OLOMOUC, CZECH REPUBLIC
9th International Workshop on Top Quark Physics
19 - 23 September 2016

Boosted top: new algorithms and perspectives

J. Caudron, Universität Bonn (DE)
on behalf of ATLAS and CMS collaborations

Introduction

This presentation focuses on top-tagging algorithms:

physics analyses results are in Konstantinos Kousouris' presentation

There are many developments and new techniques.

This presentation focuses on techniques that are likely to be used in ATLAS/CMS data analyses in near future.

Non exhaustive list of other interesting techniques is given at the end.

Outline:

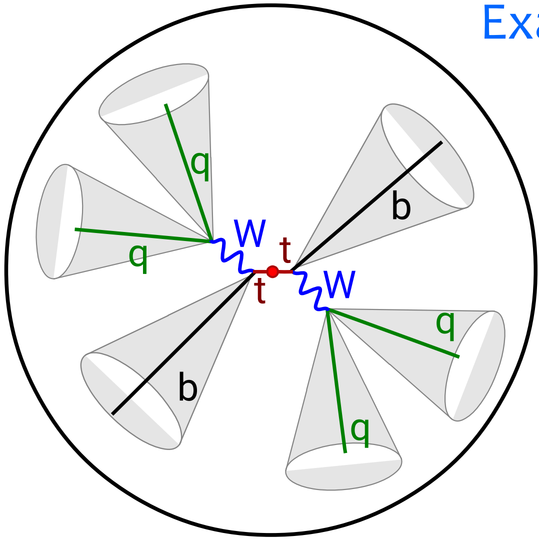
- Boosted top introduction
- Current algorithms, used in ATLAS / CMS analyses:
  - Substructure variables
  - Top-tagging algorithms: HEPTopTagger, CMSTopTagger, Shower Deconstruction
  - Comparison of the algorithms performances
- New algorithms:
  - Variable-R jets
  - HOTVR
  - PUPPI
  - Track-assisted mass

Boosted top introduction

motivation, challenges and basic concepts

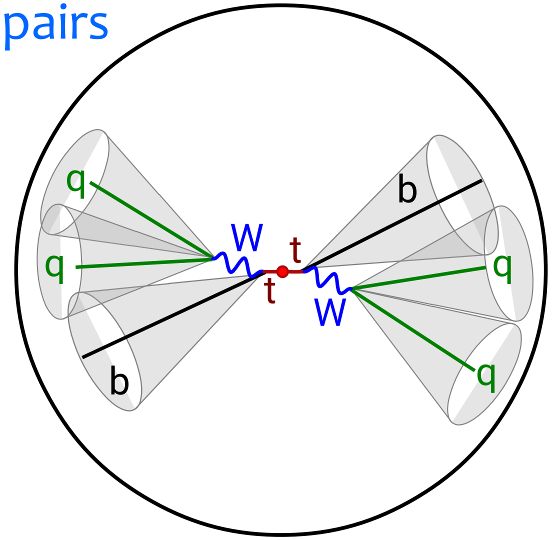
Boosted Regime

Example with top-antitop pairs



Resolved

All the jets are reconstructed



Boosted

The showers of different decay products overlap and cannot be reconstructed as individual jets

When this happens:

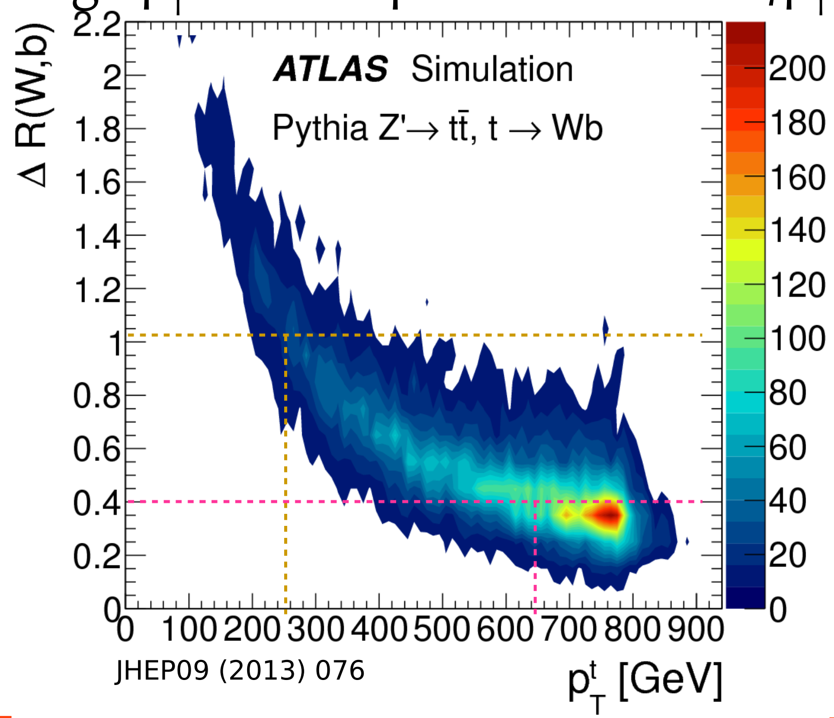
High mass particle decay:
direct searches for new physics (Z', VLQ, SUSY, ...)

Observation in specific phase-space:
new physics in precision measurements
(differential cross-section, charge asymmetry, ...)

Advantages:

- Better reconstruction/acceptance for the phase-space of interest
- Complementary with resolved regime
- Better S/B region (less QCD jets)

High p_T of initial particle: $\Delta R \approx 2m/p_T$



JHEP09 (2013) 076

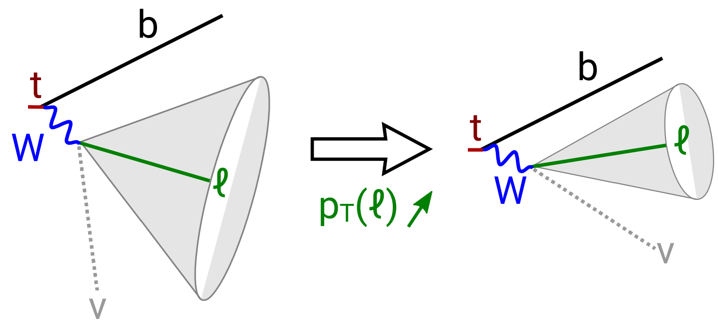
$\Delta R = \sqrt{\Delta\eta^2 + \Delta\Phi^2}$

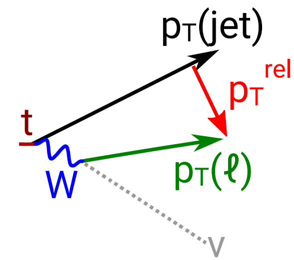
Boosted Selection

Boosted regime needs a specific selection:

- Leptonic top decay:

  ATLAS: isolation with p_T-dependant-cone CMS: cut on $\Delta R(\ell,\text{jet})$ and $p_T^{rel}(\ell,\text{jet})$





- Hadronic top decay:

  Use Large-Radius jet to capture all the top quark decay products

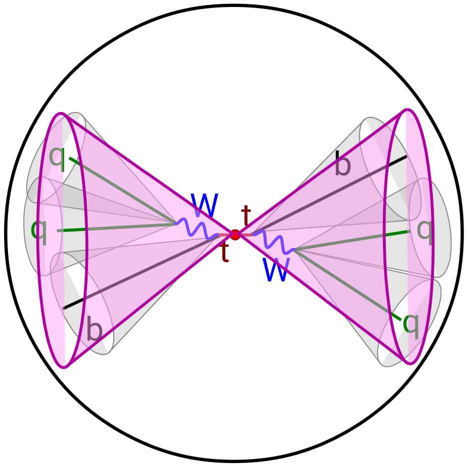
  anti-kt / CA, with large radius: 0.8, 1.0, 1.5 (ref. in backup slide)

  ATLAS: calorimeter cell clusters inputs,
  CMS: particle flow inputs

  Substructure of the large-R jet is used for top-tagging
  (using substructure variables or more complex algorithm)

  Cases using small-R jets (dense environment, low boost):
  - Re-clustering [arXiv:1606.03903] (ATLAS)
  - Resolved tagger [CMS-EXO-16-005] (CMS)

  ...



Large-R Jets

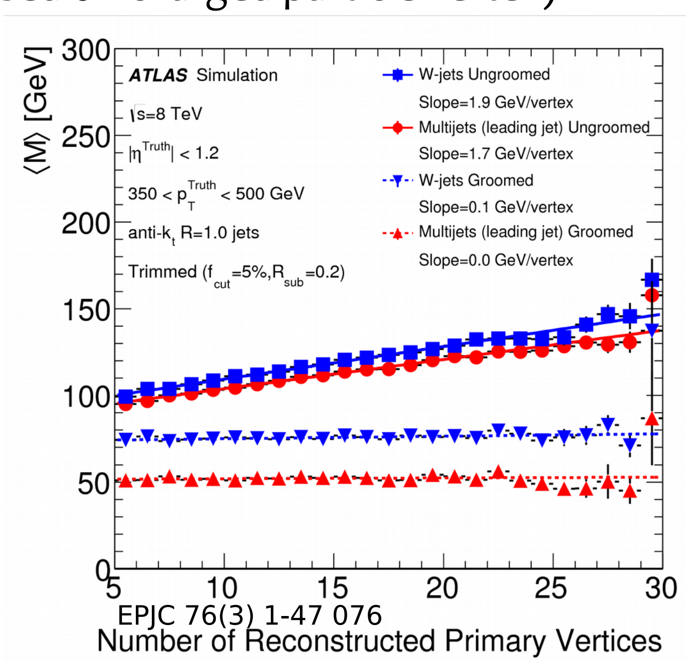
If R is large: all top decay products are contained corollary: higher contamination from Pile-Up

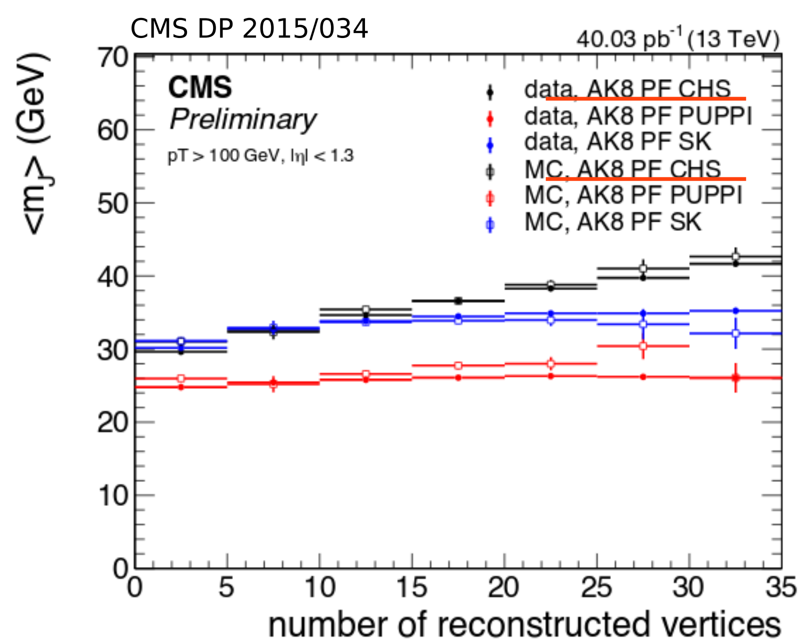
Grooming and Pile-Up:

- **Filtering:** recluster into subjets and keep the N hardest subjets [Phys. Rev. Lett. 100 (2008) 242001]
- **Trimming:** recluster into subjets and keep the subjets with $p_T / p_T^{jet} > x$ [JHEP02 (2010) 084]
- **Pruning:** ignore wide angle soft constituents during the clustering [Phys. Rev. D81 (2010) 094023]
- **Soft-drop:** decluster and remove wide angle soft constituents [JHEP05 (2014) 146]

ATLAS: Contamination from Pile-Up is reduced by grooming

CMS: Charged Hadron Subtraction [EPS-HEP2013 (2013) 433] is used (evaluation of the PU contamination based on charged particle vertex)





Current algorithms

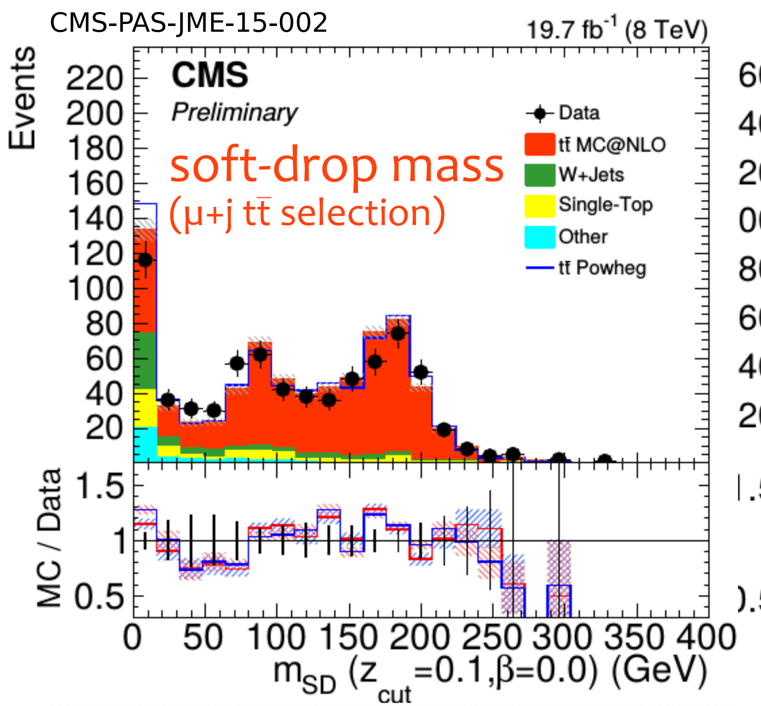
developed for few years in ATLAS/CMS

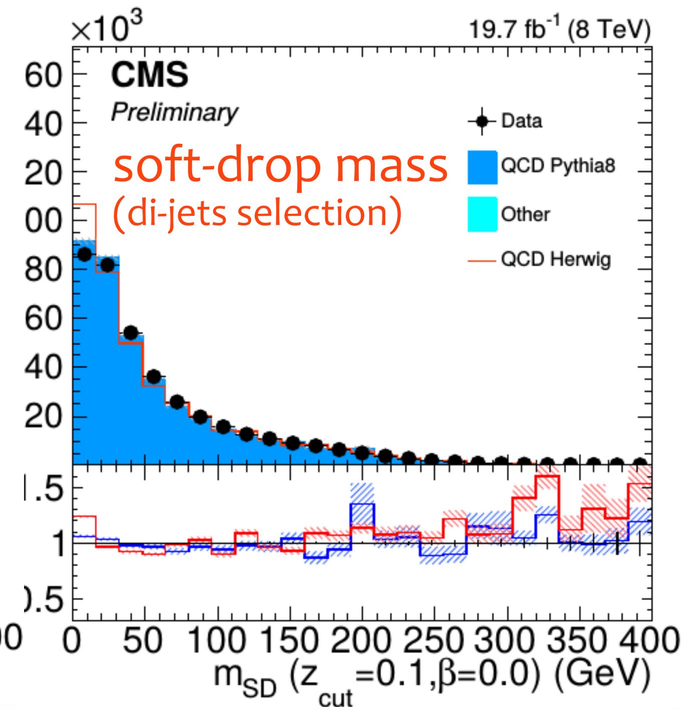
Substructure Variables

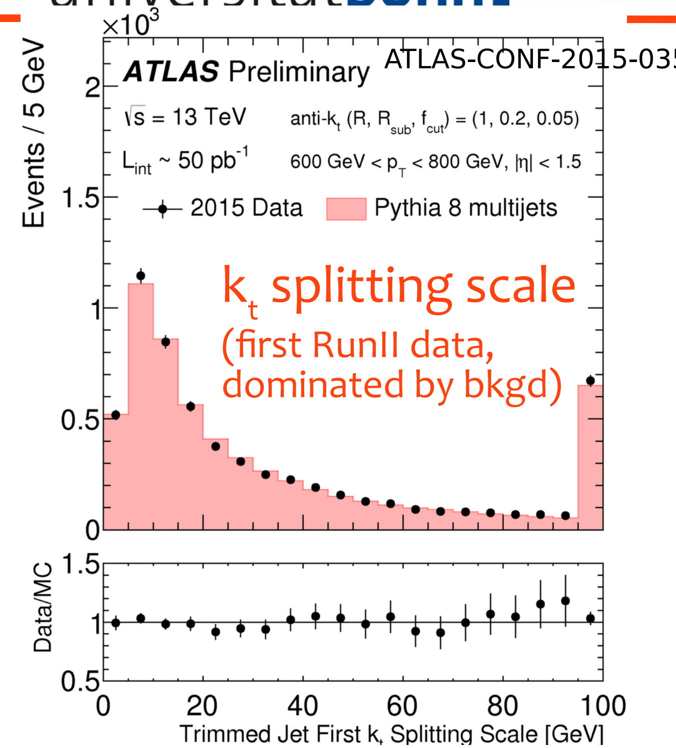
more details in backup slides

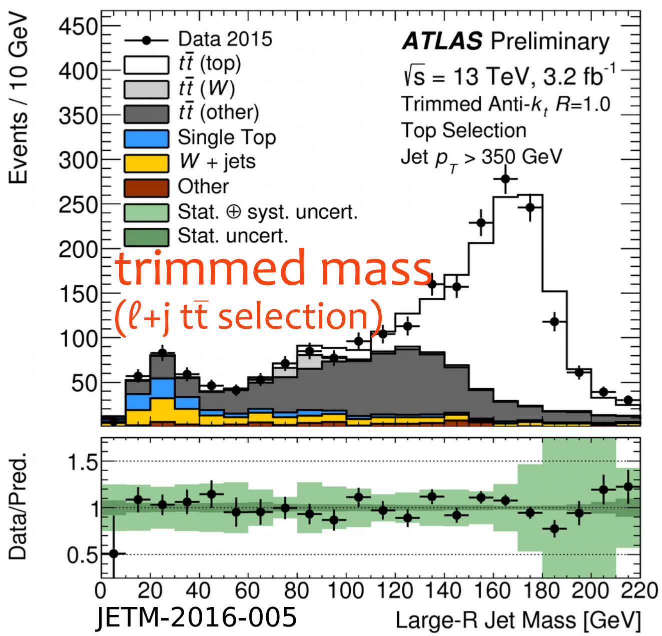
- trimmed mass [JHEP02 (2010) 084] or soft-drop mass [JHEP05 (2014) 146]
  Large-R mass after grooming
  Peak at the top mass
- k_t splitting scale [JHEP07 (2008) 092]
  Value of the jet-jet distance in the last step of the k_t algorithm clustering
  High for top-jet
- N-subjettiness ratio [JHEP03 (2011) 015]
  From N-subjettiness τ_N (formula in backup slide), $\tau_{32} = \tau_3/\tau_2$
  The lower, the more 3-prong-like
- QJet volatility [Phys. Rev. Lett. 108 182003 (2012)]
  Jet mass stability when some randomness is added to the pair-clustering decision
  Low value for top-jet
- subjet b-tagging
  Specific dense-environment resistant b-tagging applied
  CMS: standard combined b-tagging (tracks + secondary vertex) applied to subjets of a groomed large-R jet [CMS PAS JME-13-007]
  ATLAS: Multi-Variate b-tagger [ATL-PHYS-PUB-2014-014] using dense-environment-robust variables (variables associated to the primary vertex)

