Identification and tagging of double b-hadron jets from gluon splitting with the ATLAS Detector

Lic. María Laura González Silva

Tesis Doctoral en Ciencias Físicas
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Noviembre 2012

UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Física

Identification and tagging of double b-hadron jets from gluon splitting with the ATLAS Detector

Trabajo de Tesis para optar por el título de

Doctor de la Universidad de Buenos Aires en el área Ciencias Físicas

por **María Laura González Silva**

Director de Tesis: Dr. Ricardo Piegaia

Consejero de estudios: Dr. Daniel Deflorian

Lugar de Trabajo: Departamento de Física (CONICET-UBA)

Buenos Aires, 2012

AGRADECIMIENTOS

Quiero agradecer a mi director, Ricardo Piegaia, y a todos aquellos que trabajaron junto conmigo en el experimento ATLAS, Gastón Romeo, Gustavo Otero y Garzón, Hernán Reisin y Sabrina Sacerdotti. Un especial agradecimiento a Ariel Schwartzman y su equipo.

Quiero agradecer también a mis compañeros de grupo y oficina, Javier Tiffenberg, Yann Guardincerri, Pablo Pieroni y Orel Gueta.

Quiero agradecer al Experimento ATLAS, al programa HELEN y al programa e-Planet. Quiero agradecer al CONICET y a la Fundación Exactas por hacer posible la realización de esta tesis.

Quiero agradecer el apoyo de mis compañeros de la carrera, especialmente a mis amigos Cecilia Bejarano y Tomas Teitelbaum.

Quiero agradecer a los amigos que hice a lo largo de estos años en mis visitas al Laboratorio CERN, y a mis colegas y amigos de la Universidad de la Plata. Un especial agradecimiento a Fernando Monticelli.

Quiero agradecer a mis amigos de la vida por continuar a mi lado a pesar de las ausencias.

Finalmente, quiero agradecer a mi familia por su apoyo y comprensión, especialmente a Cristina Silva, Lorena González y Juan Martín Alba.

Abstract

Esta tesis describe un método que permite la identificación de jets que contienen dos hadrones b, que se originan en la división de un gluon en un par $b\bar{b}$. La técnica desarrollada explota las diferencias cinemáticas entre los llamados jets "merged" y los genuinos jets b, usando variables que describen la estructura interna y la forma de los jets, construidas a partir de las trazas asociadas a los mismos. Las variables con mayor poder discriminador son combinadas en un análisis de multivariable. Poder identificar y remover jets b que provienen de la división de un gluon es importante para la estimación y la redución del fondo a señales de física dentro del Modelo Estándar y en nueva física. El algoritmo diseñado rechaza, en eventos simulados, el 95% (50%) de los jets "merged", mientras que retiene el 50% (90%) de los jets b genuinos.

Palabras clave: Experimento ATLAS, Jets, Subestructura de Jets, Etiquetado de Jets b, *Gluon Splitting*.

Abstract

This thesis describes a method that allows the identification of double B-hadron jets originating from gluon-splitting. The technique exploits the kinematic differences between the so called “merged” jets and single B-hadron jets using track-based jet shape and jet substructure variables combined in a multivariate likelihood analysis. The ability to reject b-jets from gluon splitting is important to reduce and to improve the estimation of the b-tag background in Standard Model analyses and in new physics searches involving b-jets in the final state. In the simulation, the algorithm rejects 95% (50%) of merged B-hadron jets while retaining 50% (90%) of the tagged b-jets, although the exact values depend on the jet p_T.

Keywords: ATLAS Experiment, Jets, Jet Substructure, b-tagging, Gluon Splitting.

Contents

Chapter 1

The Multivariate Analysis

Multivariate technique multivariate classification learning algorithms They make use of training events, for which the desired output is known, to determine the mapping function that discribes a decision boundary (classification)

The following multivariate methods were explored:

- Likelihood ratio estimators
- Neural Networks (NN)
- Boosted decision Trees (BDTs)
- Fisher discriminants

The method of maximum likelihood consists of building a model out of probability density functions (PDF) that reproduces the input variables for signal and background. For a given event, the likelihood for being of signal type is obtained by multiplying the signal probability densities of all input variables and normalising this by the sum of the signal and background likelihoods. The likelihood ratio $y_L(i)$ for event i is defined by:

$$y_L(i) = \frac{L_S(i)}{L_S(i) + L_B(i)}, \tag{1.1}$$

where

$$L_{s(B)}(i) = \prod_{k=1}^{n_{var}} p_{s(B),k}(x_k(i)), \tag{1.2}$$

and where $p_{s(B),k}(x_k(i))$ is the signal (background) PDF for the kth input variable x_k. All the PDFs are normalized to one.

