Hadron-level quark/gluon tagging for ATLAS

Gareth Smith

September 12, 2018

1 Introduction

The ATLAS experiment at the LHC is a general-purpose particle physics detector used to study pp collisions at $\sqrt{s} = 13$TeV. The two main types of particles produced in the initial hard scatter are quarks and gluons, collectively called "partons". Since partons cannot exist in an isolated state, they will quickly decay into stable hadrons in the showering and hadronization process. This results collimated sprays of hadrons known as jets. One of the challenges faced by the ATLAS collaboration is to understand the type of particle which initiated a jet, given its signature in the trackers and calorimeters.

A jet in ATLAS simulations has multiple levels. A jet is called "parton-level" before it hadronizes and "hadron-level" after it hadronizes. The jet is then simulated as it enters the ATLAS calorimeters, and the energy depositions are grouped into a "reco-level" jet. We are interested in the detector response, or jet energy scale (JES), which is defined as the ratio of reco jet transverse momentum to truth jet transverse momentum $p_{\mathrm{T}}^{\mathrm{reco}}/p_{\mathrm{T}}^{\mathrm{truth}}$. This quantity measures the fraction of the energy of a jet which is captured in the calorimeters. There have been many previous efforts to understand the JES [1, 2].

It has been observed that jets initiated by different partons have different average responses. This is known as the JES flavour uncertainty. To properly correct for this effect, it is useful to be able to identify each jet with the parton which initiated it. At hadron-level, jets initiated by gluons and by light quarks (u, d, s) can be distinguished only by properties such as jet width and number of constituent hadrons; quark jets are typically narrower and have fewer constituents than gluon jets. Identifying quark and gluon jets based on their properties at hadron-level is the topic for my project.

Jets are currently labelled by the highest p_{T} parton in their volume at parton-level. However, this is not ideal for a number of reasons. Partons are not stable, and so their properties are unphysical. There are a number of different Monte Carlo (MC) generators which are used to simulate the parton showering and hadronization, and the way they handle parton-level information can be very different. Relying on unphysical parton-level information will introduce a large dependence on the MC generator used.

The goal of this project was to classify jets based on physical hadron-level topology, without using unphysical parton-level information. This classifier should take jet properties at hadron-level as input, and should output a value from -1 to $+1$ depending on how "quark-like" or "gluon-like" the hadron jet is. The relationship between the classifier and the JES was then studied.

This project is described in further detail in a full length report [3], which includes more technical information and results, and additional studies performed for this project.

2 Input variables

Following a literature search and discussions with Benjamin Nachman and Dag Gillberg, a total of 13 hadron level variables were identified as being potentially good quark/gluon discriminants.

1. n_{const}: The number of constituents (stable hadrons) within a jet volume.
2. m/p_{T}: The total mass of all jet constituents divided by the total jet p_{T}.
3. Width: The width of the jet.
$$w = \frac{\sum_i p_{\mathrm{T},i} \times \Delta R(i, \mathrm{jet})}{\sum_i p_{\mathrm{T},i}}$$
4. $p_{\mathrm{T\ hardest}}/p_{\mathrm{T}}$: The fraction of jet p_{T} carried by the highest p_{T} constituent in the jet.

5. $p_{\mathrm{T\ charged}}/p_{\mathrm{T}}$: The fraction of jet p_{T} carried by charged constituents.

6. $p_{\mathrm{T\ photons}}/p_{\mathrm{T}}$: The fraction of jet p_{T} carried by photons. This includes photons decaying from unstable hadrons, such as $\pi^0 \to \gamma\gamma$.

7. Charge: The total charge of the jet, weighted by $p_{\mathrm{T}}^{0.5}$.

$$q = \frac{\sum_i q_i \times p_{\mathrm{T},i}^{0.5}}{\sum_i p_{\mathrm{T},i}^{0.5}}$$

8. $N_{\mathrm{const}}^{\mathrm{eff}}$: An effective number of constituents.

$$N_{\mathrm{const}}^{\mathrm{eff}} = \frac{(\sum_i p_{\mathrm{T},i})^2}{\sum_i p_{\mathrm{T},i}^2}$$

$N_{\mathrm{const}}^{\mathrm{eff}} \to 1$ when all p_{T} is carried by one constituent, $N_{\mathrm{const}}^{\mathrm{eff}} \to n_{\mathrm{const}}$ when p_{T} is split equally among constituents.

