PROCEEDINGS OF SCIENCE

# Flavour tagging algorithms and performance at the ATLAS experiment

**Alessandro Calandri on behalf of the ATLAS Collaboration***
*Centre de Physique des Particules de Marseille*
*Aix-Marseille Université et CNRS/IN2P3*
*163, Avenue de Luminy - Case 902*
*13288 Marseille Cedex 09, France*
*E-mail:* alessandro.calandri@cern.ch

The identification of jets containing $b$-hadrons ($b$-tagging) is a fundamental ingredient in the physics program of the ATLAS experiment at the CERN LHC. This note presents the main $b$-tagging algorithms employed for the 2016 LHC run and is intended to update the procedure used for the 2015 run.

*Fourth Annual Large Hadron Collider Physics*
*13-18 June 2016*
*Lund, Sweden*

*Speaker.





PoS(LHCP2016)180

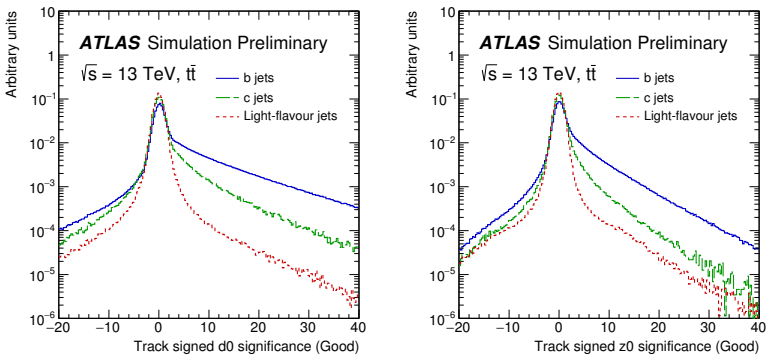
## 1. Basic Algorithms

The identification of $b$-quark jets in ATLAS (Ref. [1]), documented in Ref.[2], is based on distinct strategies carried out in three basic $b$-tagging algorithms: an impact parameter-based algorithm (Section 1.1) an inclusive secondary vertex reconstruction algorithm (Section 1.2) and a decay chain multi-vertex reconstruction algorithm (Section 1.3). The outputs of these $b$-tagging algorithms are combined in a multivariate discriminant (MV2), described in Section 2, which provides the best separation among the different flavours hypotheses.

### 1.1 IP2D and IP3D: The Impact Parameter based Algorithms

Lifetime-based algorithms rely on the specific $b$-hadron topology featuring at least one vertex displaced from the point where the hard-scatter collision takes place. Due to the long lifetime of hadrons containing a $b$-quark ($\sim$1.5 ps, $c\tau$ $\sim$450 $\mu$m) , tracks generated from $b$-hadron decay products tend to have large impact parameters enabling their contribution to be separated from the contribution of tracks from the primary vertex.

The IP2D tagger makes use of the transverse impact parameter significance, $d_0/\sigma_{d_0}$, as discriminating variable whereas IP3D uses both the transverse and the longitudinal impact parameter significance, $z_0 \sin\theta/\sigma_{z_0 \sin\theta}$, in a two-dimensional template to account for their correlation. It is possible to assign a sign to the impact parameter by determining, relatively to the jet direction, if the primary vertex is in front of the secondary vertex or behind. Probability density functions (PDF) obtained from reference histograms for the transverse and longitudinal impact parameter significances are separated into exclusive categories that depend on the hit pattern of a given track. Secondly, a log-likelihood ratio (LLR) discriminant is computed as the sum of the per-track contributions. The distributions for the transverse and longitudinal impact parameter significances are shown in Figure 1 for tracks from light, $c$- and $b$-jets for a track with typical hit pattern.




(a) (b)

Figure 1: The transverse (a) and longitudinal (b) signed impact parameter significance of tracks in $t\bar{t}$ events for $b$ (solid blue), $c$ (dashed green) and light flavour (dotted red) jets (Ref. [4]).

