster center from the nominal vertex.
`CENTER_LAMBDA`	λ_{clus}	Depth from calorimeter front face along the shower axis.
`FIRST_PHI`	$\langle\phi\rangle$	Energy-weighted first moment of the cells' ϕ distribution.
`FIRST_ETA`	$\langle\eta\rangle$	Energy-weighted first moment of the cells' η distribution.
`DELTA_PHI`	$\Delta\phi$	Azimuth of the principal shower axis relative to the vertex direction.
`DELTA_THETA`	$\Delta\theta$	Polar-angle offset of the principal shower axis vs. vertex direction.
`DELTA_ALPHA`	$\Delta\alpha$	Angle between center-of-gravity direction and shower axis.
`SECOND_R`	$\langle r^2\rangle$	Second moment of radial distances to the shower axis.
`SECOND_LAMBDA`	$\langle\lambda^2\rangle$	Second (longitudinal) moment along the shower axis.
`LATERAL`	$\langle m^2_{\text{lat}}\rangle$	Lateral energy dispersion moment.
`LONGITUDINAL`	$\langle m^2_{\text{long}}\rangle$	Longitudinal energy dispersion moment.
`ISOLATION`	f_{iso}	Topological isolation of the cluster (0 → surrounded, 1 → isolated).
`FIRST_ENG_DENS`	$\langle\rho_{\text{cell}}\rangle$	First moment of cell energy density in the cluster.

`SECOND_ENG_DENS`	$\langle\rho^2_{\text{cell}}\rangle$	Second moment of cell energy density.
`ENG_FRAC_EM`	f_{emc}	Fraction of cluster energy in the EM calorimeter.
`ENG_FRAC_MAX`	f_{max}	Most energetic cell's energy fraction.
`ENG_FRAC_CORE`	f_{core}	Energy fraction in the cluster core.
`ENG_POS`	$E^{\text{EM}}_{\text{clus,pos}}$	Sum of positive EM cell energies (alias of the positive-sum definition).
`SIGNIFICANCE`	$\zeta^{\text{EM}}_{\text{clus}}$	Cluster signal significance (energy/noise).
`CELL_SIGNIFICANCE`	$\max\{\zeta^{\text{EM}}_{\text{cell}}\}$	Maximum cell significance in the cluster.
`CELL_SIG_SAMPLING`	$\text{layer}(\arg\max_i \zeta^{\text{EM}}_{\text{cell},i})$	Sampling layer of the max-significance cell.
`PTD`	$p_T^D = \dfrac{\sum_i (E^{\text{EM}}_{\text{cell},i})^2}{\sum_i E^{\text{EM}}_{\text{cell},i}}$	Energy-concentration (jet-analogue) using cell energies.
`MASS`	$m^{\text{EM}}_{\text{clus}}$	Cluster invariant mass from cells with $E^{\text{EM}}_{\text{cell}} > 0$ (diagnostic).
`SECOND_TIME`	$\sigma^2_{t,\text{clus}}$	Second moment (spread) of the cluster's cell-time distribution.

B Monte-Carlo Samples

The following Monte-Carlo samples were used in this thesis. Each bullet point corresponds to one sample and one directory, containing 100 ROOT files each. Different samples may contain different sets of features. As an example, `jetLevel_noPU` does not contain any per-cluster or per-cell features.

jetLevel_noPU

- user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_noPU_LargeR_March12_v0_mltree_cluster_calo.root
- user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_noPU_LargeR_March12_v0_mltree_cluster_calo.root
- user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_noPU_LargeR_March12_v0_mltree_cluster_calo.root
- user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_noPU_LargeR_March12_v0_mltree_cluster_calo.root

jetLevel_withPU

- user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_LargeR_March12_v0_mltree_cluster_calo.root
- user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_LargeR_March12_v0_mltree_cluster_calo.root
- user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_LargeR_March12_v0_mltree_cluster_calo.root
- user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_LargeR_March12_v0_mltree_cluster_calo.root

clusterLevel_withPU

- JetLevel/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_LargeR_April17_v0_mltree_cluster_calo.root
- JetLevel/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_LargeR_April17_v0_mltree_cluster_calo.root
- JetLevel/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_LargeR_April17_v0_mltree_cluster_calo.root
- JetLevel/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_LargeR_April17_v0_mltree_cluster_calo.root

cellLevel_withPU

- JetLevel_Cells/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_LargeR_April25_v0_mltree_cluster_calo.root
- JetLevel_Cells/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_LargeR_April25_v0_mltree_cluster_calo.root
- JetLevel_Cells/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_LargeR_April25_v0_mltree_cluster_calo.root
- JetLevel_Cells/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_LargeR_April25_v0_mltree_cluster_calo.root

cellLevel_withPU_noSplitting

- NoSplitting_Improved.part.lnk/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_NoSplitting_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_NoSplitting_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_NoSplitting_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_NoSplitting_July25_v0_mltree_cluster_calo.root

gridsearch_450MeV_3Cells

- NoSplitting_Improved.part.lnk/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_450MeV_3Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_450MeV_3Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_450MeV_3Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_450MeV_3Cells_July25_v0_mltree_cluster_calo.root

gridsearch_450MeV_4Cells

- NoSplitting_Improved.part.lnk/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_450MeV_4Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_450MeV_4Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_450MeV_4Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_450MeV_4Cells_July25_v0_mltree_cluster_calo.root

gridsearch_450MeV_5Cells

- NoSplitting_Improved.part.lnk/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_450MeV_5Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_450MeV_5Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_450MeV_5Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_450MeV_5Cells_July25_v0_mltree_cluster_calo.root

gridsearch_500MeV_3Cells

- NoSplitting_Improved.part.lnk/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_500MeV_3Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_500MeV_3Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_500MeV_3Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_500MeV_3Cells_July25_v0_mltree_cluster_calo.root

gridsearch_500MeV_4Cells

- NoSplitting_Improved.part.lnk/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_NoSplitting_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_NoSplitting_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_NoSplitting_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_NoSplitting_July25_v0_mltree_cluster_calo.root

gridsearch_500MeV_5Cells

- NoSplitting_Improved.part.lnk/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_500MeV_5Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_500MeV_5Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_500MeV_5Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_500MeV_5Cells_July25_v0_mltree_cluster_calo.root

gridsearch_550MeV_3Cells

- NoSplitting_Improved.part.lnk/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_550MeV_3Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_550MeV_3Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_550MeV_3Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_550MeV_3Cells_July25_v0_mltree_cluster_calo.root

gridsearch_550MeV_4Cells

- NoSplitting_Improved.part.lnk/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_550MeV_4Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_550MeV_4Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_550MeV_4Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_550MeV_4Cells_July25_v0_mltree_cluster_calo.root

gridsearch_550MeV_5Cells

- NoSplitting_Improved.part.lnk/user.cdelitzs.364706.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6WithSW.mc20e_withPU_550MeV_5Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.364707.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7WithSW.mc20e_withPU_550MeV_5Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801859.Py8EG_A14NNPDF23LO_WprimeWZ_flatpT.mc20e_withPU_550MeV_5Cells_July25_v0_mltree_cluster_calo.root
- NoSplitting_Improved.part.lnk/user.cdelitzs.801661.Py8EG_A14NNPDF23LO_Zprime_tt_flatpT.mc20e_withPU_550MeV_5Cells_July25_v0_mltree_cluster_calo.root

C Sampling Layers of the ATLAS Calorimeter

Short name	$\|\eta\|$ region	EM/HAD	Remarks
Electromagnetic barrel (LAr/lead)			
PreSamplerB	< 1.52	EM	Thin LAr presampler in front of EMB to correct upstream material losses.
EMB1	< 1.475	EM	First (strip) layer with fine η segmentation for e/γ separation; variable granularity near the edge.
EMB2	< 1.475	EM	Second layer (shower-max), main EM energy measurement.
EMB3	< 1.35	EM	Back layer, improves containment/leakage control.
Electromagnetic endcap (LAr/lead)			
PreSamplerE	1.5–1.8	EM	Endcap LAr presampler correcting upstream losses.
EME1	1.375–3.2	EM	First (strip) layer with special fine-η strip geometry and varying granularity (finer up to $\|\eta\| \approx 2.5$).
EME2	1.375–3.2	EM	Second layer (shower-max) in EMEC.
EME3	1.5–2.5	EM	Back layer in EMEC.
Hadronic barrel & extended barrel (Tile/steel–scintillator)			
TileBar0	< 1.0	HAD	Barrel A-layer; steel/scintillator tiles; 64 ϕ-modules.
TileBar1	< 1.0	HAD	Barrel BC-layer.
TileBar2	< 1.0	HAD	Barrel D-layer (thickest radial layer).
TileExt0	0.8–1.7	HAD	Extended-barrel A-layer; overlaps the barrel in 0.8–1.0.
TileExt1	0.8–1.7	HAD	Extended-barrel BC-layer.
TileExt2	0.8–1.7	HAD	Extended-barrel D-layer.
Hadronic endcap (HEC; LAr/copper)			
HEC0	1.5–3.2	HAD	First longitudinal HEC sampling (LAr/Cu).
HEC1	1.5–3.2	HAD	Second HEC sampling.
HEC2	1.5–3.2	HAD	Third HEC sampling.

Short name	**$\lvert\eta\rvert$ region**	**EM/HAD**	**Remarks**
HEC3	1.5–3.2	HAD	Fourth HEC sampling (coarser granularity above $\lvert\eta\rvert = 2.5$).
Forward calorimeter (FCal; LAr with Cu/W)			
FCAL0	3.1–4.9	EM	EM forward module (LAr/Cu).
FCAL1	3.1–4.9	HAD	First HAD forward module (LAr/W).
FCAL2	3.1–4.9	HAD	Second HAD forward module (LAr/W).

