b-Tagging
(at multi-purpose LHC experiments)

Christophe Saout

CERN, Karlsruhe Institute of Technology (KIT)

- Introduction
- Track based algorithms
- Lepton based algorithms
- Vertex based algorithms
- Combined algorithms
- Mistag/Efficiency measurements
- Tracker misalignment
- Conclusions

Introduction

Why b-Tagging?

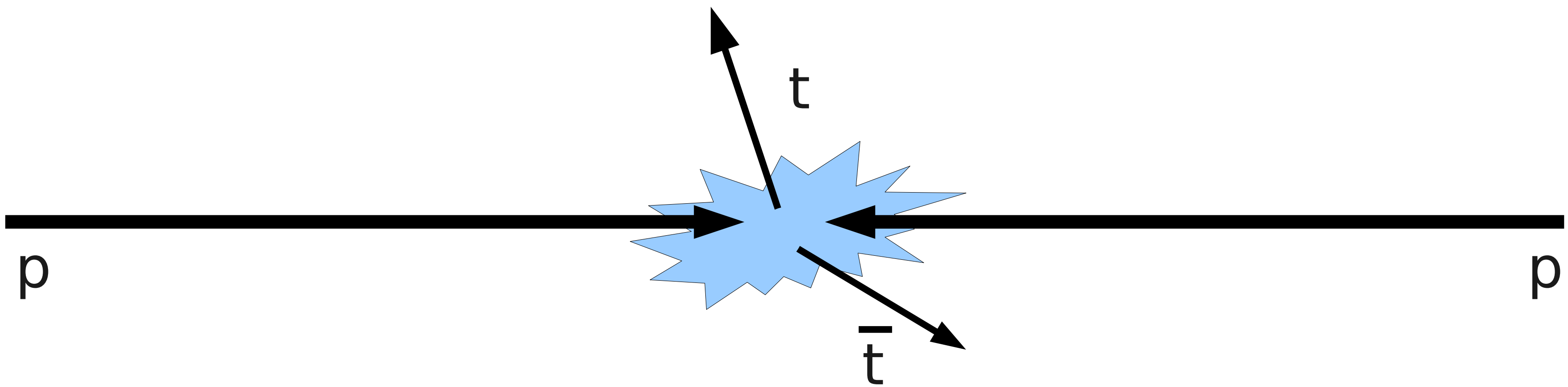
- b physics
  - b production *(cross-sections, ...)*
  - b decays *(mixing, oscillations, CP violation, CKM, ...)*
- Searches: b-tagging is of crucial importance for physics analyses that need to identify jets from **heavy quark** flavours *(b, top)*, e.g. **top, SUSY, Higgs**
  $\rightarrow$ ***background suppression!***

Where can b-Tagging be applied?

- Offline *(what we are focussing on)*
  - Full detector information
  - CPU-intense and precise reconstruction
  - Possibility to use very sophisticated algorithms
- Trigger
  - Limited information
  - Timing constraints $\rightarrow$ needs simple and fast algorithms

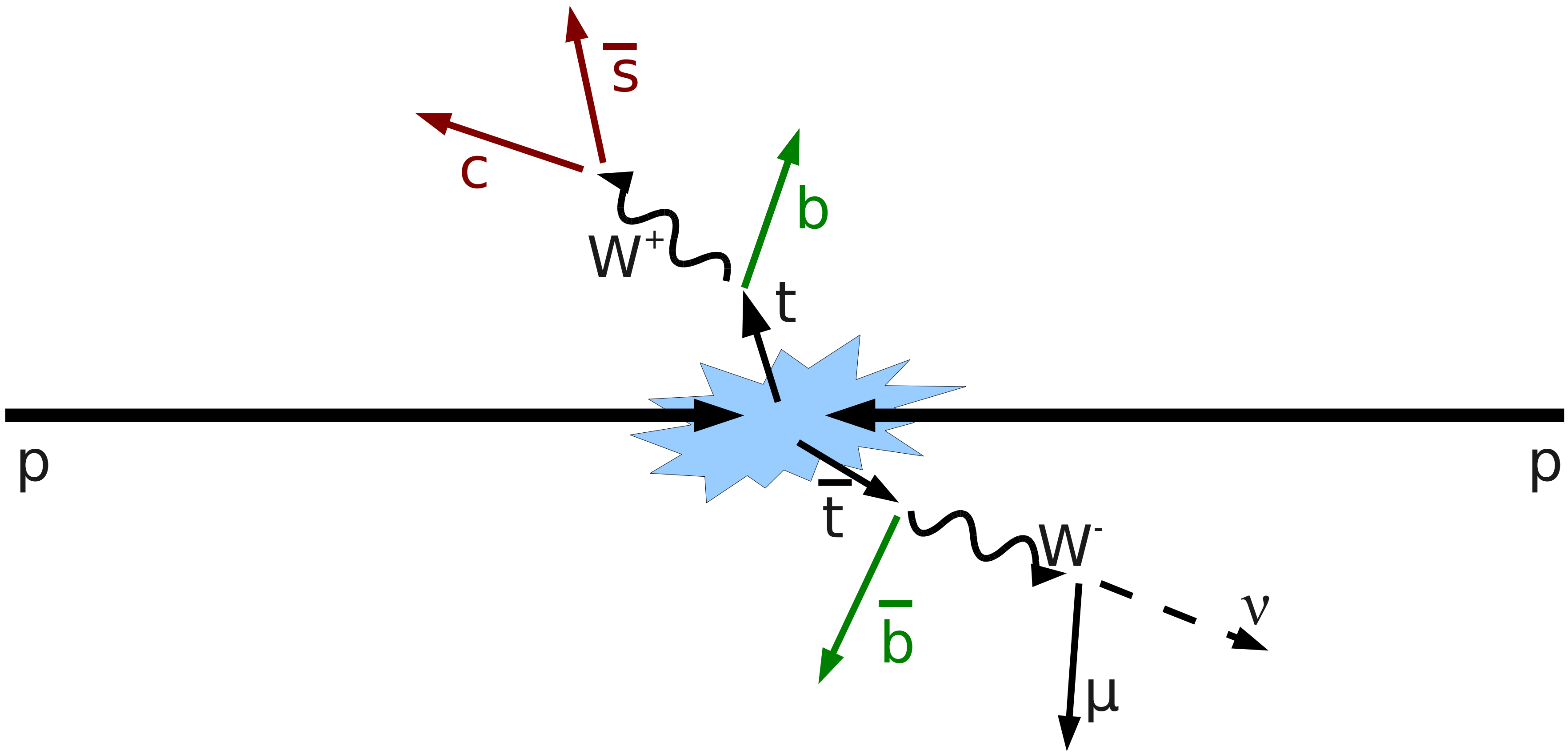
Introduction

An exemplary "big picture": $pp \to t\bar{t}$



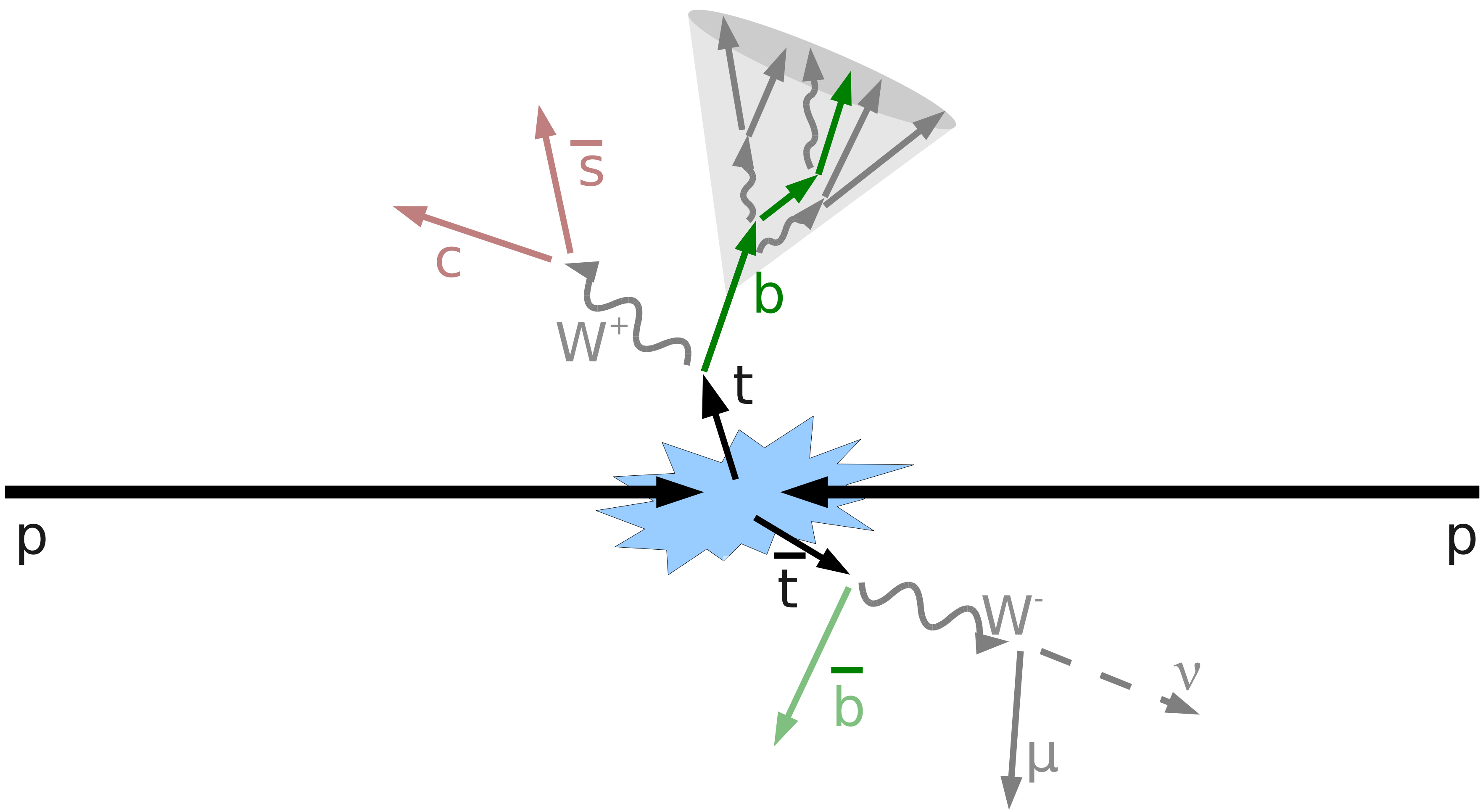
collision → heavy resonance production (here: via QCD)

Introduction



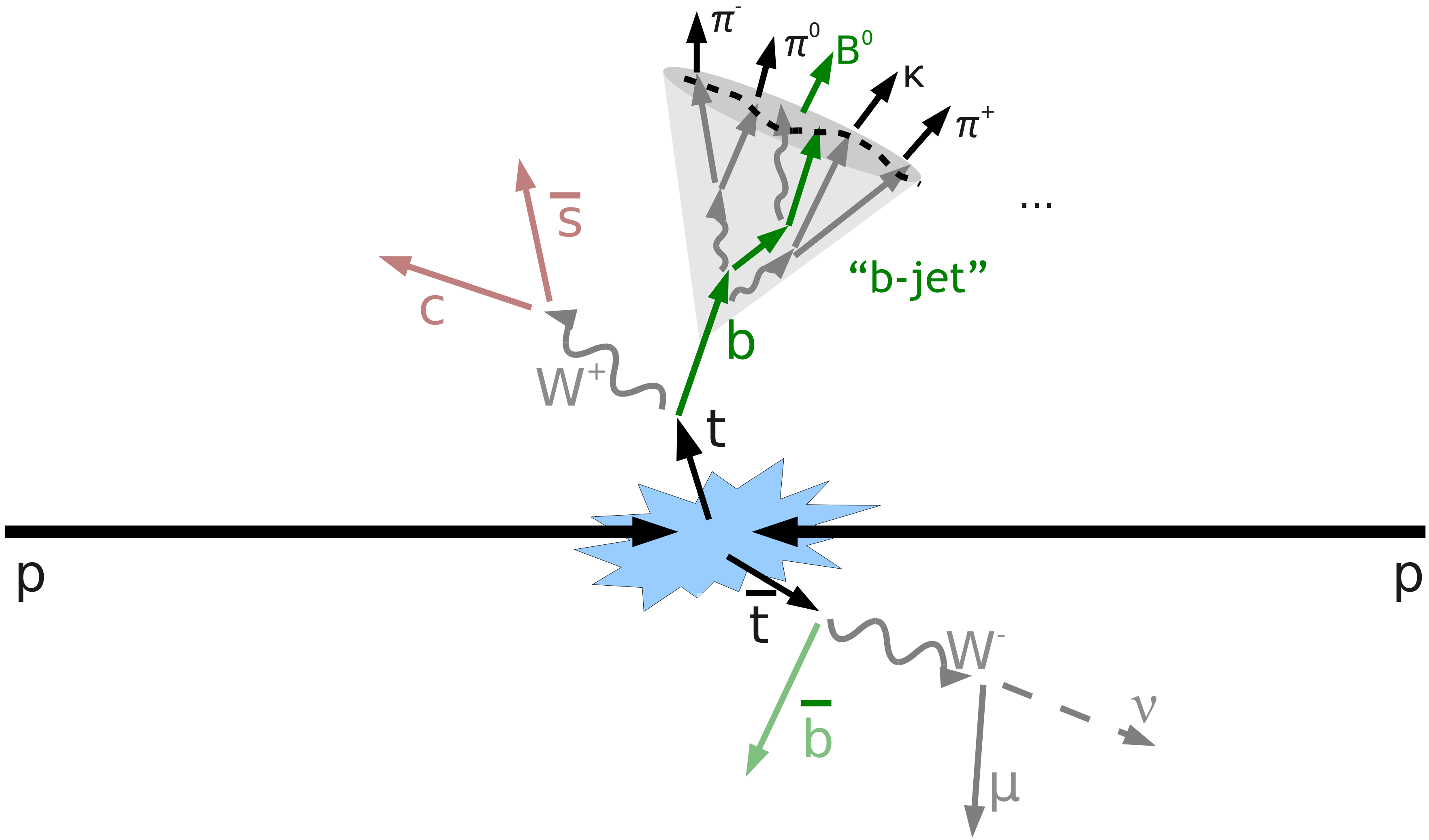
heavy resonance decay → b quark final states

