Top tagging: an analytical perspective

Marco Guzzi

Kennesaw State University

Based on **1807.04767** in collaboration with **M.Dasgupta**, **J.Rawling** and **G.Soyez**

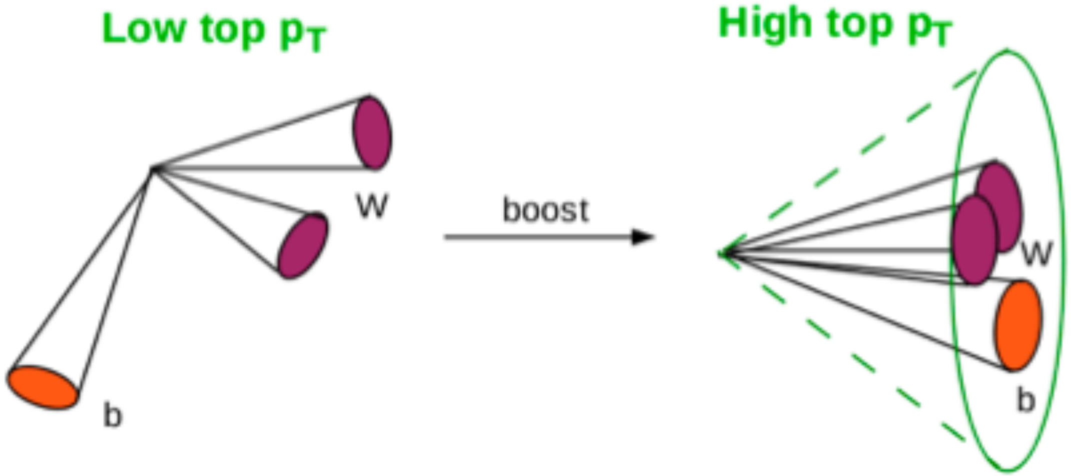
10th International Workshop on Boosted Object Phenomenology,
Reconstruction and Searches in HEP (**BOOST 2018**). 16-20 July 2018. **Paris**.

Kennesaw State University
College of Science and Mathematics
Department of Physics

Boosted objects at LHC run II

At the LHC run II electroweak scale particles, including the **top quark**, can be extremely boosted: jet masses $m \ll p_t$ and their hadronic decays will often result in a single jet.

Jet substructure studies provide key insights to effectively distinguishing signal from background as well as to improve resolution of signal mass peaks



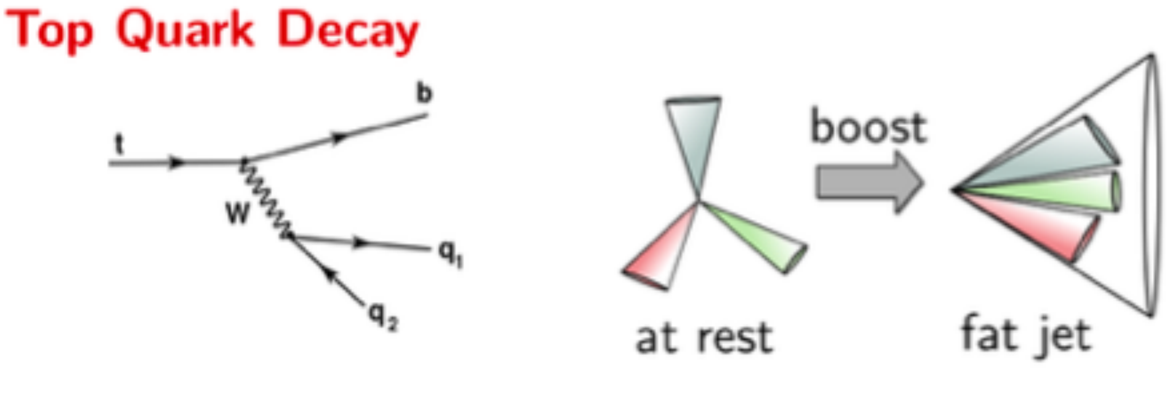


Fig. from S.Schätzel thesis

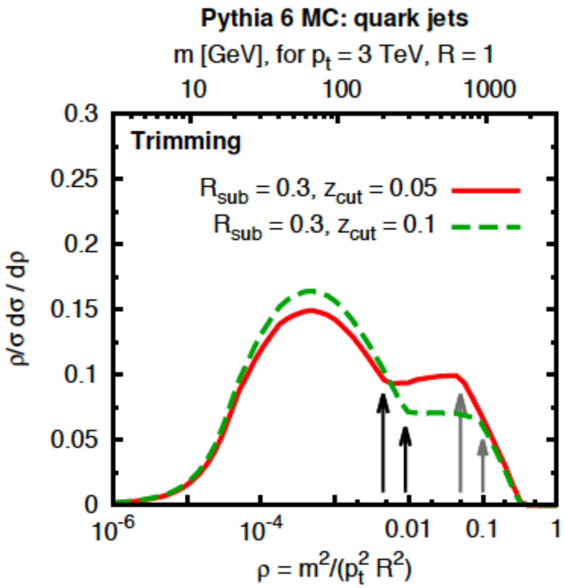
Boosted regime with jet masses $m \ll p_t$: a classic multi-scale problem with large logs in p_t/m:

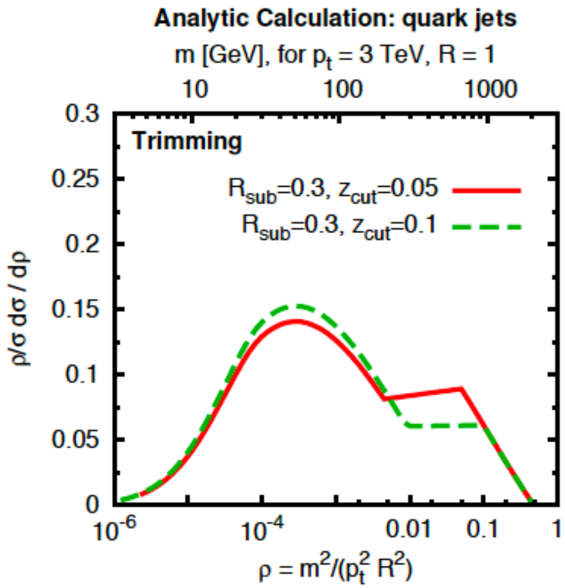
- Perturbative calculations at fixed-order in α_s might not be directly useful on their own.
- We need techniques of analytic resummation to describe substructure observables in the boosted limit

Power of analytical methods

Jet substructure techniques shed light on features that did not emerge in MC studies prior to the advent of the analytics:

- discovery of flaws (kinks and bumps in the jet mass spectrum with various taggers) → led to the emergence of improved tools,
- discovery of occasional issues with parton shower descriptions of jet substructure,
- development of observables which can be computed to high precision in QCD and which display reduced sensitivity to non-perturbative effects (Soft drop) → improved phenomenology at the LHC.





Dasgupta, Fregoso, Marzani, Salam, JHEP 2013;
Larkoski, Marzani, Soyez, Thaler, JHEP 2014

Top tagging using analytical tools

We investigate aspects of top tagging using analytic resummation as a main tool.

Studies along these lines have already been carried out for W/Z/H tagging
(Seymour 1994, Butterworth, Cox, Forshaw 2002, Butterworth, Davison, Rubin, Salam 2008…)

and top tagging techniques are current in use for analyses at the LHC:

Top tagging at the LHC

Top taggers of different kind are widely used by experimental collaborations and in theory/pheno analyses (Difficult to make a complete list here)

- CMS top tagger, based on the JH top tagger (Kaplan, Rehermann, Schwartz, Tweedie, 2008): elements are first clustered via the C-A algorithm, then a candidate jet is systematically declustered, serving two purposes: contaminating “soft” radiation is groomed away, and “hard” subjets are identified.

