Heavy flavour jet tagging at LHCb

Lorenzo Sestini
Università di Padova and INFN

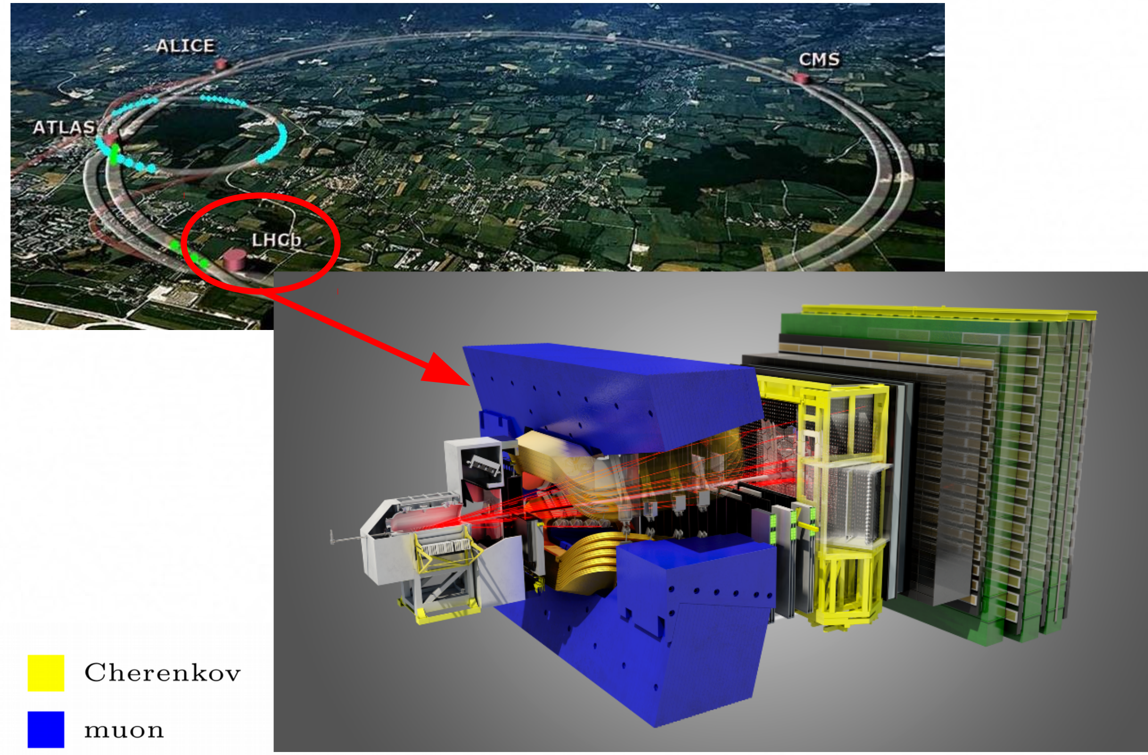
On behalf of the LHCb collaboration

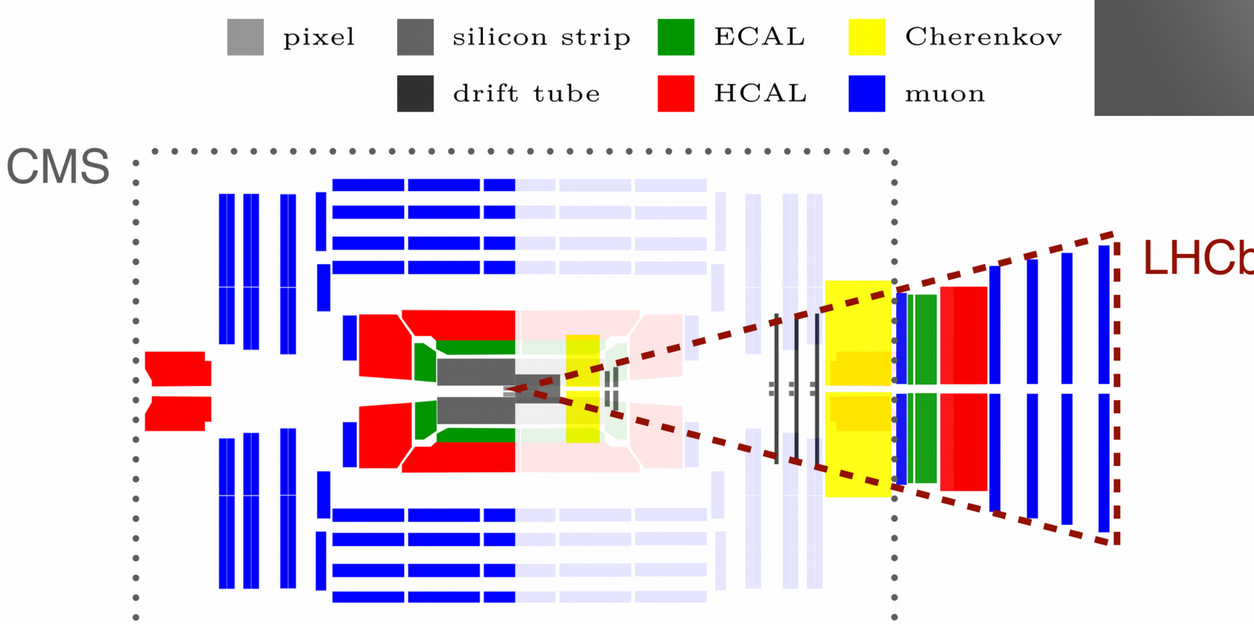
CMS Heavy Flavour Tagging Workshop, 11-04-2018, Bruxelles

LHCb detector

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

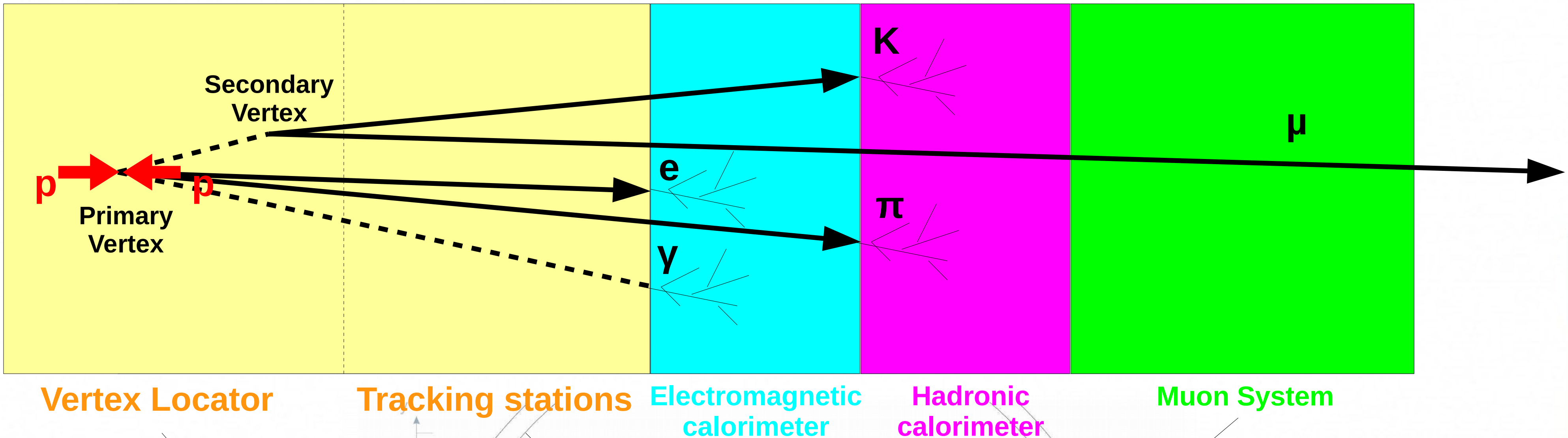
- **LHCb is a spectrometer initially designed to study b and c hadrons physics.**
- **It covers a phase space region of p-p collisions complementary to ATLAS and CMS, corresponding to $2 < \eta < 5$.**





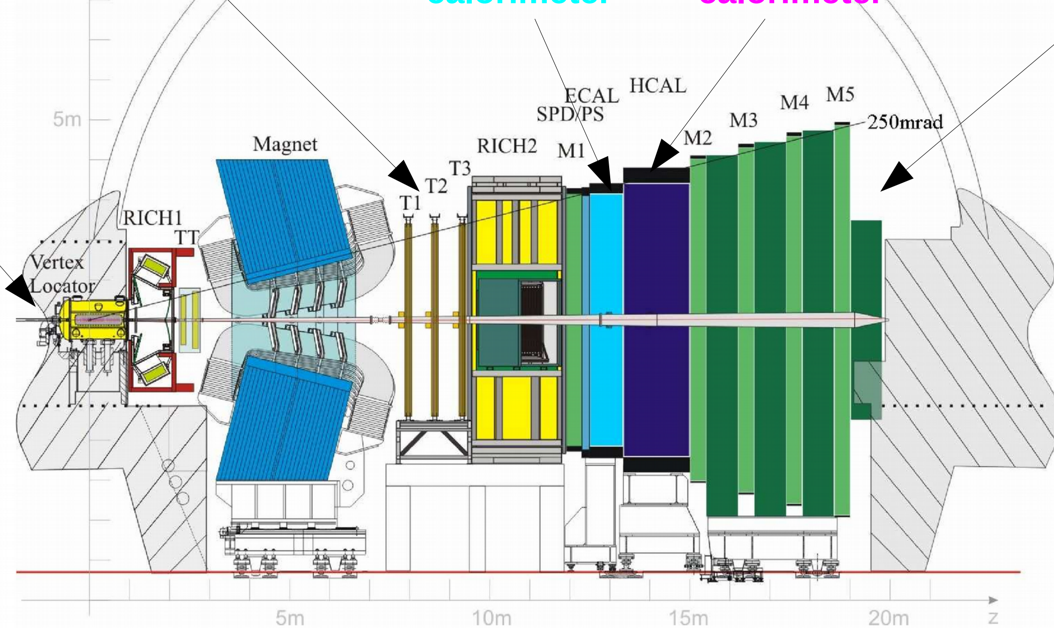
In the last years LHCb has demonstrated its capability in jet physics!