D Additional Plots

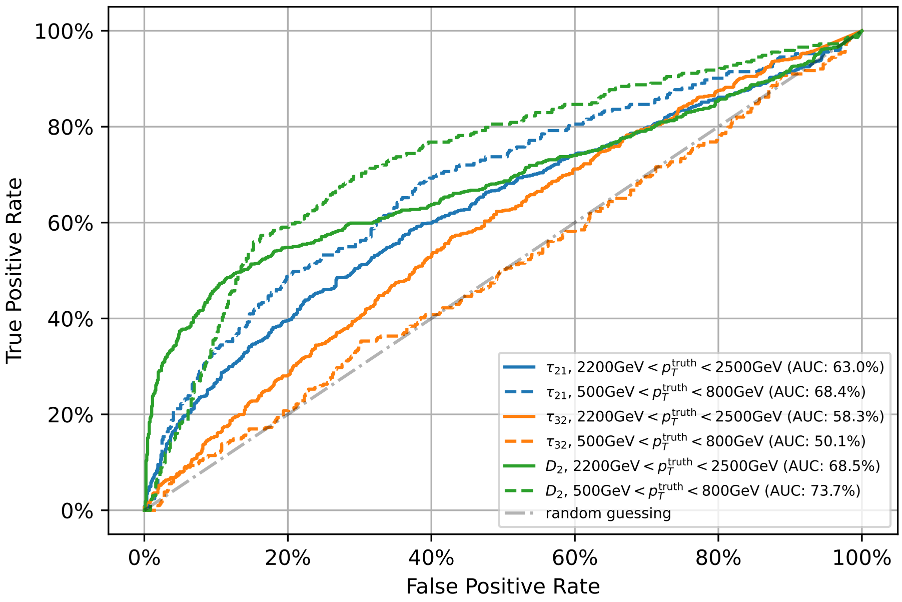


Figure A66: ROC curves for different substructure variables on truth and reconstruction level. (W' + bkg)

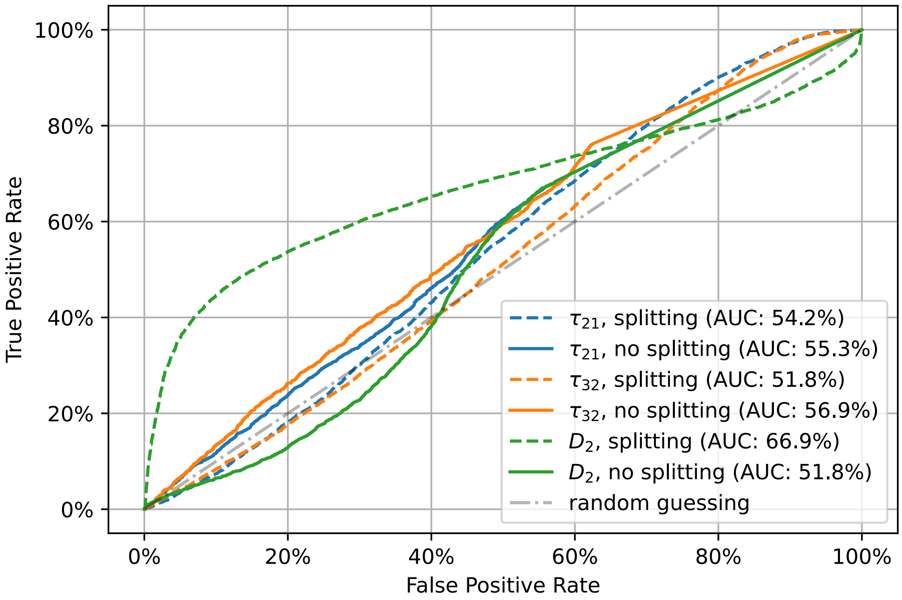


Figure A67: Comparison of ROC curves for substructure variables with and without topo-cluster splitting. (Z' + bkg)

D.1 Grid Search ROC Curves

See Section 3.8 for details.

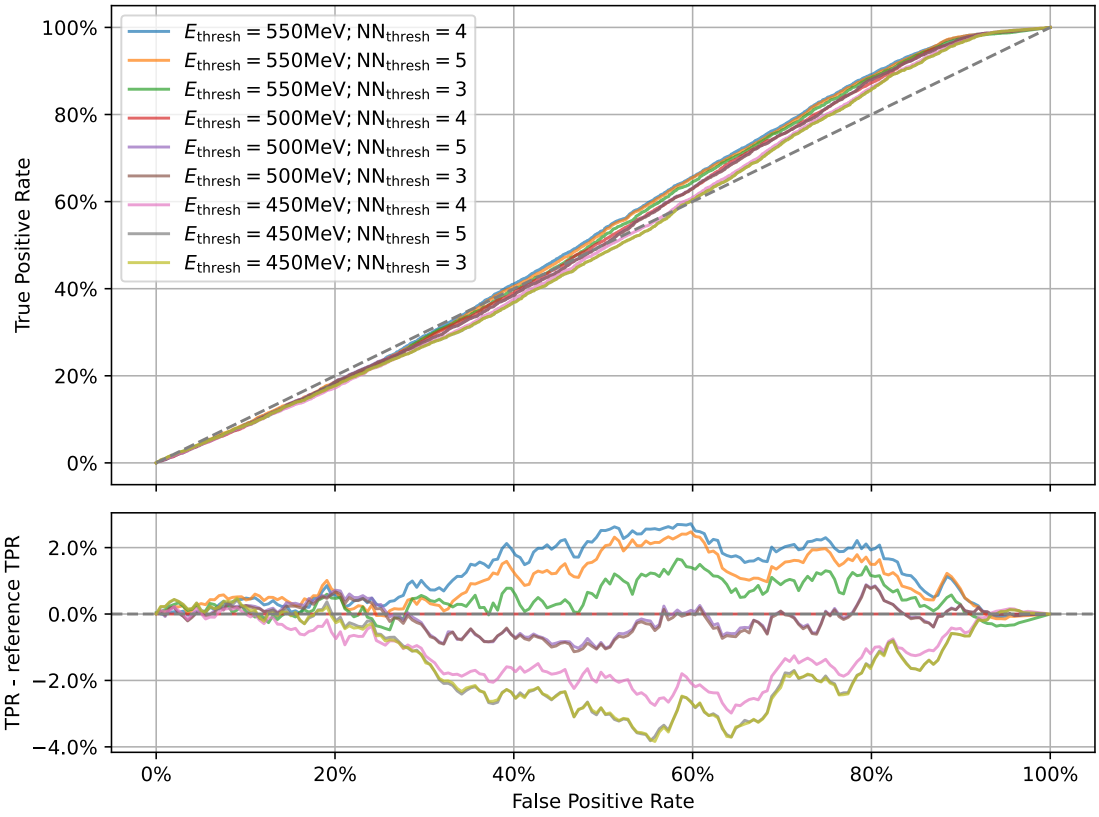


Figure A68: ROC curves for $\tau_{32}^{\mathrm{reco}}$ for different splitting hyperparameters. Residuals are shown relative to the default hyperparameters; positive residuals indicate better performance. (Z' + bkg)

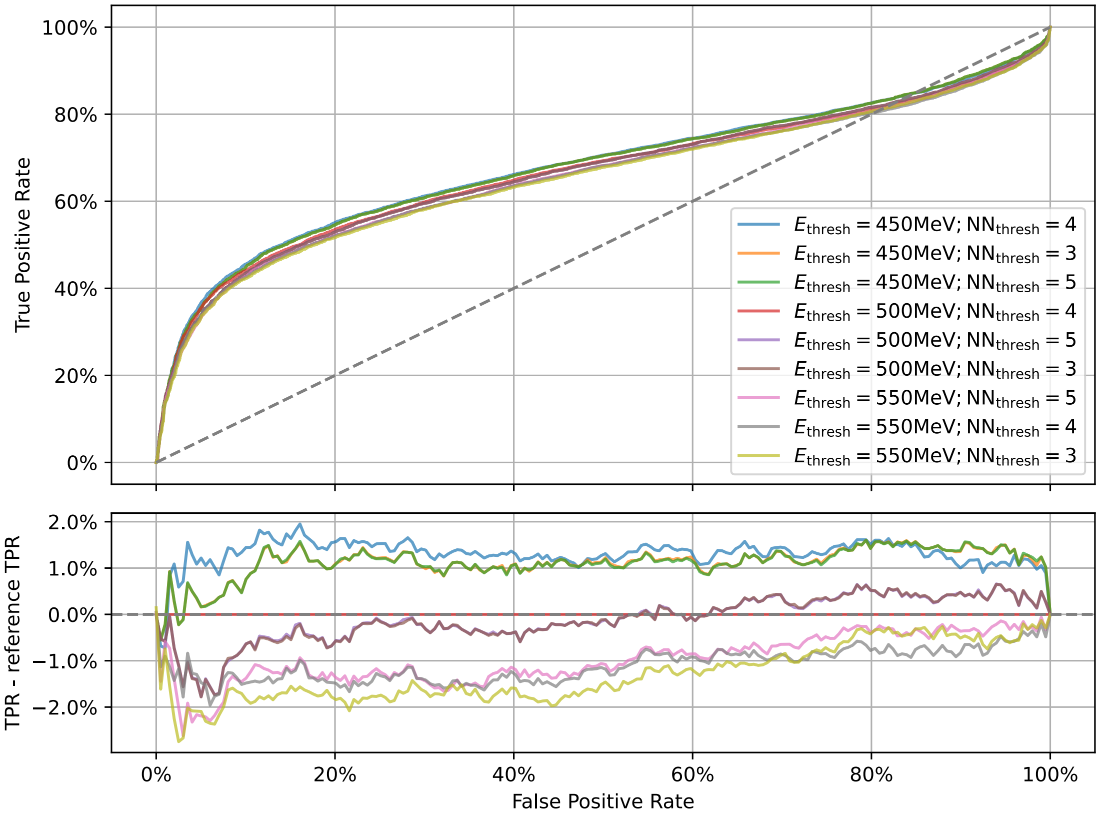
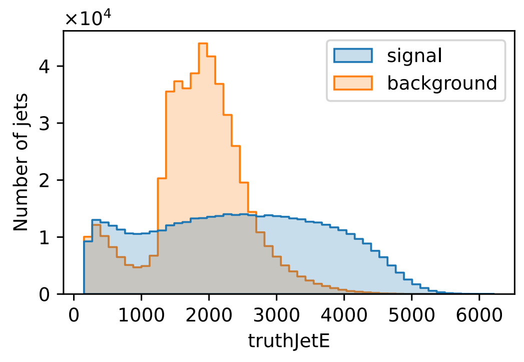


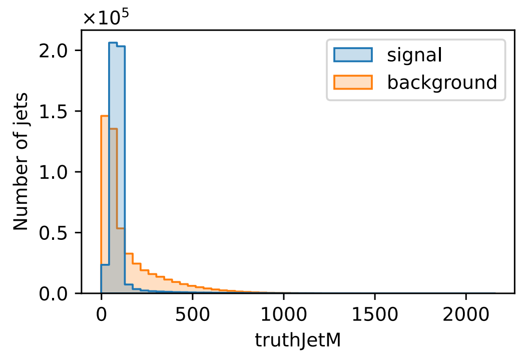
Figure A69: ROC curves for D_2^{reco} for different splitting hyperparameters. Residuals are shown relative to the default hyperparameters; positive residuals indicate better performance. (Z' + bkg)