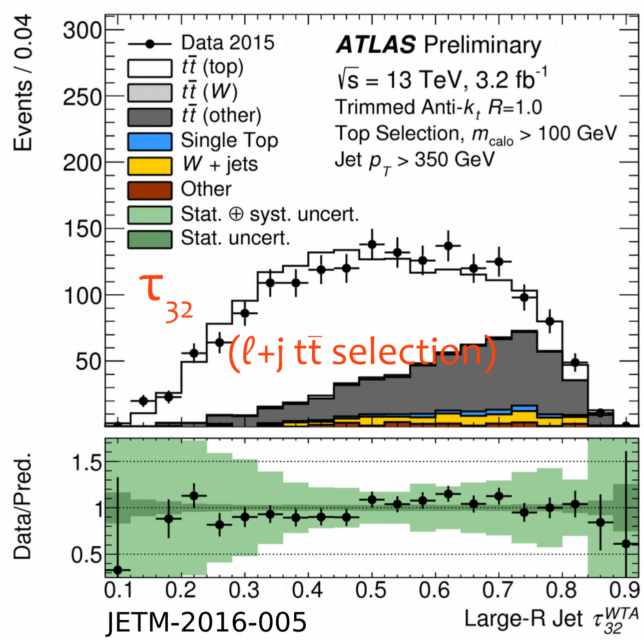
Substructure Variables

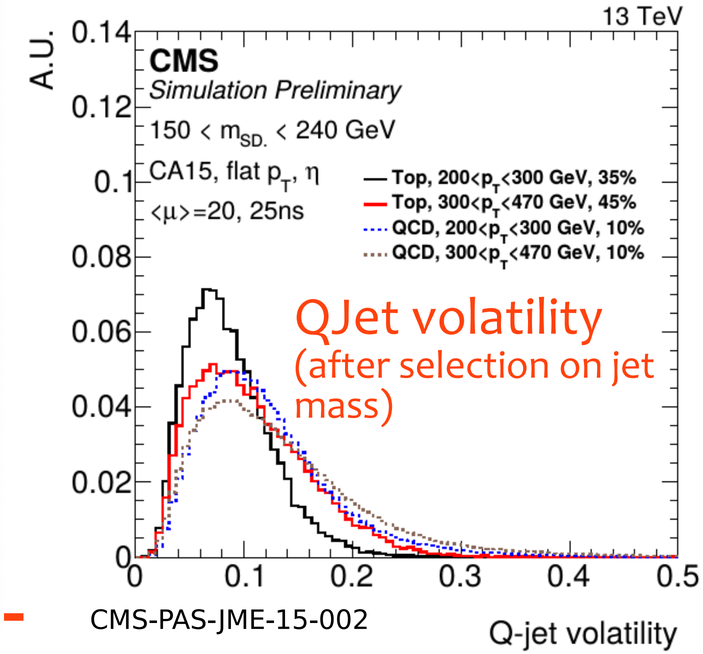












Top-tagging Algorithms

universitätbonn erc

- Substructure top-taggers

  Cuts applied on one or several substructure variables

more details in backup slides

- HEPTopTagger [JHEP 1010 (2010) 078]

  From 1.5 CA jet, decluster to hardjets and test the triplet possibilities (+filtering and reclustering) to find the three subjets corresponding to b, q and q'

  -> m_{jets}

  HEPTopTagger04 [JHEP06 (2016) 093]: use small-R 0.4 as input (helpful in dense environment)

  HEPTopTaggerV2 [arXiv:1503.05921]: minimum effective R vs expected R curve -> $R_{opt}-R_{opt}^{calc}$

- CMS Top-Tagger [CMS-PAS-JME-13-007]

  From 0.8 CA jet, decompose the jet (reject soft components, stop if subjets are too adjacent), decompose the subjets

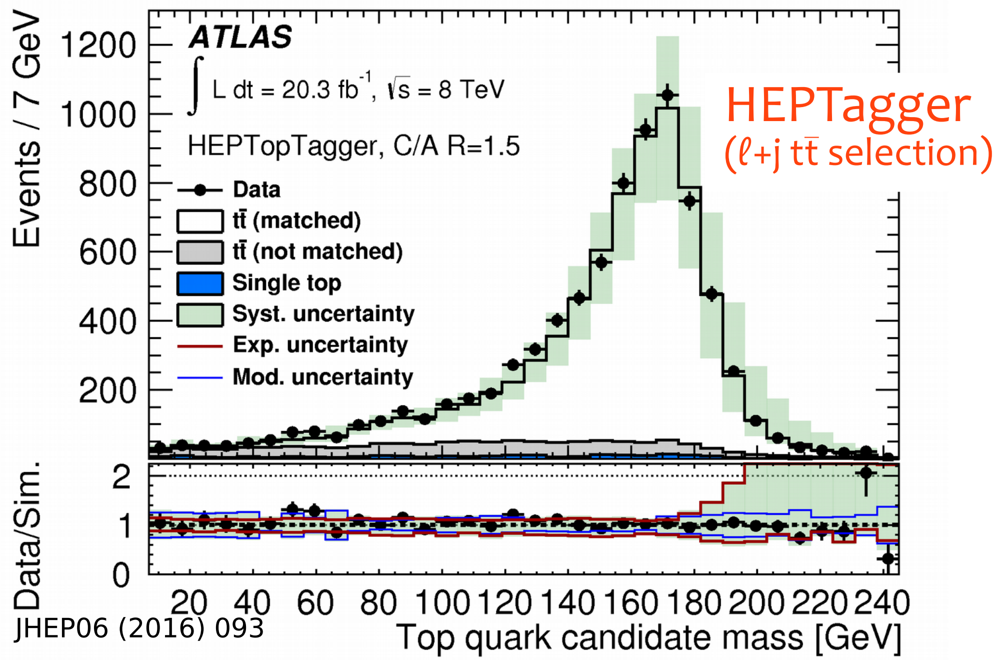
  -> $N_{subjets}$, $\min(m_{12}, m_{13}, m_{23})$, m_{jets}

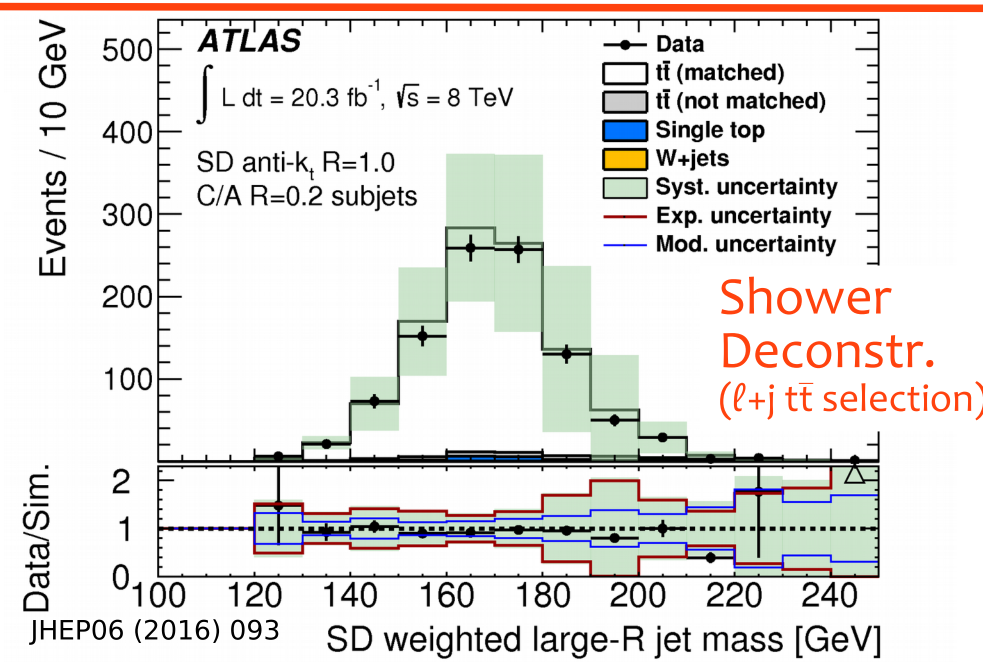
- Shower Deconstruction [Phys. Rev. D 87 054012]

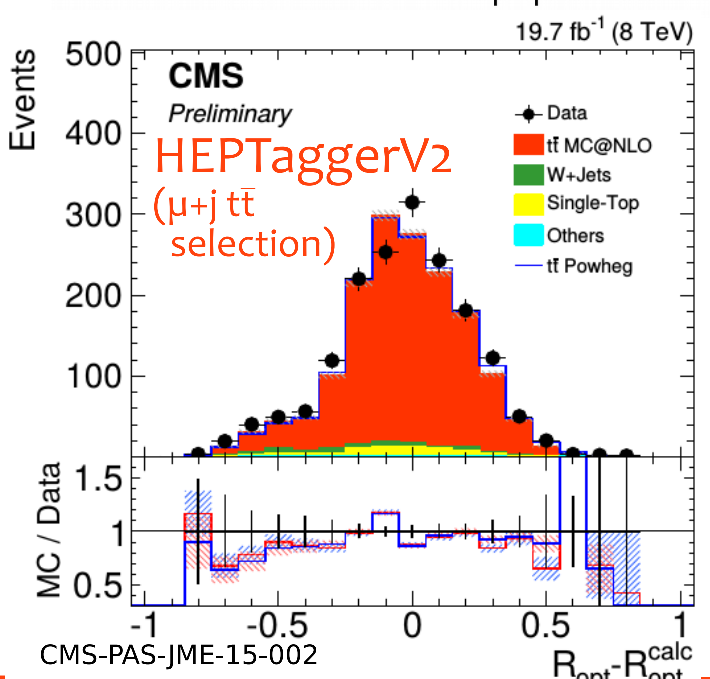
  Large-R jet (akt08, akt10, CA15) components are reclustered in micro-jets.
  Based on computed possible shower history from signal and background, the probability of the micro-jets configuration is obtained for each shower history

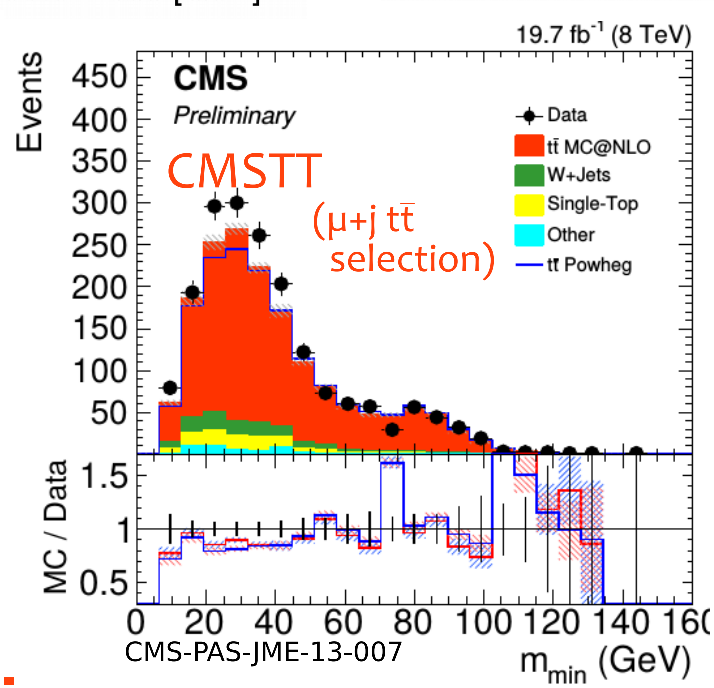
  -> χ_{SD}= P(signal)/P(bkgd)

