Resolved Top Tagger in CMS

An MVA-based tool used for tagging low to moderately boosted hadronic tops

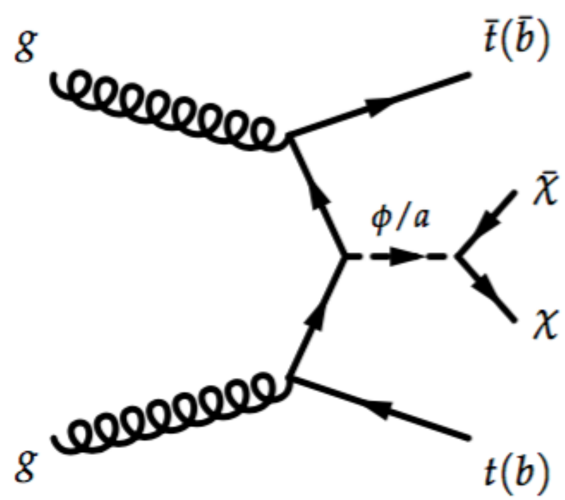
Stanislava Sevova, Northwestern University
on behalf of CMS Collaboration

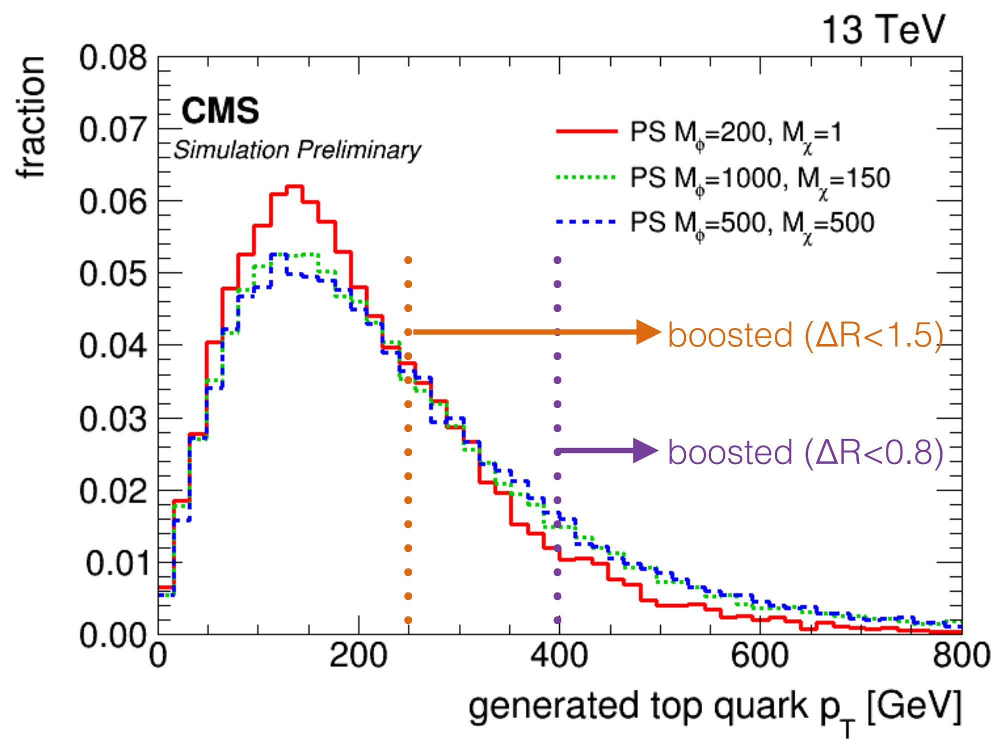
Overview

- Motivation
- Resolved Top Tagger (RTT) Overview
- Training and Inputs
- Expected Performance
- Characterization in Data and Simulation
- Systematic Uncertainties
- Application

Motivation (1)

- Many BSM theories predict low to moderately boosted tops where tops decay to three resolved jets in final state
  - e.g. tops in tt+DM models are typically not boosted