The parametric form of the PDFs is generally unknown, however it is possible to empirically approximate its shape from the training data by non-parametric functions which can be chosen individually for each variable and are either polynomial splines of various degrees fitted to binned histograms or unbinned kernel density estimators (KDE).

Both the training and the application of the likelihood are very fast operations that are suitable for very large data sets.

The method of Fisher discriminants [**?**].

An artificial Neural Network (NN) is a nonlinear discriminant, it is most generally speaking any simulated collection of interconnected neurons, with each neuron producing a certain response at a given set of input signals. . It can be viewed as a mapping from a space of input variables $x_1, ..., x_{n_{var}}$ onto, in the case of a signal-versus-background discimination problem, a one-dimensional output variable. The behaviour of an artificial neural network is determined by the layout of the neurons, the weights of the inter-neuron connections, and by the response of the neurons to the input, described by the neuron response function ÏĄ. While in principle a neural network with n neurons can have n^2 directional connections, the complexity can be reduced by organising the neurons in layers and only allowing direct connections from a given layer to the following layer. This kind of neural network is termed multi-layer perceptron. The first layer of a multilayer perceptron is the input layer, the last one the output layer, and all others are hidden layers. For a classification problem with nvar input variables the input layer consists of

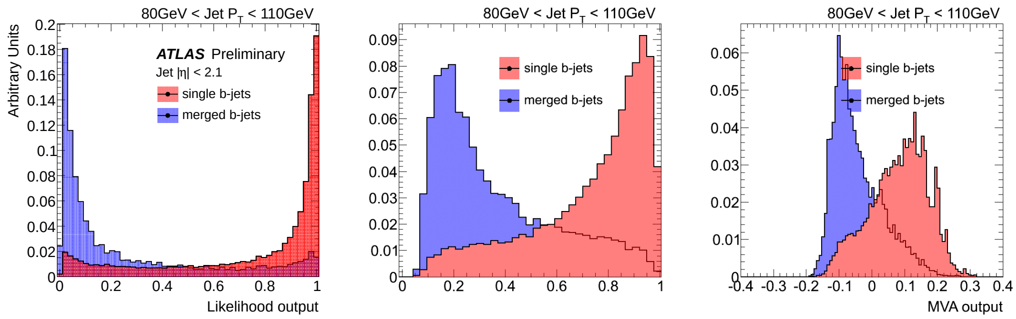

Figure 1.1: Distribution of the MVA discriminant outputs for the Likelihood (a), Neural Network (b) and Boosted Decision Trees (c) classifiers, for single and merged b-jets between 80 GeV and 110 GeV.

n_{var} neurons that hold the input values, $x_1, ..., x_{n_{var}}$, and one neuron in the output layer that holds the output variable, the neural net estimator y_{ANN} .

The NN used were built with only one hidden layer.

A decision tree (BDT) [?] is a binary tree structured classifier. Repeated yes/no decisions are taken on one single variable at a time until a stop criterion is fulfilled. The phase space is split this way into many regions that are eventually classified as signal or background, depending on the majority of training events that end up in the final $leaf$ node. The boosting of a decision tree extends this concept from one tree to several trees which form a $forest$.

Figure 1.2 a comparison of the performance of all methods.

1.1 $g \to b\bar{b}$ likelihood training and performance

A discriminant between single b-jets and merged b-jets was built by training a simple likelihood estimator in the context of the Toolkit for Multivariate Data Analysis, TMVA [1].