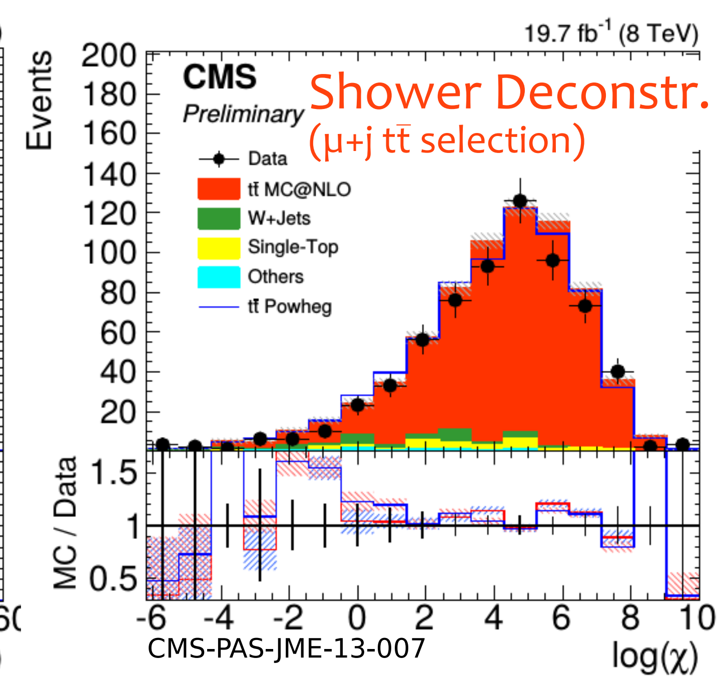
Top-tagging Algorithms



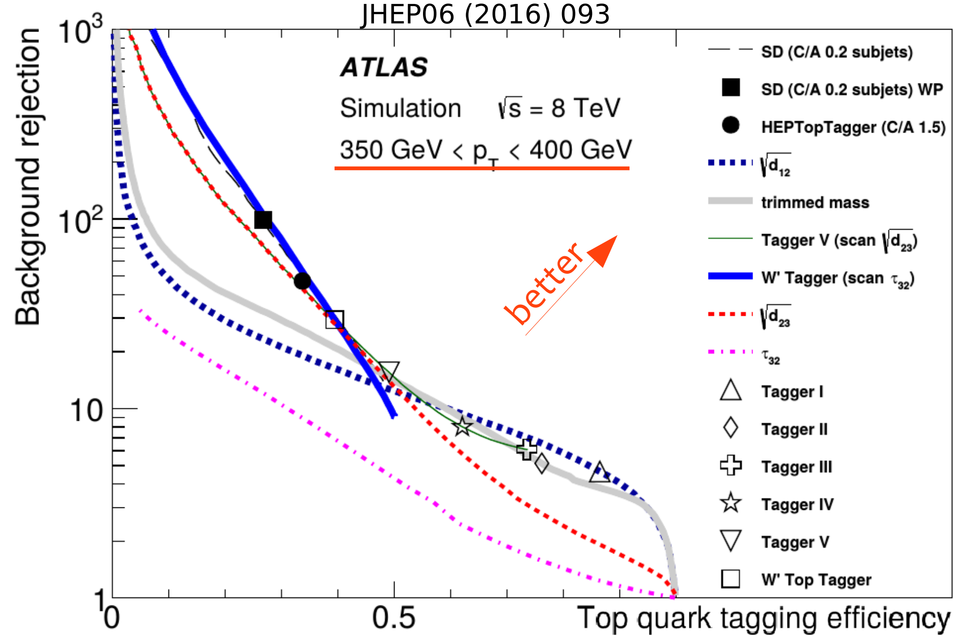


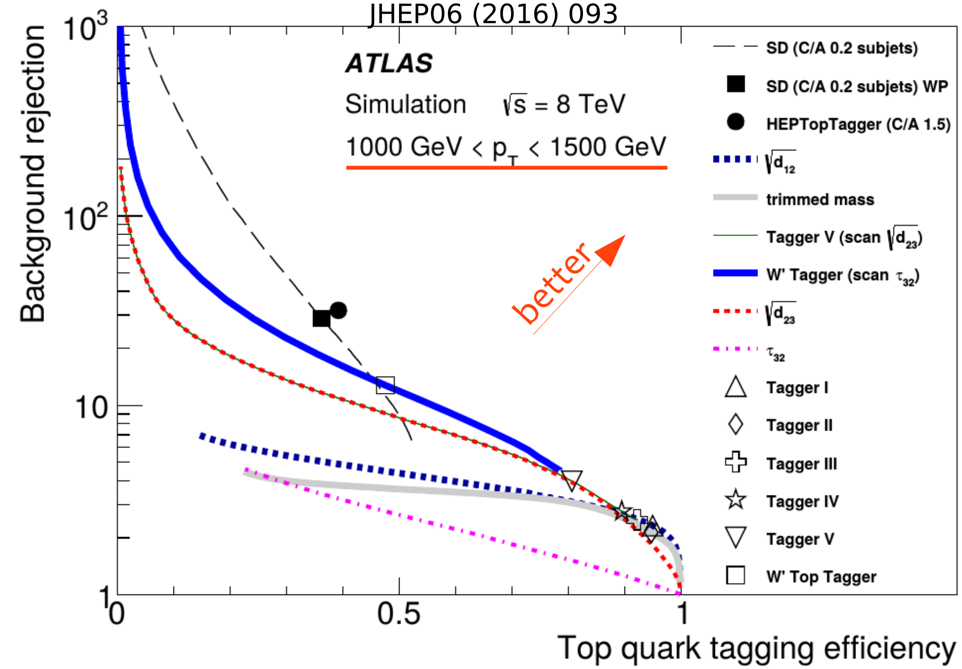


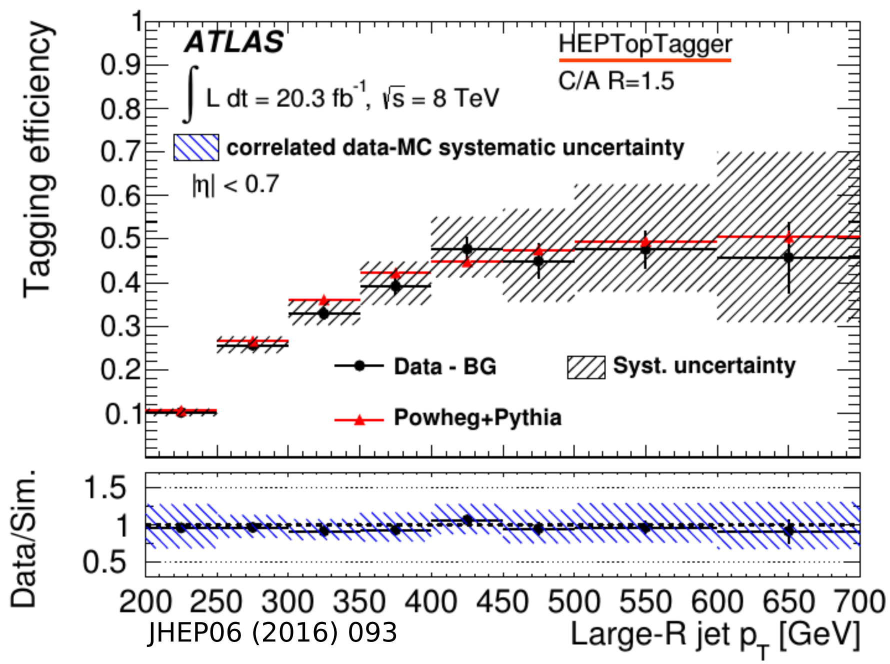


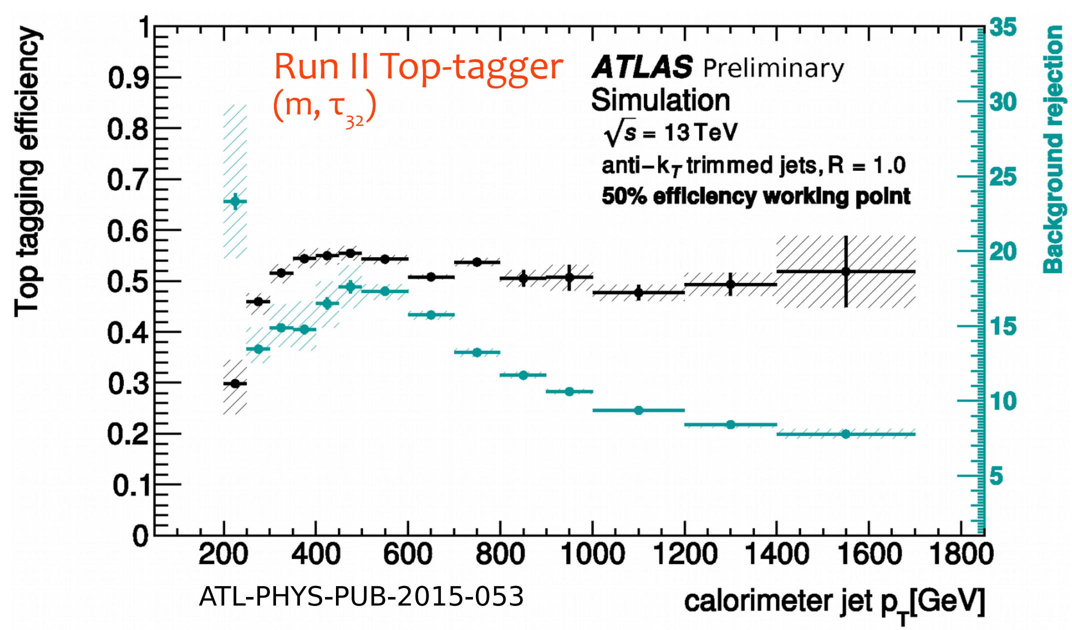


Performances in ATLAS

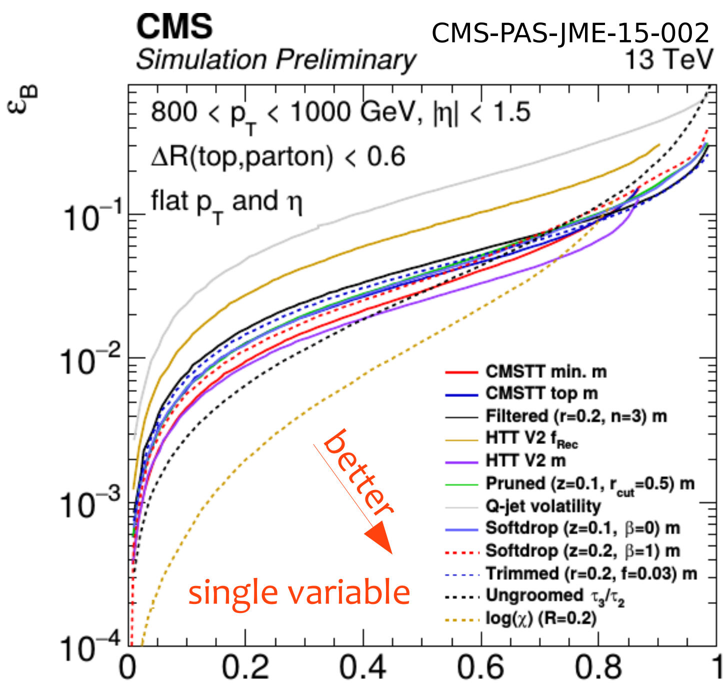


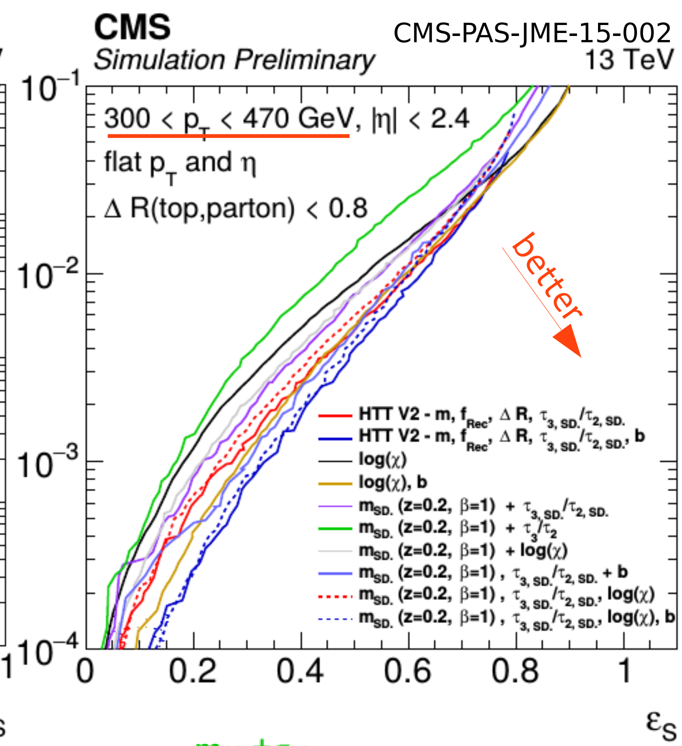


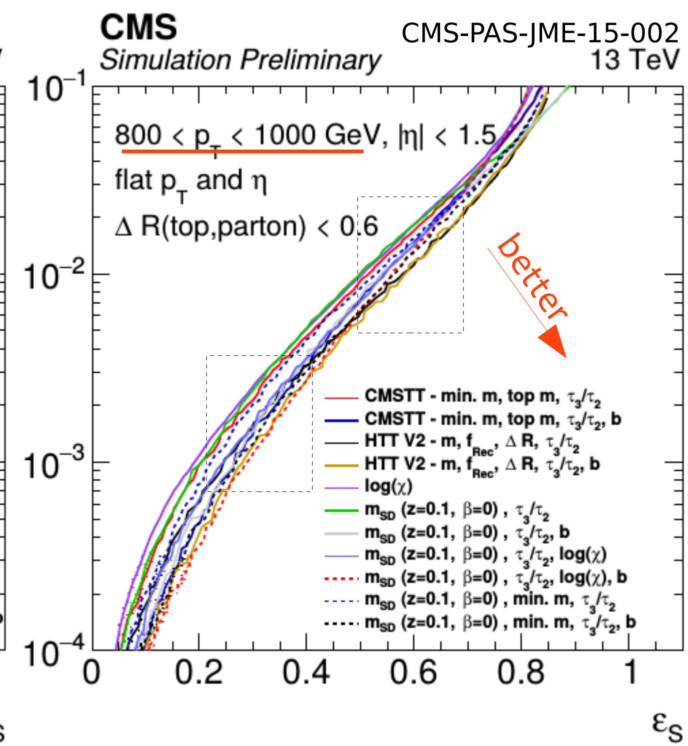


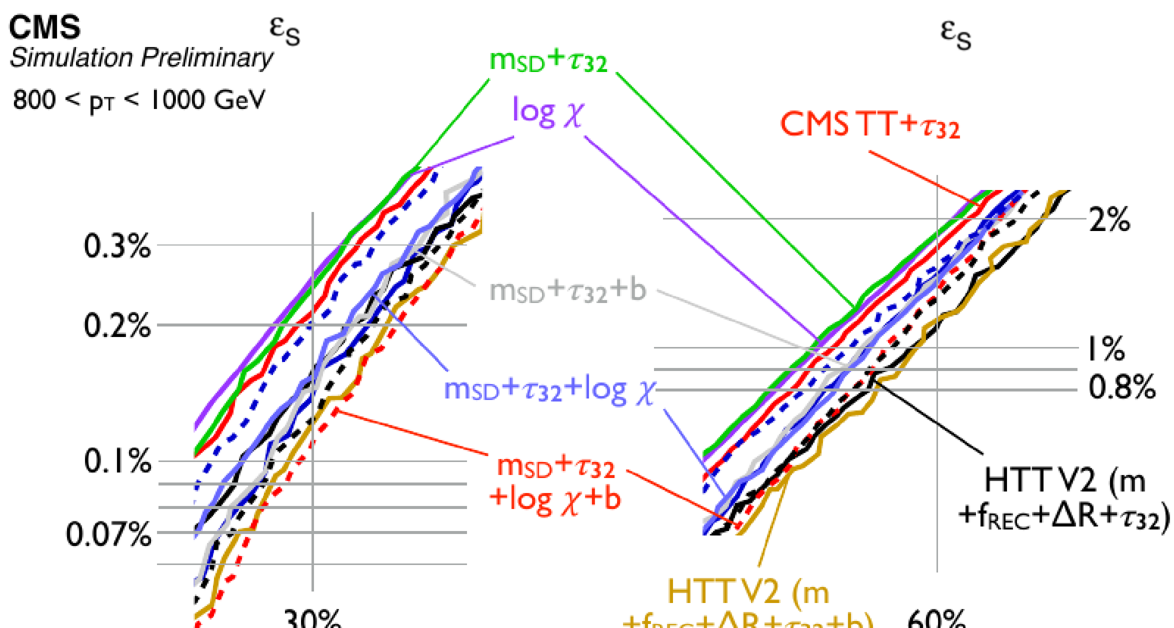


Performances in CMS









New algorithms

in advanced development in ATLAS/CMS
and new ideas

Variable-R jets in ATLAS

ATL-PHYS-PUB-2016-013

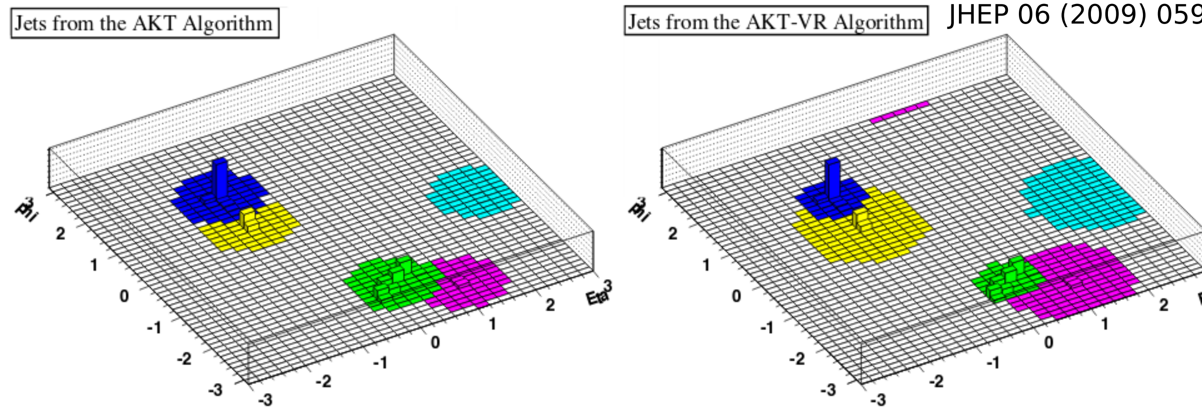
New jet clustering algorithm tested in ATLAS:

Large-R radius shrinks with p_T:

$$d_{ij} = \min(p_{Ti}^{2n}, p_{Tj}^{2n})\Delta R_{ij}^2$$

$$d_{iB} = p_{Ti}^{2n} R_{\text{eff}}^2(p_{Ti})$$

$$R_{\text{eff}}(p_T) = \begin{cases} R_{\min} & \text{if } \frac{\rho}{p_T} \leq R_{\min} \\ \frac{\rho}{p_T} & \text{if } R_{\min} < \frac{\rho}{p_T} < R_{\max} \\ R_{\max} & \text{if } \frac{\rho}{p_T} \geq R_{\max} \end{cases}$$



JHEP 06 (2009) 059

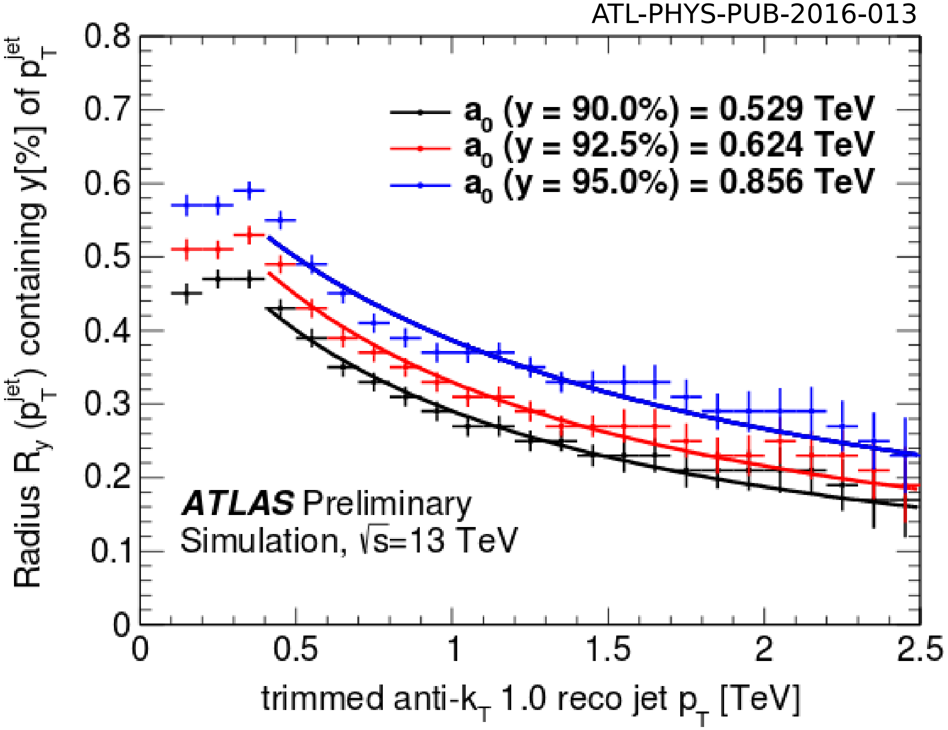
It reduces contamination from PU/UE and ISR

R_{max} = 1.0, R_{min} = 0.2, ρ = 600GeV

Compared with substructure top-tagger using $\sqrt{d_{12}}$, τ_{32} and top mass (cf. next slide)

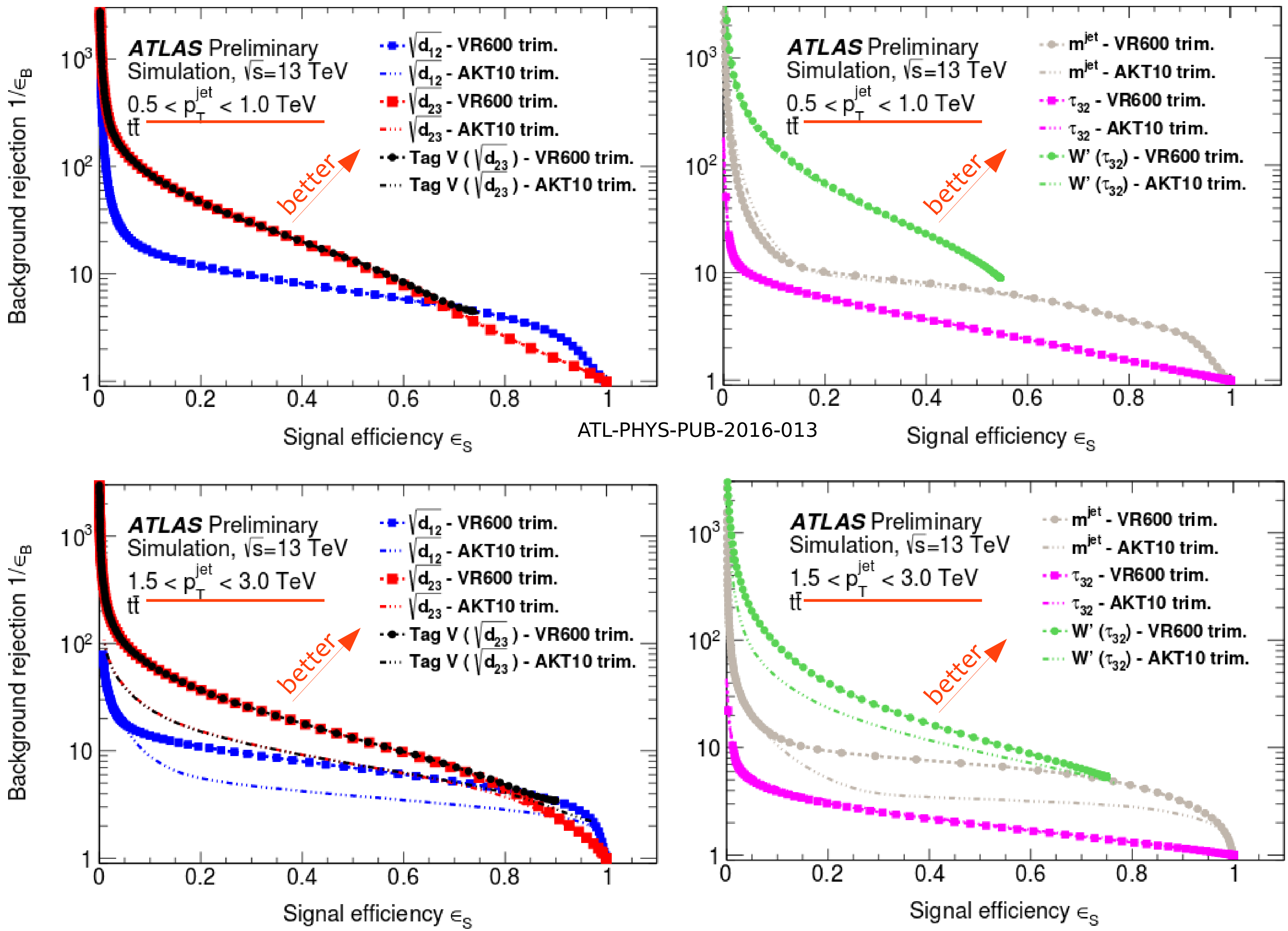
Similar performance at low p_T
Better performance at high p_T

