Double b-hadron and Quark/Gluon Jet Tagging at ATLAS

M.Laura González Silva(UBA),
on behalf of the ATLAS Collaboration

BOOST 2012

Valencia, Spain
July 27 2012

In this talk:

- Light quark initiated and gluon initiated jet discrimination.

  - Method for extracting distributions in data, tested in purified samples of quark-like and gluon-like jets

- Tool to separate b-jets containing one/two b-hadrons

  - Exploiting substructure differences between single and merged b-jets.

Light quark and gluon initiated jet identification

Introduction

Much work has gone into understanding quark- / gluon-like (q/g) jets:

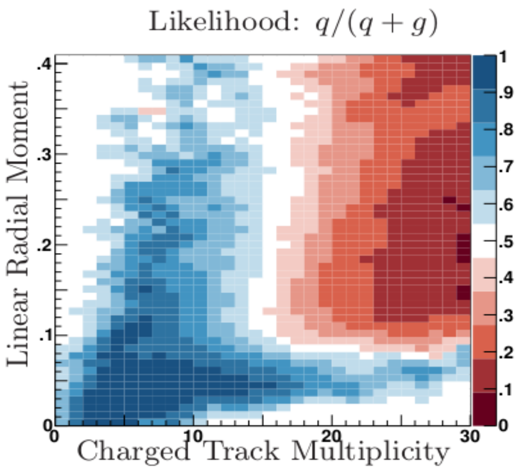
- LEP showed gluon to be broader (Phys. Lett. B 265 (1991) 462-474);
- Calorimeter response larger for light quark initiated jets (ATLAS-CONF-2011-053);
- **Theory result from Schwartz and Gallichio** (hep-ph:1106.3076) shows large differences between q/g jets.

There are several practical reasons for trying to separate these classes:

- Understanding issues with the jet energy scale as we go from analysis to analysis;
- Understanding the modeling of jet properties in the MC in more detail;
- Providing (potentially) a signal/background discriminant for use in searches.

Quark and Gluon Tagging at the LHC

J. Gallicchio and M. D. Schwartz

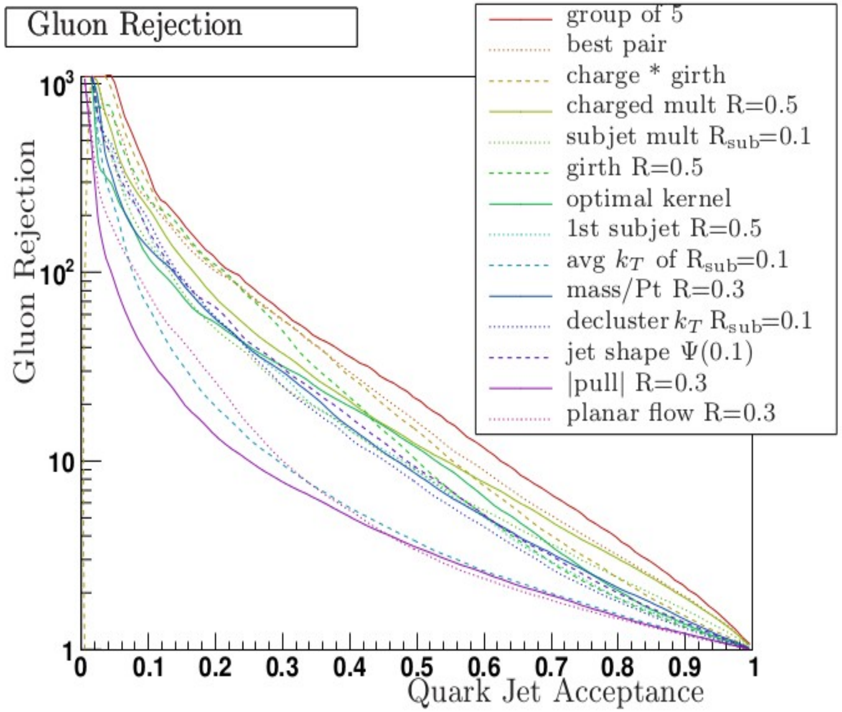


Filters 95% of the gluon jets at 50% quark-jet efficiency

Many *discrete* or *continuous* variables were studied to see which are best suited to quark/gluon tagging.

The strongest discrete observable is the number of charged particles within the jet.

The best observable in the continuous category is the linear radial moment ~ jet broadening.



Light quark and gluon initiated jets at ATLAS

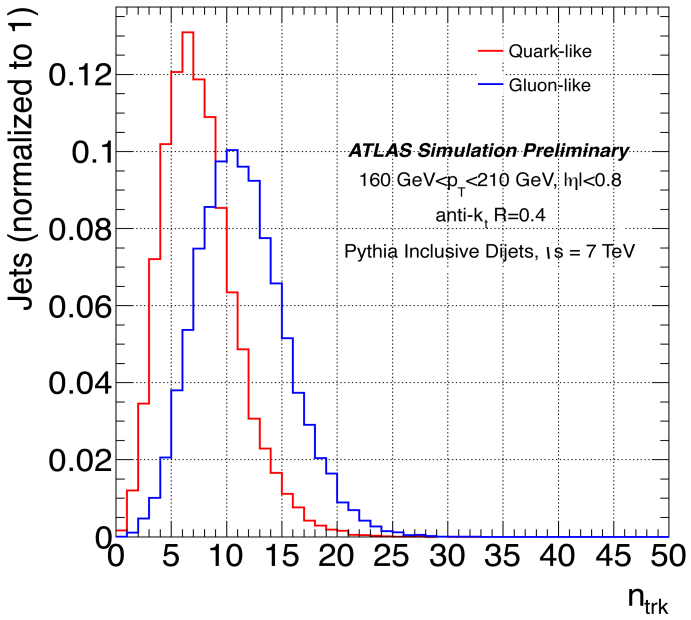
- Study discriminant variables for quark-like and gluon-like jets in MC simulation. Most promising are jet track multiplicity and jet width.
- Low agreement between data and MC leads us to derive light-quark initiated and gluon initiated jets variable distributions from data via a template method.
- Check these in situ with highly purified samples of quark and gluon initiated jets from gamma+jet and multijet events (**Schwartz and Gallichio,** hep-ph:1104.1175).
- Estimate the performance of a likelihood q/g tagger.

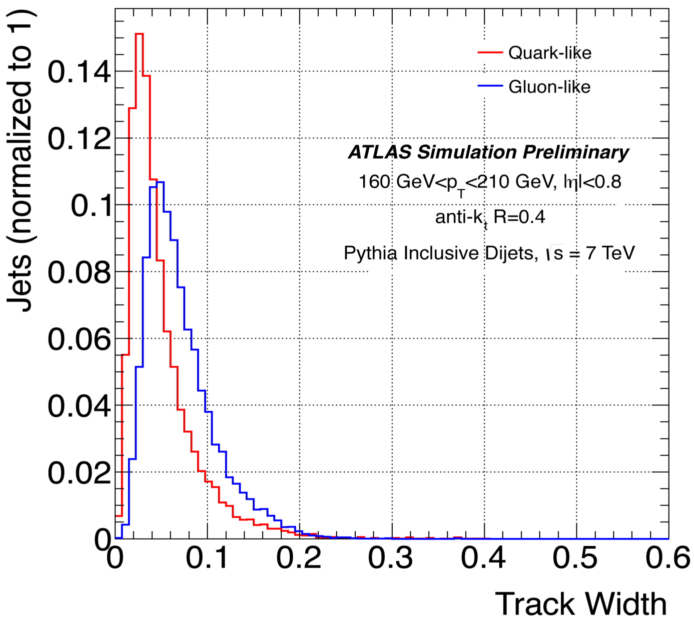
Analysis details

- Looked at isolated anti-k_T jets with distance parameter $R = 0.4$.
- Use track-based kinematic/shape variables to avoid effects from pile-up: jet width, charged multiplicity.
- Charge particle tracks with $p_T > 1$ GeV.
- q/g labeling:

  Jets were matched to the highest energy parton that lies inside the cone of the jet.

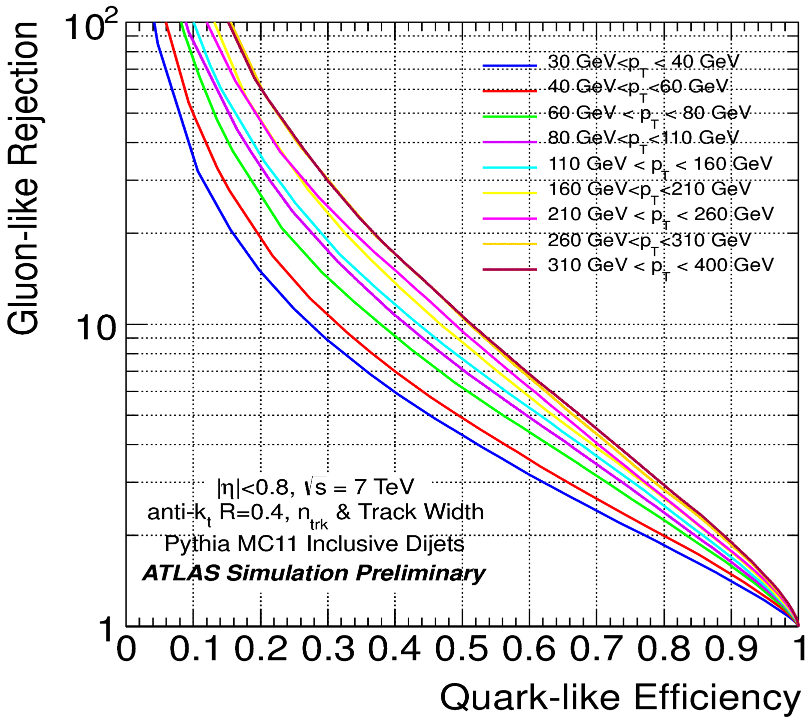
Variables for discrimination in ATLAS MC





- n_{trk}: number of good quality tracks within a cone of 0.4 in η–ϕ around the jet axis.
- Track Width: we use tracks associated to the jet, $\text{Track Width} = \frac{(\sum_i \Delta R(jet,i) p_T^i)}{\sum_i (p_T^i)}$