A sub-set of the dijet Monte Carlo sample was used for training. After the

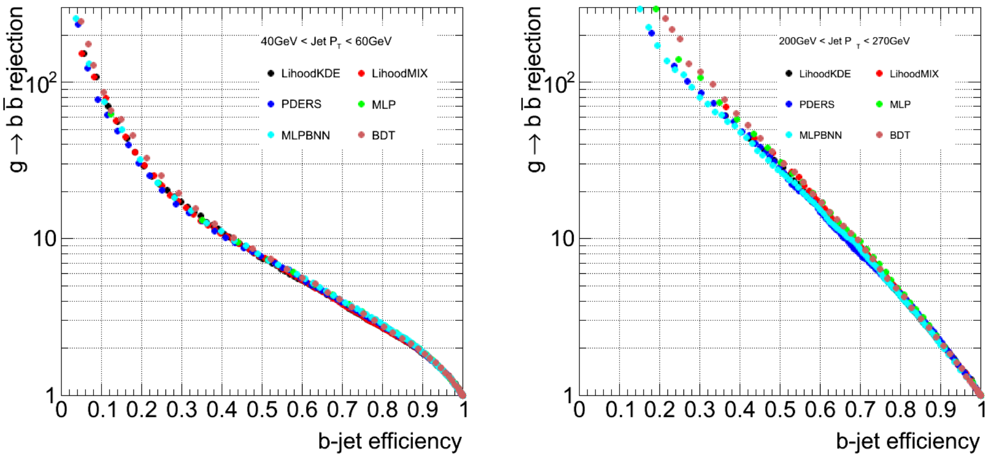

Figure 1.2: Rejection of merged b-jets as a function of single b-jet efficiency for the the different MVA methods evaluated for low and high jet p_T.

event and jet selections were performed, the b-tagged jets with $|\eta| < 2.1$ were classified as signal (single b-jets) or background (merged b). The likelihood training was done in bins of calorimeter jet p_T. Signal and background jets were not weighted by the dijet samples cross-sections to allow the contribution of subleading lower p_T jets from high p_T events, and thus increase the statistics of merged jets in the low p_T bins. For the evaluation of the method the same procedure was followed.

As mentioned in the previous section, the following combination of three variables was chosen for the multivariate analysis:

1. Jet track multiplicity
2. Track-jet width
3. ΔR between the axes of 2 k_t subjets within the jet

The distribution of the likelihood output for single and merged b-jets is shown in Fig. 1.3 for low, medium and high transverse momentum jets.

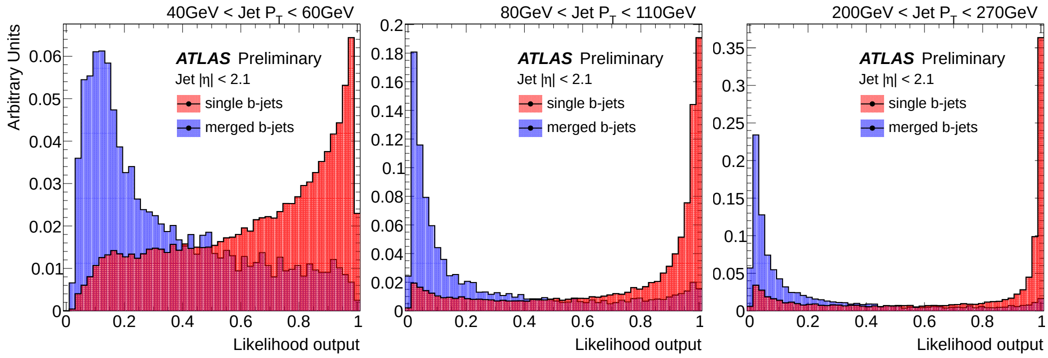

Figure 1.3: Distribution of the likelihood output for single and merged b-jets for low, medium and high p_T jets.

The performance of the tagger in the simulation can be displayed in a plot of rejection $(1/\epsilon_{bkg})$ of merged b-jets as a function of single b-jet efficiency, where ϵ_{bkg} is the probability that a double b-hadron jet passes the single b-jet tagger. This is shown in Fig. 1.4 for the eight bins of jet p_T mentioned in section **??**. The performance improves with p_T:

- $p_T > 40$ GeV: rejection above 8 at 50% eff.
- $p_T > 60$ GeV: rejection above 10 at 50% eff.
- $p_T > 200$ GeV: rejection above 30 at 50% eff.

The rejection of merged jets attained as a function of p_T for the 50% and 60% single b-jet efficiency working points are summarized in Table 1.1, together with their relative statistical error. These are propagated from the Poisson fluctuations of the number of events in the merged and single b distributions. The error is slightly lower for the 60% efficiency working point because a higher efficiency allows for a greater number of Monte Carlo events to measure the performance.