ATL-PHYS-PUB-2016-013



Variable-R jets in ATLAS

ATL-PHYS-PUB-2016-013



HOTVR algorithm

arXiv:1606.04961

Heavy Object Tagger with Variable Radius

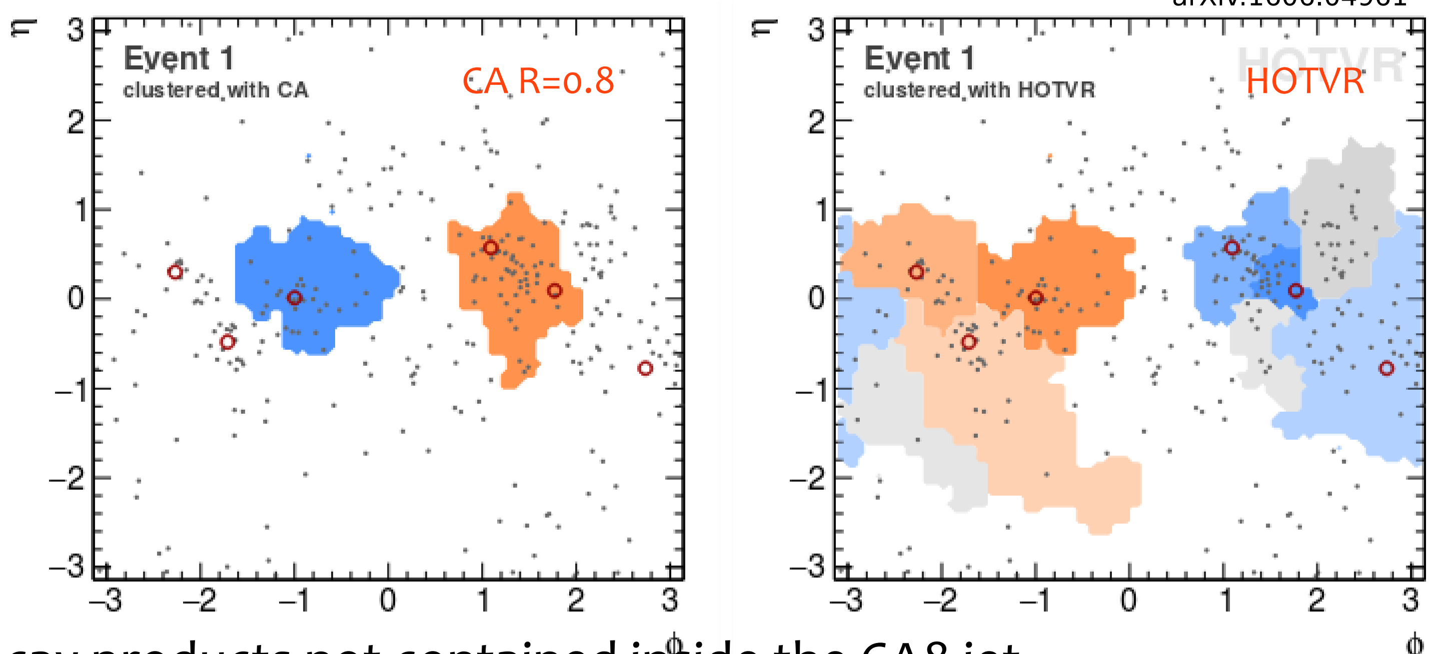
Top tagger with low complexity and good performance for a large p_T range.

New clustering algorithm that includes:

- Variable jet radius: using the same technique as in the ATLAS Variable R
- Rejection of soft components: mass drop condition in the clustering algorithm
- Subjets identification: the list of relevant subjets is built during the clustering algorithm

Low p_T ($p_T \sim 200$ GeV) case example:

arXiv:1606.04961



Top decay products not contained inside the CA8 jet
Subjets identified (colored areas) + soft components rejected (grey areas) in HOTVR

HOTVR algorithm

arXiv:1606.04961

universitätbonn

Heavy Object Tagger with Variable Radius

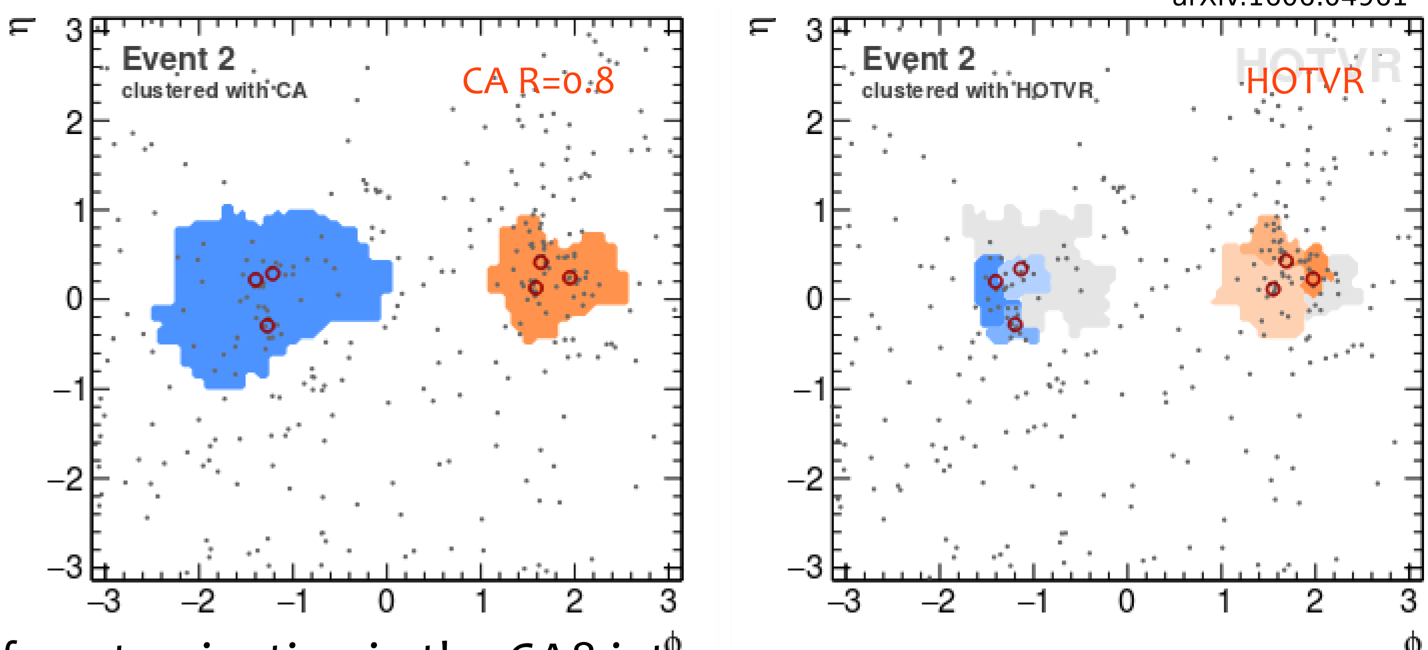
Top tagger with low complexity and good performance for a large p_T range.

New clustering algorithm that includes:

- Variable jet radius: using the same technique as in the ATLAS Variable R
- Rejection of soft components: mass drop condition in the clustering algorithm
- Subjets identification: the list of relevant subjets is built during the clustering algorithm

High p_T (p_T ~ 800 GeV) case example:

arXiv:1606.04961

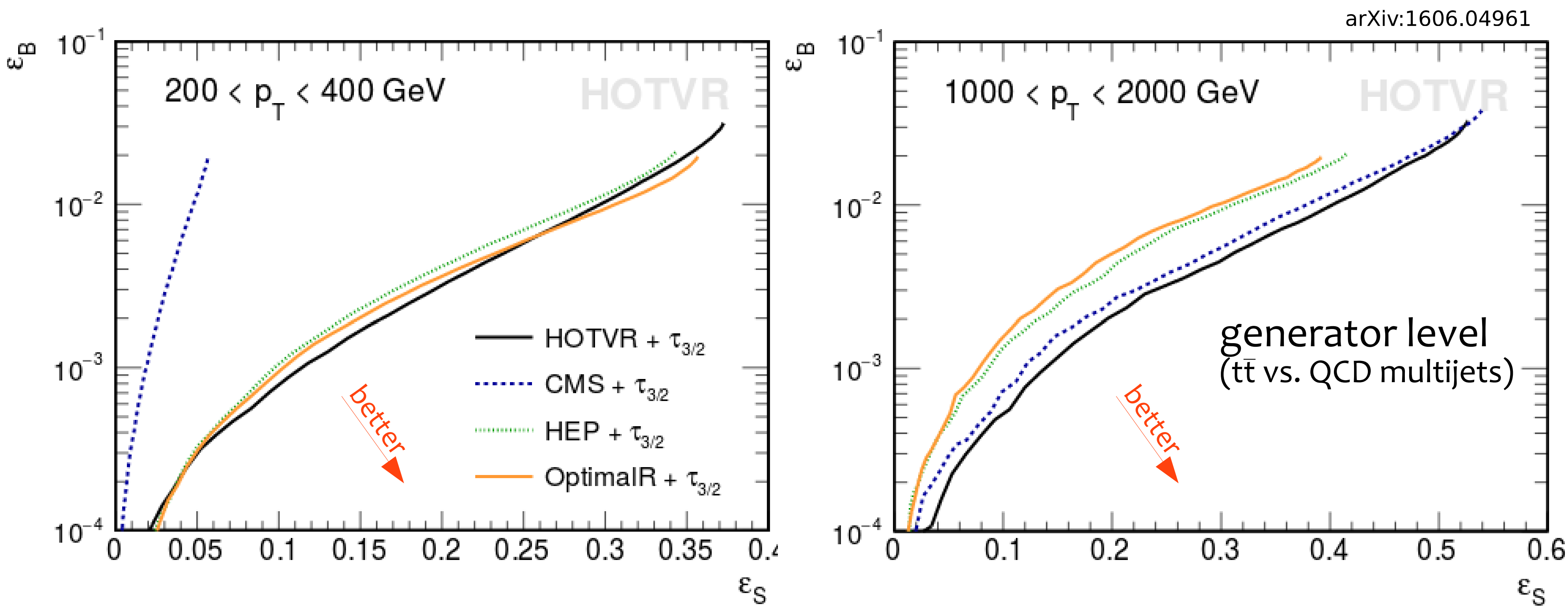


A lot of contamination in the CA8 jet

Subjets identified (colored areas) + soft components rejected (grey areas) in HOTVR

HOTVR algorithm

arXiv:1606.04961



- Competitive top-tagger
- Flat signal efficiency & background fake-rate, even at high p_T
- Low complexity -> low computational power needed

Code added into FastJet/contribs
Developed by CMS members
The concept can be extended to W/Z/H tagging

PUPPI in CMS

JHEP10 (2014) 059,
CMS DP 2015/034

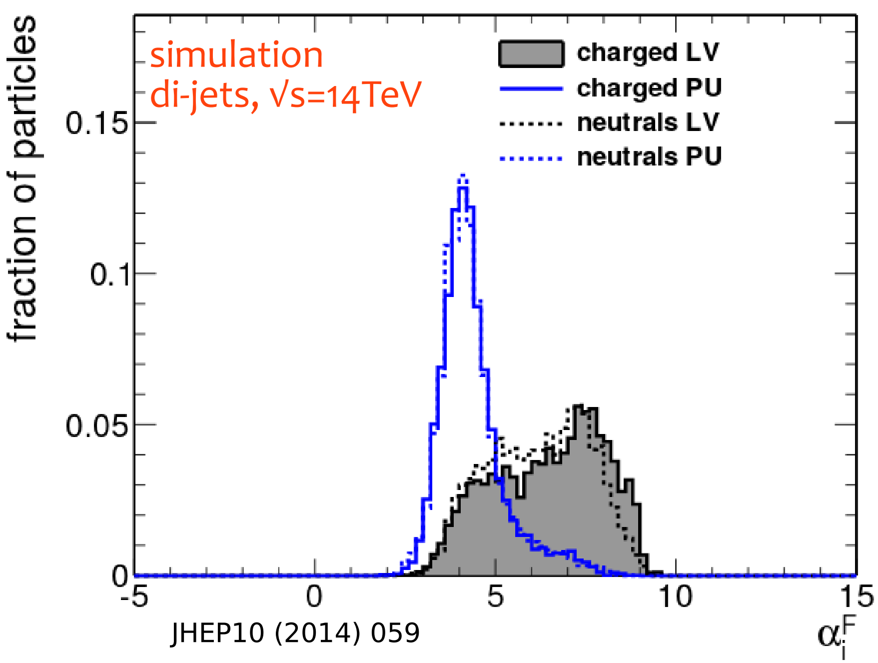
PileUp Per Particle Identification in CMS

General technique that can used in all particle-flow reconstruction.

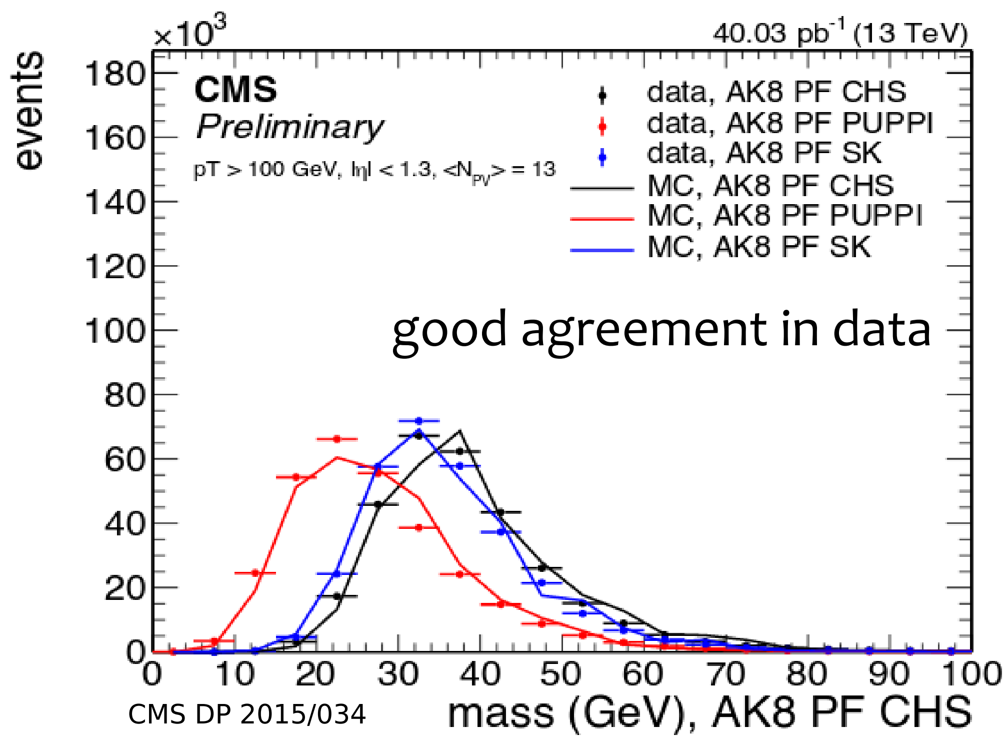
A weight is associated to every particle-flow object inputs, based on surrounding particles:

$$\alpha_i = \log \sum_{\substack{j \in Ch,PV \\ j \neq i}} \left(\frac{p_{T,j}}{\Delta R_{ij}} \right)^2 \Theta(R_0 - \Delta R_{ij})$$

-> weight(α) such that = 0 for PU
= 1 for hard scatter



JHEP10 (2014) 059

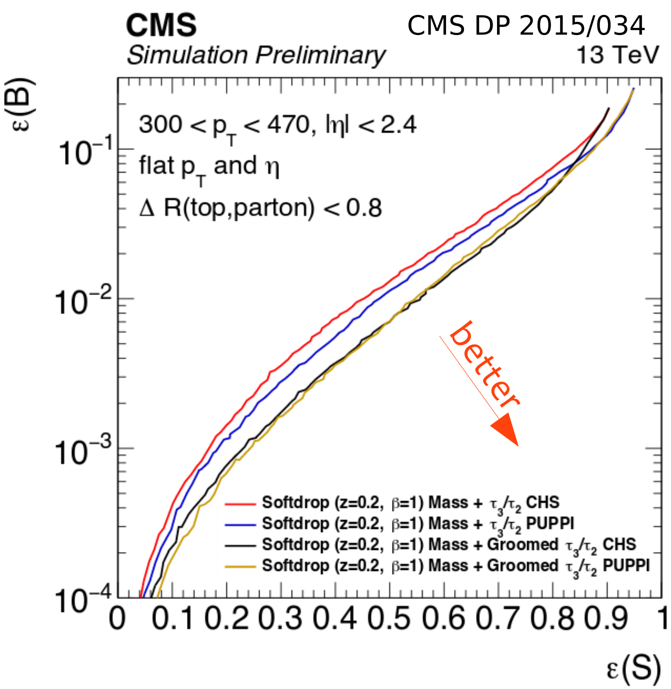


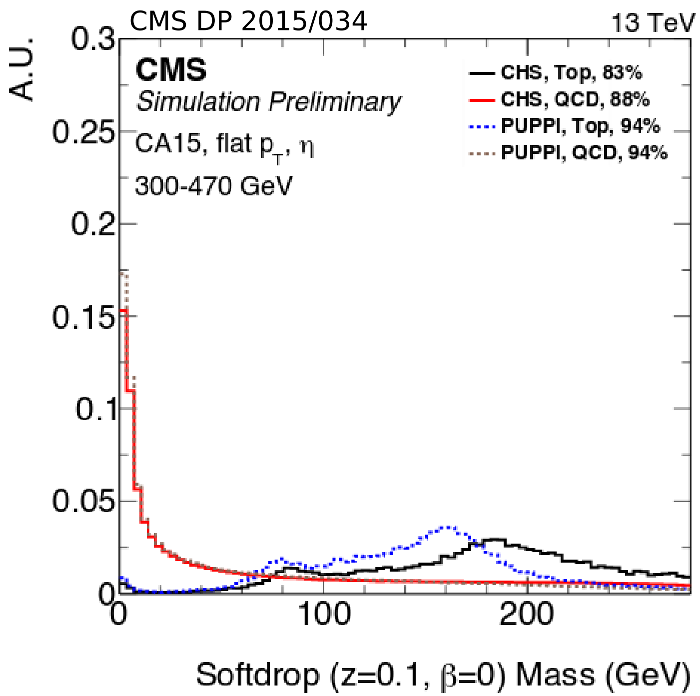
CMS DP 2015/034

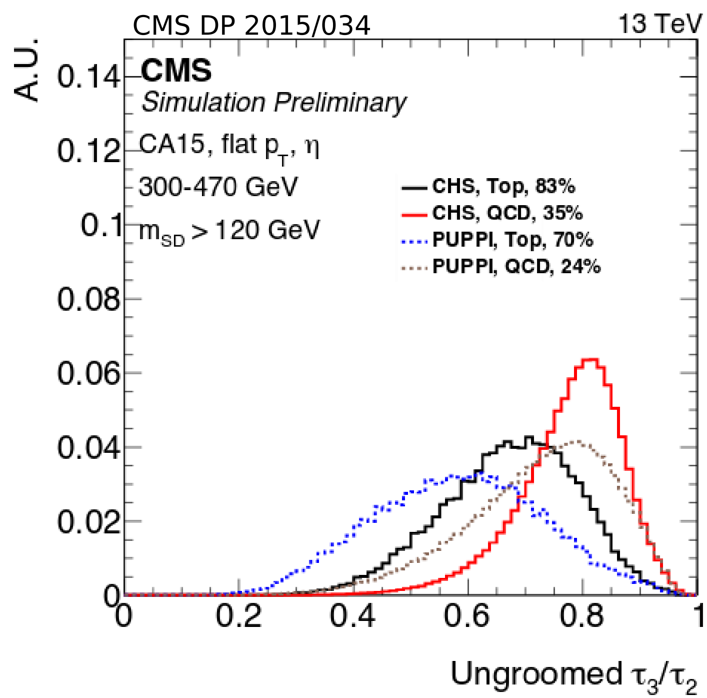
PUPPI can also be used for MET, particle isolation