CMS-PAS-JME-09-001; CMS-PAS-JME-13-007

- Y-splitter ATL-COM-PHYS-2008-001; ATL-PHYS-CONF-2008-008

- HEP top tagger uses a collection of Cambridge/Aachen jets with a distance parameter R = 1.5 (fat jets) (Plehn, Salam, Spannowsky 2009; Plehn, Spannowsky, Takeuchi, Zerwas 2010,….)

- N-subjettiness, based on a jet shape observable which measures how consistent a jet is with having N or fewer subjets. (Thaler and K. Van Tilburg 2011, 2012)

- Shower deconstruction, Soper and Spannowsky 2012,….

- Neural Network, L. de Oliveira, M. Kagan, L. Mackey, B. Nachman and A. Schwartzman (2016), P. Baldi, K. Bauer, C. Eng, P. Sadowski and D. Whiteson (2016); J. Barnard, E. N. Dawe, M. J. Dolan and N. Rajcic (2017); P. T. Komiske, E. M. Metodiev and J. Thaler (2017) ; G. Kasieczka, T. Plehn, M. Russell and T. Schell (2017)

Top tagging using analytical tools

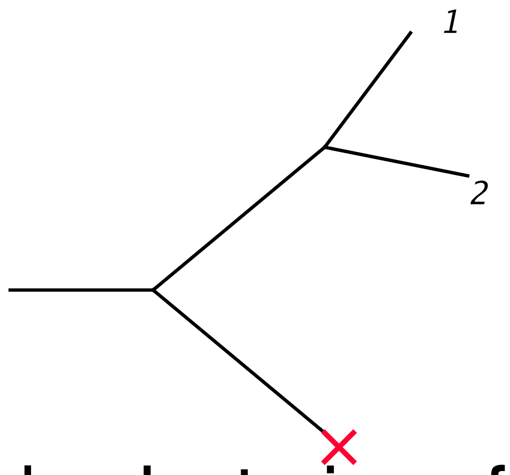
Our aim is to embark on a similar level of understanding for top tagging.
We explore methods for identifying the top quark based on its three-pronged decays (with a focus on the prong finding aspect of top taggers).

One can use:

- CMS Top tagger
- Y-splitter

CMS Top tagger

Primary Decomposition

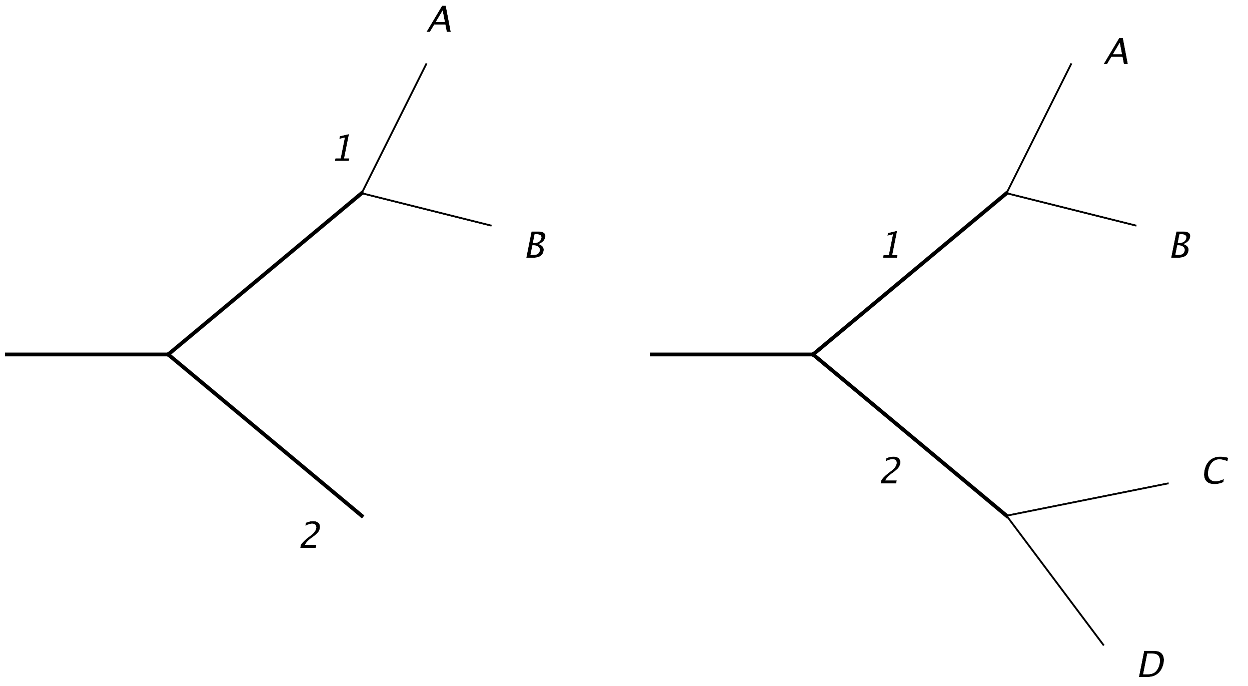


- Perform a C/A de-clustering of the jet and find two prongs.

- Use condition $p_t^{\text{prong}} > \zeta_{\text{cut}}\, p_t$ where p_t is jet rather than local p_t

CMS Top tagger

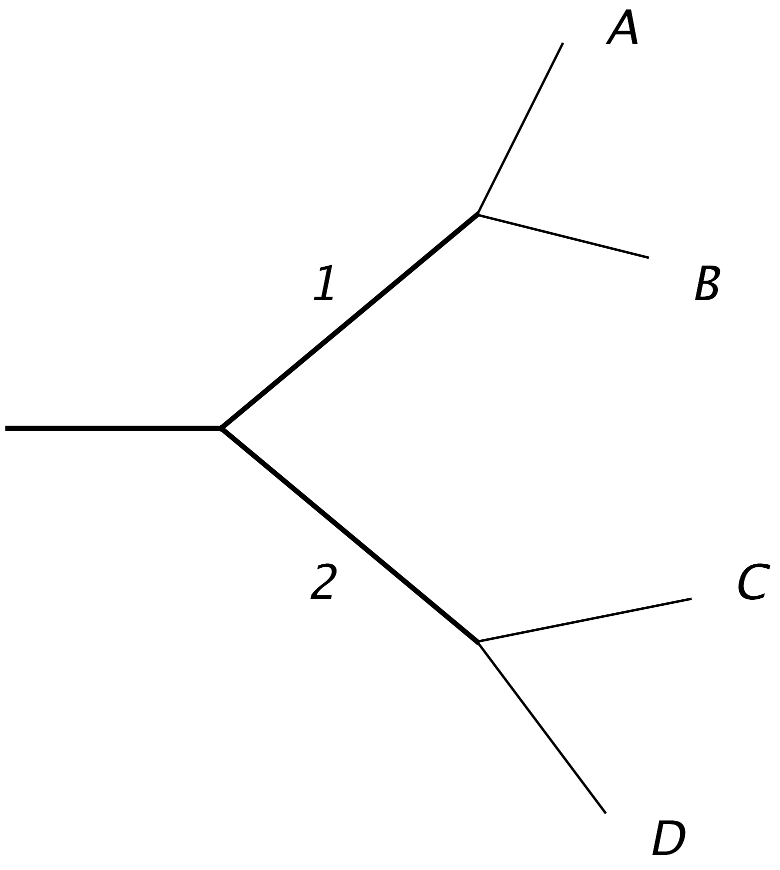
Secondary decomposition



- De-cluster both primary prongs in the same way.
- End up with 2, 3, or 4 prongs.
- Select 3 or 4 prong cases as top candidates.

