Jet reconstruction and jet flavour tagging at LHCb

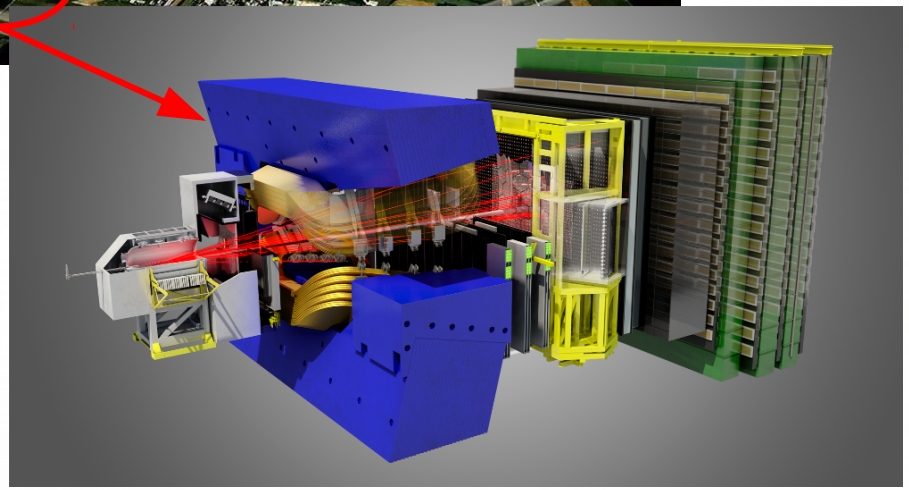
Lorenzo Sestini
Università di Padova and INFN

On behalf of the LHCb collaboration

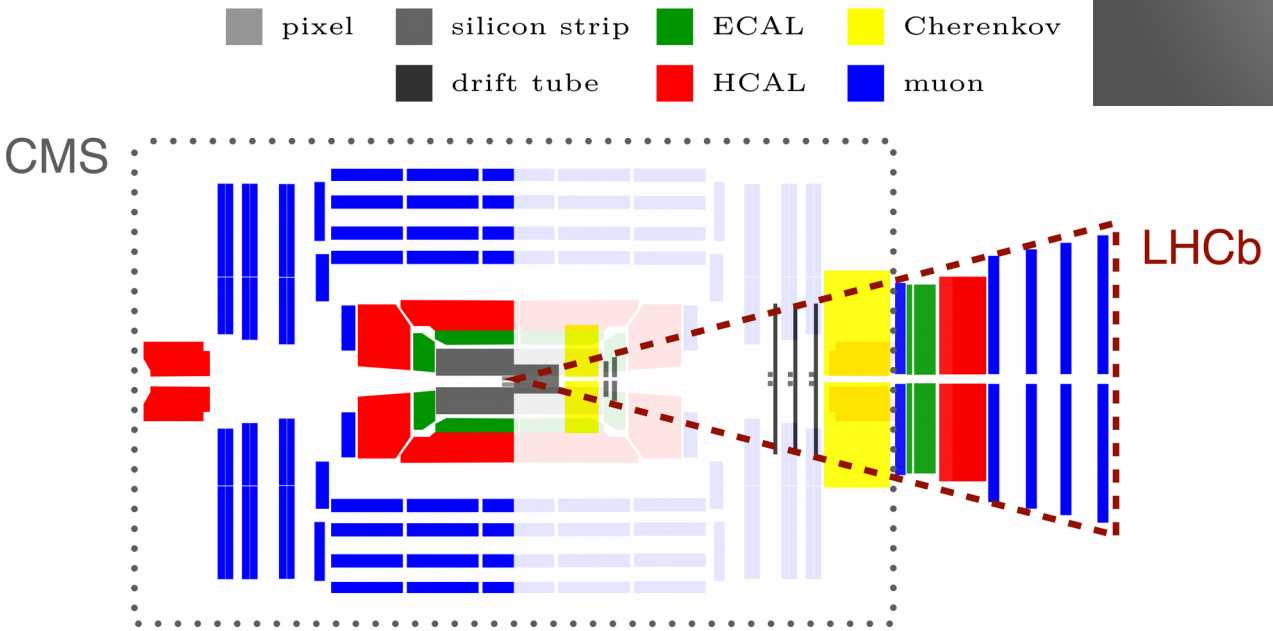
IML Machine Learning Workshop, 21-3-2017, Cern, Geneve

LHCb detector

Int. J. Mod. Phys. A 30, 1530022 (2015)

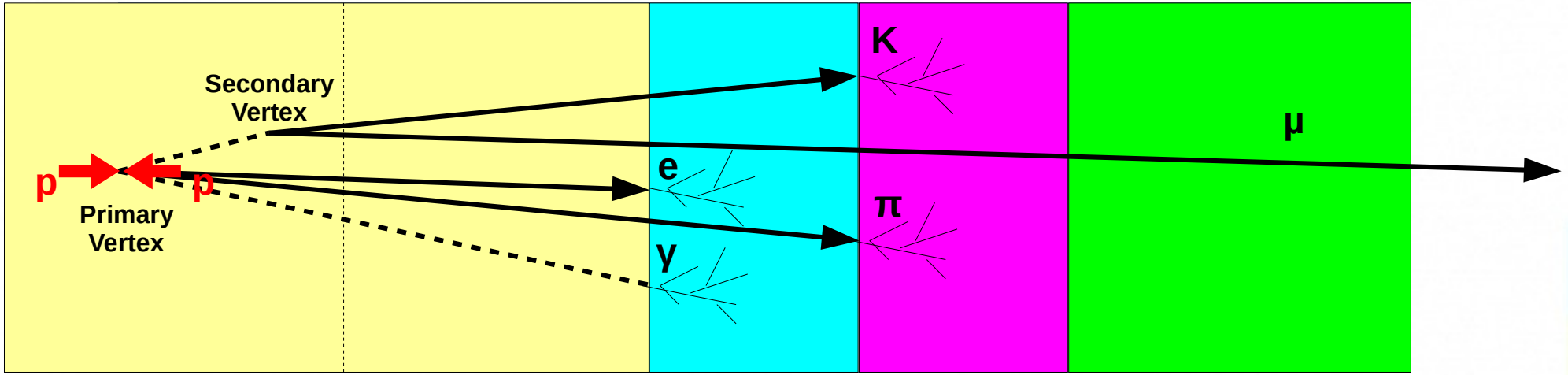


- **LHCb is a spectrometer initially designed to study b and c hadrons physics.**
- **It covers a phase space region of p-p collisions complementary to ATLAS and CMS, corresponding to $2 < \eta < 5$.**

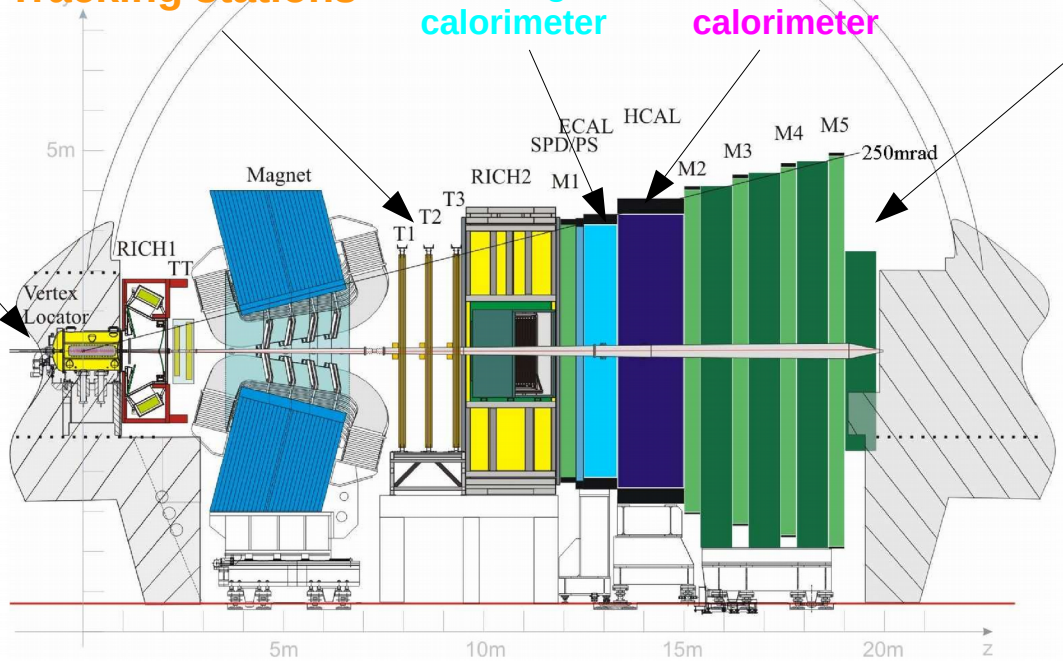


In the last years LHCb has demonstrated its capability in jets physics!