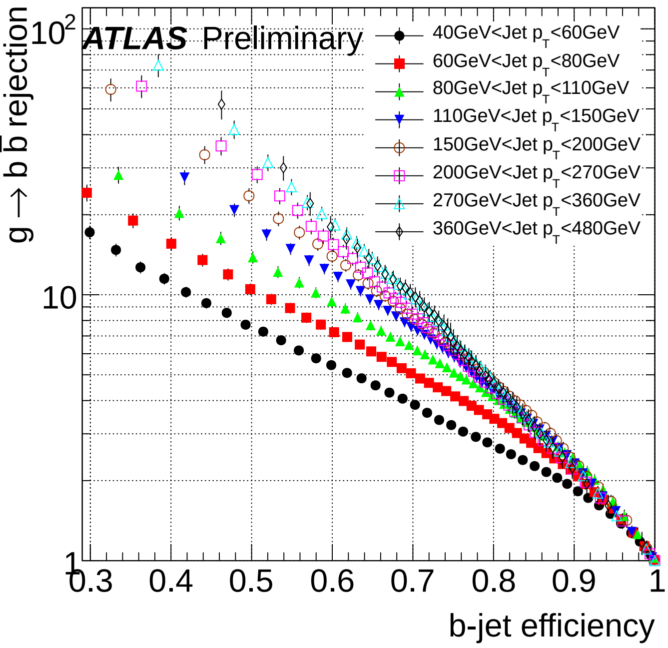

Figure 1.4: Rejection of merged b-jets as a function of single b-jet efficiency for dijet events in 8 jet p_T bins.

Jet p_T (GeV)	single b-jet efficiency 50%		single b-jet efficiency 60%	
	Rejection	stat.err.	Rejection	stat.err.
40 - 60	8	4%	5	3%
60 - 80	10	4%	7	4%
80 - 110	14	5%	9	4%
110 - 150	19	5%	12	4%
150 - 200	23	5%	14	5%
200 - 270	30	7%	16	6%
270 - 360	36	7%	19	6%
360 - 480	41	8%	18	8%

Table 1.1: The merged b-jet rejection for the 50% and 60% efficiency working points in bins of p_T.

1.2 Systematic uncertainties

The development, training and performance determination of the tagger is based on simulated events. Although the agreement between simulation and data explored in section **??** is a necessary validation condition, it is also important to investigate how the tagger performance depends on systematics relevant in the data. In particular we have considered:

- presence of additional interactions (pile-up);
- uncertainty in the b-jet tagging efficiency;
- uncertainty in the track reconstruction efficiency;
- uncertainty in the track transverse momentum resolution;
- uncertainty in the jet transverse momentum resolution;

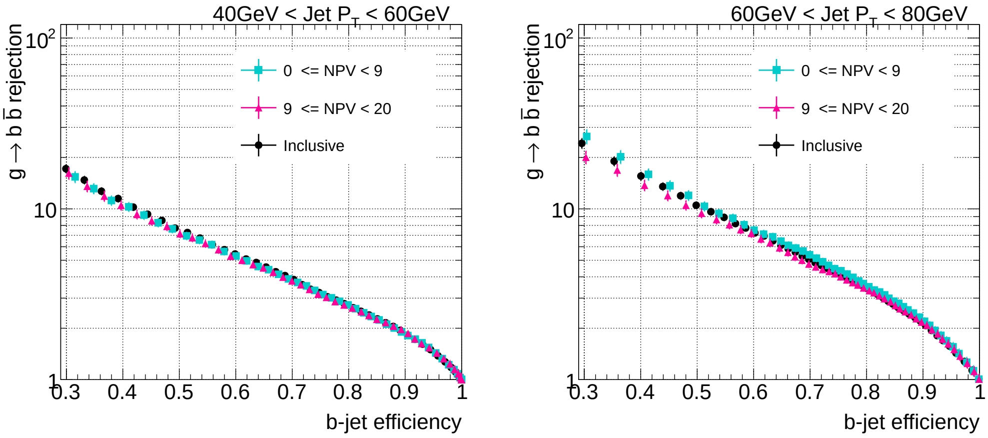

Figure 1.5: Rejection of merged b-jets as a function of single b-jet efficiency in bins of NPV for two low jet p_T bins.

- uncertainty in the b-jet energy scale.

I. Pile-up

The size of this effect was studied by comparing the performance of the likelihood discriminant with b-jets in events with small (1-9) and large (9-20) number of primary vertices. A comparison of the performance in these two sub-samples relative to the inclusive sample is shown in Fig. 1.5. As expected from the use of tracking (as opposed to calorimeter) variables no significant dependence with pile-up is observed within statistics. Performance differences between high and low number of primary vertices events are $\leq 0.3\%$ and therefore negligible compared to other sources of uncertainties. The impact of pile-up might be larger in 2012.