Efficiency vs Rejection

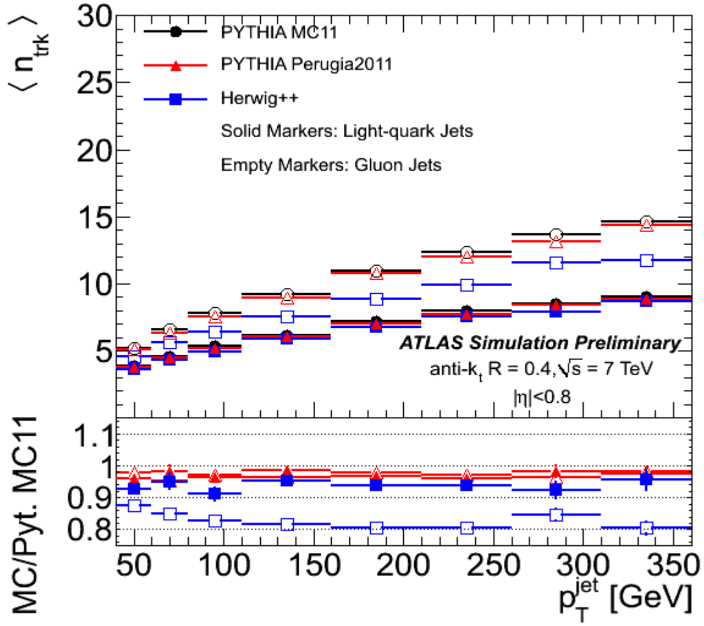


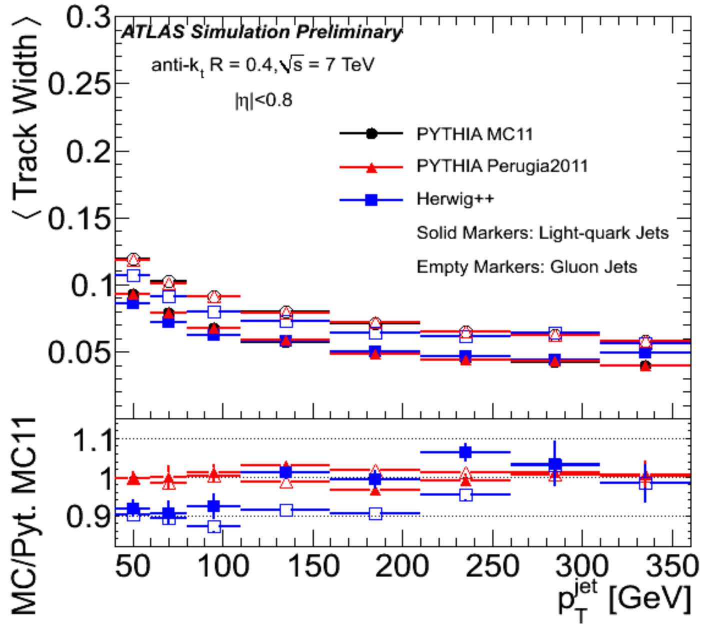
$|\eta| < 0.8$, Jet pT ~150 GeV:

Sample	Efficiency	Rejection
Pythia MC11	50%	8x
Pythia MC11	90%	5x

- Likelihood built from n_{trk} and Track Width in Pythia for isolated jets.
- The efficiency and rejection are derived using jets tagged with the generator event record.
- BUT, variables distributions differ in Pythia, Herwig++ and data.

Properties in different MCs





- Large difference in multiplicity between Herwig++ and Pythia in gluon jets.
- Some significant difference in width at low pT.

Template method

- Goal: to measure the quark/gluon shapes from data, dijet (*DJ*) and photon+jet *(γ J)* events.
- Ideally, solve for q/g (for each bin i) from:

$$h_i(DJ) = P_Q(DJ)q_i + P_G(DJ)g_i$$
$$h_i(\gamma J) = P_Q(\gamma J)q_i + P_G(\gamma J)g_i$$

P_Q = quark percentage, from MC

h = histogram value, from data

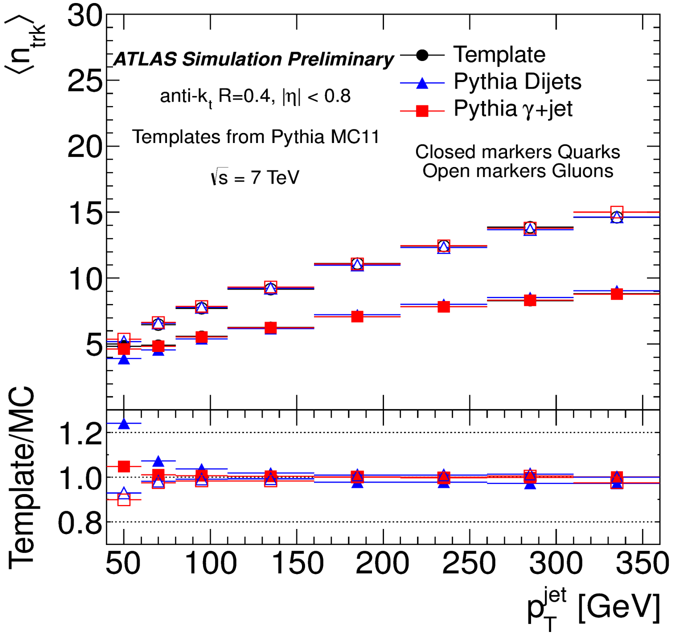
q/g = pure q/g jet distributions

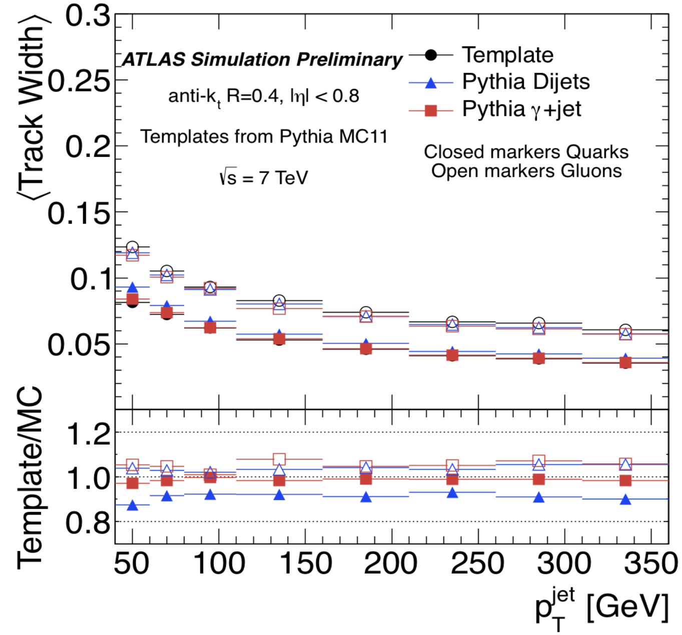
(solving for these)

- But need to account for *b* and *c* fractions (taken from MC):

$$h_i(DJ) = P_Q(DJ)q_i + P_G(DJ)g_i + P_B(DJ)b_i + P_C(DJ)c_i$$
$$h_i(\gamma J) = P_Q(\gamma J)q_i + P_G(\gamma J)g_i + P_B(\gamma J)b_i + P_C(\gamma J)c_i$$