CMS Top tagger and IRC issues

Selecting 3 prongs from 4



CMS tagger selects three hardest objects say A,B,C.

Imposes an m_{min} condition

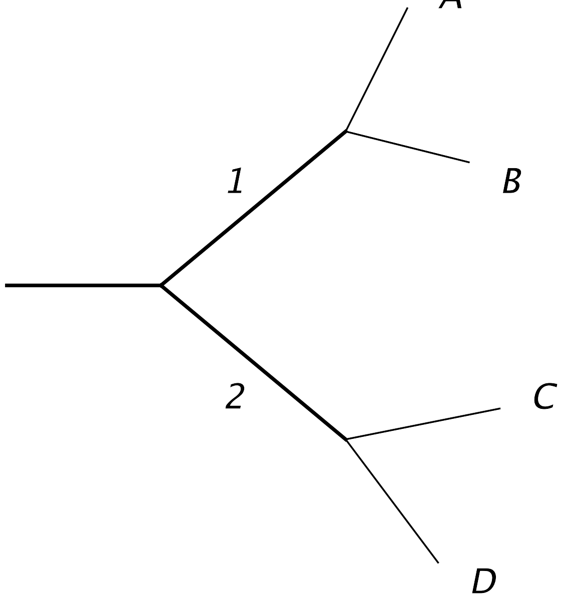
$$\min\left(m_{\text{AB}}, m_{\text{BC}}, m_{\text{CA}}\right) > m_{\text{min}}$$

This method is collinear unsafe!

The collinear unsafety arises due to the process of selecting the three hardest prongs out of four prongs (to define the m_{min} cut) which is sensitive to arbitrarily collinear hard radiation

A later CMS variant introduces a ΔR cut. But IRC unsafety still persists at large pt

CMS^{3p,mass} modified tagger



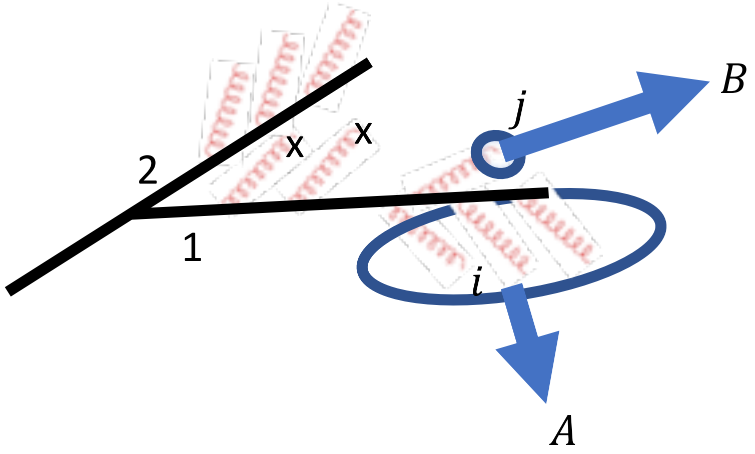
Examine the invariant masses:

$m^2_{AB} =(p_A + p_B)^2$ and $m^2_{CD} =(p_C + p_D)^2$.

If $m^2_{AB} > m^2_{CD}$ one considers the 3 prongs to be A, B and 2, and vice-versa.

We obtain 3 prongs which can be used in the m_{min} condition without any collinear unsafety issues and without a ΔR cut.

`Top-splitter` tagger



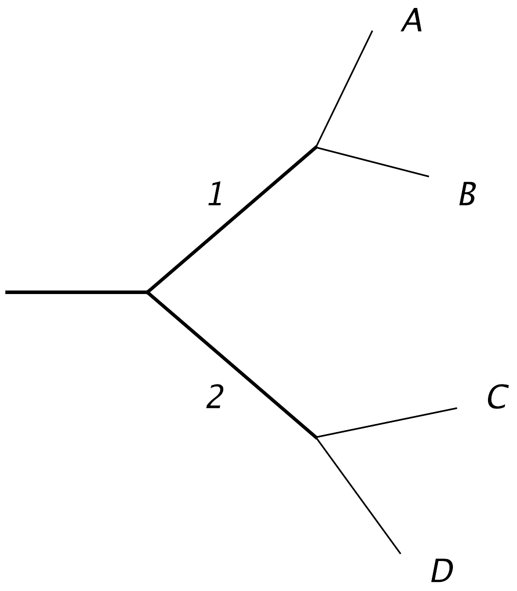
Decluster to find prong 1 and 2 (as in CMS);

- Decompose 1 and 2 until find emission i that passes the ζ_{cut},
- Consider all subsequent emissions further down the C/A tree following the hardest branch, together with emission i, and identify the emission j in this set that has the largest $p_{tj}\,\theta^2{}_j$
- take this emission to be B;
- A is the remaining object to which B is clustered in the C/A sequence, along with all emissions preceding B in the C/A tree which passed the cut condition, such as emission i.

Y_m-Splitter tagger

It uses the gen-kt (p = 1/2) distance for de-clustering:

- guarantees an ordering equivalent to mass ordering in the soft limit,
- facilitates the direct analytical understanding of the tagger behavior,
- slightly better performance compared to the standard Y-splitter

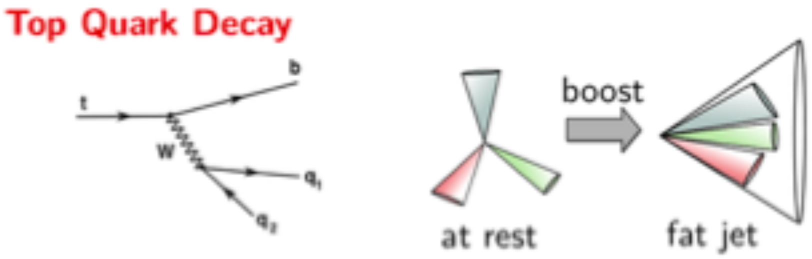


Also needs grooming

MD, Powling, Schunk, Soyez 2016
MD, Powling, Siodmok 2015

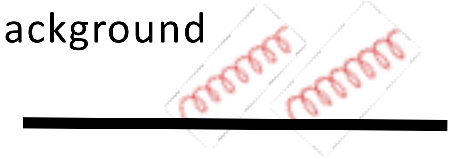
- First de-clustering based on the gen-kt (p = 1/2),
- on each of the two prongs apply the ζ_{cut} condition
- de-cluster both prongs gen-kt (p = 1/2):
  - smaller gen-kt distance prong kept unaltered
  - larger gen-kt distance prong tested for the ζ_{cut}
- Take the three prongs (i.e. the unaltered primary prong and the two secondary prongs which passed the ζ_{cut} condition) and impose the m_{min} condition on the minimum pairwise mass.

Jet mass distribution



No soft enhancement

QCD background



contains soft enhancement

We impose cuts on the soft region to reduce the background

Jet invariant mass distribution in terms of boost invariant quantities along the jet direction

$$\frac{d\sigma}{d\rho} \qquad \rho = \frac{m^2}{p_T^2 R^2} \qquad \rho_{\min} = \frac{m^2_{\min}}{p_T^2 R^2}$$

large logarithms in the jet masses ρ, $\rho_{\min}$ as well as large logarithms in ζ_{cut} , but with the former being numerically dominant over the latter.