Several refinements in the algorithm have been introduced for the new version of the IP tagger compared to the version described in Ref [3] leading to relative gains of approximately 7% in light-flavour jet rejection and 3% in $c$-jet rejection at the 77% $b$-jet efficiency working point for $t\bar{t}$ events with respect to the previous algorithm used for the 2015 analyses (Ref. [3]). They are listed below:

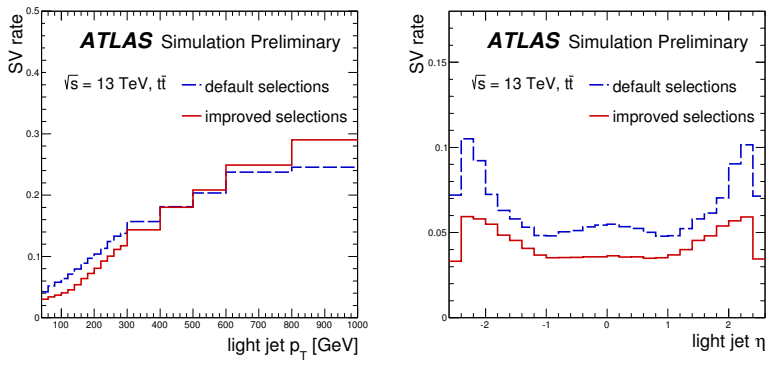
- The requirement on the number of pixel hits is relaxed from at least two to at least one. The previous requirement, used to ensure an excellent resolution, resulted in inefficiency in the high $b$-jet $p_T$ region because a significant fraction of high-$p_T$ $b$-hadrons decay after the IBL ($R$=3.3 cm) and the b-layer ($R$=5.5 cm).
- It is found that by identifying (cf. Section 1.2) and ignoring tracks originating from conversions, $\Lambda$ and $K_S$ decays, material interactions a gain of 15% in light-flavour jet rejection for a $b$-jet efficiency working point of 77% is achieved.
- Reference histograms are produced with a mixture (50%-50%) of $t\bar{t}$ and $Z' \to t\bar{t}$ (generated at the Z' mass pole of 4 TeV) for the two track categories with no hits in the first two layers, namely IBL and $b$-layer, despite them being expected (extrapolation from other measurement layer points towards the active regions in those layers). This allows us to overcome the limited statistics available in these categories in the $t\bar{t}$-only sample.

### 1.2 Secondary Vertex Finding Algorithm: SV

The secondary vertex finding algorithm (SV) explicitely reconstructs an inclusive displaced secondary vertex within the jet. All track pairs within a jet are tested for a two-track vertex hypothesis. Two-track vertices are discarded if they were likely to originate from the decay of a long-lived particle (e.g. $K_s$ or $\Lambda$), photon conversions or hadronic interactions with the detector material. Extra track requirements are used to improve the performance of the algorithm for the 2016 LHC run as follows.

- For highly-energetic jets ($p_T > 300$ GeV), the large number of tracks produced in jet fragmentation increases the probability of reconstructing fake vertices. In order to reduce the number of fake vertices, tracks are ordered according to their $p_T$ and at most 25 tracks with largest $p_T$ are used in the secondary vertex reconstruction.
- The reconstruction of tracks associated to jets in the high pseudorapidity region ($|\eta| > 1.5$) suffers from an increased amount of detector material leading to worse track parameter resolution. To increase the quality of the selected tracks, the minimal number of required hits in the silicon detectors is increased by one for tracks (from 5 to 6 hits) with $|\eta| > 1.5$.
- In order to reduce the impact of pileup, tracks with low $|d_0/\sigma_{d_0}|$ (<2) and high $|z_0/\sigma_{z_0}|$ (>6) are removed.

Figures 2 (a) and 2 (b) show the secondary vertex reconstruction rate for light-flavour jets when employing the previous track selection documented in Ref [3] and the updated requirements. Notably, the new track selection results in significantly smaller fake rate of SV reconstruction for light-flavour jets both as function of jet $p_T$ and $\eta$ while the $b$-jet efficiency is kept at approximately 80% throughout the full jet $p_T$ and $\eta$ spectra.




(a) (b)

Figure 2: Fake rate of secondary vertices for light-flavour jets as a function of jet $p_T$ (a) and $\eta$ (b) when comparing the previous track selection and the current requirements described in Section 1.2 (Ref. [4]).

### 1.3 Decay Chain Multi-Vertex Algorithm: JetFitter

The decay chain multi-vertex reconstruction algorithm, JetFitter [2], exploits the topological structure of weak $b$- and $c$-hadron decays inside the jet and tries to reconstruct the full $b$-hadron decay chain. A Kalman filter is used to find a common line on which the primary vertex and the bottom and charm vertices lie, approximating the $b$-hadron flight path, as well as their positions.

## 2. Multivariate Algorithm: MV2

To achieve a better discrimination than any of the basic algorithms can achieve individually, a Boosted Decisions Tree (BDT) algorithm is employed. It combines the outputs of the basic taggers described in the previous Sections.