Jets detection at LHCb

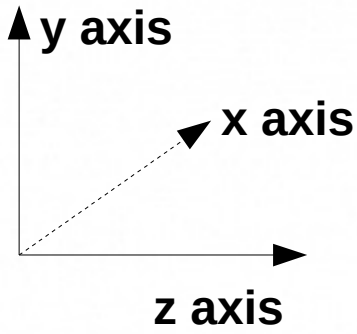


Vertex Locator | Tracking stations | Electromagnetic calorimeter | Hadronic calorimeter | Muon System



Jets reconstruction inputs:

- **Tracks**
- **Calorimeter clusters**

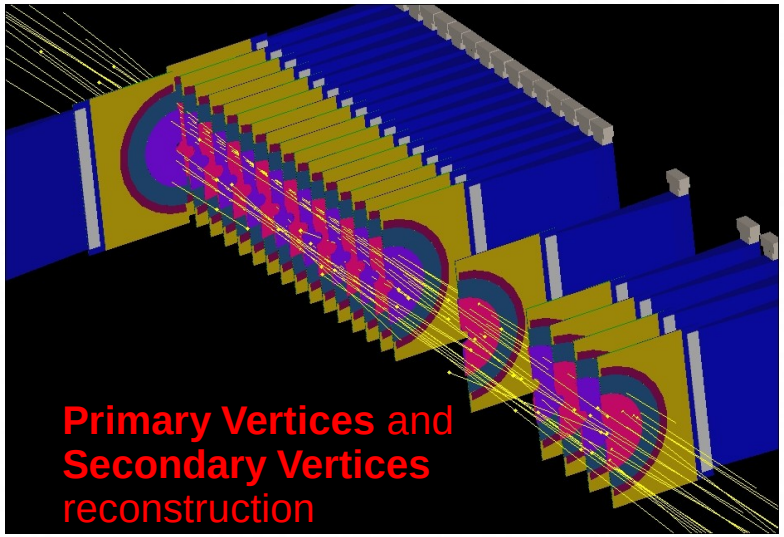


Tracking system

Tracking at LHCb: **silicon microstrip** (VELO, Inner Tracker), **drift tubes** (Outer tracker)

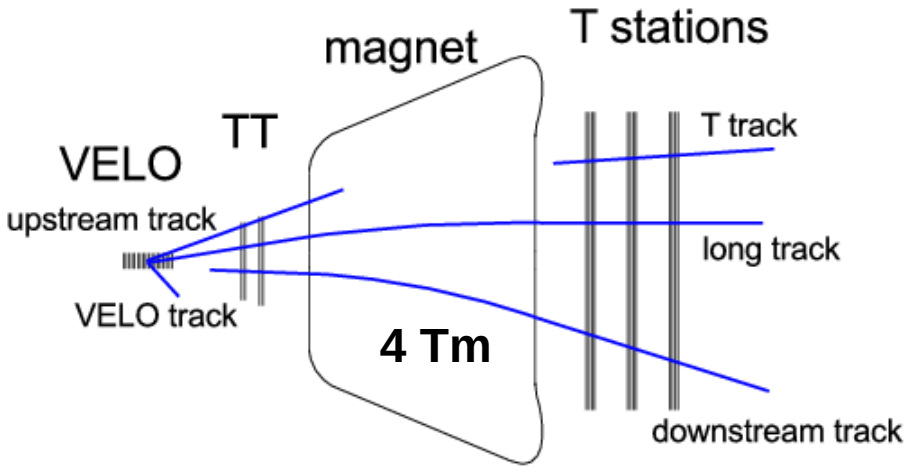
VErtex LOcator (VELO)

21 stations → (r,φ) coordinates

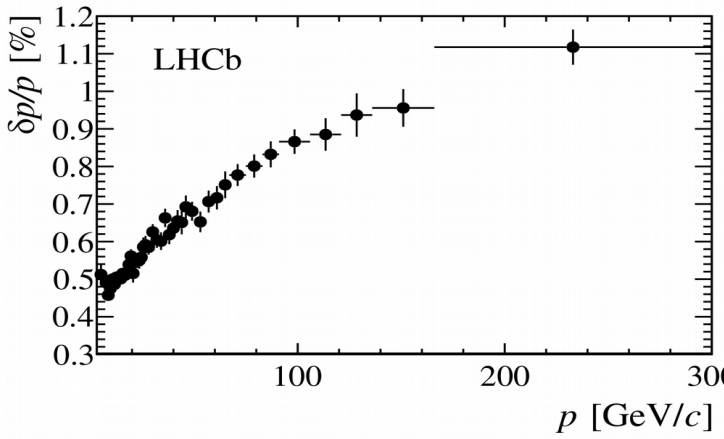


Tracking stations

4 stations → (x,y) coordinates



The **momentum** of **charged particles** is determined by measuring **the curvature of the trajectory in the magnetic field**

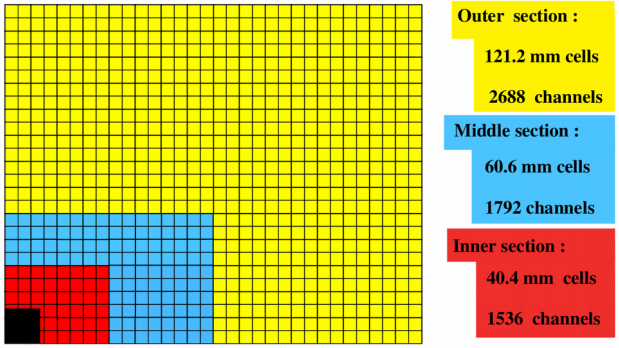


Calorimeters

Electromagnetic calorimeter

e, γ, π^0 produce electromagnetic showers in lead layers

showers are detected by layers of scintillating fibers

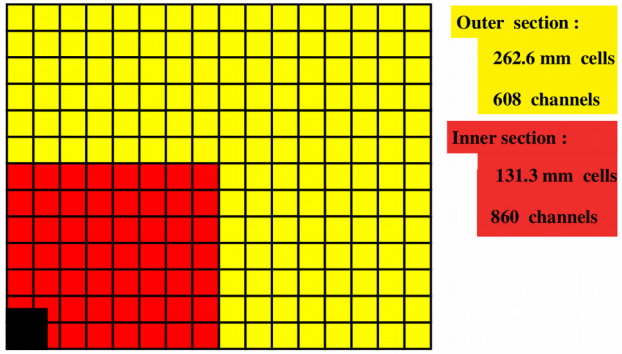


$$\frac{\sigma_E}{E} = \frac{10\,\%}{\sqrt{E}} \oplus 1\,\%$$

Hadronic calorimeter

K, π and other hadrons produce hadronic showers in iron layers

showers are detected by layers of scintillating tiles



$$\frac{\sigma_E}{E} = \frac{69\,\%}{\sqrt{E}} \oplus 10\,\%$$

Limitations due to saturation

Energy resolution

Not optimal for jets physics!

Inputs for jets reconstruction

Clusters isolated from tracks (neutral particles)

Excesses of energy nearby tracks (neutral recovery)

Jet reconstruction algorithm

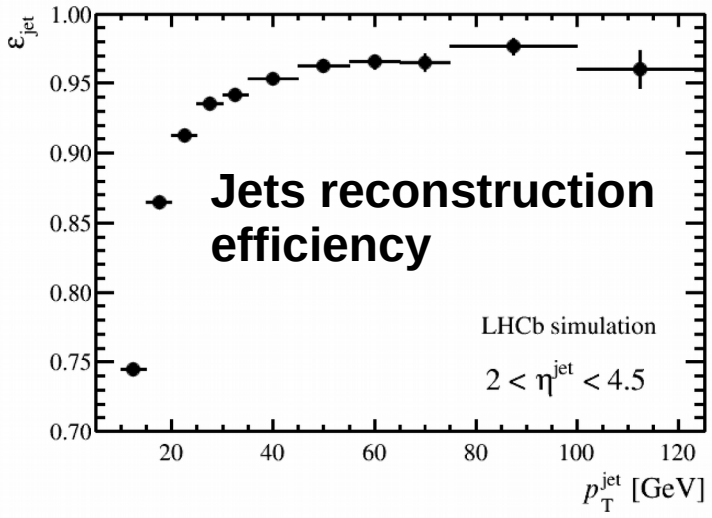
JHEP 01 (2014) 033

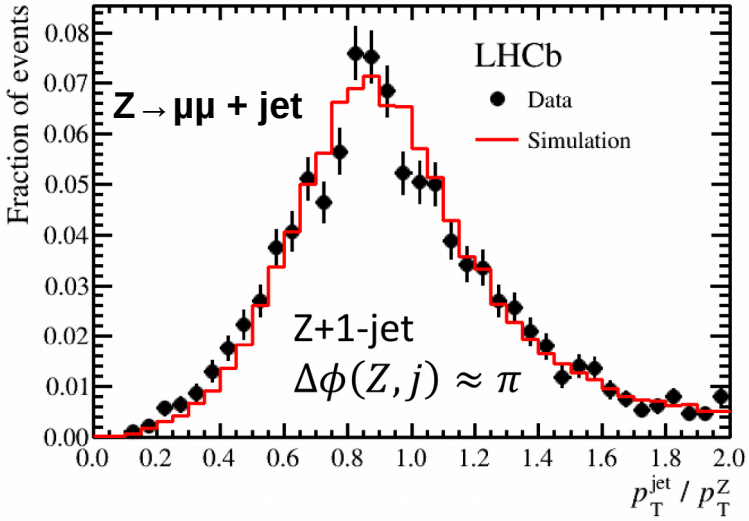
Particle Flow → **Tracks and calorimeter clusters are selected as inputs**

Anti-kt clustering → **Inputs are used to form jets of a given cone radius R=0.5**

Jet four-momentum → $E_{jet} = \sum_i E_i \quad \vec{p}_{jet} = \sum_i \vec{p}_i$

Jet energy correction → **The jet energy is corrected for detector inefficiencies (calibrated on simulation)**



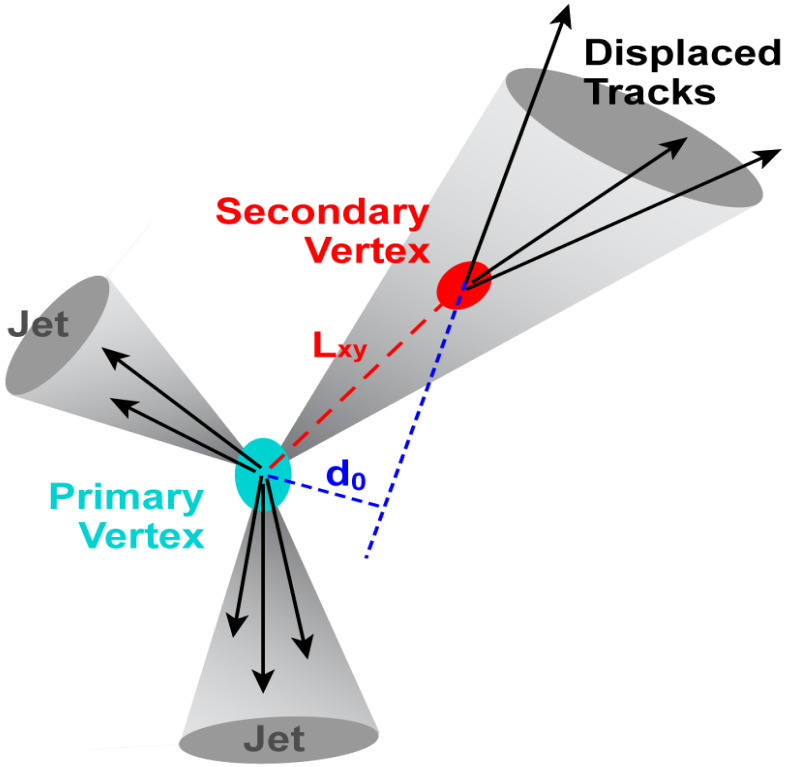


Energy resolution of final jets δE/E ≈ 10-15%

Jet tagging at LHCb

JINST 10 (2015) P06013

- **The jet tagging system takes advantage of LHCb features → precise vertex reconstruction!**
- A jet is identified to be generated from a **b** or **c** quark (**b-jet** or **c-jet**) if a **Secondary Vertex** is reconstructed within the jet cone (**ΔR <0.5**).
- Single tracks used to build the **Secondary Vertex** are **not required** to have **ΔR <0.5** with respect to the jet axis.
- Two **Boosted Decision Trees** are used to identify b and c jets.