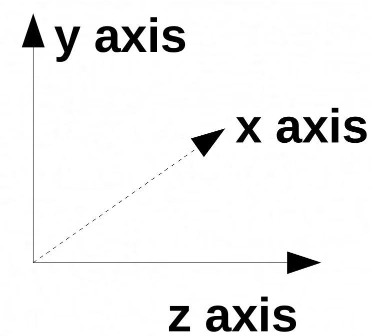
Jet detection at LHCb



Jet reconstruction inputs:

- Long tracks
- Calorimeter clusters
- Metastable particles (like K_s)



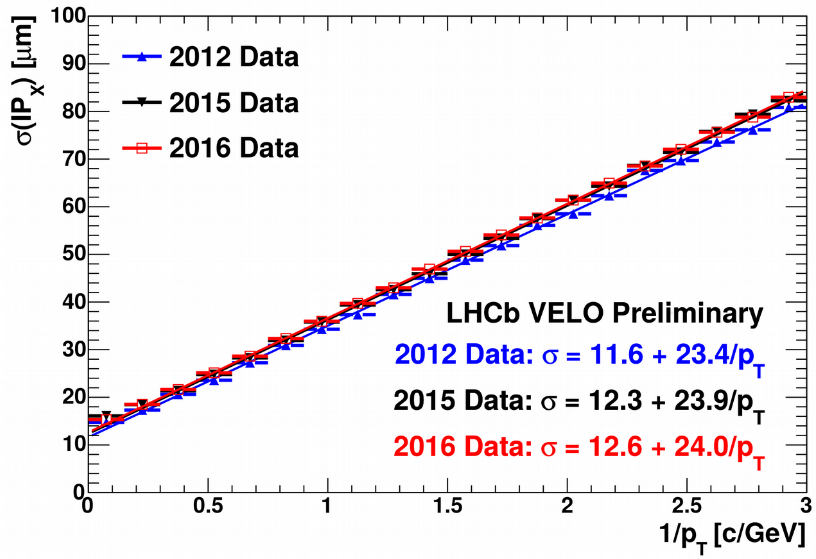


LHCb performance

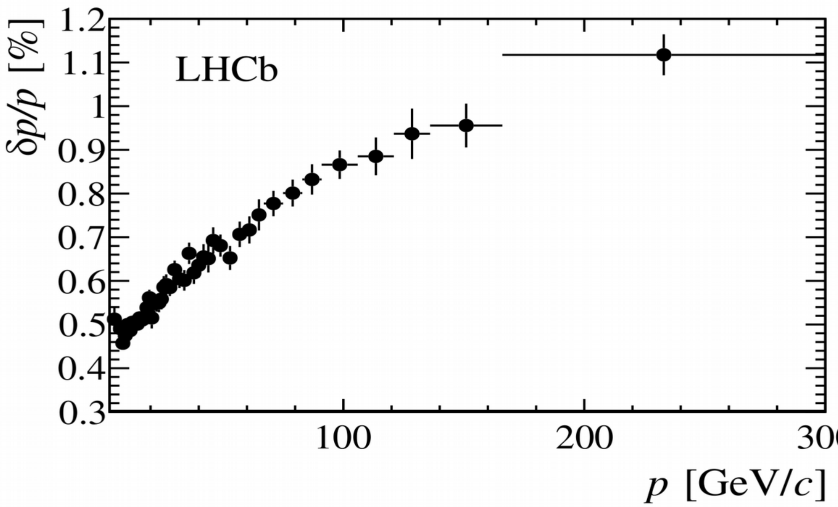
Int. J. Mod. Phys. A 30, 1530022 (2015)

Excellent IP resolution

→ very important for SV reconstruction and tagging



Tracks momentum resolution



Electromagnetic calorimeter

Hadronic calorimeter

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

Energy resolution

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

Limitations due to saturation

Calorimeter clusters in input to jets reconstruction

- **Clusters isolated from tracks (neutral particles)**
- **Excesses of energy nearby tracks (neutral recovery)**

Jet reconstruction algorithm

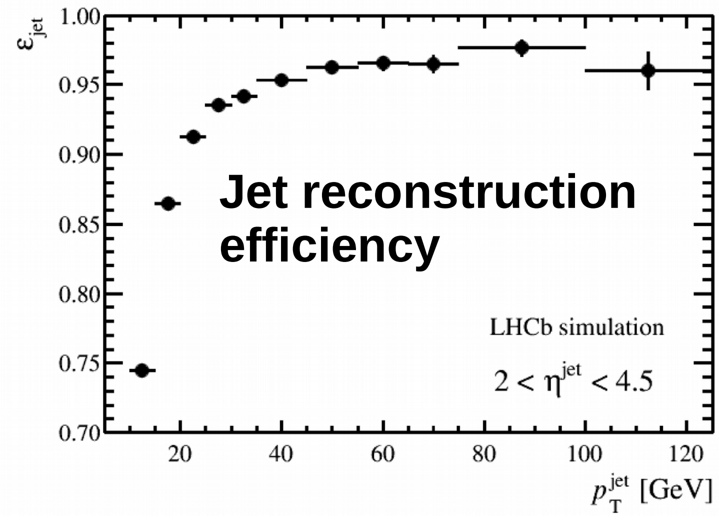
JHEP 01 (2014) 033

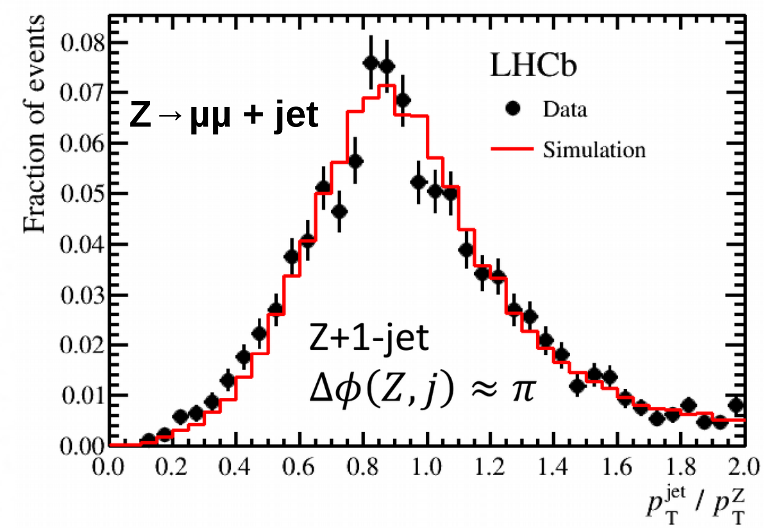
Particle Flow → **Long tracks** and **calorimeter clusters** and **metastable particles** are selected as **inputs**

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)**



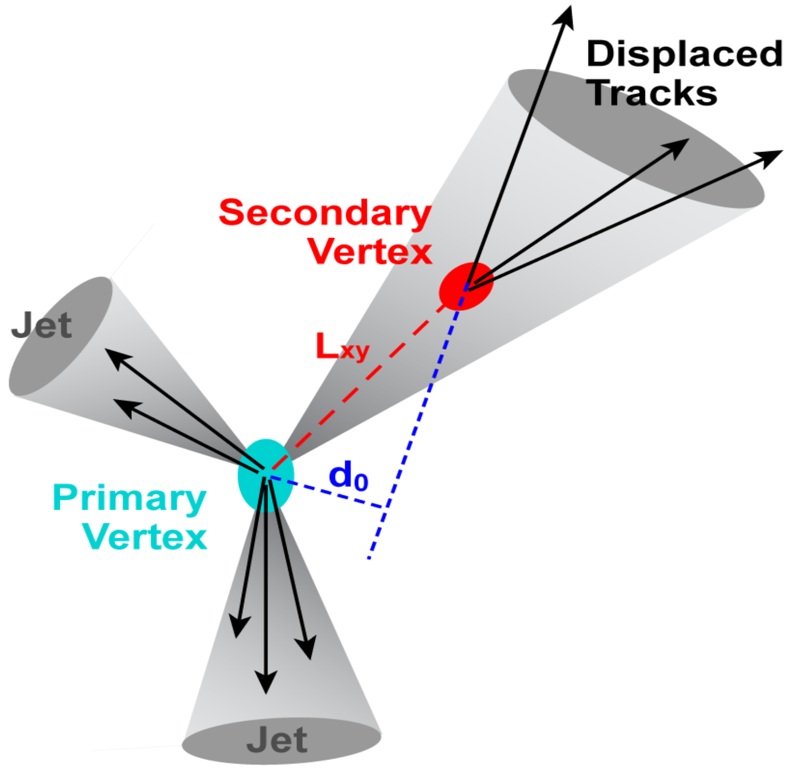


Energy resolution of final jets $\delta E/E \approx$ 10-16% in 20-100 p_T range.

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.



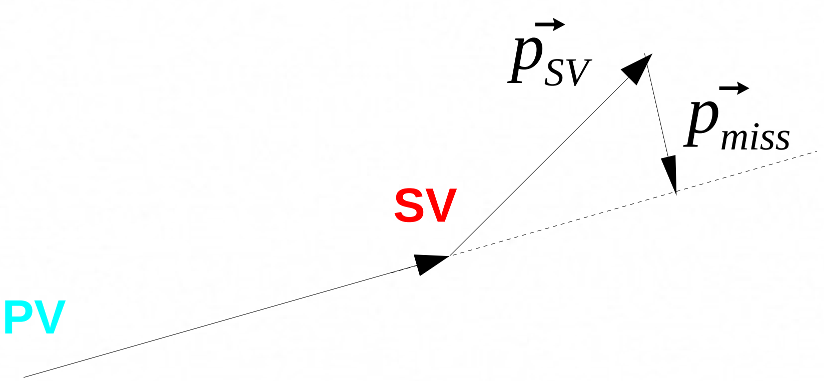
- Two **Boosted Decision Trees** are used to identify b and c jets.

