Boosted jet algorithm development

Emanuele Usai on the behalf of the CMS Collaboration

Universität Hamburg

Emmy Noether-Programm
Deutsche Forschungsgemeinschaft
DFG

August 15, 2013

Outline

Top tagging

- Data/MC comparison

W tagging

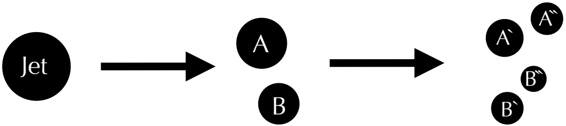
- Substructure variables
- Data/MC comparison
- Optimization
- Performance and behavior in particular scenarios

Top tagging

based on CMS-PAS-B2G-12-005

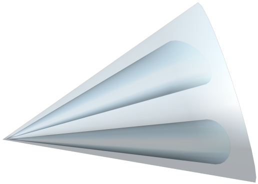
Top tagging in CMS

- Based on JHU Top Tagger (Kaplan et al.)
- Cluster a jet using **CA R=0.8**
- Decluster in two stages in order to find **up to 4 subjets**



- Subjets must satisfy two requirements
  - Momentum fraction criterion: $p_{T\ subjet} > 0.05 \times p_{T\ jet}$
  - Adjacency criterion: $\Delta R(C_1, C_2) > 0.4 - 0.0004 \times p_T(C)$
- Remove subjets which fail momentum fraction cut and try to decluster again
- Tagging variables
  - Jet mass (m_{jet})
  - Number of subjets (N_{sub})
  - Minimum pairwise mass (m_{min}) of leading 3 subjets
    $m_{min} = \min(m_{12}, m_{13}, m_{23})$

W tagging

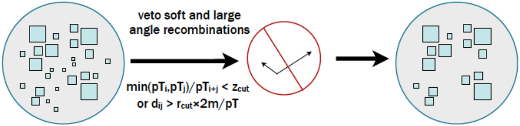


Jet collection:

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

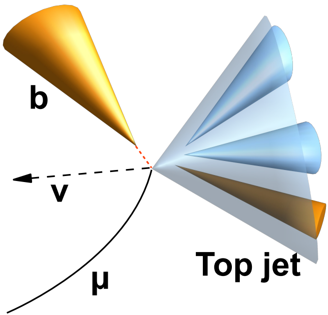
Use jet mass pruning (see Ellis, Vermillion, Walsh [arXiv:0903.5081] and CMS-PAS-SMP-12-019):

- recluster the jet using all CA8 jet particles
- for each recombination ignore the softer protojet if
  - $z = \min(p_T^i, p_T^j)/p_T^p < 0.1$, where i, j for protojets, and p_T^p for combined jet.
  - $\Delta R > D_{\text{cut}} = 0.5 \times m^{\text{orig}}/p_T^{\text{orig}}$ with respect to the previous recombination step, m^{orig} and p_T^{orig} for original CA8 jet.

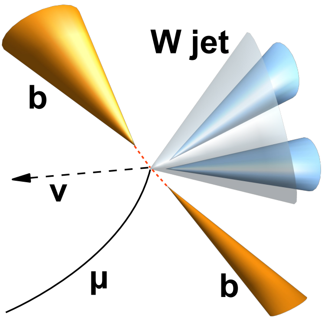


Event topologies

$t\bar{t}$ event selection: common leptonic selection



Hadronic top fully merged



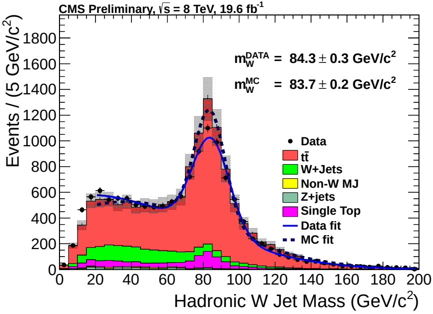
Hadronic top partially merged

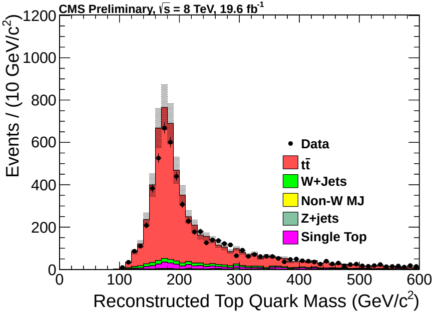
- 1 CA 0.8 jet opposite to μ

- 1 W-tagged CA0.8 jet opposite to μ
- 1 b-jet (closest jet to W-jet)

Partially merged top hadronic decay

Top candidates after a semileptonic $t\bar{t}$ selection.
Used to derive subjet energy scale uncertainty.





Mass of W candidate:

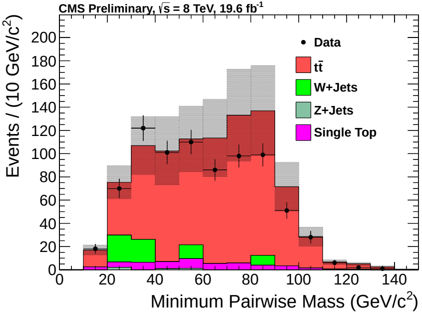
- highest-mass jet in the hemisphere opposite the identified muon.
- pruned jet mass

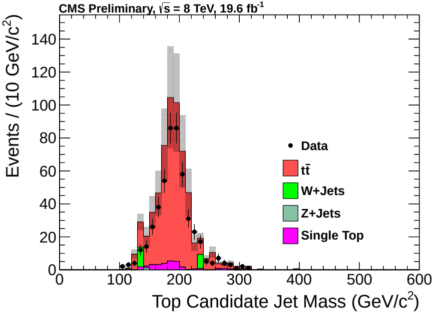
Top candidate mass

- combined invariant mass of W candidate and closest jet to W candidate
- No b-tagging on closest jet to W candidate