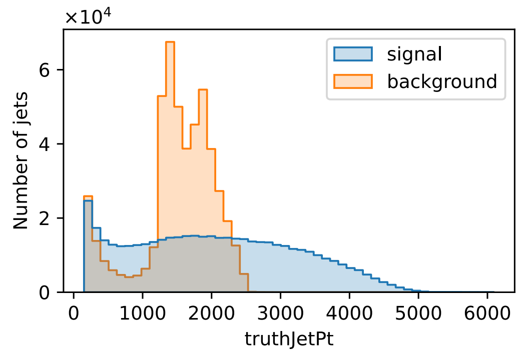
D.2 Signal/Background Distributions

D.2.1 Jet-Level

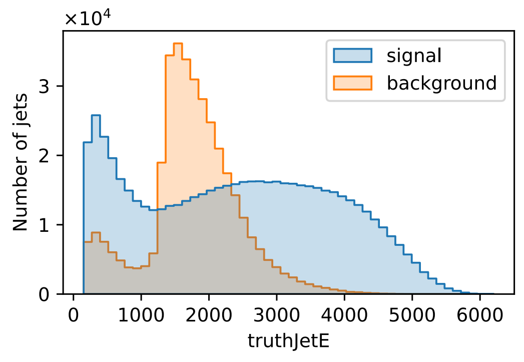
W'

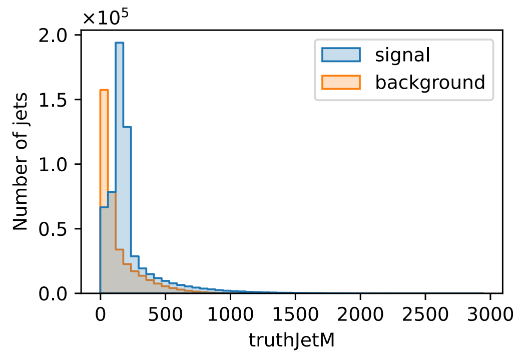


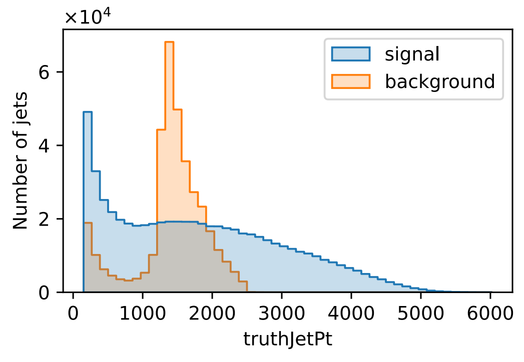




Z'

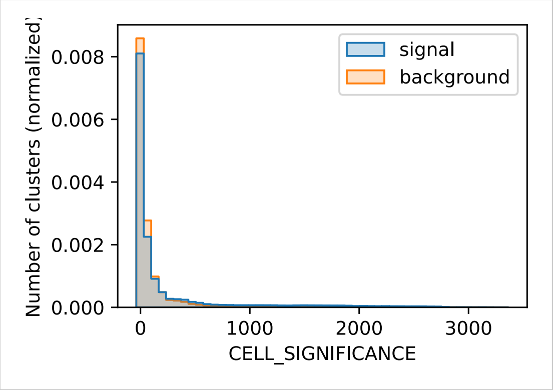


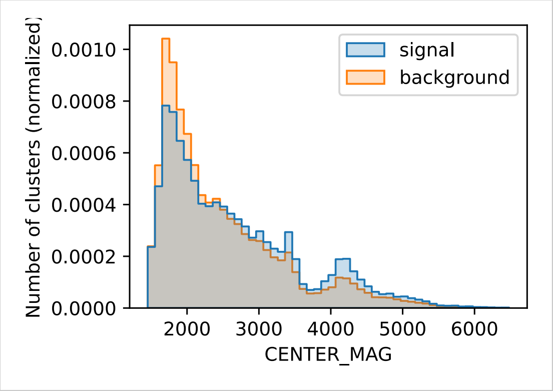


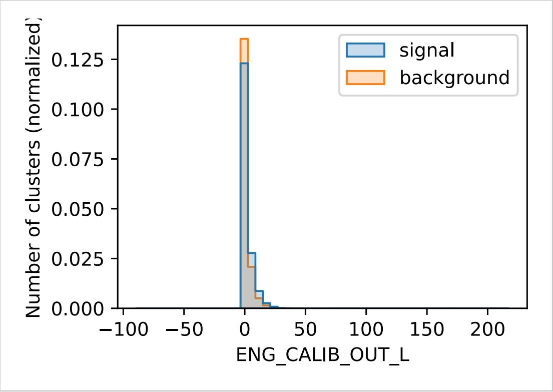


D.2.2 Cluster-Level

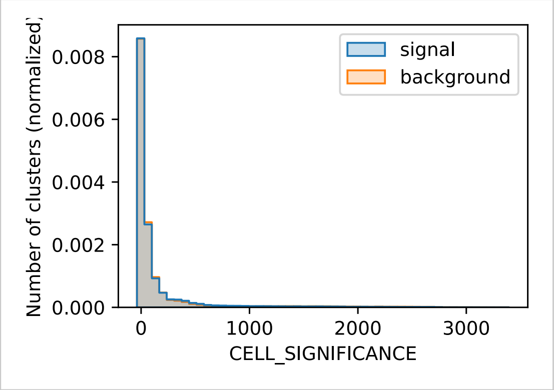
W'

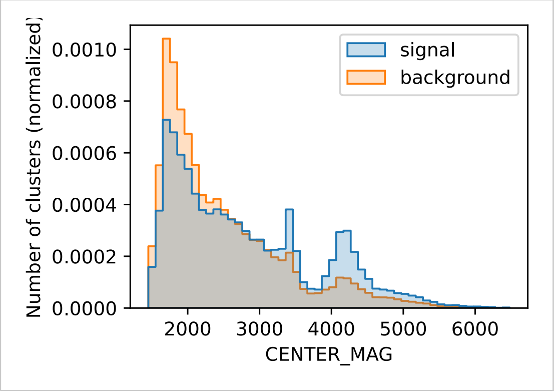


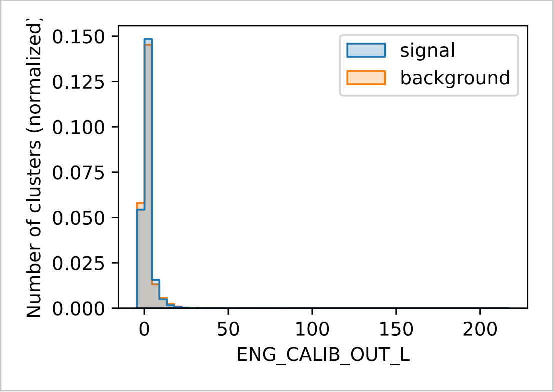




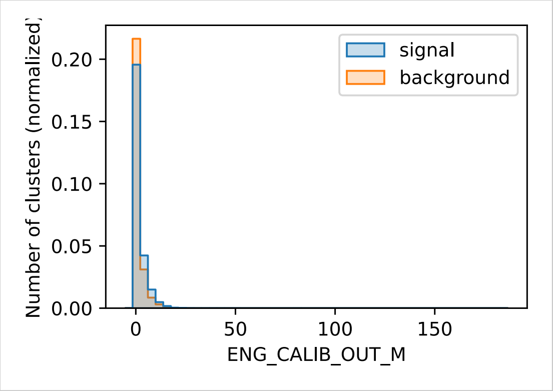
Z'

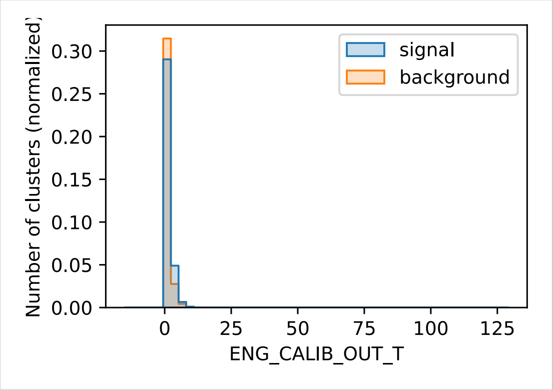


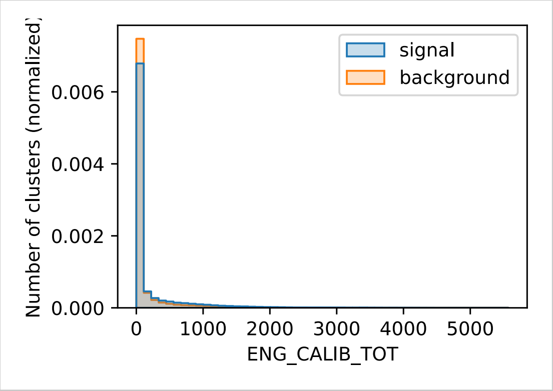


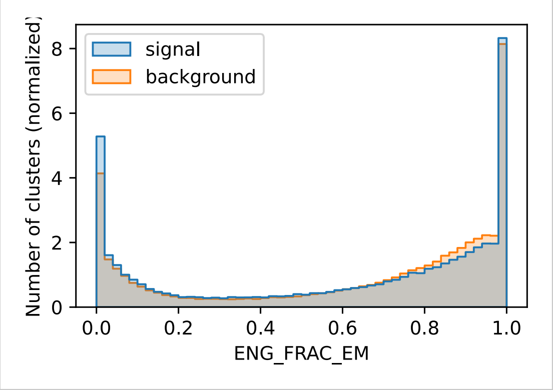


W'

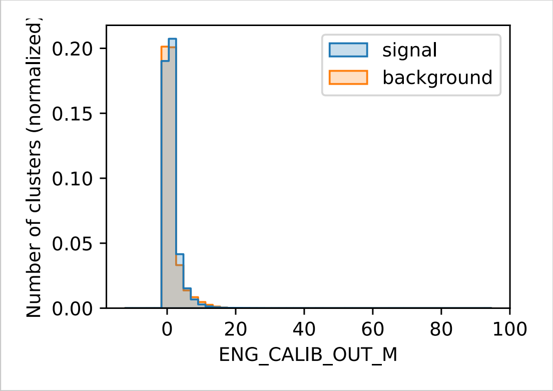

Number of clusters (normalized)
0.00
0.05
0.10
0.15
0.20
0
50
100
150
ENG_CALIB_OUT_M
signal
background