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

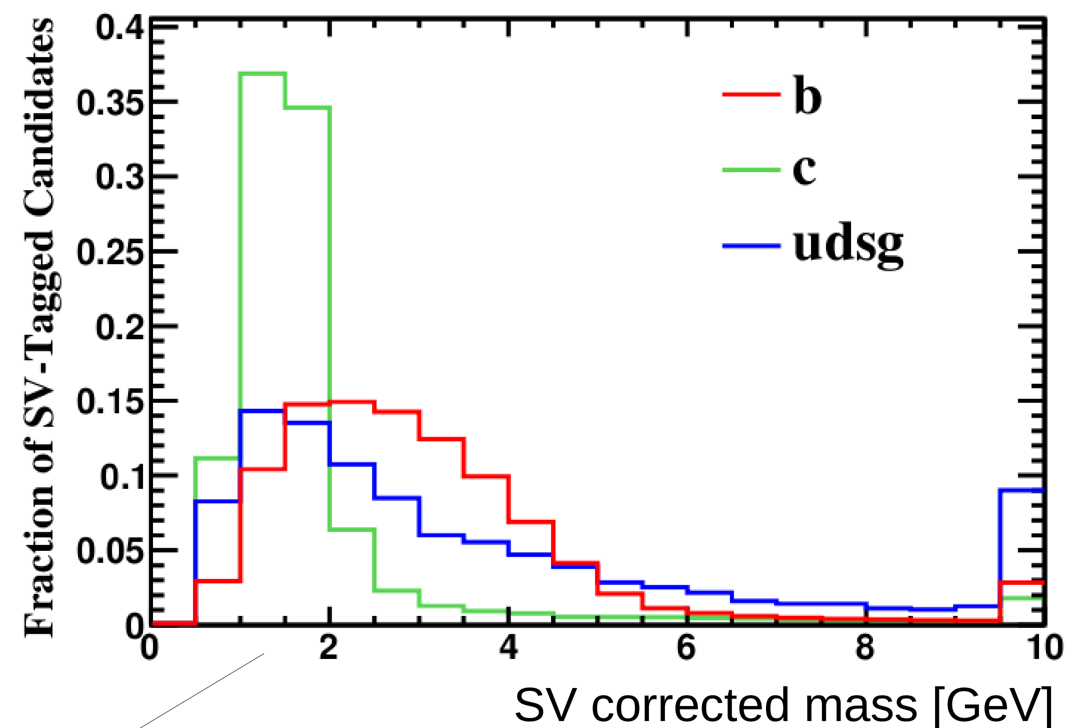
BDT(b|c)
To separate **b-jets** from **c-jets**

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**



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

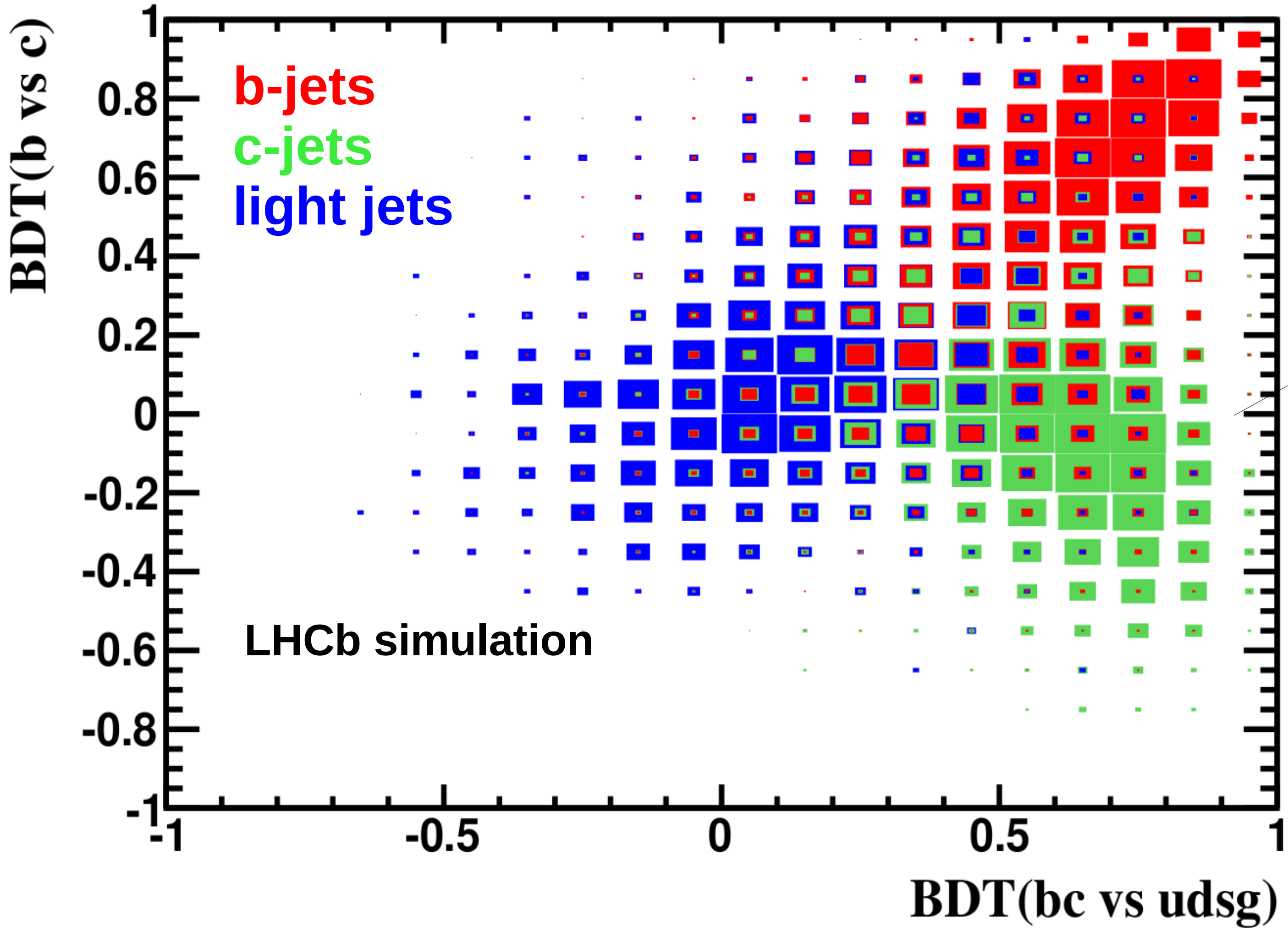


Useful to discriminate b from c

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 (Run I).**

- 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
  - **B+jet** (**b** enriched)
  - **D+jet** (**b** and **c** enriched)
  - **μ(displaced)+jet** (**b** and **c** enriched)
  - **W(→high p_T μ)+jet** (**light jets** enriched)

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

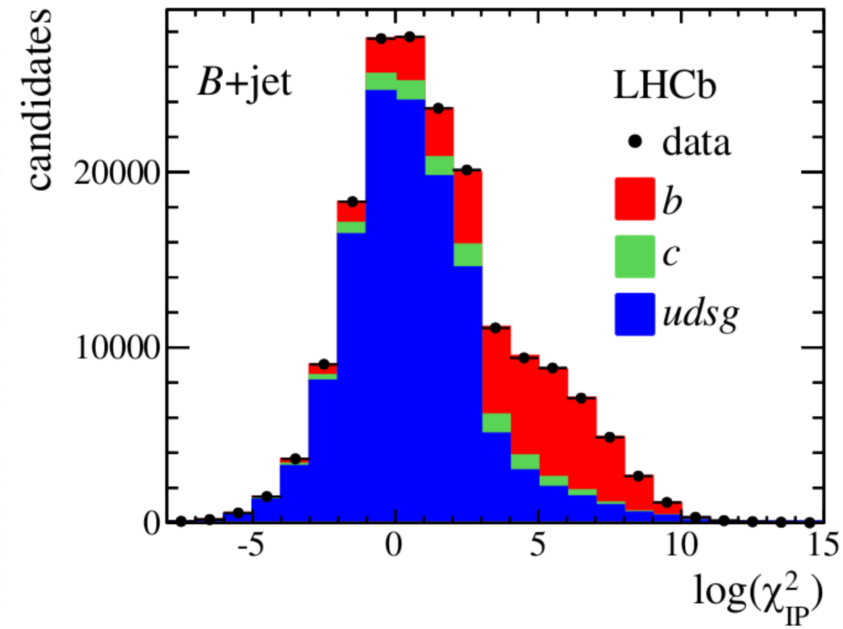
- **Two-dimensional templates are obtained from simulation**

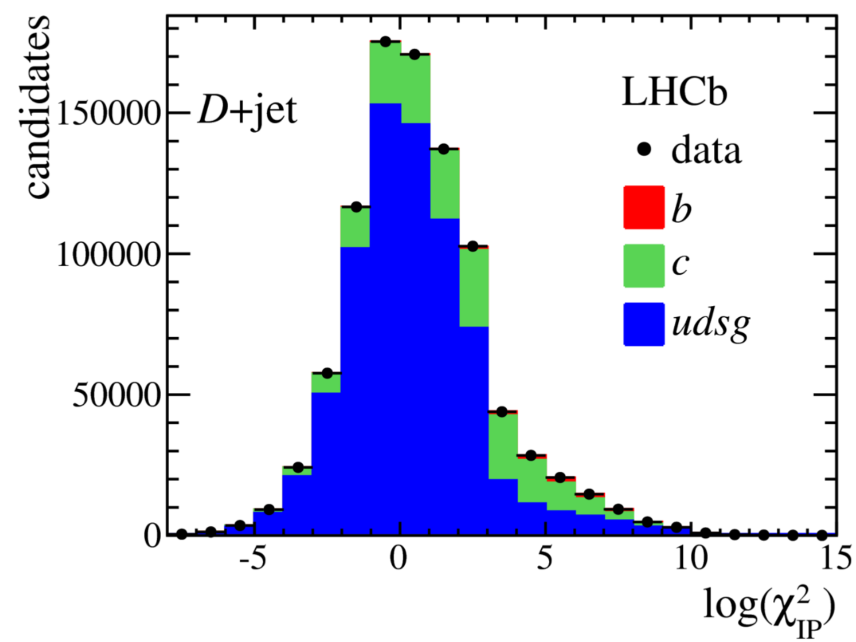
Jet tagging at LHCb

- 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.