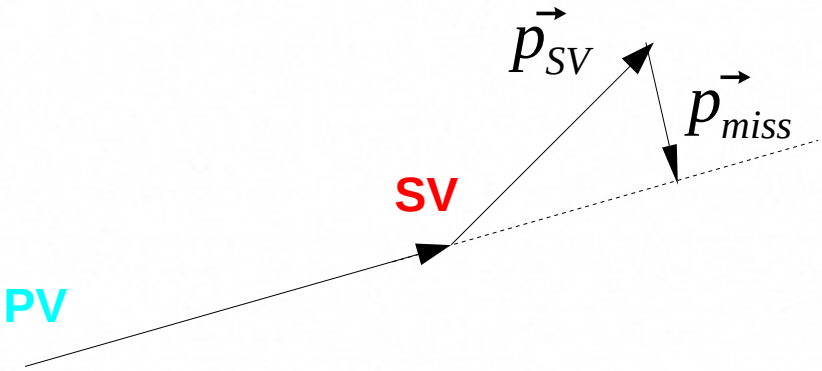


BDT(bc|udsg)
To separate **heavy flavour** jets from **light** jets

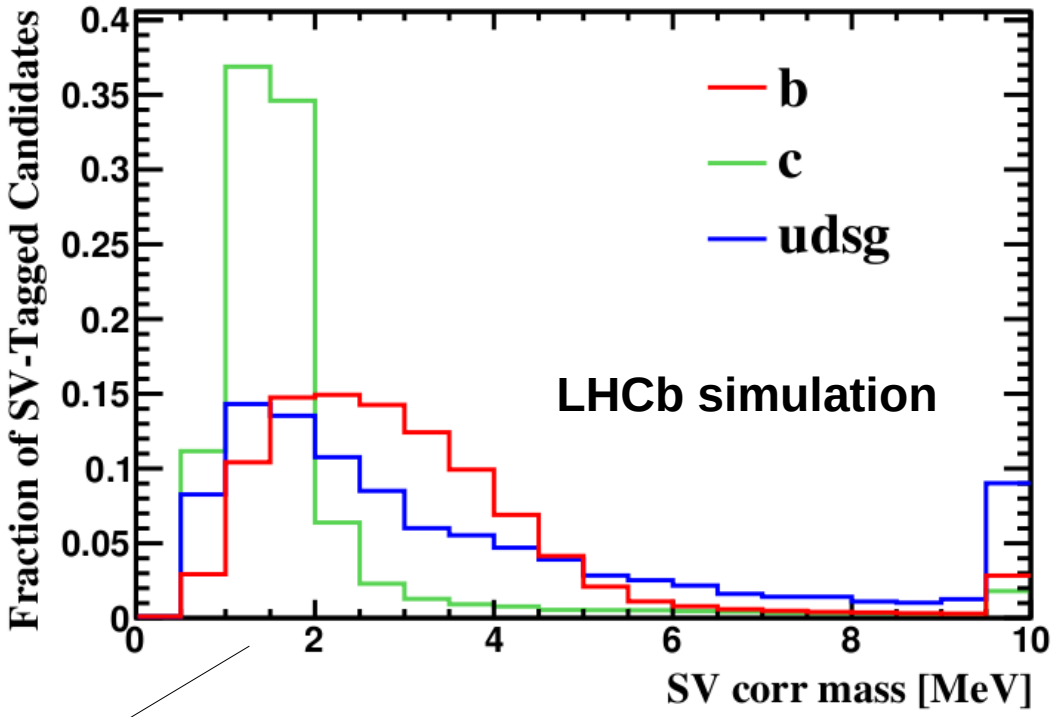
BDT(b|c)
To separate **b-jets** from **c-jets**

Jet tagging at LHCb

- **Some observables in input to the BDTs:**
  - **SV mass**
  - **SV corrected mass**
  - **Flight distance χ^2**
  - **Fraction of jet p_T taken by the SV**



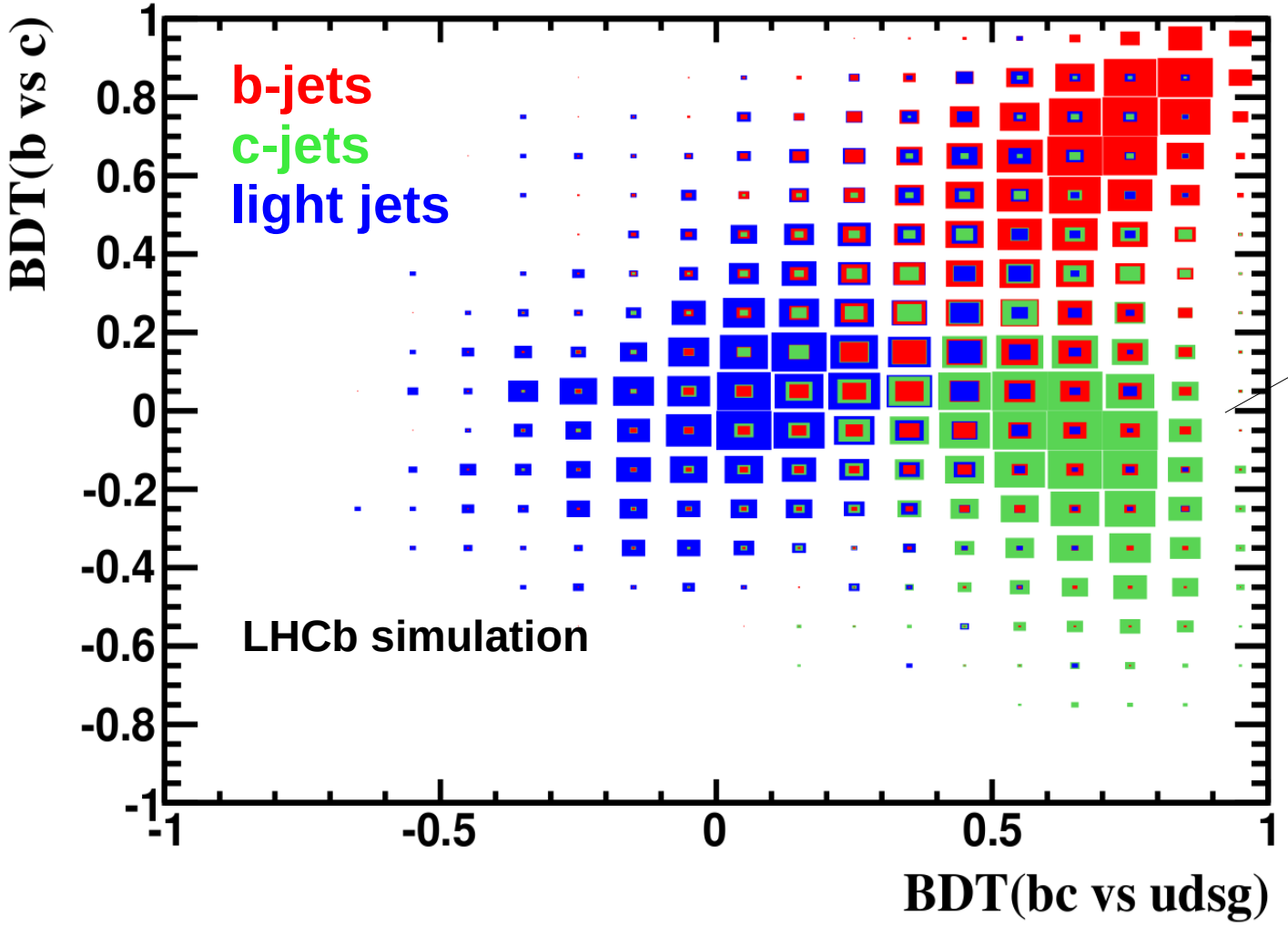
$$M_{corr} = \sqrt{M_{SV}^2 + p_{miss}^2} + p_{miss}$$



Useful to discriminate b from c

Jet tagging at LHCb

- Training samples of **b-jets**, **c-jets** and **light jets** are obtained from the Monte Carlo simulation.
- **A good discrimination power is achieved!**



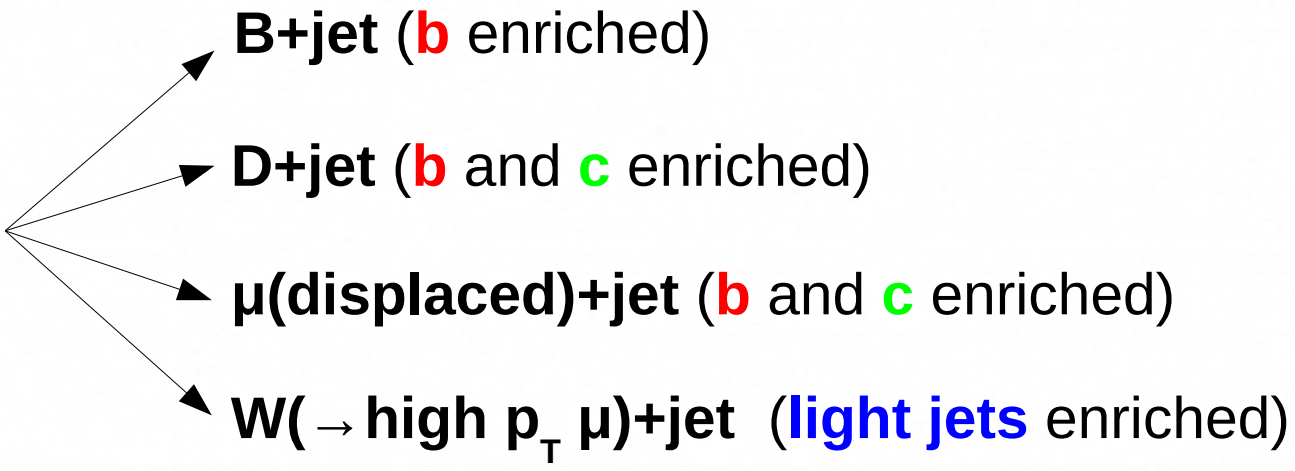
Distributions normalized to 1

Jet tagging at LHCb

- **Tagging efficiencies are measured in data.**

- Events with a **jet** and a **probe** back-to-back to the jet in the azimuthal plane are selected.