Fully merged top hadronic decay

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





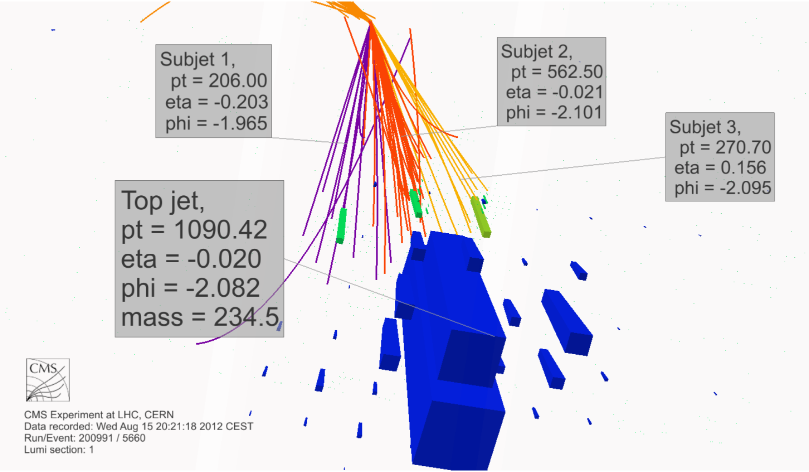
Minimum pairwise mass, W candidate

- Grayed area is MC normalization uncertainty

Top jet mass distribution

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

Event display



Top tagging: conclusions and perspectives for the future

Current status:

- CMSTopTagger is a mature and performing tool **widely used in CMS** analyses
- New developments like **b-tagging in subjets** make this tool even more performing. See Ivan Marchesini's talk.

Coming soon:

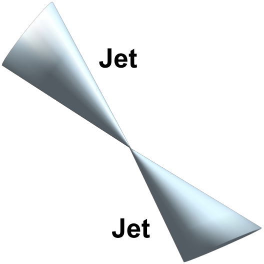
- **HEPTopTagger** is being commissioned with CMS data
- Document under review, results will be **public soon**! (CMS-PAS-JME-13-007)

W tagging

based on CMS-PAS-JME-13-006
http://cds.cern.ch/record/1577417

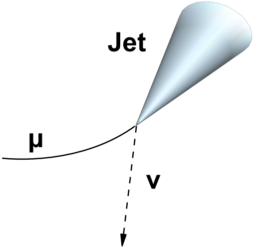
Event topologies considered

Dijet



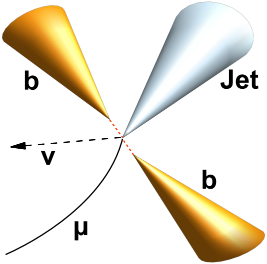
- two hard jets
- p_T = **400-600 GeV**
- accesses high p_t region
- QCD-jet dominated
- used to study fake rate

W+jets



- leptonic W + jet
- p_T = **250-350 GeV**
- accesses low p_T region
- QCD-jet dominated
- presence of non-dominant background ($t\bar{t}$, single top)
- used to study fake rate

$t\bar{t}$



- leptonic top decay + hadronic top
- highly pure sample of W-jets
- used to study efficiency

Benchmark signal: $X \rightarrow W_L W_L$, $M_X = 600\ GeV,\ 1\,TeV$

MC samples and showering models

Next slides show Data/MC comparisons for different generators and showering models.
The comparisons are made with these particular version and tuning of the generators:

Montecarlo samples

QCD

- MADGRAPH + PYTHIA6
- HERWIG++
- PYTHIA8

W+jets

- MADGRAPH + PYTHIA6
- HERWIG++

$t\bar{t}$

- POWHEG + PYTHIA6
- MC@NLO + HERWIG++

Signal ($X \to W_L W_L$)

- JHU GENERATOR + PYTHIA6

Signal ($H \to WW$)

- POWHEG + PYTHIA6

Tuning of showering models

PYTHIA6

- version 6.426, tune Z2*

PYTHIA8

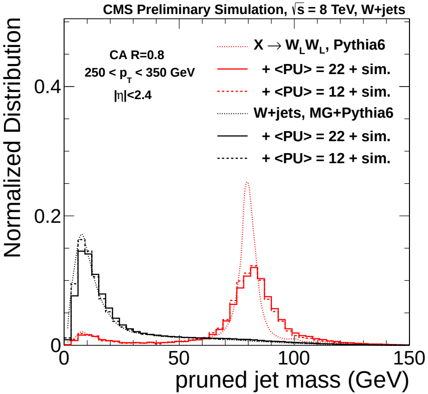
- version 8.153, tune 4C

HERWIG++

- version 2.5.0, tune 23

Pruned mass

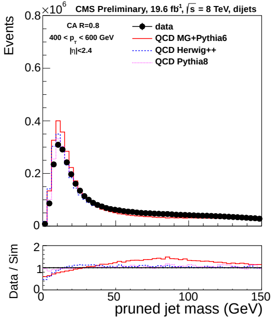
$p_T = 250 - 350$ GeV
(W+jets)



- Signal sample peaks at W mass
- QCD jets mass peaks at low masses after jet pruning
- CMS detector simulation + pileup results in broadening of W mass peak and shift towards higher mass values
- Pileup dependence for 12 and 22 average PU interactions is small due to jet pruning
- from now on a pruned jet mass cut $60 < m_{jet} < 100$ is used.