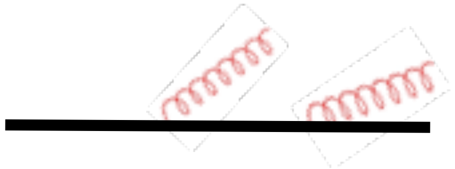
m ~ m_t and $m_{\min}$ ~ m_W at high p_t ➡ ρ, $\rho_{\min}$ << 1 and L_ρ = Log(1/ρ) >>1

No strong ordering in ρ and $\rho_{\min}$ $\quad L_\rho \sim L_{\rho_{\min}} \gg 1 \qquad \zeta_{\mathrm{cut}} \sim 0.05 \qquad L_{\rho,\rho_{\min}} \gg L_\zeta$

Tagger action

At least two emissions in addition to the hard parton that initiates the jet to pass the top-tagger conditions:

Jet mass distribution at LO in pQCD with application of top tagging, is order α_s^2



At order α_s^2 CMS^{3p,mass} and `Top-splitter` are equivalent.
The IRC unsafety issue of the CMS tagger occurs at order α_s^3

Assuming $\rho >> \rho_{\rm min}$ for simplicity

Quark init. jet

$$\frac{1}{\sigma}\left(\frac{d\sigma}{d\rho}\right)^{\rm LO,soft-collinear} = \bar{\alpha}^2 \int \frac{dz_1}{z_1}\frac{dz_2}{z_2}\frac{d\theta_1^2}{\theta_1^2}\frac{d\theta_2^2}{\theta_2^2} \times \Theta(\theta_2^2 < \theta_1^2 < 1)\,\delta(\rho - z_1\theta_1^2)\,\Theta(z_1 > \zeta_{\rm cut})\,\Theta(z_2 > \zeta_{\rm cut})\,\Theta(z_2\theta_2^2 > \rho_{\rm min}) = \frac{\bar{\alpha}^2}{\rho}\left(\boxed{\ln^2\frac{1}{\zeta_{\rm cut}}\ln\frac{\rho}{\rho_{\rm min}}} + \mathcal{O}\left(\ln^3\zeta_{\rm cut}\right)\right)$$

$$\bar{\alpha} = \frac{C_F\alpha_s}{\pi}$$

Compare to QCD jet mass distr.: $\dfrac{\bar{\alpha}^2}{\rho}\ln^3\dfrac{1}{\rho}$

Reduced background after tagging.

two large logarithms in jet mass have been replaced by logarithms of $\zeta_{\rm cut}$

Tagger action: the triple collinear limit

But $\rho >> \rho_{\min}$ is in practice not a good approximation for the case of top tagging.
Without any strong ordering between ρ and $\rho_{\min}$ we are led to a situation where the only genuinely large logarithms in the boosted limit are those in ρ or equivalently $\rho_{\min}$ but not those of $\rho/\rho_{\min}$

$\ln \frac{\rho}{\rho_{\min}},\ \ln \frac{1}{\zeta_{\mathrm{cut}}}$ not too large

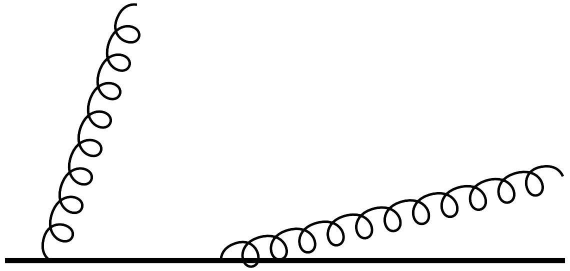
we need to lift strong ordering and soft approx.

$$\frac{\alpha_s^2}{\rho}\ln^2\frac{1}{\zeta_{\mathrm{cut}}}\ln\frac{\rho}{\rho_{\min}} \Rightarrow \frac{\alpha_s^2}{\rho}\times\mathcal{O}(1)$$

$$\frac{1}{\sigma}\left(\frac{d\sigma}{d\rho}\right)^{\mathrm{LO,triple-collinear}} = \frac{\alpha_s^2}{\rho} f_q\left(\rho, \rho_{\min}, \zeta_{\mathrm{cut}}\right)$$

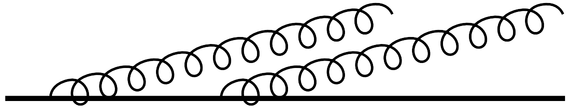
↓

f



Standard LO DGLAP or PS evolution

$\Rightarrow$

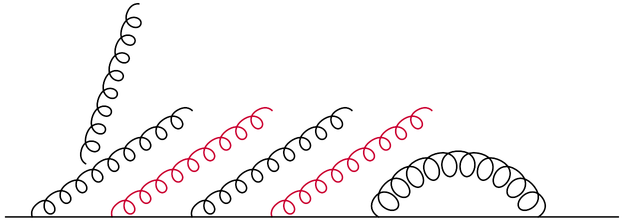


Triple collinear splitting functions

Catani and Grazzini NPB 1999

Calculation to all orders

The large logarithms of ρ or ρ_{min} need to be resummed to all orders in α_s.
For this, we need to add an arbitrary number of real or virtual soft and collinear emissions and consider the constraints on them.



⇒ **Inclusion of a Sudakov Form Factor**

- ✓ Log(ρ/ρ_{min}) and Log$(1/\zeta_{\text{cut}})$ are included in the resummation (non-negligible impact on analytical top-tagging)
- ✓ Log$(1/\rho)$, Log$(1/\rho_{\text{min}})$, Log(ρ/ρ_{min}), and Log$(1/\zeta_{\text{cut}})$ are counted on the same footing
- ✓ Hard collinear emissions described by the 1 → 3 splitting are included

Our resummation accuracy is modified LL, in which we resum all double log $\frac{1}{\rho}\alpha_s^n L^{2n-1}$ terms

We also include NLL effects from running coupling and hard collinear emissions

Calculation to all orders

$$\left(\frac{\rho}{\sigma}\frac{d\sigma}{d\rho}\right)^{\text{LO,triple}-\text{collinear}} = \left(\frac{\alpha_s}{2\pi}\right)^2 \int d\Phi_3 \, \frac{\langle \hat{P} \rangle}{s_{123}^2} \, \Theta^{\text{jet}} \, \Theta^{\text{tagger}}(\zeta_{\text{cut}}, \rho_{\text{min}}) \, \rho \, \delta\left(\rho - \frac{s_{123}}{R^2 p_t^2}\right)$$

