Understanding Jet Substructure and Top Taggers with QCD

Mrinal Dasgupta

Cambridge, 5th March 2020

Based on work with Gregory Soyez, Marco Guzzi, Jacob Rawling JHEP 1809 2018

The University of Manchester

European Research Council
Established by the European Commission

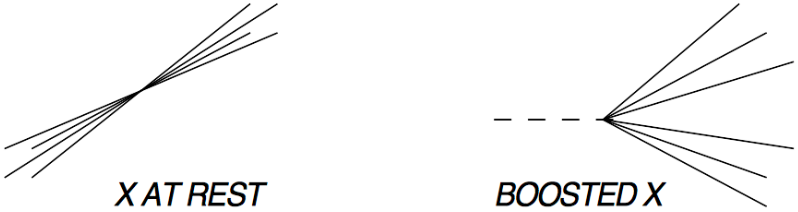
MANCHESTER 1824

The University of Manchester

Outline

- Jet substructure for boosted objects
- Tagging and grooming
- Standard vs analytic approach
- Analytic approach to top tagging
- Results for QCD and signal jets
- Challenges and prospects

Jet substructure and boosted objects



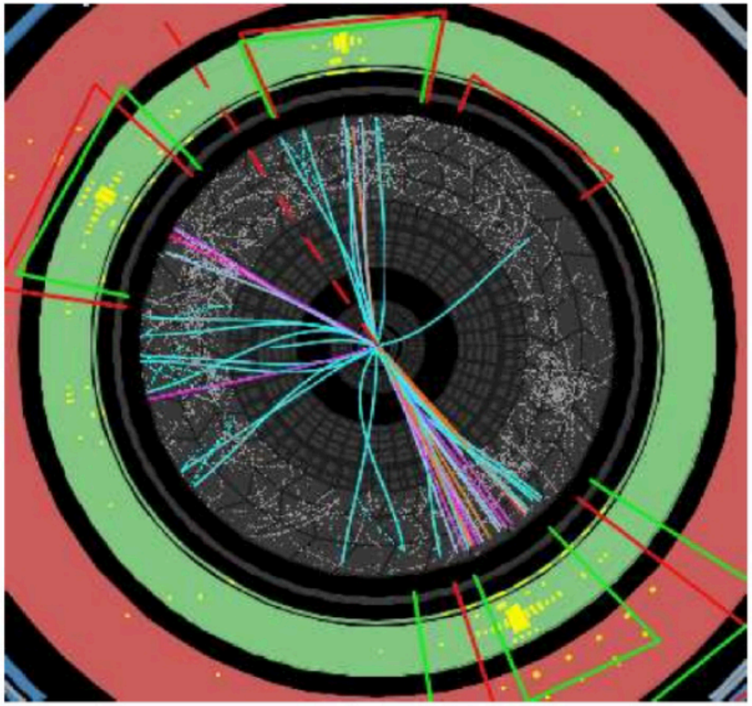
Key idea : for tagging a particle with mass M exploit boosted regime i.e. $P_T >> M$.

$$\theta^2 = \frac{M^2}{p_T^2 z(1-z)}$$

Hadronic decays reconstructed in single “fat” jet.
Use our knowledge of QCD jets to distinguish this from background.

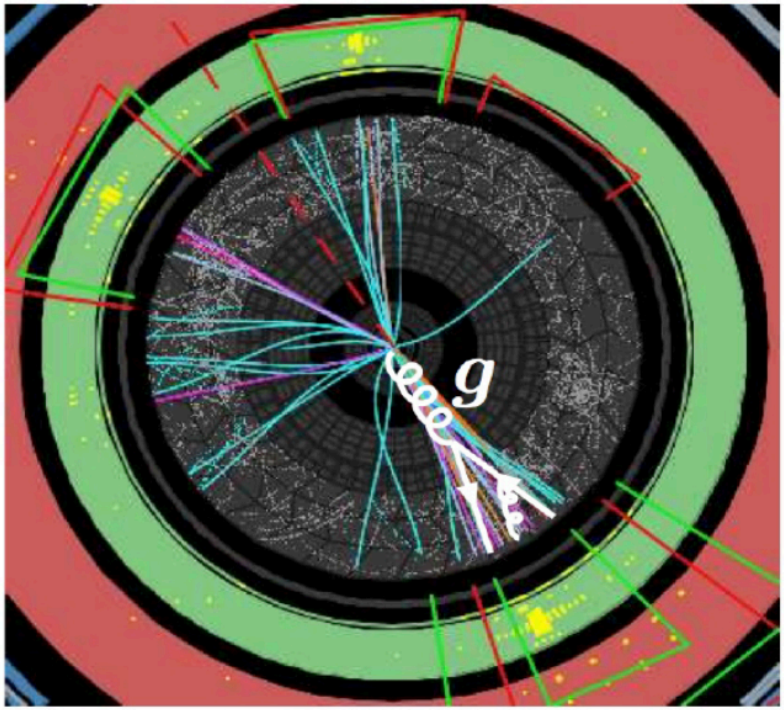
Jets from QCD vs boosted heavy particles

What jet do we have here?



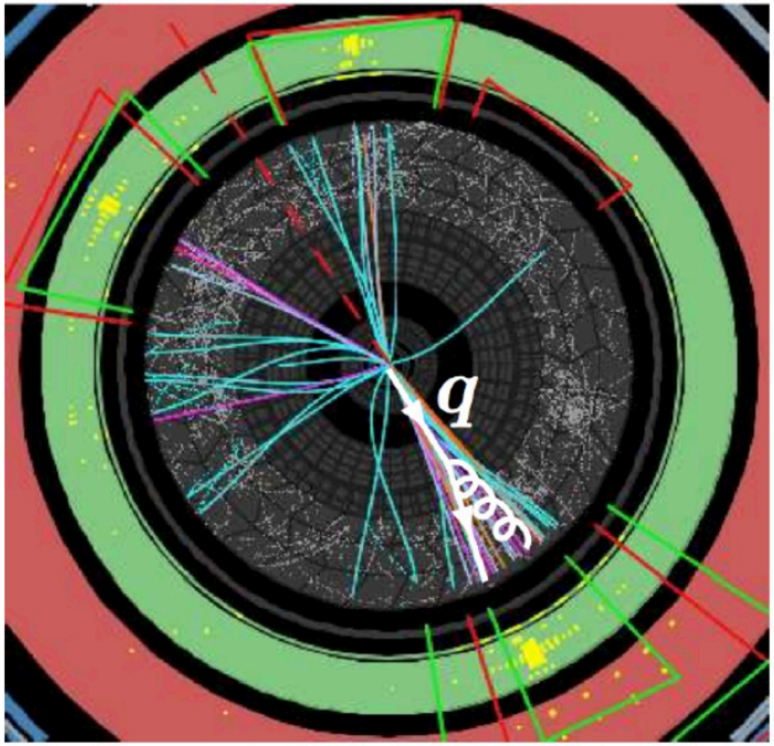
Jets from QCD vs boosted heavy particles

A gluon jet ?



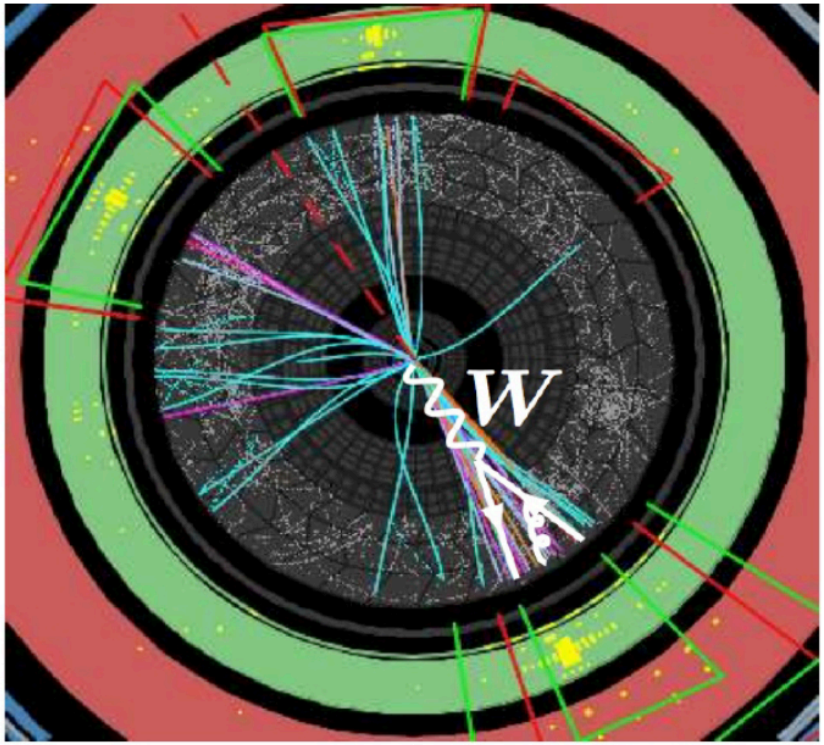
Jets from QCD vs boosted heavy particles

A quark jet ?