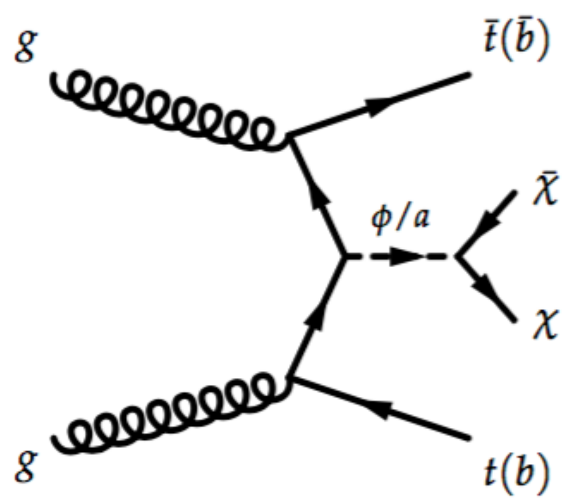


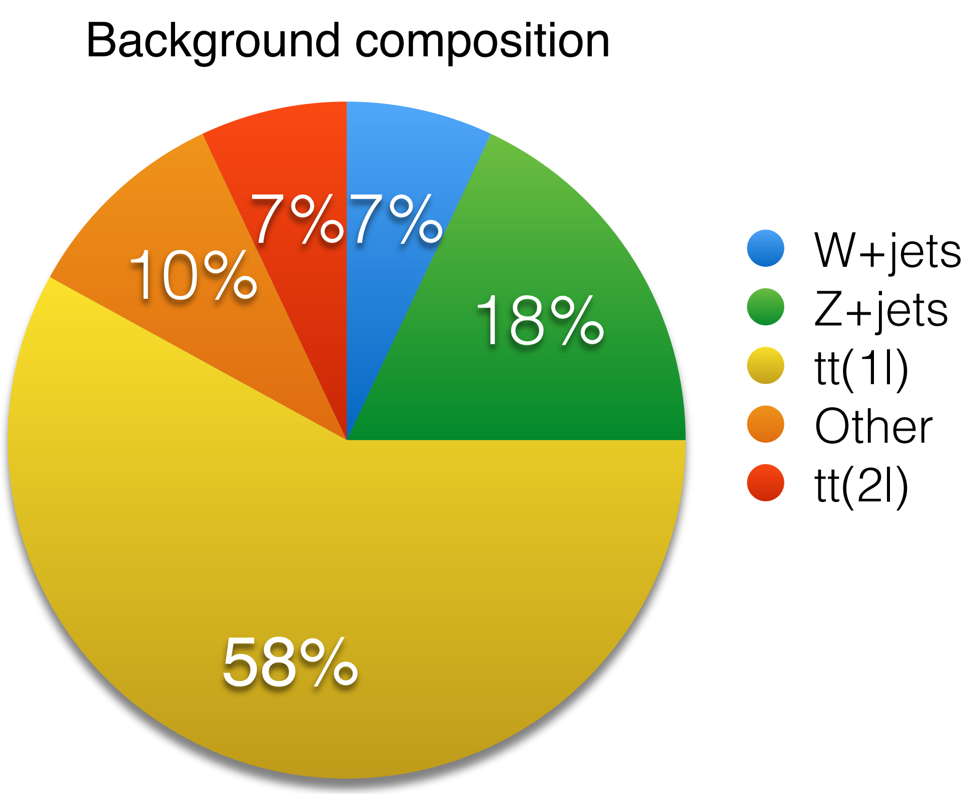
  - Aim to leverage jet properties/characterization tools to tag resolved top decays

- Similar arguments motivated earlier development of a resolved W tagger
  - Applied in Run-1 CMS monojet/mono-V Dark Matter search (EXO-12-055)
  - Presented at BOOST 2014 (JME-14-002)

Motivation (2)

