Jet Substructure, W- and Top-Tagging in CMS

Emanuele Usai

Universität Hamburg

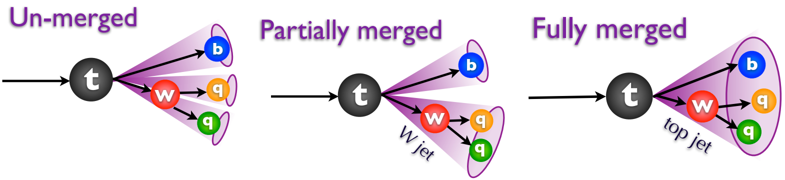
December 2, 2013

7th Workshop of the Helmholtz Alliance "Physics at the Terascale"

Team at Universität Hamburg: Daniel Gonzalez, Johannes Haller, Rebekka Höing, Roman Kogler, Vilius Kripas, Tobias Lapsien, Ivan Marchesini, Dominik Nowatschin, Jochen Ott, Thomas Peiffer, Alexander Schmidt, Emanuele Usai

Why jet substructures? An example

Decay of a top quark



low p_T moderate p_T high p_T

- Top quarks decay hadronically 68% of the times.
- as we reach higher energies, hadronic decay products start to merge.

The strategy:

- use a bigger, single jet to cluster the whole decay
- use substructure tools to filter and decluster the decay products:
  - taggers: top-taggers, W-taggers, H-taggers
  - substructure variables: N-subjettiness, mass drop, pruned mass, etc.

Outline

Substructure tools

- Current status in CMS

Top taggers

- CMS Top tagger
- HEP Top tagger

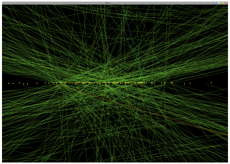
B-tagging

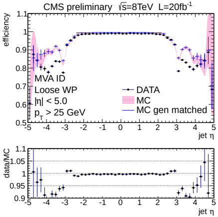
- B-tagging in boosted topologies

Jet substructure in CMS: an overview

Pileup Jet-ID:

- exploit also non-tracking quantities (jet shape) to extend PU rejection outside of the tracking acceptance
- multivariate discriminant

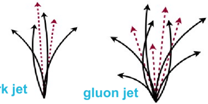


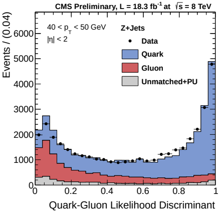


Quark/gluon discriminator
Quark- and gluon-initiated jets have different properties:

- constituents multiplicity
- jet area
- constituents energy fraction

Create a multivariate discriminant



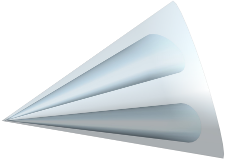


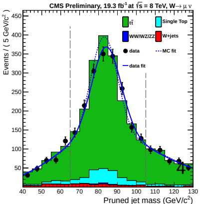
Jet grooming algorithms: get rid of soft energy/wide angle components in a jet:

- Jet filtering, trimming, pruning

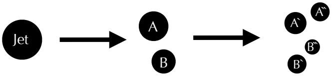
W-tagging

- cluster an hadronic W in a **CA R=0.8 jets**
- cut on pruned jet mass $60 < m_j < 100$ GeV





Top tagging in CMS

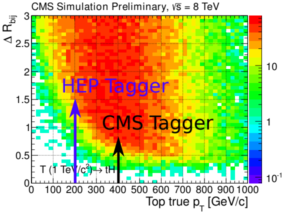


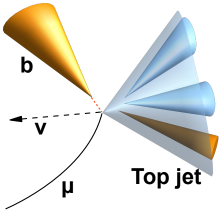
- Cluster a jet using **CA R=0.8**
- Decluster in two stages in order to find **up to 4 subjets**
- Subjets must satisfy p_T fraction and adjacency (ΔR) criteria

Tagging variables:

- Jet mass (m_{jet})
- Number of subjets (N_{sub})
- Minimum pairwise mass (m_{min}) of leading 3 subjets
  $m_{min} = \min(m_{12}, m_{13}, m_{23})$

CA8 jets start to cluster hadronic top decays from $p_T \sim$ **400 GeV**





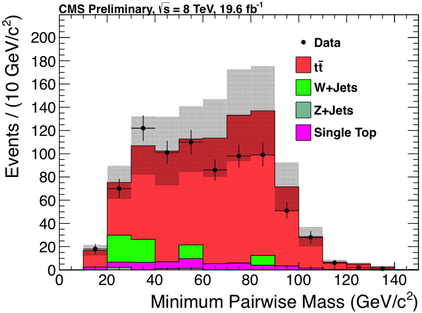
A $t\bar{t}$ **semileptonic** event topology is considered.
Tag and probe approach:, .