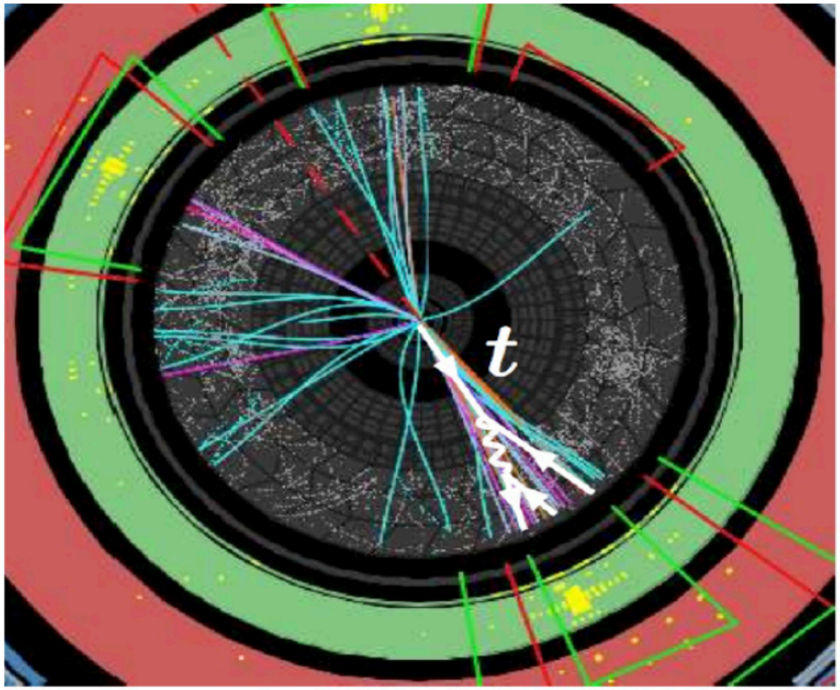
Jets from QCD vs boosted heavy particles

A W/Z/H ?



Jets from QCD vs boosted heavy particles

A top quark?

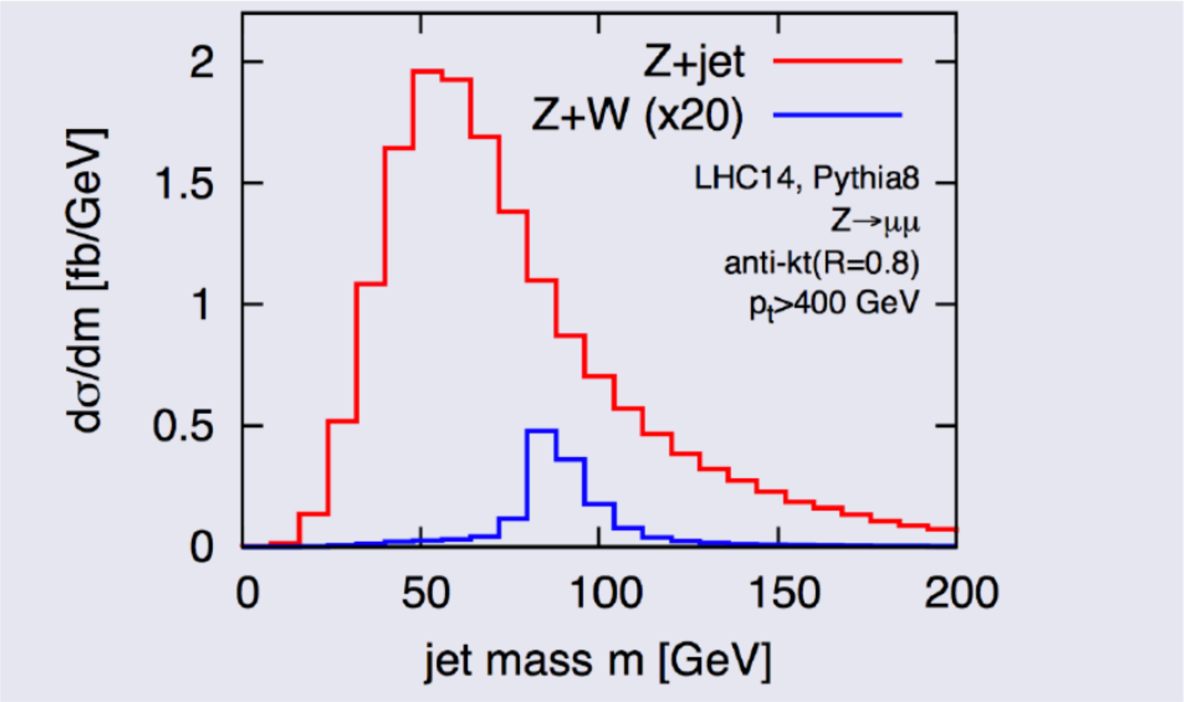


Source: An ATLAS boosted top candidate

The boosted regime implies a change in paradigm in that jets can be more than quarks and gluons.

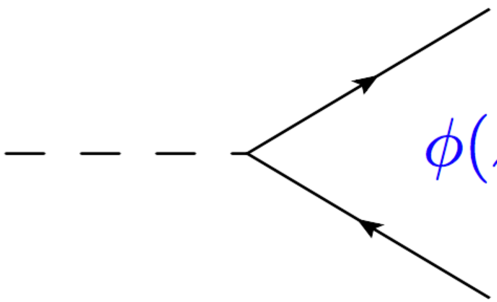
But how can we tell?

Is the jet mass a clue?

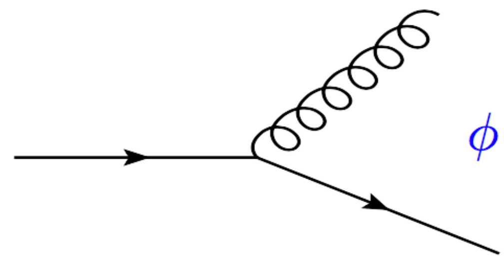


Looking at jet mass is not enough!

Substructure as a clue



$$\phi(z) \propto 1$$



$$\phi(z) \propto \frac{1+z^2}{1-z}$$

- Exploit the asymmetric nature of QCD splittings. Produce jets with single hard core or prong versus 2 pronged W/Z/H and 3 pronged t.

- Colour singlet nature of W/Z/H suppressing soft large angle radiation.

Taggers

There are two main ideas:

Idea 1:
Find $N = 2, 3, \ldots$ hard cores

Works because different splitting

QCD jets: $P(z) \propto 1/z$

$\Rightarrow$ dominated by soft emissions

$\Rightarrow$ "single" hard core

Idea 2:
Constrain radiation patterns

Works because different colours

Radiation pattern is different for

- colourless $W \to q\bar{q}$
- coloured $g \to q\bar{q}$

Taggers try and exploit the above differences.
But we also need jet grooming.

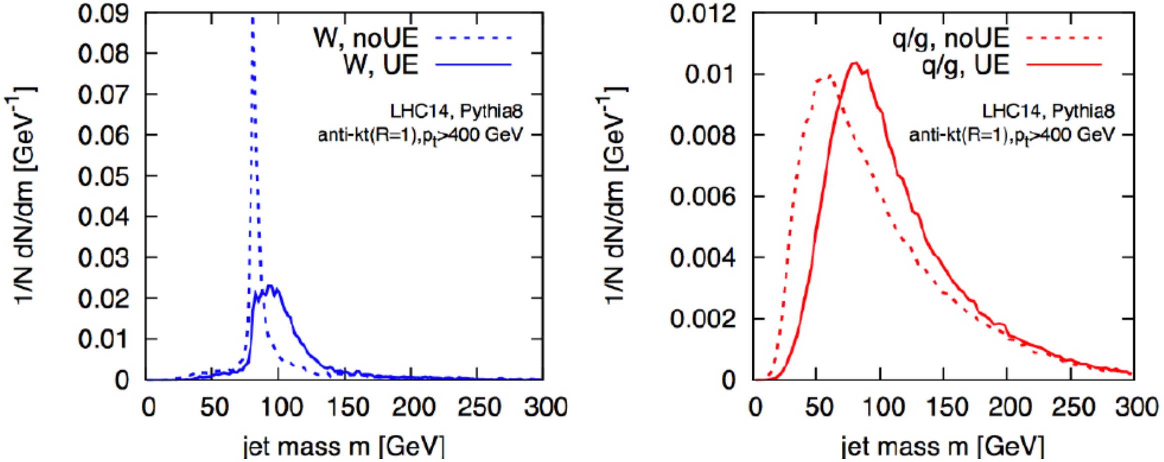
The University of Manchester

MANCHESTER 1824

Jet grooming

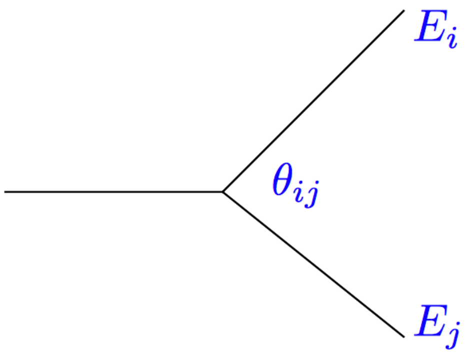
Fat Jets

One usually work with large-R jets ($R \sim 0.8 - 1.5$)
$\Rightarrow$ large sensitivity to UE (and pileup)



Grooming : removal of soft uncorrelated radiation.