II. b-tagging efficiency

The performance of heavy-flavor tagging in Monte Carlo events is calibrated to experimental data by means of the scale factors (SFs). The SFs are defined as the ratio of the heavy-flavor tagging efficiency in data over that in

Monte Carlo for the different jet flavors. They are measured by the ATLAS Flavour Tagging Working group, and their measurement carries a systematic uncertainty.

To estimate the impact of this uncertainty a conservative approach is followed: the SFs are varied in all the p_T bins simultaneously by one standard deviation both in the up and down directions. The MC distributions weighted by the varied SFs show no major deviations from the nominal. In the same manner, the effect of the b-tagging calibration uncertainty on the likelihood peformance is $< 1\%$, negligible with respect to the statistical uncertainty. This was indeed expected. The scale factors depend on the true flavor of the jet and on its p_T, but these are basically constant in the performance determination, which is based on single flavor (true b-) jets classified in p_T-bins.

See Figures 1.6 and 1.7.

III. Track reconstruction efficiency

This uncertainty arises from the limit in the understanding of the material layout of the Inner Detector. To test its impact a fraction of tracks determined from the track efficiency uncertainty was randomly removed.

The tracking efficiency systematics are given in bins of track η. For tracks with $p_{\mathrm{T}}^{\mathrm{track}} > 500$ MeV the uncertainties are independent of p_T: 2% for $|\eta^{\mathrm{track}}| < 1.3$, 3% for $1.3 < |\eta^{\mathrm{track}}| < 1.9$, 4% for $1.9 < |\eta^{\mathrm{track}}| < 2.1$, 4% for $2.1 < |\eta^{\mathrm{track}}| < 2.3$ and 7% for $2.3 < |\eta^{\mathrm{track}}| < 2.5$ [2]. All numbers are relative to the corresponding tracking efficiencies.

The tracking variables were re-calculated and the performance of the nominal likelihood was evaluated in the new sample with worse tracking efficiency. The rejection-efficiency curves show a small degradation of the

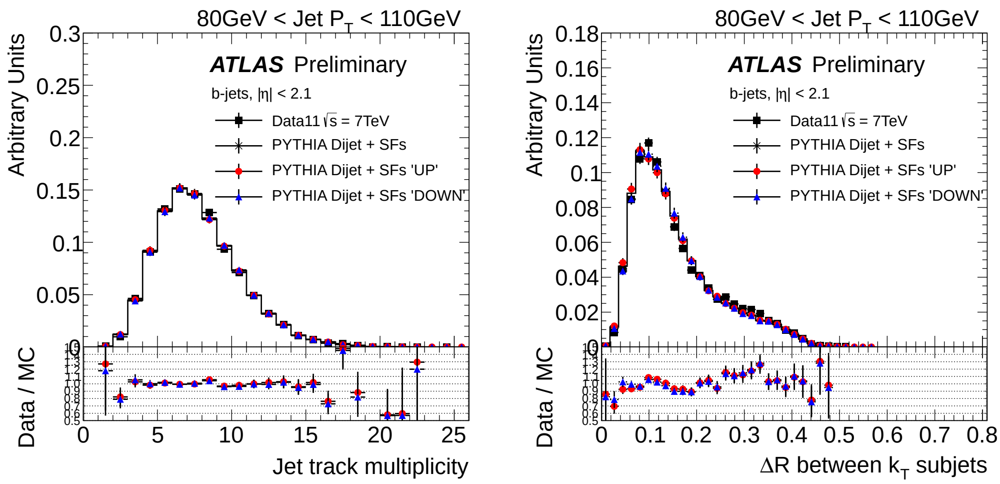

Figure 1.6: The effect of a variation in the b-tagging Scale Factors on the tracking variables distributions. Scale Factors were varied up (down) by 1-sigma to evaluate the systematic uncertainty from this source. The ratio data over MC is shown for MC PYTHIA with SFs varied up (circles) and down (triangles).