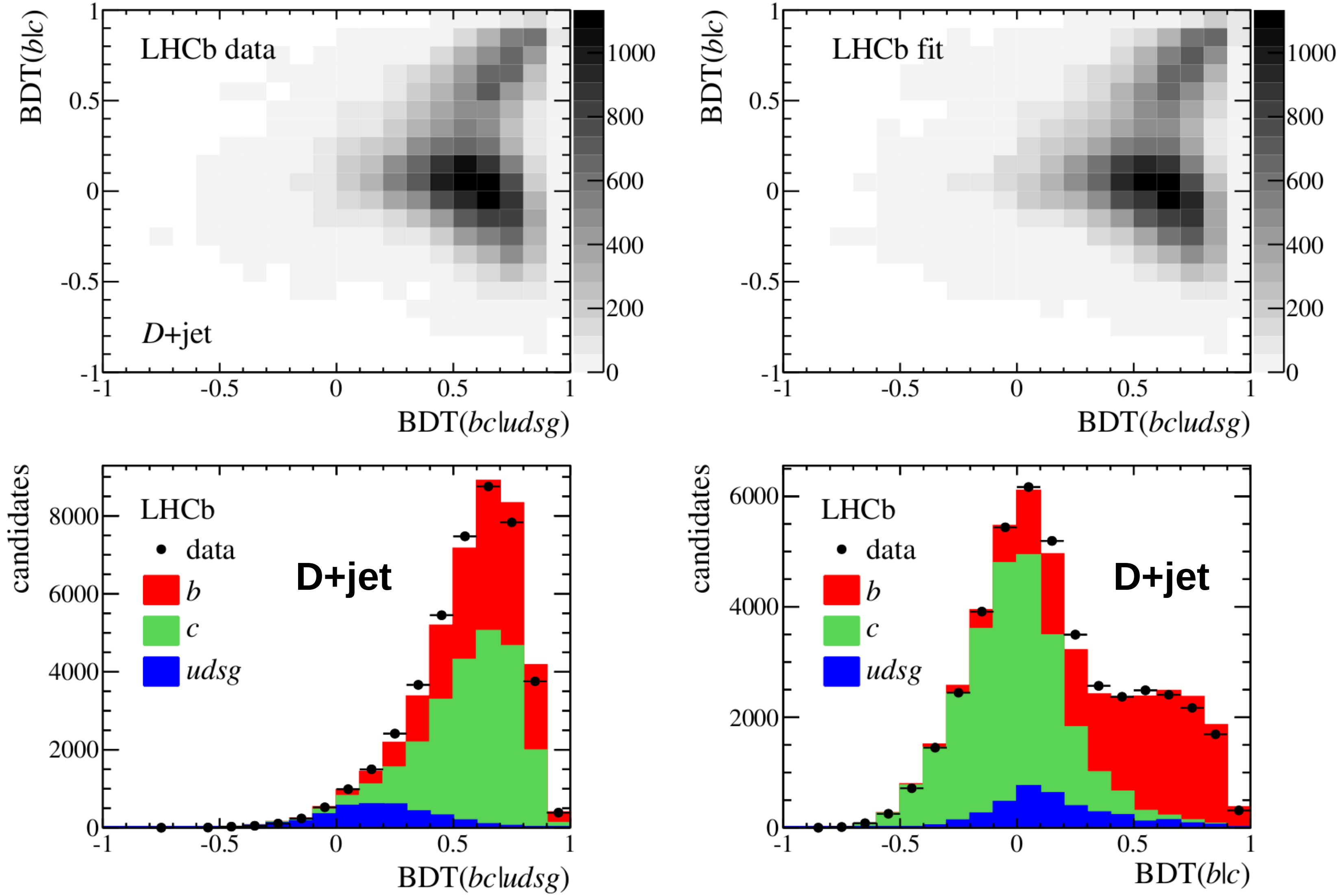
- Templates are obtained from simulation.

- Yields of b, c and light jets **after applying the SV-tag** are measured by fitting the 2-dimensional distribution of the BDTs (**next slide**).





Jet tagging at LHCb

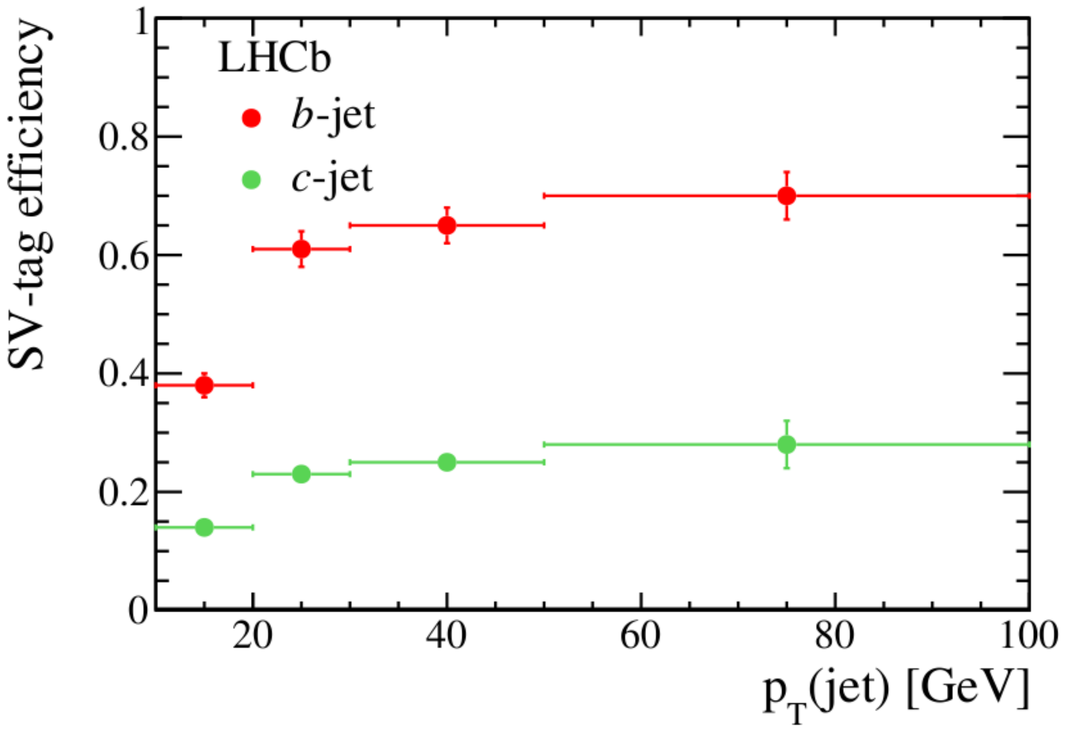


Jet tagging at LHCb

- Efficiencies obtained with: $\epsilon = \frac{N(pass)}{N(tot)}$

from BDTs fit

from χ^2_{IP} fit



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%**

- Uncertainty due to the limited statistics of the data samples and to the modeling of the templates

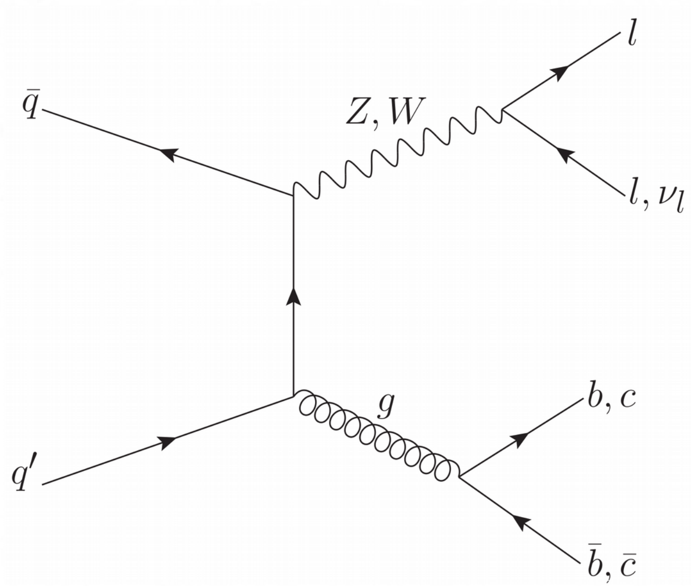
Measurement of forward W+b$\bar{b}$, W+c$\bar{c}$ and t$\bar{t}$

Phys. Lett. B767 (2017) 110

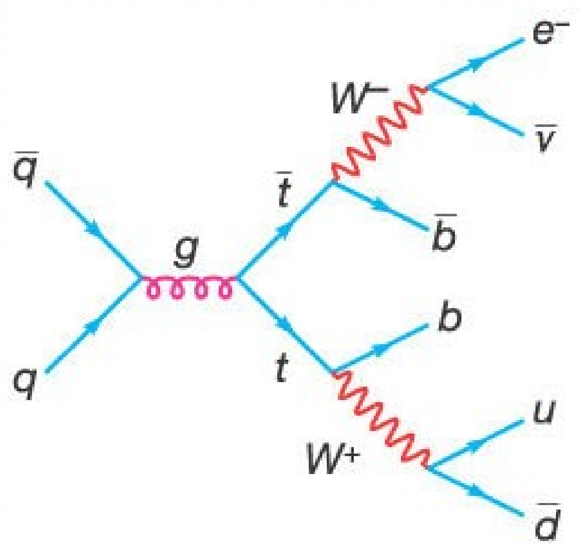
- Application of jet reconstruction and heavy flavour tagging at LHCb
- Measurement of **W+b$\bar{b}$**, **W+c$\bar{c}$** and **t$\bar{t}$** cross sections the forward region with the 8 TeV data sample (2 fb^{-1}).