An early method : Y-splitter



Butterworth, Cox, Forshaw, 2002

Decluster a jet into 2 subjets using the kt distance measure

Ask for a cut forcing prongs to be more “symmetric” i.e. a Y configuration

$$\frac{k_{t,ij}^2}{m_j^2} \approx \frac{\min(E_i, E_j)}{\max(E_i, E_j)} > y_{\mathrm{cut}} \quad \text{OR} \quad \frac{\min(E_i, E_j)}{E_i + E_j} > z_{\mathrm{cut}}$$

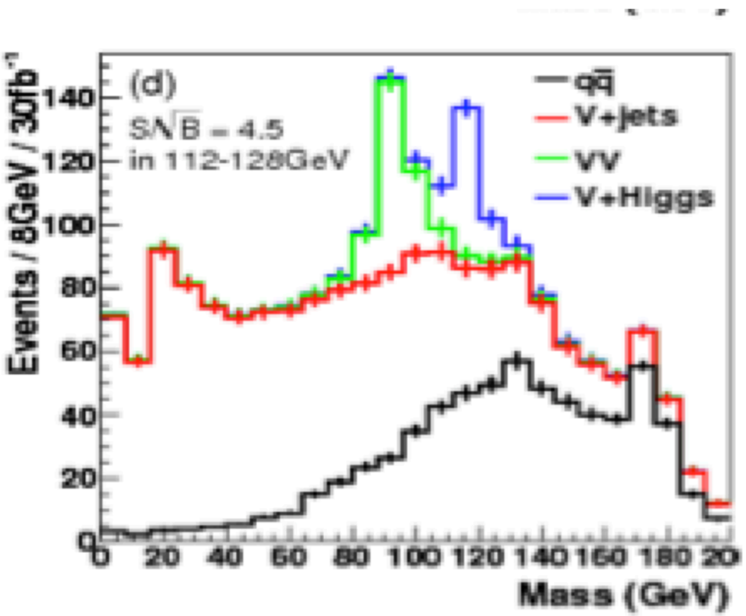
Tag jet if passes cut or discard

Note : no grooming involved

The University of Manchester

MANCHESTER 1824

Mass-drop (MDT) and filtering

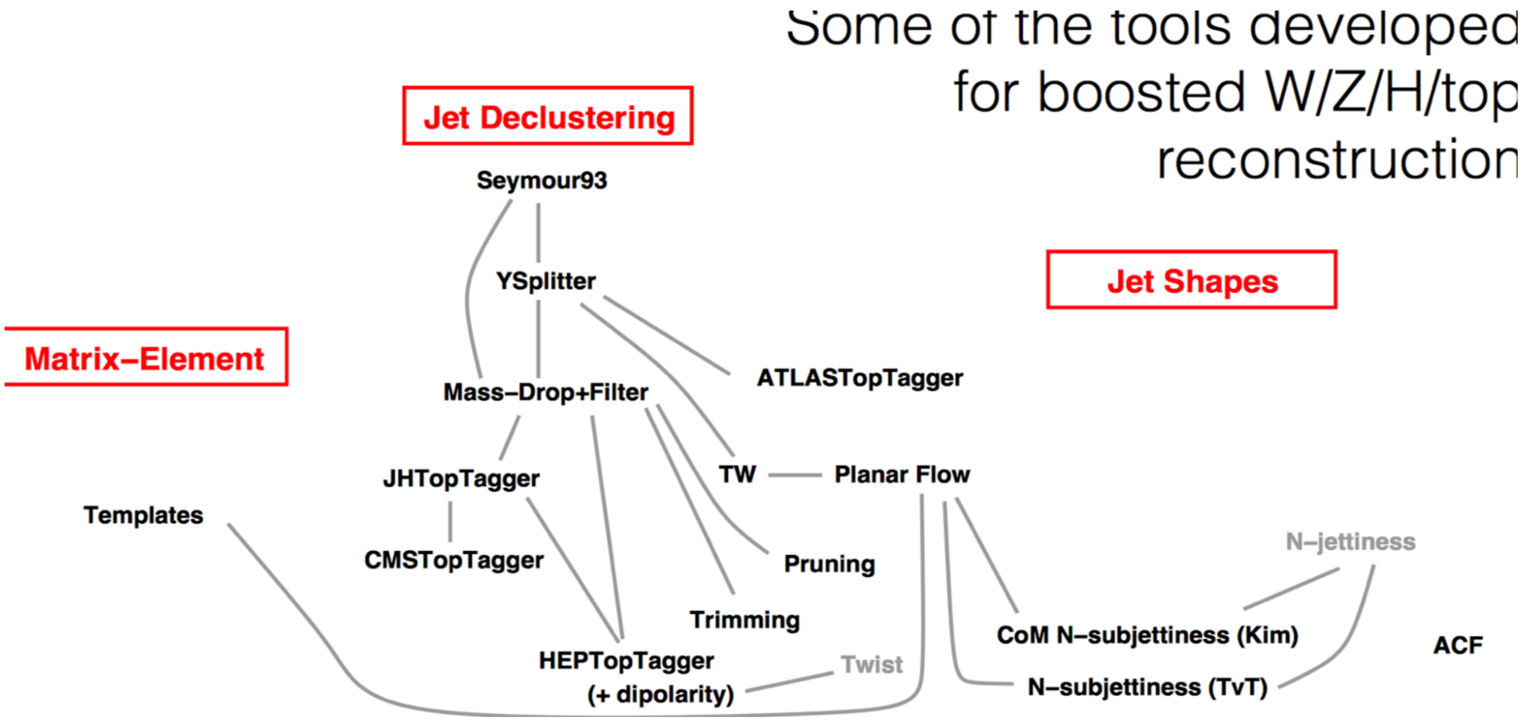


Butterworth Davison Rubin and Salam 2008

- Main motivation for current extensive applications came from BDRS Higgs studies for $pp \to VH,\ H \to b\bar{b}$
- Signal significance of $4.5\,\sigma$ in previously hopeless channel.

Grooming and tagging built into mass-drop. Filtering is pure groomer. Relevant only at moderate p_t

The tagging goldrush

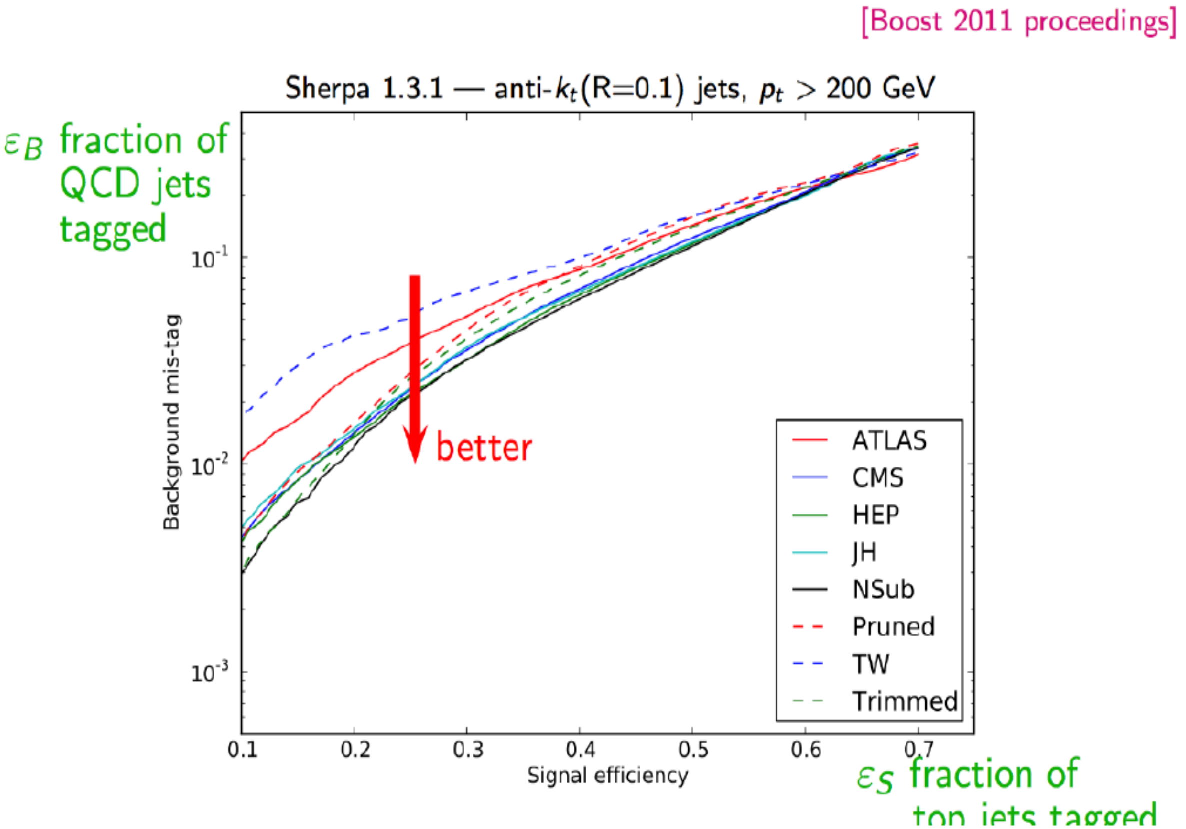


Lots of tools in short time. But still only a couple of principles. Opens up several questions.