Introduction



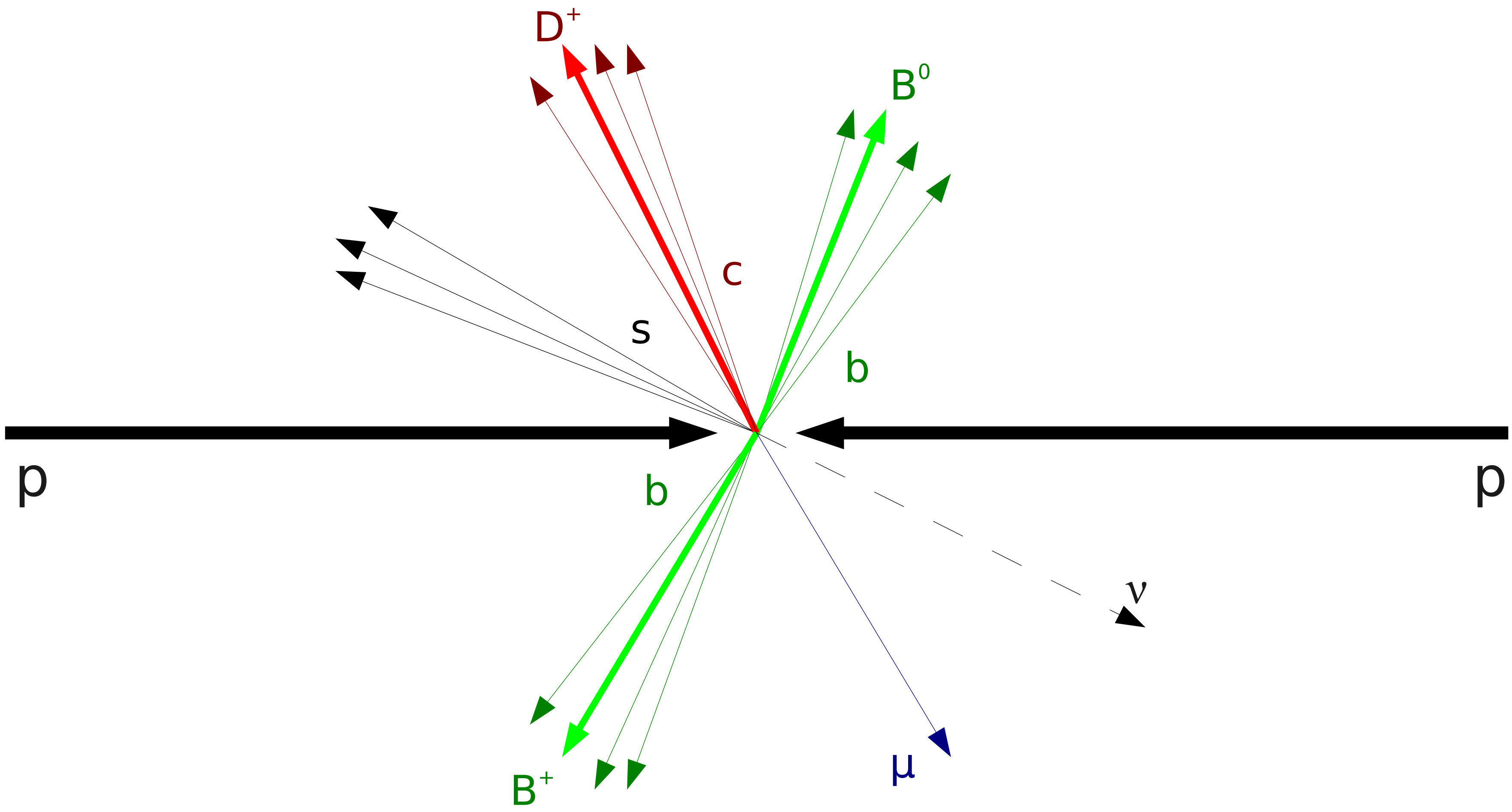
Parton Shower: *QCD radiation, gluon splitting, ...*

Introduction



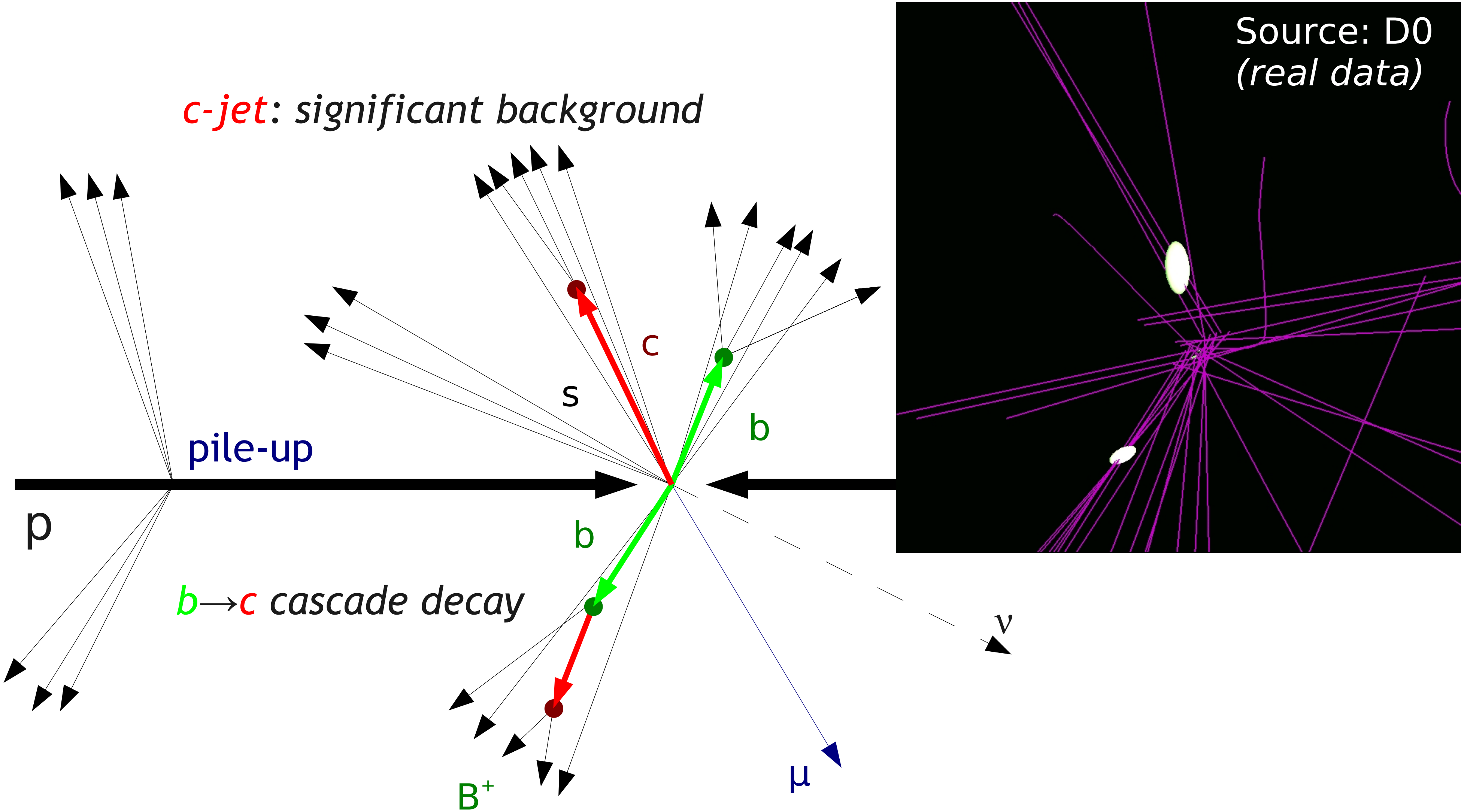
Hadronization: *jets of particles*

Introduction



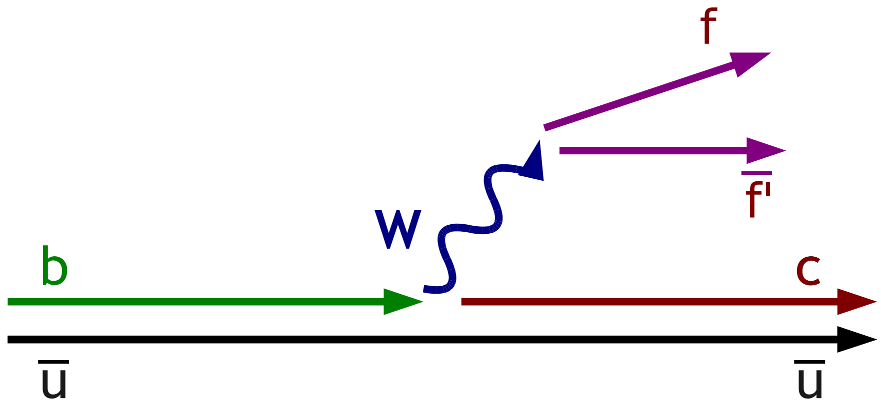
What we see in the detector: *jets of hadrons*

Introduction



heavy hadron decays: ***this is what allows us to do b-tagging***

b Decays



(a common b-hadron decay topology)

*b-hadron decays are **weak decays***

- only allowed through family change
- Strongly suppressed by CKM matrix
  → long lifetime!
  → b→c cascade decays preferred
- O(20%) of the decays occur semi-leptonically
  → lepton "in jet"

$$\begin{bmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{bmatrix} = \begin{bmatrix} 0.97419 \pm 0.00022 & 0.2257 \pm 0.0010 & 0.00359 \pm 0.00016 \\ 0.2256 \pm 0.0010 & 0.97334 \pm 0.00023 & 0.0415^{+0.0010}_{-0.0011} \\ 0.00874^{+0.00026}_{-0.00037} & 0.0407 \pm 0.0010 & 0.999133^{+0.000044}_{-0.000043} \end{bmatrix}.$$

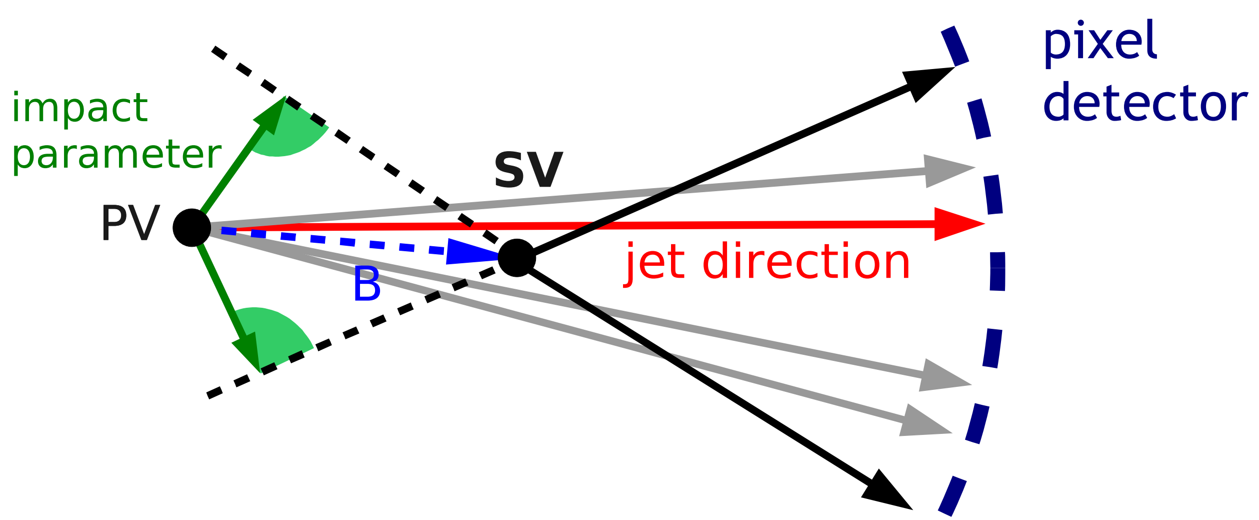
- *... but so are c-hadron decays...*

heavy hadron decays: ***this is what allows us to do b-tagging***

b Decays

- **b**-quarks significantly differ from light flavour quarks by:
  - mass: m = 4.2 GeV
  - lifetime: τ ≈ 1.5 ps → ~1.8mm *(at 20 GeV)* before decay
  - decay: weak, mostly into c-quarks *(→ 3rd decay)* → 20% into leptons
- Signatures:
  - tracks: high decay multiplicity, significant displacement *(~5 charged tracks on average)*
  - secondary vertices *(SV)*: tracks intersecting at a common vertex
  - b fragmentation: large portion of jet energy carried by b-hadron