9. $p_{\mathrm{T}}D$: The jet energy sharing value used by CMS.

$$p_{\mathrm{T}}D = \frac{\sqrt{\sum_i p_{\mathrm{T},i}^2}}{\sum_i p_{\mathrm{T},i}}$$

$p_{\mathrm{T}}D \to 1$ when all p_{T} is carried by one constituent, $p_{\mathrm{T}}D \to 0$ when p_{T} is split among infinitely many constituents [4]. These two variables are related by $N_{\mathrm{const}}^{\mathrm{eff}} = (p_{\mathrm{T}}D)^{-2}$.

10. $C(\beta)$: Energy-energy-correlation (EEC) angularity [5].

$$C(\beta) = \frac{\sum_i \sum_j p_{\mathrm{T},i} \times p_{\mathrm{T},j} \times (\Delta R(i,j))^\beta}{(\sum_i p_{\mathrm{T},i})^2}$$

Three different β values were used: 0.2, 1.0, and 2.0.

11. N^{90}: The minimum number of constituents which carry 90% of the jet p_{T}.

Classifiers were trained on two different sets of input variables. One classifier was trained on the full set of 13 input variables. A second classifier was trained on a subset of 8 input variables. The selection of these 8 variables will be described in the following section.

3 Input variable usefulness

Three different metrics were used to gauge the usefulness of each input variable.

The first is the separation of the quark and gluon input variable distributions. Given two classes of data, called signal and background, we define the separation of their distributions $\hat{y}_S$ and $\hat{y}_B$ in a variable y as:

$$\langle S^2 \rangle = \frac{1}{2} \int \frac{(\hat{y}_S(y) - \hat{y}_B(y))^2}{\hat{y}_S(y) + \hat{y}_B(y)} dy$$

This separation takes a value of 0 when the two distributions are identical, and 1 when there is no overlap in the distributions.

In our case, quark jets were chosen to be the signal, and gluon jets to be the background. The separation of each input variable is shown in Figure 1.

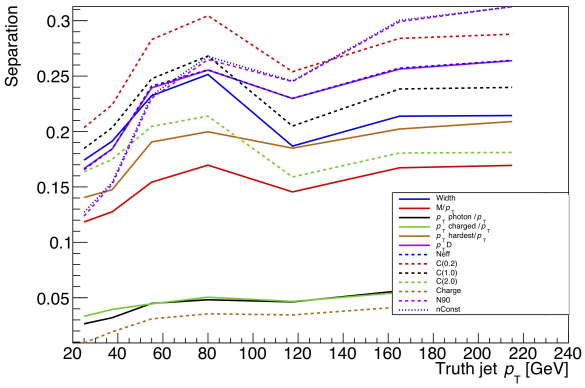

Figure 1: Quark/gluon separation of the 13 input variables.

The second metric which was examined is the signal efficiency at 80% background rejection. To find this, a cut is made on the input variable and only events to the left (or sometimes right) of this cut are kept. The cut is made at the value such that 80% of the background distribution is rejected. The same cut is then made on the signal distribution, and the fraction of events which pass the cut is called the signal efficiency. The signal efficiency at 80% background rejection of each input variable is shown in Figure 2.

Using these two metrics, we see that all variables grow in power at higher p_{T}. The weakest variables are $p_{\mathrm{T\ photons}}/p_{\mathrm{T}}$, $p_{\mathrm{T\ charged}}/p_{\mathrm{T}}$, and charge, while the strongest variables are $C(0.2)$, n_{const}, N^{90}, and

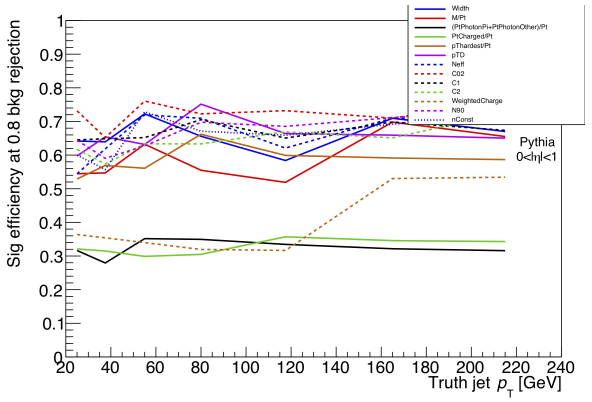

Figure 2: Variable signal efficiency with 80% background rejection.