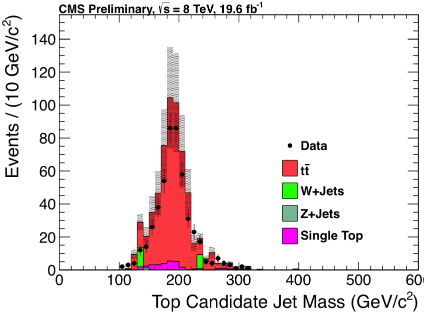
- use leptonic top to **"tag"**
- **"probe"** the hadronic side

Control Data/MC agreement, assess performance.

CMSTopTagger in action

Top candidates after a semileptonic $t\bar{t}$ selection and CMSTopTagger requirements.
Used to derive CMSTopTagger Data/MC correction.





Minimum pairwise mass, **W candidate**

- Grayed area is MC normalization uncertainty

Top jet mass distribution

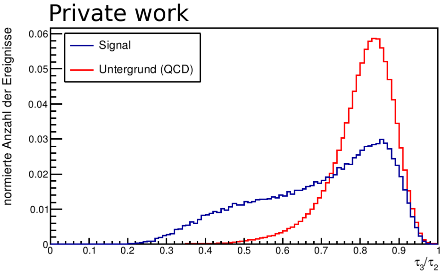
- Top tagged
- $m_{min} > 50$ GeV

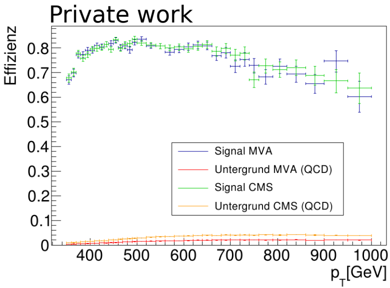
CMSTopTagger optimization

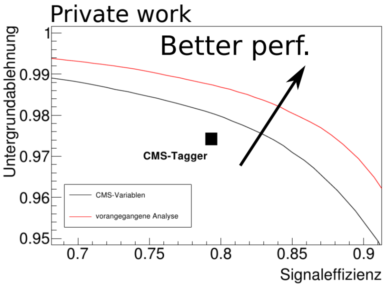
Optimize the CMSTopTagger with MVA techniques using:

- tagger's variables
- N-subjettiness
- Q-jet volatility
- many others...

N-subjettiness τ_N: encodes the likelihood of a jet to contain n subjets. For top-jets the relevant quantity is τ_3/τ_2.







In this study, the decay products of the top quark are required to have $\Delta R(\text{particle}, \text{jet axis}) < 0.8$.

The HepTopTagger

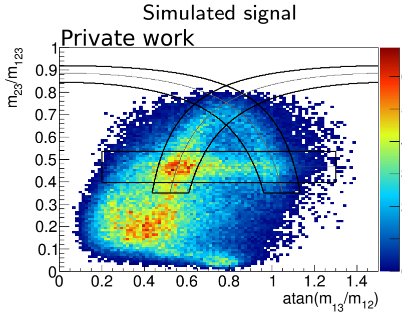
Developed by Plehn et al., arXiv:1006.2833
Iterative reclustering, starting from CA R=1.5 jet coll., forces three subjets (→ fat jet).