is extended to

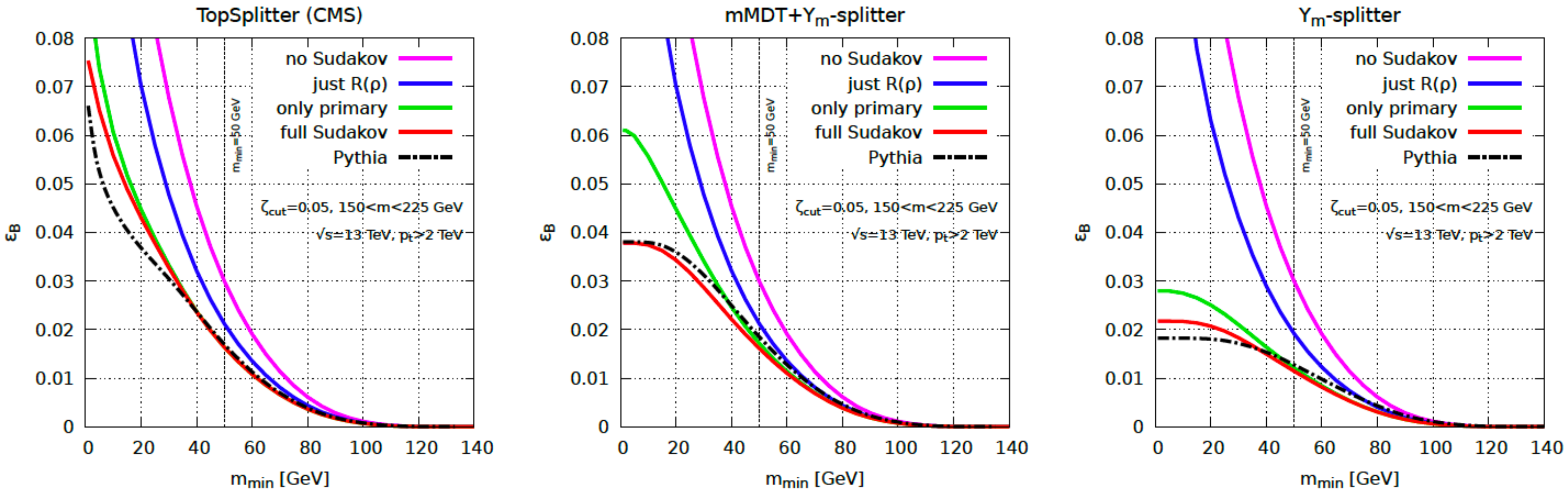
prefactor

Sudakov

$$\frac{\rho}{\sigma}\frac{d\sigma}{d\rho} = \int d\Phi_3 \, \frac{\langle \hat{P} \rangle}{s_{123}^2} \, \frac{\alpha_s(k_{t1})}{2\pi} \, \frac{\alpha_s(k_{t2})}{2\pi} \Theta^{\text{jet}} \, \Theta^{\text{tagger}} \, \delta\left(\rho - \frac{s_{123}}{R^2 p_t^2}\right) \, \mathcal{S}_{\text{tagger}}(\rho_2; \rho_1, \theta_1)$$

- Prefactor computed using triple-collinear splitting functions and phase space.
- Convoluted with a Sudakov form factor accounting for all double log terms.
- Running coupling and hard-collinear effects included.
- Matching of Sudakov to triple-collinear phase space.
- Aims to be as accurate as triple-collinear result at LO and reproduce all leading-log terms beyond.

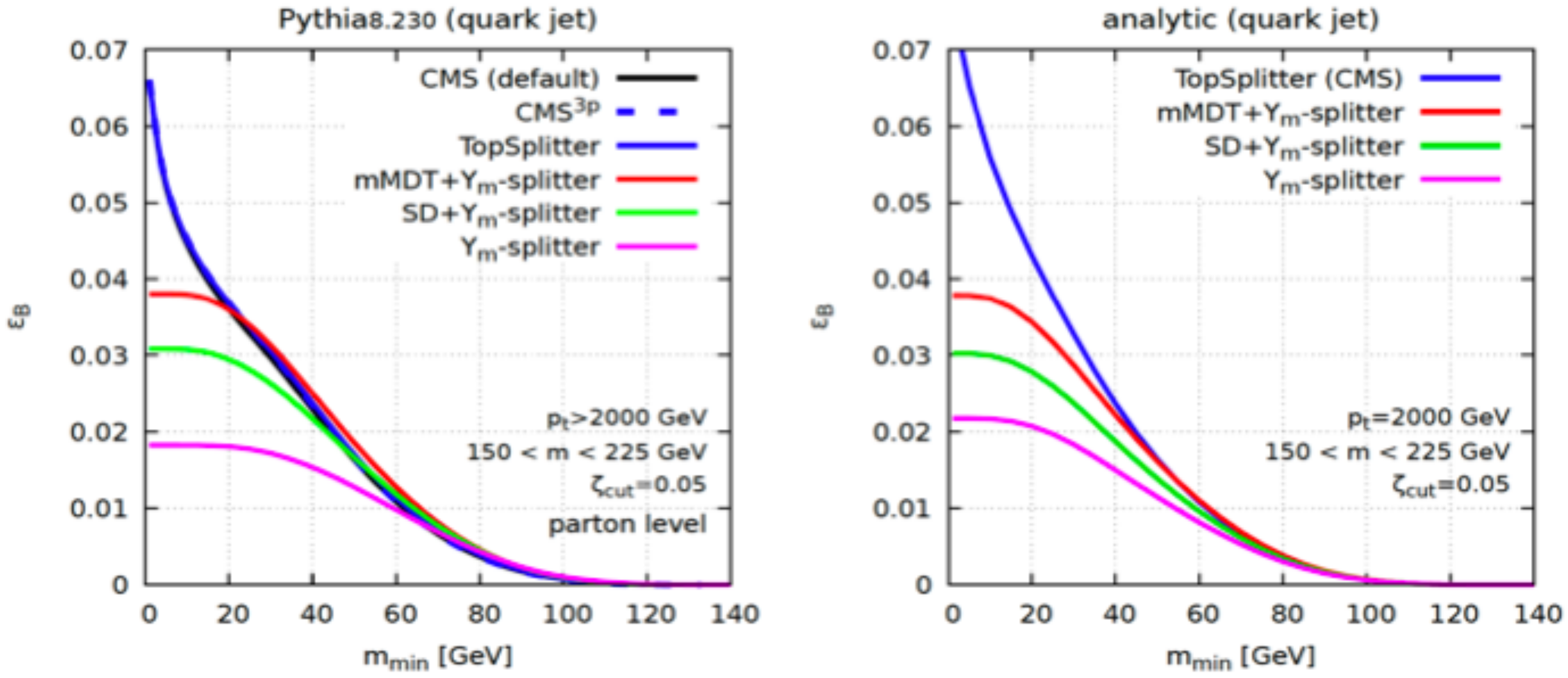
Results



Comparison between analytic results and Pythia simulations for the QCD background efficiency

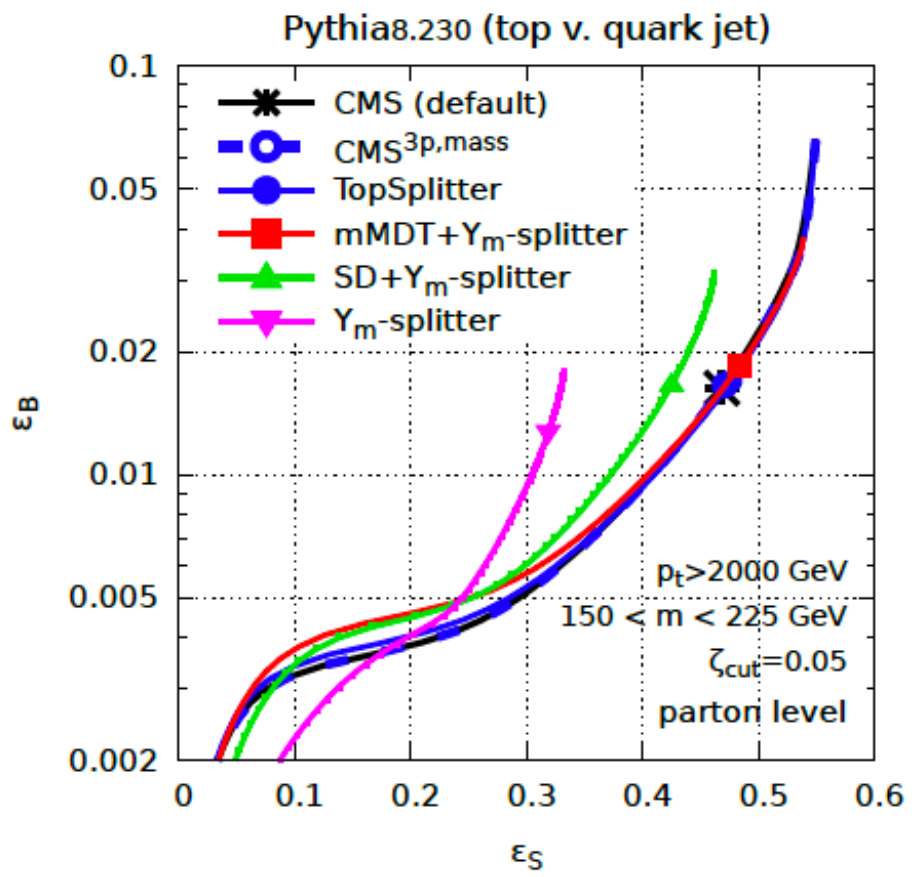
- Resummation of Log(ρ/ρ_{min}) terms does matter
- Inclusion of secondary emissions important at small m_{min}
- Overall a good agreement with PS.