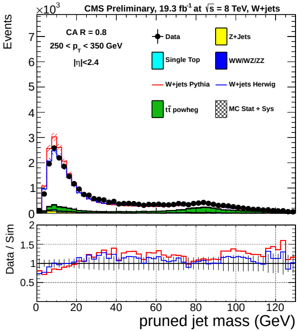
Data-montecarlo comparisons - pruned jet mass

$p_T = 400 - 600$ GeV (dijet)



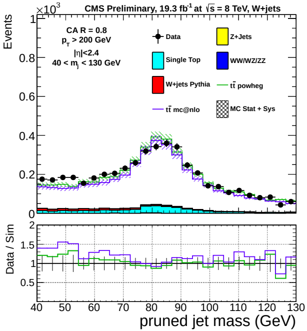
- overall good agreement
- different parton shower models
- Pythia 6 appears to be the worst
- QCD only

$p_T = 250 - 350$ GeV (W+jet)



- worse agreement at low mass
- other non-dominant background processes present ($t\bar{t}$)

W-jets ($t\bar{t}$)



- W peak visible
- single-top main background process
- Data/MC disagreement motivates Data/MC correction factor measurement

Substructure variables

- **Mass drop,** μ: Two subjets are obtained by undoing the last iteration of the pruned jet clustering. The ratio of masses of the highest mass subjet (m_1) and the total pruned jet mass is defined as the *mass drop* $\mu = \frac{m_1}{m_{\text{jet}}}$.
- **N-subjettiness,** τ_N: For N subjets of a given jet:

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \min\{\Delta R_{1,k}, \Delta R_{2,k}, \cdots, \Delta R_{N,k}\}$$

  k runs over all constituent particles, $d_0 = \sum_k p_{T,k} R_0$, and R_0 is the original jet radius. The τ_N observable is in effect, a measure of how many subjets a jet has. For boosted W identification the ratio τ_2/τ_1 is of particular interest.
- **Qjet volatility,** Γ_{Qjet}: Defined as the RMS of the mass distribution of jet trees over the average jet mass, $\text{volatility} = RMS/\langle m \rangle$. Where N_{trees} is chosen to be 50.
- **Generalized energy correlation functions,** C_2^β: The 3-point correlation function is particularly useful for W-tagging.

$$C_2^\beta = \frac{\sum_{i,j,k} p_{Ti} p_{Tj} p_{Tk} (R_{ij} R_{ik} R_{jk})^\beta \sum_i p_{Ti}}{(\sum_{i,j} p_{Ti} p_{Tj} (R_{ij})^\beta)^2}$$

- **Jet charge,** Q^κ

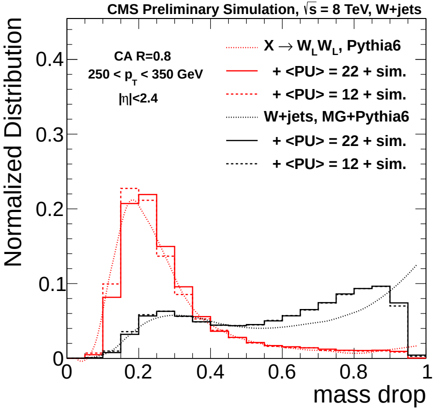
$$Q^\kappa = \frac{\sum_i q_i (p_{Ti})^\kappa}{(\sum_i p_{Ti})^\kappa}$$

  Provides additional discrimination between quark and gluon jets or between BSM signals.

All variables studied on top of the pruned jet mass cut ($60 < m_{\text{pruned}} < 100$ GeV).

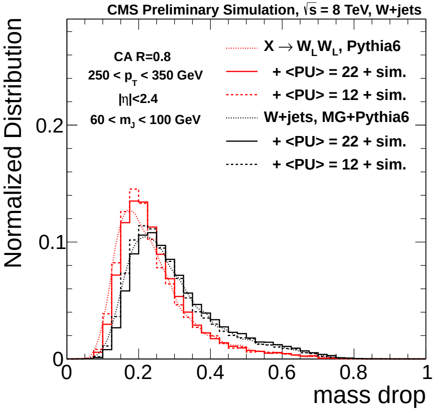
Substructure variables: mass drop, μ

$p_T = 250 - 350$ GeV
(W+jet) - no pruned mass cut



Good discrimination power

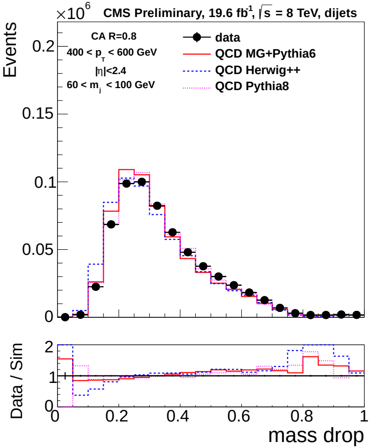
$p_T = 250 - 350$ GeV
(W+jet) - pruned mass cut



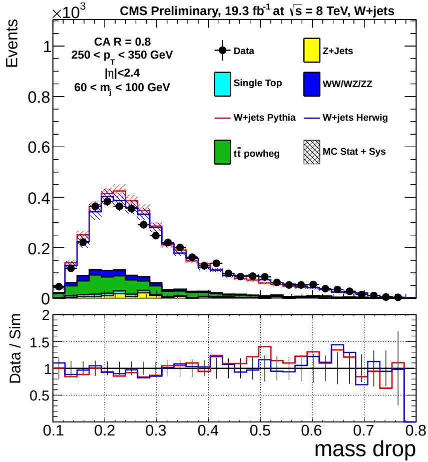
Discrimination power reduced:
correlation with mass cut

Mass drop: Data/MC comparison

$p_T = 400 - 600$ GeV
(dijet)



$p_T = 250 - 350$ GeV
(W+jet)



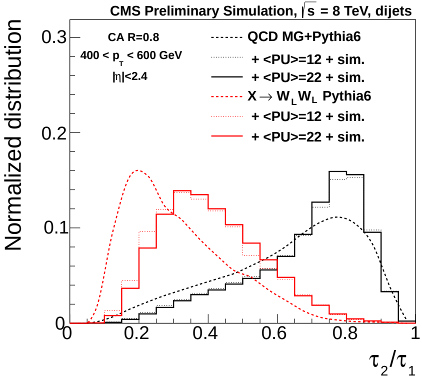
Right plot: Real W jets in green at low mass.