MANCHESTER 1824

The University of Manchester

Performance

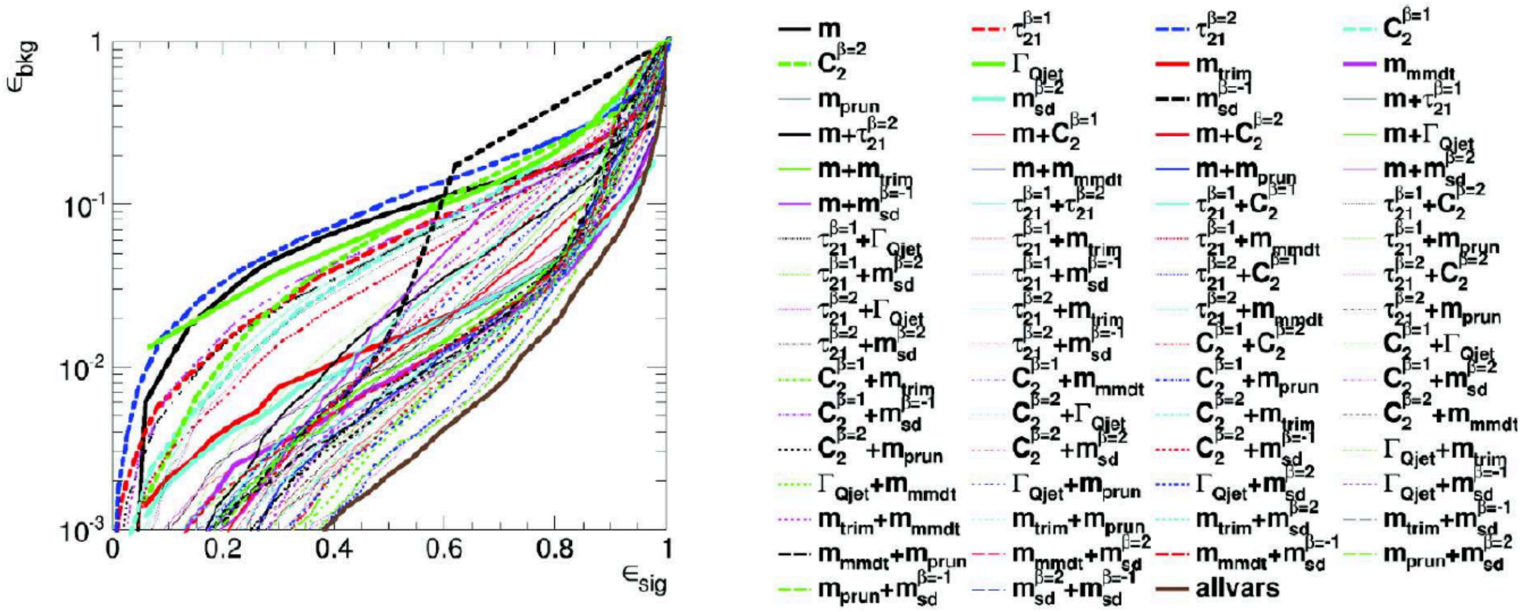


MC studies carried out for fixed parameter settings. Don't give a feel for dependence on parameters and interplay with p_T.

Combinations

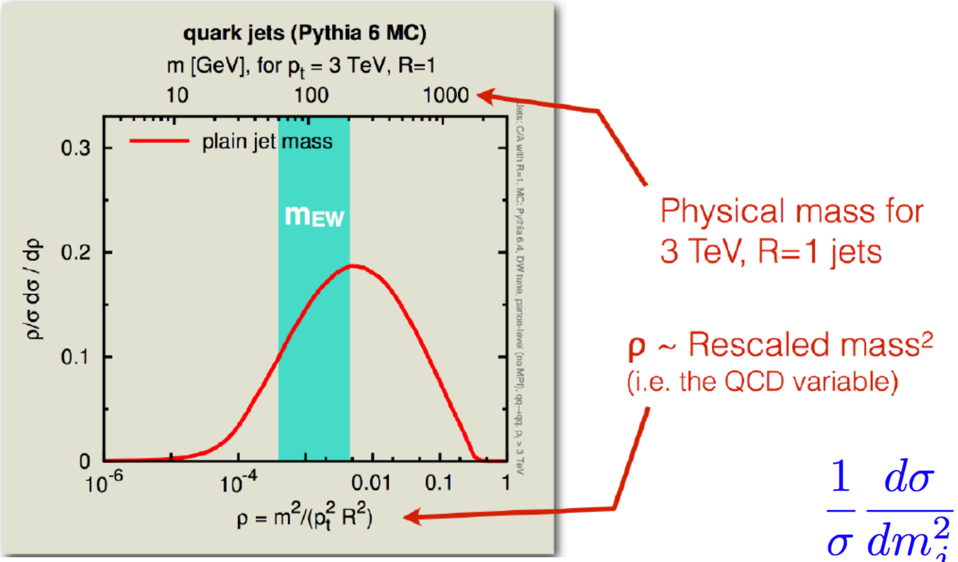
[Boost 2013 WG]

W v. *q* jets: combination of "2-core finder" + "radiation constraint"



Combinations help but details far from obvious.

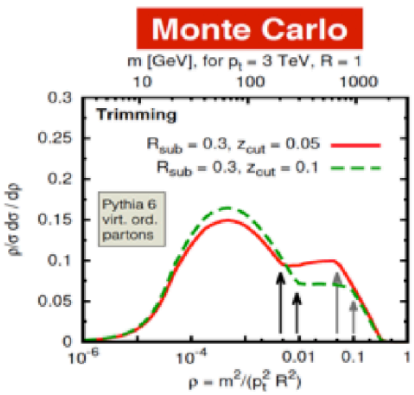
Analytical approach

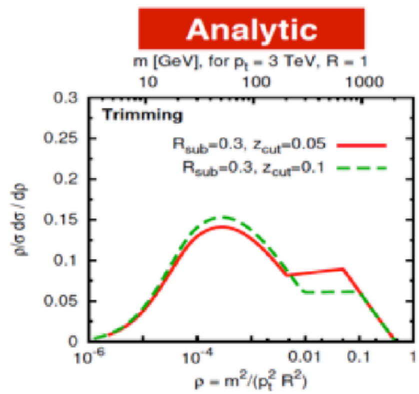


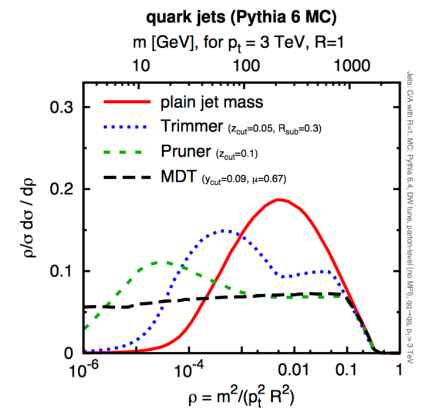
$$\frac{1}{\sigma}\frac{d\sigma}{dm_j^2} \sim \frac{1}{m_j^2}\frac{C_i\alpha_s}{\pi}\ln\left(\frac{R^2 p_t^2}{m_j^2}\right)$$

- Based on trying to understand taggers using pQCD
- We have a multiscale problem with $p_t >> m_J$
- The key tool here is resummation.

Analytical understanding



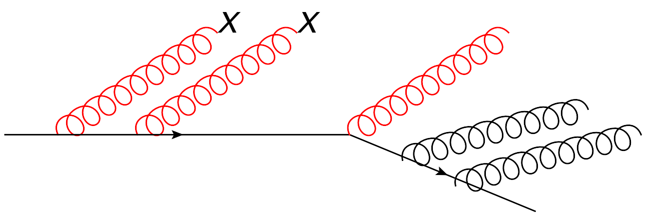


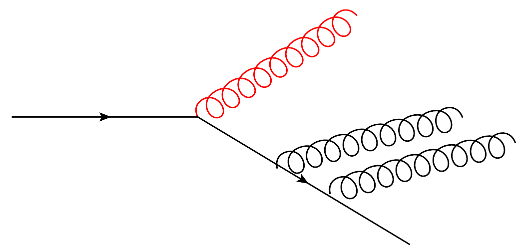


- Established analytical understanding of substructure and tools for W/Z/H decays.
- Revealed factors influencing performance.
- Revealed undesirable tagger features. Taggers can be worse than doing nothing to jet.
- Led to new tools with better properties.