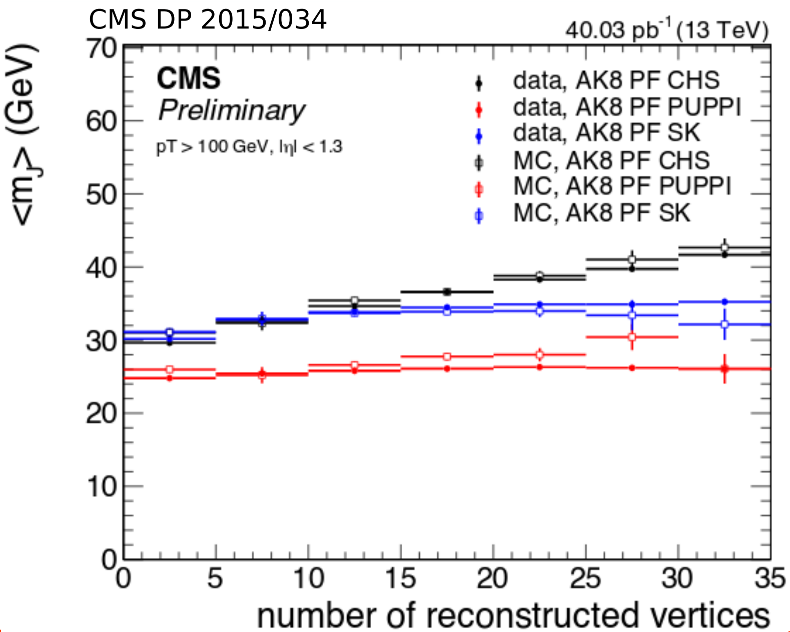
PUPPI in CMS

JHEP10 (2014) 059, CMS DP 2015/034









- Competitive performance when used in a top-tagger

CHS and PUPPI look similar in performance, but:

- PUPPI is more stable for PileUp dependence
- Variable shapes are different: better substructure resolution using PUPPI

Track-Assisted mass in ATLAS

ATLAS-CONF-2016-035

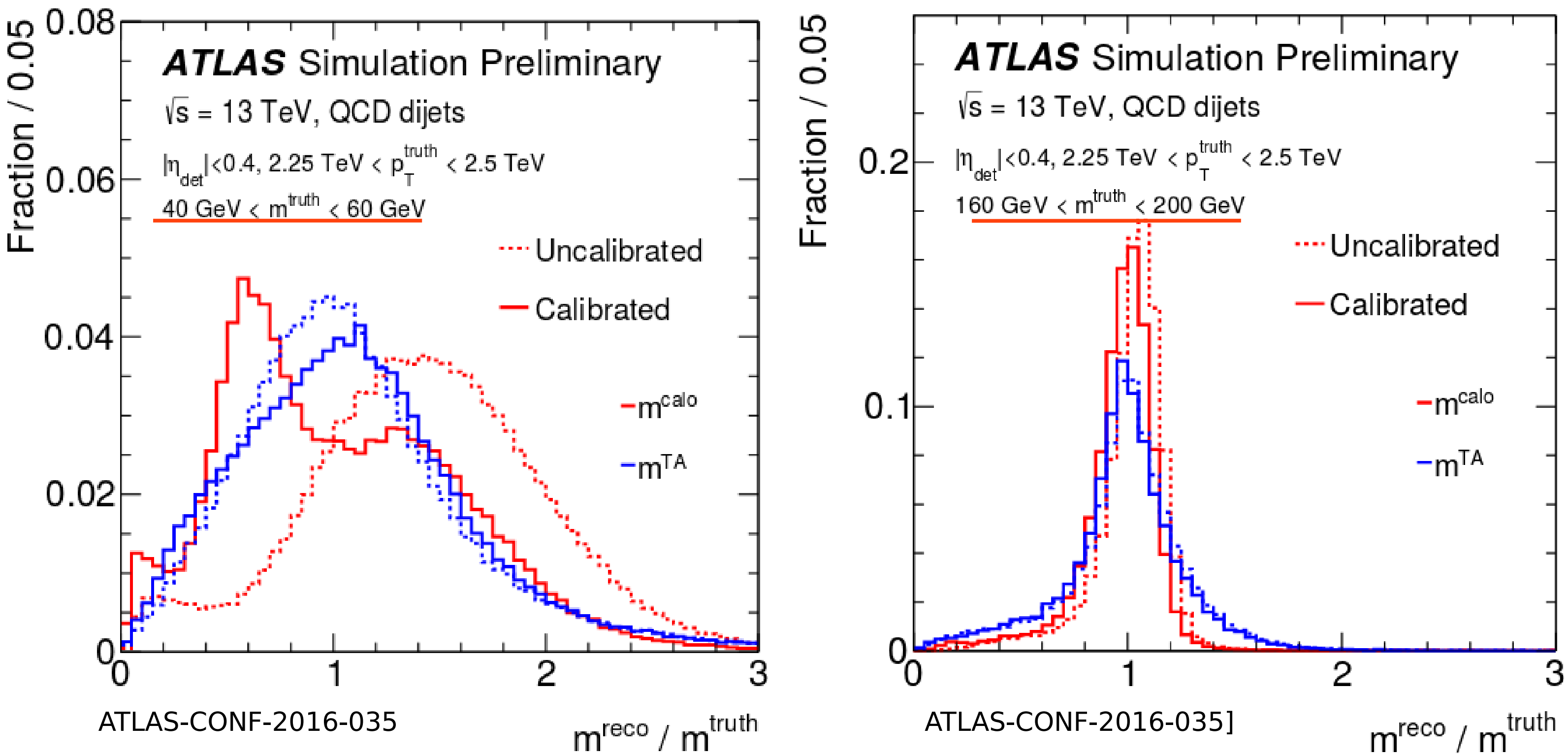
New mass variable for boosted jets

Usual mass is computed from calorimeter cell clusters.

In boosted condition, the components are closer, while the resolution is limited by the calorimeter granularity.

Tracking information can improve the granularity: $$m^{\mathrm{TA}} = \frac{p_{\mathrm{T}}^{\mathrm{calo}}}{p_{\mathrm{T}}^{\mathrm{track}}} \times m^{\mathrm{track}}$$

(but unknown neutrals smear the resolution)

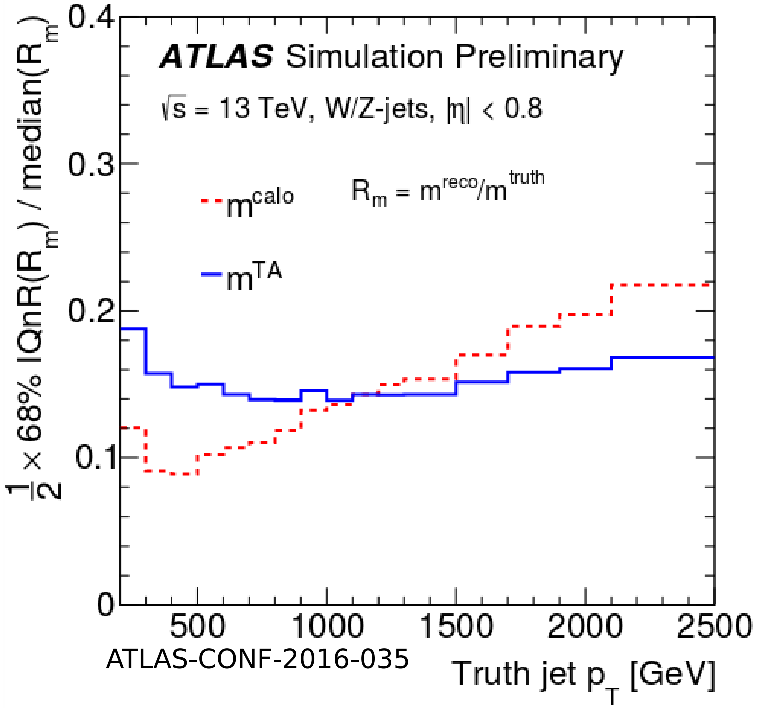


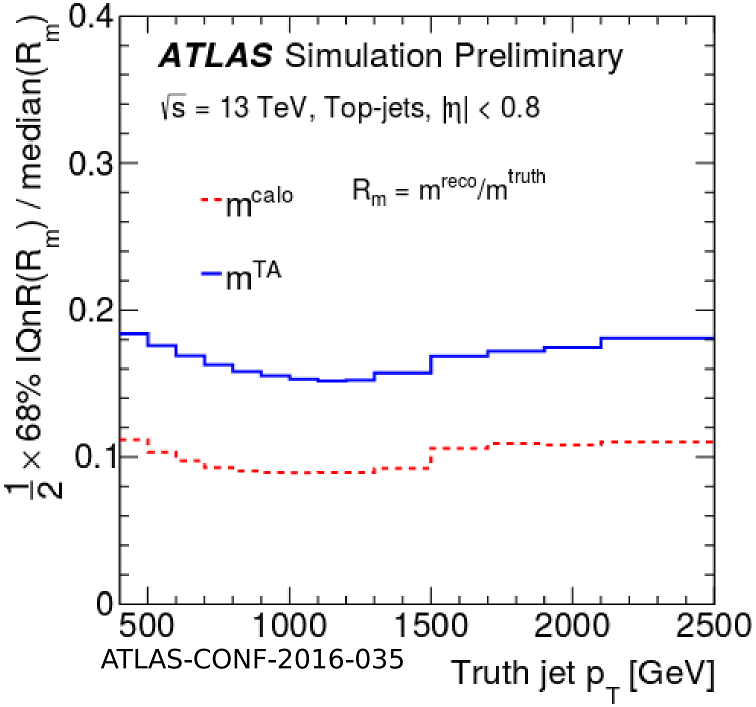
Gains when high p_T and collimated components (low mass)

Track-Assisted mass in ATLAS

ATLAS-CONF-2016-035

Resolution (half of the 68% InterQuantile Range) vs. pT for W/Z-jets and top-jets:





Apply to the large-R jet: gain for high p_T Z/W-jets, but not for top

But (as possible improvements):

- Combination with calorimeter mass
- Track-assisted subjet mass (m^{TAS}): it can profit from local charged-to-neutral corrections
  it can also be combined with calorimeter mass

-> m^{TAS} of subjets can be useful in top-tagging algorithms

Other developments

New tagger algorithms / variables:

- HEPTopTagger with BDT [arXiv:1503.05921]
- XCone jet algorithm, reconstructing exclusively N jets in the event [JHEP 11 (2015) 072]
- Image pattern recognition, using deep-neural-network [JHEP 07 (2016) 069], [Phys. Rev. D 93, 094034 (2016)]
- MVA, using event shapes in the Lorentz-boosted reference frame [arXiv:1606.06859]
- Designing Decorrelated Taggers (reducing syst. unc. by redefining substruct. var. such that there is less background mass sculpting) [JHEP 05 (2016) 156]
- ...

Also a lot of new ideas in related subjects:

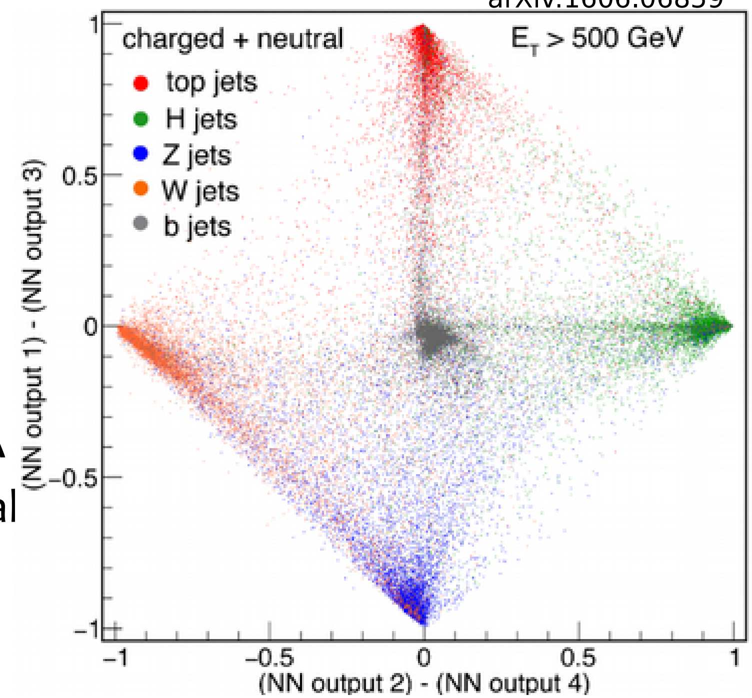
- Pile-Up suppression techniques (e.g., using wavelet analysis)
- Substructure dedicated triggers
- ...

Coming soon (non-binding):

ATLAS: work on implementing: particle-flow, soft-drop, MVA particle-based pile-up removal
shower deconstruction in analyses

CMS: PUPPI as main algorithm
implementation of new algorithms (HOTVR, DDT, MVA)

arXiv:1606.06859



Discriminating top, H, Z, W and b using a 4-output NN

Conclusion

Conclusion

universitätbonn erc

I've presented the status of the top-tagging algorithms:

1) performant and well-understood taggers are available in ATLAS / CMS

2) still a lot of developments on-going:

- solution for experimental problems (PUPPI, m^{TA})
- implementation of the new ideas (VR)
- but also new ideas from theory (cf. previous slides)

The current taggers are performing well: difficult to gain more in the ROC curve.

But possible improvements on:

- High boost / high p_T
- More universal taggers (whole p_T range)
- Systematic uncertainties considerations
- heavy object separation (either by "analysis-dependant" tagger tuning or by optimizing the tagger against another substructure object)

Backup slides

Large-R Jets

Using several jet algorithms: anti-kt / CA, with large radius: 0.8, 1.0, 1.5 (ref. in backup slide)

ATLAS: calorimeter cell clusters inputs,

CMS: particle flow inputs

Cases using small-R jets:

Re-clustering [arXiv:1606.03903] (ATLAS) :

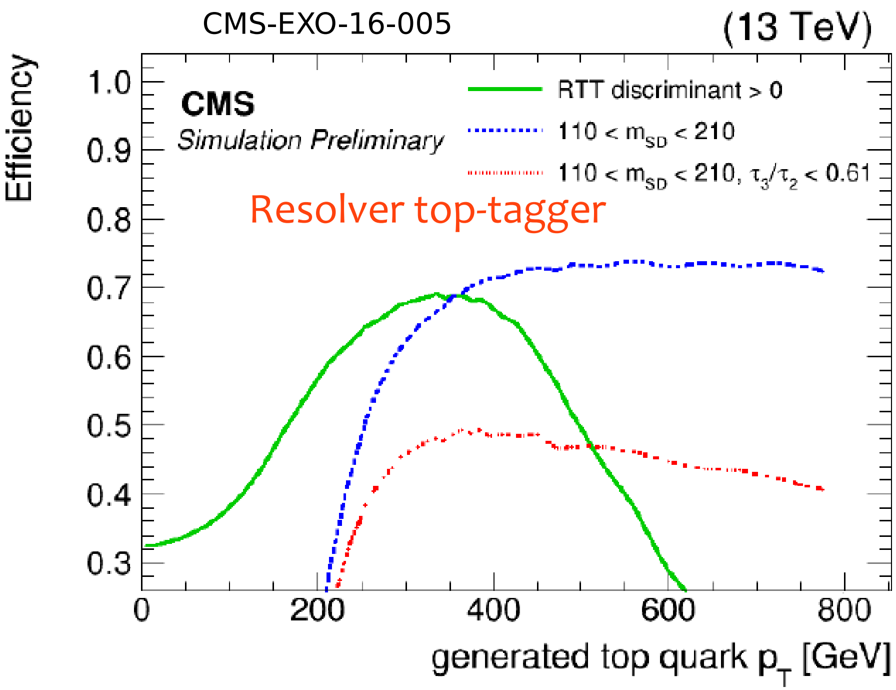
small-R jets can be used as inputs to reconstruct large-R jets

-> avoid overlaps with other contributions in dense environment (e.g., top squark search)

Resolved tagger [CMS-EXO-16-005] (CMS) :

tagger using small-R jets and MVA, identifying intermediate low boost region

-> useful for some models, such as $t\bar{t}$+DM



References for jet reconstruction

Anti-kT

M. Cacciari, G. P. Salam and G. Soyez, The Anti-kt jet clustering algorithm, JHEP 0804 (2008) 063, arXiv: 0802.1189 [hep-ph].

kT:

S. Catani, Y. L. Dokshitzer, M. H. Seymour and B. R. Webber, Nucl. Phys. B 406 (1993) 187 and refs. therein; S. D. Ellis and D. E. Soper, Phys. Rev. D 48 (1993) 3160 [hep-ph/9305266].

C/A

Y. L. Dokshitzer, G. D. Leder, S. Moretti and B. R. Webber, JHEP 9708, 001 (1997) [hep-ph/9707323]; M. Wobisch and T. Wengler, hep-ph/9907280.

Calorimeter clusters:

ATLAS Collaboration, Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1, submitted to Eur. Phys. J C (2016), arXiv: 1603.02934 [hep-ex].

Particle Flow:

CMS Collaboration, "Particle-flow event reconstruction in CMS and performance for jets, taus, and E miss T ", CMS Physics Analysis Summary CMS-PAS-PFT-09-001, CERN, 2009.

