Measuring b-quark jet structure at the LHC

PHYS4022P

Author Student ID: 2547473m
Supervisor: Dr. Andy Buckley

2024-03-28

Contents

Introduction

While the Standard Model remains highly successful in particle physics, theories beyond it, known as Beyond the Standard Model (BSM), often predict an abundance of b-quarks [1]. Thus, understanding the composition and behavior of these b-quarks is crucial, forming the foundational framework for a comprehensive understanding of particle physics.

Jets, characterized as collimated sprays of hadrons resulting from high-energy collisions [1], serve as a primary focus in this project. The primary source of bottom quarks in this project stems from the decay pattern of top and antitop quark pairs ($t\bar{t}$). This choice is deliberate, given its distinct decay channel of b-jets plus dileptonic [12], offering a complex yet distinguishable experiment.

This project delves into the in-depth measurement of bottom quark jet structures in two parts: simulation data and real-life ATLAS data. The analysis encompasses various key characteristics of these jets, including Les Houches Angularity (LHA), N-Subjettiness ratio (τ_{21}), Energy Correlation Functions (ECF), C2, and D2 correlations [1]. A significant emphasis lies in establishing solid selection criteria to identify valid b-jets. The "Tag and Probe" method will be implemented for the analysis of real data, evaluating the efficiency and purity of these events. The objective is to showcase the b-jet structures from collision events at the LHC, providing insights into particle interactions at high-energy scales.

The simulation employs Rivet and Pythia systems (C++) and is informed by prior ATLAS analyses, specifically "ATLAS_2022_I2152933" [10] and "ATLAS_2019_I1724098" [4]. Analysis code for real data is crafted anew using Python. All simulation models and analysis code are available on GitHub: (https://github.com/HowaiMak/ATLAS_2023_BJETS). Detailed instructions on implementations and startup guidelines are noted in the accompanying "README" file.

Monte-Carlo Collision Model

Monte Carlo (MC) event generation, fundamental in particle physics, utilizes iterative random sampling and statistical techniques to simulate complex collision events [1].

Our simulation employs the Rivet and Pythia systems. During our experimentation phase involving 200,000 events, optimal outcomes were achieved by sampling events with either 2 or 3 jets for ttbar events, termed the signal (SIG) mode. This ensures the presence of at least 2 jets, crucial for $t\bar{t}$ events, while accommodating random jet production at LHC with a third jet [12]. While this additional jet may influence reconstruction, it doesn't impact b-tagging accuracy. Conversely, the background (BKG) mode requires events to possess at least 1 jet, ensuring the presence of light and gluon jets for noise considerations. Mode toggling and decay pathway adjustments are facilitated using a separate C++ file, "ttbar-dilep.cmnd".

The "ATLAS_2023_BJETS.cc" serves as the primary analysis simulation model, allowing convenient exploration of diverse structures and filtering conditions. This includes constraining jet production within $|\eta| < 3.5$, with a slightly extended detection range of $|\eta| < 4.0$. Lepton production is limited within $|\eta| < 2.5$. The minimum p_T value for particles to be recognized as either a lepton or jet is set at $25GeV$. The code structure draws inspiration from two previous ATLAS analyses, "ATLAS_2022_I2152933" [10] and "ATLAS_2019_I1724098" [4].

Sources of Bottom Jets

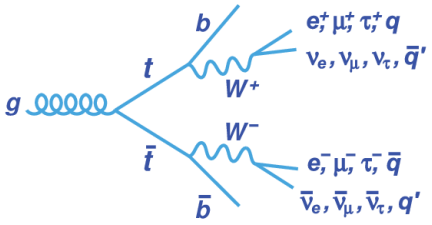


Figure 1: Top and anti-Top decay channel [2]

Bottom quarks (b quarks) can be produced through various radiational processes at the LHC, each with its own set of advantages and disadvantages:

- $t\bar{t}$ decay: The $t\bar{t}$ channel exhibits a significant incidence rate, characterized by a natural double b-jet purity, thus enabling the generation of purer samples for analysis[12]. This channel is notably prominent in decay processes observed at high-energy colliders such as the LHC[1]. Nevertheless, the presence of random accompanying jets can pose challenges in accurately identifying and distinguishing b-jets. The reconstruction process is further complicated by fragmentation and hadronisation phenomena.

- Higgs bosons: Higgs bosons are directly associated with bottom quarks, enabling investigations into b-quark interactions during Higgs decay processes [3]. They provide valuable insights into the properties of the Higgs boson and its interactions with other particles. However, their production rate is lower compared to ttbar events, and they may involve additional decay channels, complicating the identification of b-jets.

- Gluon jets: Gluon jets provide valuable insights into gluon splitting and fragmentation processes, offering a diverse range of energies and multiplicities for studying b-jet characteristics. However, they generally exhibit a lower occurrence rate of b-jets compared to ttbar events. Analysis and interpretation are complicated by background contamination from non-b-jets.