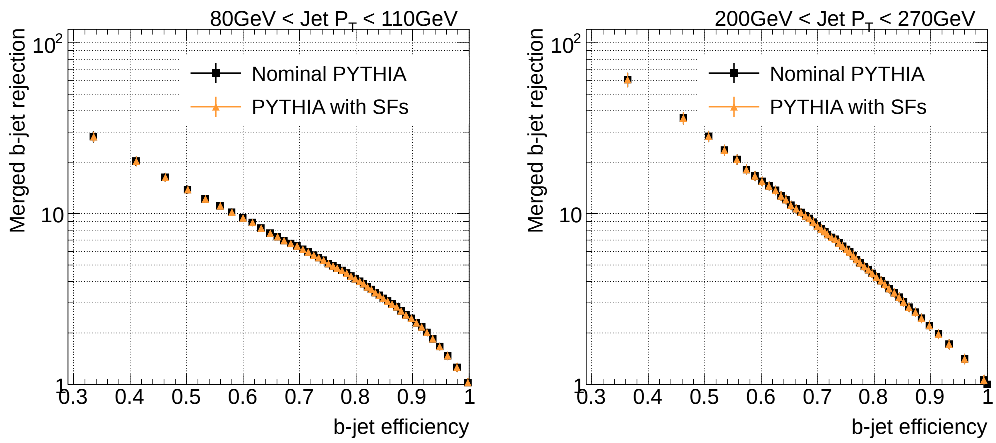

Figure 1.7: Rejection of merged b-jets as a function of single b-jet efficiency with and without scale factors.

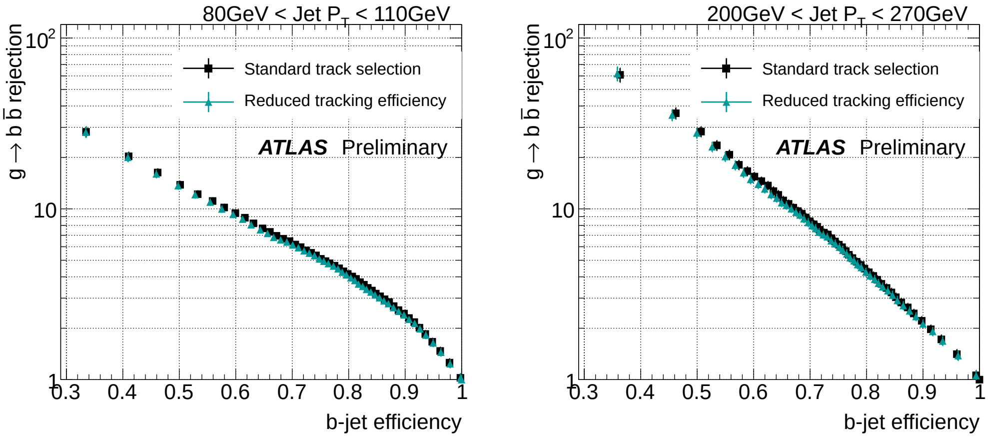

Figure 1.8: Rejection of merged b-jets as a function of the single b-jet efficiency showing shift in likelihood performance caused by a reduction in the tracking efficiency.

performance which is comparable to the statistical uncertainty. The effect is however systematically present over all 16 p_T bin/working points, without a clear p_T dependence. We have thus taken the average over p_T, and obtained a global systematic uncertainty of 4% both for the 50% and 60% efficiency working points. The performance comparison is shown in in Fig. 1.8 for two p_T bins.

IV. Track momentum resolution

The knowledge of the track momentum resolution is limited by the precision both in the material description of the Inner Detector and in the mapping of the magnetic field. Its uncertainty propagates to the kinematic variables used in the double b-hadron jet tagger. In order to study this effect, track momenta are over-smeared according to the measured resolution uncertainties, before the track selection cuts are applied. The actual smearing is done in $1/p_T$, with an upper bound to the resolution uncertainty given by $\sigma(1/p_T) = 0.02/p_T$ [3].

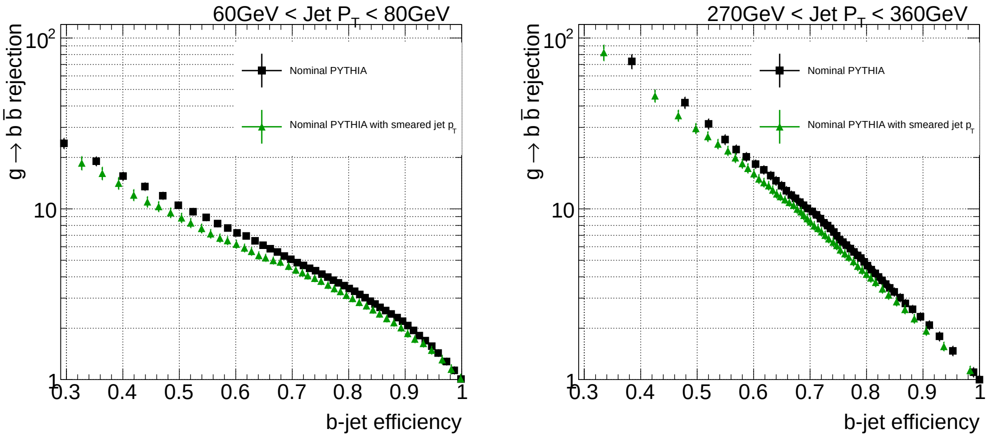

Figure 1.9: Rejection of merged b-jets as a function of single b-jet efficiency for jets with smeared p_T.