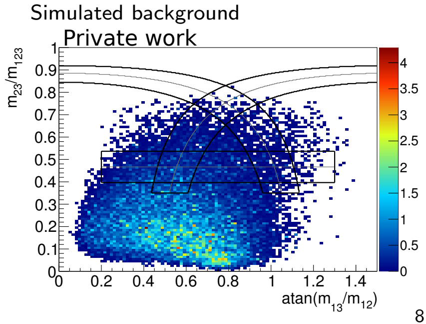
- $p_T > 200$ GeV
- $140 < m_{jet} < 250$ GeV

Define: m_{ij} = inv. mass of i-th and j-th p_T leading subjet.
Select signal using tagging variables:

- **arctan $\frac{\mathbf{m_{13}}}{\mathbf{m_{12}}}$**
- $\frac{\mathbf{m_{23}}}{\mathbf{m_{123}}}$

Define this jet a “top tagged jet”.

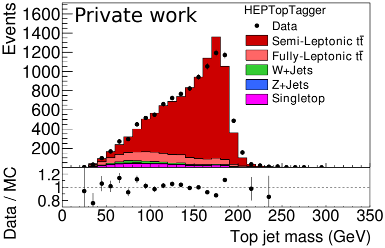




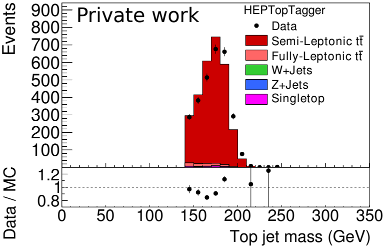
Performance and Data/MC comparison

Before the HEPTopTagger cuts | After the HEPTopTagger cuts

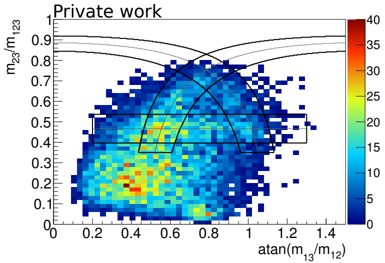
TopJet mass



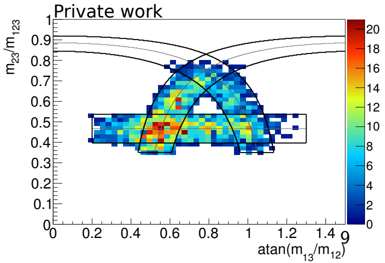
Top tagger selection



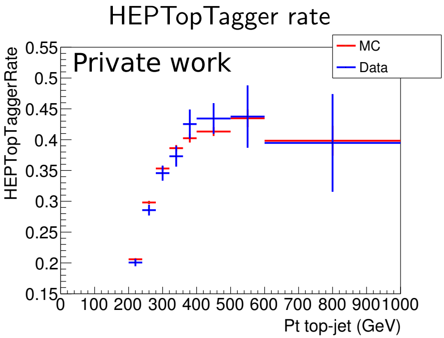
Data

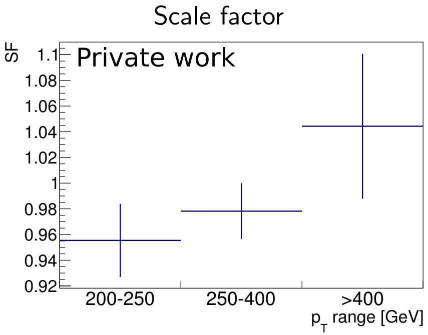


Top tagger selection



Performance and calibration





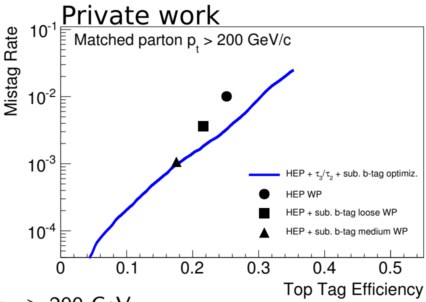
- $\epsilon = \frac{N_{evts \text{ after HepTopTagger}}}{N_{evts \text{ before HepTopTagger}}}$
- unmerged topjets contamination at low p_T

- $SF = \frac{\epsilon_{\text{Data-Bkg}}}{\epsilon_{\text{MC Signal}}}$

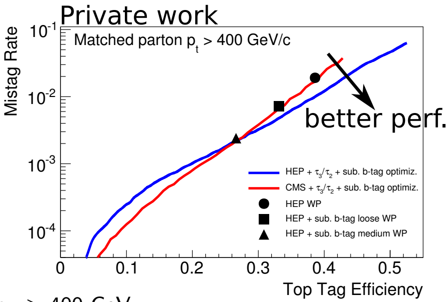
Performance comparison with CMSTopTagger

Many search analyses have moderately boosted top jets, below 600 GeV.