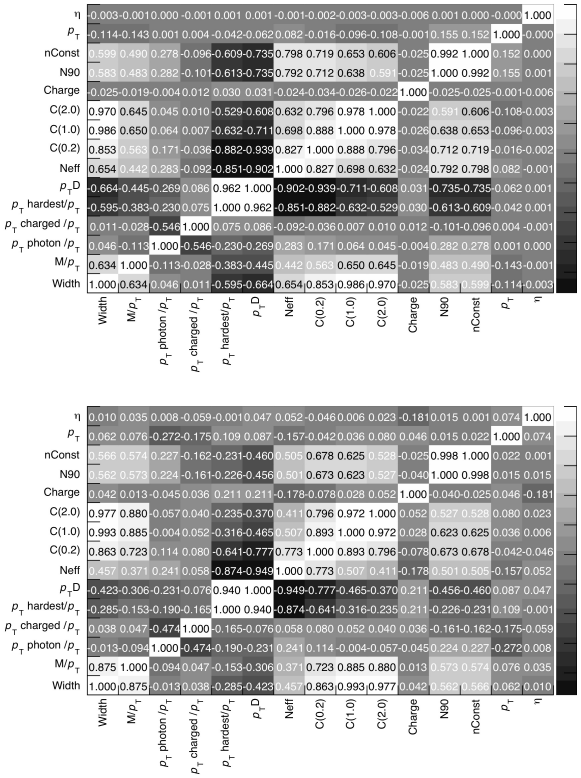

Figure 3: Input variable linear correlation coefficients in Pythia at $20 < p_T < 30$ and $190 < p_T < 240$ respectively.

$N^{\text{eff}}_{\text{const}}$. It also shows that $N^{\text{eff}}_{\text{const}}$ and $p_{\text{T}}D$ are redundant as expected, and also that n_{const} and N^{90} are redundant. There is also some systematic effect at work which is producing a dip at 120 GeV for most of the variables. This effect is not currently understood.

Finally, the correlation between each of these input variables was also studied. Consider multiple copies of the strongest variable; they would all have high separation and signal efficiency, but would not give any new information. This is reflected by the linear correlation factors. The linear correlation factors between the input variables, as well as the kinematic variables p_{T} and η, are given in Figure 3.

A second classifier was trained on the 8 input variables which were deemed to be the most important. The variables m/p_{T}, $p_{\text{T}}D$, N^{90}, $C(1.0)$, and $C(2.0)$ were chosen to be dropped as they are highly correlated with other variables, and have the lower separation and signal efficiency than the variables they are correlated with.

4 Classifier training

A classifier is a function which takes values for each of the input variables and returns a value between -1 and $+1$. It is trained using two classes of data, signal and background. An output close to -1 means an event is very similar to a background (gluon) event, while an output close to $+1$ means the event is very similar to a signal (quark) event.

As well as training classifiers on two different sets of input variables, two different MVA methods were also used, leading to a total of four classifiers being trained. A separate classifier was trained for each p_{T} bin. The bins used are 20-30, 30-45, 45-65, 65-95, 95-140, 140-190, and 190-240 GeV. In the future, these classifiers could be merged together in a continuous fashion to create a classifier which works across any p_{T}.

Dijet samples from Pythia, Herwig, and Sherpa were used for training. The difference between generators was studied but will not be addressed here. Only the leading and subleading jets in a central η region were used. On average about 60000 background events and 15000 signal events were used for training.

The two MVA methods used were Boosted Decision Trees (BDT) and Fisher discriminants. A BDT is composed of a series of nodes in a tree-like pattern. At each node, the data is split into two groups according to a cut on the input variable which offers the highest separation. Each group is passed on to a new node. The ending nodes, or

leaves, are then labelled as either signal or background. A Fisher discriminant is a lot simpler; it is simply a linear combination of all input variables, where coefficients are determined by training:

$$y_F(i) = F_0 + \sum_{k=1}^{n_{\text{var}}} F_k x_k(i)$$

5 Classifier power

The classifiers are trained such that they will maximize the separation between the output distributions for signal and background. As such, the power of a classifier can be judged in the same manner as the power of a variable; namely, using separation values and signal efficiency at fixed background rejection. The separation of the classifiers is shown if Figure 4 and the signal efficiency at 80% background rejection is shown in Figure 5.