The effect is found to be negligible.

V. Jet transverse momentum resolution

The jet momentum resolution was measured for 2011 data and found to be in agreement with the predictions from the PYTHIA-based simulation [4]. The precision of this measurement, determined in p_T and η bins,is typically 10%. The systematic uncertainty due to the calorimeter jet p_T resolution was estimated by over-smearing the jet 4-momentum in the simulated data, without changing jet η or ϕ angles. The performance is found to globally decrease by 6%, without a particular p_T dependence.

See Fig. 1.9.

VI. Jet energy scale for heavy flavour jets

The jet energy scale (JES) uncertainty for light jets reconstructed with the anti-k_t algorithm with distance parameter $R = 0.4$ and calibrated to the EM+JES scale is between ∼4% at low p_T and ∼2.5% for jets with

p_T >60 GeV in the central region [?]. In the case of b-jets, and additional uncertainty arising from the modelling of the b-quark production mechanism and the b-quark fragmentation was determined from systematic variations of the Monte Carlo simulation. The resulting fractional additional JES uncertainty for b-jets has an upper bound of 2% for jets with $p_T \leq$100 GeV and it is below 1% for higher p_T jets. To obtain the overall b-jet uncertainty this needs to be added in quadrature to the light JES uncertainty.

The systematic uncertainty originating from the jet energy scale is obtained by scaling the p_T of each jet in the simulation up and down by one standard deviation according to the uncertainty of the JES. The effect on the likelihood performance is an average variation of 5% for the 50% and 60% efficiency working points.

The different contributions to the systematic uncertainty on the merged b-jet rejection are summarized in Table 1.2.

Systematic source	Uncertainty
pile-up	neglible
b-tagging efficiency	neglible
track reconstruction efficiency	4%
track p_T resolution	neglible
jet p_T resolution	6%
jet energy scale	5%

Table 1.2: Systematic uncertainities in the merged b-jet rejection (common to both the 50% and the 60% efficiency working points).

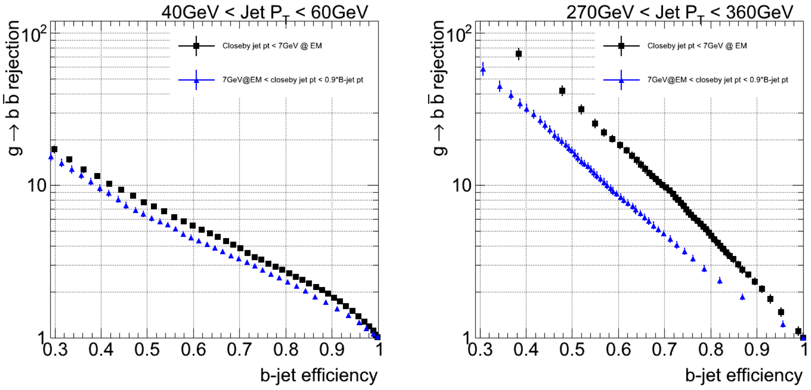

Figure 1.10: Rejection of merged b-jets as a function of single b-jet efficiency for two different isolation cuts.

Although the tagger was derived with isolated jets it can also be applied to non-isolated jets. Studies were performed to evaluate the likelihood rejection in b-jets with close-by jet with p_T between 7 GeV at electromagnetic scale scale and 90% of the b-jet p_T. The results can be seen in Fig. 1.10. The presence of close-by jets with a susbtancial fraction of the b-jet pt worsens the performance in more than 50% at very high p_T.

1.3 Other Monte Carlo generators

The development, training and performance determination of the tagger has been done using Monte Carlo events generated with the PYTHIA event simulator, interfaced to the GEANT4 based simulation of the ATLAS detector. An immediate question is what the performance would be if studied with a different simulation. In this section we investigate this question for the PYTHIA Perugia tune and the HERWIG++ event generators.