New tools from analytics

Two distinct types:





mMDT uses CA declustering. Recurses through jet until finds splitting with $\frac{\min(E_i, E_j)}{E_i + E_j} > \zeta_{\rm cut}$

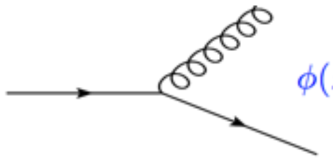
Descendent of MDT
The same as SoftDrop for $\beta = 0$

Ym-splitter uses gen-kt (p=1/2) i.e. mass declustering and examines 1st emission only.

Descended from Y-splitter. Add grooming to improve performance

mMDT versus Ym-splitter

$$\left(\rho\frac{d\sigma}{d\rho}\right)^{\mathrm{LO}} = \frac{C_F\alpha_s}{\pi}\ln\frac{1}{\zeta_{\mathrm{cut}}} \qquad \rho = \frac{m^2}{p_T^2}$$



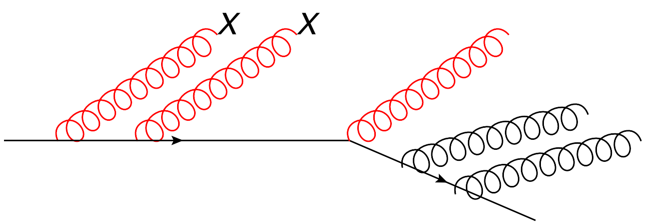
$$\phi(z) \propto \frac{1+z^2}{1-z}$$

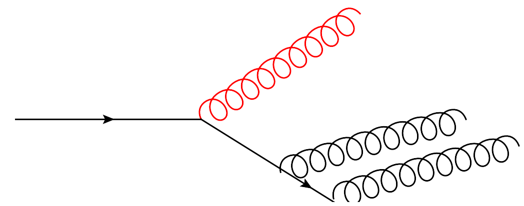
$$z, 1-z > \zeta_{\mathrm{cut}}$$

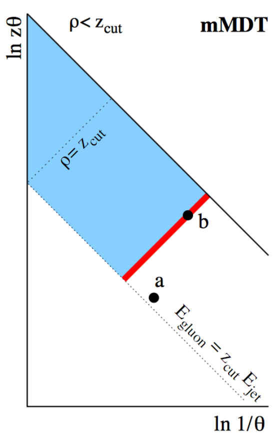
Methods coincide at leading order.
Reduce background by eliminating large log in m/p_t

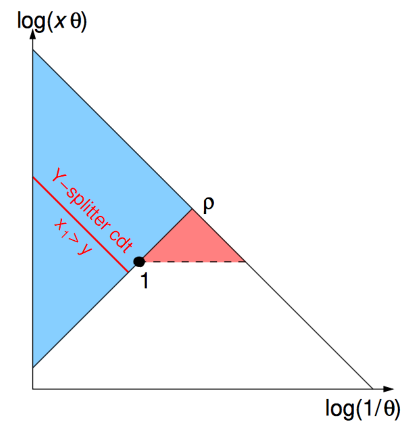
Beyond LO constrain emissions differently.

mMDT/SoftDrop vs Ym-splitter







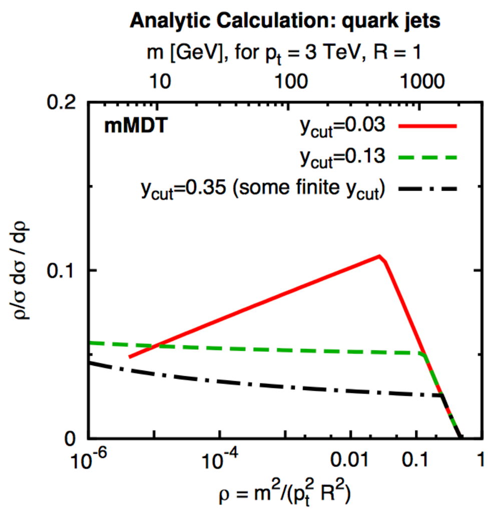


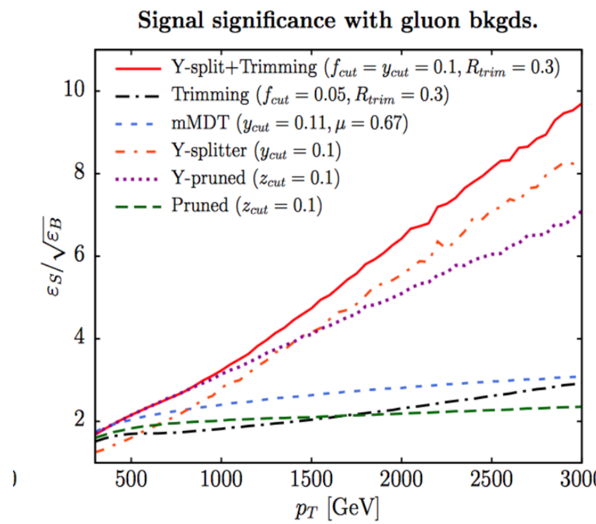
Large Sudakov suppression

$$\rho \frac{d\sigma}{d\rho}^{\rm mMDT} = \frac{C_F \alpha_s}{\pi} \ln \frac{1}{\zeta_{\rm cut}} \exp\left[-\frac{C_F \alpha_s}{\pi} \ln \frac{1}{\zeta_{\rm cut}} \ln \frac{1}{\rho}\right]$$

$$\rho \frac{d\sigma}{d\rho}^{\rm Ym-splitter} = \frac{C_F \alpha_s}{\pi} \ln \frac{1}{\zeta_{\rm cut}} \exp\left[-\frac{C_F \alpha_s}{2\pi} \ln^2 \frac{1}{\rho}\right]$$

Performance





MD, Powling Siodmok 2016

- Ym-splitter needs to be supplemented by grooming to improve signal efficiency.
- Gives important performance gains relative to other methods due to Sudakov for W/Z/H.
- mMDT less performant but more robust. Can give flat background and has much lower NP effects (10% compared to 40%).

Extension to Top Tagging

MANCHESTER 1824

The University of Manchester

Analytics for top taggers

- Want to identify the main relevant physics effects. Start with the CMS tagger and Y-splitter (used in early ATLAS top tagger).

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

ATL-COM-PHYS-2008-001

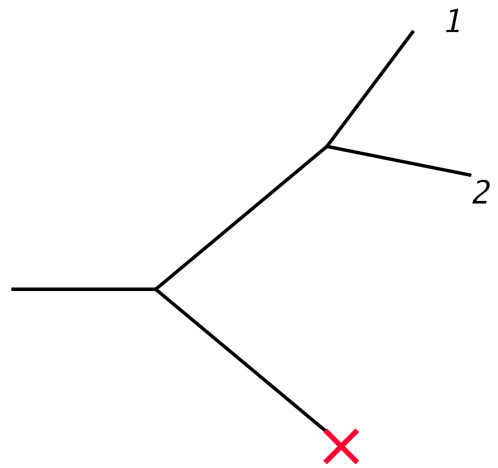
- CMS tagger descends from JH top tagger.

Kaplan, Rehermann, Schwartz and Tweedie 2008

- Both CMS tagger and Y-splitter offer ways of identifying three prongs relevant to top decays.