Among these, the top and anti-top $t\bar{t}$ decay channel stands out as a dominant means of probing bottom quarks at the LHC. Despite challenges posed by accompanying random jets, event simulation leveraging double b-jet purity within the $t\bar{t}$ channel enables efficient and targeted investigations into the structural characteristics of b-jets.

W^+W^- Boson Decay

The $t\bar{t}$ decay channel inherently produces two b-quarks and W^+, W^- bosons. However, diverse decay modes exist for the W bosons, with a 9% probability of decaying into a lepton pair (comprising electrons, muons, or tau in any two-part combination), a 45% likelihood of decay into one lepton plus a jet (originating from a light quark), and a 46% chance of complete hadronic decay (involving jets only) [9]. To exclusively attain a final state in the $t\bar{t}$ events featuring jets of bottom quarks (dijets case), it is imperative to configure the simulation such that the W bosons are constrained to decay solely into dileptonic pairs. This framework establishes a nuanced yet distinguishable context for studying bottom quarks within a sufficiently complex yet distinguishable channel.

Background Contamination

Several decay channels, such as double light jet emissions from Quantum Chromodynamics (QCD) and Weak Boson Fusion (WBF), among others, can produce excessive light and gluon jets, contributing to background noise [1]. However, these channels are intricate, involving complex particle interactions that are not thoroughly investigated and pose challenges for accurate modeling and simulation. In our model, the light and gluon jets are generated through the decay of double bosons W^+, W^-, Z, or the jet plus single Z boson decay channel. This focused selection ensures a more manageable and targeted investigation of the desired b-jet characteristics.

Simulation Results

Bottom Jet Identification

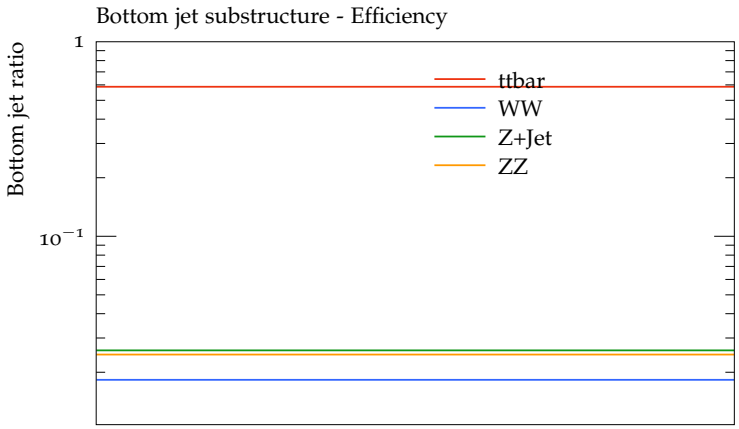


Figure 2: Bottom Jet Natural Occurrence Rate

Particle collision experiments encounter challenges in accurately identifying b-jets and implementing effective quark flavor tagging, primarily due to mechanical constraints and the complex experimental environment, including limitations in the detection angle $|\eta|$ [1]. These factors contribute to suboptimal jet reconstruction from final states. Computational simulations offer a crucial solution to address these challenges by providing a controlled environment with perfect flavor tagging. Within this controlled setting, parameters can be adjusted, enabling the exploration of scenarios that may present difficulties in replication under real experimental conditions.

Figure 2 illustrates the natural occurrence rate of b-jets in the collision model. Within the $t\bar{t}$ decay channel, 58.7% of the total jets are identified as b-jets. In contrast, within the background, 1.83% originate from the W^+W^- channel, 2.59% from Z+Jet, and 2.47% from ZZ. These results align with expectations, demonstrating a higher detection rate of bottom jets in $t\bar{t}$ events compared to background events, which typically involve numerous light and gluon jets.

Jet Variables

Les Houches Angularity (LHA)

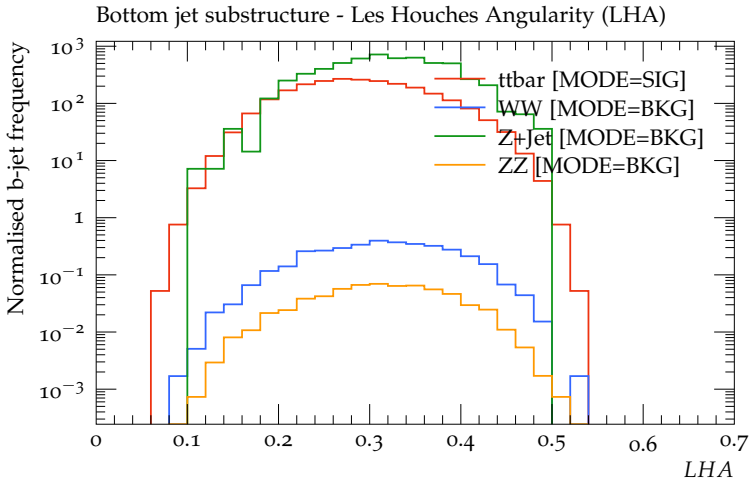


Figure 3: Scaled angular and momentum distribution of jet

The Les Houches Angularity (LHA) is a specific case within the family of observables known as "generalised angularities" [1]. This family of angularities is defined by two parameters, κ and β, both of which are non-negative.