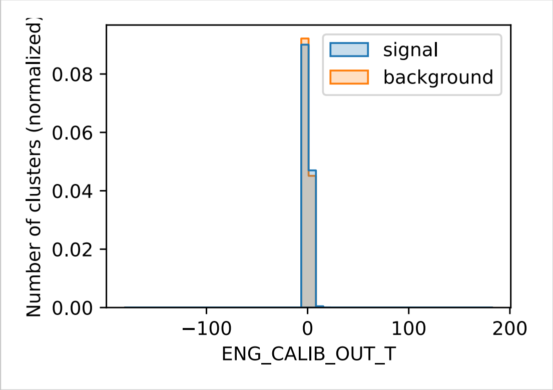

Number of clusters (normalized)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
25
50
75
100
125
ENG_CALIB_OUT_T
signal
background

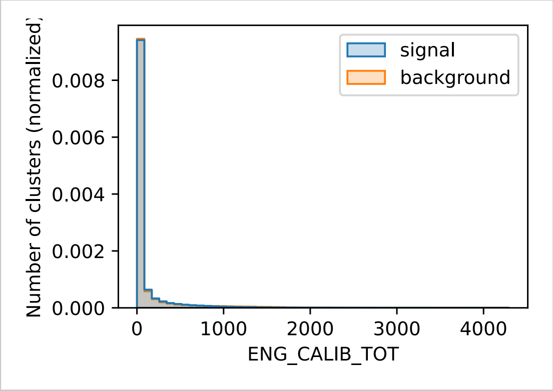

Number of clusters (normalized)
0.000
0.002
0.004
0.006
0
1000
2000
3000
4000
5000
ENG_CALIB_TOT
signal
background

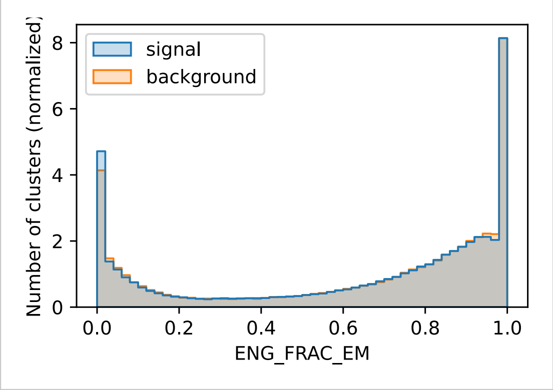

Number of clusters (normalized)
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
ENG_FRAC_EM
signal
background

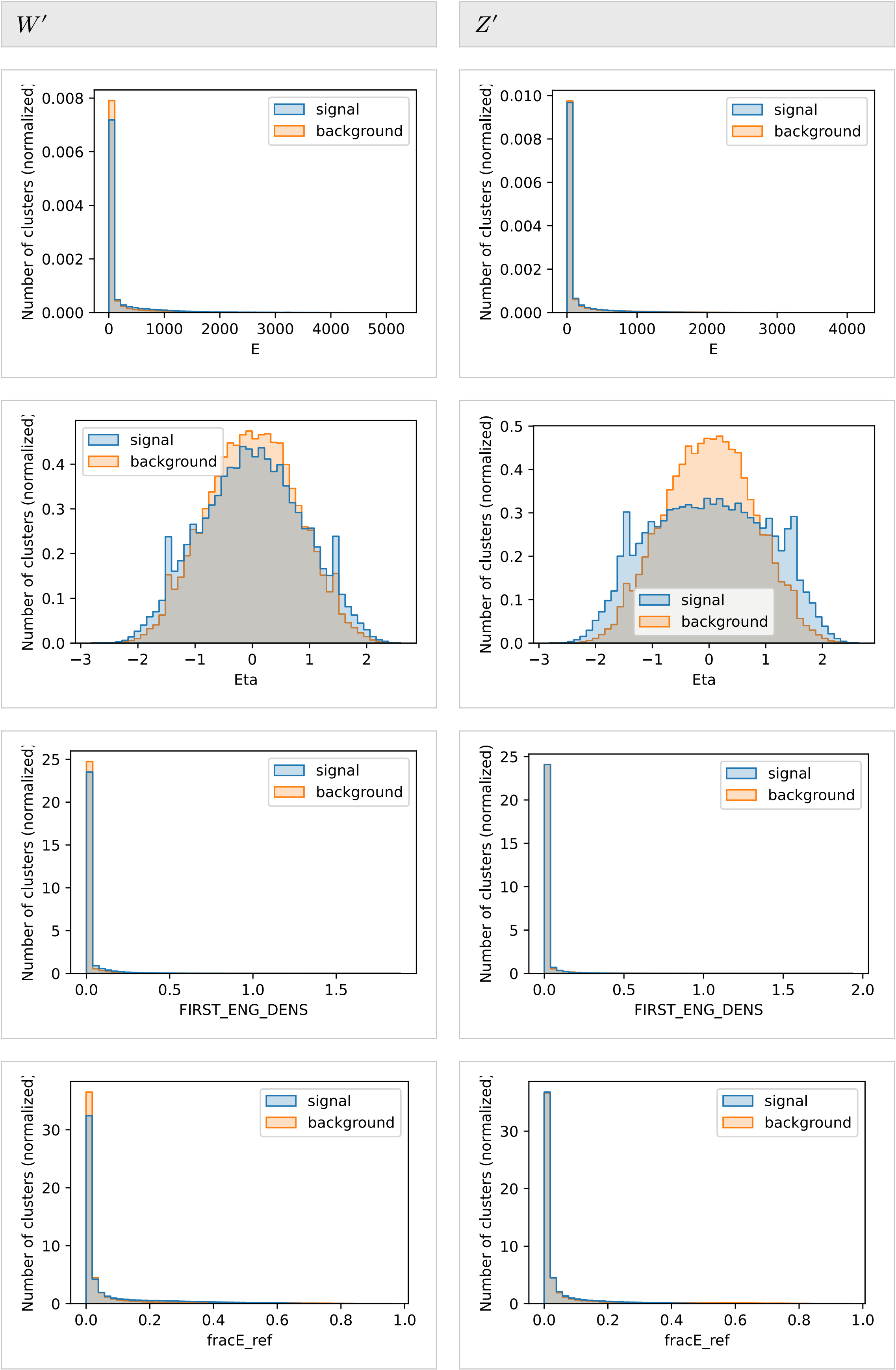
Z'

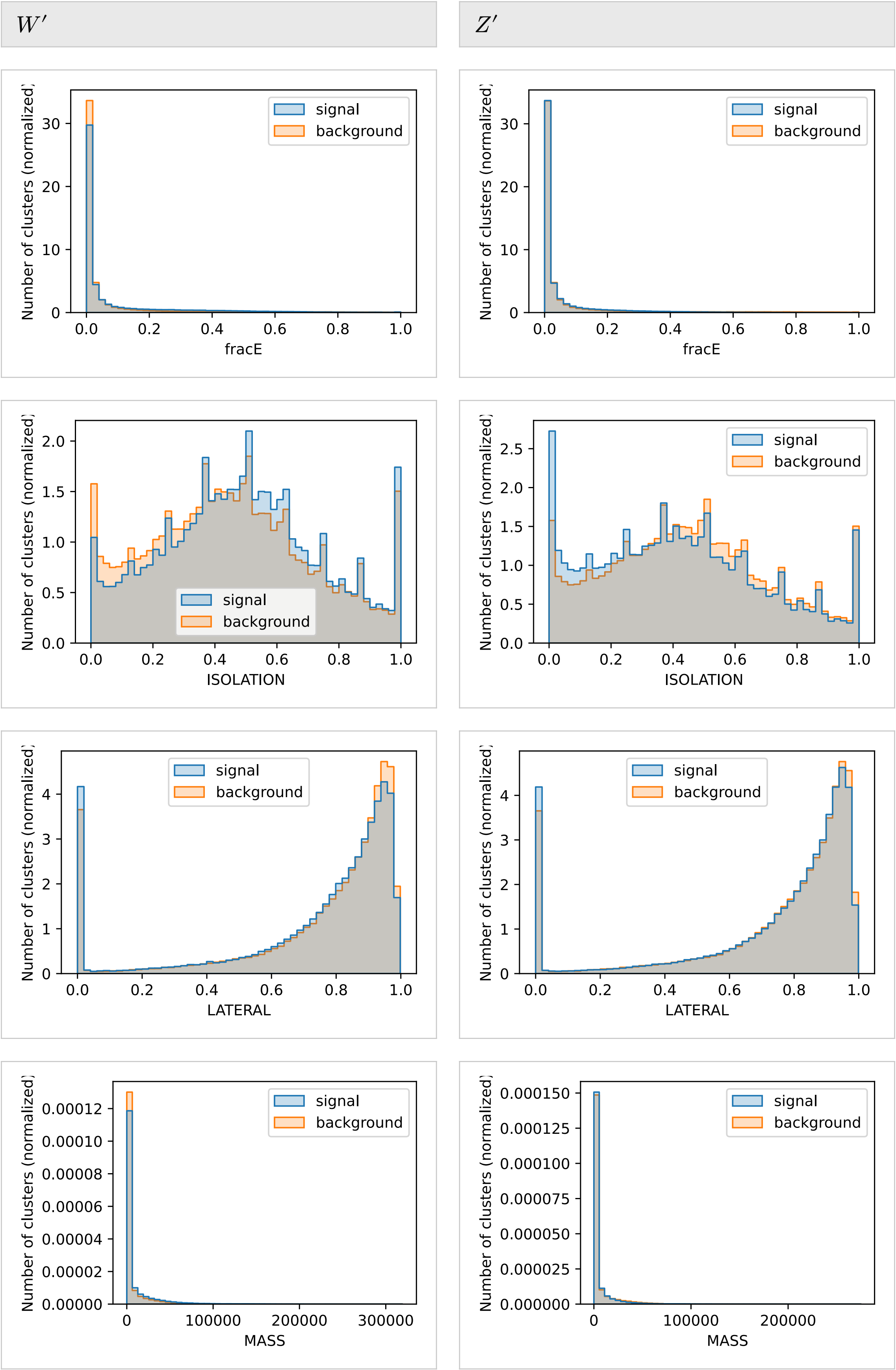

Number of clusters (normalized)
0.00
0.05
0.10
0.15
0.20
0
20
40
60
80
100
ENG_CALIB_OUT_M
signal
background