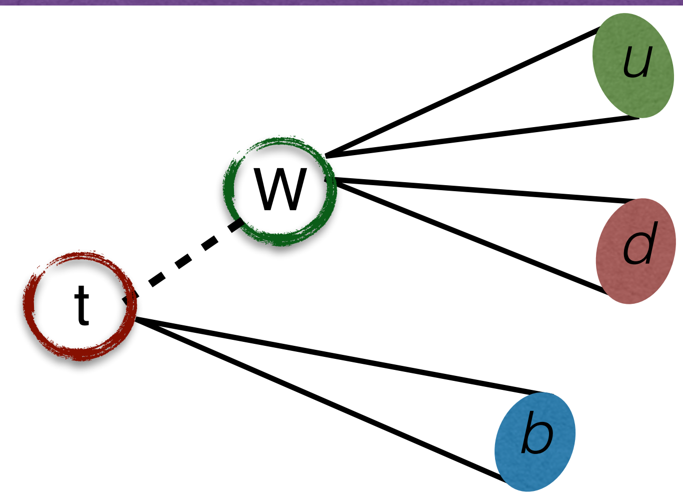
- Search for tt+DM in semileptonic and hadronic channels
  - hadronic channel background has either no tops (i.e. V+jets) or one hadronically decaying top (i.e. tt(1ℓ))





- MVA training performed in simulated tt sample
  - **Signal training sample**: each jet in trijet combo is matched (R<0.3) to generator level quarks from a hadronic top decay
  - **Background training sample**: any trijet combo where *at least one jet is not matched* to a generator level quark from hadronic top decay