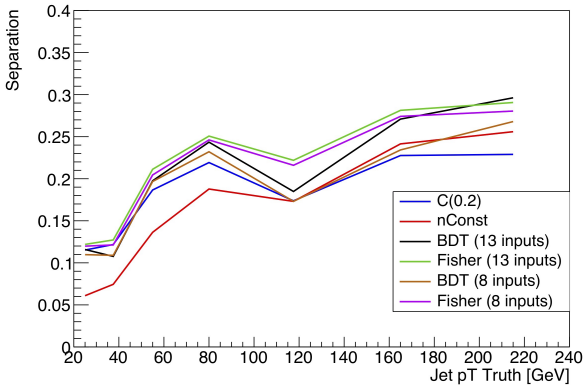

Figure 4: Separation of four MVA classifiers trained on Pythia.

Firstly, despite being simpler, the Fisher classifiers almost universally outperform the BDTs. This is unexpected and could be explained by insufficient training data.

Secondly, it can be seen that by excluding the five weakest and most correlated variables, some power is lost in the classifier. Further study should be performed into which of the input variables can be dropped with no loss of classifying power.

Finally, the classifiers are not so much more powerful than the strongest input variables, $C(0.2)$ and n_{const}. This suggests that using a classifier may not be necessary; perhaps a simple cut on $C(0.2)$ or n_{const} could be used instead as a hadron-level quark/gluon classifier.

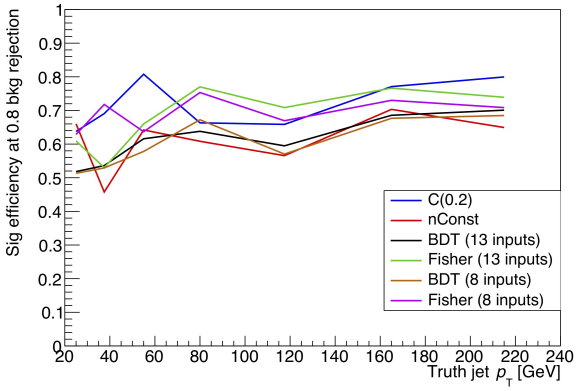

Figure 5: Signal efficiency of four MVA classifiers trained on Pythia at 80% background rejection.

6 Relationship with JES response

The response $R = p_{\text{T reco}}/p_{\text{T truth}}$ of quark jets is significantly higher than that of gluon jets, and understanding this discrepancy is one of the motivations for this project.

Instead of a binary label of "quark" or "gluon", a classifier gives each jet a continuous value from $+1$ ("quark-like") to -1 ("gluon-like"). Thus it is reasonable to hope that the classifier output value should be monotonically related to the average response. Figure 6 shows that this is indeed the case. The relation is in fact very linear, which gives a good indication that the "quark-gluon axis" defined by the classifiers is the same axis in our variable space which determines the response offset.

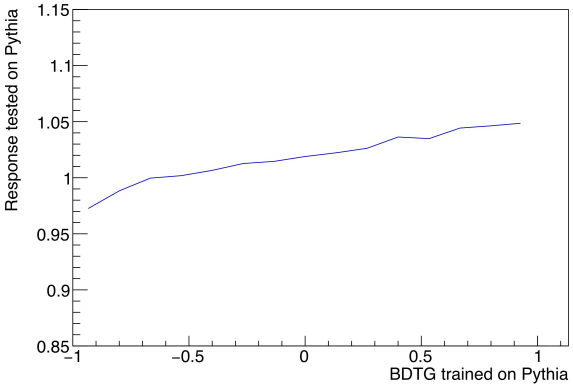

Figure 6: Average response of jets with with a certain classifier output.

Another way of seeing this is to split the data into four quartiles of classifier output, such that the first quartile represents the most "gluon-like" events, and the fourth quartile represents the most "quark-like" events. The response of these two data classes are plotted in Figure 7, alongside the response of the pure quark and gluon parton-labelled classes.