Provide constraints to Parton Distribution Functions

Dominant backgrounds in the W/Z+H(→b$\bar{b}$) search



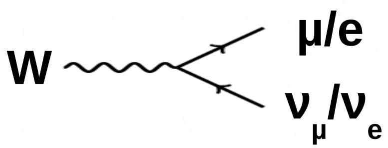
W+b$\bar{b}$
W+c$\bar{c}$



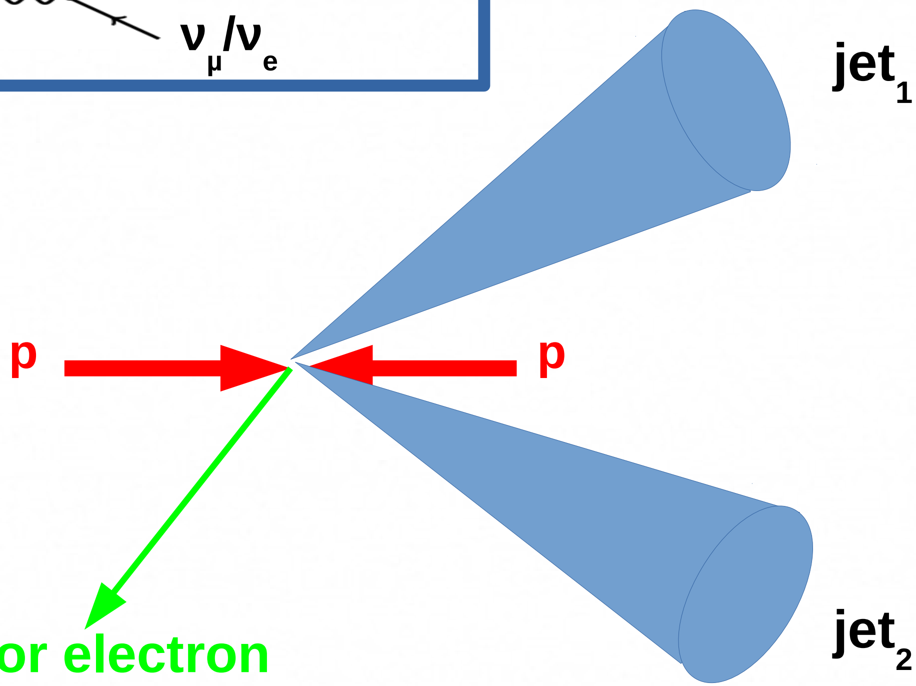
t$\bar{t}$

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



SV tagged

BDT(bc|udsg) > 0.2

12.5 GeV < p_T < 100 GeV

2.2 < η < 4.2

Selected by the trigger

Isolated from other jets

p_T>20 GeV

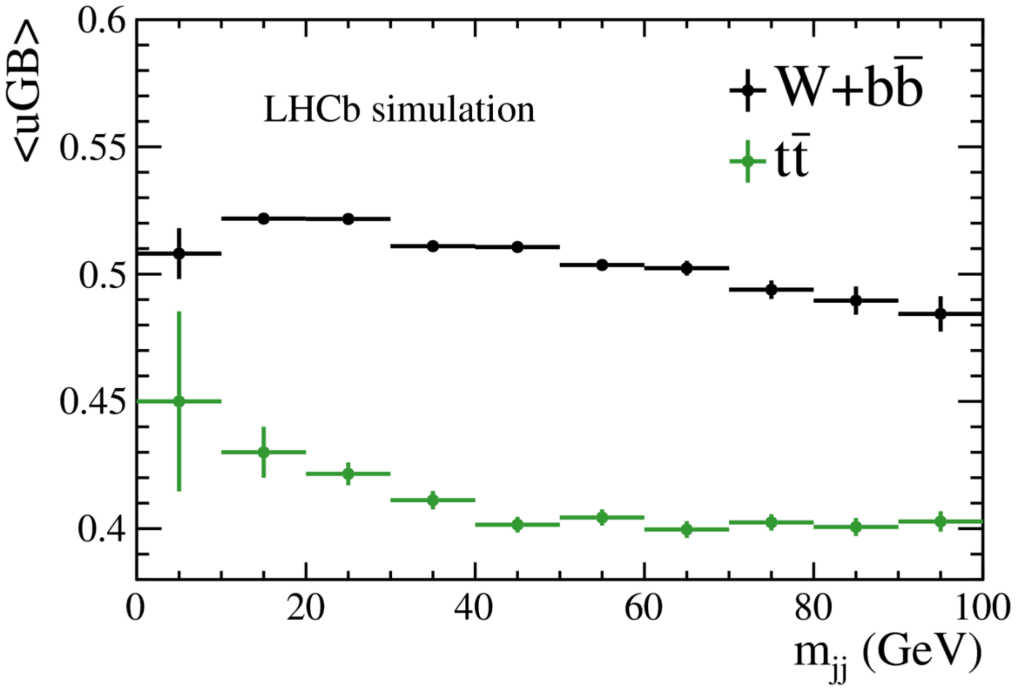
2 < η < 4.5 (4.25 for electrons)

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

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.**
- Uncorrelation with the dijet invariant mass is required**: the BDT is trained with the Uniform Gradient Boost technique.**
- $W+b\bar{b}$ and $t\bar{t}$ Monte Carlo samples are used in the training.
- **Some observables in input to the BDT**
  - **lepton transverse momentum**
  - **jets transverse momenta**
  - **jets masses**
  - **lepton pseudorapidity**

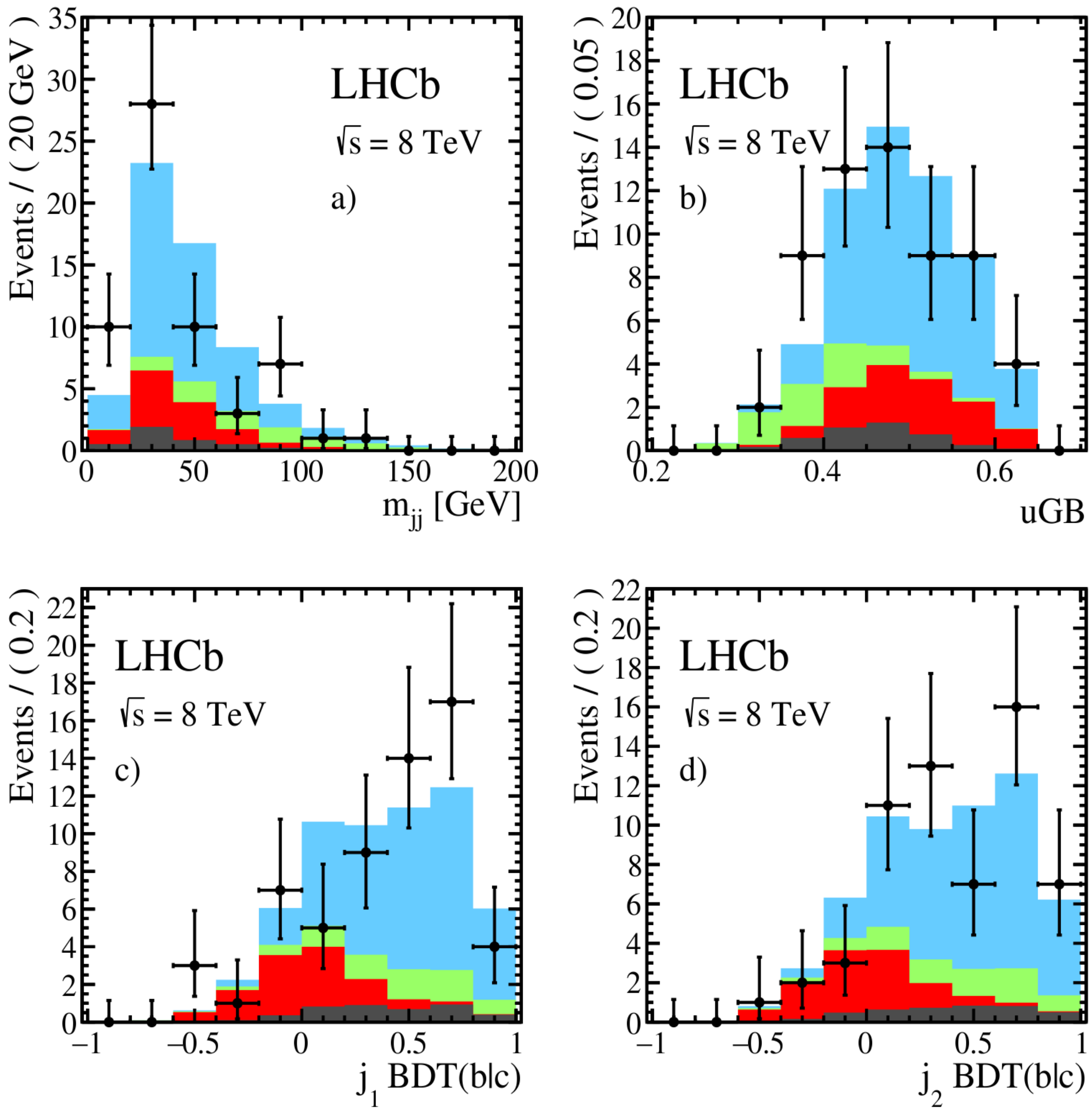


Average uGB response in different intervals of dijet invariant mass

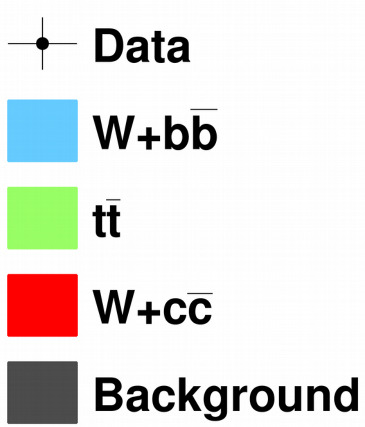
The uncorrelation is achieved to reduce systematics in the final fit