CMS top tagger

Primary Decomposition



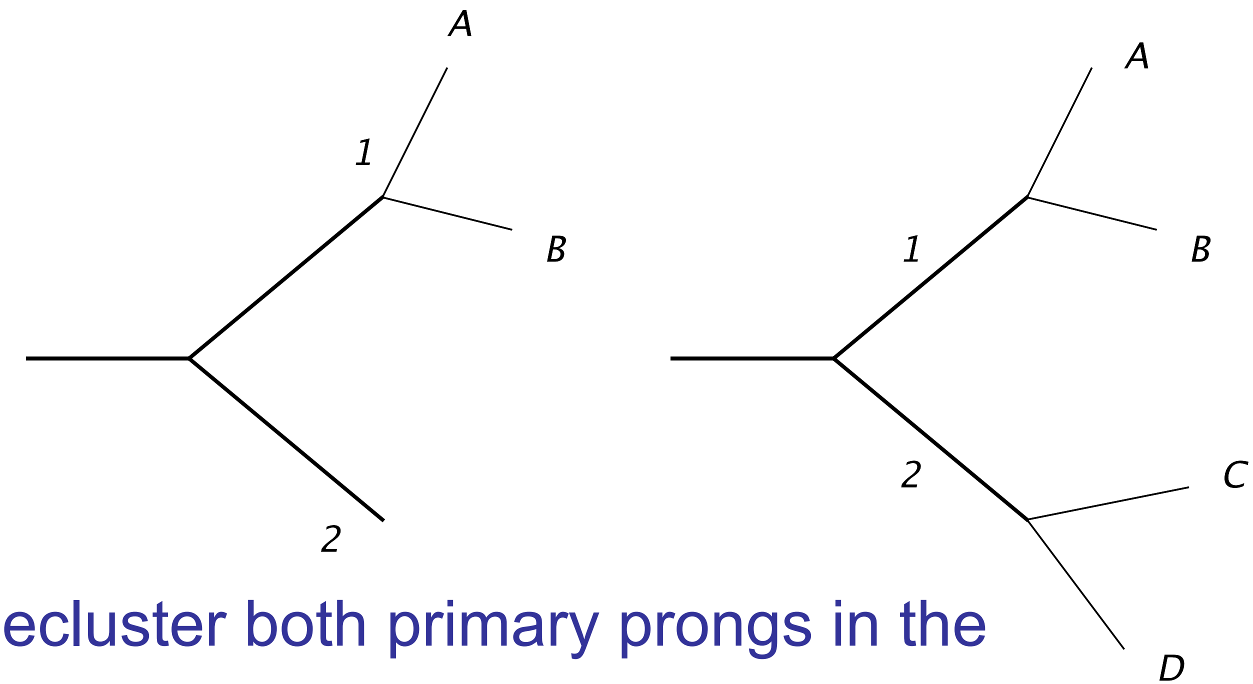
- Perform a C/A de-clustering of the jet and find two prongs.
- Use condition $p_t^{\text{prong}} > \zeta_{\text{cut}}\, p_t$ where p_t is jet rather than local p_t

The University of Manchester

MANCHESTER 1824

CMS top tagger

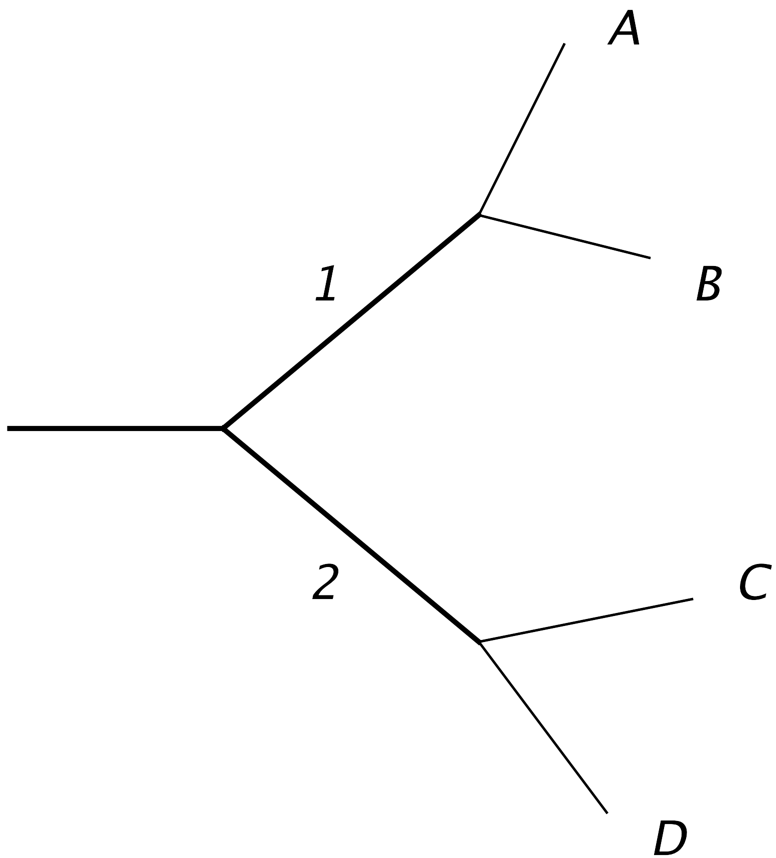
Secondary decomposition



- Decluster 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

Selecting 3 prongs from 4



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

Imposes an m_{min} condition

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

This method is collinear unsafe!

CMS tagger with angular cut

Original CMS tagger suffers from collinear unsafety CMS-PAS-JME-09-001

A later version introduces an angular cut in addition to the $\zeta_{\rm cut}$

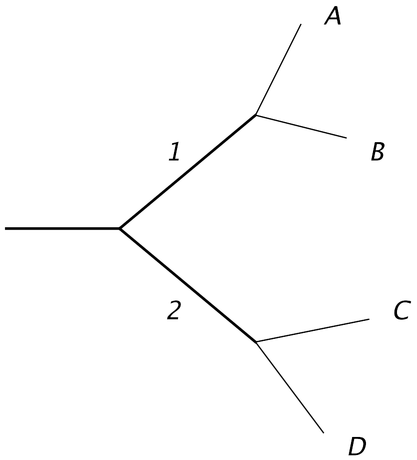
$$\Delta R_{ij} > 0.4 - A p_T$$

CMS-PAS-JME-13-007

with A = 0.0004 GeV^{-1}. Cuts off collinear divergence but vanishes at 1 TeV.

Modified taggers

IRC unsafe tagger may not be reliable so create modified tools



- CMS^{3p,mass} finally selects only the larger invariant mass de-clustering. This restores collinear safety with no ΔR

MD, Guzzi, Rawling, Soyez

MANCHESTER 1824

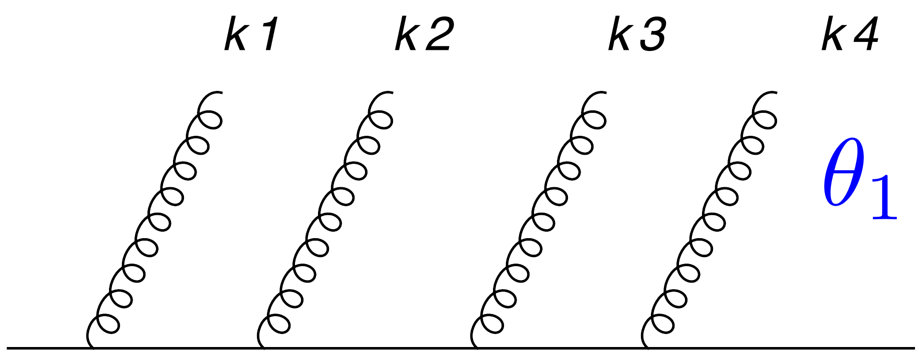
The University of Manchester

Modified taggers

Another method : TopSplitter

MD, Guzzi, Rawling, Soyez 2018

Take not largest angle emission but emission that "dominates prong mass" as product of declustering.



$$\theta_1 > \theta_2 > \theta_3 > \theta_4$$

$$z_1 > \zeta_{\mathrm{cut}}$$

Follow hardest branch and go all the way down C/A tree to find largest $pti\,\theta_i^2$ emission.

MANCHESTER
1824

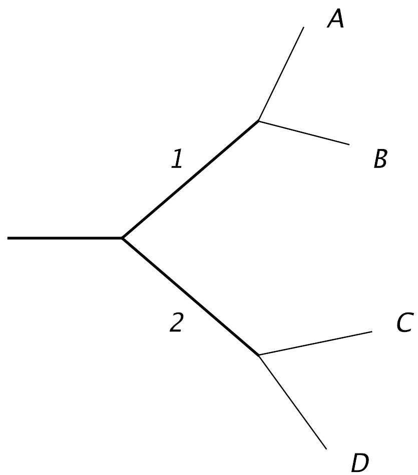
The University of Manchester

Y_m-splitter

- Uses gen-kt (p=1/2) algorithm for de-clustering. Equivalent to mass ordering in soft limit.

MD, Guzzi, Rawling, Soyez 2018

- Not recursive but continue to use $\zeta_{\rm cut}$



Consider prong with larger gen-kt value as declustered if $\zeta_{\rm cut}$ passes.

Also needs grooming

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

Analytics for QCD jets

We calculate jet mass distribution after application of taggers.

Define

$$\rho = \frac{m^2}{p_T^2 R^2}$$ and $$\rho_{\min} = \frac{m_{\min}^2}{p_T^2 R^2}$$

Compute $\frac{d\sigma}{d\rho}$ for fixed $\rho_{\min}$ and related quantities.