$$\lambda_\beta^\kappa = \sum_i z_i^\kappa \theta_i^\beta = \sum_i \left(\frac{(p_T)_i}{(p_T)_{\text{jet}}} \right)^\kappa \left(\frac{\Delta R_i}{R} \right)^\beta, \tag{1}$$

where the sum runs over the constituents i of the jet. Here, $z_i \in [0, 1]$ is the fraction of the transverse momentum p_T of the i-th constituent relative to the total p_T of the jet, and $\theta_i = \frac{\Delta R_i}{R}$ represents the angular separation of the i-th constituent from the jet axis, normalized by the jet radius R (typically set to $R = 1$).

The specific case of $\lambda_{0.5}^1$, also known as LHA, is notable for its sensitivity to both the momentum and angular structure of the jet. Setting $\kappa = 1$ ensures that the observable is infrared and collinear (IRC) safe, meaning it remains stable in the presence of soft emissions and collinear splittings, making it suitable for perturbative calculations [5]. The choice of $\beta = 0.5$ introduces a moderate angular dependence, emphasizing both momentum and angular scaling, though with less sensitivity to angular structure than momentum [1].

Figure 3 shows the normalized distribution of LHA values for b-jets across different decay channels in our simulation, with the LHA value scaled by a factor of 10^3. A clear increase in the number of detected b-jets is observed, particularly in the $t\bar{t}$ and Z+jet channels, consistent with the well-known tendency of these processes to produce bottom quarks.

The peak LHA value across all channels is approximately 320, suggesting that b-jets tend to have LHA values centered around this number. The distribution of LHA values spans from approximately 100 to 500, providing a range of LHA values characteristic of b-jets in these channels.

N-subjettiness Ratio (τ_{21})

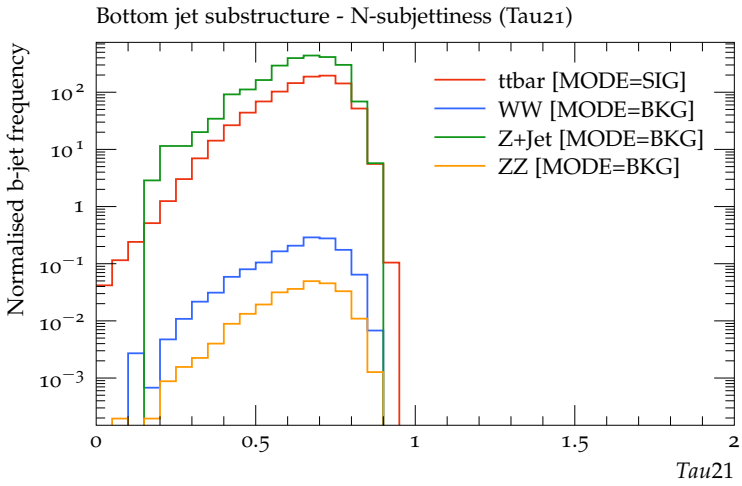


Figure 4: 2-prong bjet against 1-prong bjet ratio frequency

The N-subjettiness (τ_N) is a quantitative measure of the degree to which a jet can be represented as consisting of N subjets. It is computed by hypothesizing N subjets from a reconstructed jet using a jet algorithm and is given by the formula:

$$\tau_N = \frac{1}{\sum_i (p_T)_i R_0} \sum_i (p_T)_i \min(\Delta R_{1,i}, \Delta R_{2,i}, \ldots, \Delta R_{N,i}), \tag{2}$$

where the index i iterates over the constituent particles within a given jet, $(p_T)_i$ denotes their transverse momenta, and $\Delta R_{J,i} = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ is the distance in the rapidity-azimuth plane between a candidate subjet J and the constituent particle i. The parameter R_0 represents the characteristic jet radius (commonly set to $R = 1$) used in the original jet clustering algorithm [11].