Number of clusters (normalized)
0.00
0.02
0.04
0.06
0.08
−100
0
100
200
ENG_CALIB_OUT_T
signal
background


Number of clusters (normalized)
0.000
0.002
0.004
0.006
0.008
0
1000
2000
3000
4000
ENG_CALIB_TOT
signal
background

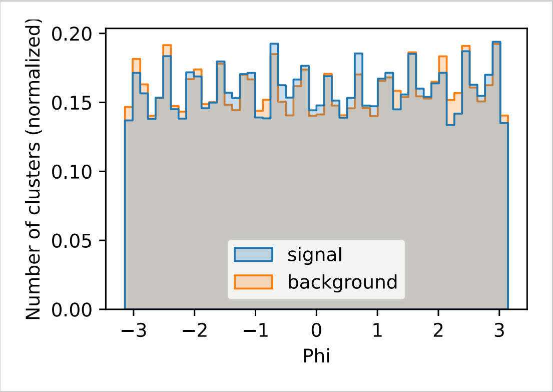

Number of clusters (normalized)
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
ENG_FRAC_EM
signal
background


W'
Z'
Number of clusters (normalized)
signal
background
0.000
0.002
0.004
0.006
0.008
0.010
0
1000
2000
3000
4000
5000
E
0.0
0.1
0.2
0.3
0.4
0.5
−3
−2
−1
0
1
2
Eta
0
5
10
15
20
25
0.0
0.5
1.0
1.5
2.0
FIRST_ENG_DENS
0
10
20
30
0.0
0.2
0.4
0.6
0.8
1.0
fracE_ref

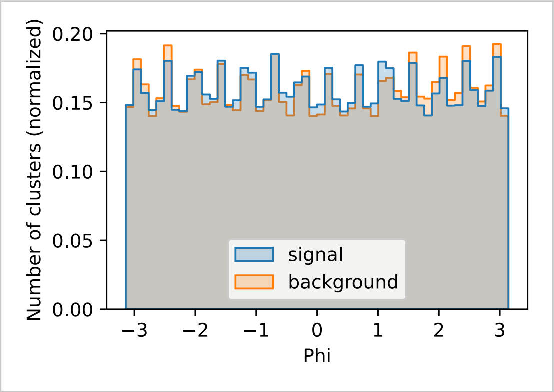

W'
Z'
Number of clusters (normalized)
signal
background
fracE
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
ISOLATION
0.0
0.5
1.0
1.5
2.0
2.5
LATERAL
0
1
2
3
4
MASS
0
100000
200000
300000
0.00000
0.00002
0.00004
0.00006
0.00008
0.00010
0.00012
0.000000
0.000025
0.000050
0.000075
0.000100
0.000125
0.000150

W'

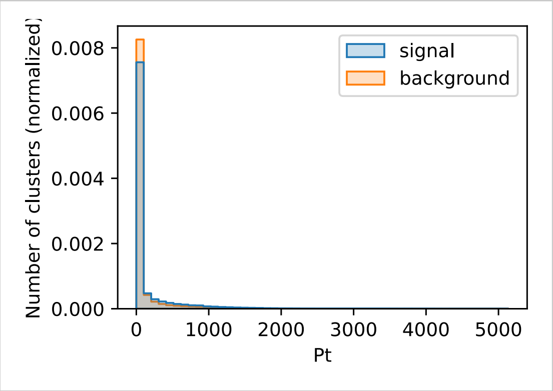
Z'



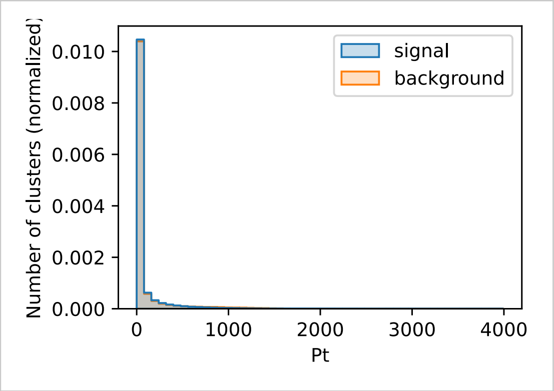
Number of clusters (normalized)
0.20
0.15
0.10
0.05
0.00
−3
−2
−1
0
1
2
3
Phi
signal
background



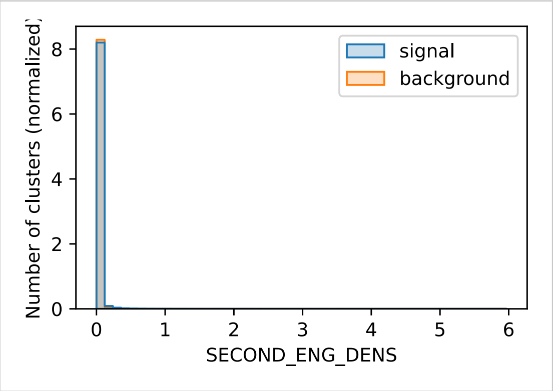
Number of clusters (normalized)
0.20
0.15
0.10
0.05
0.00
−3
−2
−1
0
1
2
3
Phi
signal
background



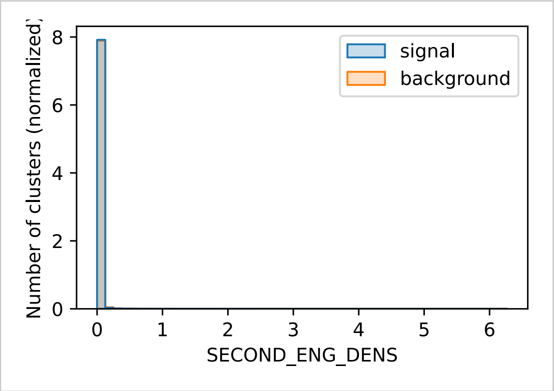
Number of clusters (normalized)
0.008
0.006
0.004
0.002
0.000
0
1000
2000
3000
4000
5000
Pt
signal
background



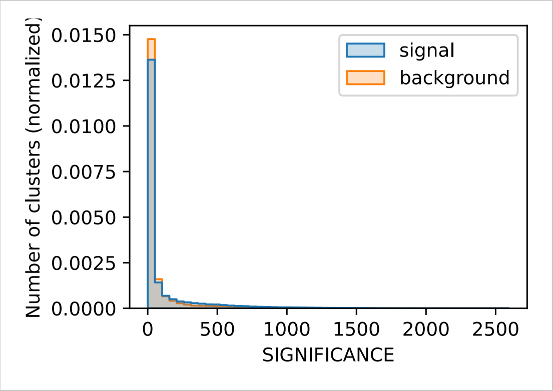
Number of clusters (normalized)
0.010
0.008
0.006
0.004
0.002
0.000
0
1000
2000
3000
4000
Pt
signal
background



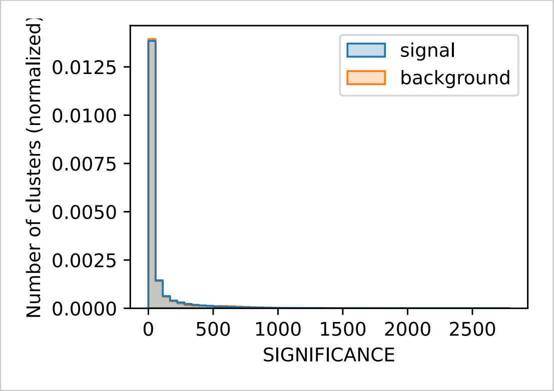
Number of clusters (normalized)
8
6
4
2
0
0
1
2
3
4
5
6
SECOND_ENG_DENS
signal
background



Number of clusters (normalized)
8
6
4
2
0
0
1
2
3
4
5
6
SECOND_ENG_DENS
signal
background



Number of clusters (normalized)
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
500
1000
1500
2000
2500
SIGNIFICANCE
signal
background



Number of clusters (normalized)
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
500
1000
1500
2000
2500
SIGNIFICANCE
signal
background

E Additional Tables

Table A9: 90-percentile of cell pitches in η and ϕ per sampling layer.

Sampling	$\Delta\eta$	$\Delta\phi$/ rad
0	0.0001	0.0932
1	9.2625×10^{-05}	2.3841×10^{-07}
2	2.0027×10^{-05}	4.7683×10^{-06}
3	1.7285×10^{-05}	6.3881×10^{-06}
4	0.0250	0.0899
5	6.8664×10^{-05}	7.2121×10^{-06}
6	1.5020×10^{-05}	3.0994×10^{-05}
7	1.4781×10^{-05}	6.4611×10^{-05}
8	0.2000	0.0981
9	0.1946	0.0981
10	0.1946	0.0981
11	0.1461	0.0981
12	0.1000	0.0981
13	0.1000	0.0981
14	0.2000	0.0981
15	1.9146	0.0981
16	1.7139	0.0981
17	0.9665	0.0981
18	0.5429	0.0981
19	0.5026	0.0981
20	1.6502	0.0981