The HEPTopTagger can be a better choice in many cases.



$p_T > 200$ GeV



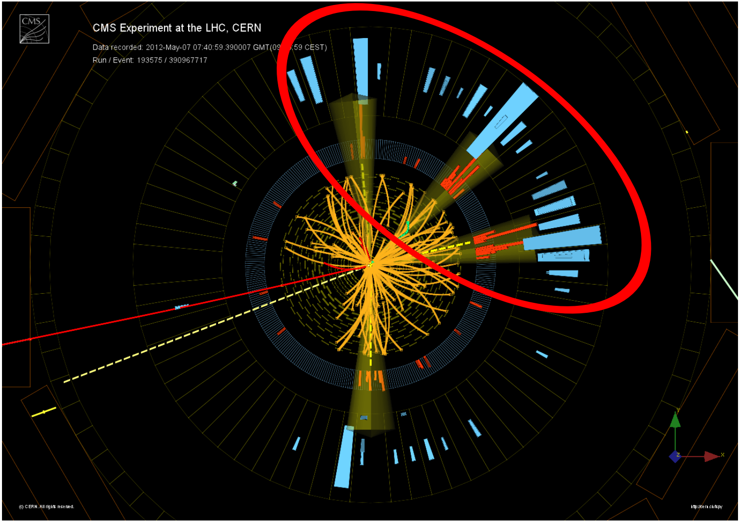
$p_T > 400$ GeV

- The p_T range 200-400 GeV is not covered by CMSTopTagger.
- HEPTopTagger developed specifically for moderately boosted regimes.

- HEPTopTagger still working well.
- For $p_T > 600$ GeV the CMSTopTagger performs better.

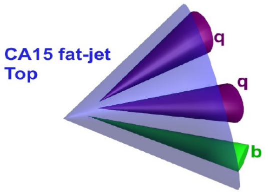
HEPTopTagger: event display

Private work

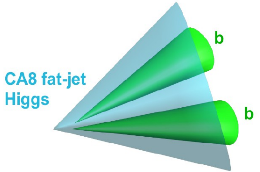


B-tagging in boosted topologies [CMS-PAS-BTV-13-001]

- B-tagging at CMS traditionally developed on isolated AK5 jets, mostly suitable for the non-boosted regime.
- First study at LHC dedicated to b-tagging in the boosted regime. Benchmark topologies:



Boosted top, hadronic decay: selected using HEPTopTagger, CA15 jet collection



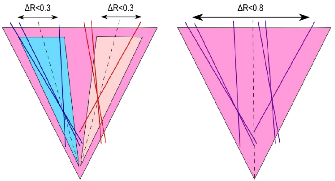
Boosted Higgs$\rightarrow b\bar{b}$: studies based on pruned CA8 jets

- Boosted studies based on the Combined Secondary Vertex CSV tagger: likelihood ratio combination of secondary vertex + single track information.

B-tagging in boosted topologies [CMS-PAS-BTV-13-001]

Two scenarios considered:

- **subjet CSV**:
  CSV b-tagger applied to subjets (2 b- tags for Higgs-tagging, ≥ 1 for top- tagging)
- **fat-jet CSV**:
  CSV b-tagger applied to the Higgs/top candidate fat-jet



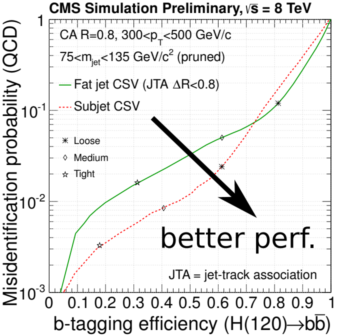
subjet b-tagging: based on subjet tracks

fat-jet b-tagging: based on all fat-jet tracks

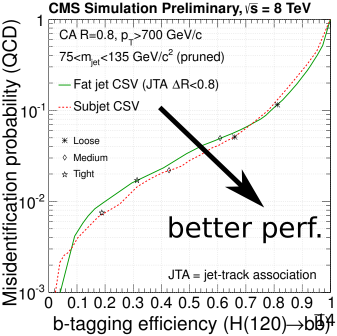
Subjet b-tagging generally performs better: chosen as default technique.