→ need good tracking resolution at impact point! *(→ alignment!)*

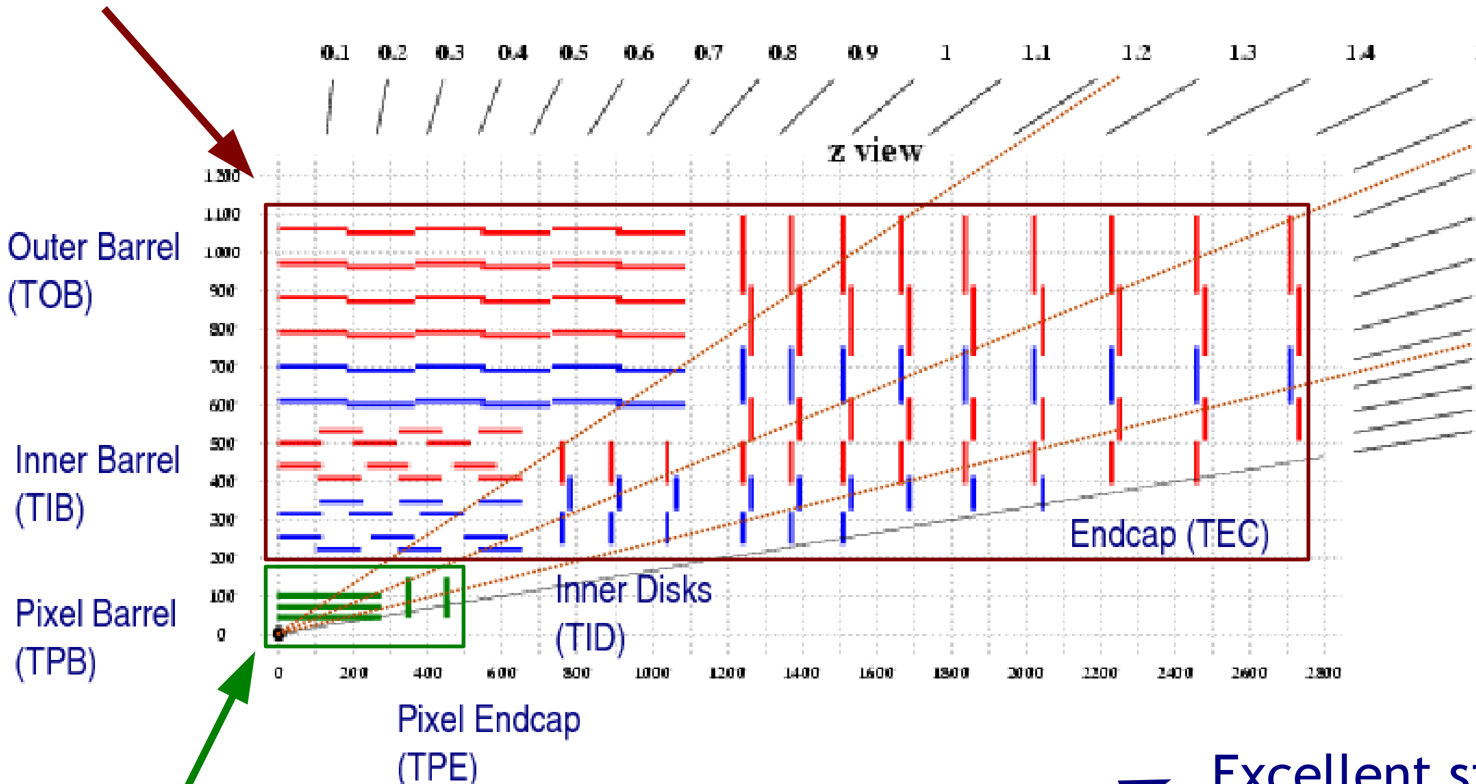


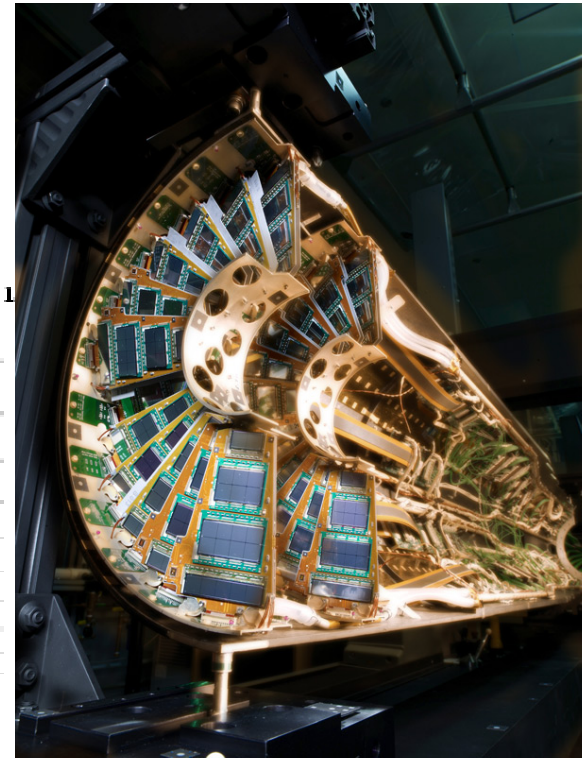
Limitations & Backgrounds

- Track reconstruction → **key role in b-tagging!**
  - impact parameter resolution
  - measurement errors
  - tracking in jets difficult
  - Efficiency / minimum p_T limitations
  - can only reconstruct charged hadrons
  - bad / fake tracks *(pattern recognition / combinatorics)*
- Real lifetime:
  - c-jets
  - Long-lived decays: κ_s^0, Λ^0, hyperons
- Leptons
  - Muon identification *(punch-through)*
  - Real muons *(e.g. muons from in flight π decays)*
  - electron ID in jets difficult (no calorimeter isolation)
- Other
  - Pile-up → *more than one collision → need z measurement!*
    → 2D pixel detector
  - "jet overlap" → *neighboring tracks isolation criteria*
  - QCD: $g \to b\bar{b}$ / $g \to c\bar{c}$ *splitting*

Tracking Detector

for instance: the CMS tracking system





part of the CMS pixel detector

Three(*) layers of pixel detectors:

- 768 modules
- Inner ring at r = 4.4cm
- 100 µm × 150 µm pixel size

Excellent single-point resolution:

- 10µm in r-φ, 20µm in z

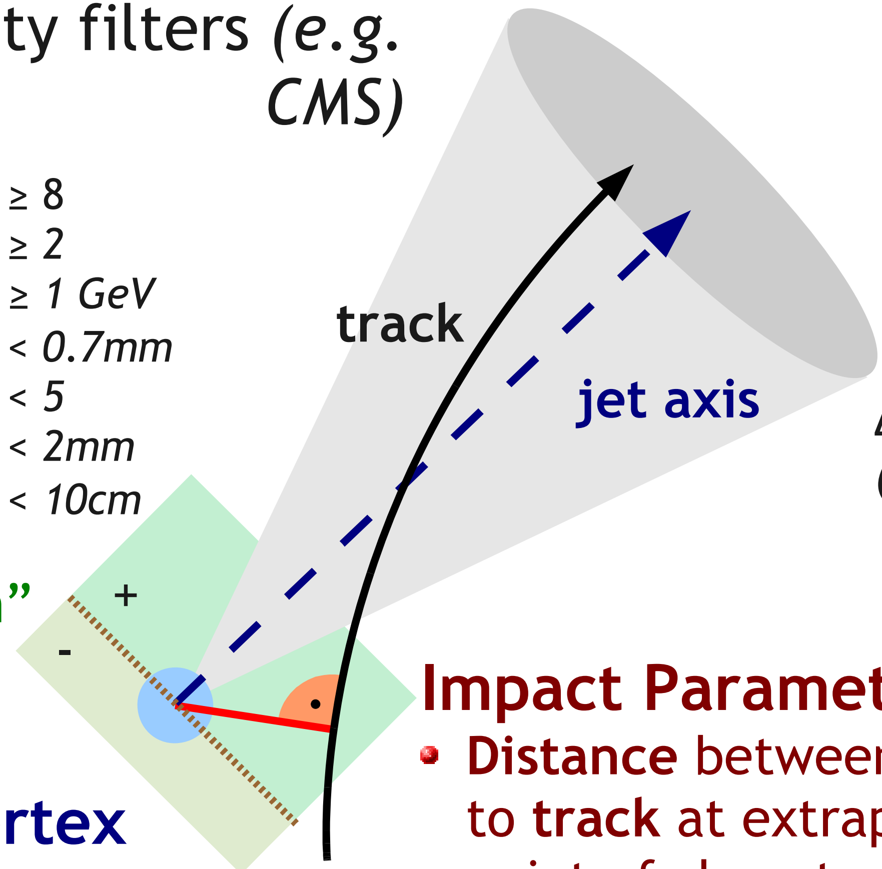
→ *good for b-tagging*

(*) in the central detector

Impact Parameters

Track quality filters *(e.g. CMS)*

total tracker hits	≥ 8
pixel hits	≥ 2
P_T	≥ *1 GeV*
jet axis dist.	< *0.7mm*
$\chi^2/ndof$	< *5*
IP_{xy}	< *2mm*
decay length	< *10cm*



ΔR cone around jet axis
(or something better)

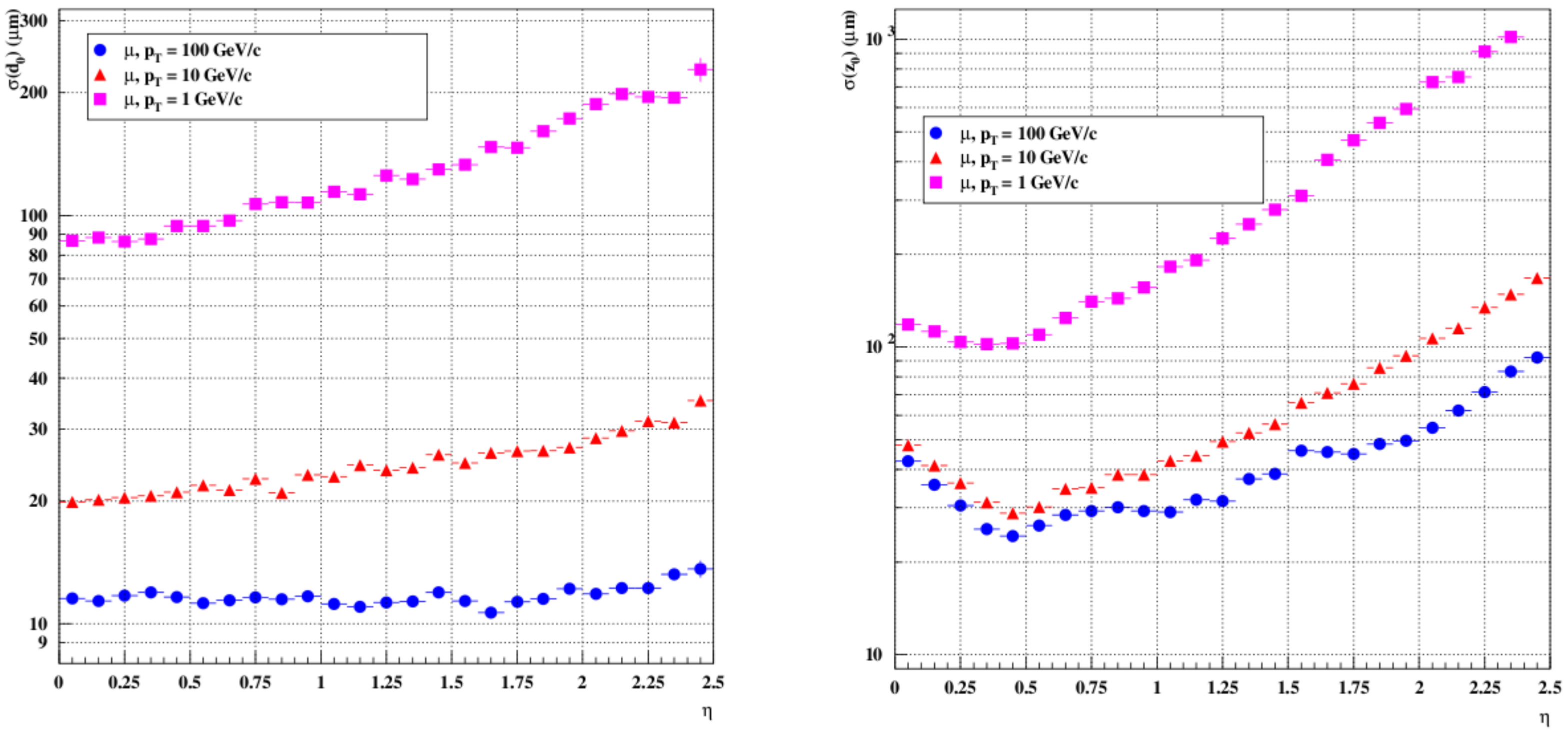
Primary Vertex

reconstructed using tracks in event

Impact Parameter

- **Distance** between **Primary Vertex** to **track** at extrapolated point of closest approach
- **Signed**
- Transverse **r-φ** or full **3D** value
- **Significance:** distance / error
  using full PV fit and track extrapolation covariance matrices

Impact Parameter Resolution

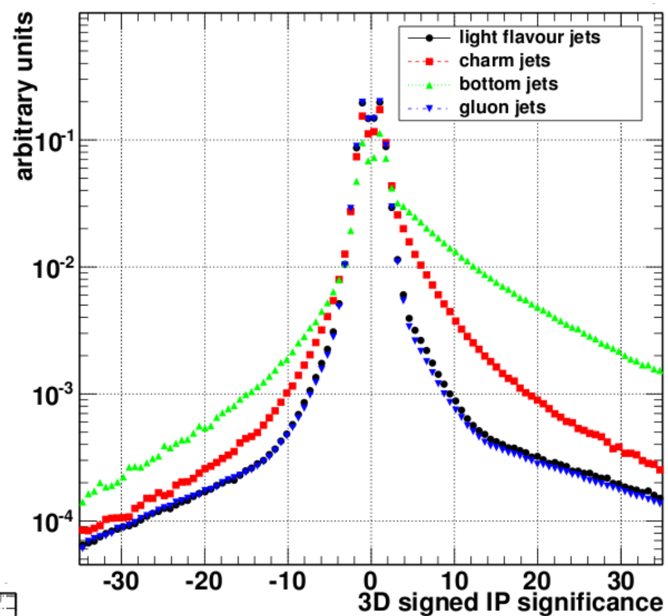


transverse and longitudinal
Impact Parameter resolution *(CMS)*

(multiple scattering, nuclear interactions, ...)

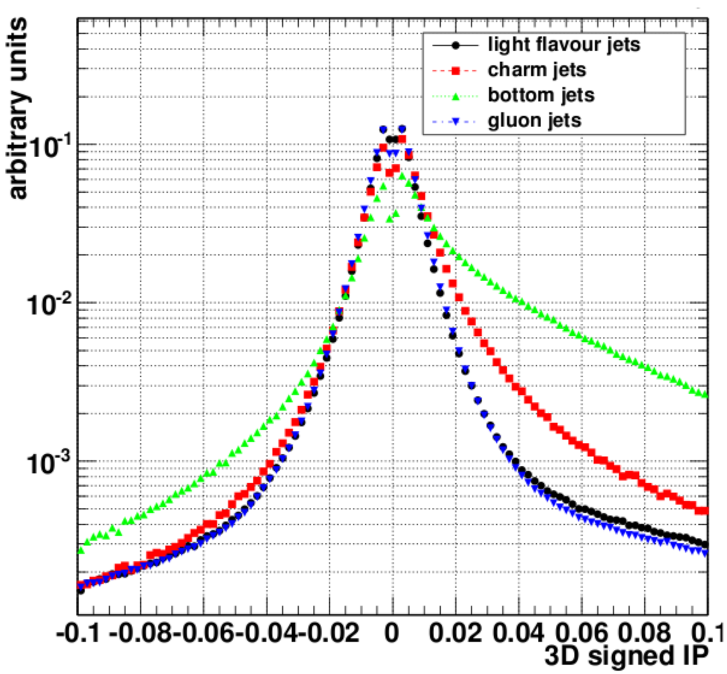
But also: *Tracking detector alignment!*