F Complete Tables

Table A10: Normalized EMDs between cluster-level variables with high/low $p_{T\text{jet}}^{\text{truth}}$. Extended version of Table 1. (W' + bkg)

variable	unit	EMD	mean	mean
			($p_{T\text{jet}}^{\text{truth}} < 800\,\text{GeV}$)	($p_{T\text{jet}}^{\text{truth}} > 2\,000\,\text{GeV}$)
CENTER_MAG	mm	0.594	2897.664	2352.156
CELL_SIGNIFICANCE		0.566	54.704	340.264
MASS	GeV	0.558	1198.655	13 324.379
ENG_CALIB_TOT	GeV	0.551	28.808	310.518
SIGNIFICANCE		0.548	27.426	161.036
ENG_POS	GeV	0.543	29 320.573	294 640.097
sumCellE	GeV	0.543	29.220	294.534
E	GeV	0.543	28.738	293.833
Pt	GeV	0.541	18.524	259.304
FIRST_ETA		0.510	0.026	0.010
Eta		0.510	0.026	0.010
CENTER_Z	mm	0.508	58.703	18.861
ENG_CALIB_DEAD_TOT	GeV	0.436	6.283	34.031
ENG_CALIB_OUT_T	GeV	0.409	0.918	1.339
ENG_CALIB_OUT_M	GeV	0.387	1.182	2.126
FIRST_ENG_DENS	GeV/mm^3	0.355	0.004	0.031
ENG_CALIB_OUT_L	GeV	0.346	1.433	2.702
fracE_ref		0.331	0.043	0.091
fracE		0.303	0.058	0.109
N_BAD_HV_CELLS		0.292	5.124	10.066
OOC_WEIGHT		0.270	1.268	1.174
SECOND_LAMBDA	mm^2	0.268	75 815.091	113 157.814
EM_PROBABILITY		0.254	0.220	0.150
ISOLATION		0.236	0.522	0.463
ENG_FRAC_EM		0.229	0.671	0.587
ENG_CALIB_FRAC_EM		0.224	0.229	0.153
AVG_TILE_Q		0.211	5.441	8.679
CELL_SIG_SAMPLING		0.210	4.159	4.587
DELTA_PHI	rad	0.201	0.001	−0.000
CENTER_LAMBDA	mm	0.192	469.676	570.335
SECOND_ENG_DENS	(GeV/mm^3)^2	0.172	0.000	0.018
SECOND_R	mm^2	0.169	19 116.705	27 976.856
ENG_CALIB_FRAC_HAD		0.147	0.433	0.452
DM_WEIGHT		0.140	1.216	1.168
ENG_FRAC_CORE		0.139	0.452	0.480
time		0.132	−0.368	0.233
ePerSampling		0.132	1026.354	10 490.382
ENG_CALIB_FRAC_REST		0.132	0.334	0.394
N_BAD_CELLS_CORR		0.131	0.055	0.176
N_BAD_CELLS		0.131	0.055	0.176
CENTER_X	mm	0.116	−55.841	3.722
HAD_WEIGHT		0.107	1.069	1.070
PTD		0.105	0.391	0.408
ENG_FRAC_MAX		0.105	0.309	0.332
CENTER_Y	mm	0.091	18.406	0.212
DELTA_THETA	rad	0.086	−0.001	−0.000
ENG_BAD_HV_CELLS		0.082	2493.803	9595.973
LONGITUDINAL		0.075	0.660	0.641
SECOND_TIME		0.064	11.488	6.450

variable	**unit**	**EMD**	**mean** ($p_{T\,\mathrm{jet}}^{\mathrm{truth}} < 800\,\mathrm{GeV}$)	**mean** ($p_{T\,\mathrm{jet}}^{\mathrm{truth}} > 2\,000\,\mathrm{GeV}$)
DELTA_ALPHA		0.061	0.206	0.219
LATERAL		0.059	0.731	0.744
AVG_LAR_Q		0.049	557.054	394.404
ENG_BAD_CELLS		0.025	6.630	41.408
BAD_CELLS_CORR_E		0.025	6.630	41.408
Phi	rad	0.025	0.020	−0.004
FIRST_PHI	rad	0.025	0.020	−0.004
BADLARQ_FRAC		0.006	0.021	0.010

Table A11: Normalized EMDs between cluster-level variables with high/low $p_{T\,\mathrm{cluster}}^{\mathrm{reco}}$. Extended version of Table 2. (W' + bkg)

variable	**unit**	**EMD**	**mean**	**mean**
			($p_{T\,\mathrm{cluster}}^{\mathrm{reco}} < 1\,\mathrm{GeV}$)	($p_{T\,\mathrm{cluster}}^{\mathrm{reco}} > 40\,\mathrm{GeV}$)
ENG_CALIB_OUT_L	GeV	1.673	0.113	5.796
ENG_CALIB_OUT_T	GeV	1.377	0.252	2.298
fracE		1.224	0.000	0.279
ISOLATION		1.208	0.660	0.326
LATERAL		1.206	0.483	0.859
fracE_ref		1.201	0.000	0.229
OOC_WEIGHT		1.196	1.500	1.008
ENG_CALIB_OUT_M	GeV	1.164	0.142	4.324
LONGITUDINAL		1.150	0.401	0.755
SECOND_LAMBDA	mm^2	1.125	12 594.243	171 239.244
CELL_SIGNIFICANCE		1.098	5.612	680.960
CENTER_LAMBDA	mm	1.092	205.074	903.851
SIGNIFICANCE		1.086	3.002	306.981
CELL_SIG_SAMPLING		1.009	1.525	6.830
ENG_CALIB_TOT	GeV	0.983	0.380	547.577
MASS	GeV	0.978	15.995	22 917.014
ENG_POS	GeV	0.977	810.669	524 480.806
sumCellE	GeV	0.977	0.792	524.384
E	GeV	0.976	0.638	523.321
Pt	GeV	0.922	0.453	439.465
PTD		0.892	0.573	0.378
HAD_WEIGHT		0.875	1.035	1.080
EM_PROBABILITY		0.857	0.305	0.072
time		0.807	−1.165	0.523
ENG_FRAC_CORE		0.794	0.642	0.465
ENG_FRAC_MAX		0.774	0.498	0.301
N_BAD_HV_CELLS		0.757	1.279	15.474
ENG_CALIB_DEAD_TOT	GeV	0.745	0.568	59.739
SECOND_R	mm^2	0.713	2877.340	37 029.272
AVG_TILE_Q		0.663	0.386	10.146
CENTER_MAG	mm	0.606	2250.191	2771.710
FIRST_ENG_DENS	$\mathrm{GeV/mm}^3$	0.587	0.000	0.057
ENG_FRAC_EM		0.499	0.683	0.483
DELTA_THETA	rad	0.387	−0.001	−0.000
DM_WEIGHT		0.299	1.141	1.109
ENG_CALIB_FRAC_HAD		0.295	0.491	0.414
CENTER_Y	mm	0.285	6.695	−2.737
CENTER_X	mm	0.280	−6.317	−3.099
N_BAD_CELLS		0.270	0.004	0.280
N_BAD_CELLS_CORR		0.270	0.004	0.280
ENG_CALIB_FRAC_REST		0.240	0.365	0.372
DELTA_ALPHA		0.223	0.161	0.207
ePerSampling		0.221	22.796	18 683.715
AVG_LAR_Q		0.221	897.382	80.629
SECOND_ENG_DENS	$(\mathrm{GeV/mm}^3)^2$	0.203	4.046×10^{-06}	0.027
ENG_CALIB_FRAC_EM		0.182	0.141	0.213
Eta		0.172	0.011	0.004
FIRST_ETA		0.172	0.011	0.004
ENG_BAD_HV_CELLS		0.146	283.132	18 465.283
DELTA_PHI	rad	0.134	0.000	0.000
SECOND_TIME		0.115	6.042	0.372

variable	**unit**	**EMD**	**mean**	**mean**
			($p_{T\,\text{cluster}}^{\text{reco}} < 1\,\text{GeV}$)	($p_{T\,\text{cluster}}^{\text{reco}} > 40\,\text{GeV}$)
BAD_CELLS_CORR_E		0.046	0.064	71.603
ENG_BAD_CELLS		0.046	0.064	71.603
CENTER_Z	mm	0.034	22.567	9.947
BADLARQ_FRAC		0.010	0.044	0.000
Phi	rad	0.007	0.015	0.008
FIRST_PHI	rad	0.006	0.015	0.008

Table A12: Normalized EMDs between cluster-level variables for jets with high/low number of clusters. (Z' + bkg)

variable	**unit**	**EMD**	**mean**	**mean**
			($n_{\text{clus}} \leq 8$)	($n_{\text{clus}} \geq 40$)
ENG_CALIB_OUT_T	GeV	1.189	1.653	0.825
CENTER_MAG	mm	0.768	2401.060	3161.950
fracE_ref		0.766	0.150	0.016
fracE		0.764	0.179	0.021
MASS	GeV	0.709	14 230.989	1560.402
Pt	GeV	0.692	269.680	24.105
ENG_CALIB_TOT	GeV	0.684	324.952	40.047
CELL_SIGNIFICANCE		0.682	402.808	64.568
ENG_POS	GeV	0.671	309 418.851	41 332.852
sumCellE	GeV	0.671	309.292	41.247
E	GeV	0.670	308.383	40.877
SIGNIFICANCE		0.634	173.253	36.627
Eta		0.629	−0.008	0.090
FIRST_ETA		0.629	−0.008	0.090
CENTER_Z	mm	0.613	−20.857	203.435
ENG_CALIB_OUT_M	GeV	0.611	2.674	1.062
ENG_CALIB_OUT_L	GeV	0.578	3.392	1.319
ENG_CALIB_DEAD_TOT	GeV	0.490	35.036	10.576
N_BAD_HV_CELLS		0.435	12.331	4.847
FIRST_ENG_DENS	GeV/mm^3	0.422	0.035	0.005
SECOND_LAMBDA	mm^2	0.399	134 607.566	76 368.737
ISOLATION		0.362	0.484	0.392
CELL_SIG_SAMPLING		0.295	5.098	4.221
LATERAL		0.293	0.788	0.704
CENTER_LAMBDA	mm	0.280	632.274	474.859
DM_WEIGHT		0.273	1.154	1.244
EM_PROBABILITY		0.262	0.132	0.204
DELTA_PHI	rad	0.252	−0.002	0.000
HAD_WEIGHT		0.242	1.076	1.062
SECOND_R	mm^2	0.235	30 336.795	18 631.993
SECOND_ENG_DENS	(GeV/mm^3)^2	0.227	0.018	0.001
ENG_FRAC_CORE		0.177	0.460	0.458
N_BAD_CELLS_CORR		0.176	0.206	0.047
N_BAD_CELLS		0.176	0.206	0.047
AVG_TILE_Q		0.174	8.639	5.725
ENG_FRAC_EM		0.168	0.574	0.636
PTD		0.164	0.385	0.403
ePerSampling		0.158	11 013.194	1459.886
CENTER_Y	mm	0.156	11.123	−4.952
ENG_FRAC_MAX		0.154	0.308	0.325
CENTER_X	mm	0.145	11.452	−14.074
ENG_CALIB_FRAC_HAD		0.143	0.458	0.425
OOC_WEIGHT		0.139	1.138	1.198
LONGITUDINAL		0.135	0.682	0.642
time		0.130	0.377	−0.056
DELTA_THETA	rad	0.117	−0.001	−0.000
DELTA_ALPHA		0.106	0.215	0.193
ENG_CALIB_FRAC_EM		0.099	0.191	0.217
SECOND_TIME		0.084	5.346	10.371
ENG_BAD_HV_CELLS		0.076	9669.030	4126.919
AVG_LAR_Q		0.064	298.900	509.723

variable	**unit**	**EMD**	**mean**	**mean**
			($n_{\text{clus}} \leq 8$)	($n_{\text{clus}} \geq 40$)
ENG_CALIB_FRAC_REST		0.056	0.349	0.356
ENG_BAD_CELLS		0.046	24.259	6.848
BAD_CELLS_CORR_E		0.046	24.259	6.848
BADLARQ_FRAC		0.028	0.005	0.014
FIRST_PHI	rad	0.024	0.028	−0.008
Phi	rad	0.023	0.028	−0.007