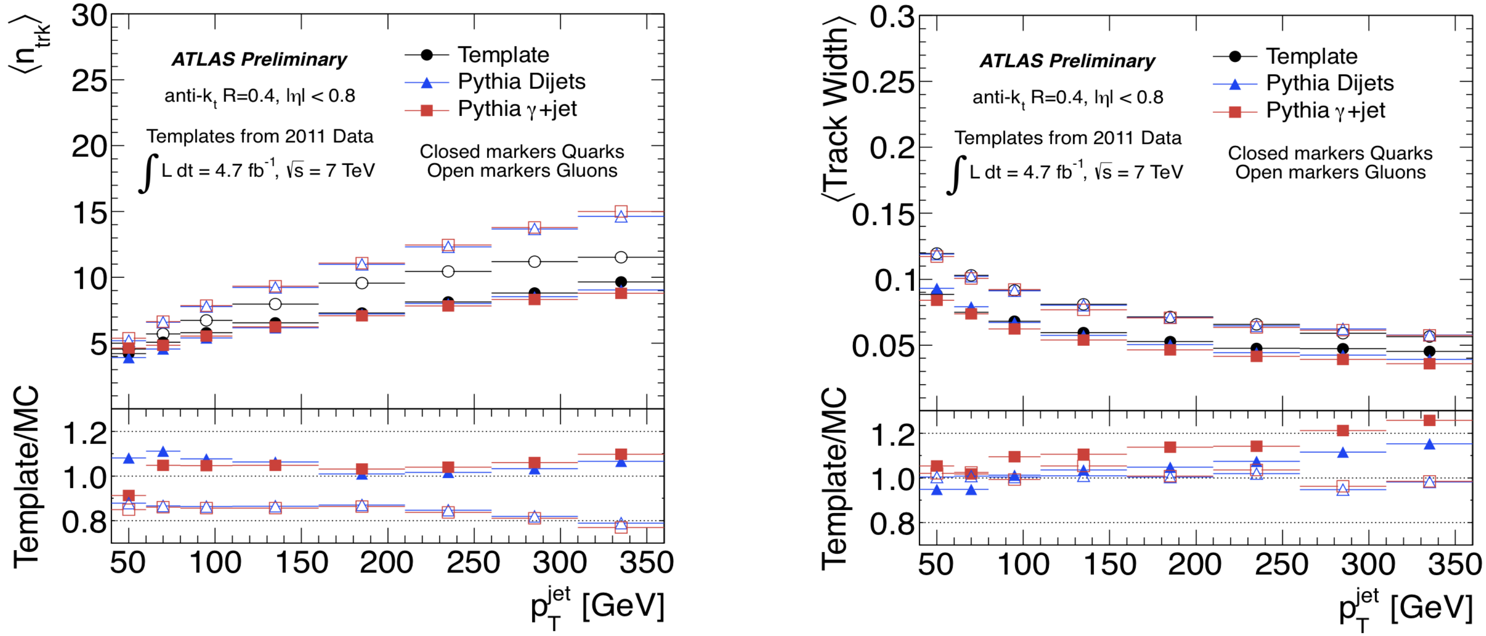
Template method: testing in MC





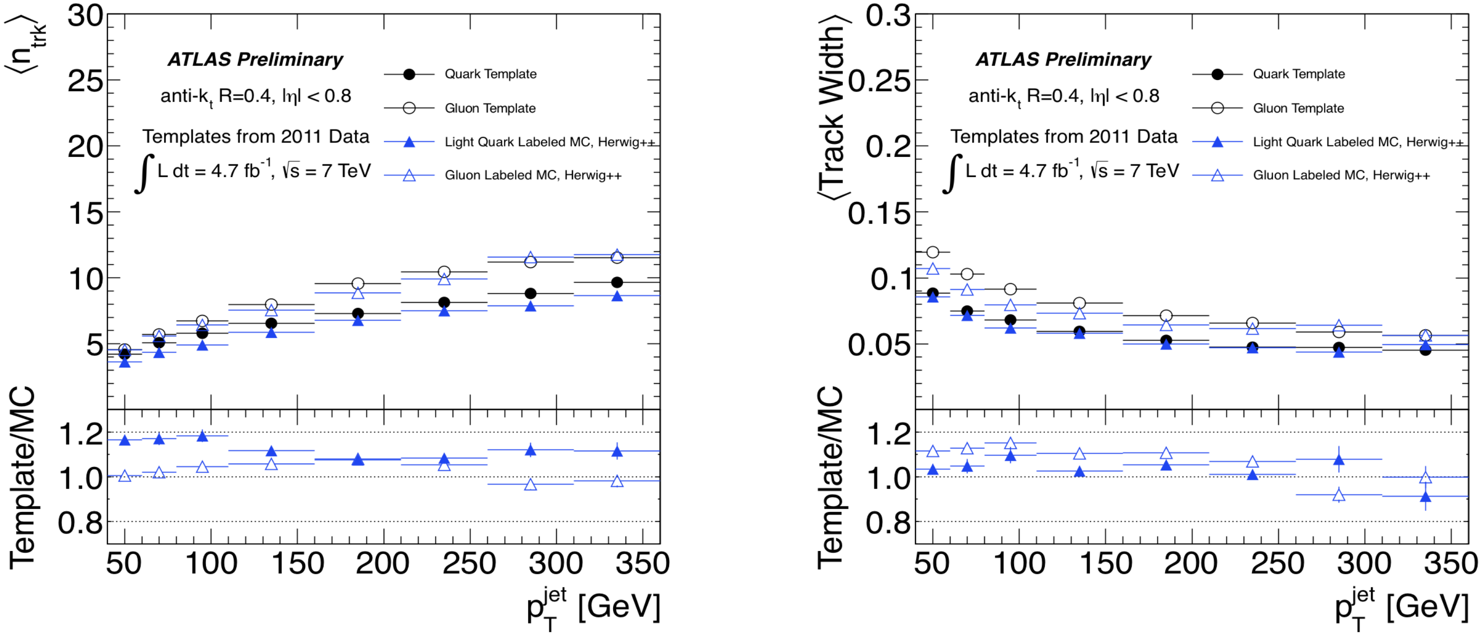
- Small differences in track width between different samples, even among the labeled jets, mean closure uncertainties in the method.
- Track multiplicity looks excellent, except in the lowest pT bin.
- The extraction does about as well as one can reasonably expect given those differences.

Template method: Data measurement



- Relative to the last set, only the template has changed (from MC sim to data)
- Track width shows good agreement.
- Gluon induced jet templates for n_{trk} show disagreement between data and MC simulation, demonstrating a MC mis-modeling of the gluon induced jet properties.

Template method: Data compared to Herwig++



- Track multiplicity for gluon induced jets is better described in Herwig++.
- Agreement is poorer for Track Width compared to Pythia.

Purified samples

- Purified samples can provide cross-check for templates.
  - Multijet sample with:

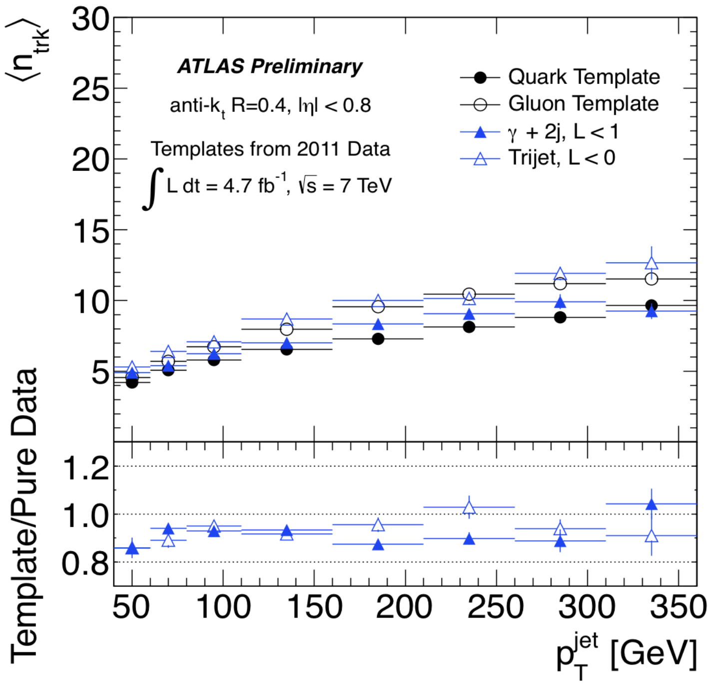
    $$L_q = \eta_\gamma \eta_{j1} + \Delta R_{\gamma j2} \quad < 0$$

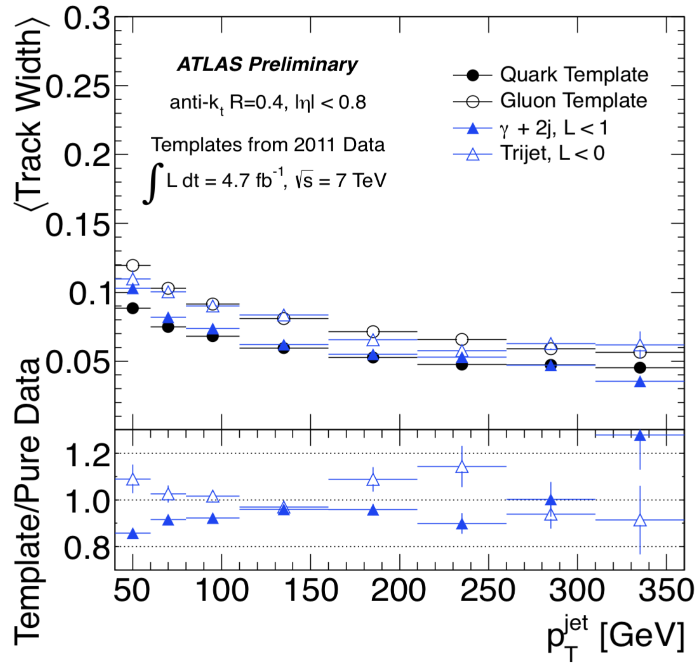
    gives > 90% pure gluon jet samples;
  - Photon+Jets samples with:

    $$L_g = |\eta_{j3}| - |\eta_{j1} - \eta_{j2}| \quad < 1$$

    gives ~90% pure quark jet samples.

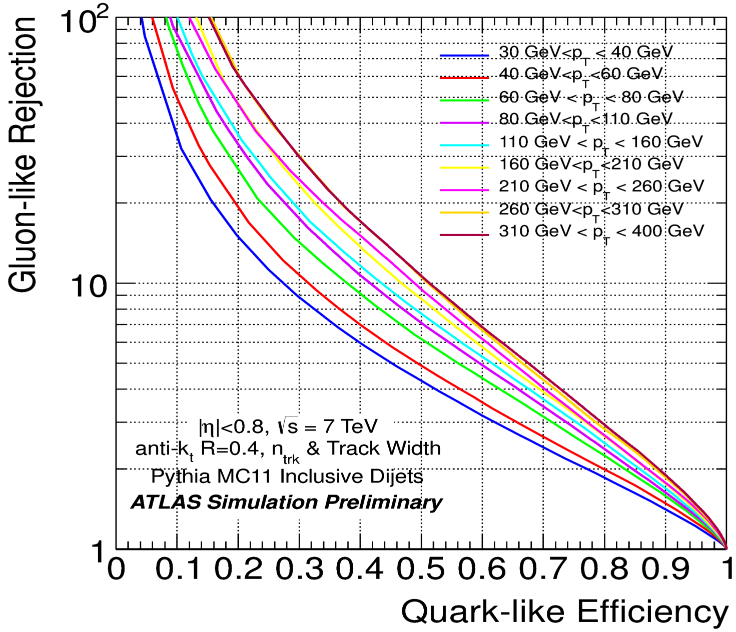
- With the statistics available, the purified samples show good agreement with extracted templates.





q/g Summary

- **Significant differences** amongst the MC simulations in the gluon jet properties was observed.
- And a significant disagreement between data and MC simulation in extracted gluon templates.
- The difference is validated by a method from purified samples.
- Scale factors and careful understanding of sample dependence is needed for use in physics analyses, currently deriving such scale factors for ATLAS analyses.
- Likelihood performance using distributions from data is reduced:



|η| <0.8, Jet pT ~150 GeV:

Sample	Efficiency	Rejection
Pythia MC11	50%	8x
Data 2011	50%	4x

Double b-hadron Jet Tagging

Introduction

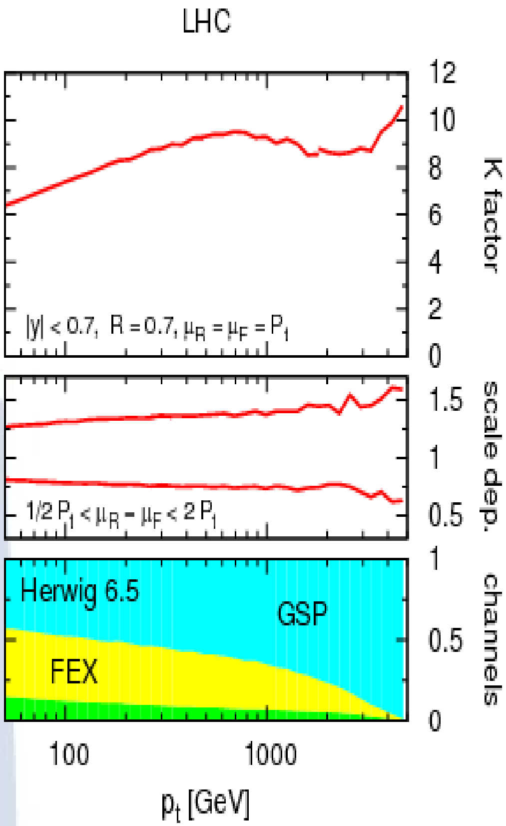
B-tagging algorithms do not provide information on the number of b-hadrons within a jet.

- The identification of close by b-hadron pairs has been approached using vertexing **(CDF Collaboration, arxiv:hep-ex/0412006)**
- We developed an alternative method that exploits the substructure differences between single and merged b-jets.