Substructure variables: N-subjettiness

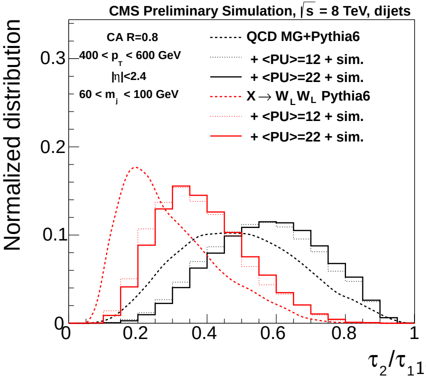
Three variants considered:

- τ_2/τ_1: one step optimization of the k_T subjet axes
- τ_2/τ_1 k_T axes: no optimization
- pruned τ_2/τ_1: uses only pruned constituents + one pass optimization.

$p_T = 400 - 600$ GeV (dijet) - no pruned mass cut

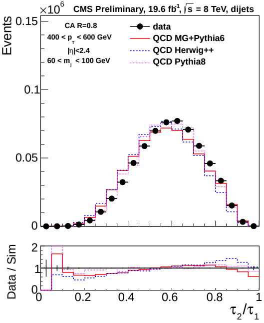


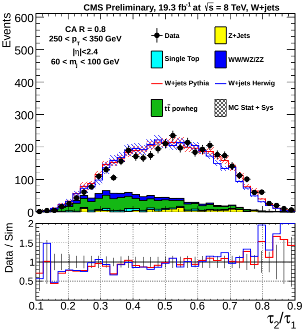
$p_T = 400 - 600$ GeV (dijet) - pruned mass cut

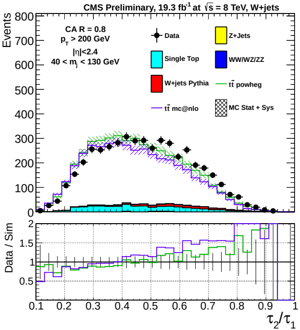


τ_2/τ_1: Data/MC comparison

$p_T = 400 - 600$ GeV (dijet)	$p_T = 250 - 350$ GeV (W+jet)	W-jets ($t\bar{t}$)







Common ascending trend in the Data/simulation ratio in function of τ_2/τ_1

Disagreement motivates measurement of Data/MC correction factor

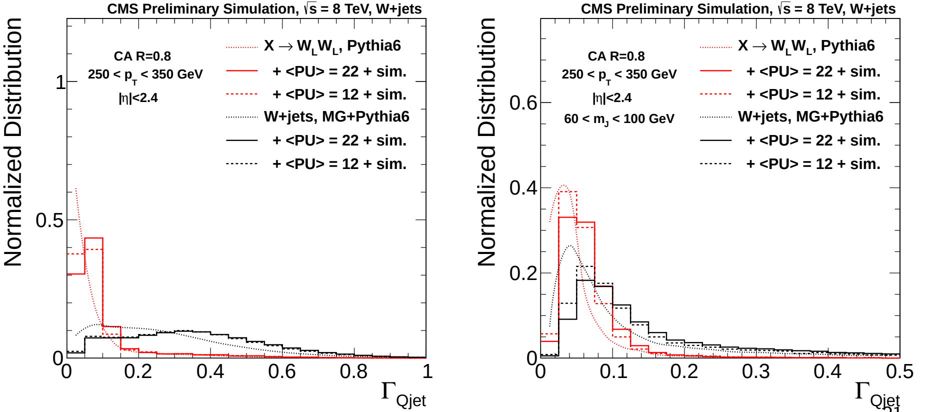
Right plot: Pythia8 best modeling, Herwig worst agreement

Substructure variables: Qjet volatility, Γ_{Qjet}

$N_{\text{trees}} = 50$
$N_{\text{preclustered components}} = 35$
$\alpha = 0.1$

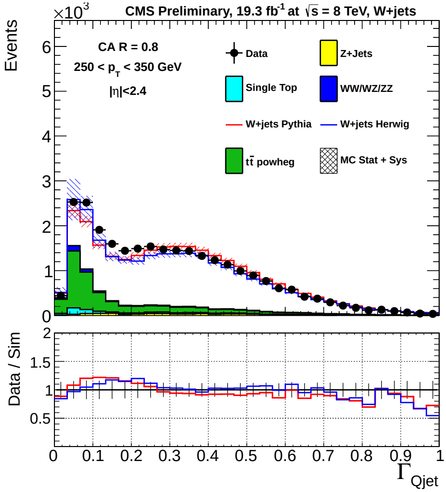
$p_T = 250 - 350$ GeV
(W+jets) - no pruned mass cut