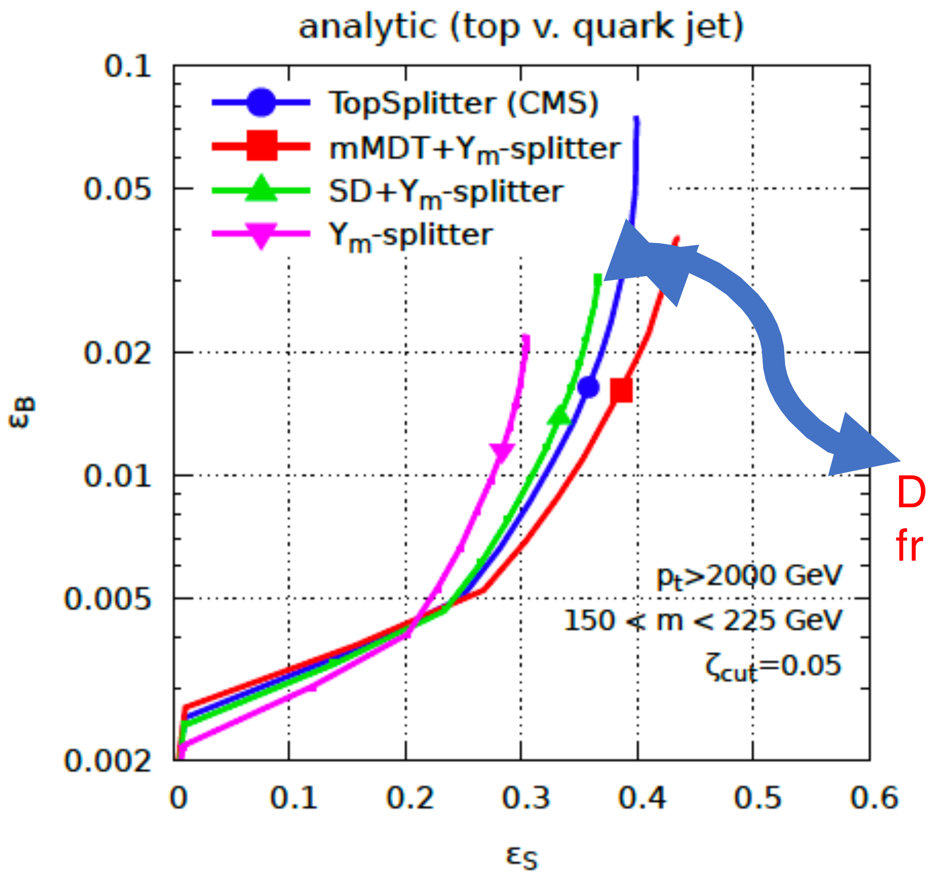
Results: quark Jets



- MC and analytics agree on comparative performance
- Y_m-splitter best at suppressing QCD jets
- CMS and variants are basically identical for performance
- Groomed Y_m-splitter comparable with CMS. Differences largely due to secondary emissions.

Results: Signal vs Background



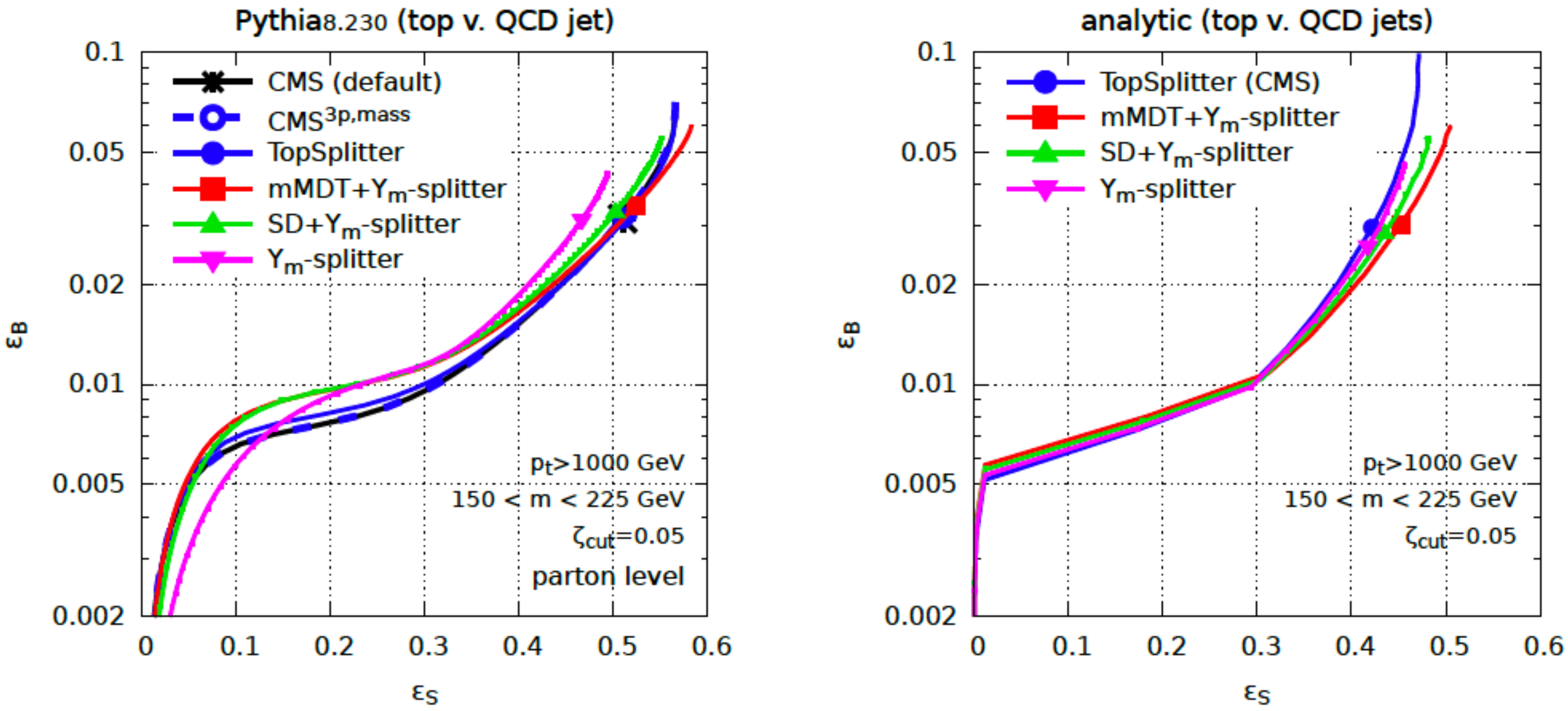


Differences come from the signal

Large Sudakov suppression is not necessarily beneficial for the case of top tagging in contrast to the tagging of color singlet electroweak and Higgs bosons.

the performance difference is driven (mostly) by the worse description of the signal:
we ultimately aim at a better understanding of the signal

Results: Signal vs Background



Sudakov effects are smaller and the differences between the taggers becomes less important in both MC and the analytics for pt ~1 TeV

Conclusions

A first analytic study of aspects of top tagging from first principles of QCD using the methods of analytic resummation have been carried out:

- Motivated by IRC safety issues we proposed variants of the CMS top tagger that are explicitly IRC safe even at large p_t (CMS^{3p,mass} and Top-Splitter)
- analytic vs MC: we appear to have very good analytic control over top taggers we developed when applied to QCD jets
- Signal jets: large Sudakov effects not necessarily desirable and hurt signal efficiency: we ultimately aim at a better understanding of the signal
- CMS tagger becomes potentially unsafe at high p_t. Potentially harmful for precision studies. Easy to design safe variants with no change in performance.
- It will be interesting to investigate combinations with jet shape variables as next step.