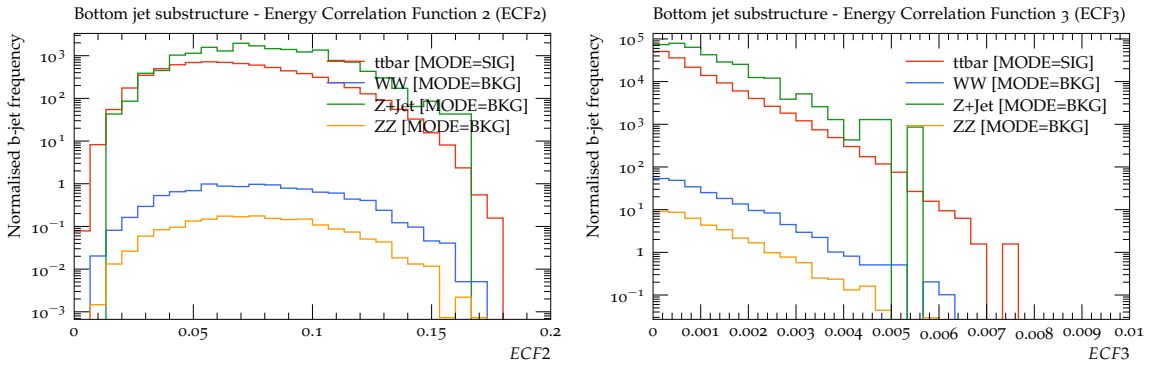
The N-subjettiness ratio τ_{NM} becomes particularly powerful when employed similarly to a likelihood-ratio discriminant between different hypotheses. It is defined as the ratio of N-subjettiness values for different hypothesized values of N:

$$\tau_{nm} = \frac{\tau_n}{\tau_m}. \tag{3}$$

In our simulation, we focus on the specific ratio $\tau_{21} = \frac{\tau_2}{\tau_1}$ to evaluate whether the events exhibit a 1-jet-like or 2-jet-like configuration. Figure 4 presents the distribution of τ_{21} values across all events. A value of $\tau_N = 0$ indicates a clear N-jet-like event; thus, a higher τ_{21} suggests a greater likelihood of the event resembling a 1-subjet-like structure, whereas a lower value indicates a preference for a 2-subjet-like configuration.

Across all decay channels, the τ_{21} value does not exceed 1.0, suggesting a strong tendency towards 2-subjettiness. With a peak at approximately 0.67, this indicates that events are roughly 30% more likely to be characterized by 2 jets. Notably, a distinct 2-subjettiness feature, indicated by $\tau_{21} = 0$, is observed in the $t\bar{t}$ channel, which is absent in the background channels. This distinction can be attributed to the enhanced production of light quarks and gluons in the background processes.

Energy Correlation Coefficient (ECF)



(a) ECF2

(b) ECF3

Figure 5: Likely-hood of 2 or 3 particles event comparison

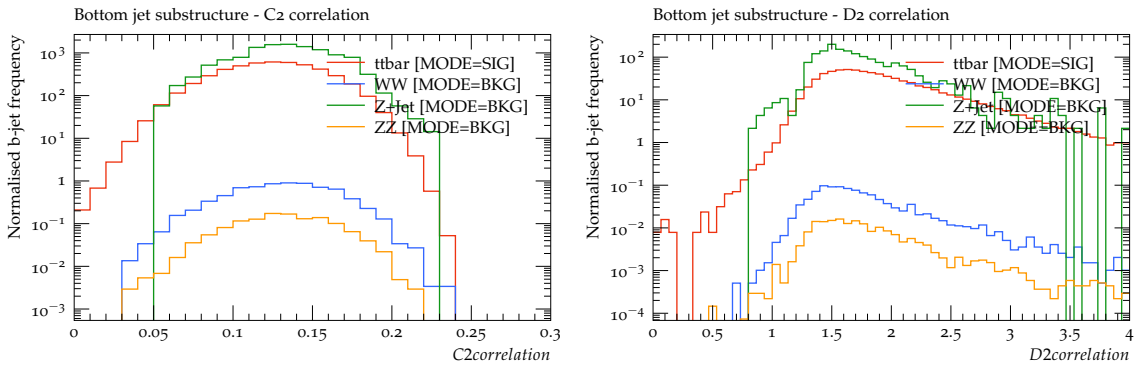
The energy correlation function (ECF) quantifies the correlations between the transverse momentum products of particle pairs within a jet J. This function sums the products of the transverse momenta of each pair of particles, denoted by indices i and j, and is weighted by the β-dependent distance ΔR in the rapidity-azimuth plane, as previously defined. Unlike several prior jet substructure methodologies, these correlation functions do not require the explicit identification of subjet regions [8].

$$ECF(2,\beta) = \sum_{i<j\in J} (p_T)_i (p_T)_j (\Delta R_{ij})^\beta \tag{4}$$

$$ECF(3,\beta) = \sum_{i<j<k\in J} (p_T)_i (p_T)_j (p_T)_k (\Delta R_{ij}\Delta R_{ik}\Delta R_{jk})^\beta \tag{5}$$

Figure 5 displays the Energy Correlation function for both 2-particle and 3-particle likelihoods of bottom quark events. Notably, the $ECF(2)$ values exhibit a relatively uniform distribution across their range, while the $ECF(3)$ values show a peak at 0, gradually decreasing as the value increases. This suggests that the majority of events are predominantly 2-particle configurations, with approximately 10^1 events at the midpoint of $ECF(3)$ potentially involving 3 particles. Additionally, the $t\bar{t}$ channel demonstrates a higher propensity for generating 3-particle events compared to the background channels, as indicated by its $ECF(3)$ values extending to 0.0075.