MANCHESTER 1824
The University of Manchester

Analytics for QCD jets

With m ~ m_t and m_{min} ~ m_w at high p_t :

$$\rho, \rho_{\rm min} \ll 1$$

$$L_\rho = \ln \frac{1}{\rho} \gg 1$$

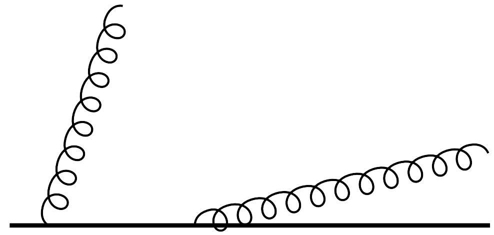
$\Rightarrow$ **Resum large logs**

Also we have no strong ordering in these masses.

$$L_\rho \sim L_{\rho_{\rm min}} \gg 1$$

and $\zeta_{\rm cut} \sim 0.05$ $\quad L_{\rho,\rho_{\rm min}} \gg L_\zeta$

Leading order calculation



Two real emissions to pass the tagger so starts with $\mathcal{O}\left(\alpha_s^2\right)$. For simplicity take limit $\rho \gg \rho_{\text{min}}$

$$\frac{d\sigma}{d\rho} \sim \frac{\alpha_s^2}{\rho} \ln^2 \frac{1}{\zeta_{\text{cut}}} \ln \frac{\rho}{\rho_{\text{min}}}$$

soft **strong-ordered**

Compare to QCD jet $\frac{\alpha_s^2}{\rho} \ln^3 \frac{1}{\rho}$

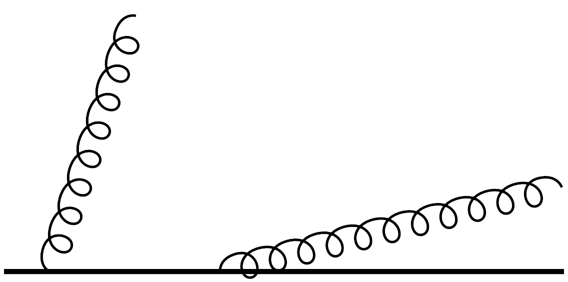
Reduced background after tagging.

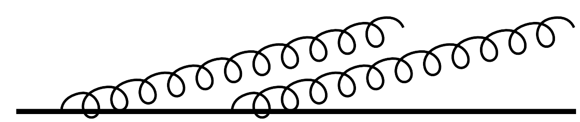
Triple collinear limit for QCD jet

But $\ln\dfrac{\rho}{\rho_{\min}},\ \ln\dfrac{1}{\zeta_{\rm cut}}$ are not too large.

Need to lift strong ordering and soft approx.

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

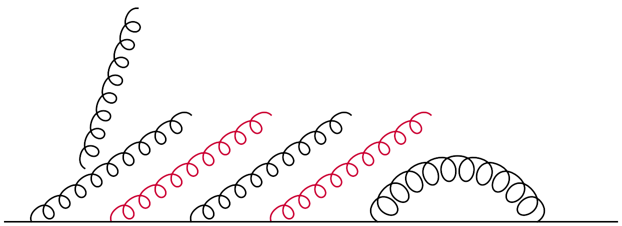




Standard LO DGLAP or PS evolution

Triple collinear splitting functions

All orders



- Beyond leading order : constraints on real emissions arise from ρ and $\rho_{\rm min}$ conditions.

  ⇨ **Sudakov form factors**

- Our resummation accuracy is modified LL. Resums all double logs $\frac{1}{\rho}\alpha_s^n L^{2n-1}$

- Counts $\ln\frac{1}{\rho}, \ln\frac{1}{\rho_{\rm min}}, \ln\frac{1}{\zeta_{\rm cut}}, \ln\frac{\rho}{\rho_{\rm min}}$ all on same footing

- Also includes NLL effects from running coupling and hard collinear emissions.

Results

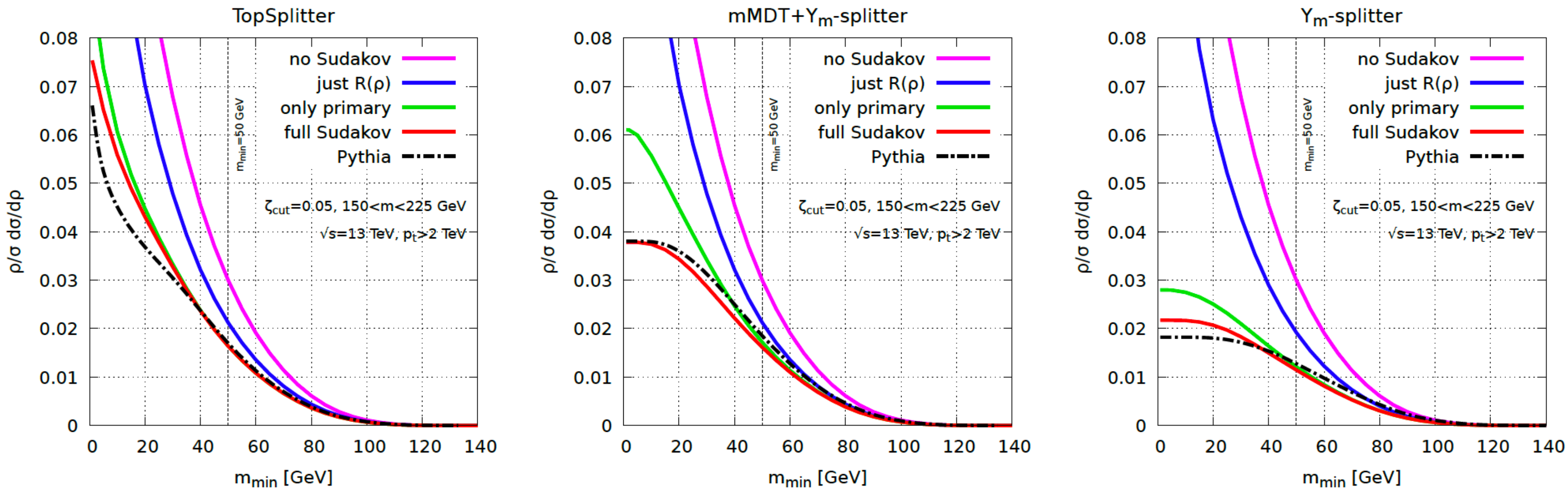
General form:

$$\rho \frac{d\sigma}{d\rho} = \int d\Phi_3 \frac{\hat{P}}{s_{123}^2} \frac{\alpha_s\left(k_{t1}\right)}{2\pi} \frac{\alpha_s\left(k_{t2}\right)}{2\pi} \Theta^{\mathrm{jet}} \Theta^{\mathrm{tagger}} \delta\left(\rho - \frac{s123}{p_t^2 R^2}\right) \times e^{-R}$$

- Prefactor computed using triple-collinear splitting functions and phase space
- Convoluted with a Sudakov form factor accounting for all leading 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.

Campbell and Glover 1997, Catani and Grazzini 1998

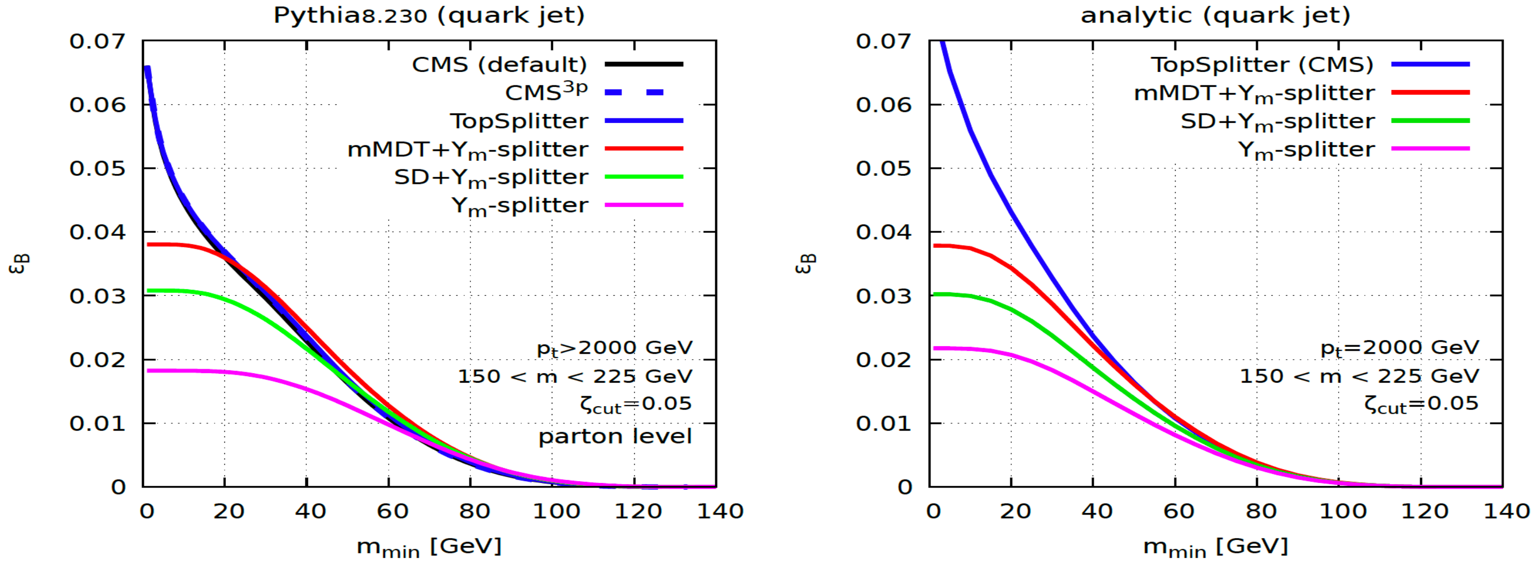
Results and comparisons to PS



MD, Guzzi, Rawling, Soyez. Preliminary

- Plots reflect that resummation of $\ln \frac{\rho}{\rho_{\min}}$ terms does matter
- Inclusion of secondary emissions important at small m_{min}
- Overall a good agreement with PS.