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

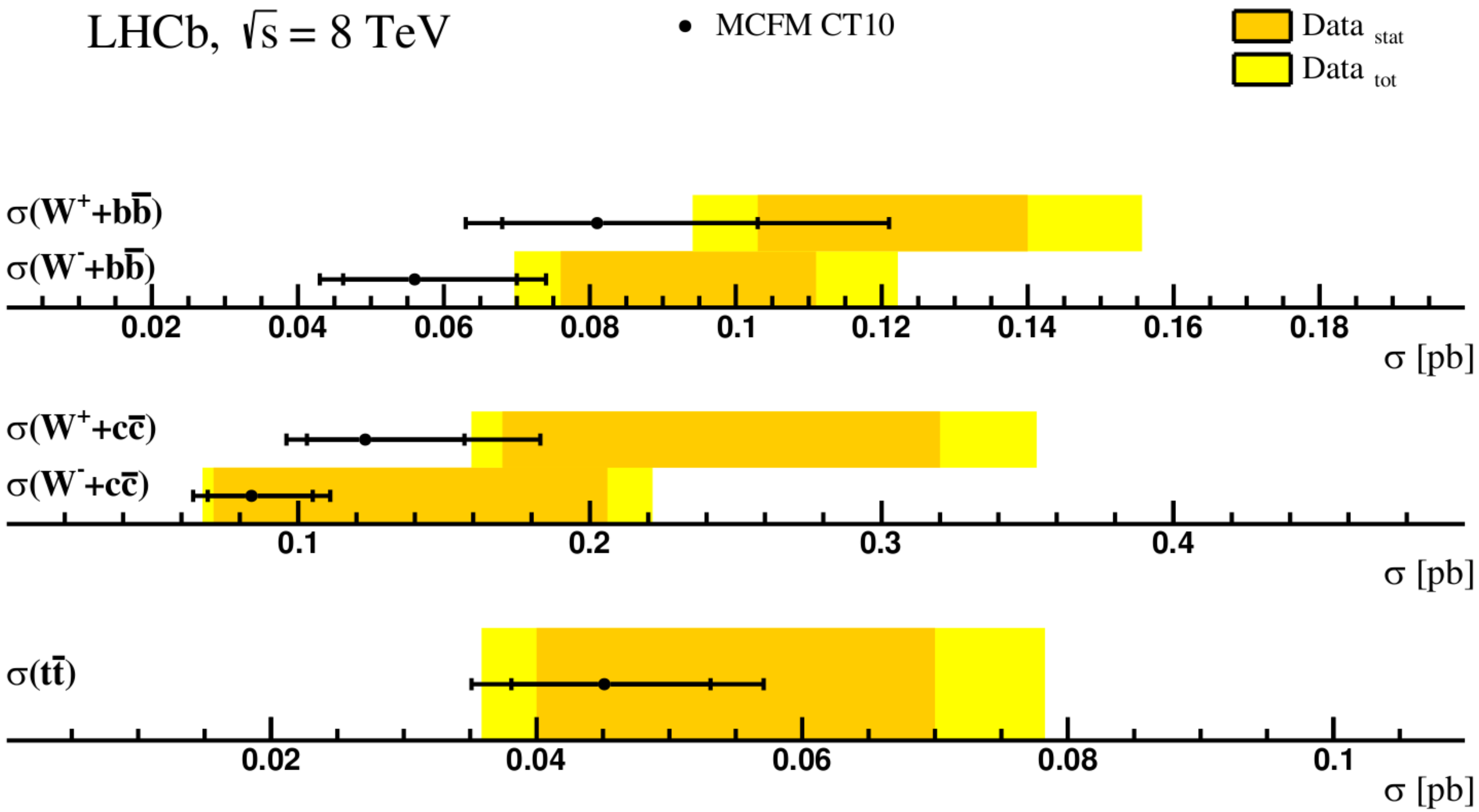
μ^+ sample



- 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.
- Backgrounds: QCD, Z+b etc.



$W+b\bar{b}$, $W+c\bar{c}$ and $t\bar{t}$ cross sections



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

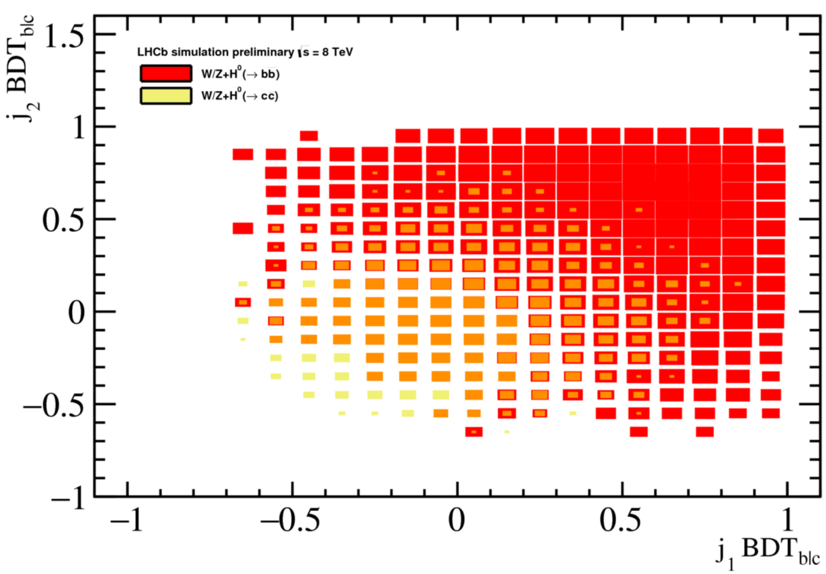
The measured cross sections are compatible with the Standard Model predictions within the errors.

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**
    - 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 →$c\bar{c}$

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)

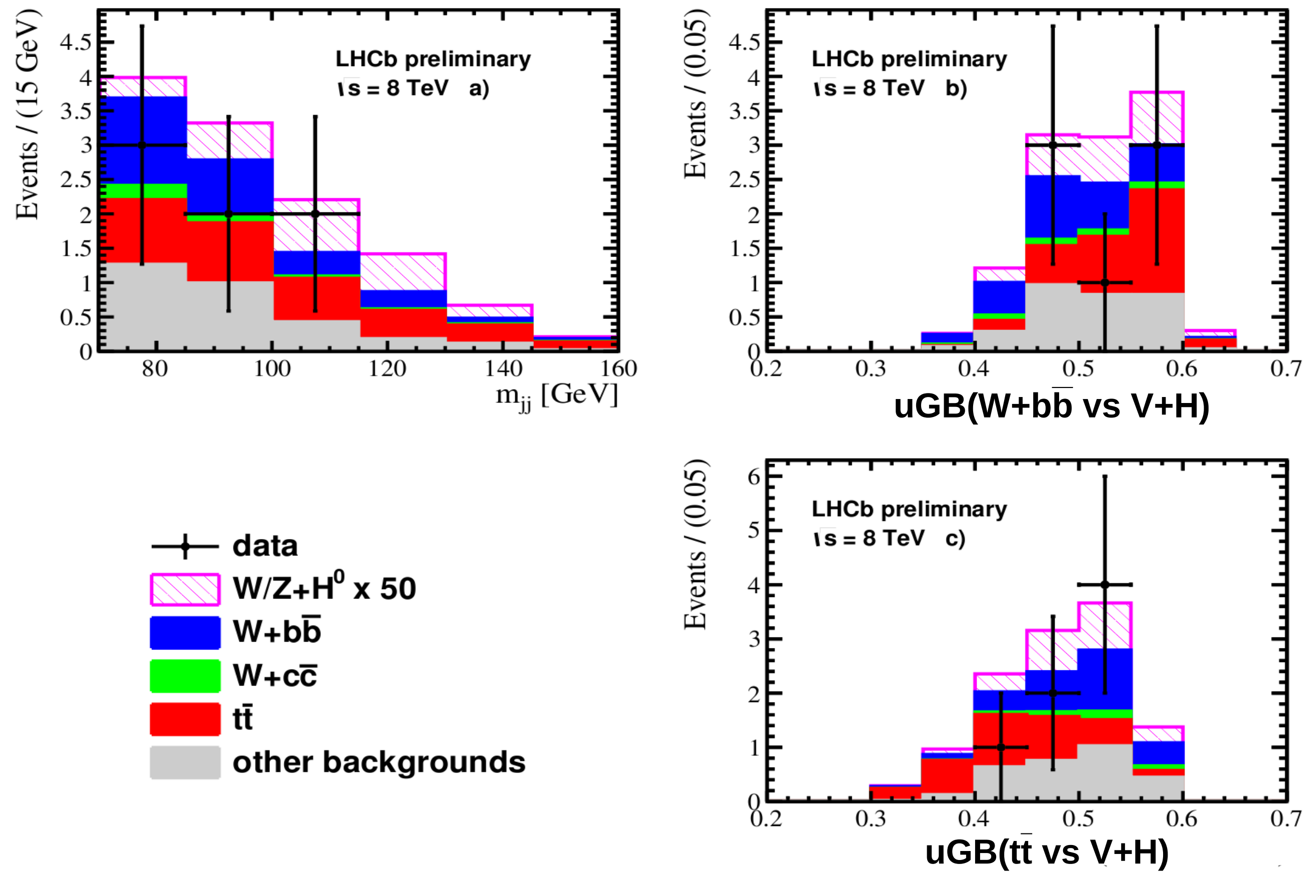
Data compatible with the background only hypothesis → **We can set an upper limit on the cross section**

Input distributions in the CL_s computation

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

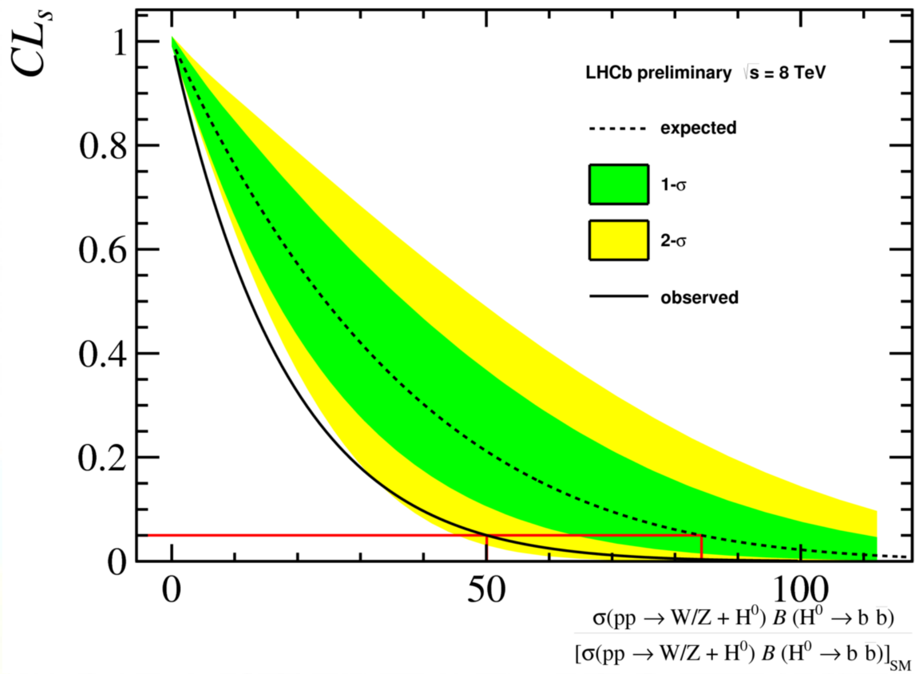
- **Background prediction obtained from:**
  - Simulation
  - SM cross sections
  - Data-driven techniques (for QCD)