C2 & D2 correlation



(a) C2 correlation

(b) D2 correlation

Figure 6: Subjet-number sensitivity

Equations 4 and 5 illustrate that if a jet contains fewer than N particles, the corresponding Energy Correlation Function (ECF) returns a value of 0. Therefore, in a system with N subjets, the magnitude of $ECF(N+1,\beta)$ is expected to be notably smaller than that of $ECF(N,\beta)$ [1]. This observation prompts an investigation into the energy correlation function ratio, specifically the C_2 and D_2 correlation.

$$C_2 = \frac{ECF(3) \times ECF(1)}{(ECF(2))^2} \tag{6}$$

$$D_2 = \frac{ECF(3) \times (ECF(1))^3}{(ECF(2))^3} \tag{7}$$

A key distinction between N-subjettiness and the energy correlation function ratio is that C_2 and D_2 do not require a separate procedure (such as minimization, as seen in Equation 2) for their determination. This feature simplifies the analysis of subjet structures when employing reconstruction algorithms.

The values of C_2 in Figure 6a are scaled down by a factor of 10^2, while the values of D_2 in Figure 6b are scaled up by a factor of 10^1. Both ratios effectively identify 2-body structures [7]; their values increase as the number of subjets within the large-R jet rises. Notably, the C_2 value peaks around 125 and maintains a relatively uniform distribution in the range from 0 to 250. In contrast, the D_2 value exhibits a rapid increase within the range from 0 to 0.1, followed by a gradual decrease up to 0.4, where substantial noise is observed, complicating the data interpretation.

Real Data from ATLAS at LHC

Event Selections

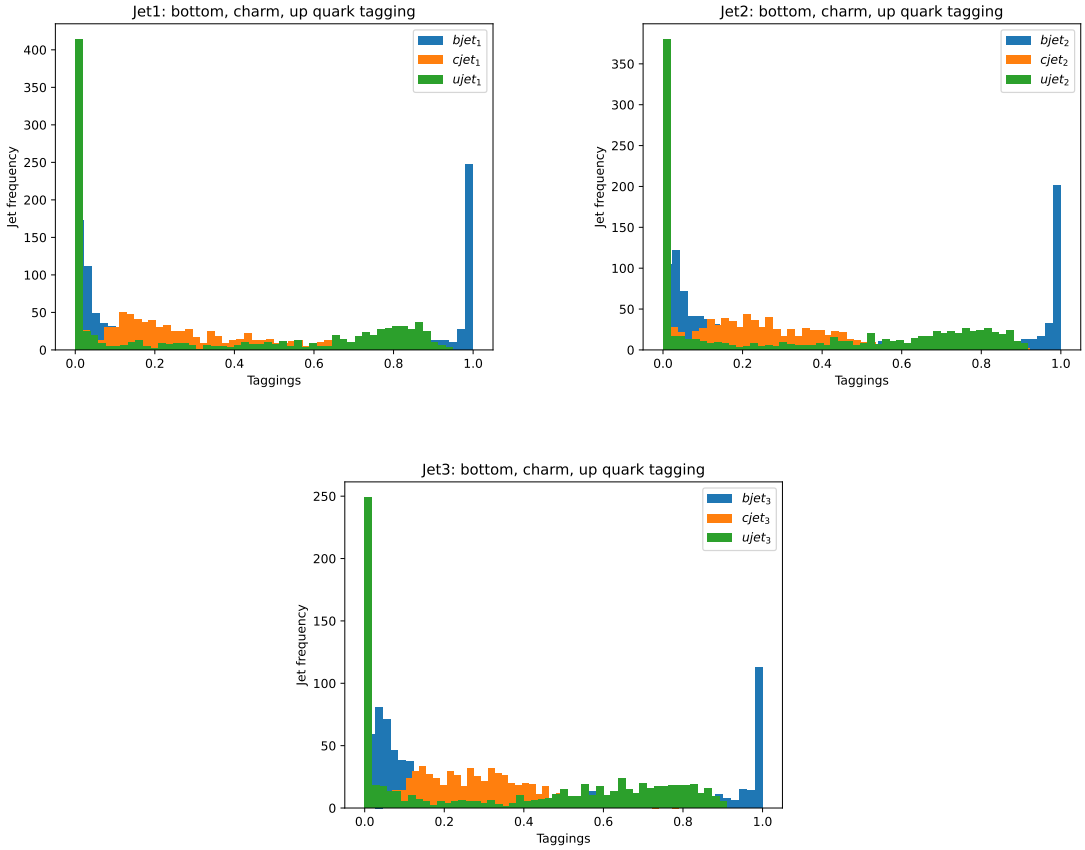


Figure 7: Probabilities of jets being a bottom, charm, or up quark in real collision events