$p_T = 250 - 350$ GeV
(W+jets) - pruned mass cut

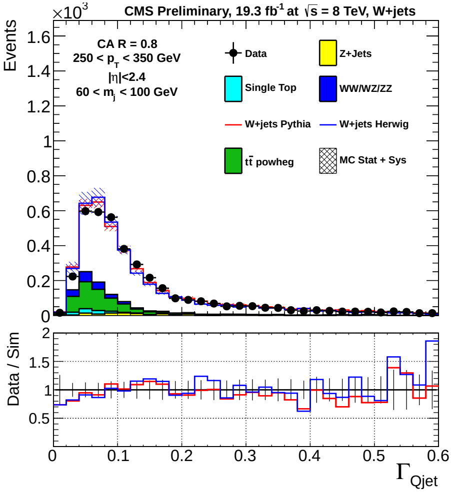


Γ_{Qjet}: Data/MC comparison

$p_T = 250 - 350$ GeV
(W+jets) - no pruned mass cut



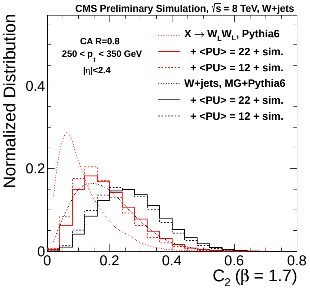
$p_T = 250 - 350$ GeV
(W+jets) - pruned mass cut



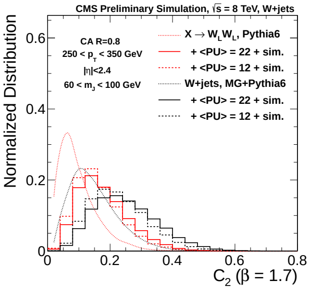
W from $t\bar{t}$ peaks at low values.
Retains good discrimination power also after mass cut

Substructure variables: generalized energy correlation function, C_2^{β}

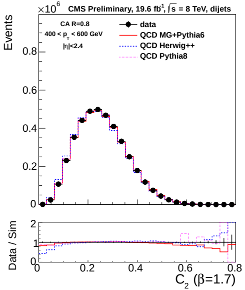
Simulation
$p_T = 250 - 350$ GeV
(W+jet) - no pruned mass cut



Simulation
$p_T = 250 - 350$ GeV
(W+jet) - pruned mass cut



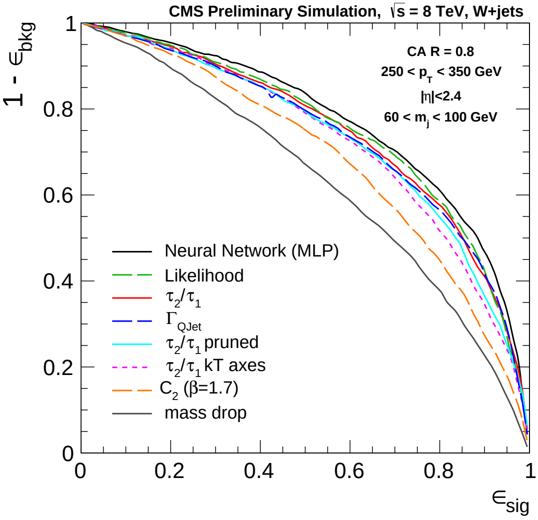
Data/MC
$p_T = 400 - 600$ GeV
(dijet) - no pruned mass cut



Mass cut has small impact on the discriminating power of this variable.

Optimization

$60 < m_{\mathrm{pruned}} < 100$ GeV



Mistag vs efficiency

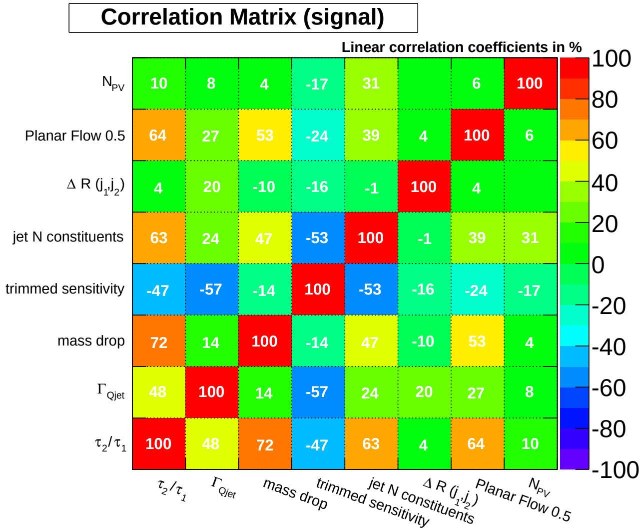
- background: QCD, signal: $H \to WW$
- pruned jet mass cut $60 < m < 100$ GeV
- τ_2/τ_1 best single variable discriminator
- neural network trained with TMVA shows improvement over τ_2/τ_1
- other variables correlated with τ_2/τ_1

MVA variables

- Pruned mass drop m_{pr}^{drop}
- Q-jet volatility Γ_{Qjet}
- N-subjettines τ_2/τ_1
- Planar flow R $= 0.5$
- Number of jet constituents
- Subjet ΔR
- Trimmed grooming sensitivity
- Number of primary vertices N_{PV}

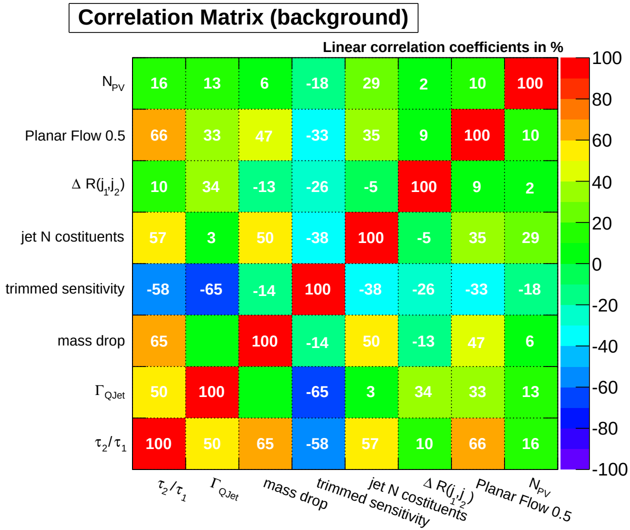
Correlation matrix (signal)