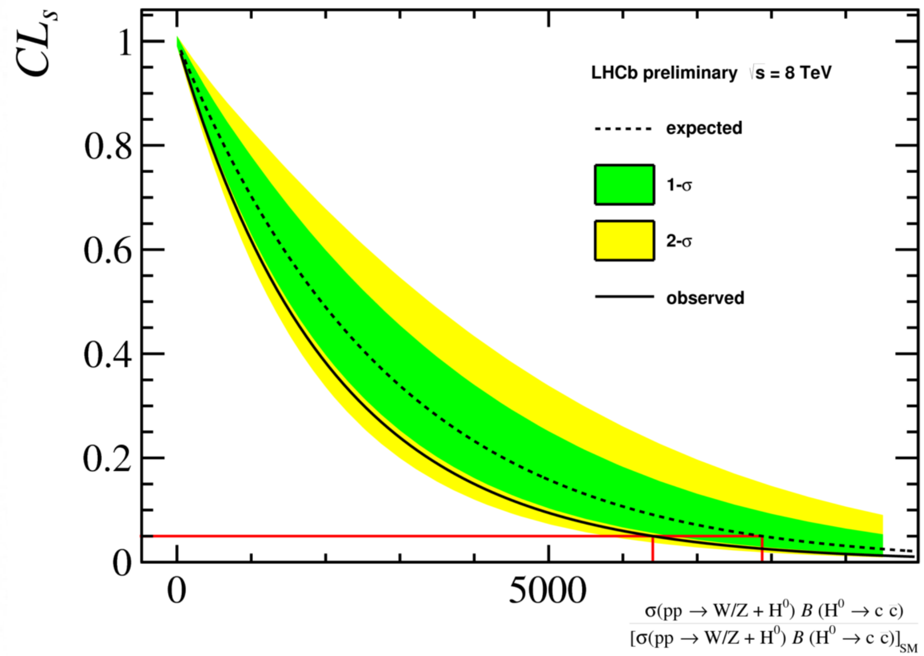
- **Higgs model obtained from simulation, assuming a mass of 125 GeV**



Electron sample

$V+H(\rightarrow b\bar{b}/c\bar{c})$ cross section upper limit





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

Upper limit at 95% Confidence Level

$\sigma(V+H \rightarrow b\bar{b}) < 50\ \sigma_{SM}$ — **$\sigma(V+H \rightarrow c\bar{c}) < 6200\ \sigma_{SM}$**

$\sigma(V+H \rightarrow b\bar{b}) < 1.6$ pb — **$\sigma(V+H \rightarrow c\bar{c}) < 9.4$ pb**

But not as good as ATLAS result

Future prospects on $H \to c\bar{c}$

- In the HL-LHC phase LHCb plans to collect of about 300 fb^{-1} of integrated luminosity.

- **We expect several improvements in the future**:
  - Dedicated c-tagging algorithms.
  - Improved electron reconstruction for W/Z selection.
  - Improvements in the jet energy resolution.

- We estimated that an observation of VH($\to c\bar{c}$) is likely out of reach, due to a lack of signal events.

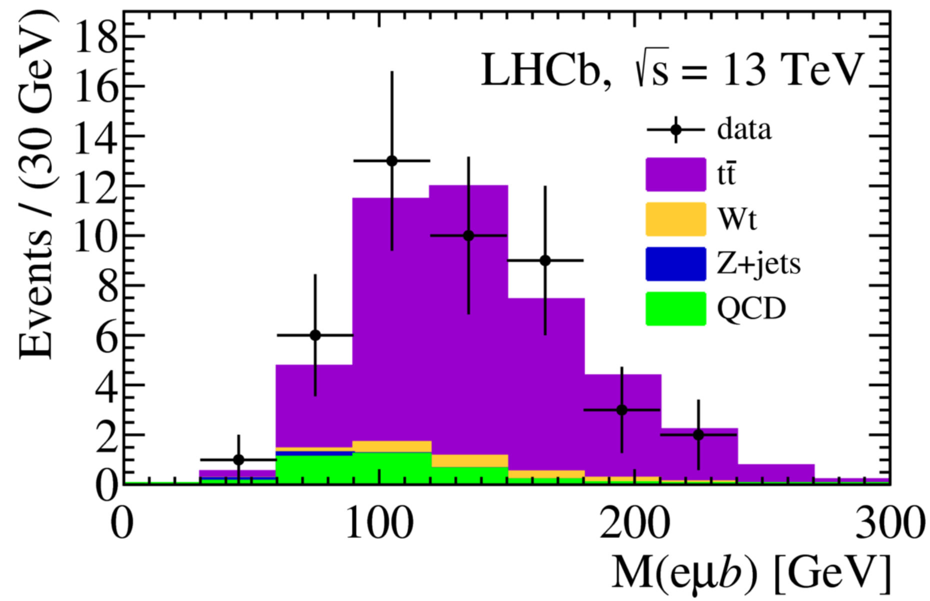
- But a limit below 5-10 x σ_{SM} (2-3 x SM on the Yukawa coupling) seems plausible with 300 fb^{-1}.

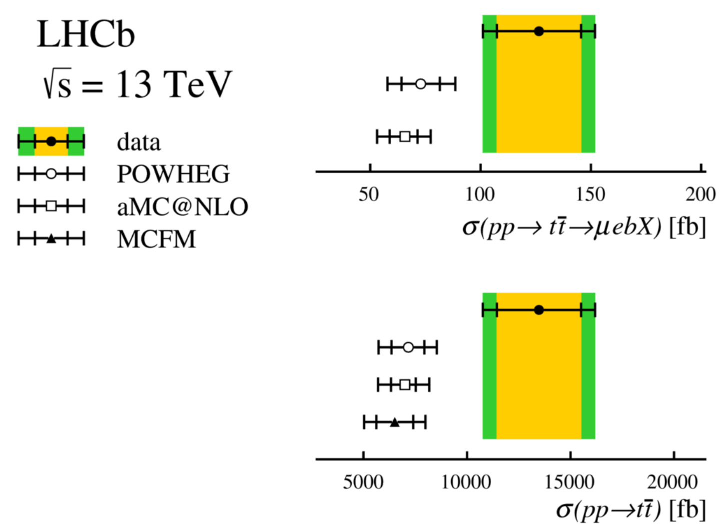
- For more details take a look at https://agenda.infn.it/getFile.py/access?contribId=36&sessionId=4&resId=0&materialId=slides&confId=12253

$t\bar{t}$ production in Run II

LHCb-PAPER-2017-050

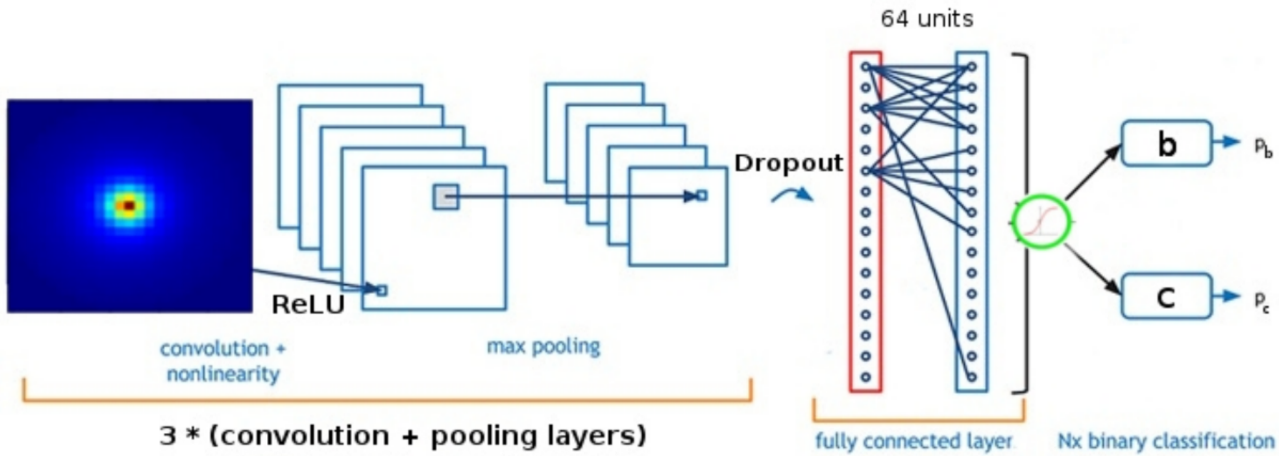
- **Application of LHCb flavour tagging in Run II** (1.93 fb^{-1} at 13 TeV).
- **$t\bar{t}$ → μeb** final state: one heavy flavour tagged b-jet, one muon and one electron with opposite charge in the LHCb acceptance.
- The two leptons must have p_T> 20 GeV, 2.0<η<4.5 and a separation between them of ΔR>0.1
- The jet must have p_T > 20, 2.2<η<4.2, and separated from leptons of ΔR>0.1
- QCD background obtained from same charge lepton candidates, other backgrounds from simulation.
- The measured cross section is less than 2σ from the theoretical predictions.
- **Cross section systematic dominated by the knowledge of the tagging efficiency.**





Improving the tagging algorithms

- Using SV + jets images.
- **Jet image**: transverse energy distribution in the η-φ space, divided in pixels.



CNN image processing:

- Layers weigths regularized with L2 norm.
- Batch normalization after convolution.
- Used Adam optimizer.
- Kernel size (5,5) (3,3) (5,5)
- Convolution layers size: 96, 128, 48