Impact Parameters



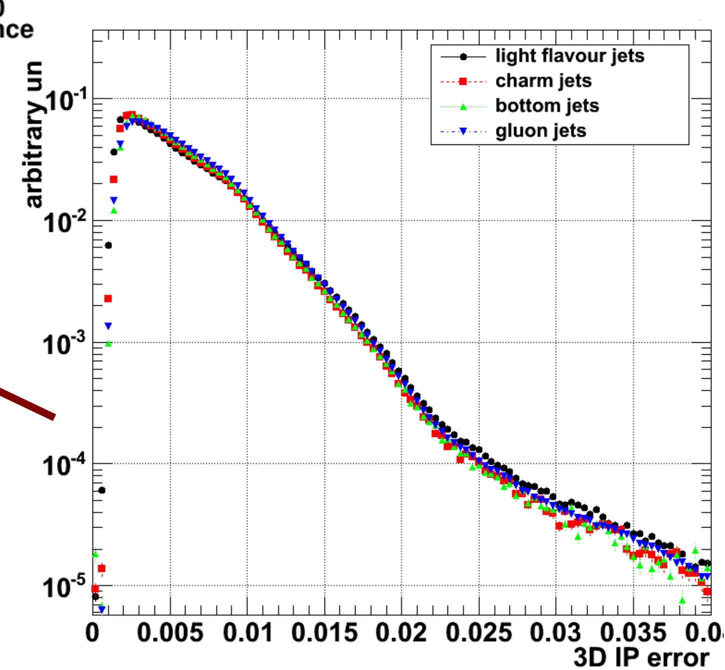
All tracks in jets

(> 20 GeV "corrected" Calorimeter jets)



significance

= value / error

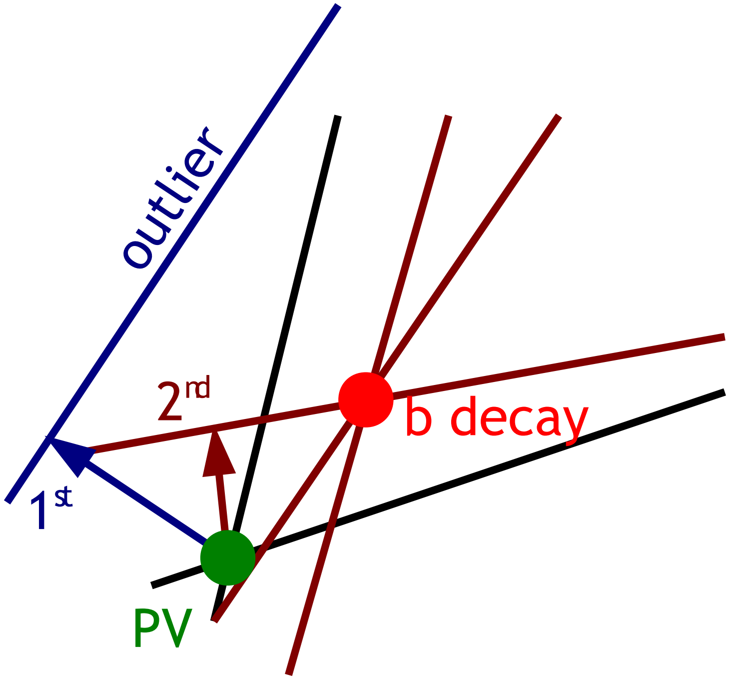


Simple I.P. based Algorithm

- Compute Impact Parameters for all tracks in jet
- Sort tracks by descending **Signed IP Significances** *(2D or 3D)*
- Select nth **track** *(hence called "Track Counting" in CMS)*
  - 2nd track → "high efficiency" tag
  - 3rd track → "high purity" tag
- Use IP significance as **discriminator**
- Simple, fast → suitable for online trigger

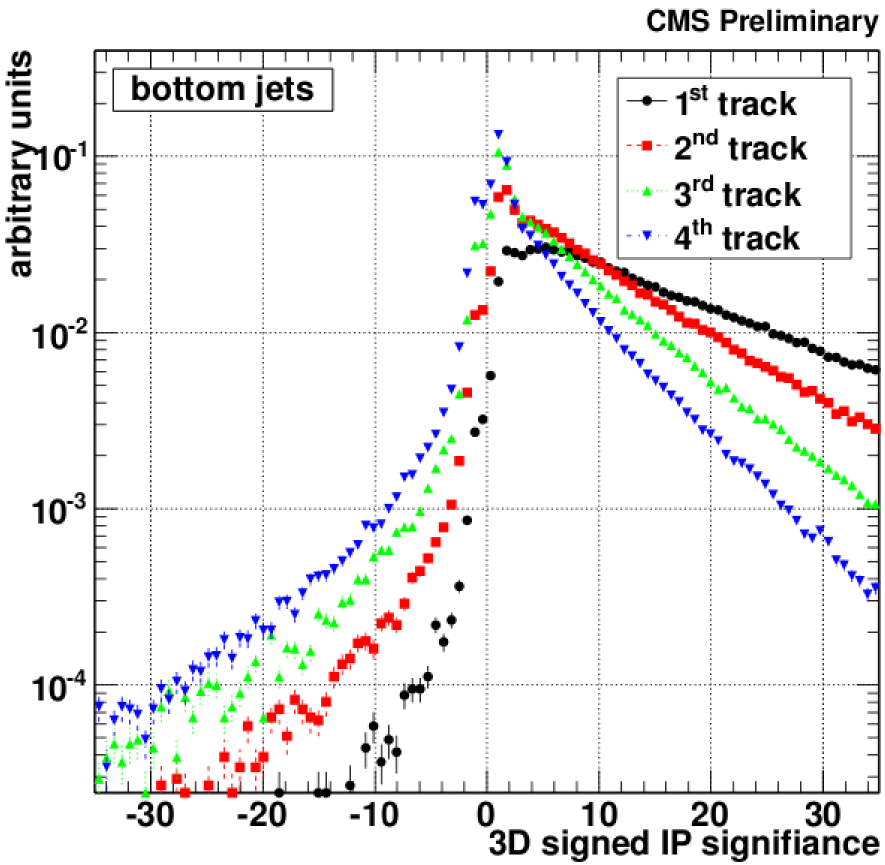
Eliminate non-b decay outliers

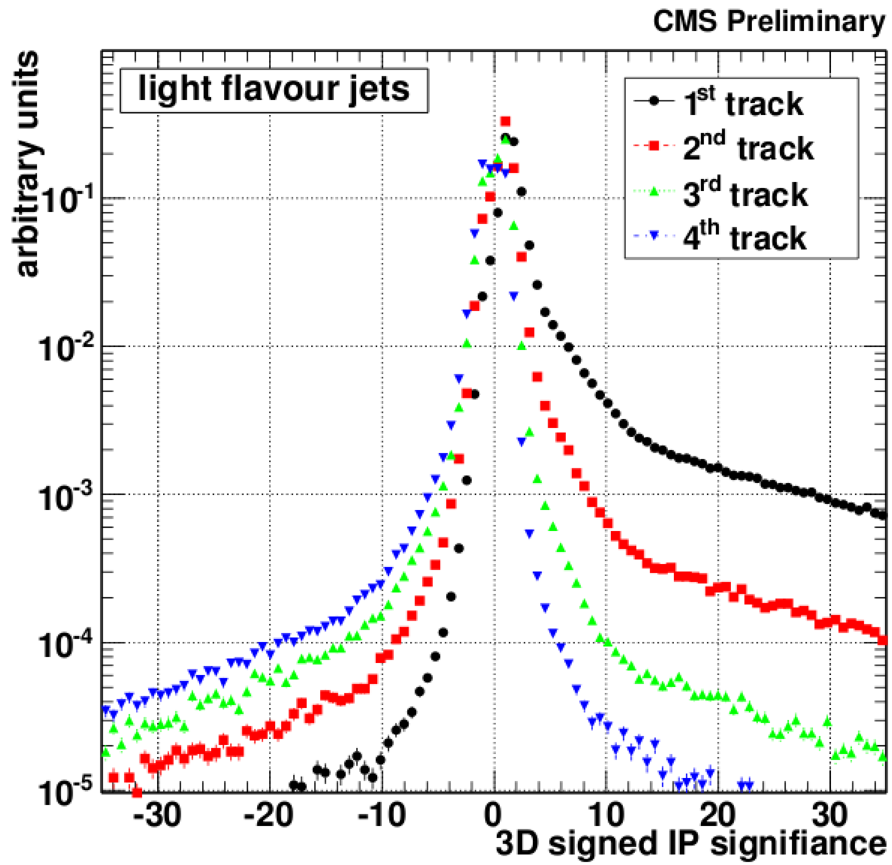
- Fake tracks
- "V^0" decays
- ...



simple & suitable for early data

Simple I.P. based Algorithm





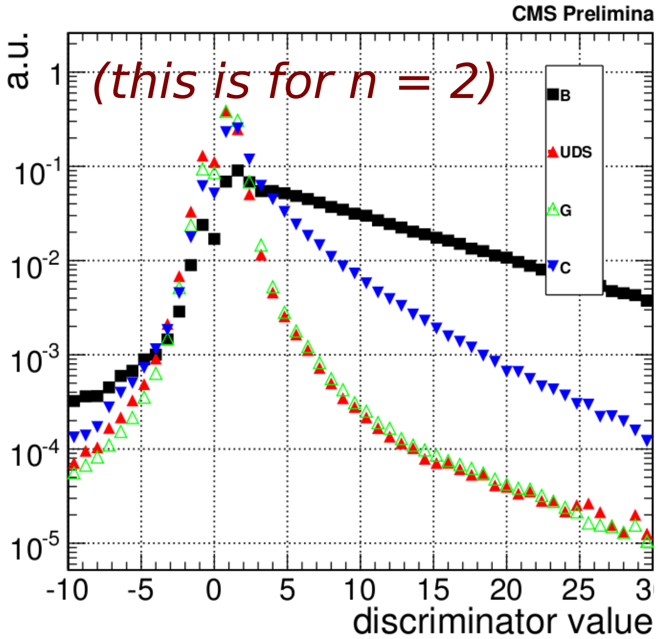
Choosing the optimal "n"-th track.

→ *trade-off between*

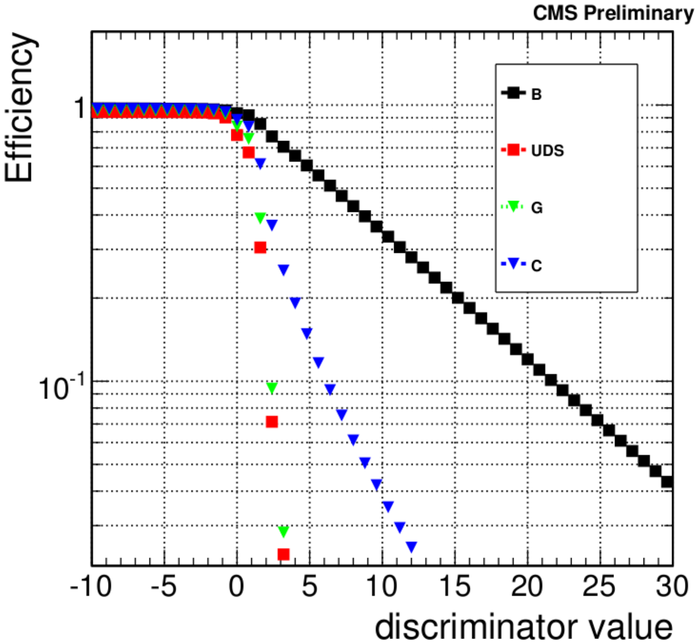
background reduction

b-tagging efficiency sacrifice

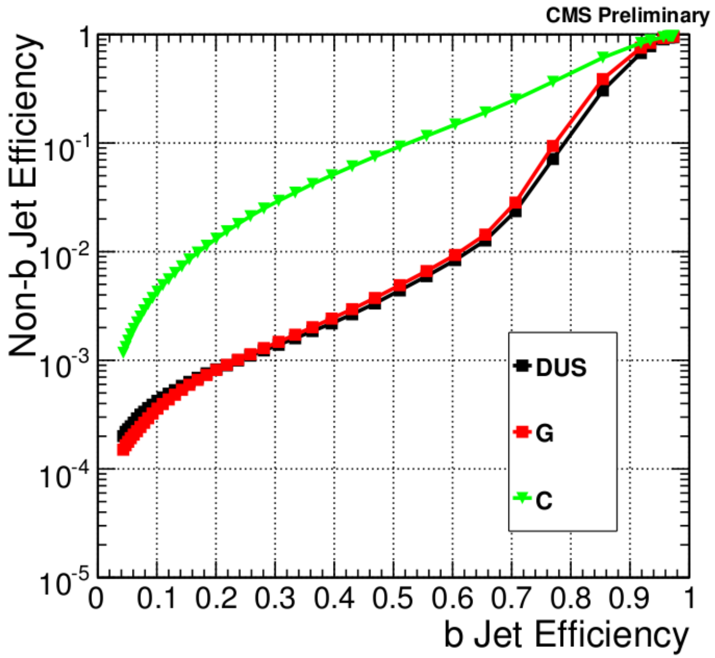
Simple I.P. Algo Performance



discriminator distribution



discriminator vs efficiency



efficiency vs. mistag rate

We can interprete the discriminator distributions with MC truth to determine working points and performance:

- We choose a working point, e.g. discr cut = 2.5
- We call a jet, which has a discr > 2.5 tagged
- The percentage of b-jets we correctly tag, is our efficiency
- The percentage of non-b jets we tag, is our mistag rate
- We can scan all working points and collect efficiencies and mistag rates

Track Probabilities

Method originally from LEP *(ALEPH)*:

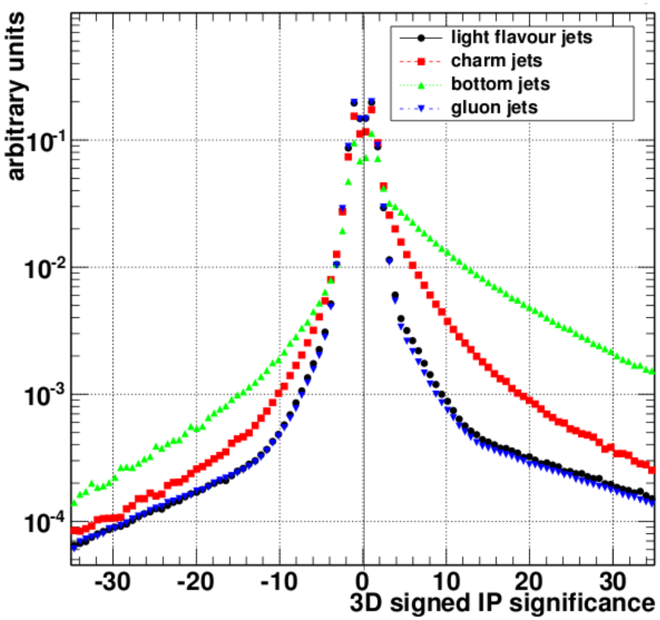
- For each *track*, determine the *probability* that it comes from the PV

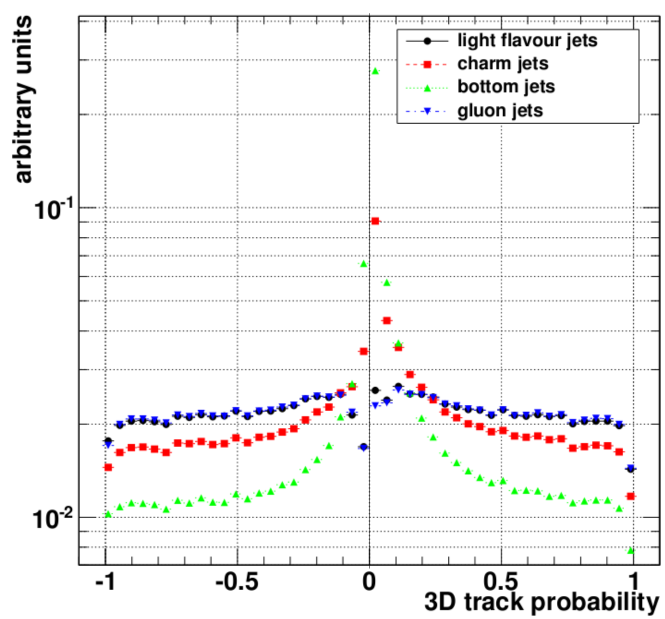
$$\rightarrow P(S') = \int_{S'}^{\infty} pdf(S)\, dS \qquad \textit{(pdf being the normalized IP Sig. distr.)}$$