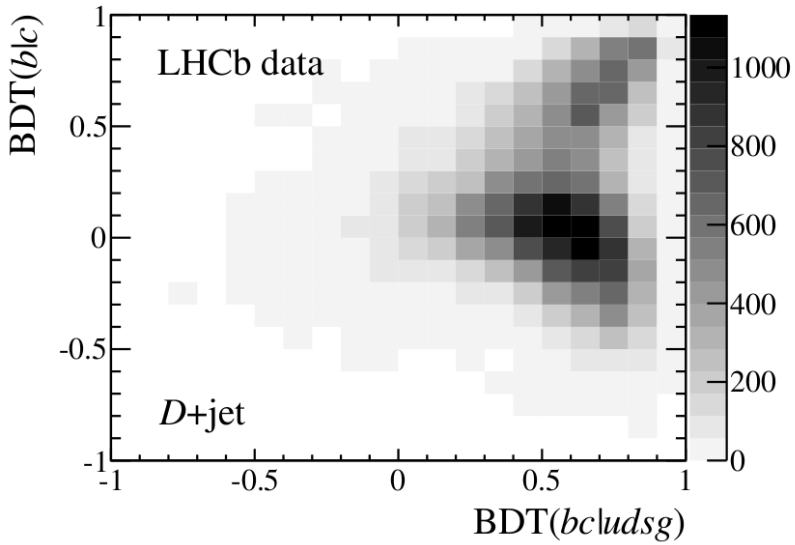
- The **probe** identifies the jet flavour

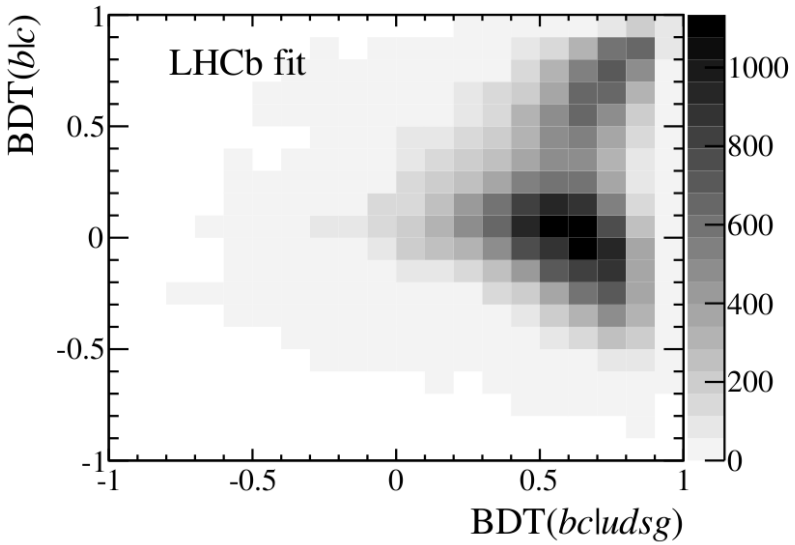


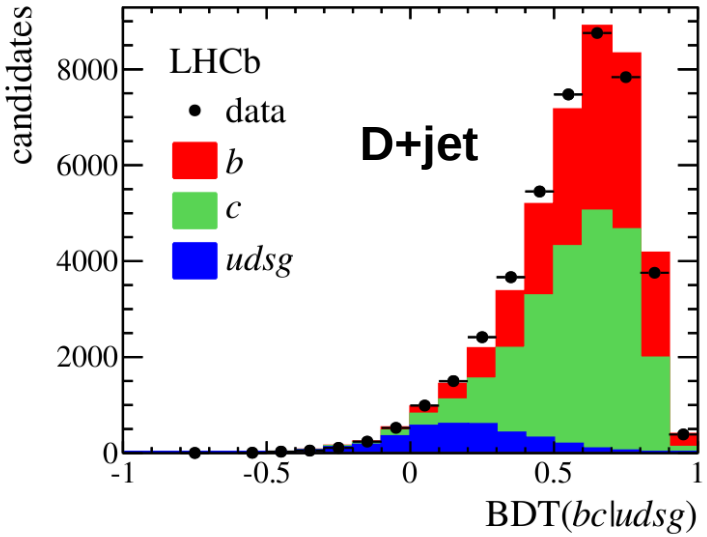
- Yields of **b**, **c** and **light** jets with a SV-tag are measured with a **two-dimesional templates fit to the BDTs distributions**.

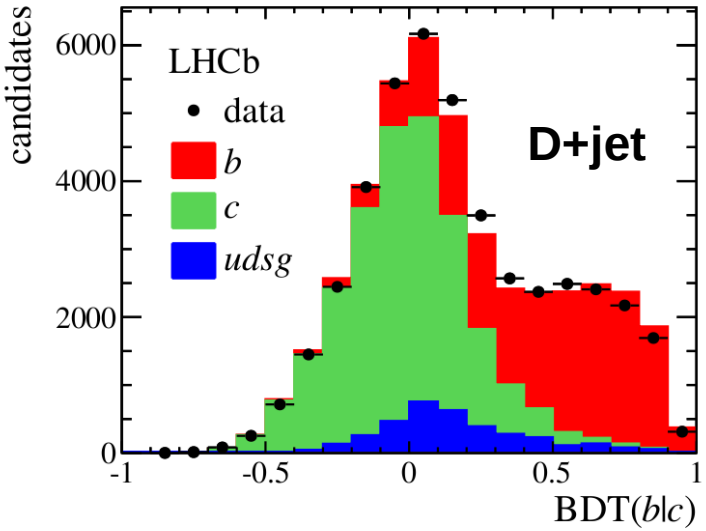
- **Two-dimensional templates are obtained from simulation.**

Jet tagging at LHCb





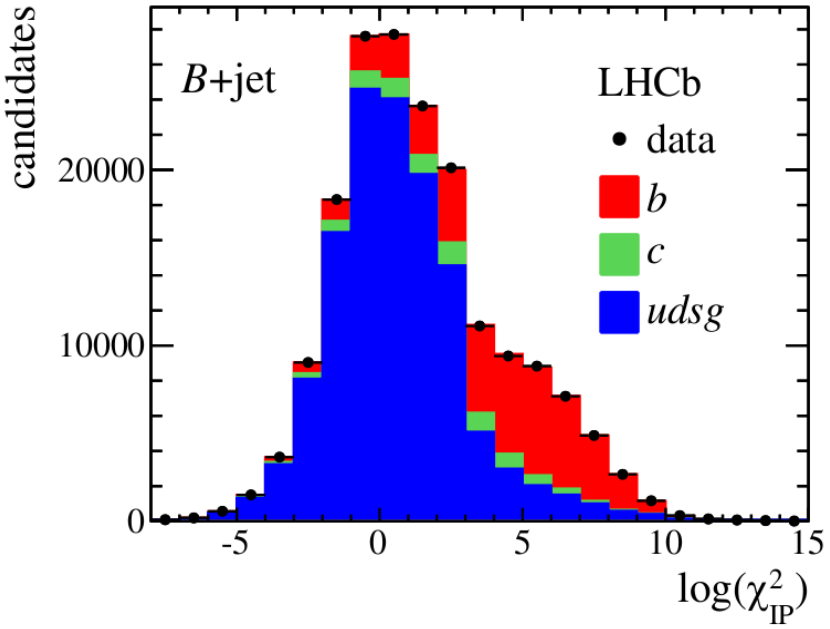


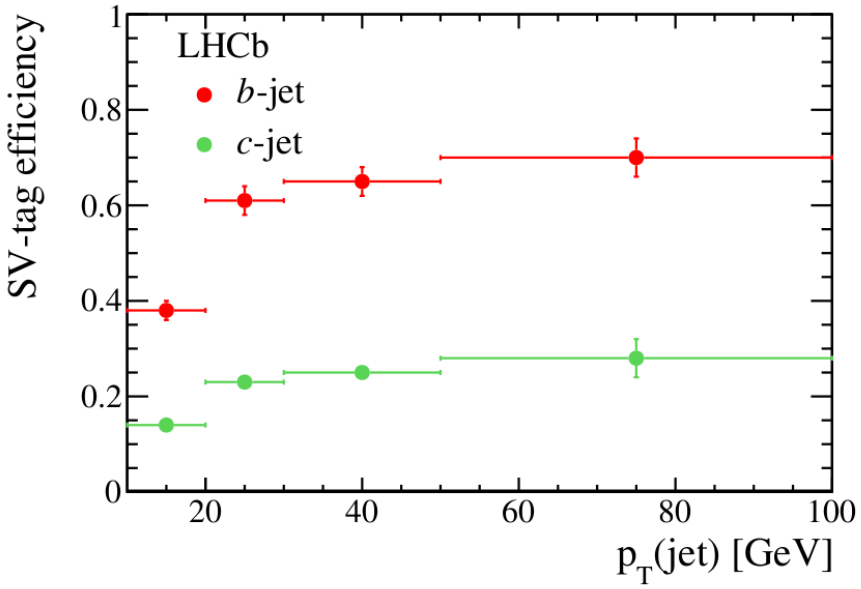


Jet tagging at LHCb

- Yields of **b**, **c** and **light** jets **prior to apply the SV-tag** are measured by fitting the distribution of the χ^2_{IP} associated to the highest p_T tracks in the jet.