Correlation matrix for the input variables given to the MVA evaluated on signal sample ($gg \to H$ at $m_H = 600$ GeV)



Correlation matrix (background)

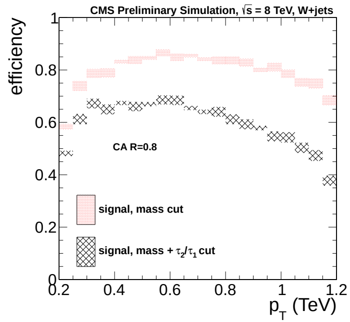
Correlation matrix for the input variables given to the MVA evaluated on background (W+jets Pythia $p_T > 180$ GeV)



Performance in function of p_T

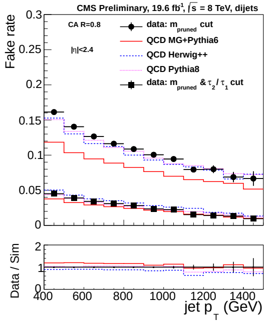
Performance studied for: $60 < m_{jet} < 100$ GeV + $\tau_2/\tau_1 < 0.5$

Efficiency vs p_T (W+jets topology)



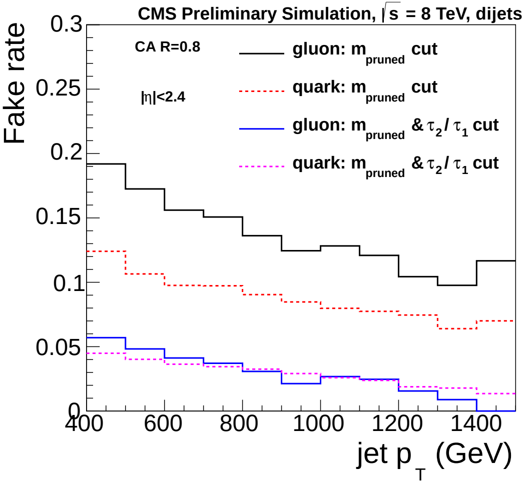
- low p_T: W decay products begin to be reconstructed inside CA8 jets
- high p_T: detector resolution for jet substructures degrades, pruning remove too much of the mass of the W

Fake rate vs p_T (dijet topology)



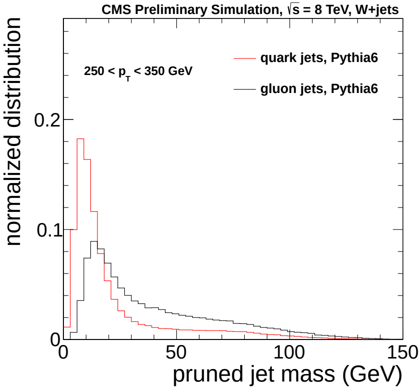
- drops at high p_T similarly to efficiency

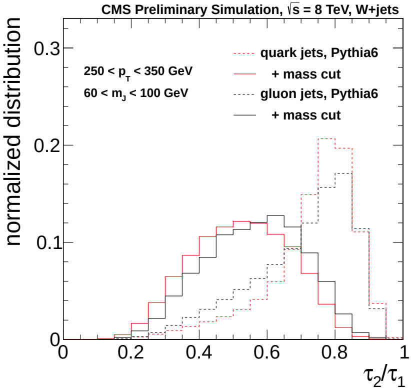
Quark vs gluons - fake rate



- gluon jets have on average higher mass → fake rate higher with mass only cuts
- quark jets tend to have lower τ_2/τ_1 → cut fake rate is similar

Quark vs gluons



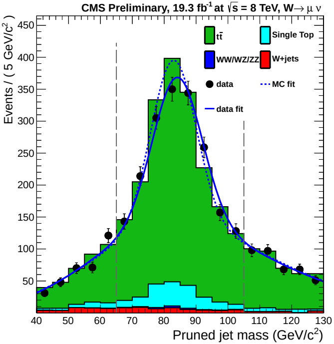


- gluon jets tend to have larger mass
- before mass cut gluon jets appear more W-jets-like with respect to τ_2/τ_1
- after mass cut quark jets appear more W-jets-like with respect to τ_2/τ_1

Jet mass scale and resolution

Performance studied for the following working point: $60 < m_{jet} < 100$ GeV + $\tau_2/\tau_1 < 0.5$

Muon channel



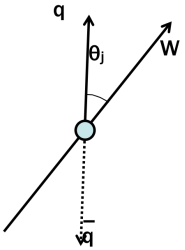
Extract:

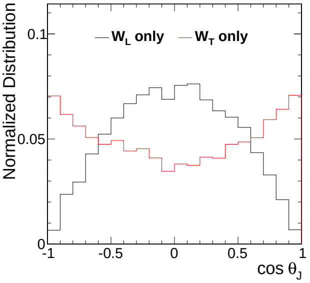
- Data/MC correction for W-tagging efficiency
- W-jet mass scale
- W-jet mass resolution

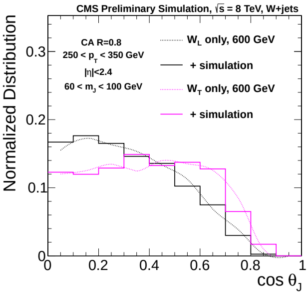
	data	MC	scale factor / shift
efficiency $200<p_T<265$ GeV			* 0.96 +- 0.08
efficiency $265<p_T<600$ GeV			* 0.89 +- 0.10
mass peak position	84.5 +- 0.4 GeV	83.4 +- 0.4 GeV	+1.1 +- 0.4 GeV
mass peak width	8.7 +- 0.6 GeV	7.5 +- 0.4 GeV	+16% +- 9%