Possible applications,

- Measurement of QCD beauty production in the framework of the proposed flavour-k_T jet algorithm, that discerns between single and merged b-jets **(Banfi, Salam and Zanderighi, arxiv:hep-ph/0601139)**.
- Rejection of QCD/W+jets background in BSM searches dominated by single b-jets

Accurate QCD predictions for heavy-quark jets at the Tevatron and LHC, A. Banfi, G. P. Salam, and G. Zanderighi, JHEP 0707:026



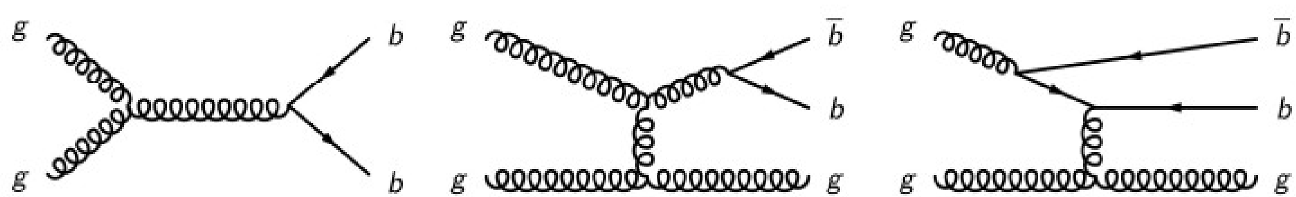
Inclusive b-jet spectrum has large theoretical uncertainties

K-factor (NLO/LO) as obtained with MCFM: K~ 6-10

Scale dependence is large: 50%

At LO only FCR is present, at NLO, 2 new channels open up: FEX and GSP.

Largest uncertainties are associated with channel with most logarithms: GLUON SPLITTING



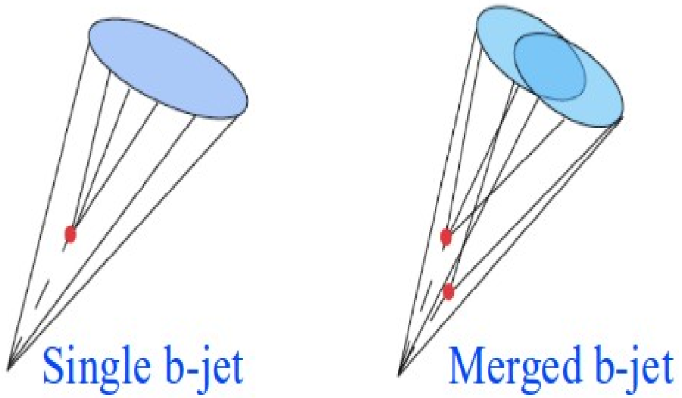
Proposal: Use the Flavour-kt algorithm: K ~ 1.2-1.4, scale dependence 10%

Jets containing equal number of b & $\bar{b}$ considered to be a light jet

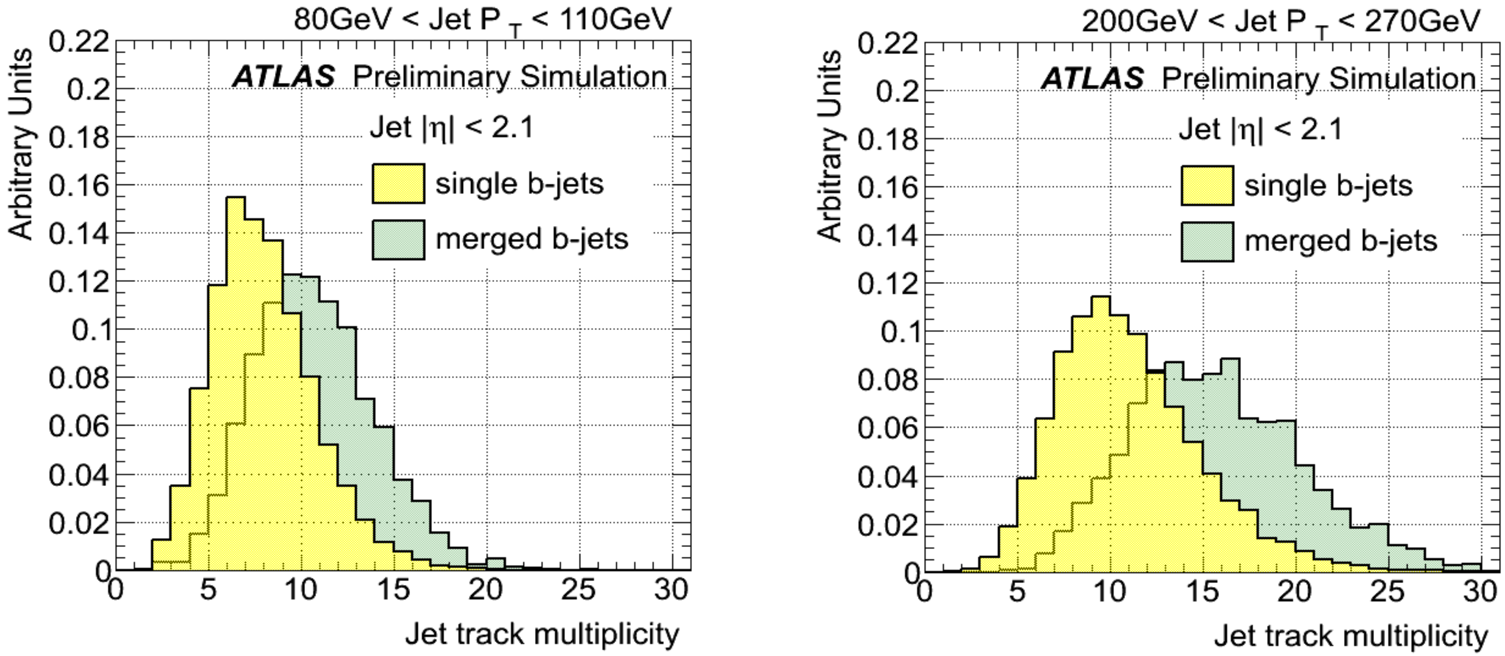
Analysis details

- Looked at isolated anti-k_T jets with distance parameter R = 0.4.
- Use track-based kinematic/shape variables to avoid pile-up effects: jet width, charged multiplicity, sub-jets.
- Charge particle tracks with p_T > 1 GeV.
- Double b-hadron jets:

  Jets were labeled as “merged” if they contained two b-hadrons within a radius of 0.4.
- Jets were tagged using ATLAS MV1 tagging algorithm.

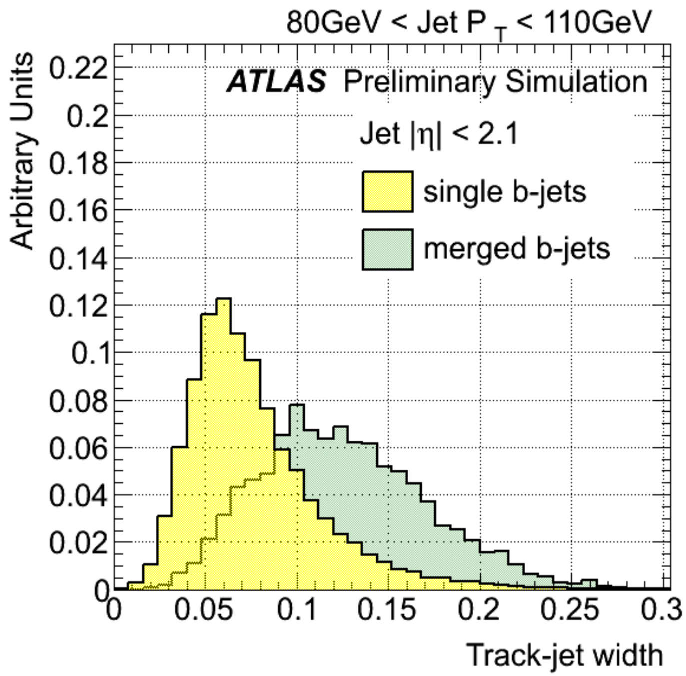


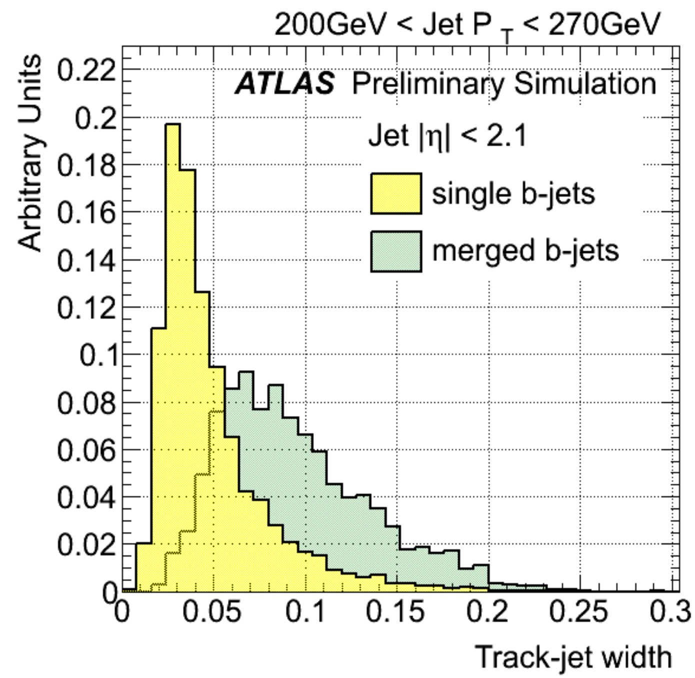
Single/Double b-hadron jets: discriminating variables



- Jet track multiplicity: same as in q/g n_{trk}.
- Merged b-jets contain on average 50% (70%) more tracks than single b-jets at low (high) jet pT.