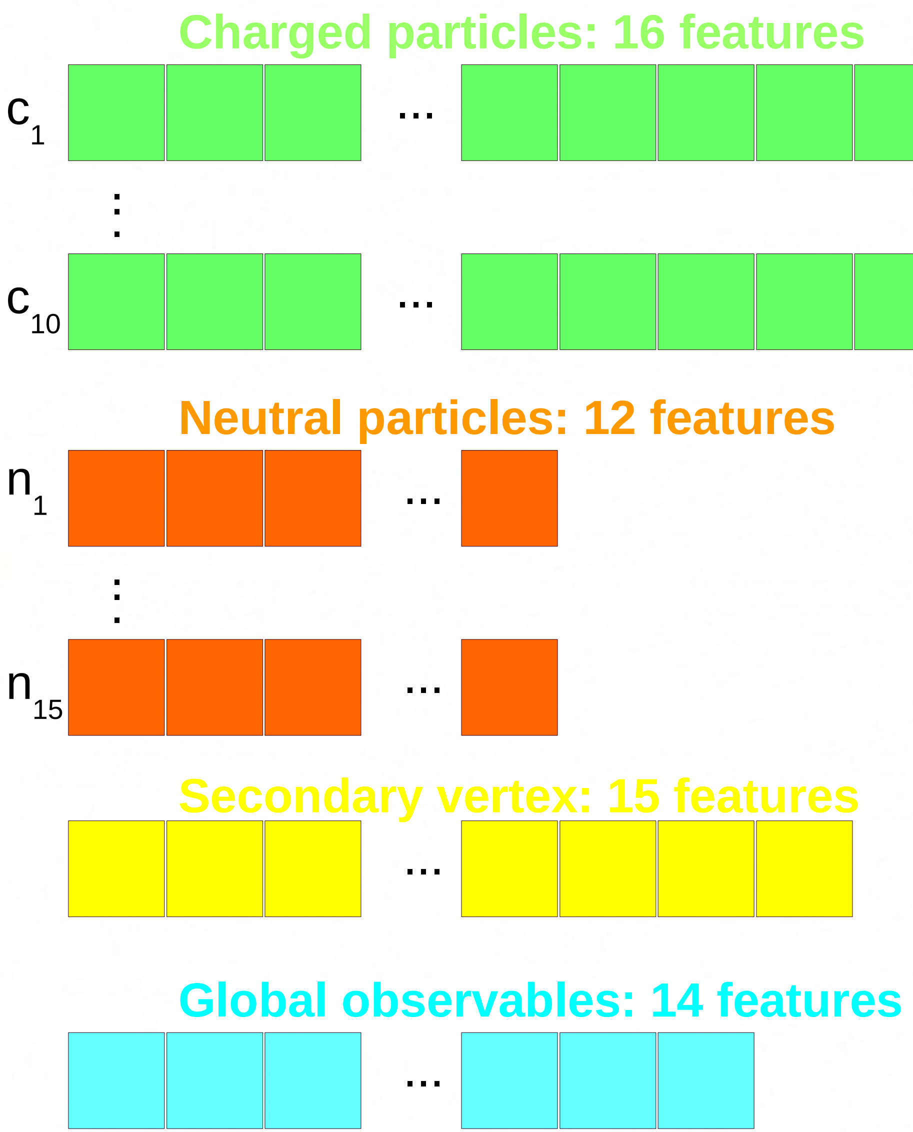
Forward CNN output to DNN:

- Added 13 variables SV-related.
- 4 * (dense + dropout) layers (96,128,96,48)
- ReLU activation.
- Used Adam optimizer.

- Keras and Tensorflow are used in the computation.
- We can use this technique to tag merged b- fat jets.

Improving the tagging algorithms

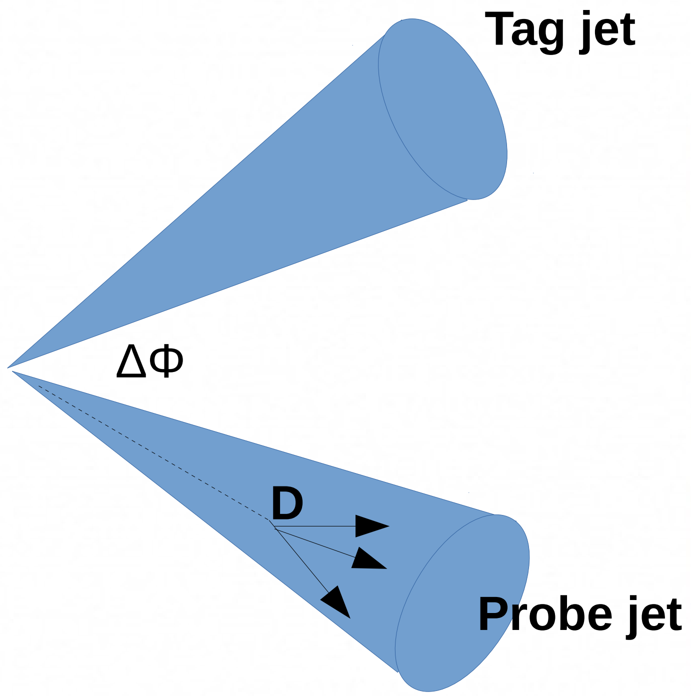
- **Matrix with jets constituents as input (inspired by CMS Deep Tagging)**



- Tracks, calorimeter clusters, and SV. **Particle ID informations included**
- Fixed number of particles: when a charged, neutral or SV is not present all its variables are set to 0.
- LSTM tecnique is used to exploit correlations among particles.
- Output of LSTM is given to a DNN.
- Ternary classification with 3 probabilities p_b, p_c , p_q.
- Keras and Tensorflow are used.
- **First results are promising, network optimization ongoing.**

Calibration and uncertainties

- In Run I the tagging efficiency uncertainty has been one the main systematic source of b- and c-jets analysis.
- The plan is to use Run II data to calibrate the tagging algorithm and to reduce the uncertainty.
- We can use use a dijet tag and probe tecnique.
- For b-tag calibration: the probe jet is a jet that contains a displaced J/Ψ .
- For c-tag calibration: the probe jet is a jet that contains a reconstructed prompt D meson.



Conclusions

- **LHCb** capability in **jet 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**.

- **Work in progress to improve the jet tagging algorithm and performances!**

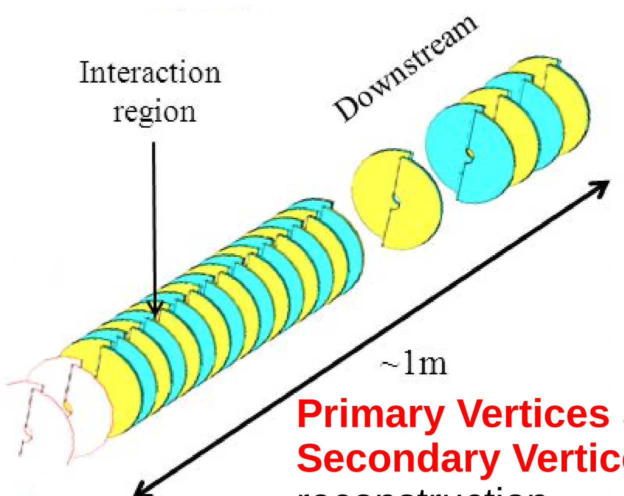
Backup slides

Tracking system

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

VErtex LOcator (VELO)

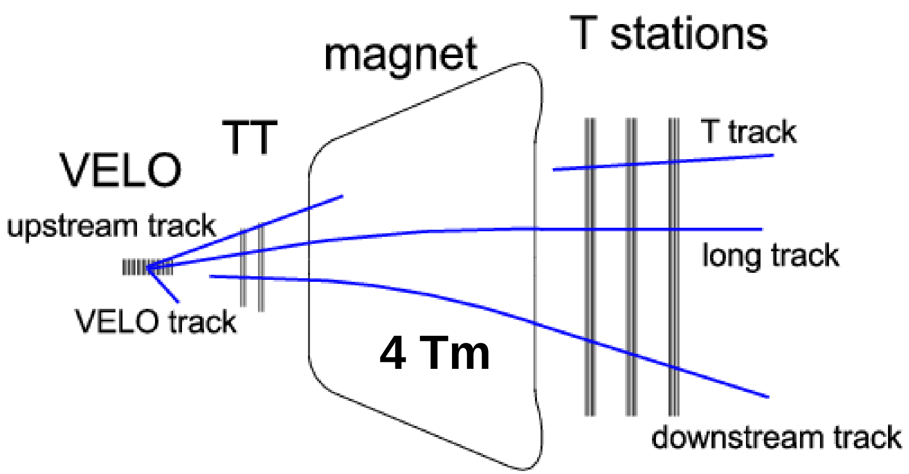
21 stations → (r,φ) coordinates



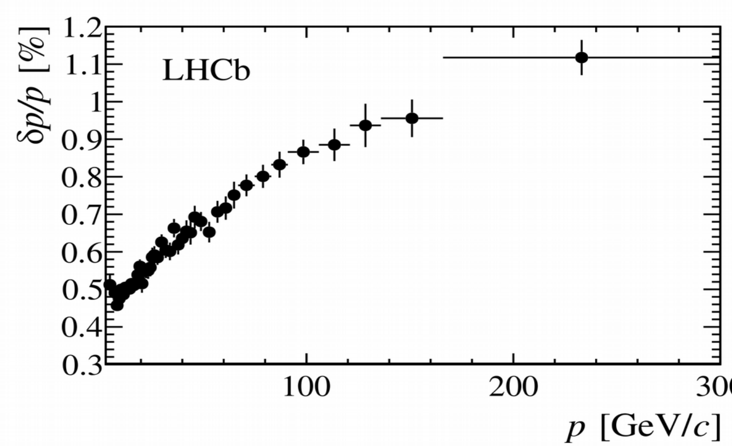
Primary Vertices and **Secondary Vertices** reconstruction

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**

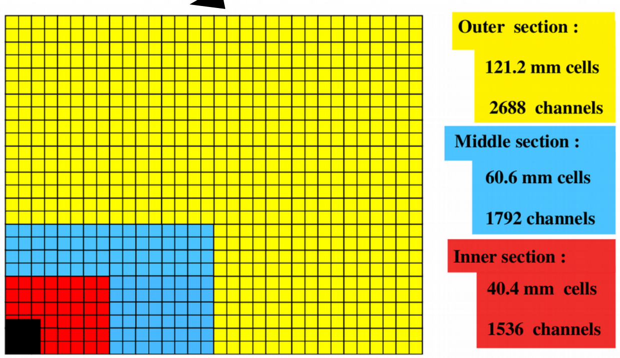


Calorimeters

Electromagnetic calorimeter

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

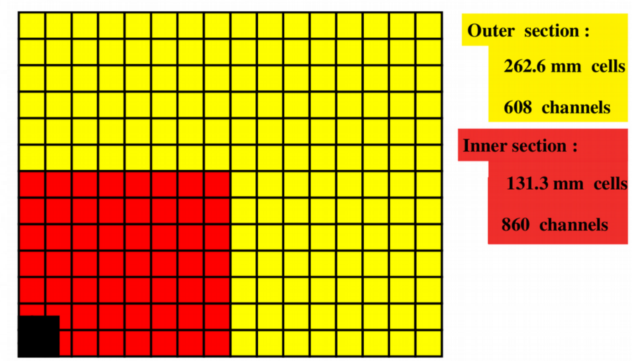
showers are detected by layers of scintillating fibers



Hadronic calorimeter

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

showers are detected by layers of scintillating tiles



Limitations due to saturation

Energy resolution

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

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

Not optimal for jets physics!

Inputs for jets reconstruction

- **Clusters isolated from tracks** (neutral particles)
- **Excesses of energy nearby tracks** (neutral recovery)

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.**
- **Uncorrelation with the dijet invariant mass is required**
- 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.