- Efficiencies obtained with: $\epsilon = \frac{N(SV)}{N(\chi^2_{IP})}$





Probability for a **b-jet** to be selected ~ **65%**

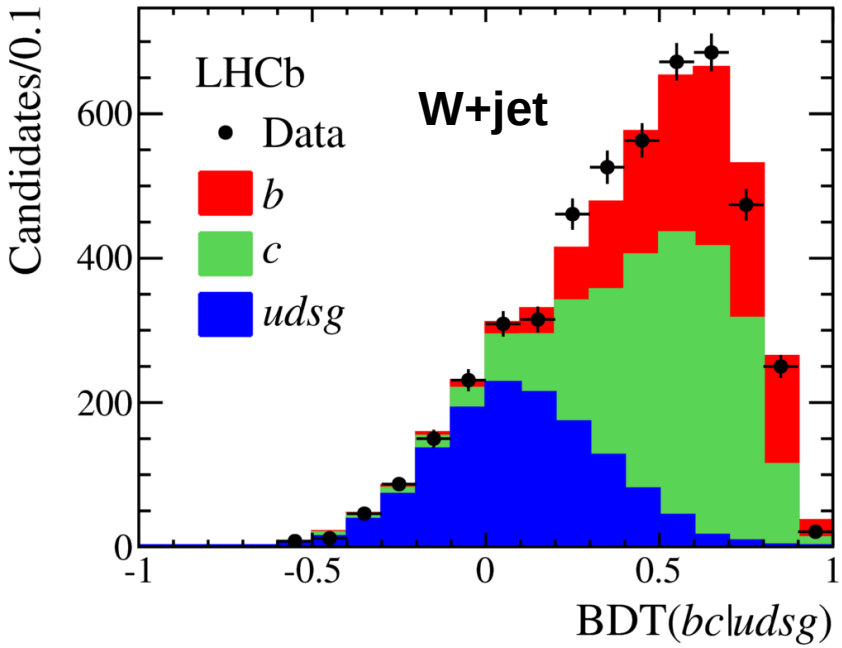
Probability for a **c-jet** to be selected ~ **25%**

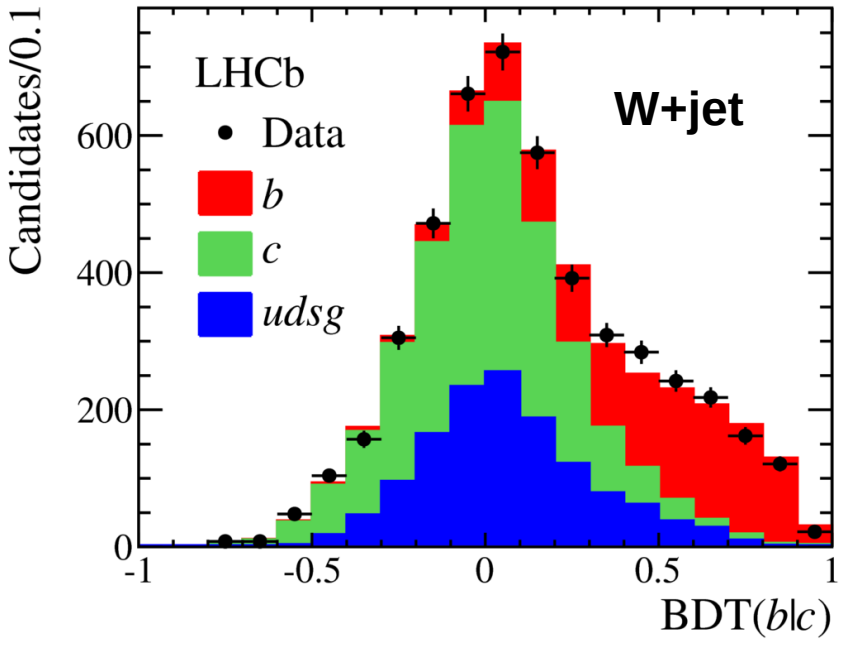
Probability to wrongly select a **light jet (g,u,d,s) ~ 0.3%**

Measurement of forward W+b and W+c at 8 TeV

Phys. Rev. D 92, 052001 (2015)

- **Application of jets reconstruction and heavy flavour tagging at LHCb.**
- Events with one **SV-tagged jet** and **one high energy muon** (**to identify the W decay**) are selected.
- The composition of the **W+jet sample** is determined by fitting the **BDT(bc|udsg)** and **BDT(b|c)** distributions.



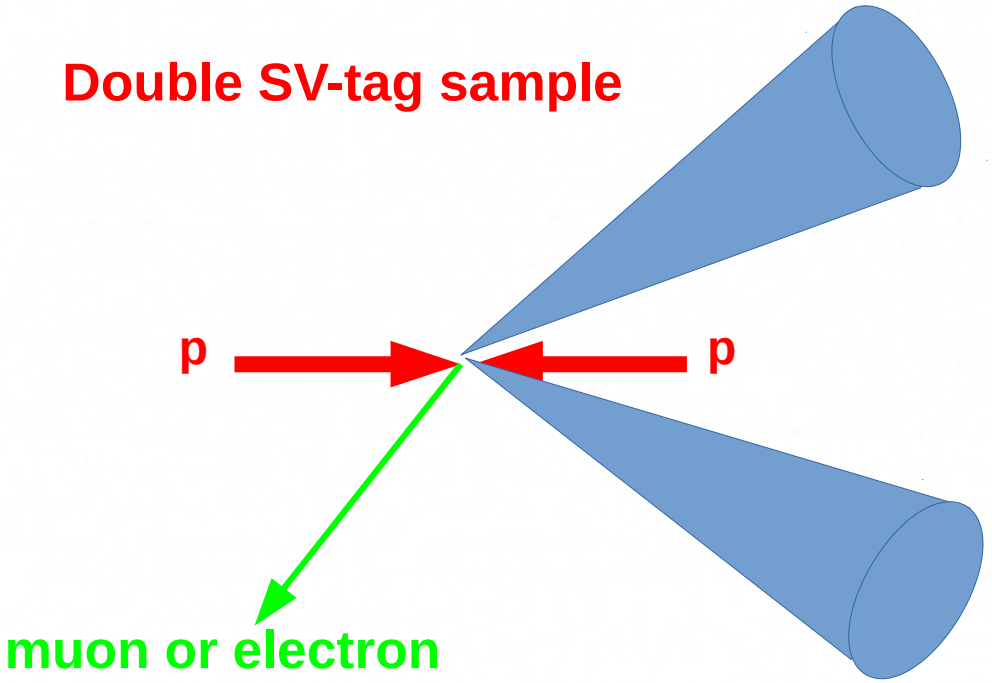


- **All the results are in agreement with the Standard Model prediction.**

Measurement of forward $W+b\bar{b}$, $W+c\bar{c}$ and $t\bar{t}$ at 8 TeV

Phys. Lett. B767 (2017) 110

- Measurement of **$W+b\bar{b}$**, **$W+c\bar{c}$** and **$t\bar{t}$** cross sections the forward region.
- **First LHCb measurement with the double SV-tag sample.**



Uniform Gradient Boost for BDT

A. Rogozhnikov et al. JINST 10 (2015) T03002

- **W+$b\bar{b}$/$t\bar{t}$ separation obtained with a BDT (top decays in b+W).**
- **Uncorrelation with the dijet invariant mass is required**

↓

Necessary to reduce systematics in the simultaneous (mass,BDT) fit

- The BDT is trained with the **Uniform Gradient Boost technique.**

At each step of the training, the weights of the trees are determined by minimizing a loss function