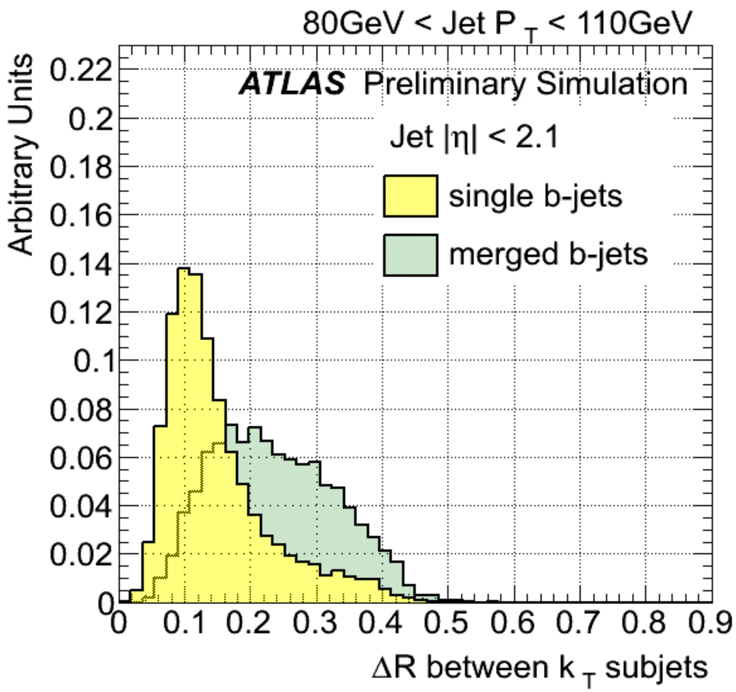
Single/Double b-hadron jets: discriminating variables

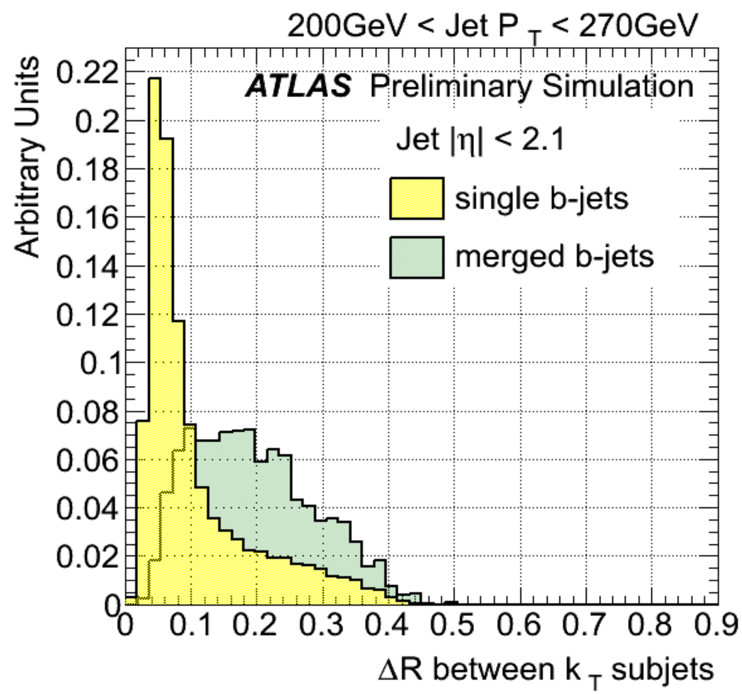




- Track-jet width: p_T weighted average of the ΔR distance between the associated tracks and the jet axis, as in q/g Track Width.
- As expected, merged b-jets are wider than single b-jets.

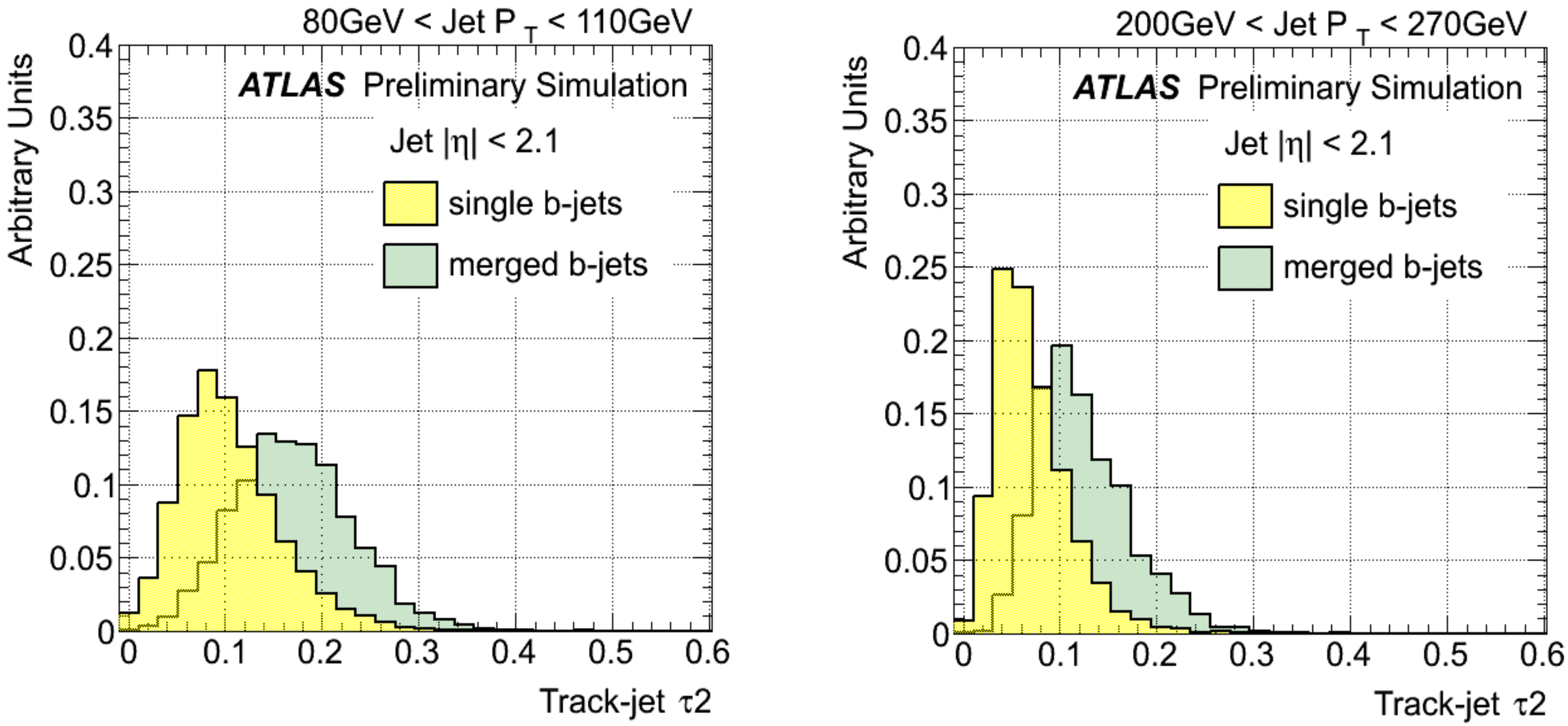
Single/Double b-hadron jets: discriminating variables





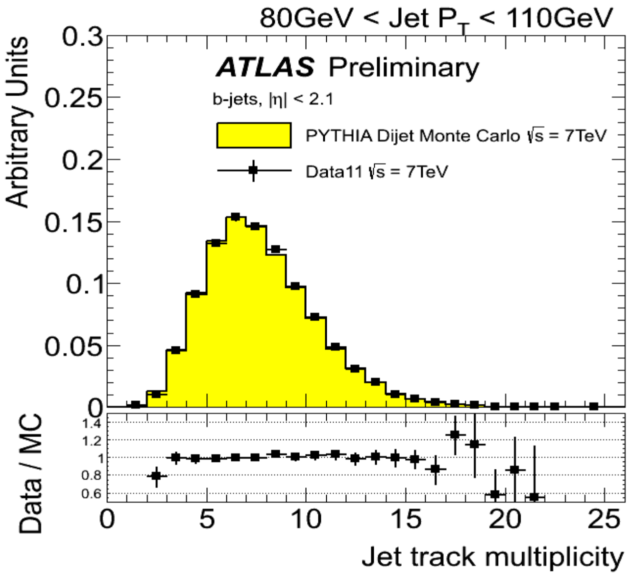
- ΔR between k_t subjets: k_t algorithm is used to cluster all the tracks associated to the jet, stopping the clustering at exactly two jets.

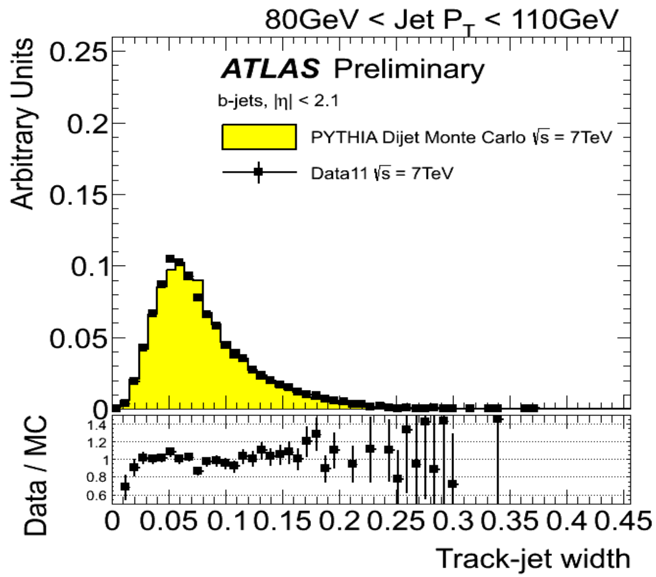
Single/Double b-hadron jets: discriminating variables

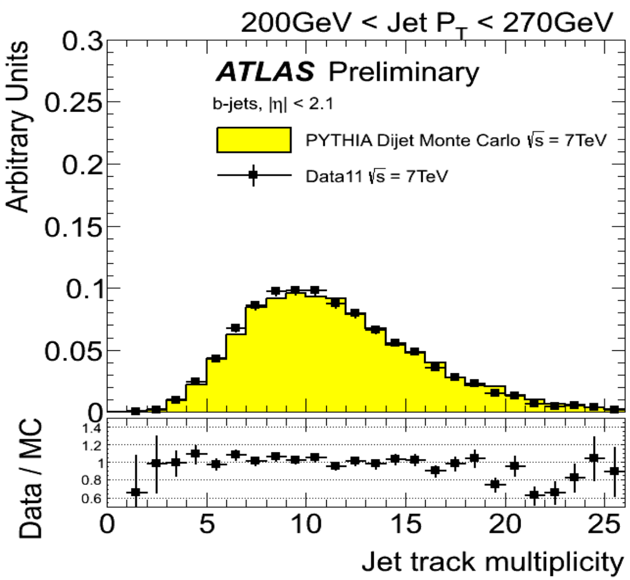


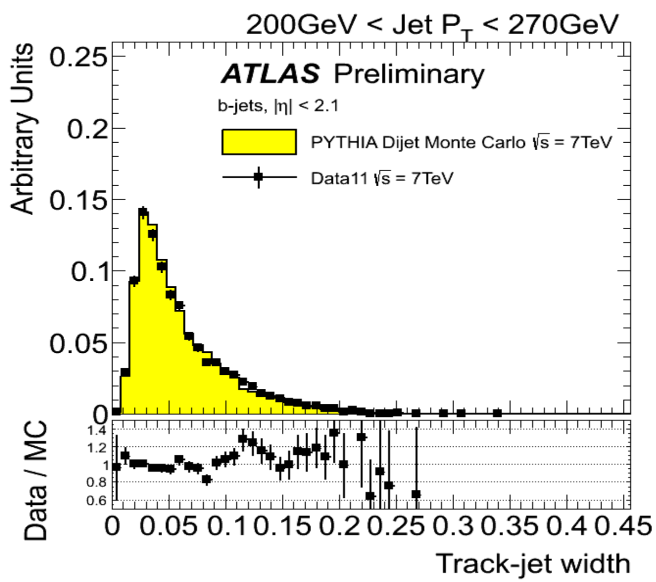
- N-subjettiness: proposed for massive boosted jet studies by **Thaler & Van Tilburg, arxiv:1011.2268v3.**
- Tau2 quantifies to what degree a jet can be regarded as composed of 2 subjets.
- For b-jets (no mass scale) tau 2 is larger for merged than for single jets, and correlated with width