Fig. 1.11 shows a comparison of the likelihood rejection, at the 50% efficiency working point, between nominal PYTHIA and the alternative simulations as a function of the jet p_T . The larger errors are due to the reduced statistics available, which are even lower for the Perugia case than for HERWIG.

The performance in HERWIG shows a systematic trend, with agreement at low p_T and increasingly poor performances compared to PYTHIA as p_T grows. For the Perugia tune, on the other hand, there is no definite behavior, with the performance fluctuating above or below the nominal simulation for different p_T bins consistently with the statistical uncertainties.

The reason for the systematic difference observed between the performances of PYTHIA and HERWIG can be traced to the extent with which jets are accurately modelled. Fig. 1.12 compares the measured jet track multiplicity distributions in b-tagged jets and the prediction from both simulations, for low and high p_T jets. It is observed that indeed HERWIG++ does not correctly reproduce the data, particularly at high p_T. The level of agreement is found to be better for track-jet width and the ΔR between the axes of the two k_t subjets in the jet, the two other variables used for discrimination.

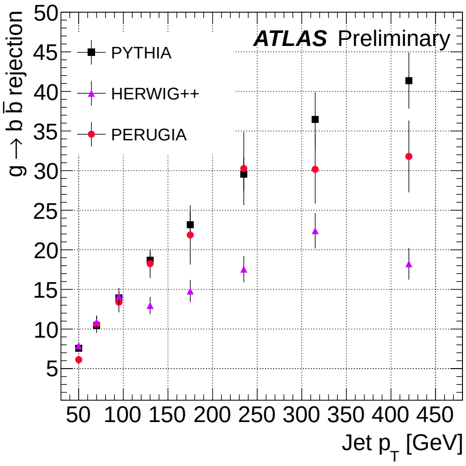

Figure 1.11: Rejection of $g \to b\bar{b}$ merged b-jets as a function of jet p_T for different Monte Carlo generators, at the 50% efficiency working point.

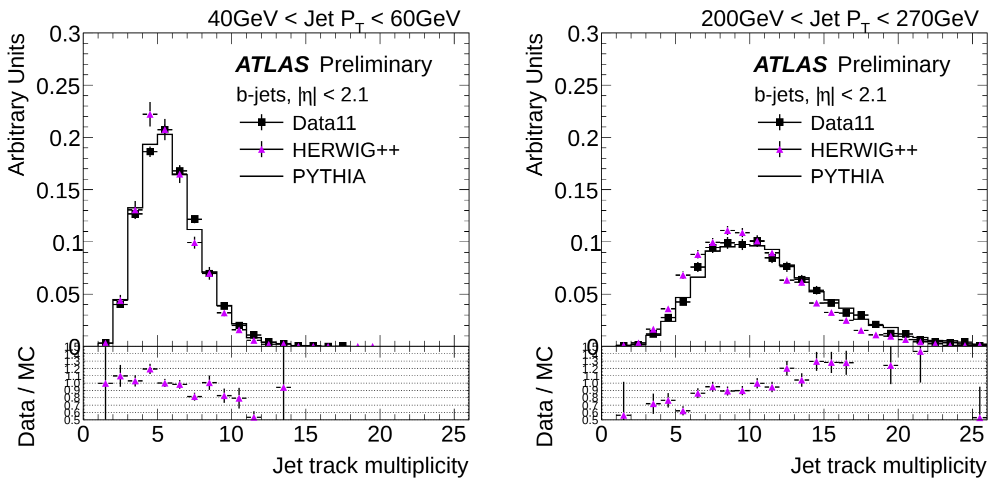

Figure 1.12: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected during 2011 (solid black points) and HERWIG++ events (solid violet triangules). The ratio data over HERWIG++ simulation is shown at the bottom of the plot. PYTHIA distribution is also shown for reference.

Bibliography

[1] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[2] G. Aad et al. Charged-particle multiplicities in pp interactions measured with the ATLAS detector at the LHC. *New J.Phys.*, 13:053033, 2011.

[3] ATLAS Collaboration. Estimating Track Momentum Resolution in Minimum Bias Events using Simulation and K_s in $\sqrt{s} = 900$ GeV collision data. *ATLAS-CONF-2010-009*, 2010.

[4] G. Romeo, A. Schwartzman, R. Piegaia, T. Carli, and R. Teuscher. Jet energy resolution from in-situ techniques with the atlas detector using proton-proton collisions at a center of mass energy $\sqrt{s} = 7$ tev. *ATL-COM-PHYS-2011-240*, 2011.