Table A13: Normalized EMDs between cluster-level variables with and without topo-cluster splitting enabled. Extended version of Table 4. (W' + bkg)

variable	unit	EMD	mean	mean
			(no splitting)	(splitting)
nCells_tot		1.480	701.013	137.022
nCells		1.460	539.010	106.664
fracECalib_ref		1.363	0.443	0.077
ENG_CALIB_OUT_L	GeV	1.358	11.535	1.998
fracE_ref		1.356	0.385	0.065
fracECalib		1.354	0.461	0.080
fracE		1.339	0.461	0.080
MASS	GeV	1.242	64 481.297	6426.914
sumCellECalib	GeV	1.160	955.366	167.641
ECalib	GeV	1.157	950.753	166.866
ENG_CALIB_TOT	GeV	1.148	868.934	150.104
ENG_POS	GeV	1.148	834 479.445	144 174.022
sumCellE	GeV	1.147	834.345	144.074
E	GeV	1.145	830.368	143.512
PtCalib	GeV	1.121	793.346	138.978
Pt	GeV	1.110	692.734	119.676
ENG_CALIB_OUT_T	GeV	1.101	3.445	1.083
ENG_CALIB_DEAD_TOT	GeV	1.037	110.442	19.197
N_BAD_HV_CELLS		1.035	37.961	7.814
ISOLATION		0.936	0.711	0.465
CELL_SIGNIFICANCE		0.865	694.440	193.960
SIGNIFICANCE		0.693	248.233	89.954
SECOND_LAMBDA	mm^2	0.675	212 259.981	99 359.814
ENG_CALIB_OUT_M	GeV	0.444	−1.000	1.590
CELL_SIG_SAMPLING		0.443	2.062	4.315
DELTA_ALPHA		0.436	0.125	0.212
FIRST_ENG_DENS	GeV/mm^3	0.420	0.050	0.015
N_BAD_CELLS_CORR		0.402	0.612	0.115
N_BAD_CELLS		0.402	0.612	0.115
HAD_WEIGHT		0.379	1.047	1.070
LATERAL		0.333	0.687	0.743
CENTER_LAMBDA	mm	0.313	486.009	520.934
ePerSampling		0.304	29 651.117	5116.480
AVG_TILE_Q		0.272	3.900	7.311
DELTA_THETA	rad	0.267	−0.000	−0.000
PTD		0.265	0.436	0.394
ENG_BAD_HV_CELLS		0.246	29 872.849	6058.678
ENG_FRAC_MAX		0.242	0.368	0.316
ENG_FRAC_CORE		0.241	0.518	0.460
DM_WEIGHT		0.225	1.135	1.191
DELTA_PHI	rad	0.225	0.001	0.000
ENG_FRAC_EM		0.224	0.562	0.627
LONGITUDINAL		0.207	0.604	0.652
SECOND_ENG_DENS	(GeV/mm^3)^2	0.167	0.027	0.007
SECOND_R	mm^2	0.146	34 605.457	24 377.216
ENG_CALIB_FRAC_HAD		0.126	0.411	0.454
EM_PROBABILITY		0.120	0.141	0.174
CENTER_MAG	mm	0.111	2528.488	2544.688
time		0.083	−0.221	0.042
ENG_BAD_CELLS		0.061	117.447	20.329
BAD_CELLS_CORR_E		0.061	117.447	20.329

variable	**unit**	**EMD**	**mean**	**mean**
			(no splitting)	**(splitting)**
ENG_CALIB_FRAC_EM		0.059	0.246	0.187
OOC_WEIGHT		0.052	1.177	1.202
CENTER_Y	mm	0.037	−10.455	5.474
FIRST_ETA		0.035	0.008	0.007
EtaCalib		0.034	0.008	0.007
Eta		0.034	0.008	0.007
CENTER_Z	mm	0.032	16.033	14.401
ENG_CALIB_FRAC_REST		0.029	0.338	0.357
SECOND_TIME		0.029	4.232	8.347
CENTER_X	mm	0.027	−9.183	−14.466
AVG_LAR_Q		0.009	414.041	426.207
PhiCalib	rad	0.008	0.005	0.018
Phi	rad	0.008	0.006	0.018
FIRST_PHI	rad	0.008	0.006	0.018
BADLARQ_FRAC		0.004	0.013	0.016

Table A14: Normalized EMDs between cluster-level variables for non-split clusters matching 1 or ≥ 2 split clusters. Extended version of Table 5. (Z' + bkg)

variable	unit	EMD	mean ($m_i = 1$)	mean ($m_i \geq 2$)
fracE		1.885	0.013	0.924
fracE_ref		1.882	0.011	0.779
SECOND_LAMBDA	mm^2	1.552	16 065.308	452 754.603
ENG_POS	GeV	1.526	16 259.747	1 744 003.767
sumCellE	GeV	1.526	16.231	1743.671
ENG_CALIB_TOT	GeV	1.525	14.446	1798.484
E	GeV	1.524	16.032	1733.561
CELL_SIGNIFICANCE		1.523	28.925	1278.453
SIGNIFICANCE		1.521	15.880	433.215
Pt	GeV	1.473	13.230	1340.184
ENG_CALIB_OUT_L	GeV	1.462	0.353	29.343
PTD		1.419	0.653	0.200
LATERAL		1.408	0.399	0.961
LONGITUDINAL		1.406	0.343	0.871
ENG_FRAC_CORE		1.342	0.699	0.297
ENG_FRAC_MAX		1.326	0.597	0.132
N_BAD_HV_CELLS		1.304	2.024	99.875
MASS	GeV	1.220	695.322	178 633.243
ENG_CALIB_DEAD_TOT	GeV	1.219	4.316	279.555
ENG_CALIB_OUT_T	GeV	1.138	0.476	6.498
ISOLATION		1.104	0.561	0.825
CENTER_LAMBDA	mm	1.013	320.671	690.126
EM_PROBABILITY		0.943	0.246	0.006
ENG_FRAC_EM		0.922	0.485	0.605
SECOND_R	mm^2	0.843	4616.604	103 860.578
HAD_WEIGHT		0.838	1.029	1.063
FIRST_ENG_DENS	GeV/mm^3	0.805	0.003	0.074
OOC_WEIGHT		0.673	1.302	1.014
time		0.610	−0.575	0.295
N_BAD_CELLS		0.584	0.017	1.390
N_BAD_CELLS_CORR		0.584	0.017	1.390
ENG_CALIB_FRAC_HAD		0.571	0.412	0.370
ENG_CALIB_FRAC_EM		0.555	0.255	0.252
ENG_BAD_HV_CELLS		0.483	2102.052	77 702.778
ENG_CALIB_FRAC_REST		0.481	0.328	0.377
CELL_SIG_SAMPLING		0.444	1.702	2.596
AVG_TILE_Q		0.429	1.313	7.336
ePerSampling		0.407	572.597	61 909.086
CENTER_MAG	mm	0.399	2574.880	2776.923
DM_WEIGHT		0.353	1.202	1.061
DELTA_THETA	rad	0.234	0.000	−0.000
SECOND_ENG_DENS	(GeV/mm^3)^2	0.229	0.001	0.035
CENTER_Y	mm	0.225	−7.153	−4.255
DELTA_PHI	rad	0.216	−0.002	0.001
FIRST_ETA		0.212	-9.303×10^{-05}	0.000
Eta		0.212	-9.445×10^{-05}	0.000
AVG_LAR_Q		0.207	778.154	35.265
CENTER_X	mm	0.204	−13.425	−2.855
BADLARQ_FRAC		0.177	0.023	0.001
DELTA_ALPHA		0.176	0.127	0.138
SECOND_TIME		0.160	6.801	0.757

variable	**unit**	**EMD**	**mean**	**mean**
			($m_i = 1$)	($m_i \geq 2$)
BAD_CELLS_CORR_E		0.152	1.471	252.702
ENG_BAD_CELLS		0.152	1.471	252.702
CENTER_Z	mm	0.113	0.745	0.691
Phi	rad	0.010	-0.006	0.001
FIRST_PHI	rad	0.010	-0.006	0.001
ENG_CALIB_OUT_M	GeV	0.000	-1.000	-1.000

Table A15: Normalized EMDs between $N_{\text{matching}} \leq 5$ and $N_{\text{matching}} \geq 20$. Extended version of Table 6. (convex hull method) (Z' + bkg)

variable	**unit**	**EMD**	**mean**	**mean**
			($N_{\text{matching}} \leq 5$)	($N_{\text{matching}} \geq 20$)
ENG_CALIB_OUT_T	GeV	2.690	0.604	1.754
ENG_CALIB_OUT_M	GeV	1.438	0.544	3.350
ENG_CALIB_OUT_L	GeV	1.369	0.571	4.568
fracE		1.272	0.006	0.194
fracE_ref		1.247	0.005	0.158
CELL_SIGNIFICANCE		1.226	18.454	515.279
MASS	GeV	1.169	404.816	17 571.809
SIGNIFICANCE		1.166	11.096	235.392
ENG_CALIB_TOT	GeV	1.153	8.491	403.457
ENG_POS	GeV	1.142	8822.642	387 454.358

Table A16: Normalized EMDs between $N_{\text{matching}} \leq 5$ and $N_{\text{matching}} \geq 20$. Extended version of Table 7. (pitch-aware method) (Z' + bkg)

variable	**unit**	**EMD**	**mean**	**mean**
			($N_{\text{matching}} \leq 5$)	($N_{\text{matching}} \geq 20$)
ENG_CALIB_OUT_T	GeV	2.360	0.479	1.449
ENG_CALIB_OUT_L	GeV	1.092	0.593	2.926
OOC_WEIGHT		1.079	1.347	1.041
SECOND_LAMBDA	mm^2	1.070	17 920.171	175 924.693
ENG_CALIB_OUT_M	GeV	1.054	0.505	2.275
CELL_SIG_SAMPLING		1.041	1.579	6.936
CENTER_LAMBDA	mm	0.972	211.208	802.358
ISOLATION		0.934	0.552	0.304
LATERAL		0.865	0.617	0.862
AVG_TILE_Q		0.851	0.319	14.113

Table A17: Overview of the performance of different splitting hyperparameters in terms of the AUC of different substructure variables as well as their improvement relative to the default hyperparameters. (W' + bkg)

E_{thresh}/ MeV	N_{thresh}	AUC(τ_{21}^{reco})	AUC(τ_{32}^{reco})	AUC(D_2^{reco})	ΔAUC(τ_{21}^{reco})	ΔAUC(τ_{32}^{reco})	ΔAUC(D_2^{reco})	⟨ΔAUC⟩
550	4	0.646	0.589	0.653	−0.001	−0.003	+0.008	0.001
550	5	0.646	0.589	0.653	−0.001	−0.004	+0.008	0.001
550	3	0.644	0.589	0.654	−0.003	−0.003	+0.009	0.000
500	4	0.647	0.593	0.644	+0.000	+0.000	+0.000	0.000
500	3	0.649	0.584	0.649	+0.002	−0.008	+0.004	−0.000
500	5	0.649	0.584	0.649	+0.002	−0.008	+0.004	−0.000
450	4	0.652	0.582	0.643	+0.004	−0.010	−0.001	−0.002
450	3	0.652	0.582	0.643	+0.004	−0.010	−0.001	−0.002
450	5	0.652	0.582	0.643	+0.004	−0.010	−0.001	−0.002

Bibliography

[1] T. J. Berners-Lee, Information management: a proposal, 1989.