↓

$$L_{loss} = L_{ada} + \alpha L_{flat}$$

AdaBoost function for the discrimination

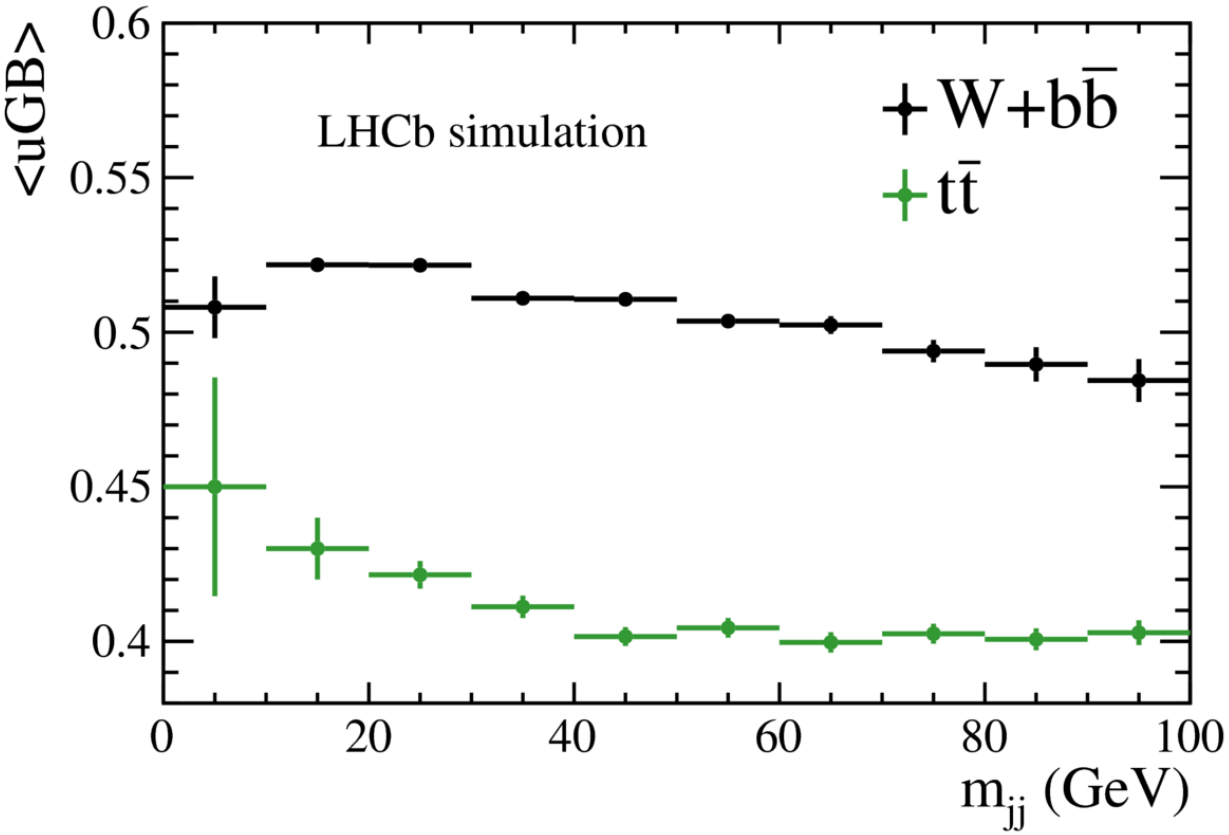
parameter to be tuned

flatness function for the uncorrelation with observables NOT in input to the BDT

- **W+$b\bar{b}$** and **$t\bar{t}$** Monte Carlo samples are used in the training.

Uniform Gradient Boost for BDT

- **Some observables in input to the BDT**
  - **lepton transverse momentum**
  - **jets transverse momenta**
  - **jets masses**
  - **lepton pseudorapidity**



Average uGB response in different intervals of dijet invariant mass

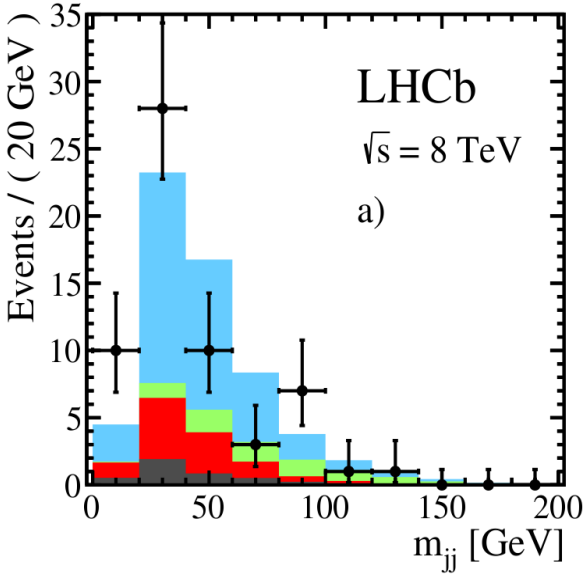
↓

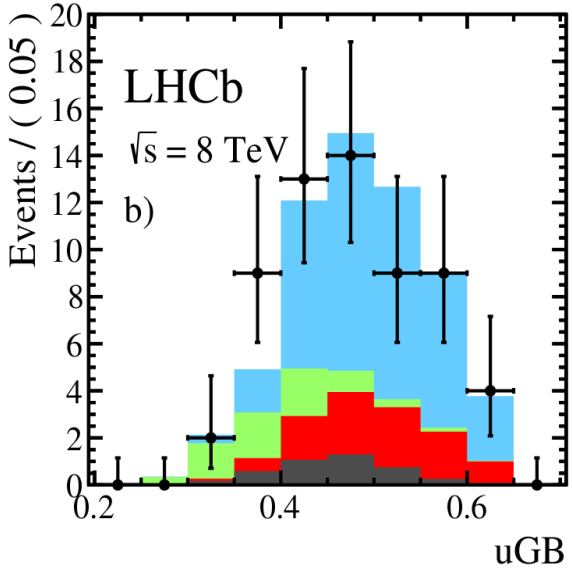
The uncorrelation is achieved to reduce systematics in the final fit

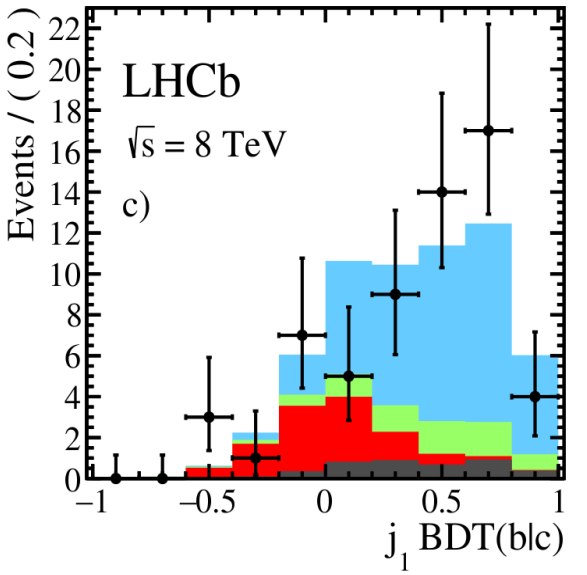
Repository with the BDT reweighter https://github.com/anaderi/lhcb_trigger_ml

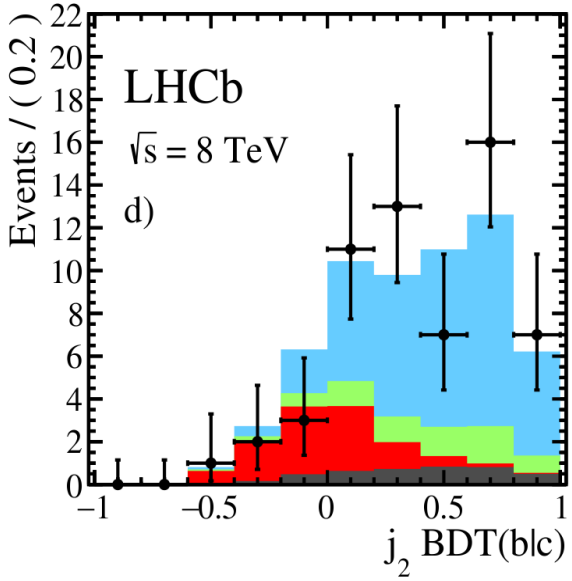
Measurement of forward $W+b\bar{b}$, $W+c\bar{c}$ and $t\bar{t}$ at 8 TeV

μ^+ sample

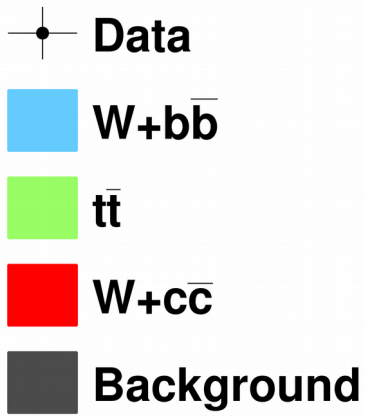




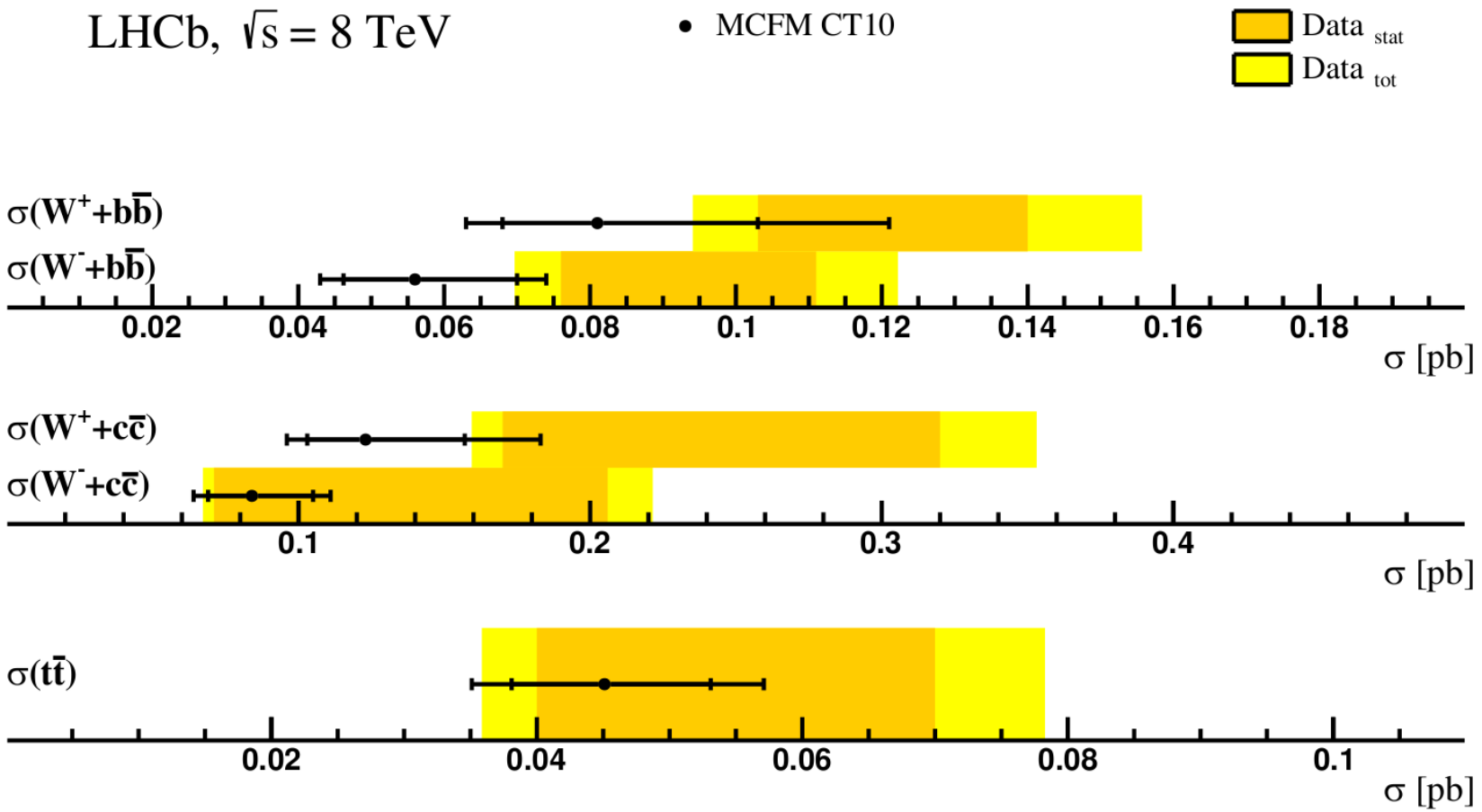




- The data sample is splitted in **4 sub-samples (μ^+, μ^-, e^+, e^-)** that are fitted simultaneously.
- $W^+b\bar{b}$, $W^-b\bar{b}$, $W^+c\bar{c}$, $W^-c\bar{c}$ and $t\bar{t}$ **normalization factors with respect to SM prediction** are free parameters.