Validation of discriminating variables with data







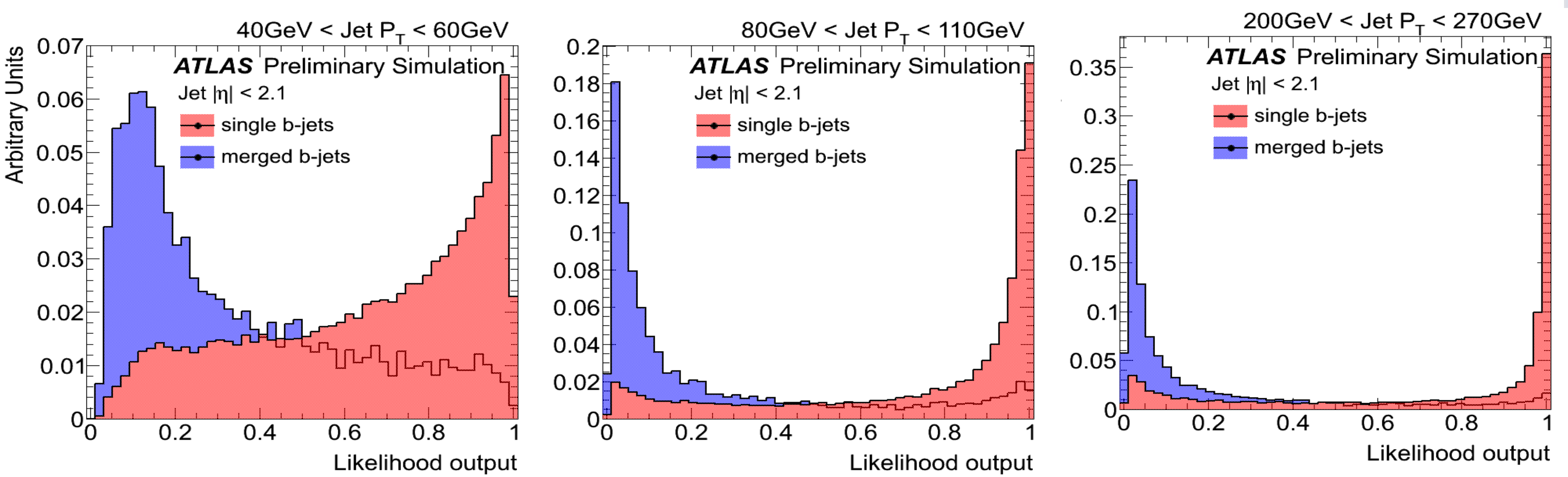


- Very good agreement between data and simulation.

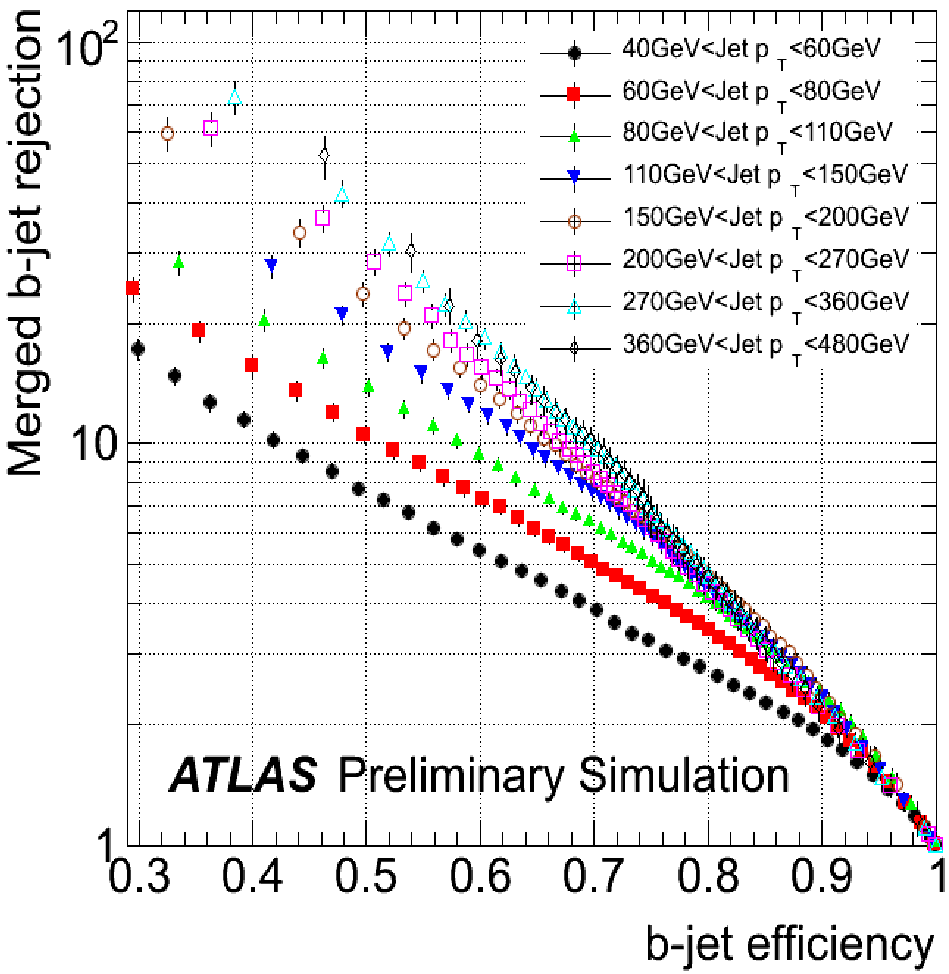
Multivariate Analysis

- A discriminant between single b-jets and merged b-jets was built training a likelihood estimator in the context of TMVA.
- After balancing discrimination power, pile-up independence and correlations, we kept three variables for our multivariate analysis:

  1. Jet track multiplicity;
  2. Track-jet width, and;
  3. ΔR between the axes of the two exclusive k_t subjets.



Multivariate Analysis: Performance



- Performance improves with Pt
  - Pt > 40GeV: rejection above 8 at 50% b-jet efficiency
  - Pt > 200GeV: rejection above 30 at 50% efficiency
- Only statistical errors are shown

Systematic uncertainties

- The following sources of systematic uncertainties were considered:
  1. The presence of additional interactions (pile-up);
  2. The b-tagging efficiency;
  3. The track reconstruction efficiency;
  4. The track transverse momentum resolution;
  5. The jet transverse momentum resolution (JER);
  6. The jet energy scale.
- The main contributions are the uncertainties in track reconstruction efficiency, jet energy scale, and jet energy resolution.
- Summary table:

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%

Summary

- A multivariate discriminant to identify *b*-tagged jets containing two *B*-hadrons was presented.
- The method exploits shape and substructure differences between single b-jets and merged b-jets, produced for instance from gluon splitting.
- The Monte Carlo distributions of the explored discriminant variables were validated using experimental data recorded by ATLAS during 2011.
- The agreement between data and simulation is excellent.
- The performance of the tagger in Monte Carlo was studied as a function of jet p_T, achieving, at 50% single b-jet efficiency, a merged b-jet rejection of over 30 (8) for $p_T > 200$ GeV ($p_T > 40$ GeV).
- This tool has applications in measurement of QCD beauty production, rejection of QCD/W+jets background in searches dominated by single b-jets and substructure studies in heavy boosted jets (Z->bb, H->bb)

Back-up slides

Jet reconstruction and calibration

- Jets are reconstructed using anti-K_T jet algorithm with R=0.4, using calorimeter topoclusters as inputs.
- Jets are calibrated using three different calibration schemes: “EM+JES”, “LCW+JES” and “GS”.
- The total uncertainty on the JES is smaller than ±10%.

Tracks & vertices

- The charged-particle tracks with pseudorapidity $|\eta|<2.5$ are reconstructed in the ID.
- Tracks with $p_T^{track} > 400$ MeV are associated in primary vertices (PV).
- The PV must be reconstructed from at least five tracks.

- Several PV can be reconstructed per event due to the presence of pile-up.
- The one with the largest $\sum_{trk}(p_T^{trk})^2$, is selected as the one associated to the hard interaction.