Simulation studies have provided valuable insights into the characteristic features of b-jets. Leveraging this knowledge, the tagging algorithm can be applied to real data to estimate the probabilities

associated with each quark type. To adhere to the constraints established in our initial investigations, which focus on studying b-quarks via the $t\bar{t}$ channel, we select real collision events that contain one electron and one muon, each with a transverse momentum (p_T) of 20,000 GeV. Furthermore, we consider jets occurring in quantities of 2 or 3, with $p_T > 30,000$ GeV.

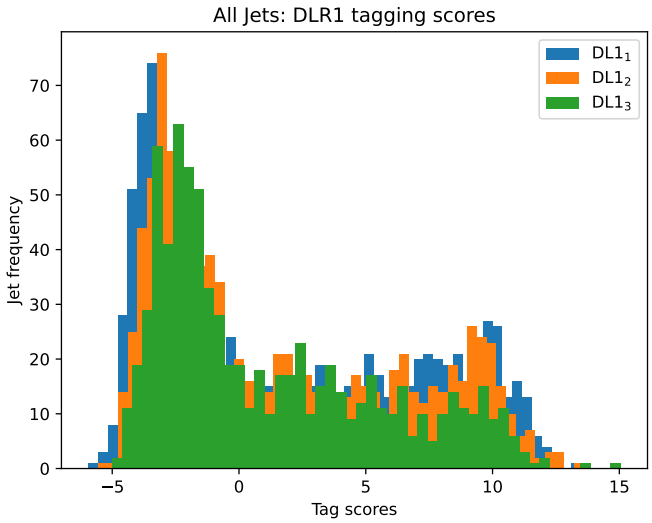


Figure 8: DL1 tag scores for all jets

In particle detection, it is crucial to strike a balance between efficiency and purity. High efficiency can be easily achieved by classifying all detections as valid (i.e., considering every particle as a bottom quark). However, this approach is impractical, particularly in collider physics, where jet structures consist of various particles.

The DL1 score quantifies the probability of detecting a specific quark type within a jet, serving as a metric for evaluating detection efficiency standards for the desired quark. Equation 8 presents the DL1 score for b-tagging [6].

$$DL1_b = \ln\left(\frac{p_b}{f_c p_c + (1 - f_c) p_u}\right) \tag{8}$$

In this equation, p_b, p_c, and p_u represent the tagging probabilities for bottom, charm, and up quarks, respectively. The parameter f_c denotes the fraction of charm quarks within each event, typically defined as a fixed value; in our experiment, we set $f_c = 0.03$.

Figure 7 illustrates the distribution of jets against their flavor tagging probabilities (p_b, p_c, p_u). Jet1 corresponds to the leading p_T in each event, followed by Jet2 and Jet3. It is evident that the b-taggers display relatively high certainty, with b-jet identification peaking at values around 1.0 or 0.0, indicating 100% and 0% tagging probabilities, respectively. The identification of up-jets predominantly clusters towards the lower end of the scale, reflecting the reduced expectation of light jets after the initial selections. The production of charm jets is minimal, as indicated by the previously

established f_c value, rendering them less significant in our analysis.

After processing all p values through Equation 8, Figure 8 displays the resulting DL1 scores for the jets. The overall efficiency for b-tagging is set at approximately 77%, corresponding to a DL1 score of 2.23 [6]. In our investigation, we aim for a DL1 score greater than 2.96, which corresponds to efficiencies of 70% or lower. This choice of jet events with lower efficiency is strategic, as it allows us to explore the purity of bottom jet structure measurements through a specialized method known as Tag & Probe, thereby mitigating tagging bias.

Tag and Probe

The Tag and Probe method (T&P) is utilized as a technique to mitigate bias in particle detection. This strategy leverages the properties of double b-jets from the $t\bar{t}$ channel to investigate the tagging bias of bottom quarks [1]. The procedure is as follows:

- **Tag:** Define a well-identified bottom quark, referred to as the "b-tag," which successfully passes the DL1 score threshold.
- **Probe:** Instead of directly measuring the "b-tag," assess the substructure of its paired b-jet from the $t\bar{t}$ decay channel.

The analysis of bottom quarks will encompass both the "b-tag" and its accompanying jet, which we will refer to as the direct and T&P methods, respectively. To compare variations between these methods, the χ^2 test will be employed, and the data will be normalized using equal binning. A χ^2 value of 0 indicates perfect symmetry between the measurements.