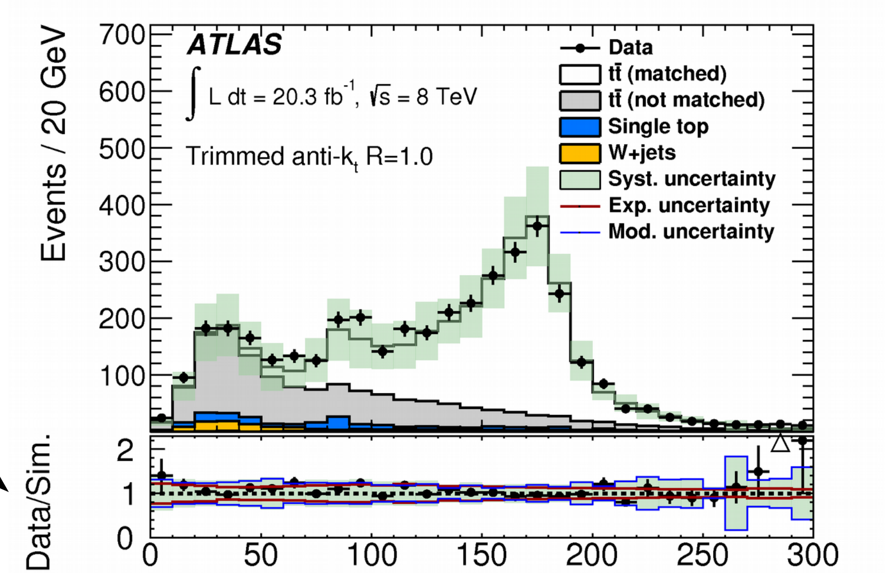
CMS Collaboration, "Commissioning of the particle-flow reconstruction in minimum-bias and jet events from pp collisions at 7 TeV", CMS Physics Analysis Summary CMS-PAS-PFT-10-002, CERN, 2010.

Substructure Variables

Trimmed and Soft-drop masses:

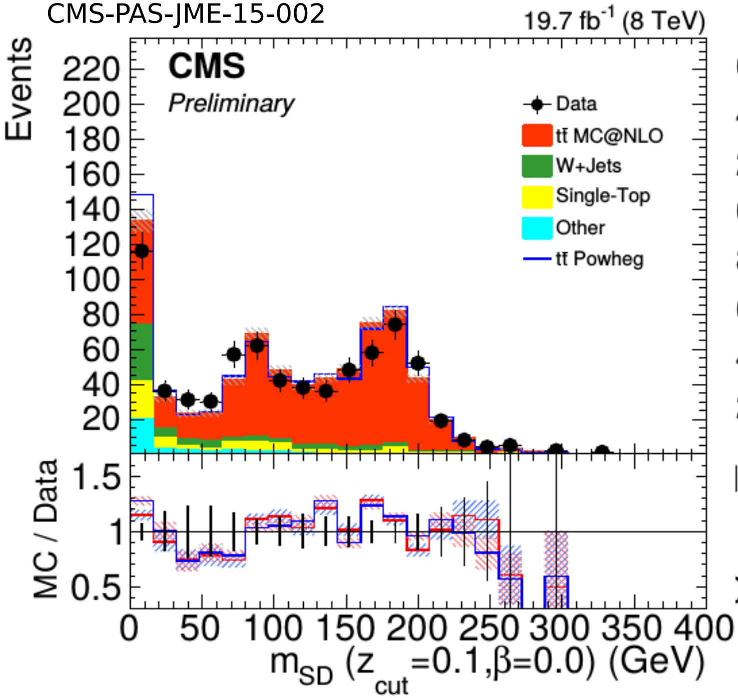
Large-R mass after grooming

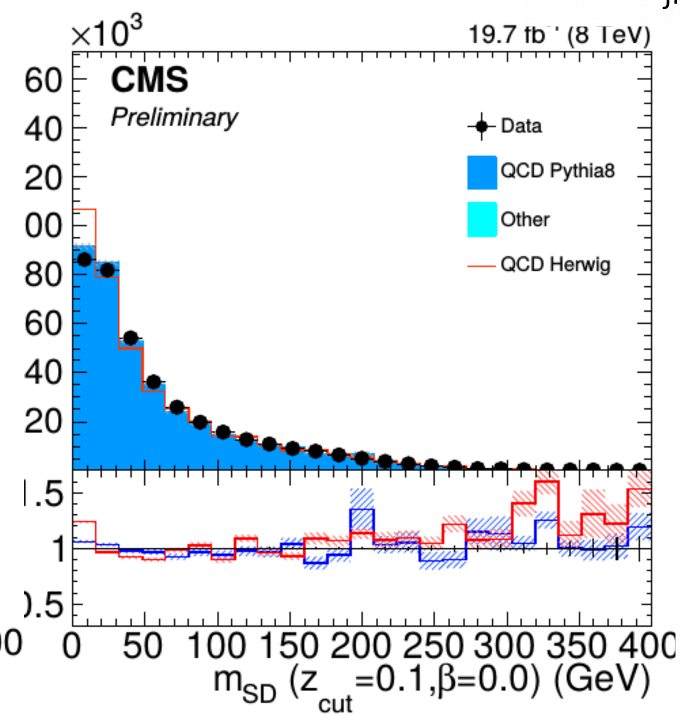
- trimmed mass [JHEP02 (2010) 084]
- soft-drop mass [JHEP05 (2014) 146]

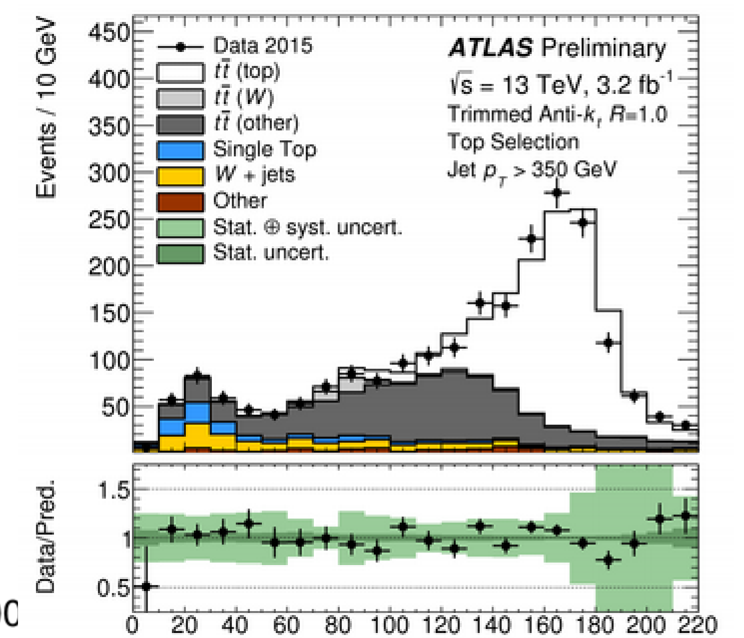


JHEP06 (2016) 093

CMS-PAS-JME-15-002







JETM-2016-005

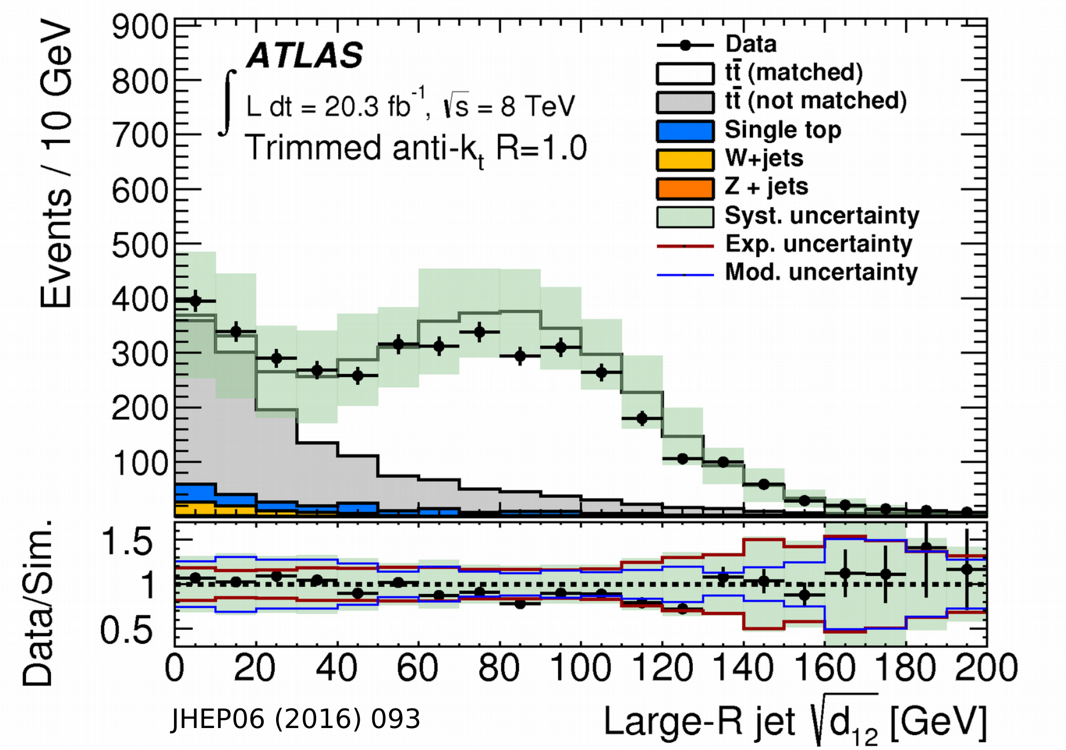
Substructure Variables

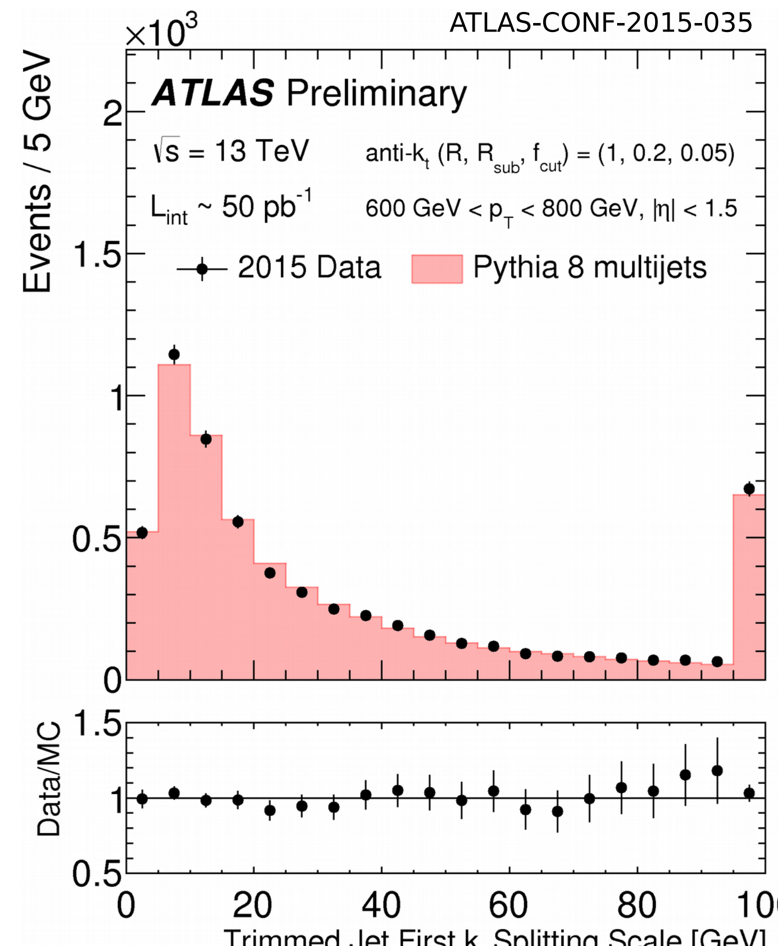
k_t splitting scale: [JHEP07 (2008) 092]

value of the jet-jet distance in the k_t algorithm clustering

$$d_{\rm cut} = \min(p_{TA}^2, p_{TB}^2)\Delta R_{AB}^2, \qquad \Delta R_{AB}^2 \equiv (\phi_A - \phi_B)^2 + (\eta_A - \eta_B)^2,$$

for the two proto-jets (A and B) in the last step





Substructure Variables

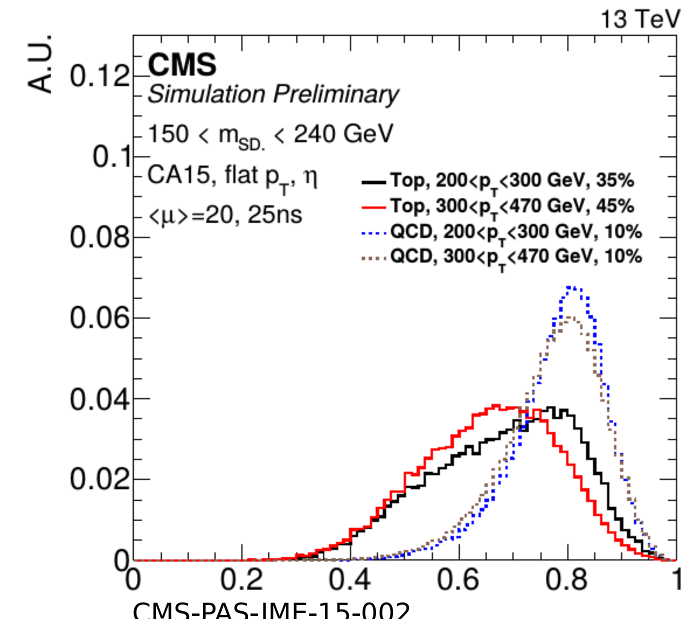
N-subjettiness ratio: [JHEP03 (2011) 015]

The N-subjettiness describes how well the jet contains N or fewer subjets

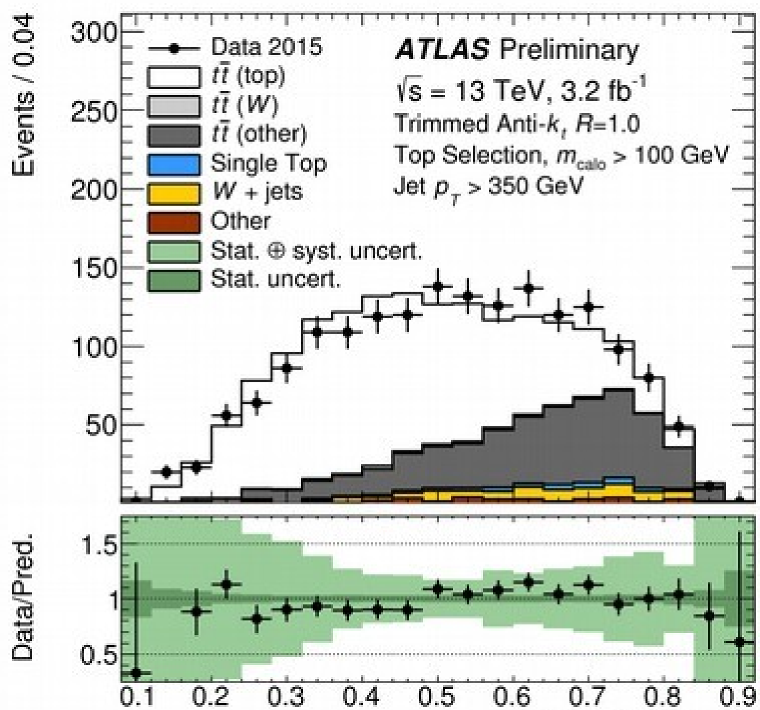
$$\tau_N = \frac{1}{d_0} \sum_k p_{\mathrm{T}k} \times \min(\delta R_{1k}, \delta R_{2k}, \ldots, \delta R_{Nk})$$

$$d_0 = \sum_k p_{\mathrm{T}k} \times R$$

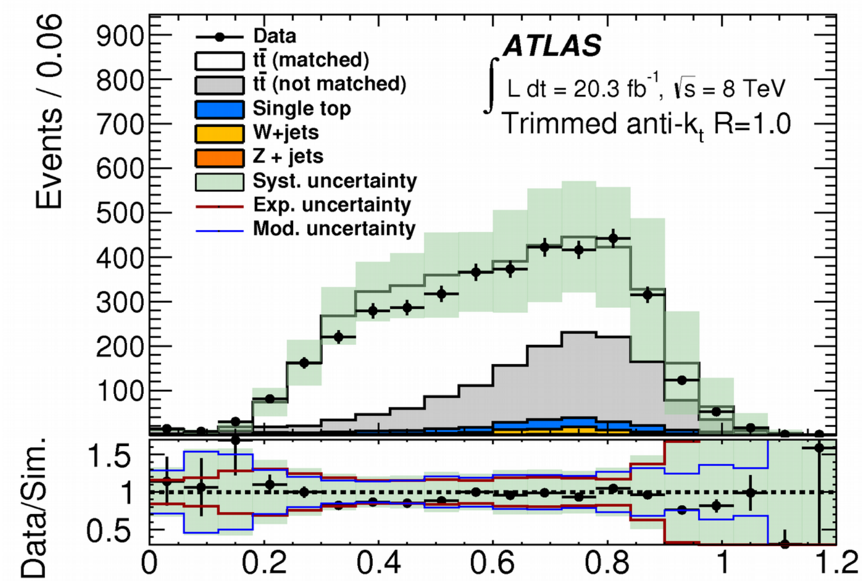
The lower $\tau_{32} = \tau_3/\tau_2$, the more 3-prong the jet is



CMS-PAS-JME-15-002



JETM-2016-005



JHEP06 (2016) 093

Substructure Variables

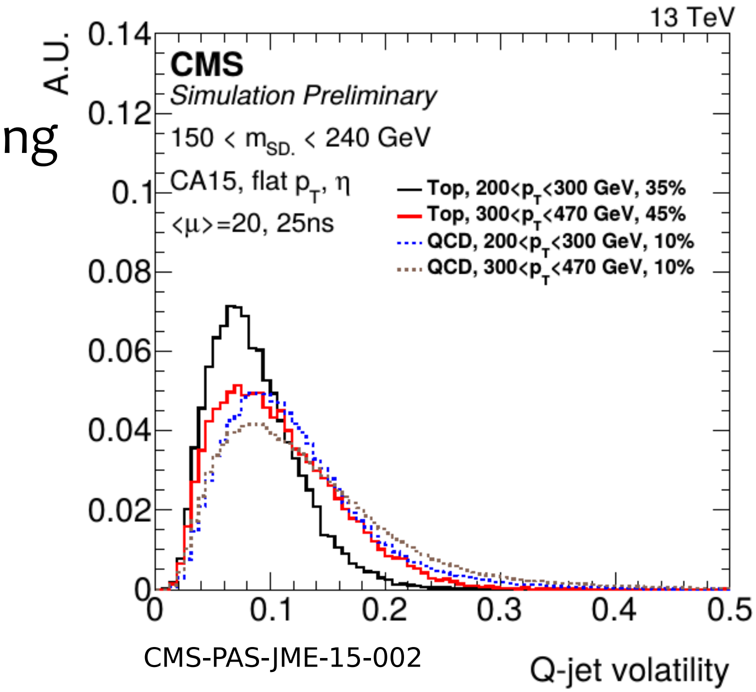
QJet volatility: [Phys. Rev. Lett. 108 182003 (2012)]

Evaluation of the jet mass stability when the pair-clustering is randomly chosen.
The probability to choose the pair ij is given by:

$$\omega_{ij}(\alpha) = \exp\left(-\alpha \frac{(d_{ij} - d^{min})}{d^{min}}\right)$$, with the rigidity α = 0.1

After 50 re-clustering of the jet, the volatility is given by:

$$V = \frac{\sqrt{\mathrm{Var}(m)}}{\langle m \rangle}$$



CMS-PAS-JME-15-002

subjet b-tagging:

Checking if one subjet is b-tagging is a good discriminating variable against QCD jets.