B-Tagging algorithms

- Algorithms to identify heavy flavor content in reconstructed jets.
- Impact parameter of tracks in jets:

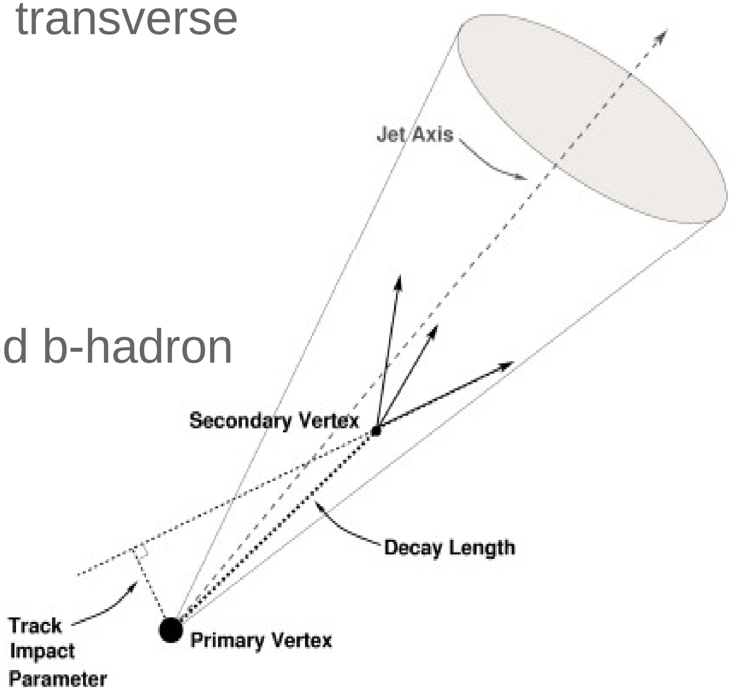
  IP3D: uses track weights based on longitudinal and transverse IP significance.
- Displaced secondary vertex:

  SV1: reconstructs inclusive displaced vertex

  JetFitter: reconstructs multiple vertices along implied b-hadron line of flight.
- Advanced NN taggers:

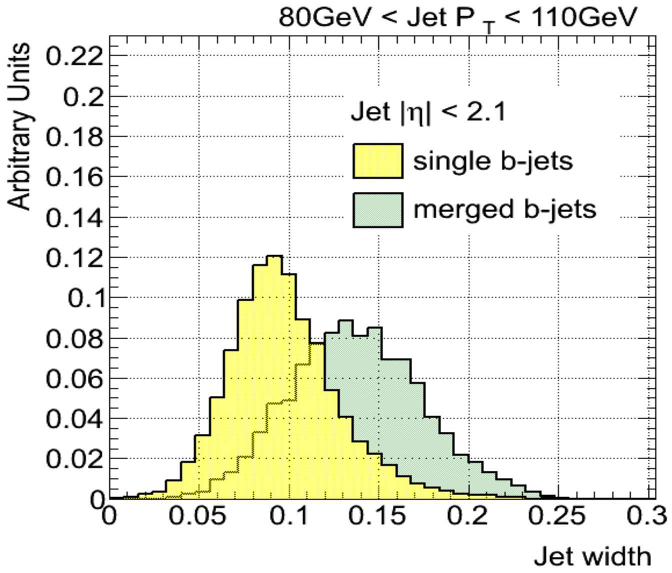
  JetFitterCOMBNN: IP3D+JetFitter

  MV1: IP3D+JetFitter+SV1

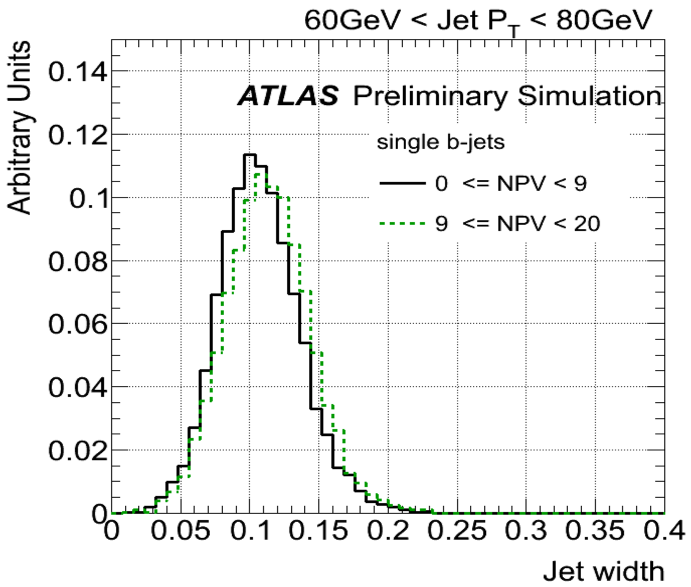


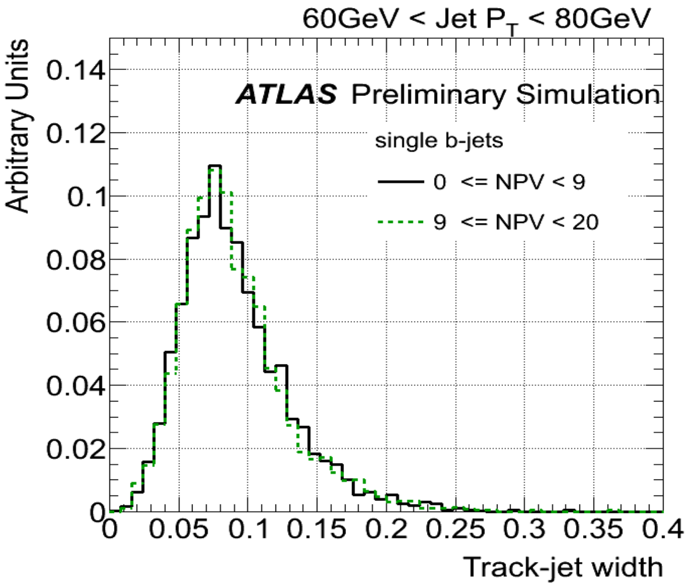
- The b-tagging performance is determined using a simulated tt sample and is calibrated using experimental data with jets containing muons and with a sample of tt events.

Single/Double b-hadron jets: discriminating variables

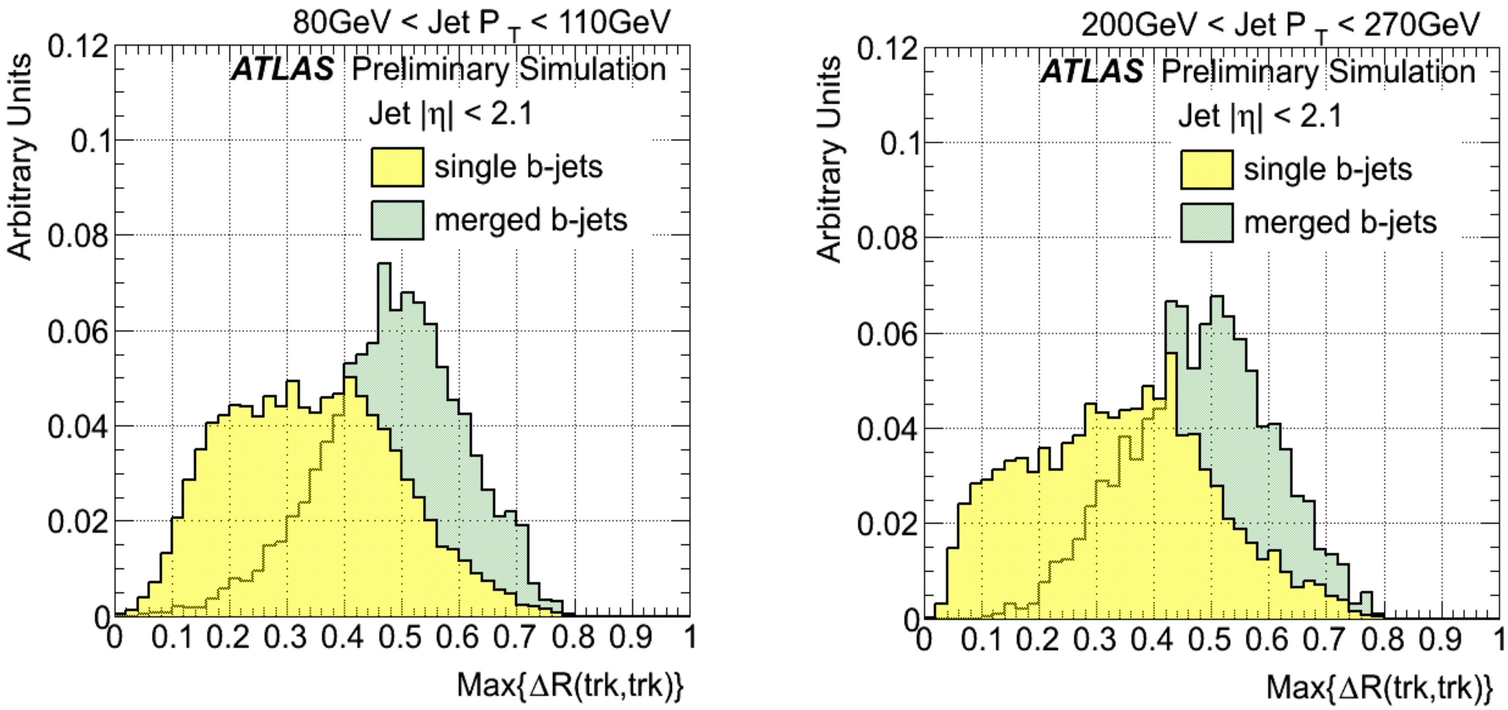


- Jet width: uses calorimeter constituents (topoclusters) instead of the associated tracks.
- It provides very good discrimination, but it is more sensitive to the amount of pile-up in the event than its track-based counterpart.