BACK UP

CMS Top tagger and IRC issues: angular cut

A second version of the CMS tagger described in

CMS-PAS-JME-13-007

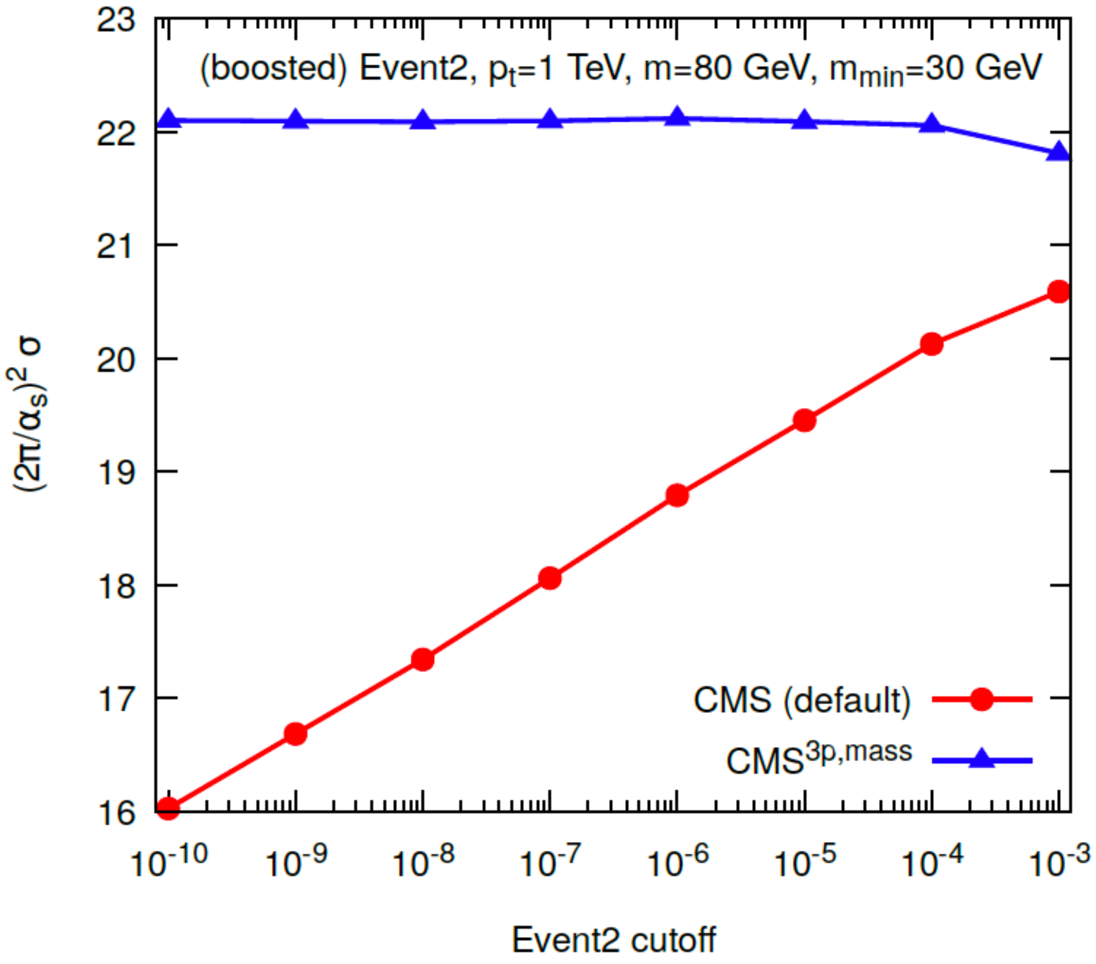
imposes an angular cut in addition to $\zeta_{\rm cut}$: when examining the decomposition into two prongs i and j

$$\Delta R_{ij} > 0.4 - A p_t^S \qquad \Delta R_{ij} = \sqrt{\Delta y_{ij}^2 + \Delta\phi_{ij}^2}$$

and pt^S refers to the transverse momentum of the subjet. **The default value for A is 0.0004 GeV^{-1}**.

As one progresses towards high pt values the R cut becomes smaller and eventually vanishes which means that the default CMS tagger will again be collinear unsafe at 1TeV and for larger pt.

CMS top-tagger: IRC safety issues



Cross-section for passing the CMS tagger as a function of the Event2 cut-off.

For the default CMS top tagger, we see an obvious logarithmic dependence on the cut-off as a result of the collinear unsafety of the tagger. Switching instead to the $\mathrm{CMS}^{3p,\mathrm{mass}}$ tagger, the cross-section converges rapidly when the cut-o is decreased, showing that the collinear unsafety has been cured.

Calculation to all orders

$$\left(\frac{\rho}{\sigma}\frac{d\sigma}{d\rho}\right)^{\text{LO,triple}-\text{collinear}} = \left(\frac{\alpha_s}{2\pi}\right)^2 \int d\Phi_3 \, \frac{\langle \hat{P} \rangle}{s_{123}^2} \, \Theta^{\text{jet}} \, \Theta^{\text{tagger}}(\zeta_{\text{cut}}, \rho_{\text{min}}) \, \rho \, \delta\left(\rho - \frac{s_{123}}{R^2 p_t^2}\right)$$

is extended to

Sudakov

$$\frac{\rho}{\sigma}\frac{d\sigma}{d\rho} = \int d\Phi_3 \, \frac{\langle \hat{P} \rangle}{s_{123}^2} \, \frac{\alpha_s(k_{t1})}{2\pi} \, \frac{\alpha_s(k_{t2})}{2\pi} \, \Theta^{\text{jet}} \, \Theta^{\text{tagger}} \, \delta\left(\rho - \frac{s_{123}}{R^2 p_t^2}\right) \, \mathcal{S}_{\text{tagger}}(\rho_2; \rho_1, \theta_1)$$

Where:

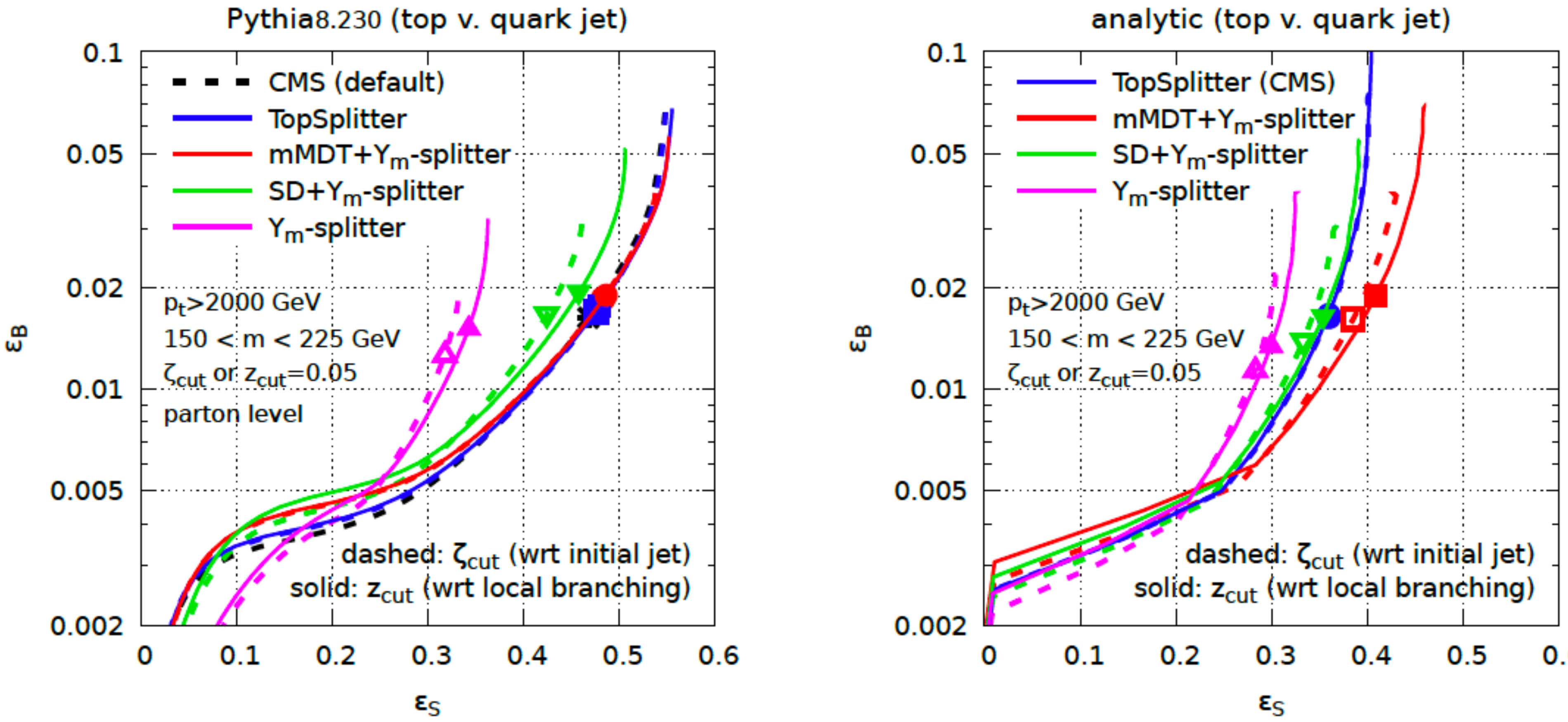
- Prefactor computed using triple-collinear splitting functions and phase space.
- Convoluted with a Sudakov form factor accounting for all double log terms.
- Running coupling and hard-collinear effects included.
- Matching of Sudakov to triple-collinear phase space.
- Aims to be as accurate as triple-collinear result at LO and reproduce all leading-log terms beyond.

$$\Theta^{\text{jet}} = \sum_{i>j\neq k} \Theta\left(d_{ij}^{(\text{anti-}k_t)} < \min(d_{ik}^{(\text{anti-}k_t)}, d_{jk}^{(\text{anti-}k_t)})\right) \Theta(\theta_{ij} < R) \Theta(\theta_{(i+j)k} < R)$$

$$\Theta^{\text{tagger}}(\zeta_{\text{cut}}, \rho_{\text{min}}) = \sum_{i>j\neq k} \Theta\left(d_{ij}^{(\text{tagger})} < \min(d_{ik}^{(\text{tagger})}, d_{jk}^{(\text{tagger})})\right) \Theta\left(\min(z_k, 1-z_k) > \zeta_{\text{cut}}\right) \times$$

$$\times \, \Theta\left(\min(z_i, z_j) > \zeta_{\text{cut}}\right) \Theta\left(\min(\rho_{ij}, \rho_{jk}, \rho_{ki}) > \rho_{\text{min}}\right)$$

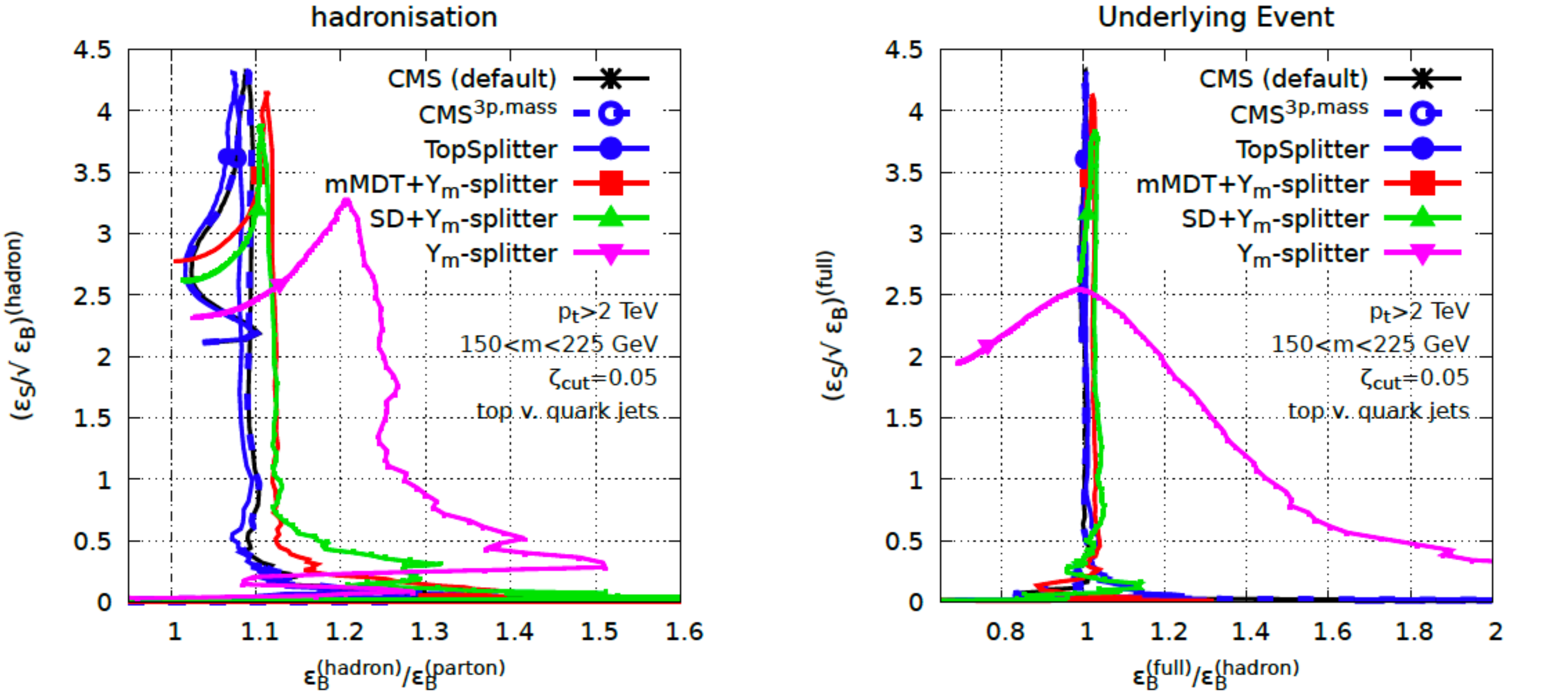
Taggers performance



Comparison of the taggers performance when using a ζ_{cut} condition (solid lines) compared to the default cut condition (dashed lines). **Left**: Pythia simulations, **Right**: our analytic calculation.

little differences between the two variants, in particular, for the CMS-related taggers

Nonperturbative effects



Both plots show how the sensitivity to non-perturbative effect (x axis) and the discriminating power (y axis) evolve when varying the cut on m_{min} for different taggers. **Left**: effects observed when switching on hadronisation, i.e. going from parton level to hadron level. **Right**: effects observed when including the Underlying Event. The symbols correspond to m_{min} = 50 GeV.