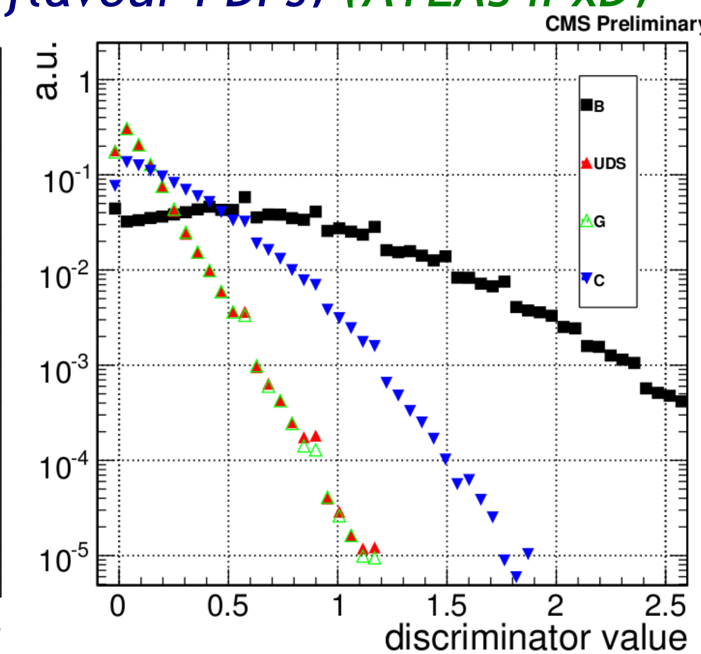
- Can use the measured negative signed IP distribution for calibration!
- Use information from all tracks by combining the probabilities

$$P_N = \pi \cdot \sum_{j=0}^{N-1} \frac{-\log \pi}{j\,!} \quad \text{with} \quad \pi = \prod_{i=1}^{N} P(S^i) \quad \text{and using} \quad discr = -\log P_N^+$$

- Alternative: Likelihood Ratio *(using both b and light flavour PDFs) (ATLAS IPxD)*







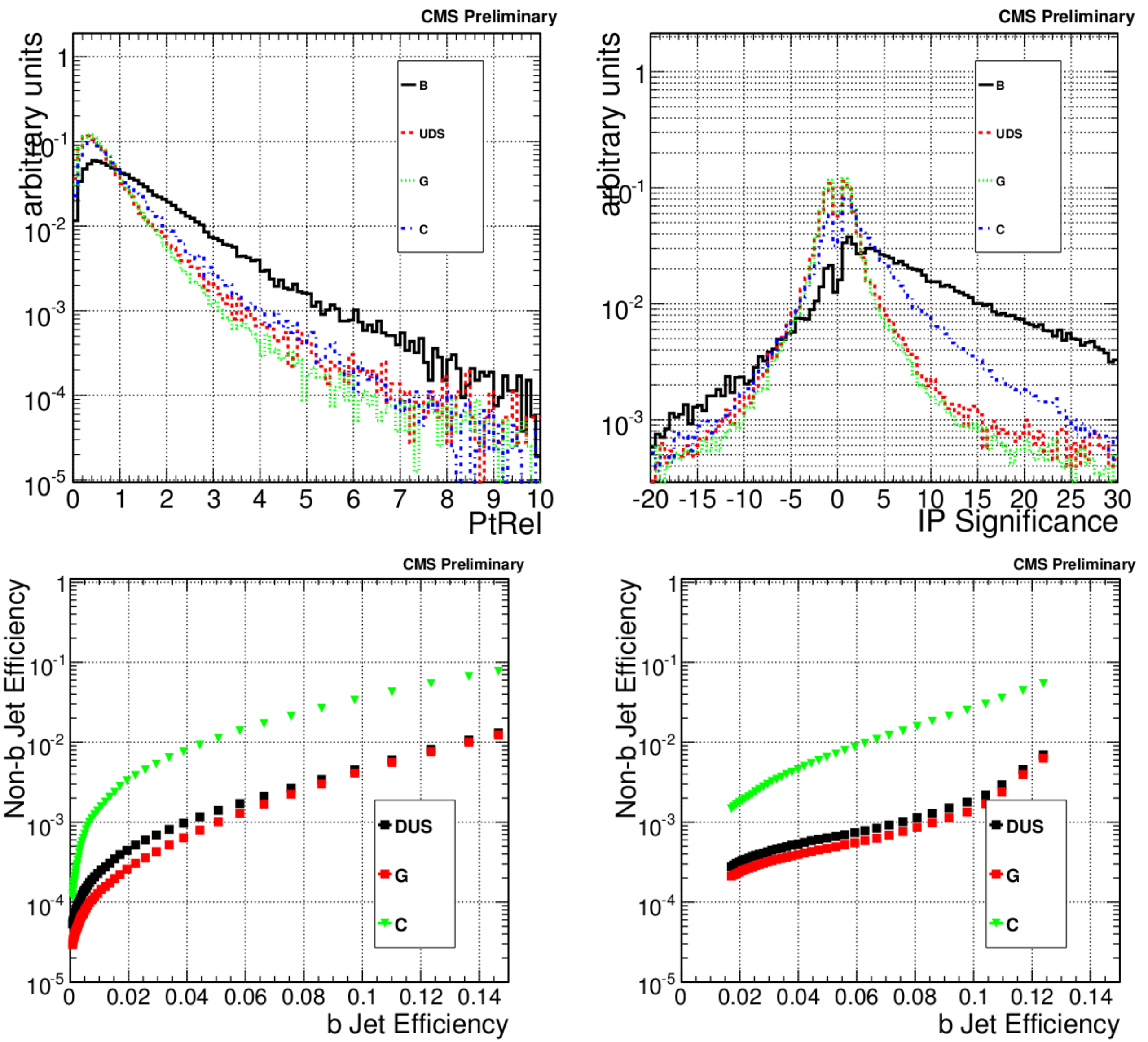
Soft Lepton b-Tagging

- As seen earlier, the branching ratio $b \to l$ *($l = e, \mu$)* is of the order of 20%
- This allows to use reconstructed leptons in jets as a complementary way of doing b-tagging
- This is particularly useful for b-tagging efficiency measurement methods
- Or can, combined with lifetime-based methods, improve efficiency

- Leptons from b-decays are *soft (as apposed to leptons from hard process)*
- They are within the jet and non-isolated *(typically $\Delta R < 0.4$)*
- The b-quark mass gives them a significant transverse momentum with respect to the jet direction *(called "p_T^{rel}")*
- As for other tracks from the b-hadron, the tracks are displaced

- Reconstruction of electrons in jets is particularly tricky *(Bremsstrahlung, calorimeter-track matching, calorimeter isolation, …)*
- Background: photon conversion, decays in flight

Soft Lepton b-Tagging



More complex alternatives:

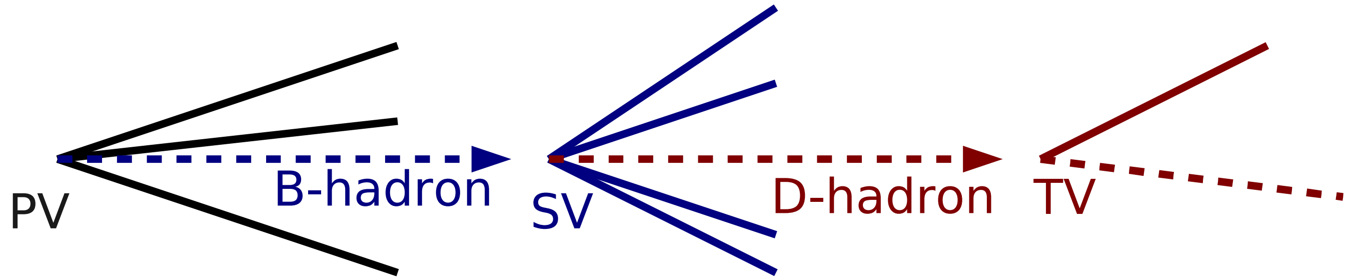
- Combine multiple variables using MVA techniques:
  - Pt rel
  - IP significance
  - Other kinematic variables *(angle, momentum ratios)*
  - or lepton ID discriminator
  - ...

For the two “simple” CMS soft muon taggers

Secondary Vertices

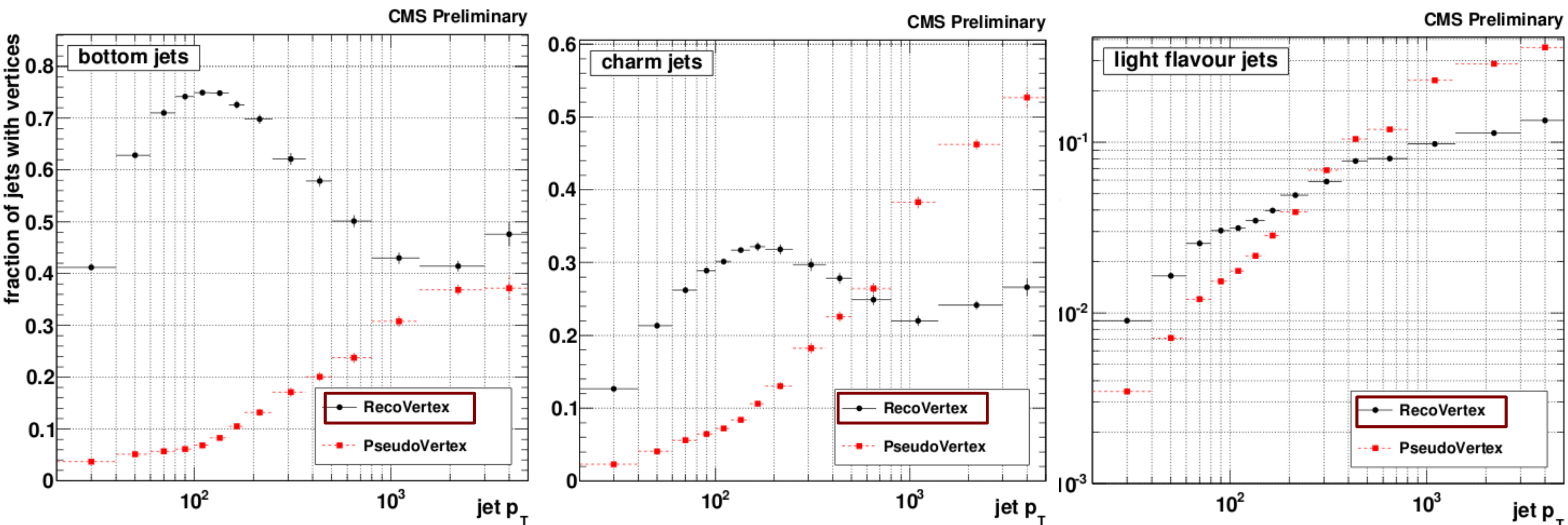
- Requiring the reconstruction of secondary vertices is an excellent indication for a heavy particle decay:
  → not only requiring displaced tracks, but also
  → displaced tracks intersecting at a common vertex

- Different approaches to vertex finding
  - Repeated fitting with outlier removal
  - Finding track pairs that form secondary vertices
  - Seeding vertices from compatible with a b-hadron direction hypothesis
  - ...



Secondary Vertices

- The reconstruction efficiency is strongly depending on the jet kinematics
  - low energy tracks → strongly reduced impact parameter resolution
  - low energy jet → jet is wieder → jet / track association efficiency loss
  - high energy jet → high track density → track reconstruction problems → bad / fake tracks

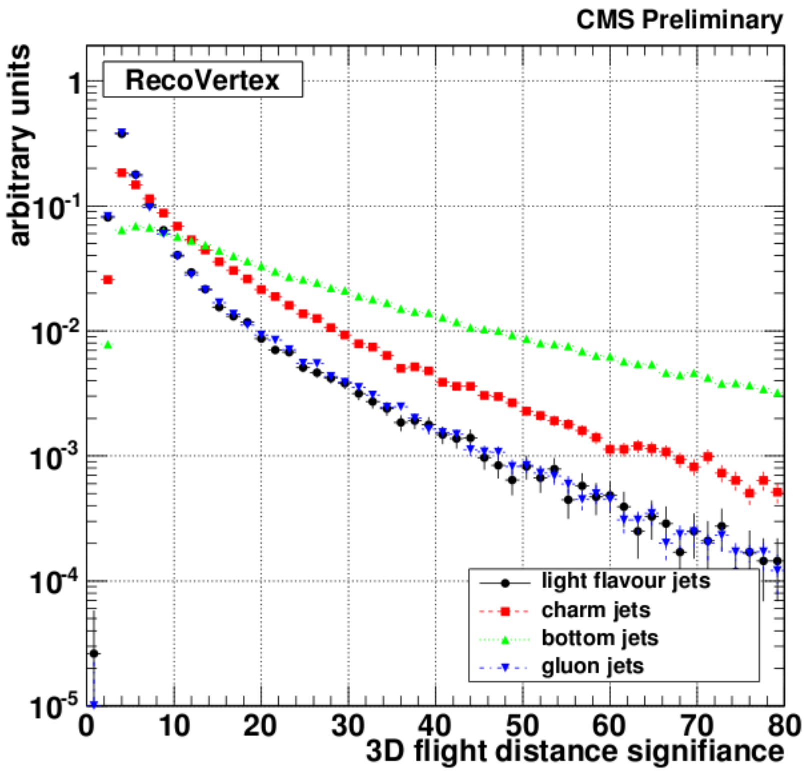


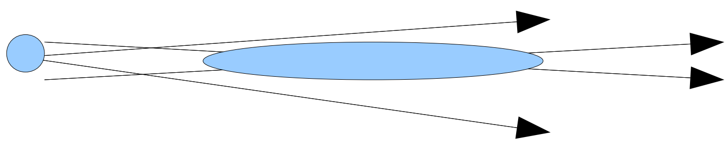
(PseudoVertex: no real vertex, but ≥ 2 displaced tracks)

calorimeter cone jets with ΔR ≤ 0.5 tracks associated to jet if ΔR ≤ 0.3

Secondary Vertices

- The vertex position can be used to probe the lifetime:
  - The flight distance: $D^{3D} = |\overrightarrow{SV} - \overrightarrow{PV}|$ *(in 2D or 3D)*
  - Similarly to the impact parameters: flight distance significance
  - (Can also define a flight distance sign and use vertices behind the primary vertex to measure light flavour fake vertex distributions)