Fat-jet b-tagging suitable at very high pT where subjets start to merge.

Higgs channel, **moderate** p_T



Higgs channel, **very high** p_T



B-tagging in boosted topologies [CMS-PAS-BTV-13-001]

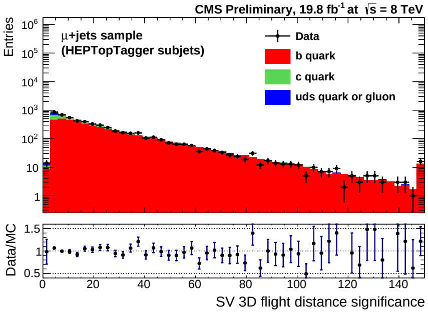
Control samples

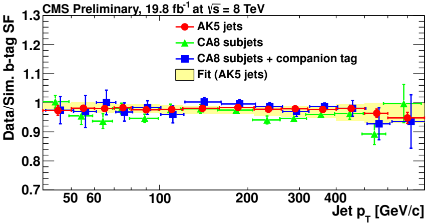
Boosted top:

- μ+jets, semileptonic ttbar

Boosted Higgs: challenging definition of the control sample

- similar topology: gluon splitting jets, two closeby b's

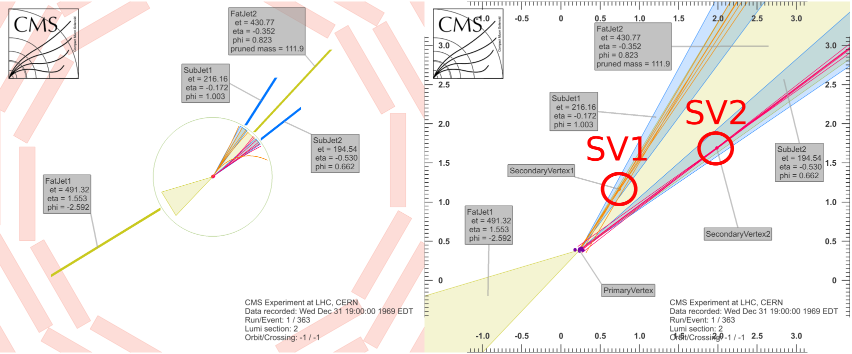




- Good data/MC agreement for b-tagging observables.
- All observables cross-checked.
- SF$\sim$ 1, compatibly with SF for standard b-tagging in the non-boosted regime, for both channels.
- Nothing pathological in the boosted regime.

B-tagging in boosted topologies: event display

Simulated Radion→HH→4b event at 8 TeV containing a boosted H→ $b\bar{b}$ jet reconstructed as a CA8 fat jet



The Radion mass is set to 1.5 TeV/c^2 and the Higgs boson mass is set to 125 GeV/c^2

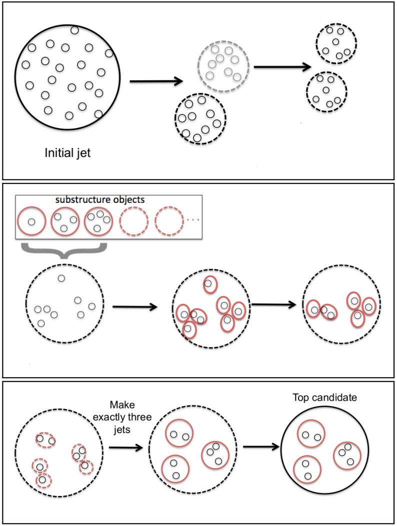
Thank you for the attention!

Would you like to know more?