Specific b-tagging algorithms, efficient in dense boosted environment, are used.

In CMS, standard combined b-tagging (tracks + secondary vertex) applied to subjets of a groomed large-R jet [CMS PAS JME-13-007]

In ATLAS: Multi-Variate b-tagger [ATL-PHYS-PUB-2014-014] using dense-environment-robust variables (variables associated to the primary vertex)

Top-tagging in ATLAS

From JHEP06 (2016) 093

tagger	jet algorithm	grooming	radius parameter	p_T range	$\|\eta\|$ range
Tagger I–V W' top tagger Shower Deconstruction	anti-k_t	trimming	$R = 1.0$	> 350 GeV	< 2
Shower Deconstruction	C/A	none	$R = 1.2$	> 350 GeV	< 2
HEPTopTagger	C/A	none	$R = 1.5$	> 200 GeV	< 2

Tagger I	$\sqrt{d_{12}} > 40$ GeV
Tagger II	m > 100 GeV
Tagger III	m > 100 GeV && $\sqrt{d_{12}} > 40$ GeV
Tagger IV	m > 100 GeV && $\sqrt{d_{12}} > 40$ GeV && $\sqrt{d_{23}} > 10$ GeV
Tagger V	m > 100 GeV && $\sqrt{d_{12}} > 40$ GeV && $\sqrt{d_{23}} > 20$ GeV
W' Top Tagger	$\sqrt{d_{12}} > 40$ GeV && $0.4 < \tau_{21} < 0.9$ && $\tau_{32} < 0.65$

HEPTopTagger

HEPTopTagger: [JHEP 1010 (2010) 078]

The hard subjets are obtained from Mass-drop + filtering algorithm:

- Declustering of the large-R jet
- Subjets with low mass w.r.t. to parent's mass are excluded
- iterate until $m_{subjets} < m_{cut}$

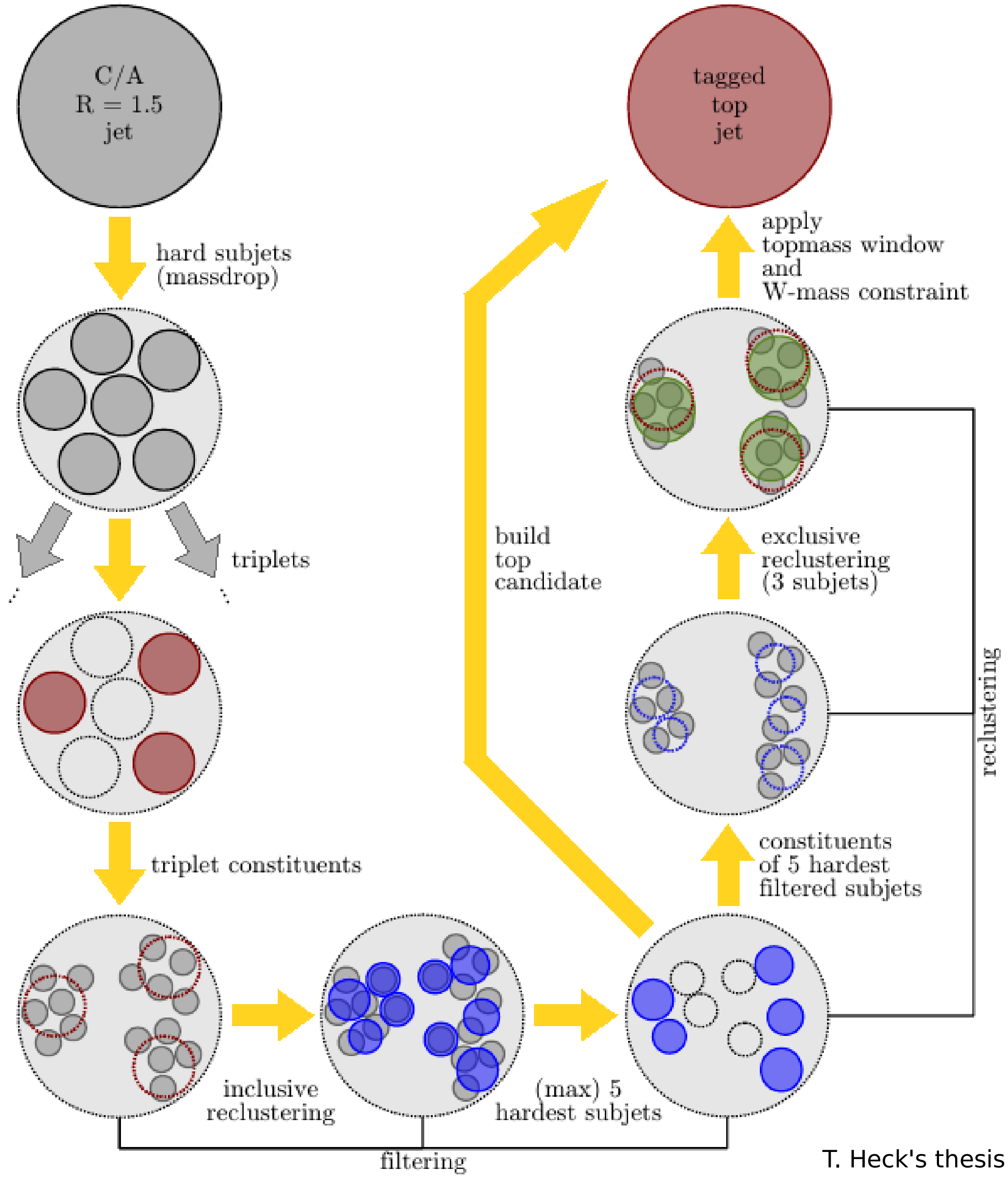
The W-mass constraint is based on the dijet masses of the 3 subjets

HEPTopTagger v2: [arXiv:1503.05921]

The procedure is done for all R, from 1.5 to 0.5

R_{opt} = smallest R with $m(R_{opt}) > 0.8\ m(R=1$

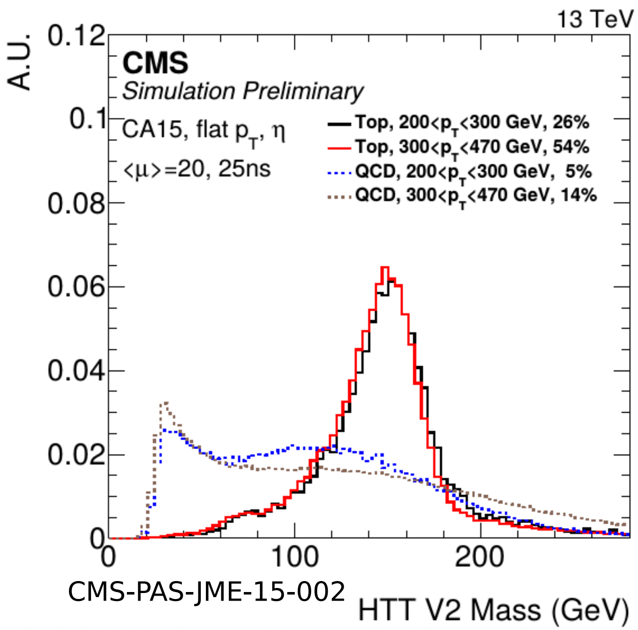
$\Delta R_{opt} = R_{opt}^{expt}(p_T) - R_{opt}$



T. Heck's thesis

HEPTopTagger

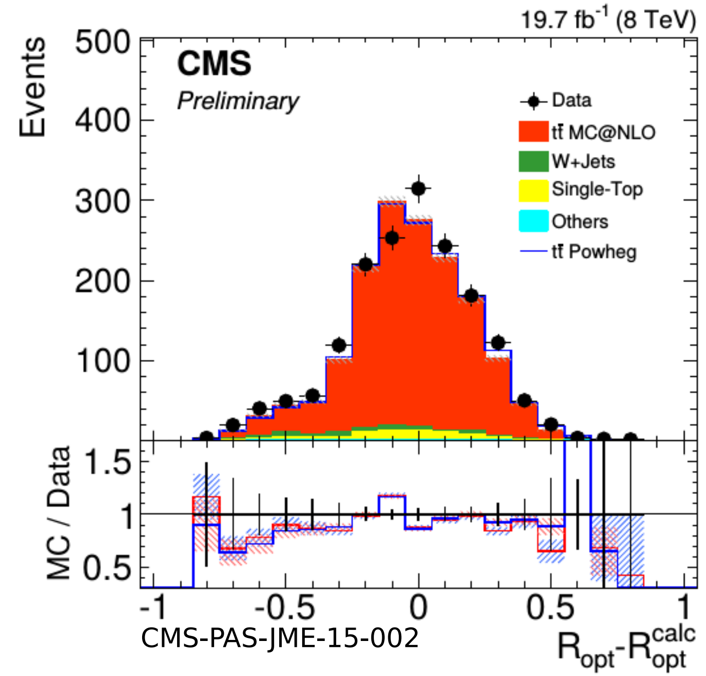
[JHEP 1010 (2010) 078]

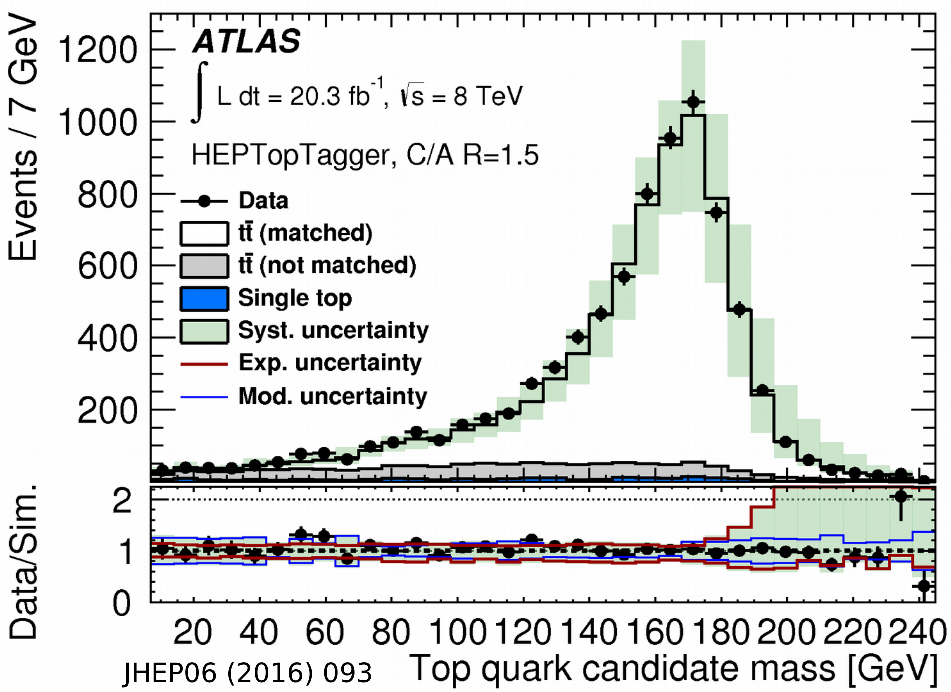


1) the Large-R C/A 1.5 jet is declustered in hard subjets (mass-drop filtering)

2) all the possible triplet of hard subjets are filtered and tested for top-like kinematics

3) Mass obtained from the reclustered subjets in the best triplet candidate

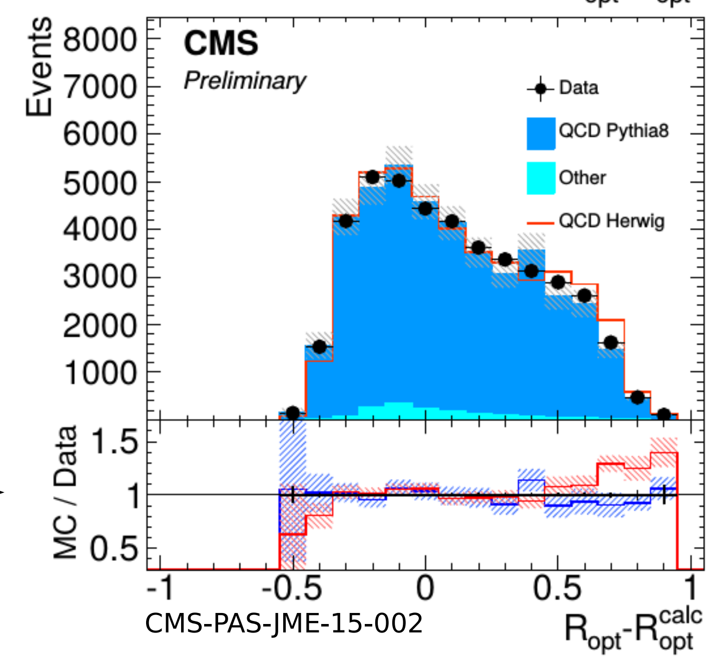




In HTTv2 [arXiv:1503.05921]:

4) minimum effective radius is compared with expected radius curve

$R_{opt}-R_{opt}^{calc}$



CMS Top Tagger

CMS Top Tagger: [CMS-PAS-JME-13-007]

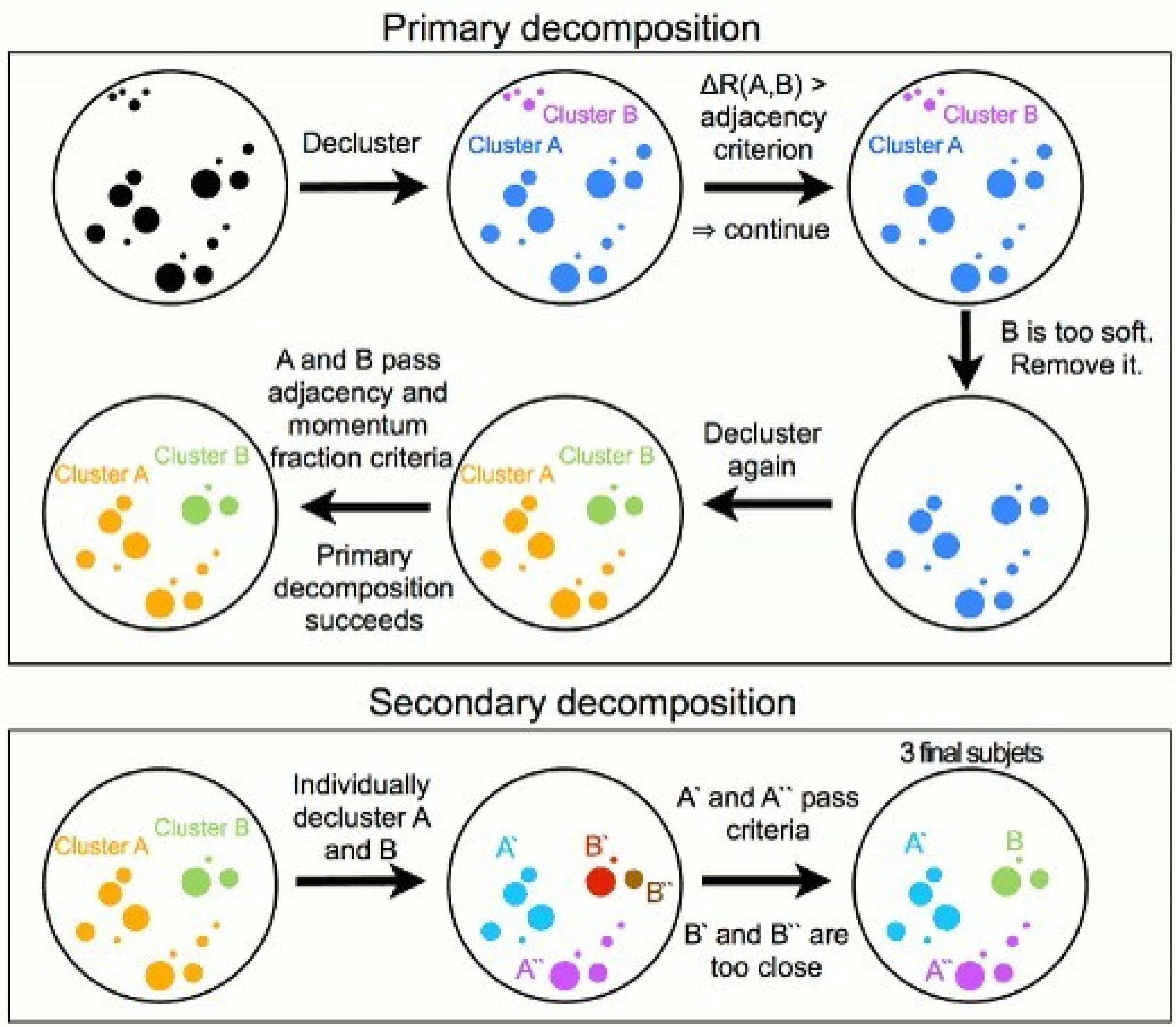
Decomposition algorithm:

1) Decluster the jets into the 2 subjets
2) the ΔR should be > D_{cut}
3) if one of the two subjets is too soft, it is discarded and the procedure restarts with the remaining subjet

The decomposition algorithm is applied twice.