Polarization studies

- Polarization can affect substructure distribution
- Sample used: scalar $X \to W^L_{lept} W^L_{had}$ and $X \to W^T_{lept} W^T_{had}$

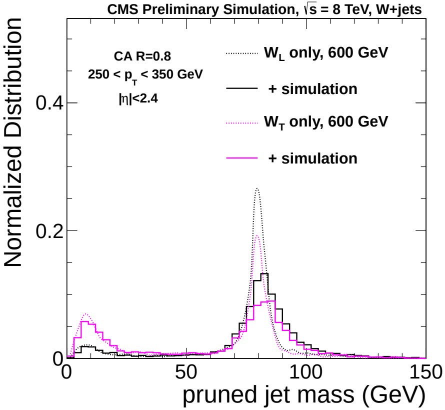


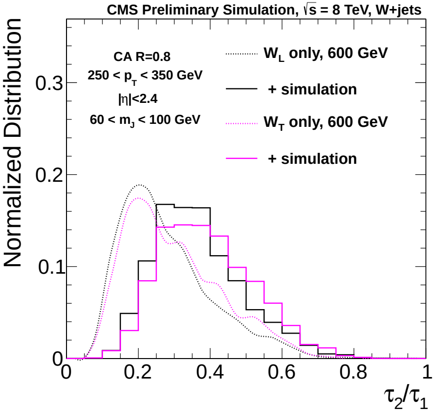




- parton level helicity angle for hadronic W
- observable helicity angle from subjets

Polarization studies - τ_2/τ_1



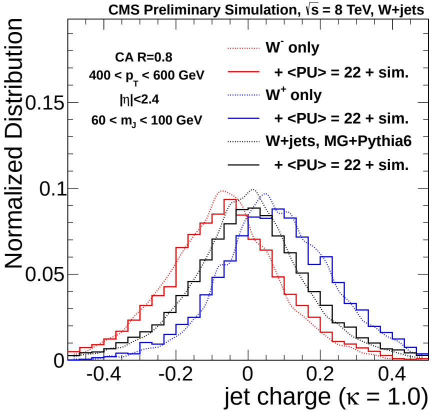


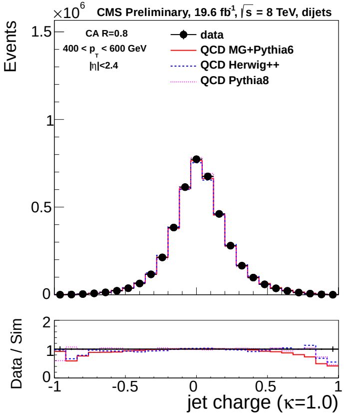
- pruned jet mass acceptance different for W_L and W_T
- ΔR between partons smaller on average for W_L
- W_L more likely to be accepted by CA8 jet
- in W_T topology p_T of the subjets is more asymmetric, thus more QCD-like

Jet charge, Q^κ

$$Q^\kappa = \frac{\sum_i q_i (p_{Ti})^\kappa}{(\sum_i p_{Ti})^\kappa}$$

Used to discriminate between W+ and W-



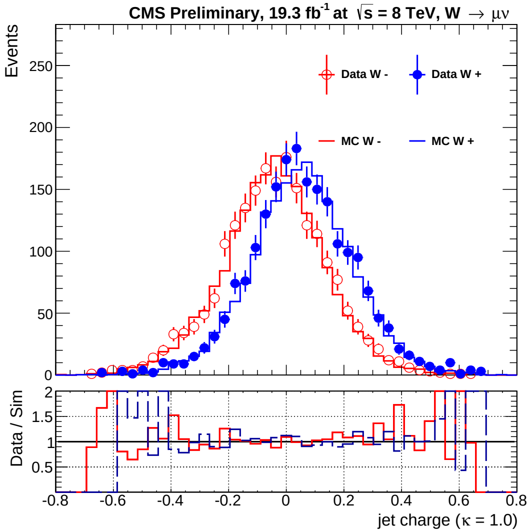


Right plot, note: $<$**jet charge**$>\neq$ **0**

Jet charge distribution

$t\bar{t}$ sample for W^+ and W^- jets in simulation and data.

Simulated distributions are a sum of all processes.



$t\bar{t}$ semileptonic selection
By selecting on the lepton charge, we can isolate W^+ from W^- jets.

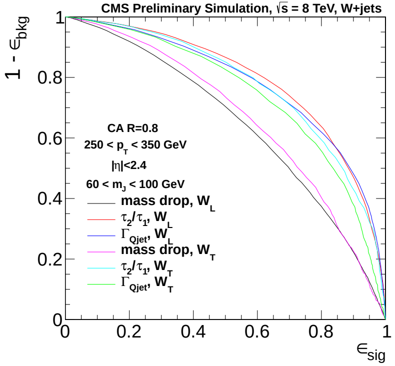
W-tagging: conclusions

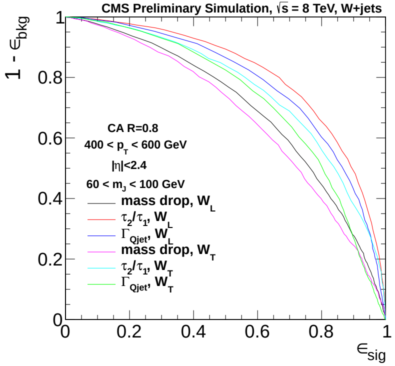
- W-tagging studied in association with several substructure variables
- Performance assessed for these variables (ROC curves)
- Behavior studied in various scenarios (polarization, quark and gluons, etc.)
- Comparison with 8TeV data for different showering models
- Jet charge variable has encouraging discrimination power for W-jet charge in Data

Thank you for the attention!