- Jet grooming: JHEP 05 (2013) 090
- Quark/gluon discriminator: CMS PAS JME-13-002
- Pileup jet identification: CMS PAS JME-13-005
- W tagger: CMS PAS JME-13-006
- CMS top tagger: CMS PAS B2G-12-005
- B-tagging in boosted topologies: CMS PAS BTV-13-001
- Top tagging: CMS PAS JME-13-007 (to appear soon)

Backup

The HepTopTagger



Developed by Plehn, Tilman et al., arXiv:1006.2833

Start from jet collection: **CA R=1.5**

Iterative declustering with mass drop criterium

Filtering of the constituents

Reclustering of the filtered constituents forcing exactly three subjets

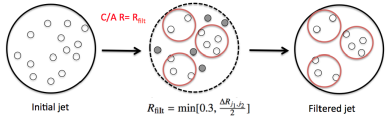
Define this jet a "top-jet" or "fat-jet".

arxiv:1306.4945v1

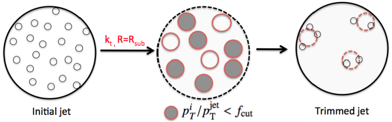
Jet grooming techniques

Get rid of softer components in a jet from UE or pileup and leave the constituents from the hard process

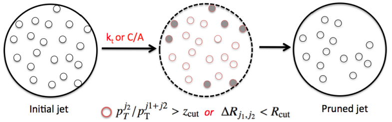
Filtering

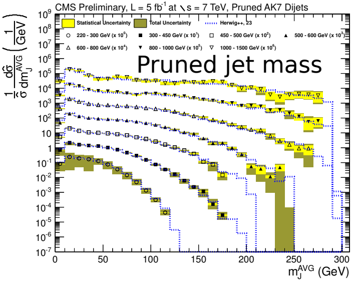


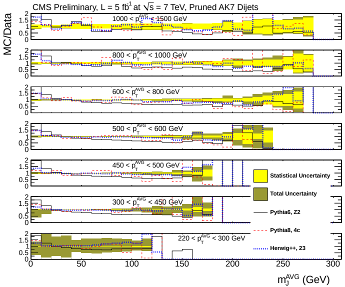
Trimming



Pruning





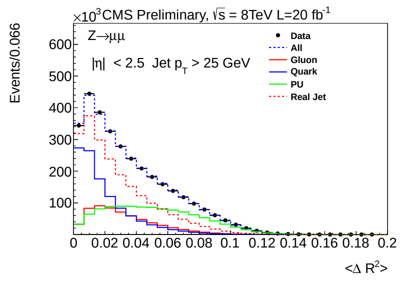


arxiv:1306.4945v1

Pileup jet identification

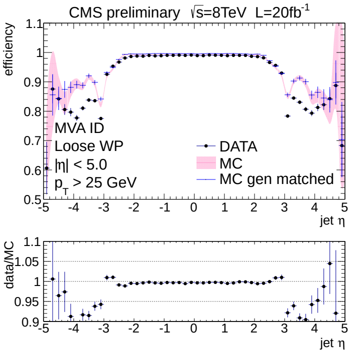
- Traditional PU subtraction: subtract charged particles not pointing to the primary vertex.
- PU Jet-ID:
  - exploit also non-tracking quantities (jet shape) to extend PU rejection outside of the tracking acceptance
  - multivariate discriminant

Z ($\rightarrow \mu\mu$) + jets events (PU)



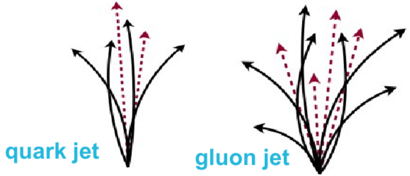
e.g. non-tracking observable: radial distribution of Particle-Flow jet-constituents

MVA discriminator efficiency



Quark/gluon jet discriminator

Quark and gluon have different colour interaction



Variables

Multiplicity

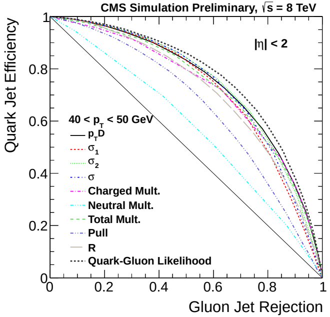
- charged, neutral, total

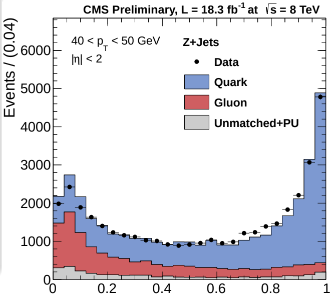
Spread

- $\eta - \phi$ spread
- major/minor $\eta - \phi$ matrix axes σ_1, σ_2

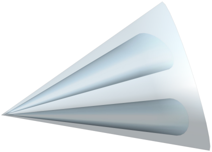
Energy sharing

- energy fraction ot the hardest component
- $p_T D = \frac{\sqrt{\sum_i p_{T,i}^2}}{\sum_i p_{T,i}}$