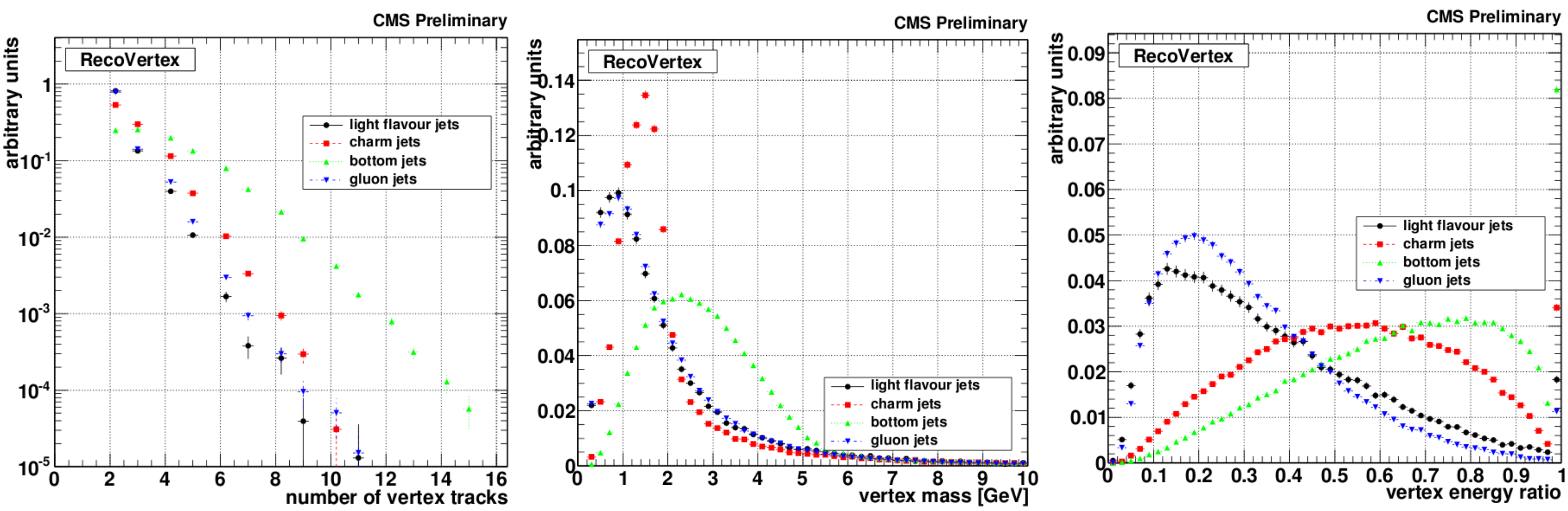




Note that the error on the vertex position longitudinally along the jet is usually very large

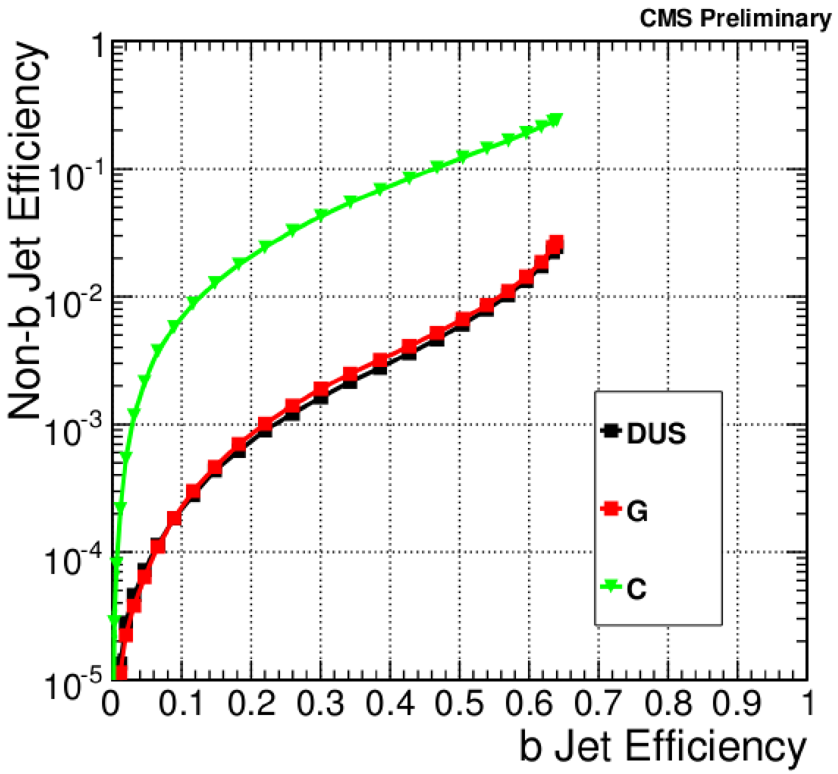
Secondary Vertices

- In addition to the existence one or more secondary vertices, one gets an additional amount of observables: One...
  - can obtain the charged decay multiplicity
  - can compute an invariant vertex mass *(from charged tracks only)*
  - can probe the fragmentation function: energy fraction carried by vertex



Secondary Vertex b-Tagging

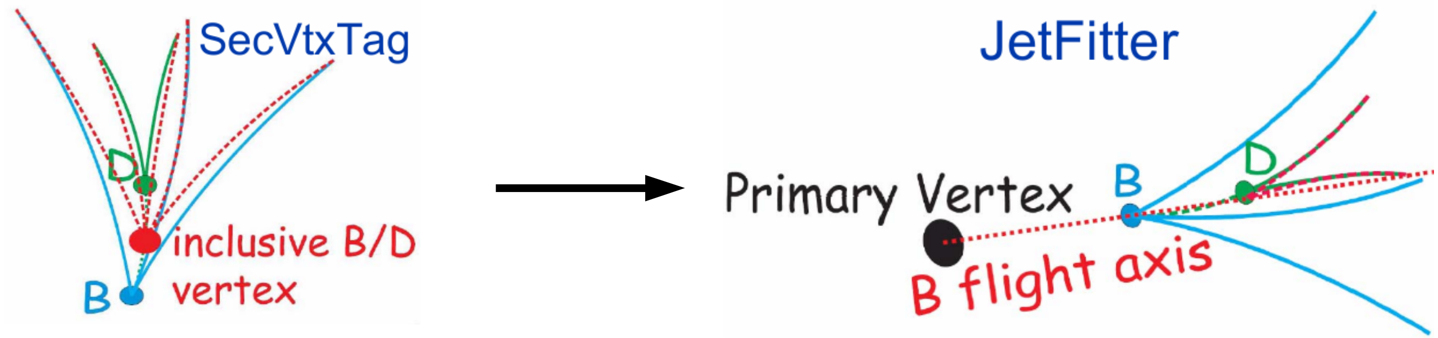
- The fact that a secondary vertex has been reconstructed can already be used as a b-tag *(e.g. CDF at Tevatron)*
- One can define different working points by tightening requirements
  - minimum nuber of tracks
  - track qualitites
  - vertex fit quality
  - vertex flight distance significance → e.g. as discriminator to cut on:



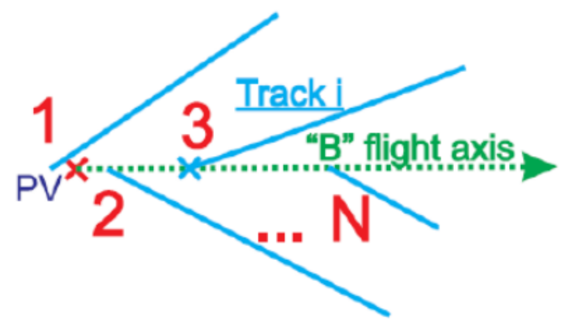
ATLAS “Jet Fitter”

First presented by SLD under the acronym “Ghost Track” b-tagging:

- Instead of finding vertices directly, first attempt to reconstruct the b-hadron flight line *(takes into account b-c decay chain)*



- This is done by first fitting a straight line (“the ghost track”) that minimizes the distance to all track:



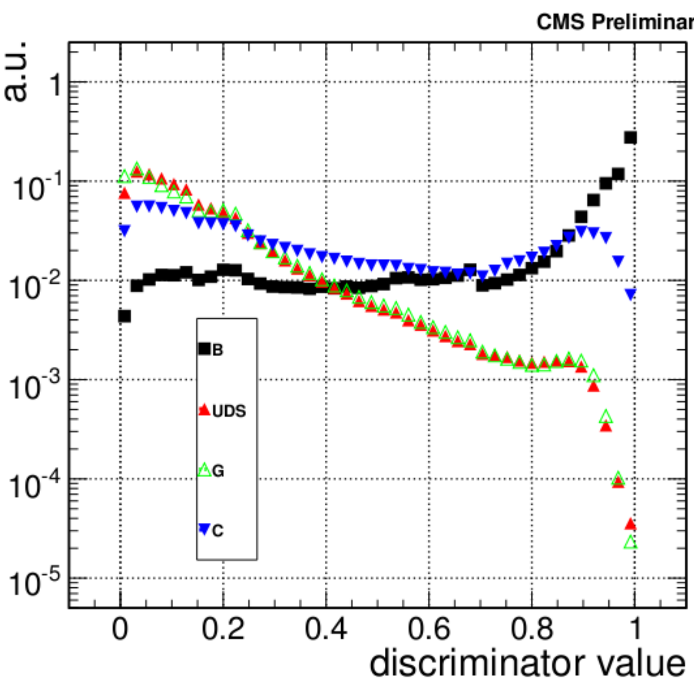
- Then tracks can be clustered along the “ghost track” to form vertices

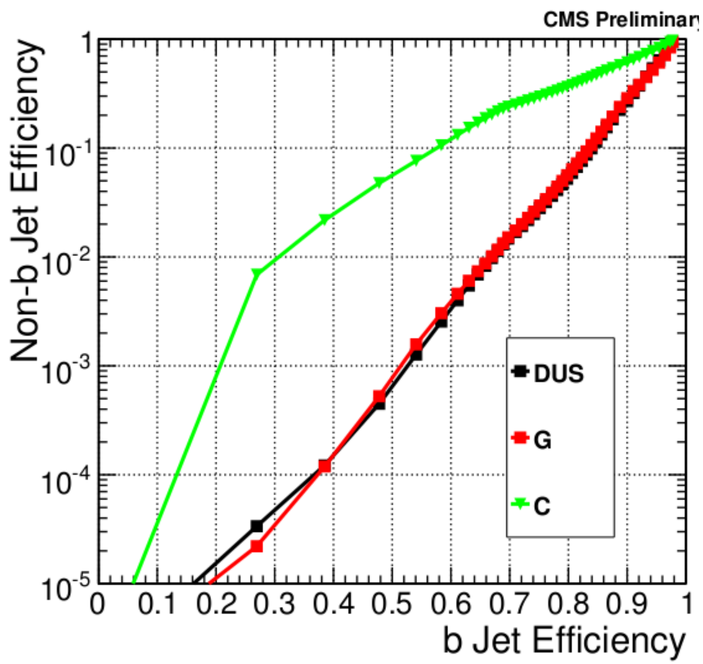
→ b-c decay chains can be reconstruced, including “1-track vertices”
decay topology as additional observable → higher efficiency/purity

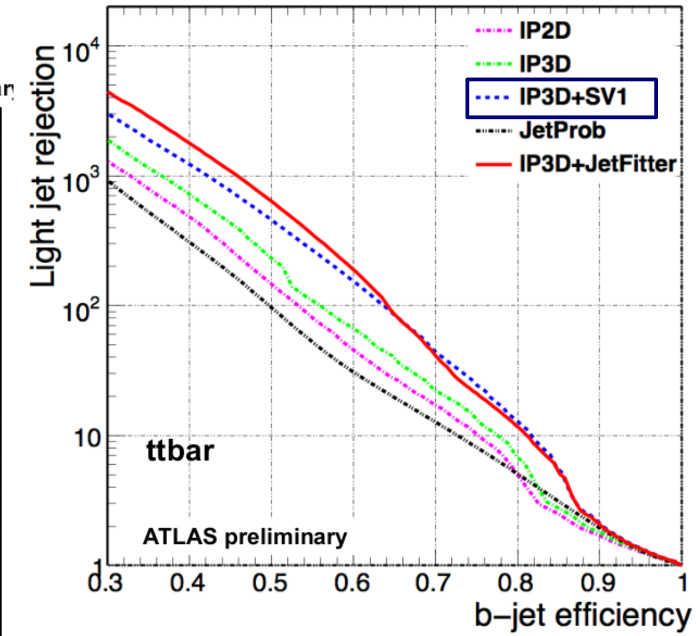
Combined Algorithms

- Instead of using individual observables → combine variables
- "track probability" such an example *(uses background hypothesis only)*
- More common: Growing family of "Multivariate Analysis Techniques"
- e.g. Likelihood Ratio or more advanced: Neural Networks, Boosted Decision Trees, …

  → machine-learning *(from signal & background samples)*







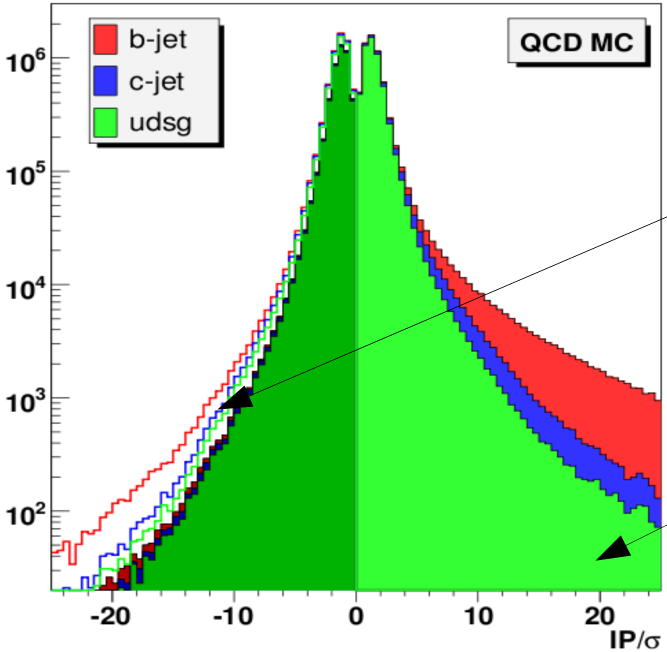
e.g. CMS "Combined Secondary Vertex" algo *(lifetime information, up to 10 variables)*
In 3 categories: IP only, IP + PseudoVertex, IP + SV → whole b-tag eff. range covered

Mistag Measurement (on Data)

Efficiency/Mistag measurements on data are much more complicated
Also, unfortunately, there are no simple "tag & probe" type methods

Light flavour jet observables approximately symmetric around zero
→ sign of b-tagging observables *(signed IP, signed distance)*
→ mistag measurement using negative side

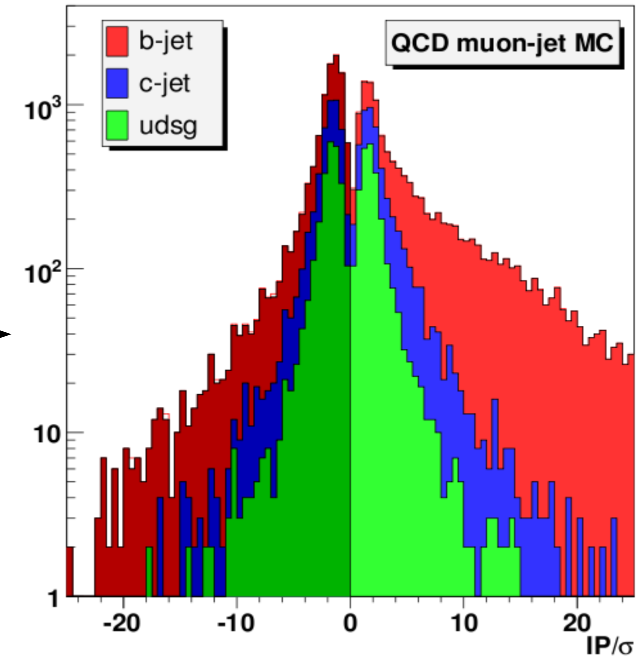
→ **Idea:** flip the sign and you get a b-tagging algorithm that will behave *as if the jet was always a light flavour jet*
(in principle works for IP + SV observables and even combined taggers)