Backup

Optimization

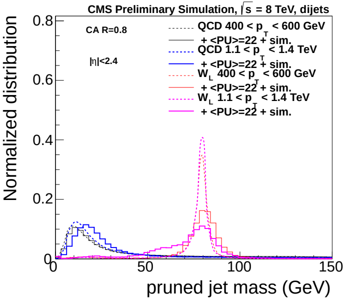


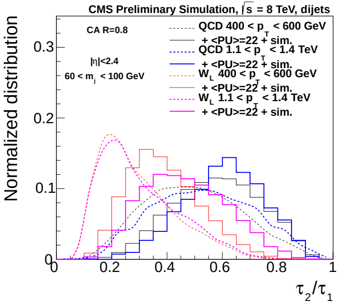


Mistag vs efficiency

- at high p_T discrimination power is better for W_L
- W_T are more QCD-like at more boosted regimes
- τ_2/τ_1 works the best in both cases

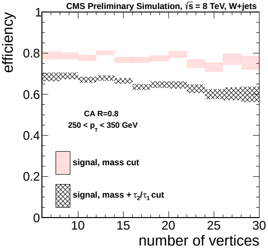
High p_T behavior

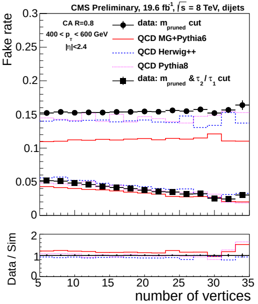




- At generator level, pruning still performs very well at high p_T
- Degradation of detector level substructure at high p_T
- pruning rejects too many particles in the W jet
- pruned W jet mass peaks between 40 and 60 GeV
- τ_2/τ_1 discrimination power is also reduced at high pt

Performance in function of number of vertices





Efficiency vs Nvtx (W+jets topology)

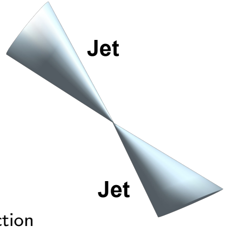
- slight degrade of performance
- jet pruning fails to remove all soft contributions

Fake rate vs Nvtx (dijet topology)

- constant behavior with respect to Nvtx

Event topologies considered

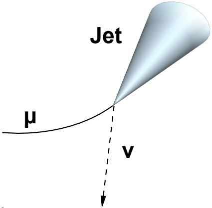
Dijet



Selection

- dijet mass > 890 GeV
- $|\eta| < 2.4$, $|\eta_1 - \eta_2| < 1.3$
- $400 < p_T(jet) < 600$ GeV (msignal $= 1$ TeV)

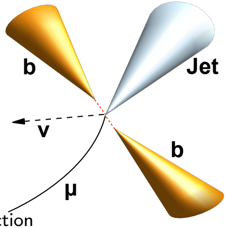
W+jets



Selection

- Lepton pT $> 50/90$ GeV, $|\eta| < 2.1/2.4$
- Jet $|\eta| < 2.4$
- $\Delta R_{l,j} > \pi/2$, $\Delta\phi_{V,j} > 2$, $\Delta\phi_{MET,j} > 2$
- MET $> 50/70$ GeV
- Leptonic W pT > 200 GeV
- anti-CSVM-btag

$t\bar{t}$



Selection

- Similar to W+jet case
- $\geq$ AK5 b-jet
- choose the highest mass CA8 opposite to the lepton
- for mu:$pt > 50 GeV$, $ET > 50 GeV$
- for electrons: $pt > 90 GeV$, $ET > 80 GeV$

Signal $X \rightarrow WW$, $M_X = 600\ GeV/c^2$, $1 TeV/c^2$

Top tagging in CMS

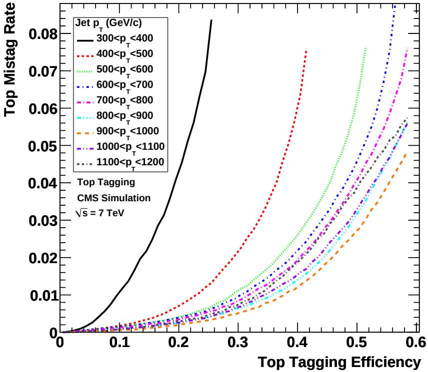
CMS Top tagger, based on algorithm by Kaplan et al.

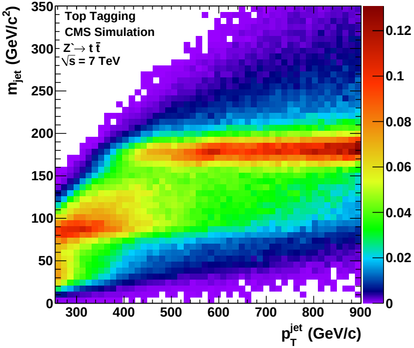
- Cambridge-Aachen R = 0.8 ($R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$) jets (hard-jets) are used as input
- the primary decomposition: attempts to split the hard jet into two subjets by reversing the pairwise clustering sequence
- continue to the next step if the two subclusters satisfy $R^2 > 0.4 - 0.0004 \times p_T^{\text{orig. subcl.}}$
- if the two subclusters satisfy the momentum fraction criterion $p_T^{\text{cluster}} > 0.05 \times p_T^{\text{hardjet}}$, then the decomposition succeeds ("jet grooming")
- if only one subcluster satisfy the momentum fraction criterion the decomposition is repeated on the passed cluster.
- secondary decomposition: repeat the decomposition on the subclusters passing the primary decomposition.

Additional cuts:

- $140 < m_{\text{jet}} < 250$ GeV/c^2
- $N_{\text{subjets}} \geq 3$
- Minimum pairwise mass, $m_{\text{min}} > 50$ GeV/c^2

CMSTopTagger performance in 7TeV data





CMSTopTagger additional plots