Tagger comparisons for QCD jets



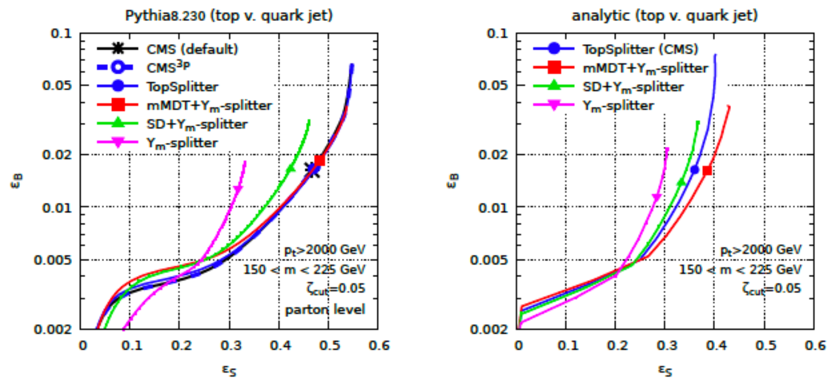
MD, Guzzi, Rawling and Soyez, 2018

- 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.

MANCHESTER 1824

The University of Manchester

Signal jets



MD, Guzzi, Rawling, Soyez. 2018

- For W/Z/H decays impact on background key to final performance. Taggers like Y-splitter are high-performance owing to large Sudakov
- For coloured top this is not the case due to signal Sudakov suppression. Also analysed signal jets with a basic Sudakov. Groomed Y-splitter comparable to CMS and variants.

Conclusions

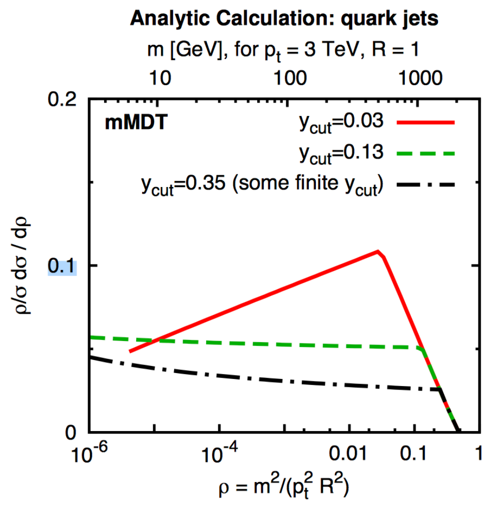
- A first analytic study of aspects of top taggers carried out.
- Shows analytic control over basic features
- Large Sudakov effects not necessarily desirable and hurt signal efficiency.
- CMS tagger become potentially unsafe at high $p_{t..}$ Potentially harmful for precision studies. Easy to design safe variants with no change in performance.
- Plan to investigate combinations with jet shape variables like τ_{32} as next step.

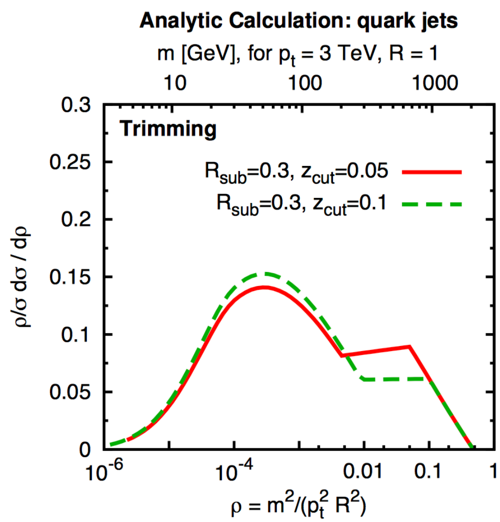
MANCHESTER 1824

The University of Manchester

BACK UP MATERIAL

Analytical insight



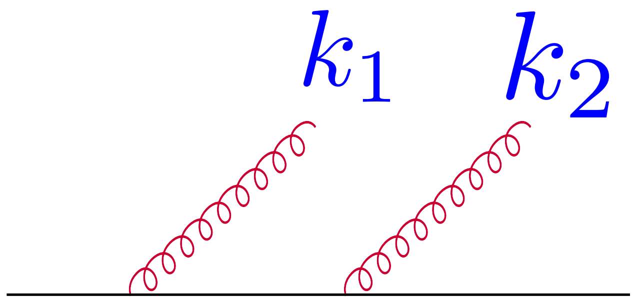


- Traditional approach : Construct taggers on simple intuitive ideas. Leave details to MC studies. Lots of freedom to create many new tools.

- Analytical approach : Worry about details. Get main physics principles. Then construct optimal tools.

MD, Fregoso, Marzani and Salam 2013

Results



$$\otimes \quad \exp[-R]$$

$$\rho_2 = z_2\theta_2^2 < z_1\theta_1^2$$

- The key differences between taggers come from the Sudakov.
- Y_m-splitter has a plain jet mass double log Sudakov in ρ_2
- TopSplitter and safe variants of CMS have an mMDT style single-log Sudakov
- mMDT/SoftDrop grooming + Y_m-splitter inherits grooming Sudakov structure. MD, Fregoso, Marzani and Salam 2013. Larkoski, Marzani, Thaler and Soyez 2014.

Top tagging methods

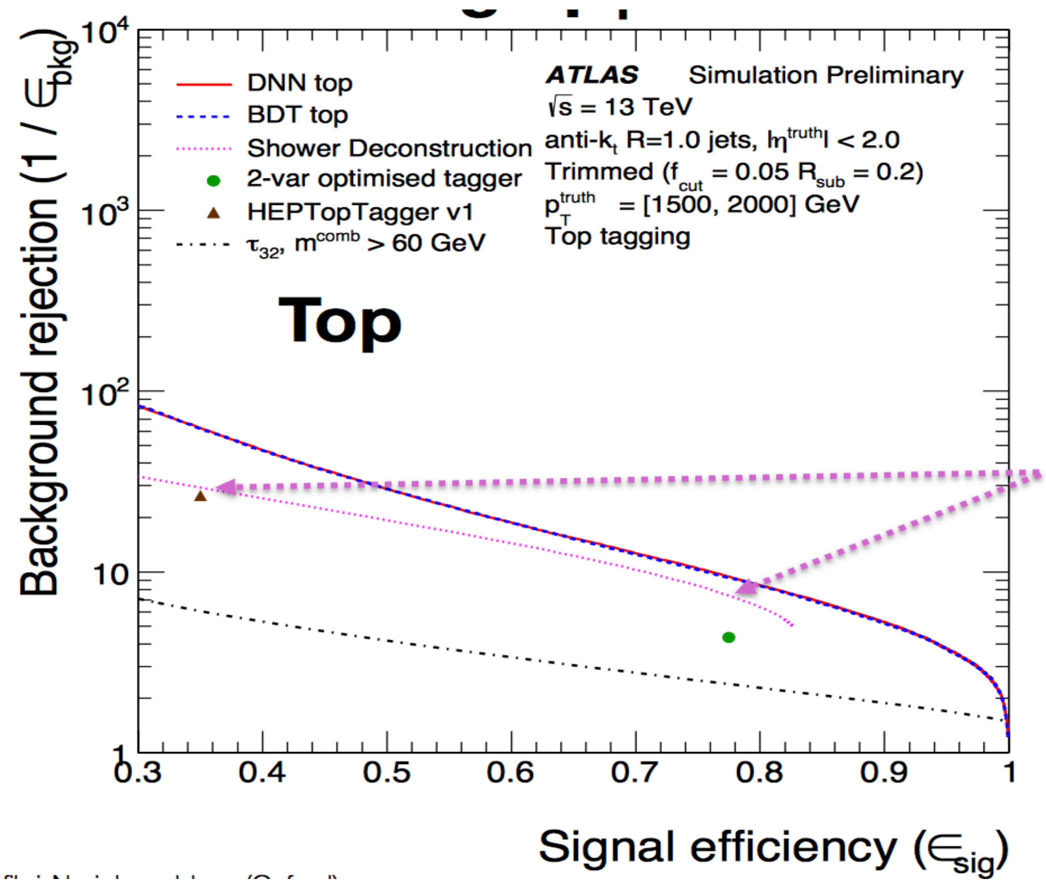


Figure from talk by N.Norjoharuddin on behalf of ATLAS, Boost 2017