### 2.1 Description of the algorithm and training procedure

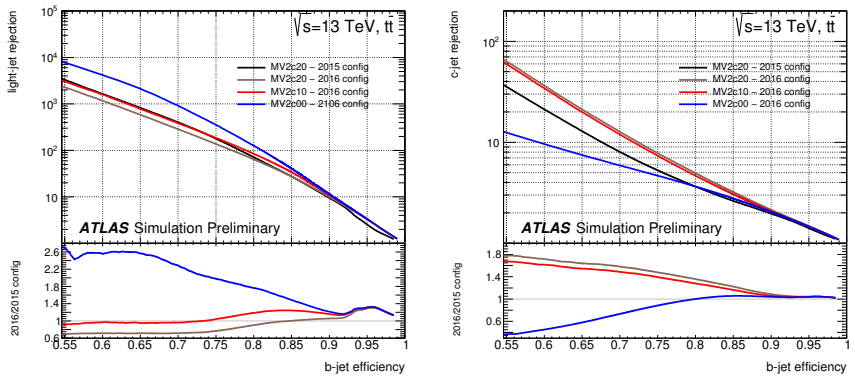
The list of input variables that are found to be discriminating among the different flavour hypotheses have not been modified with respect to the previous version of the MV2 algorithm documented in Ref. [3]. The training of the multivariate classifier is performed on jets from $t\bar{t}$ events. The kinematic properties of the jets, namely $p_T$ and $|\eta|$, are included in the training in order to take advantage of the correlations with the other input variables. However, to avoid differences in the kinematic distributions of signal and background being interpreted as discriminating by the training, the signal jets are reweighted in $p_T$ and $|\eta|$ to match the spectrum of the light-flavour background jets. The main refinements of the new version of the MV2 algorithms are listed below.

- Each tagging algorithm can fail to produce a result for a jet (e.g. if it does not reconstruct a secondary vertex in the case of SV or JetFitter or if no tracks fulfills the quality criteria defined for the impact parameter-based tagger); in the previous version of the algorithm, jets failing to produce results in any of the algorithms were given a penalty weight ($10^{-6}$) in

the training procedure. This approach was found to be sub-optimal; therefore in the current configuration, the penalty factor ($10^{-6}$) is applied if all (0.6% in the case of light-flavour jets and 0.1% for $c$- and $b$-jets) the three underlying tagger algorithms are found to be invalid.

- It is possible to modify the balance between light and charm rejection by changing the fraction of $c$-jets in the training. Given that the majority of physics analyses are presently more limited by the $c$- rather than the light-flavour jet rejection, the $c$-jet background fraction in the training has been chosen in such a way (7% $c$-jet fraction in the training for the released MV2 algorithm, MV2c10) to enhance the charm rejection by keeping a similar light-flavour jet rejection compared to the previous approach.
- The training parameters (depth, number of trees, minimum node size) of the BDT have also been optimized in order to take into account the new training conditions and the modified phase space.

The performance of the optimized MV2c00, MV2c10 and MV2c20 $b$-tagging algorithms is shown in Figure 3 for the light and $c$-jet rejection as a function of the $b$-jet efficiency in comparison to the 2015 MV2c20 configuration. The current MV2c10 (2016 configuration) discriminant, recommended for analyses on 2016 data, provides a similar light-flavour jet rejection (improvement of approximately 4% at 77% $b$-jet efficiency) to the 2015 MV2c20 configuration, but a significantly better $c$-jet rejection (+40%) at the 77% $b$-jet efficiency working point. Furthermore, as a consequence of the choice of the $c$-jet fraction in the training for MV2c10, the $\tau$-rejection has increased by approximately a factor 2 with respect the 2015 algorithm.




(a) (b)

Figure 3: Light-flavour jet (a) and $c$-jet (b) rejection versus $b$-jet efficiency for the previous (2015 config) and the current configuration (2016 config) of the MV2 $b$-tagging algorithm evaluated on $t\bar{t}$ events (Ref. [4]).

## 3. Conclusions

The current performance of the ATLAS baseline flavour-tagging algorithm for the 2016 LHC run has been described in this proceeding with an overview of the new training procedure for MV2

that leads to a gain of approximately 40% in $c$-jet rejection at 77% $b$-jet efficiency with respect to the previous algorithm used for 2015 analyses.

## References

[1] ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider. *JINST 3 S08003 (2008)*.

[2] ATLAS Collaboration, Performance of $b$-jet identification in the ATLAS Experiment. *JINST-11-P04008 (2016)*.

[3] ATLAS Collaboration, Expected performance of the ATLAS $b$-tagging algorithms in Run-2, *ATLAS-PUB-2015-022 (https://cds.cern.ch/record/2037697)*.

[4] ATLAS Collaboration, Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run. *ATLAS-CONF-2016-012 (https://cds.cern.ch/record/2160731)*.