[2] L. Evans and P. B. (Eds.), LHC Machine, JINST **3**, 8001 (2008).

[3] J. Thaler and K. Van Tilburg, Identifying boosted objects with N-subjettiness, Journal of High Energy Physics **2011**, 1 (2011).

[4] A. J. Larkoski, G. P. Salam, and J. Thaler, Energy correlation functions for jet substructure, Journal of High Energy Physics **2013**, 1 (2013).

[5] A. J. Larkoski, I. Moult, and D. Neill, Power counting to better jet observables, Journal of High Energy Physics **2014**, 1 (2014).

[6] The ATLAS Experiment at the CERN Large Hadron Collider, JINST **3**, S8003 (2008).

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

[8] S. L. Glashow, Partial Symmetries of Weak Interactions, Nucl. Phys. **22**, 579 (1961).

[9] S. Weinberg, A Model of Leptons, Phys. Rev. Lett. **19**, 1264 (1967).

[10] A. Salam, Weak and Electromagnetic Interactions, Conf. Proc. C **680519**, 367 (1968).

[11] G. 't Hooft and M. J. G. Veltman, Regularization and Renormalization of Gauge Fields, Nucl. Phys. B **44**, 189 (1972).

[12] ATLAS Collaboration, Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Physics Letters B **716**, 1 (2012).

[13] CMS Collaboration and others, Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, Arxiv Preprint Arxiv:1207.7235 (2012).

[14] D. O. et al., *Approaches to Quantum Gravity* (Cambridge University Press, 2009).

[15] R. N. Mohapatra and A. Y. Smirnov, Neutrino mass and new physics, Annu. Rev. Nucl. Part. Sci. **56**, 569 (2006).

[16] A. D. Sakharov, Violation of CP in variance, C asymmetry, and baryon asymmetry of the universe, Physics-Uspekhi **34**, 392 (1991).

[17] G. Bertone, D. Hooper, and J. Silk, Particle dark matter: Evidence, candidates and constraints, Phys. Rept. **405**, 279 (2005).

[18] Cush, Standard Model of Elementary Particles, (2019).

[19] F. Englert and R. Brout, Broken Symmetry and the Mass of Gauge Vector Mesons, Phys. Rev. Lett. **13**, 321 (1964).

[20] P. W. Higgs, Broken symmetries, massless particles and gauge fields, Phys. Lett. **12**, 132 (1964).

[21] P. W. Higgs, Broken Symmetries and the Masses of Gauge Bosons, Phys. Rev. Lett. **13**, 508 (1964).

[22] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, Global Conservation Laws and Massless Particles, Phys. Rev. Lett. **13**, 585 (1964).

[23] P. W. Higgs, Spontaneous Symmetry Breakdown without Massless Bosons, Phys. Rev. **145**, 1156 (1966).

[24] T. W. B. Kibble, Symmetry breaking in non-Abelian gauge theories, Phys. Rev. **155**, 1554 (1967).

[25] G. Altarelli, B. Melé, and M. Ruiz-Altaba, Searching for new heavy vector bosons in p colliders, Zeitschrift Für Physik C Particles and Fields **45**, 109 (1989).

[26] CERN, About CERN, (n.d.).

[27] G. Arnison et al., Experimental observation of isolated large transverse energy electrons with associated missing energy at s= 540 GeV, Physics Letters B **122**, 103 (1983).

[28] G. t. Arnison et al., Experimental observation of lepton pairs of invariant mass around 95 GeV/c^2 at the CERN SPS collider, Physics Letters B **126**, 398 (1983).

[29] ALICE Collaboration, The ALICE experiment at the CERN LHC, JINST **3**, 8002 (2008).

[30] CMS Collaboration, The CMS Experiment at the CERN LHC, JINST **3**, 8004 (2008).

[31] LHCb Collaboration, The LHCb Detector at the LHC, JINST **3**, 8005 (2008).

[32] CERN LHC Operations, Longer term LHC schedule, (2024).

[33] ATLAS Collaboration and others, Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC, Arxiv Preprint Arxiv:2212.09379 (2022).

[34] G. Apollinari, O. Brüning, T. Nakamoto, and L. Rossi, High luminosity large hadron collider HL-LHC, Arxiv Preprint Arxiv:1705.08830 (2017).

[35] ATLAS Collaboration, Letter of Intent for the Phase-II Upgrade of the ATLAS Experiment, 2012.

[36] S. Mehlhase, ATLAS detector slice (and particle visualisations), (2021).

[37] ATLAS Collaboration, Jet energy scale and resolution measured in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Eur. Phys. J. C **81**, 689 (2021).

[38] 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).

[39] Jet reconstruction and performance using particle flow with the ATLAS Detector, The European Physical Journal C **77**, 466 (2017).

[40] Improving jet substructure performance in ATLAS using Track-CaloClusters, 2017.

[41] Optimisation of large-radius jet reconstruction for the ATLAS detector in 13 TeV proton–proton collisions, The European Physical Journal C **81**, 334 (2021).

[42] K. T. Greif, Constituent-Based Top-Quark Tagging with the ATLAS Detector, 2022.

[43] M. Cacciari, G. P. Salam, and G. Soyez, The anti-kt jet clustering algorithm, Journal of High Energy Physics **2008**, 63 (2008).

[44] Y. Dokshitzer, G. Leder, S. Moretti, and B. Webber, Better jet clustering algorithms, Journal of High Energy Physics **1997**, 1 (1997).

[45] M. Wobisch and T. Wengler, Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering, (1999).

[46] S. D. Ellis and D. E. Soper, Successive combination jet algorithm for hadron collisions, Physical Review D **48**, 3160 (1993).

[47] Identification of hadronically-decaying top quarks using UFO jets with ATLAS in Run 2, (2021).

[48] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber, New clustering algorithm for multijet cross sections in e^+e^- annihilation, Physics Letters B **269**, 432 (1991).

[49] D. Krohn, J. Thaler, and L.-T. Wang, Jet trimming, Journal of High Energy Physics **2010**, 1 (2010).

[50] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft drop, Journal of High Energy Physics **2014**, 1 (2014).

[51] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, Recombination algorithms and jet substructure: Pruning as a tool for heavy particle searches, Physical Review D **81**, (2010).

[52] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, Jet substructure as a new Higgs-search channel at the large hadron collider, Physical Review Letters **100**, 242001 (2008).

[53] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, Towards an understanding of jet substructure, Journal of High Energy Physics **2013**, (2013).

[54] ATLAS Collaboration and others, Jet energy measurement with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7$ TeV, Arxiv Preprint Arxiv:1112.6426 (2011).

[55] ATLAS Collaboration and others, Jet energy measurement and its systematic uncertainty in proton–proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector, The European Physical Journal C **75**, (2015).

[56] I. W. Stewart, F. J. Tackmann, and W. J. Waalewijn, N-Jettiness: An Inclusive Event Shape to Veto Jets, Physical Review Letters **105**, (2010).

[57] ATLAS Collaboration and others, Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC, The European Physical Journal C **79**, (2019).

[58] Y. Rubner, C. Tomasi, and L. J. Guibas, *A Metric for Distributions with Applications to Image Databases*, in *Sixth International Conference on Computer Vision (IEEE Cat. No. 98ch36271)* (1998), pp. 59–66.

[59] P. Virtanen et al., SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods **17**, 261 (2020).

[60] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition **30**, 1145 (1997).

[61] F. Pedregosa et al., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research **12**, 2825 (2011).

[62] R. Munroe, XKCD 2270: Picking Bad Stocks, (2020).

[63] J. Pivarski, I. Osborne, I. Ifrim, H. Schreiner, A. Hollands, A. Biswas, P. Das, S. Roy Choudhury, N. Smith, and M. Goyal, Awkward Array, (2018).

[64] J. D. Hunter, Matplotlib: A 2D graphics environment, Computing in Science & Engineering **9**, 90 (2007).

[65] M. L. Waskom, seaborn: statistical data visualization, Journal of Open Source Software **6**, 3021 (2021).

[66] Performance of W/Z taggers using UFO jets in ATLAS, (2021).

[67] C. Bierlich et al., A comprehensive guide to the physics and usage of PYTHIA 8.3, Scipost Physics Codebases 8 (2022).

[68] R. D. Ball et al., Parton distributions with LHC data, Nuclear Physics B **867**, 244 (2013).

[69] ATLAS Pythia 8 tunes to 7 TeV data, 2014.

[70] S. Agostinelli et al., Geant4—a simulation toolkit, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **506**, 250 (2003).

[71] The Pythia 8 A3 tune description of ATLAS minimum bias and inelastic measurements incorporating the Donnachie-Landshoff diffractive model, 2016.

[72] ATLAS Collaboration, Public ATLAS Online Luminosity Plots for Run-2 of the LHC, (n.d.).

[73] P. F. Åkesson and E. Moyse, Event Data Model in ATLAS, (2005).

[74] ATLAS Collaboration, Athena, (2019).

Index of Figures

Index of Tables