If < 3 subjets -> top-tagging fails

$\min(m_{12}, m_{13}, m_{23})$, $N_{subjets}$, m_{jets} are used as discriminative variables



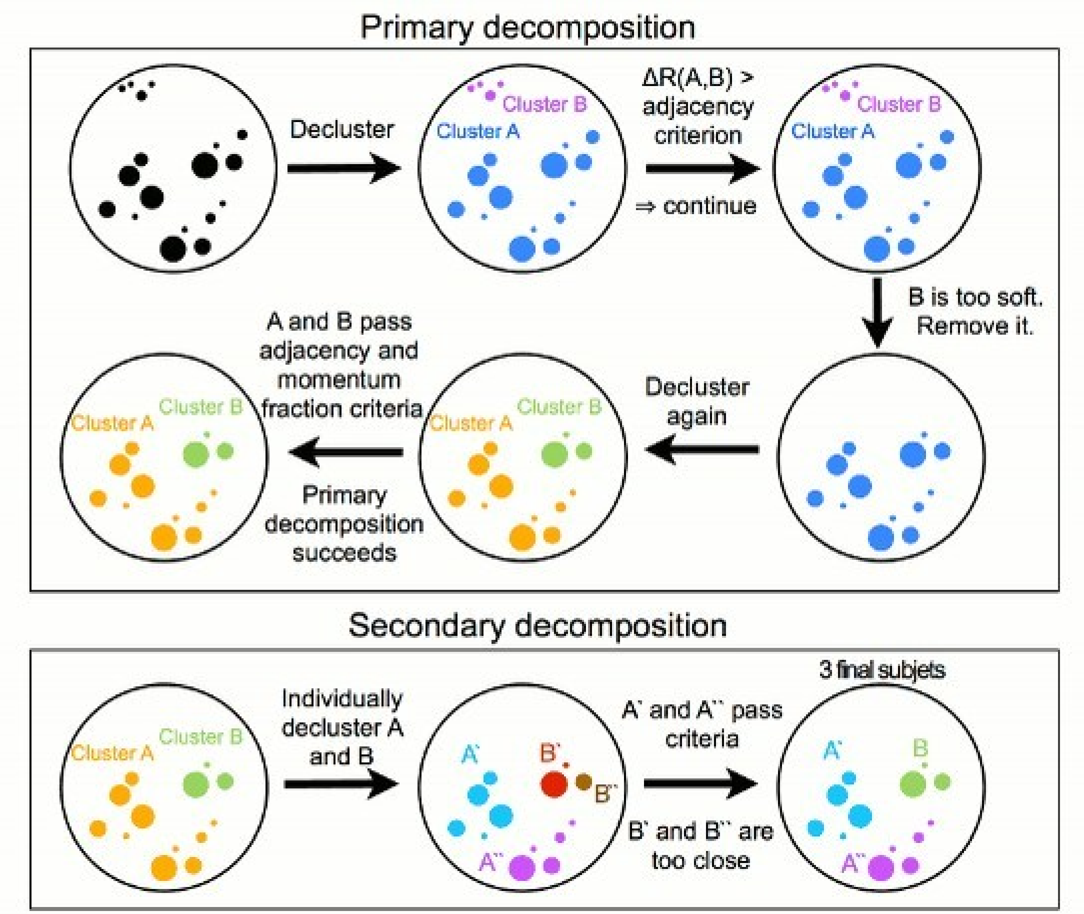
CMS-PAS-JME-13-007

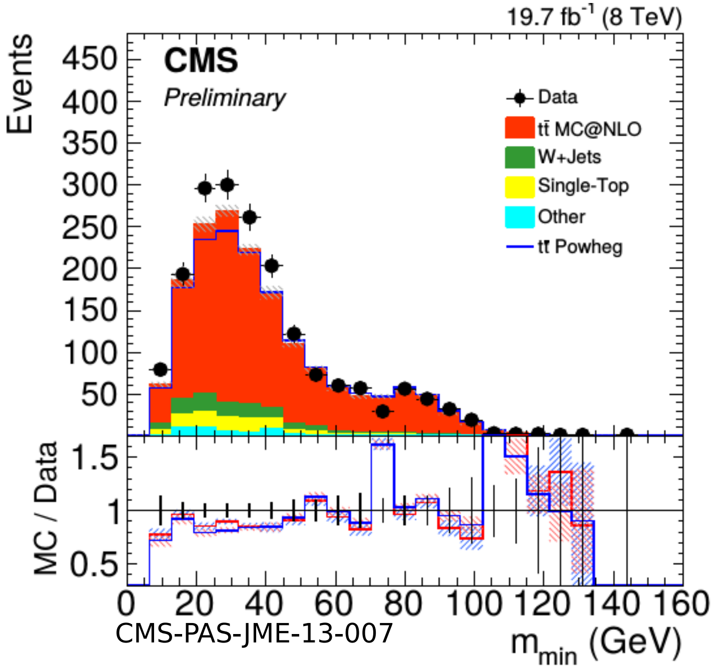
CMS Top Tagger [CMS-PAS-JME-13-007]

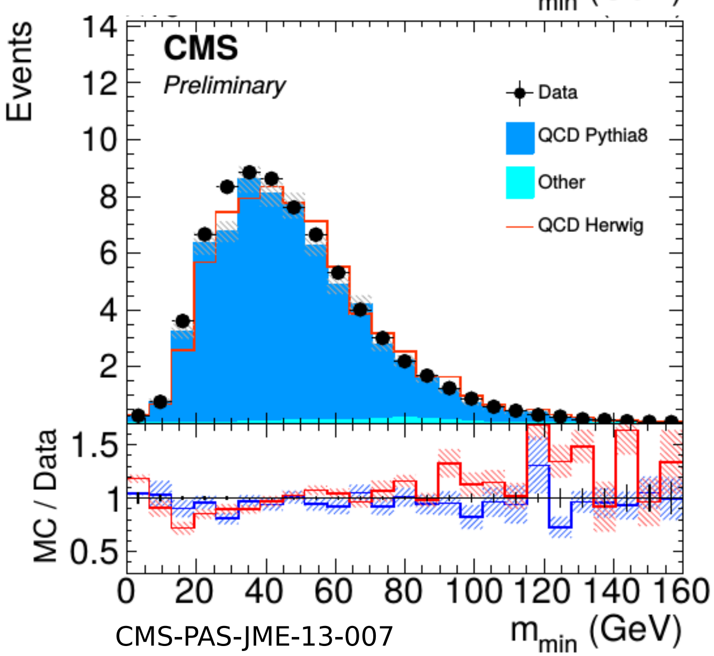
1) the large-R CA 0.8 jet is decomposed a first time
(soft components are rejected, if too adjacent, the decomposition fails)

2) the results of this decomposition is decomposed a second time

3) several discriminating variables: min(m_{12}, m_{13}, m_{23}), $N_{subjets}$, m_{jets}







Shower Deconstruction

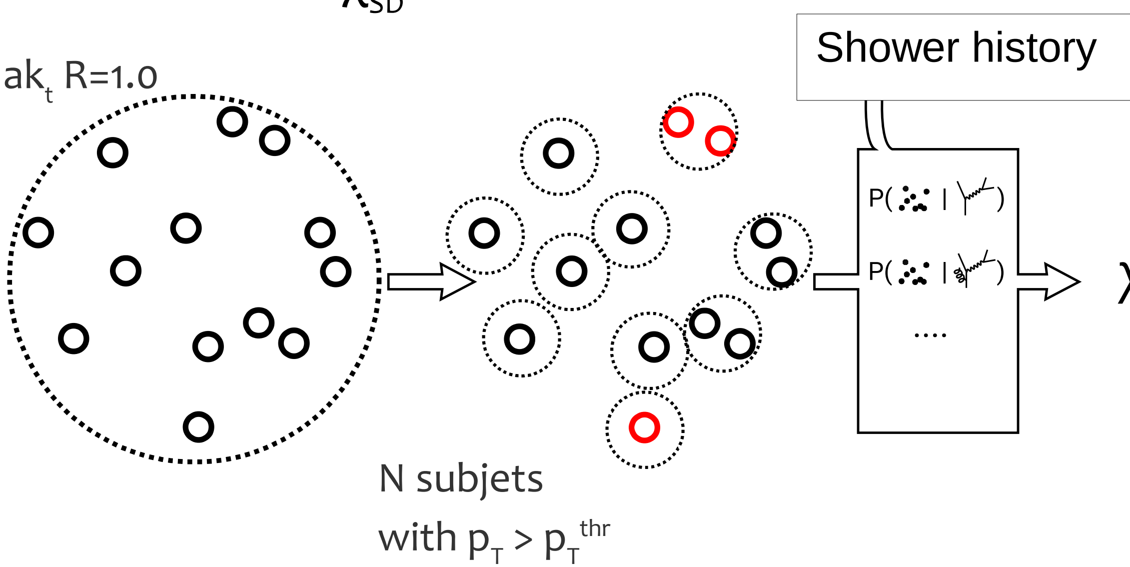
Shower Deconstruction: [Phys. Rev. D 87 054012]

1) Compute possible shower histories for signal and backgrounds (based on Sudakov form factors and splitting functions)

signal

backgrounds

2) Likelihood ratio χ_{SD}:



$$\chi_{SD} = \frac{P(..|Signal)}{P(..|Bkgd)}$$

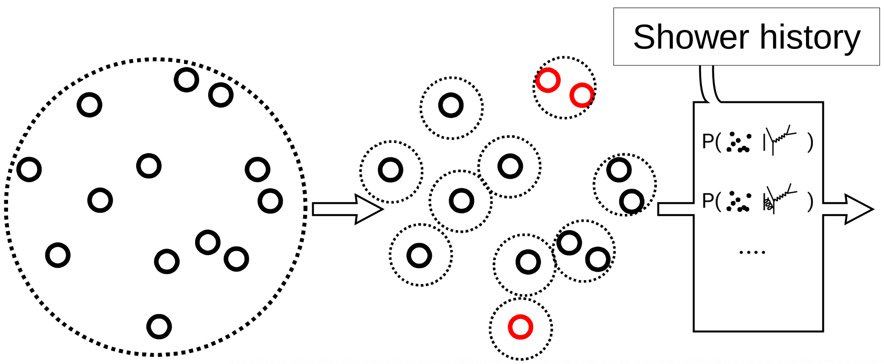
➡ cut on χ_{SD}

Shower Deconstruction [Phys. Rev. D 87 054012]

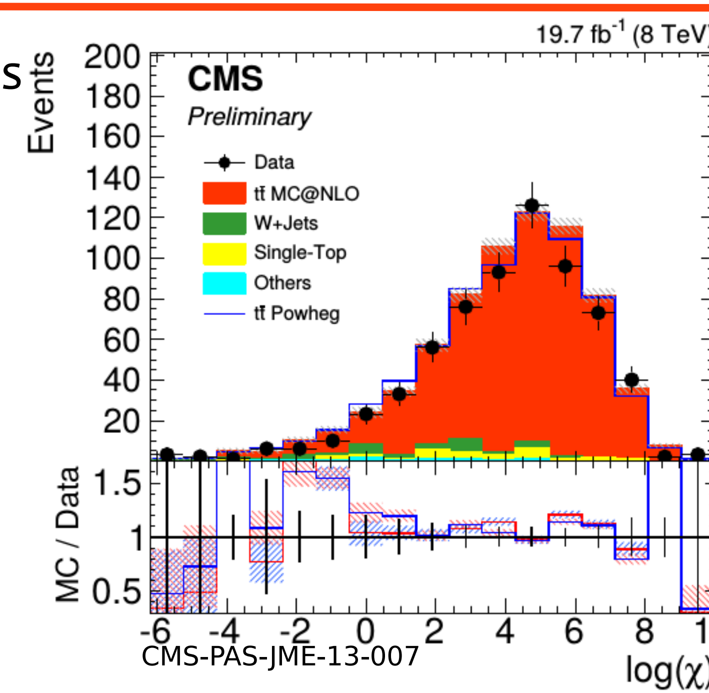
1) shower histories are computed for signal and backgrounds
(based on Sudakov form factors and splitting functions)

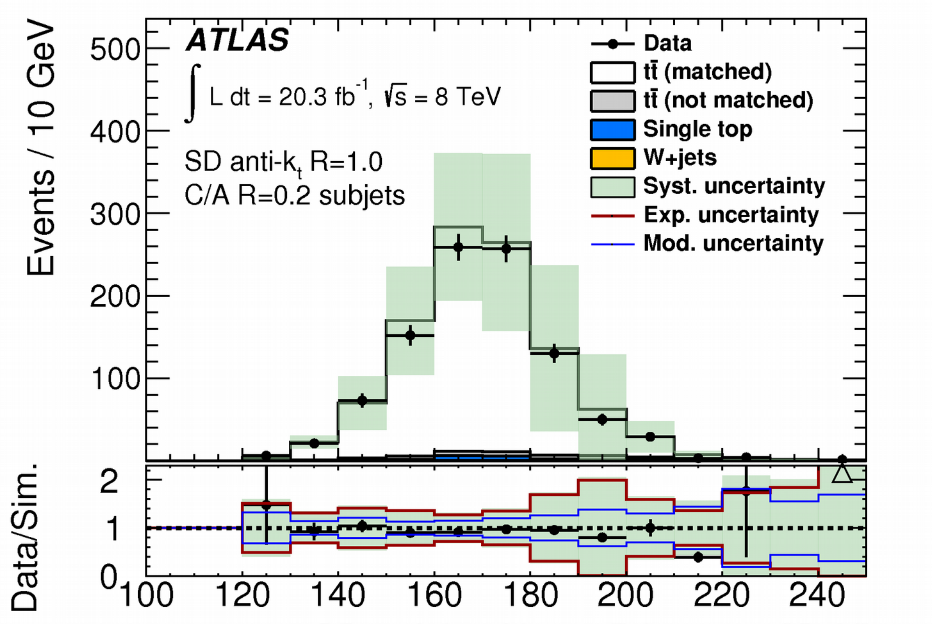
2) the large-R jet is reclustered in small-R microjets

3) the likelihood ratio χ_{SD} is computed

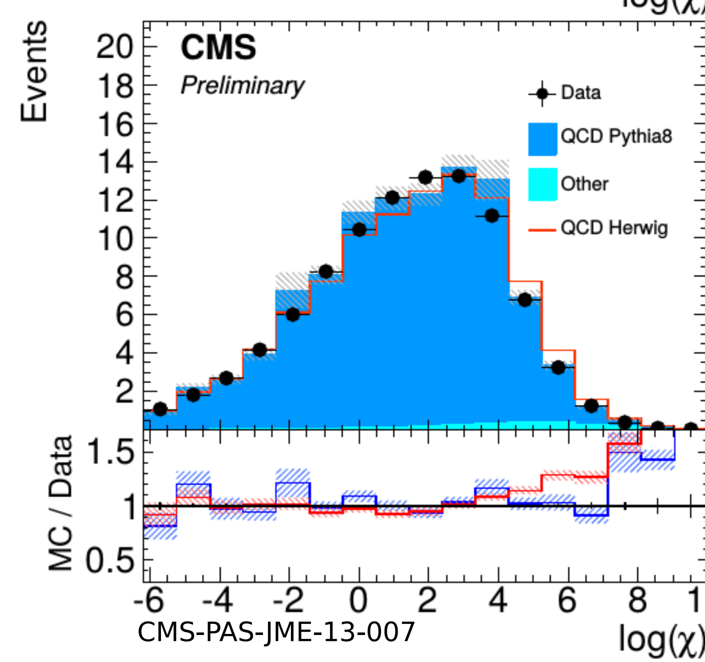


$$\chi_{SD} = \frac{P(..|Signal)}{P(..|Bkgd)}$$





JHEP06 (2016) 093

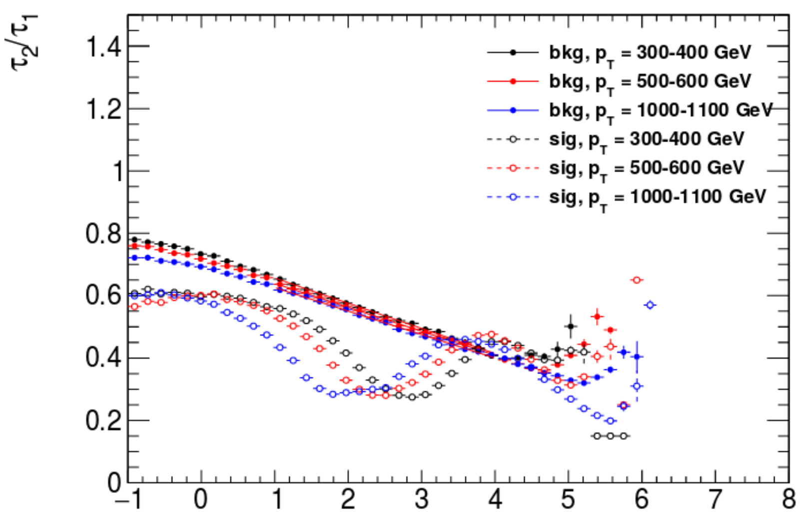


Designing Decorrelated Taggers

Usual substructure variables are correlated to the mass, in a p_T depend way.

It leads to a sculpting of the mass for the background when cutting on a variable, which leads to higher systematic uncertainties.

Defining new variables [JHEP 05 (2016) 156] with same discriminating power, but reducing the systematic uncertainties based on theoretical considerations.



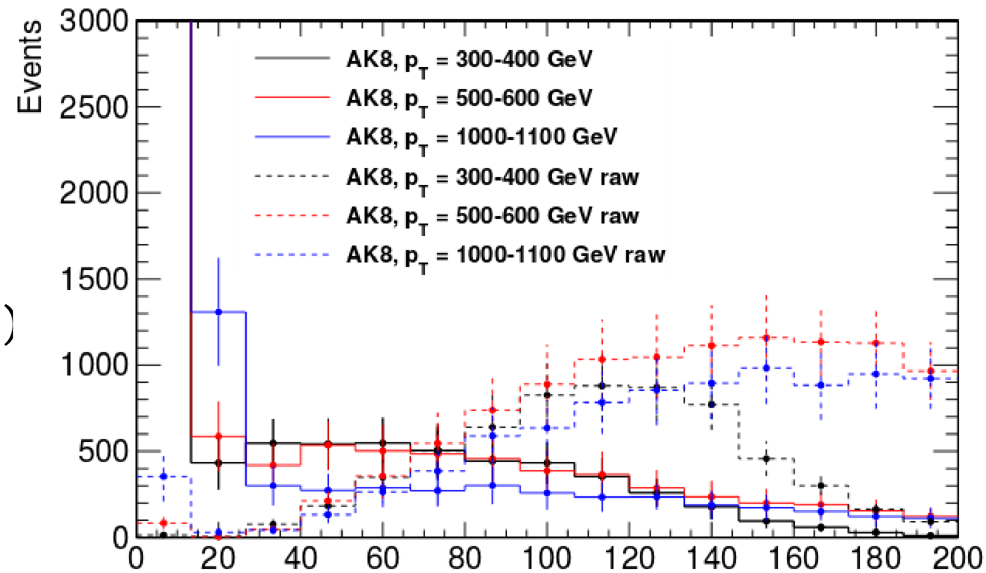
JHEP 05 (2016) 156

$$\rho' = \log\left(\frac{m^2}{p_T \mu}\right) \qquad \mu \sim 1 \text{ GeV}$$

new variable: $\tau'_{21} = \tau_2/\tau_1 - M \times \rho'$

slope from the fit from the bkgd curve

Applying the same efficiency cut on τ_{21} (dashed) and τ_{21}' (solid), flat background band for τ_{21}' for the whole p_T range



JHEP 05 (2016) 156

Very high luminosity

https://twiki.cern.ch/twiki/bin/view/AtlasPublic/JetSubstructureECFA2014