IP of all tracks

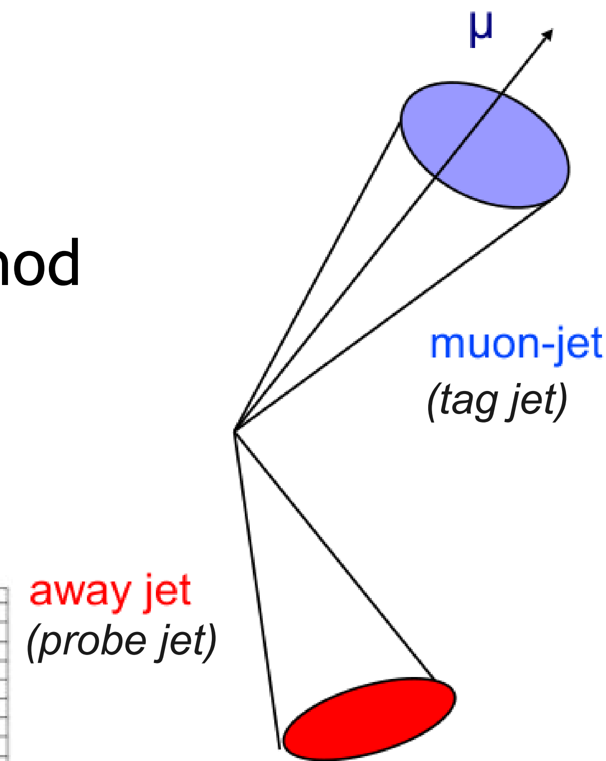
"Sign flip" effect can be estimated from b-enriched QCD *(muon in jet)*

Remaining asymmetry: "R_{light}" factor...
(reco effects, conversions, Long-lived decays, ...)



Efficiency Measurement

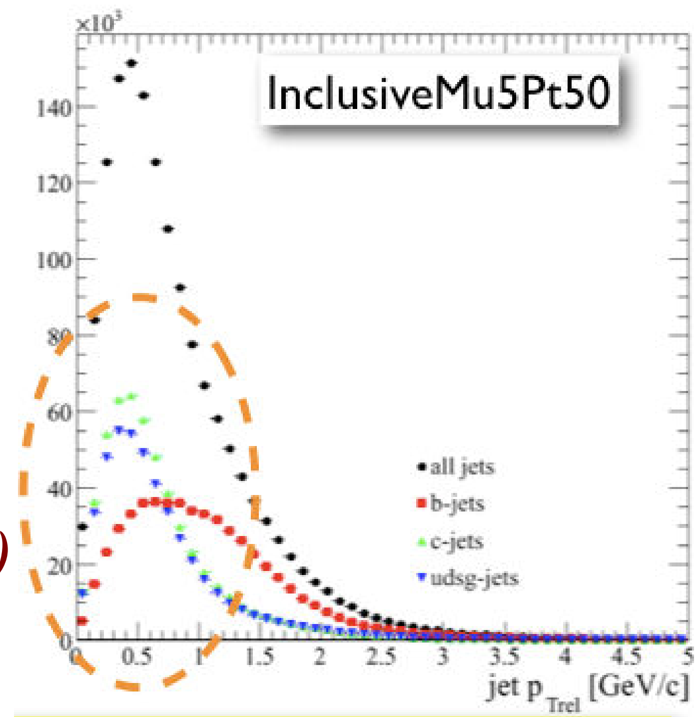
A simple tag & probe method with flavour content fit



- use di-jet events
- assume b-jets are produced in pairs
- require moft muon tag in one jet
- probe the b-tag in the other jet

“p_T^{rel} ” method

- Fit heavy flavour content using muon p_T^{rel} templates



- Pro: simple
- Con: MC dependence
  (templates easily biased)

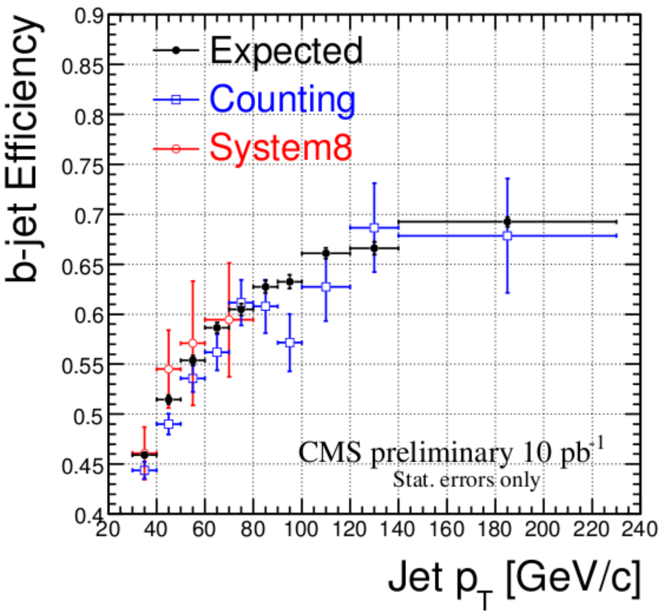
Efficiency Measurement

You can overcome these efficiencies by solving for both tag & probe taggers simulatenously

→ System of 8 equations → System 8 method (developed at D0)

- Only slight MC dependency: "probe" / "tag" tagger correlations

  → use uncorrelated taggers *(e.g. soft muon & lifetime)*



muon-in-jet+away-jet
muon-in-jet+tagged-away-jet

$$\begin{cases} n = n_b + n_{cl} \\ p = p_b + p_{cl} \end{cases}$$

"probe" tagger

$$\begin{cases} n^{\text{tag}} = \varepsilon_b^{\text{tag}} n_b + \varepsilon_{cl}^{\text{tag}} n_{cl} \\ p^{\text{tag}} = \beta\, \varepsilon_b^{\text{tag}} p_b + \alpha\, \varepsilon_{cl}^{\text{tag}} p_{cl} \end{cases}$$

"tag" tagger

$$\begin{cases} n^{p_{Trel}} = \varepsilon_b^{p_{Trel}} n_b + \varepsilon_{cl}^{p_{Trel}} n_{cl} \\ p^{p_{Trel}} = \delta\, \varepsilon_b^{p_{Trel}} p_b + \gamma\, \varepsilon_{cl}^{p_{Trel}} p_{cl} \end{cases}$$

"tag"+"probe" tagger

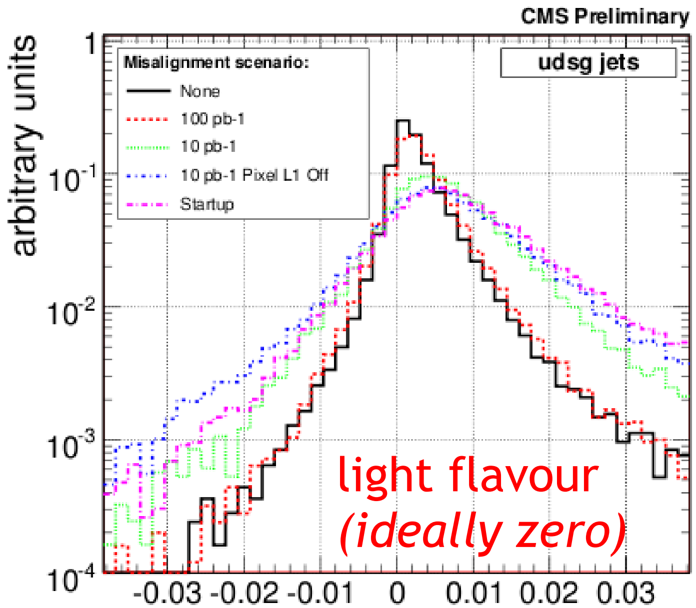
$$\begin{cases} n^{\text{tag},p_{Trel}} = \kappa_b\, \varepsilon_b^{\text{tag}} \varepsilon_b^{p_{Trel}} n_b + \kappa_{cl}\, \varepsilon_{cl}^{\text{tag}} \varepsilon_{cl}^{p_{Trel}} n_{cl} \\ p^{\text{tag},p_{Trel}} = \kappa_b\, \beta\, \delta\, \varepsilon_b^{\text{tag}} \varepsilon_b^{p_{Trel}} p_b + \kappa_{cl}\, \alpha\, \gamma\, \varepsilon_{cl}^{\text{tag}} \varepsilon_{cl\,Trel}^{p} p_{cl} \end{cases}$$

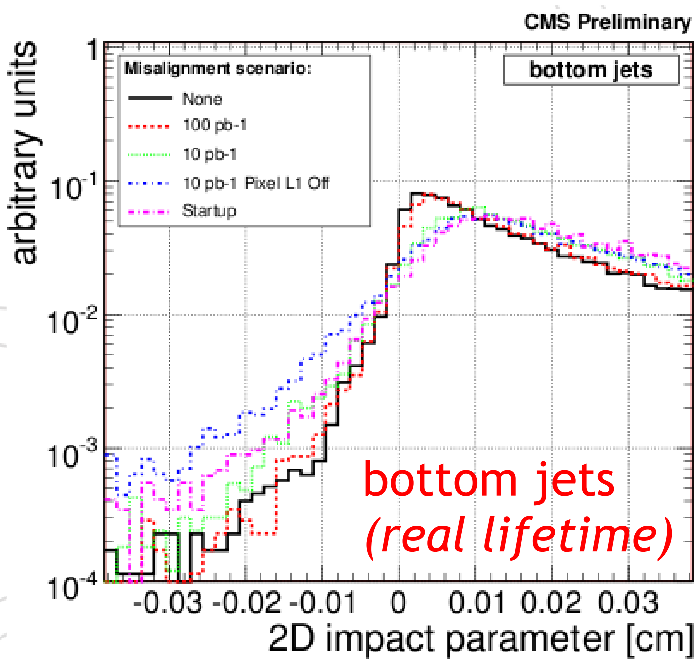
Solutions: $\varepsilon_b^{tag}, \varepsilon_{cl}^{tag}, \varepsilon_b^{\mu}, \varepsilon_{cl}^{\mu}, n_b, n_{cl}, p_b, p_{cl}$

Tagger efficiency

SLT efficiency

Tracker Misalignment

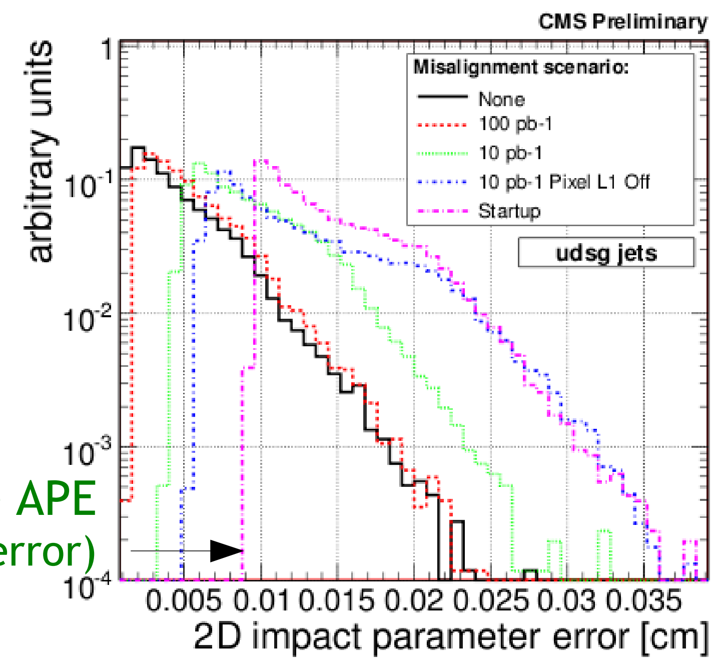




Distributions for the 2nd-highest signed impact parameter in jets:
(→ "track counting" b-tag algorithm)

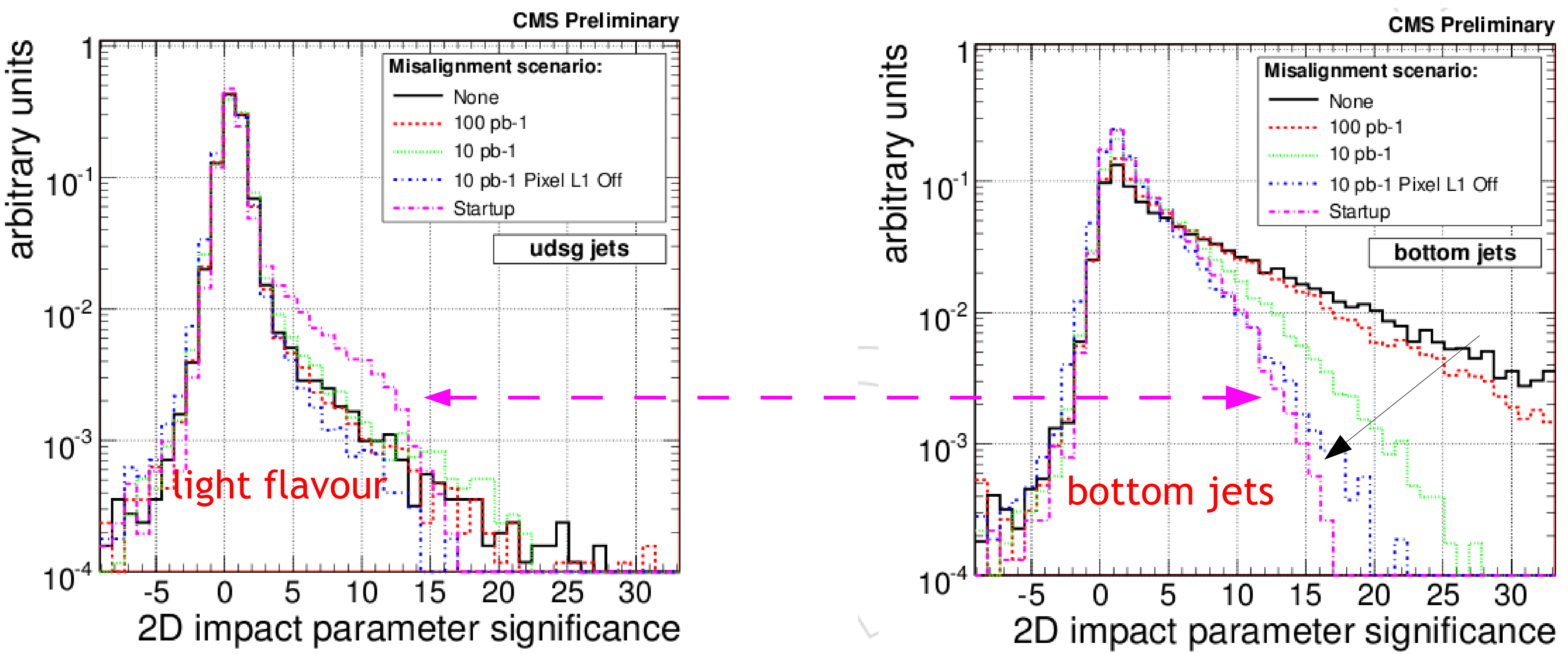
error on the IP measurement
dominated by pixel hit resolution, extrapolation from innermost hits