Measurement of forward $W+b\bar{b}$, $W+c\bar{c}$ and $t\bar{t}$ at 8 TeV



First $W+c\bar{c}$ observation

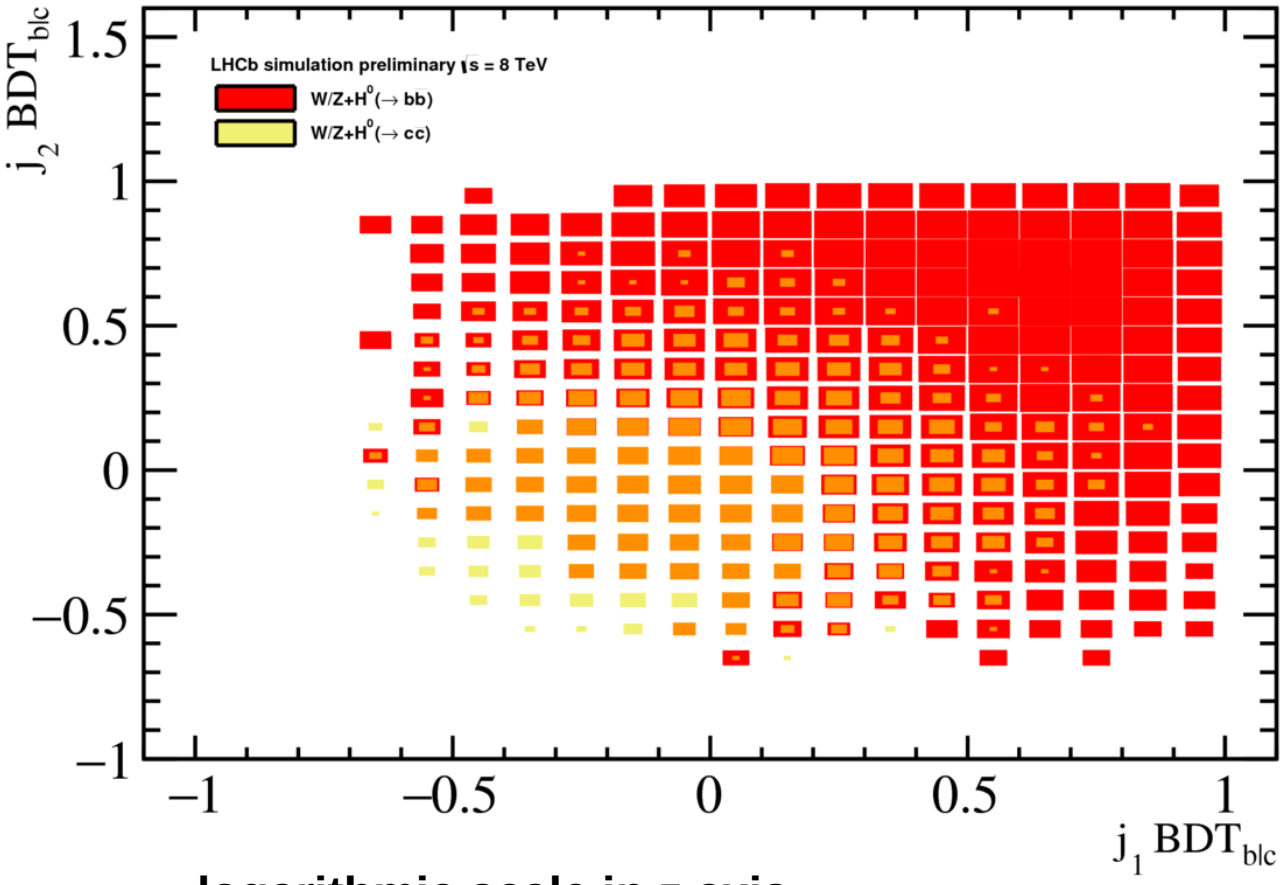
Great opportunities with c tagging at LHCb!

The measured cross sections are compatible with the Standard Model predictions within the errors.

Search for H→b$\bar{b}$ and H→c$\bar{c}$ in association with a W or Z in the forward region (8 TeV)

LHCb-CONF-2016-006

- **W/Z + H → (b$\bar{b}$/c$\bar{c}$)** candidates selection: **two SV-tagged jets and one high energy lepton.**



logarithmic scale in z axis

H → b$\bar{b}$ vs **H → c$\bar{c}$** using **BDT(b|c)**

Extra cuts in H → c$\bar{c}$ search:

- **Jet$_1$ BDT(b|c) < 0.2**
- **Jet$_2$ BDT(b|c) < 0.2**

Normalized to Standard Model yields.

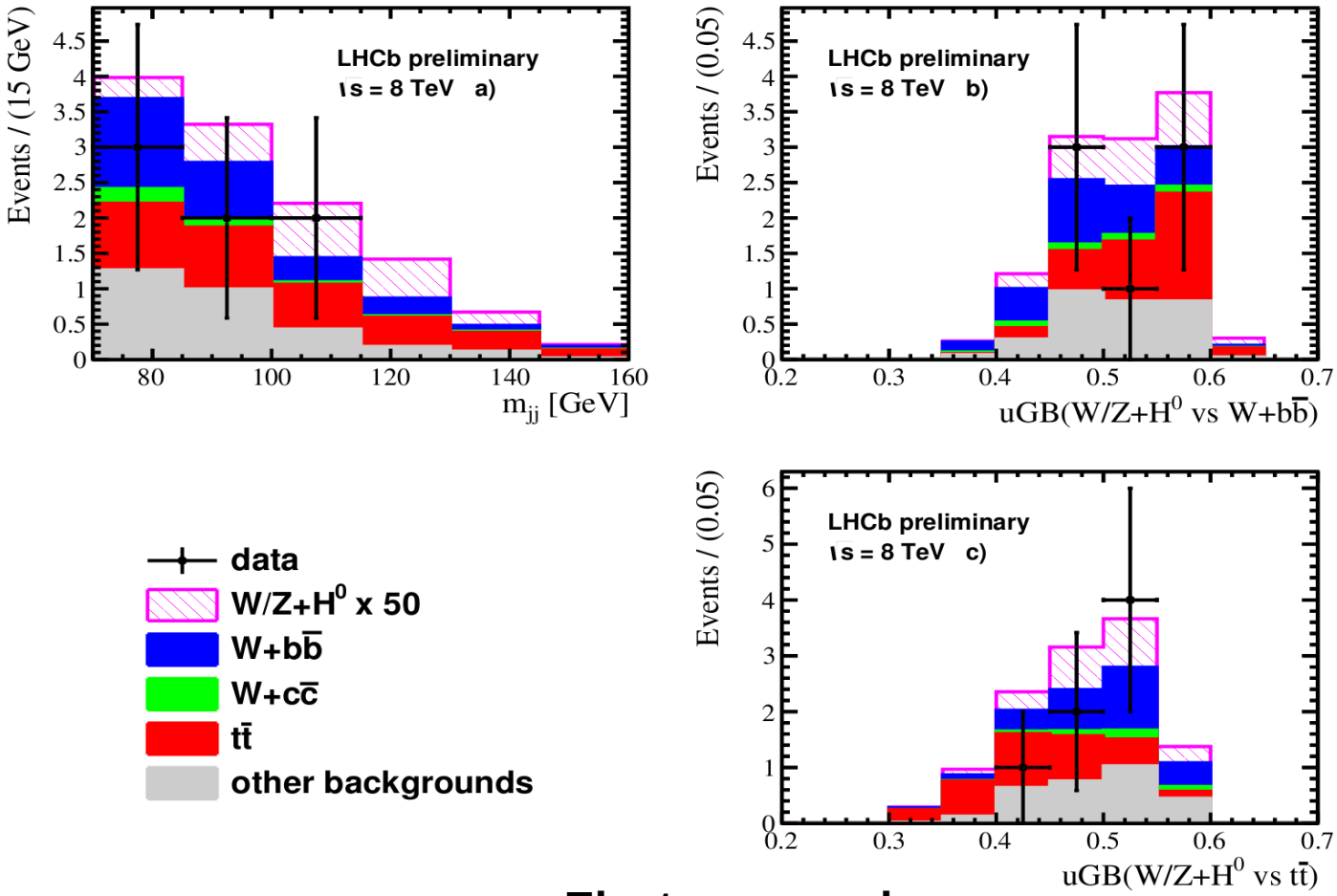
Search for H→b$\bar{b}$ and H→c$\bar{c}$ in association with a W or Z in the forward region (8 TeV)