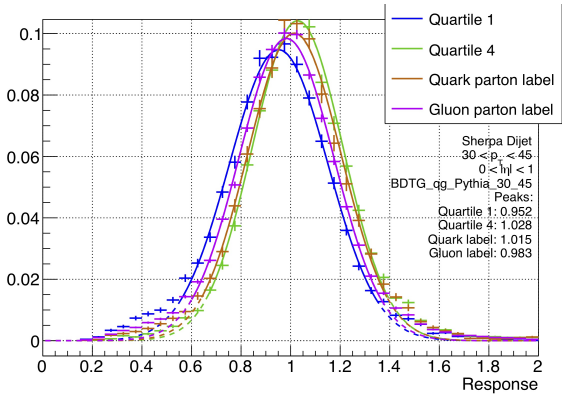

Figure 7: Response of jets with quark/gluon parton label and quark/gluon-like classifier output.

It is interesting that the average response for jets in the most gluon-like classifier quartile is lower than for jets with the gluon parton label, and the average response for jets in the most quark-like classifier quartile is higher than for jets with the quark parton label. This means that the classifier is doing a good job at highlighting the response difference between quarks and gluons.

7 Further work

Some further studies were also performed on this project which are not discussed here. These include insight into the generator dependance of the classifiers, by testing and training using data from different MC generators. Preliminary results indicate that this method is not able to fully remove the dependance on the MC generator used. Some initial investigations into other methods which could avoid this generator dependance were also carried out. These include using an MVA regression directly on the JES response, and using unsupervised learning methods such as CWoLa [6] and topic modelling[7]. The results of these studies, and some more technical details of this project, are described in the full report [3].

8 Conclusions

For the purpose of studying the JES flavour uncertainty, it is useful to have a method of tagging jets as quark-initiated or gluon-initiated at hadron-level. This should not depend on unphysical parton-level information which introduces unwanted dependance on the MC generator. Using TMVA, classifiers were trained on the parton label using a number of hadron-level input variables. These classifiers were able to achieve a signal efficiency of up to 75% at a background rejection rate of 80%. This is slightly higher than the strongest input variable.

Both a Fisher discriminant and BDTs were trained. The Fisher discriminant was found to be as powerful as the BDTs. Since it is much simpler to understand in terms of an axis in hyperspace, and simpler to train, the Fisher method should be preferred in the future.

The classifiers trained were shown to have a strong dependence on the average JES response. The "quark-like" and "gluon-like" classes of jets coincide very well with the high response and low response classes of jets.

References

[1] ATLAS Collaboration. "Single hadron response measurement and calorimeter jet energy scale uncertainty with the ATLAS detector at the LHC". In: *Eur. Phys. J. C* 73 (2013), p. 2305. DOI: 10.1140/epjc/s10052-013-2305-1. arXiv: 1203.1302 [hep-ex].

[2] ATLAS Collaboration. "Jet energy scale measurements and their systematic uncertainties in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Rev. D* 96 (2017), p. 072002. DOI: 10.1103/PhysRevD.96.072002. arXiv: 1703.09665 [hep-ex].

[3] Gareth Smith. *Hadron-level quark/gluon tagging for ATLAS*. 2018. URL: https://cds.cern.ch/record/2636108.

[4] Tom Cornelis. "Quark-gluon Jet Discrimination At CMS". In: (2014). arXiv: 1409.3072 [hep-ex].

[5] ATLAS Collaboration. "Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector". In: *Eur. Phys. J. C* 74 (2014), p. 3023. DOI: 10.1140/epjc/s10052-014-3023-z. arXiv: 1405.6583 [hep-ex].

[6] Eric M. Metodiev, Benjamin Nachman, Jesse Thaler. "Classification without labels: Learning from mixed samples in high energy physics". In: *JHEP* 10 (2017), p. 174. DOI: 10.1007/JHEP10(2017)174. arXiv: 1708.02949 [hep-ph].

[7] Eric M. Metodiev, Jesse Thaler. "On the Topic of Jets: Disentangling Quarks and Gluons at Colliders". In: *Phys. Rev. Lett.* 120 (2018), p. 241602. DOI: 10.1103/PhysRevLett.120.241602. arXiv: 1802.00008 [hep-ph].