Effect of the APE
(alignment position error)



Tracker Misalignment

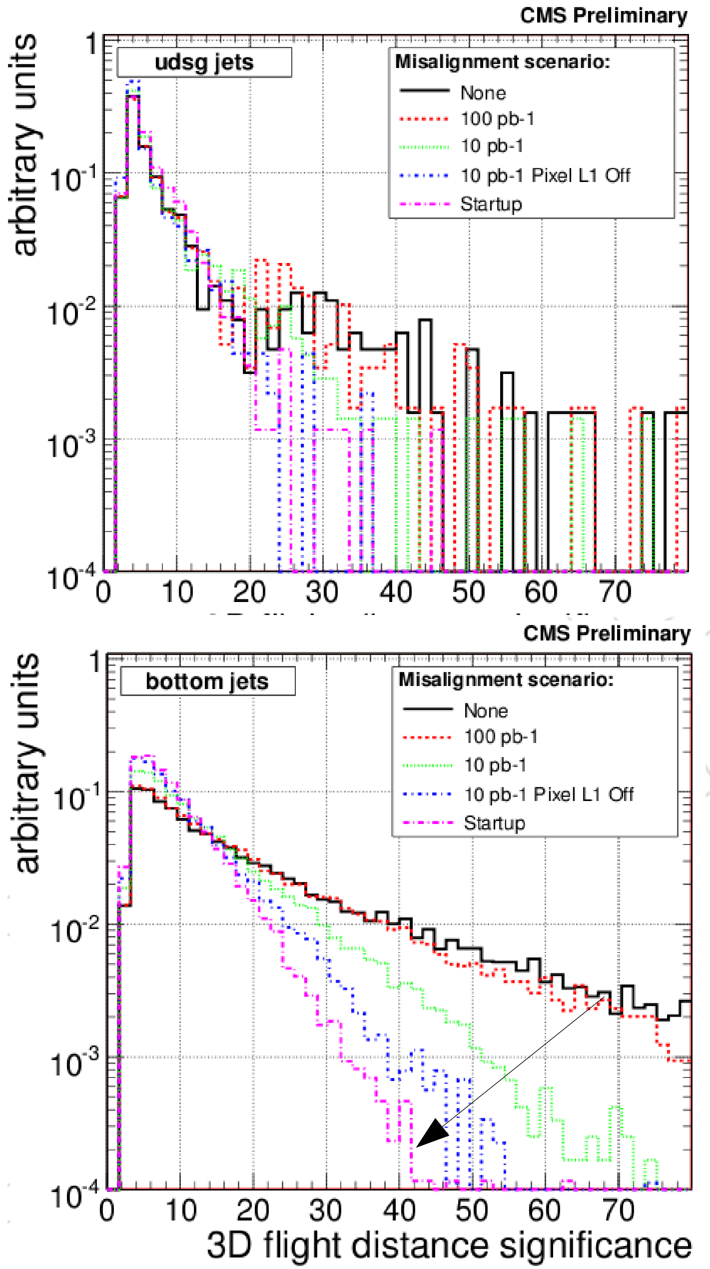
Effect on Impact Parameters



The errors are chosen such that the pull distributions have a width of 1!

With the old “Startup 2008” scenario b-tagging would have been useless!

Tracker Misalignment

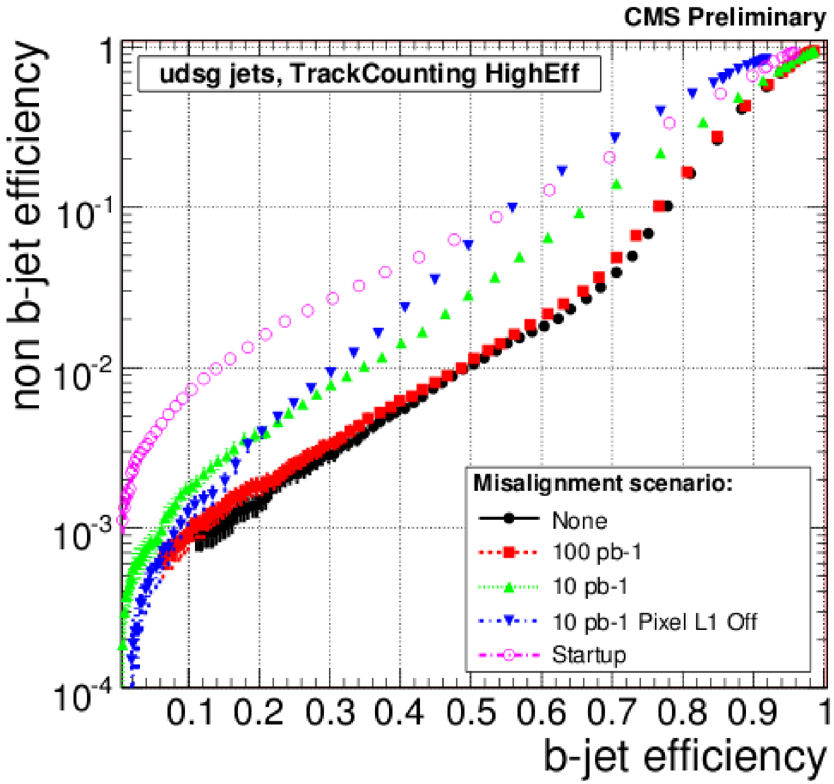


Effect on Secondary Vertex Reconstruction

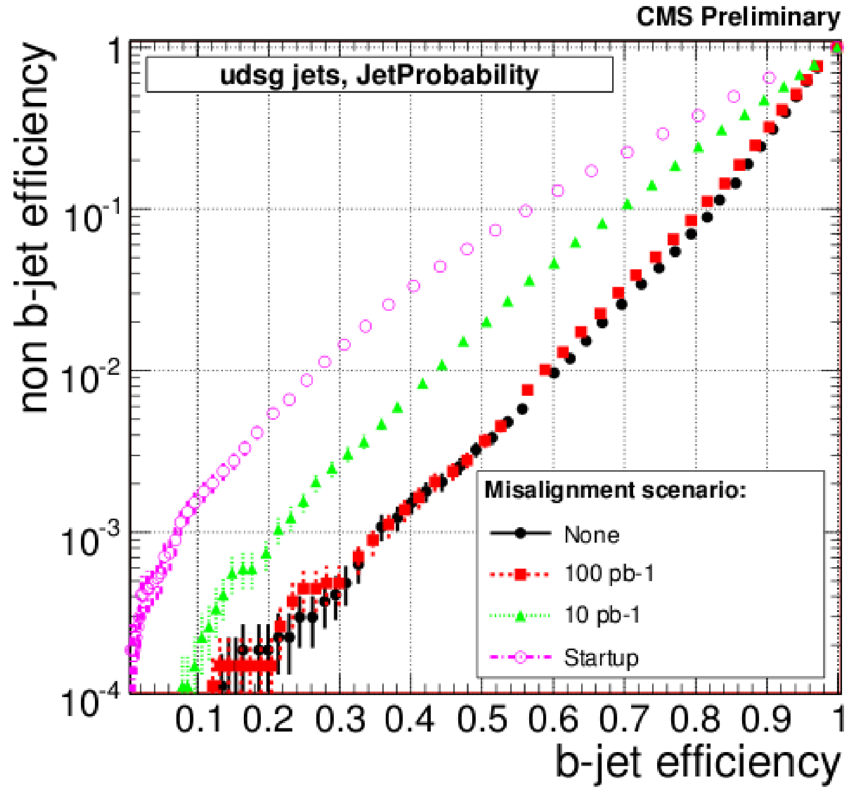
Since essentially, candidate tracks for SV fit are those incompatible with the PV *(related to significance)*, the SV finding efficiency also decreases with the tracker misalignment:

Misalignment scenario	Secondary vertex fraction [%] b-jets	c-jets	udsg-jets
No Misalignment	62.6	22.0	2.7
100 pb^{-1} Misalignment	62.1	19.6	2.4
10 pb^{-1} Misalignment	53.0	12.7	2.9
10 pb^{-1} Pixel L1 Off Misal.	39.2	7.6	1.9
Startup Misalignment	37.8	7.7	3.5

Tracker Misalignment



"Track Counting":
2[rd] -highest IP significance



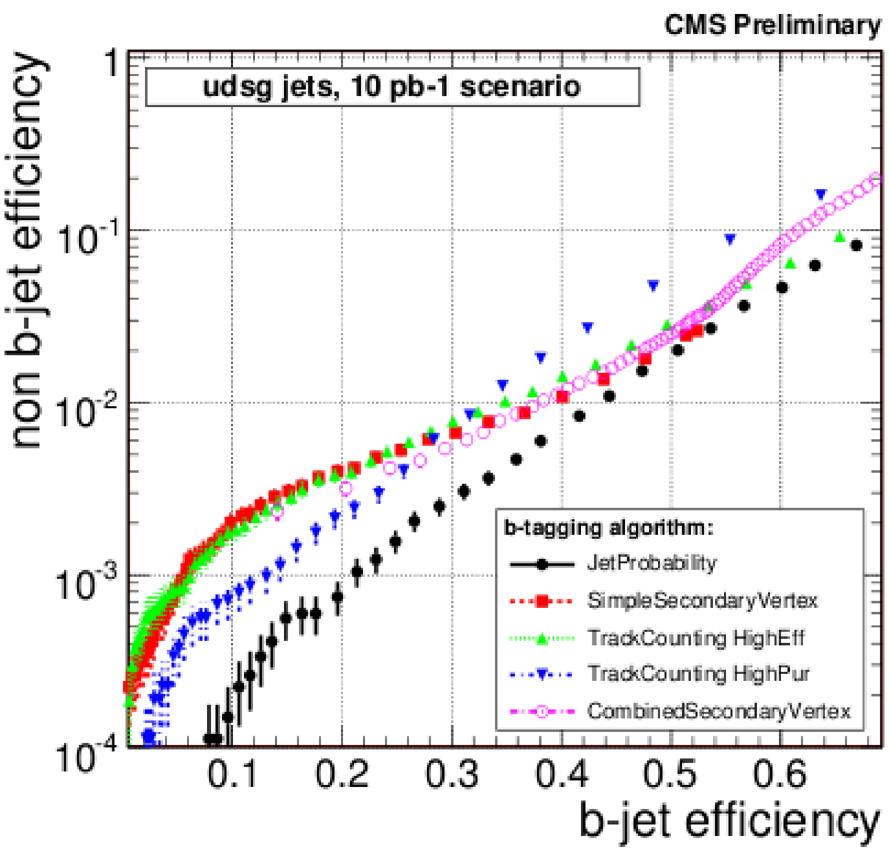
"Jet Probability":
Combination of all "Track Probabilities"
(per-track IP sig. pdf's)

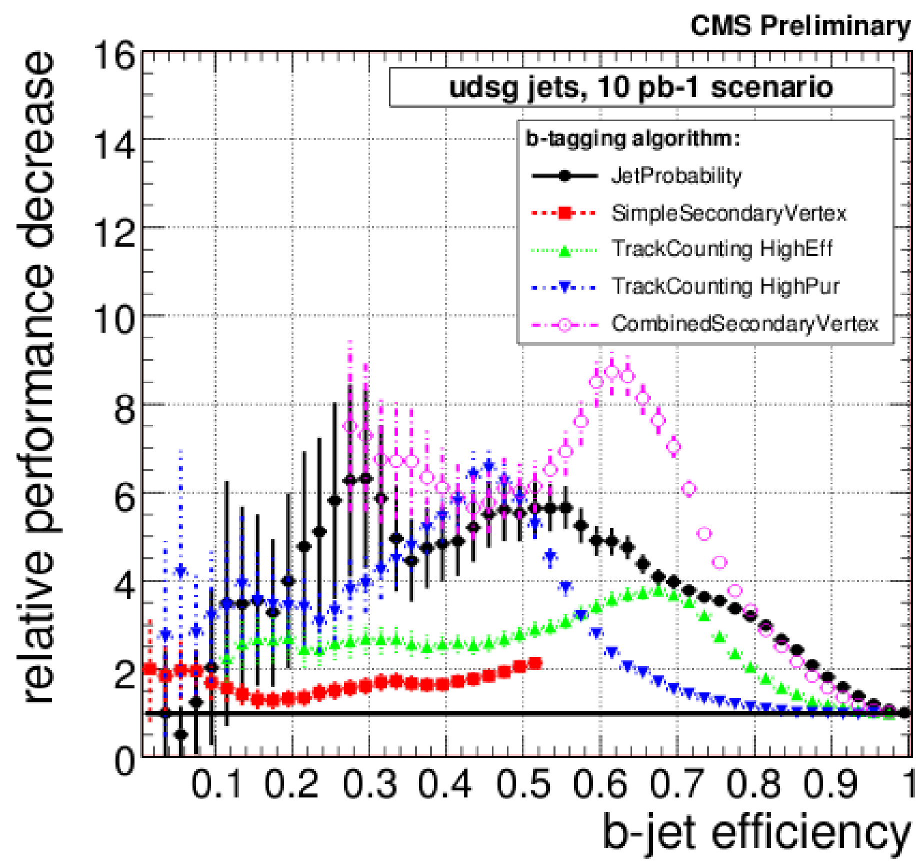
data points represent light flavour mistag vs. b-tagging efficiency
for different working points *(discriminator cuts)*

Tracker Misalignment

General observation:

The more complicated an algorithm (and the more efficient), the more sensitive it is to misalignment





The simple “Track Counting” and “Simple SV” algorithm are presumably easiest To get under control with early data. They do not need any “training” on MC.

→ with $10pb^{-1}$ of data, b-tagging will already be usable

Conclusions

- Many exciting discovery channels, as well as well-known SM channels have b-jets in the final state *(tt→bWbW, H→bb, SUSY, ...)*
- Both CMS and ATLAS have high-resolution state-of the art tracking detectors, with pixels allowing for impact parameter measurements in the r-φ and z plane
- Both CMS and ATLAS have an exhaustive set of b-tagging algorithms in place, both suitable for the startup phase and the long-term discovery phase, as well as for online triggers
- Methods and infrastructure to measure tagging performance and mistag rates on data are in place
- **Track reconstruction is key ingredient**
  → *will be challenging to keep up with alignment progress!*
- Given the current state of alignment from cosmics b-tagging should be usable very early
- *A lot of work and exciting first years ahead of us!*