- **Two uGBs are trained to discriminate:**
  - **W+b$\bar{b}$** from **V+H**
  - **t$\bar{t}$** from **V+H**

uncorrelation from the dijet invariant mass

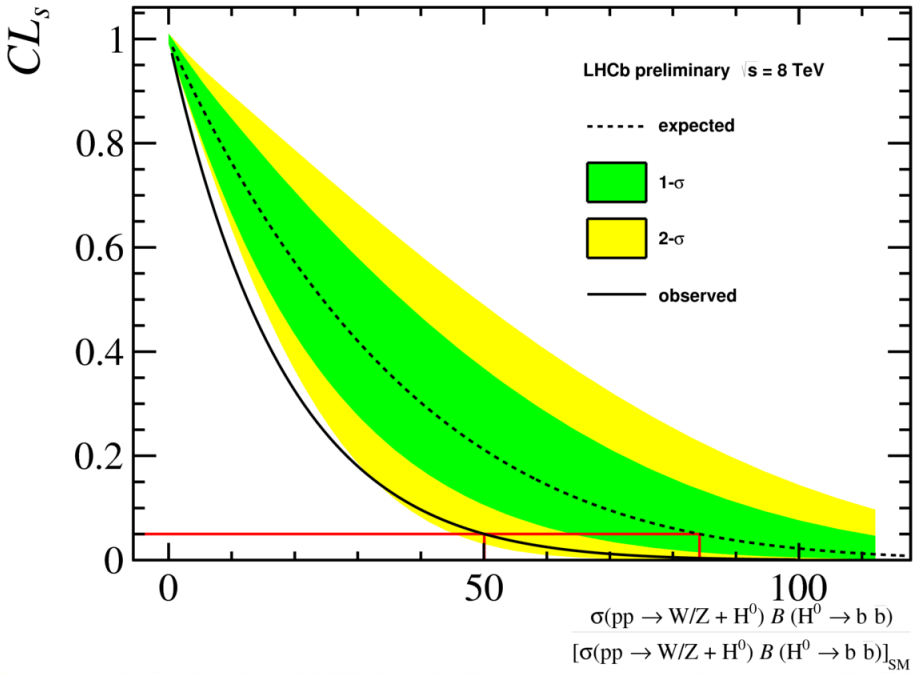
- **V+H (mass=125 GeV), W+b$\bar{b}$, t$\bar{t}$ training samples obtained from Monte Carlo simulation.**

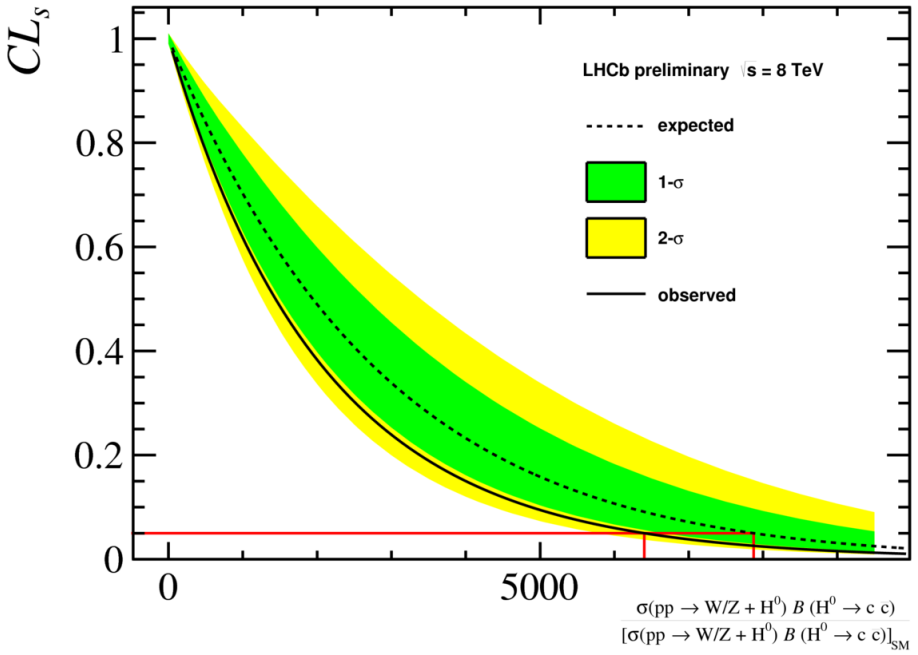
Input distributions in the Higgs upper limit computation



Electron sample

Search for $H \to b\bar{b}$ and $H \to c\bar{c}$ in association with a W or Z in the forward region (8 TeV)





Upper limit at 95% Confidence Level

$\sigma(V+H \to b\bar{b}) < 50\ \sigma_{SM}$ $\quad$ $\sigma(V+H \to c\bar{c}) < 6200\ \sigma_{SM}$

$\sigma(V+H \to b\bar{b}) < 1.6$ pb $\quad$ $\sigma(V+H \to c\bar{c}) < 9.4$ pb

First direct experimental upper limit on $H \to c\bar{c}$

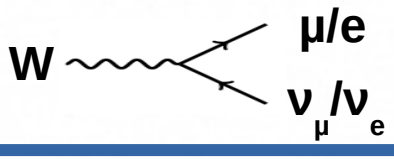
Conclusions

- **LHCb** capability in **jets physics** has been demonstrated.
- Thanks to the LHCb unique features an **excellent heavy flavour tagging system** has been developed.
- **Boosted Decision Trees** are used to separate **heavy flavour jets** from **light jets** and **b-jets** from **c-jets**.
- **Heavy flavour jets analyses** in Run I benefit from modern machine learning techniques (i.e. **uniform Gradient Boost**).
- **Work in progress to improve the jet tagging algorithm and performances.**

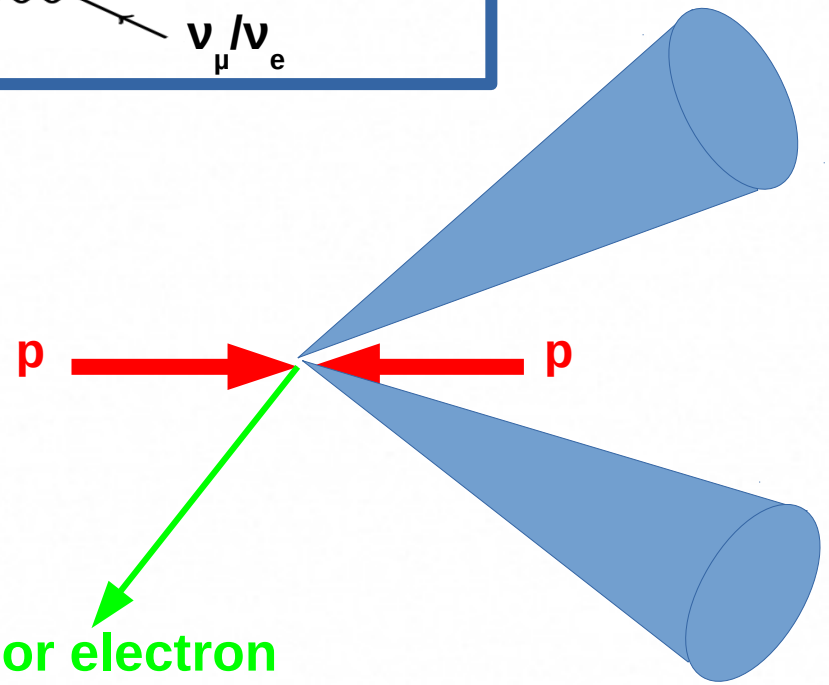
Backup slides

W+$b\bar{b}$, W+$c\bar{c}$ and $t\bar{t}$ candidates selection

The signature of W decays is a high momentum, isolated lepton.



Double SV-tag sample



jet_1

SV tagged

BDT(bc|udsg) > 0.2

12.5 GeV < p_T < 100 GeV

2.2 < η < 4.2

jet_2

Selected by the trigger

Isolated from other jets

p_T>20 GeV

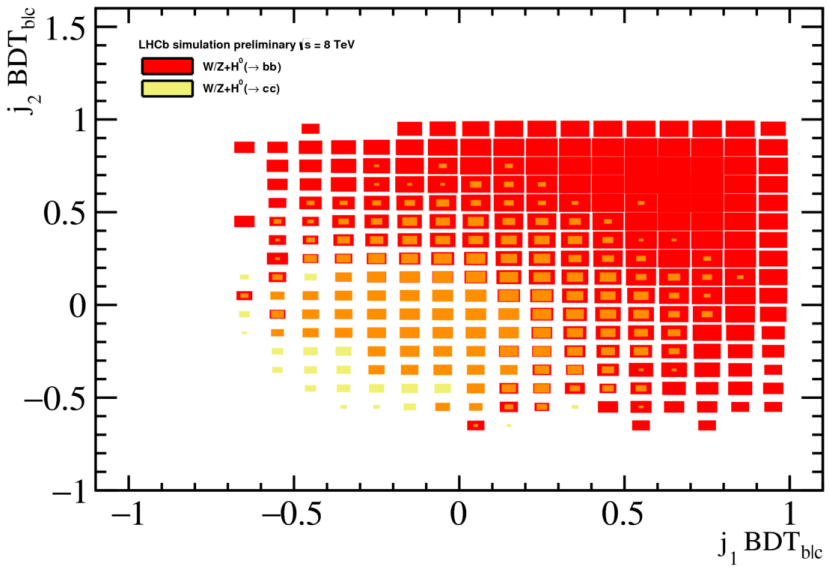
2 < η < 4.5 (4.25 for electrons)

The lepton requirement reduces almost all the combinatorial $b\bar{b}$ background!

Search for H→b$\bar{b}$ and H→c$\bar{c}$ in association with a W or Z in the forward region (8 TeV)

LHCb-CONF-2016-006

- **W/Z + H→(b$\bar{b}$/c$\bar{c}$)** candidates selection:
  - **Two SV-tagged jets**
    - 20 GeV < p_T < 100 GeV
    - 2.2 < η < 4.2
    - BDT(bc|udsg)>0.2
  - **One muon or electron**
    - p_T > 20 GeV
    - 2 < η < 4.5 (4.25 for electron)
    - Isolated from jets



H →b$\bar{b}$ vs **H →c$\bar{c}$** using **BDT(b|c)**

↓

Extra cuts in H →c$\bar{c}$ search:

- **Jet_1 BDT(b|c) < 0.2**
- **Jet_2 BDT(b|c) < 0.2**

↓

Remove 90% of H →b$\bar{b}$ and retain 60% of H →c$\bar{c}$