Bottom Jet Structures

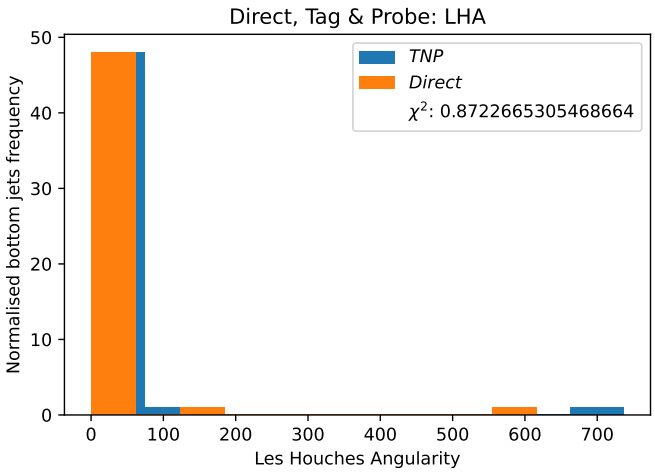


Figure 9: Les Houches Angularity

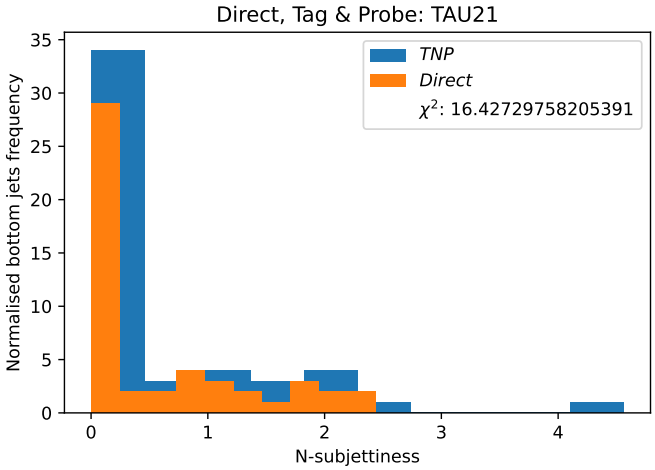


Figure 10: N-subjettiness ratio

Figure 9 illustrates the measured Linear Hadron Activity (LHA) values from the selected b-jets that passed the tagging criteria. The two methods demonstrate close agreement, as indicated by the calculated $\chi^2 = 0.87$. The LHA distribution is more concentrated within the range of 0 to 100, in contrast to the broader distribution predicted in Figure 3. This observation suggests that the actual distribution of LHA values deviates from the model's expectations, favoring more tightly clustered measurements in this region.

Figure 10 presents the measured τ_{21} values from the selected b-jets. While the peak of the distributions is observed between 0 and 1, consistent with the simulation results shown in Figure 4, there is a notable discrepancy between the two methods, as indicated by $\chi^2 = 16.43$. Specifically, the T&P method identifies a population of high τ_{21} b-jets that had been previously overlooked, suggesting that these jets are approximately four times more likely to exhibit a 1-prong substructure than a 2-prong substructure in this region.

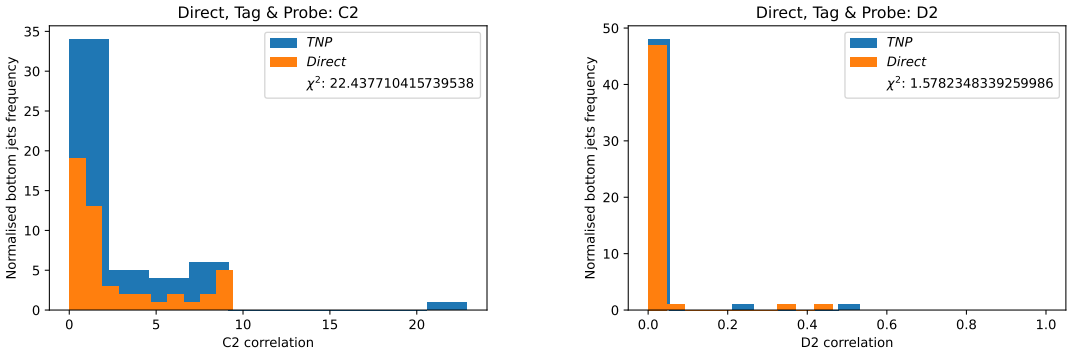


Figure 11: C2 and D2 correlation

Comparing Figure 11 with Figure 6, a noticeable shift towards lower values is observed in both energy correlation ratios. This shift suggests that the model has overestimated the presence of substructure in the b-jets, predicting a higher number of subjets than what was observed in the experimental data [7]. This discrepancy indicates an overestimation of the complexity of the b-jet structure by the model.