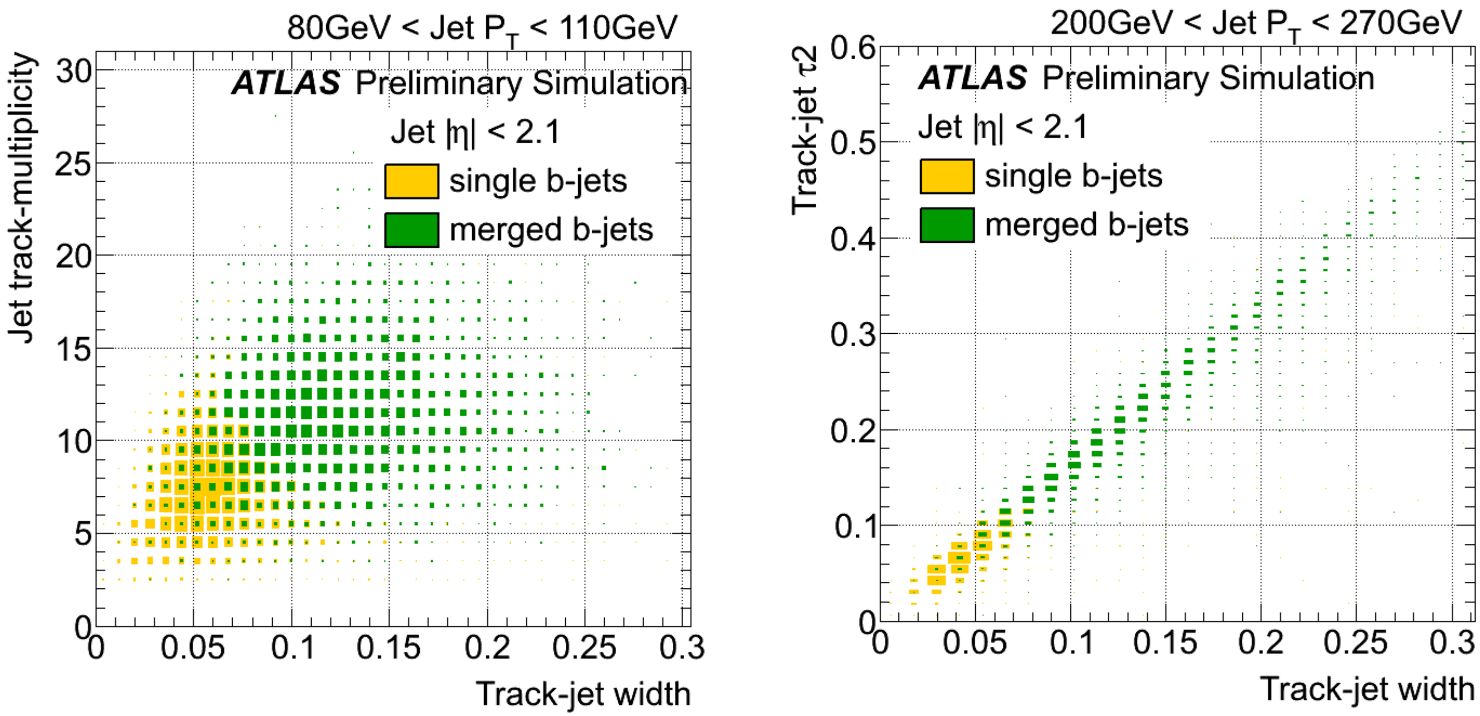


Single/Double b-hadron jets: discriminating variables



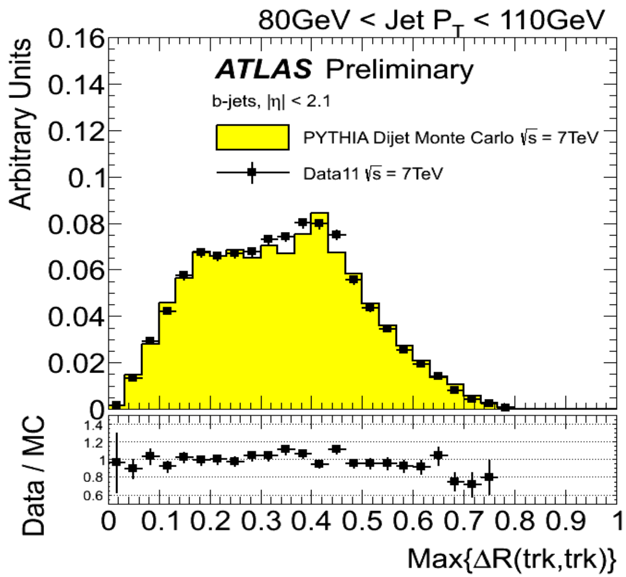
- Max{ΔR(trk,trk)}: maximum distance in the η–ϕ plane between track pairs in the jet.
- Although it shows good discrimination between single and merged b-jets, we looked for alternatives to Max{ΔR(trk,trk)} as it is not an infrared safe observable and sensitive to the presence of non-relevant soft tracks in the jet periphery.

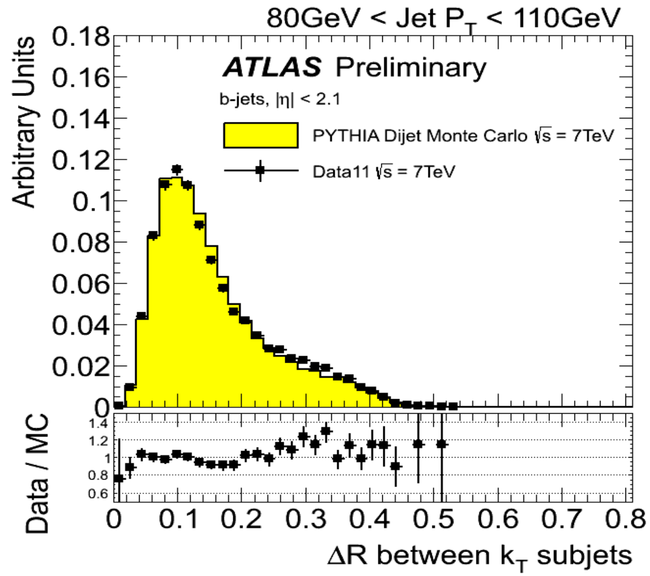
Single/Double b-hadron jets: correlations

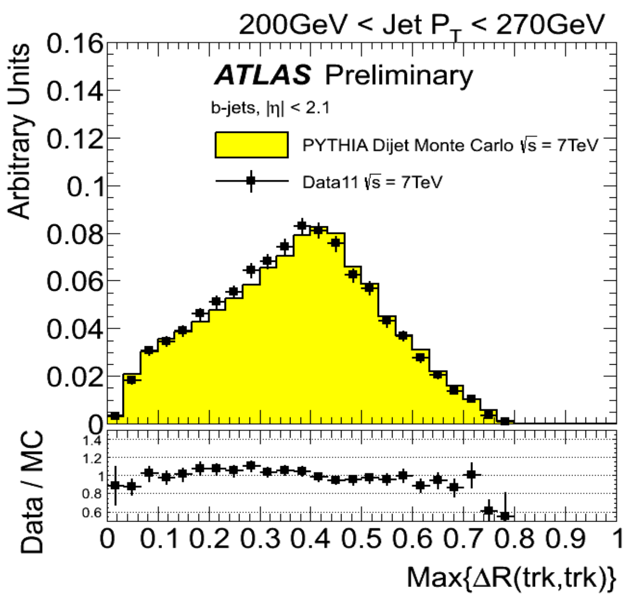


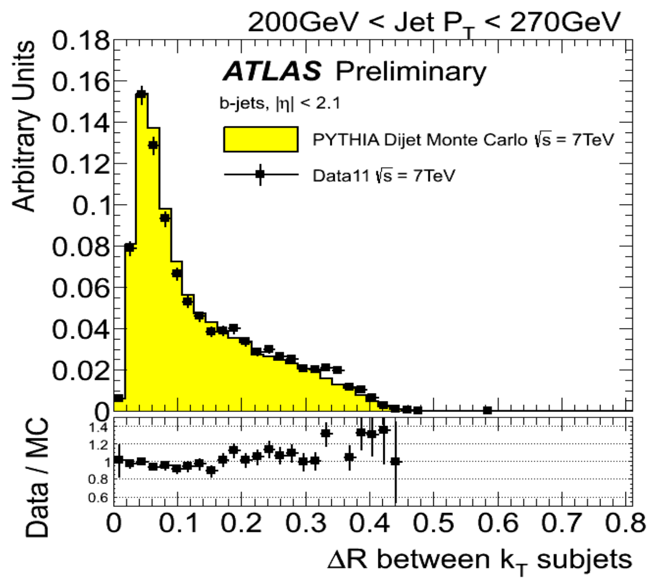
- Studied correlation between variables to avoid using strongly correlated pairs, as illustrated here for τ_2 and track-jet width.

Validation of discriminating variables with data



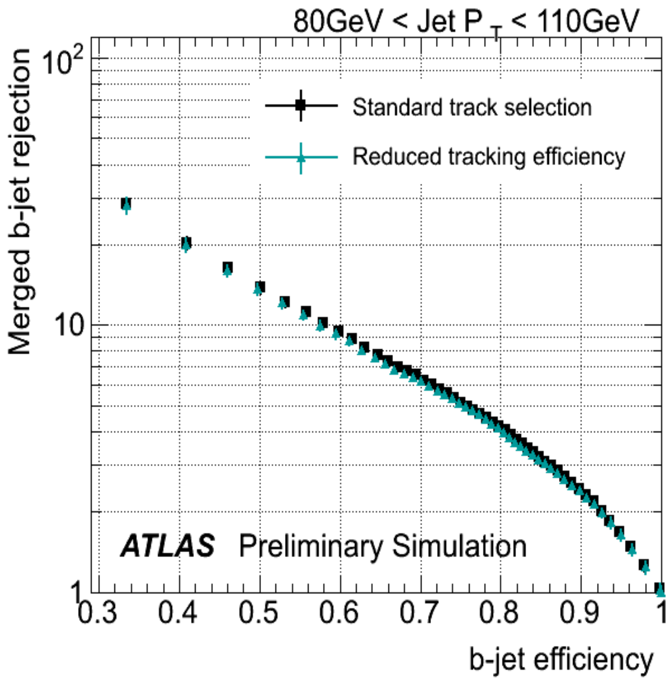


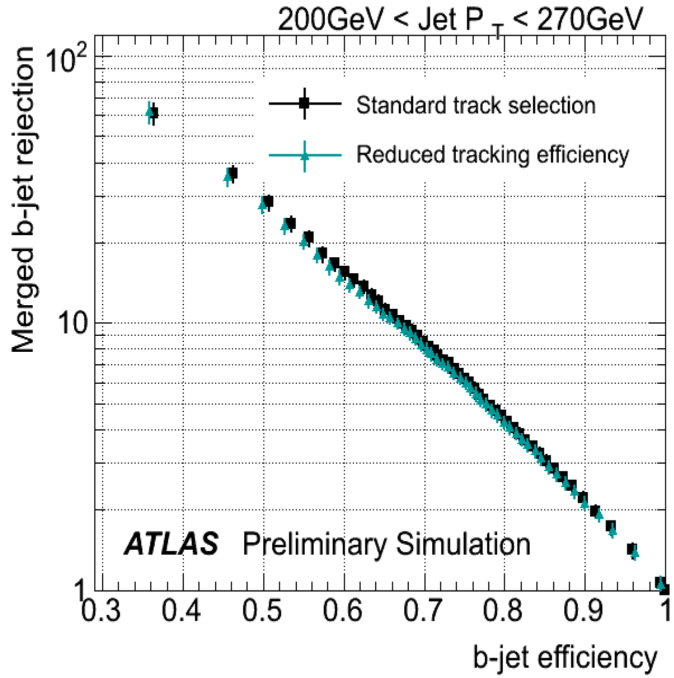




- Very good agreement between data and simulation.

3. Uncertainty in the track reconstruction efficiency





- An uncertainty arises from our limit in the understanding of the ID material layout
- To test its impact a fraction of tracks determined from the track efficiency uncertainty was randomly removed.
- A systematic degradation of the performance of 4% is present over all pt bins / 2 working points considered.

Conclusions

- We have presented results on the study of the properties of light-quark and gluon jets.
- JES differences can be large, especially at low pT.
- Track-based variables provide good discrimination in Monte Carlo.
    - Suggests potential for JES corrections, tagger, etc.
- Data/MC disagreement required use of data-driven techniques: extract pure templates from γ+jet and dijet samples.
- Template method has good closure in MC, but poor performance in data.
- “Purified” samples provide alternative, possibly pure samples of q/g in data: confirm template results.