W tagger

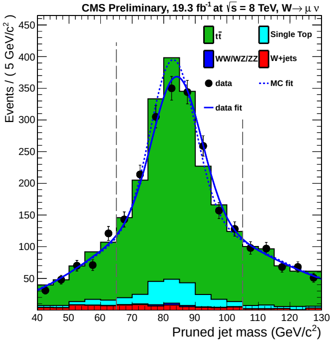


Cluster a boosted hadronic W decay in a single jet

- start from **CA R=0.8 jets**

Use jet mass pruning

- cut on pruned jet mass $60 < m_j < 100$ GeV



Extract:

W-jet mass scale (peak position):

- data: 84.5 ± 0.4 GeV
- MC: 83.4 ± 0.4 GeV

W-jet mass resolution:

- Data: 8.7 ± 0.6 GeV
- MC: 7.5 ± 0.4 GeV

data/MC correction for W-tagging efficiency (SF):

- 0.905 ± 0.08

(operating point: m_{pruned} cut + N-subjettiness $\tau_2/\tau_1 < 0.5$)

W tagger: optimization

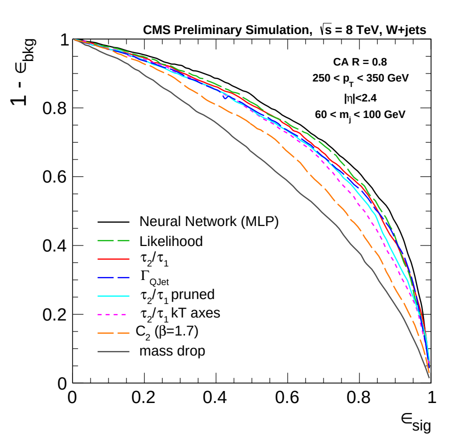
Pruning can be combined with additional observables:

- mass-drop μ
- N-subjettiness τ_n: τ_2/τ_1 used for W- tagging
- also examined: Qjet volatility Γ_{Qjet}, generalized energy correlation function C_2^β

N-subjettiness shows the best single discriminating power.

Observables are correlated: moderate improvement with multivariate combination using TMVA.

$60 < m_{pruned} < 100$ GeV

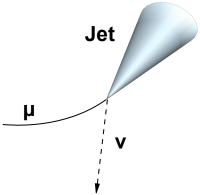


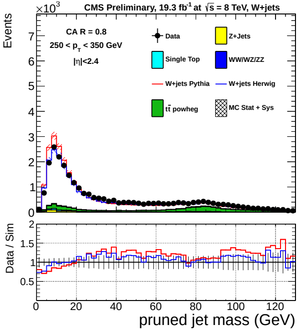
W-tagging ROC curves: single substructure observables and their combinations

W-tagger

Pruned mass - QCD jets

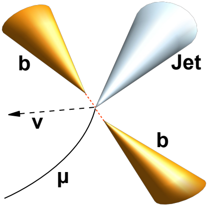
from W_{lept} + Jet_{QCD}
$p_T = 250 - 350$ GeV

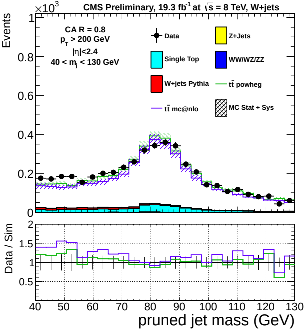




Pruned mass - W-jets

from $t_{had} \rightarrow \mathbf{W_{had}} + b$
$t\bar{t}$ semileptonic selection





Jet charge for W-jets

W^+ and W^-

- $Q^\kappa = \frac{\sum_i q_i (p_{Ti})^\kappa}{(\sum_i p_{Ti})^\kappa}$
- W-tagged jets
- $t\bar{t}$ semileptonic topology
- "true" W charge determined by looking at the lepton charge