Conclusion

In conclusion, our investigation into b-jet substructures through computational simulations and real ATLAS data analysis provides valuable insights into jet tagging and flavor identification in high-energy particle collisions. Simulations offer a robust framework for studying b-jet behavior in a controlled environment, enabling ideal tagging and enhancing our understanding of key jet observables such as Linear Hadron Activity (LHA), N-subjettiness (τ_{21}), and energy correlation functions (ECF). These variables are critical for differentiating b-jets from light and gluon jets, particularly in complex decay channels like $t\bar{t}$.

While real data from the ATLAS experiment generally aligns with simulation results, some notable discrepancies are observed, particularly in the τ_{21} distributions and the energy correlation ratios ($C2$, $D2$). These differences suggest that the simulations may oversimplify certain aspects of jet substructure, such as the frequency of 1-prong versus 2-prong configurations and the number of subjets in b-jets. These findings highlight the inherent challenges in fully capturing the complexity of real collision environments. The Tag-and-Probe (T&P) method proved effective in reducing bias and improving the tagging accuracy of b-jets, especially in regions where traditional methods tend to overlook high τ_{21} jets.

Looking forward, there is room for improvement in both the simulation and analysis approaches. Refining models to better account for the observed deviations in subjet structures and energy correlations will help bridge the gap between simulated and experimental results. Additionally, incorporating more sophisticated machine learning techniques or data-driven corrections could further enhance the accuracy of b-jet identification and tagging performance.

Overall, this study underscores the indispensable role of simulations in preparing for collider data analysis, while also emphasizing the need for ongoing model refinement to reflect the complexity of real-world collisions. Continuous improvements in this area are crucial for enhancing the precision of b-jet identification, ultimately advancing our understanding of quark flavor tagging and jet substructure in high-energy physics.

Bibliography

[1] Andy Buckley, Christopher White, and Martin White. *Practical Collider Physics*. 2053-2563. IOP Publishing, 2021. ISBN: 978-0-7503-2444-1. DOI: 10.1088/978-0-7503-2444-1. URL: https://dx.doi.org/10.1088/978-0-7503-2444-1.

[2] The CDF et al. "Combination of the top-quark mass measurements from the Tevatron collider". In: *Physical Review D* 86 (July 2012). DOI: 10.1103/PhysRevD.86.092003.

[3] ATLAS Collaboration. *Higgs boson observed decaying to b quarks*. URL: https://atlas.cern/updates/briefing/higgs-observed-decaying-b-quarks.

[4] Christian Gutschow Deepak Kar Amal Vaidya. *Jet substructure at 13 TeV Experiment: ATLAS (LHC)*. 2019. URL: https://rivet.hepforge.org/analyses/ATLAS_2019_I1724098.html.

[5] Philippe Gras et al. "Systematics of quark/gluon tagging". In: *Journal of High Energy Physics* 2017.7 (July 2017). ISSN: 1029-8479. DOI: 10.1007/jhep07(2017)091. URL: http://dx.doi.org/10.1007/JHEP07(2017)091.

[6] Marie Lanfermann. "Deep Learning in Flavour Tagging at the ATLAS experiment". In: Jan. 2018, p. 764. DOI: 10.22323/1.314.0764.

[7] Andrew J. Larkoski, Ian Moult, and Duff Neill. "Power counting to better jet observables". In: *Journal of High Energy Physics* 2014.12 (Dec. 2014). ISSN: 1029-8479. DOI: 10.1007/jhep12(2014)009. URL: http://dx.doi.org/10.1007/JHEP12(2014)009.

[8] Andrew J. Larkoski, Gavin P. Salam, and Jesse Thaler. "Energy correlation functions for jet substructure". In: *Journal of High Energy Physics* 2013.6 (June 2013). ISSN: 1029-8479. DOI: 10.1007/jhep06(2013)108. URL: http://dx.doi.org/10.1007/JHEP06(2013)108.

[9] Izaak Neutelings. *Branching fraction matrix of pair decays*. 2021. URL: https://tikz.net/sm_decay_matrix/.

[10] Andrea Helen Knue Shayma Wahdan Dominic Hirschbuhl. *Measurement of observables sensitive to colour reconnection in ttbar dileptonic emu channel at 13 TeV Experiment: ATLAS (LHC)*. 2022. URL: https://rivet.hepforge.org/analyses/ATLAS_2022_I2152933.html.

[11] Jesse Thaler and Ken Van Tilburg. "Identifying boosted objects with N-subjettiness". In: *Journal of High Energy Physics* 2011.3 (Mar. 2011). ISSN: 1029-8479. DOI: 10.1007/jhep03(2011)015. URL: http://dx.doi.org/10.1007/JHEP03(2011)015.

[12] Timothee Theveneaux-Pelzer. "Measurement of ttbar with additional jets with the ATLAS detector". In: (2019). URL: https://cds.cern.ch/record/2687376.