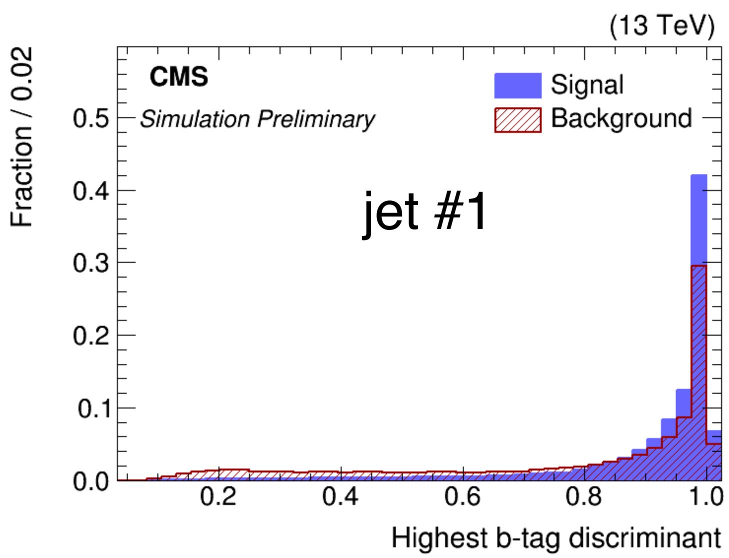
RTT Training

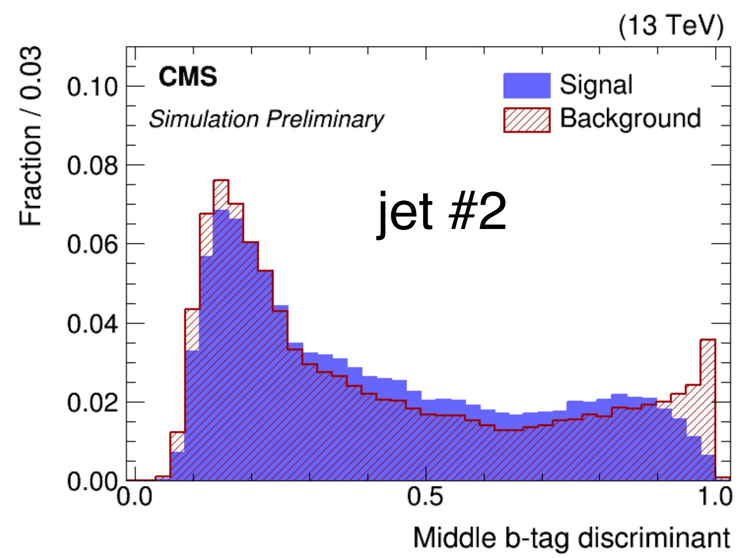


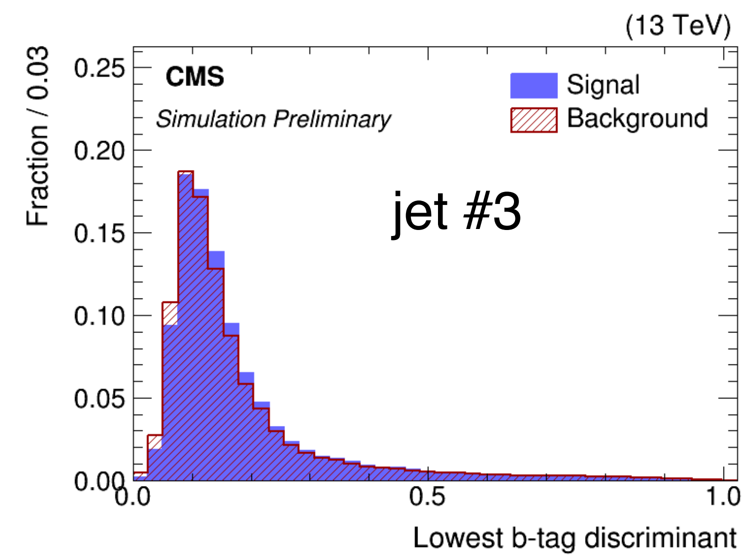
- MVA discriminant used to identify jet triplets from top decays
  - **Self-contained**: uses only interrelations and properties from jet triplets, and nothing from global event
  - **Inputs**: b-tag discriminant, QGL, angles between jets, kinematic fit probability
- Jet selection: anti-k_T R=0.4, p_T > 30 GeV, $|\eta| < 2.4$
  - Top mass window: 100 GeV < M_{jjj} < 300 GeV
- Trained using BDT with gradient descent method on 100,000 signal and 200,000 background events
  - 1000 trees used for training
  - Training depth = 3
  - Shrinkage parameter = 0.05 ➪ small parameter means more decision trees

Inputs (1)

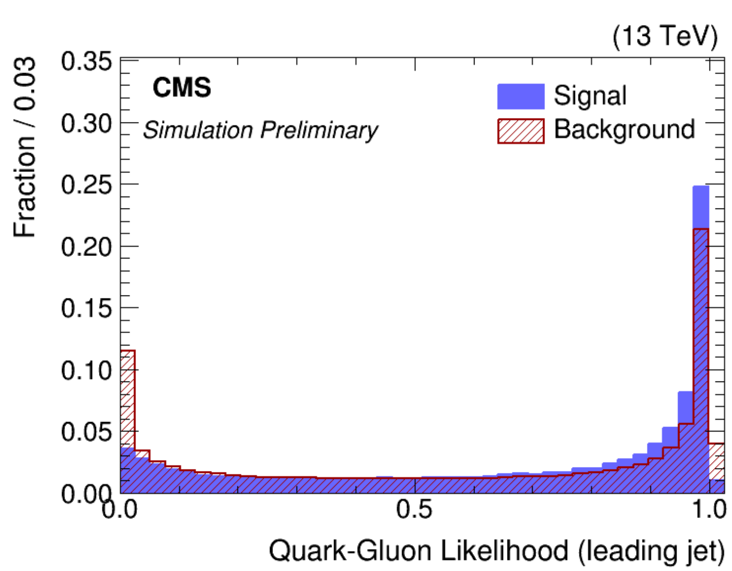
- **b-tag discriminant:** no explicit b-tag requirement for jets fed into BDT
    - jet #1 designated as “b-jet” (highest b-tag discriminant)
    - jet #2 and jet #3 are considered to be coming from W decay

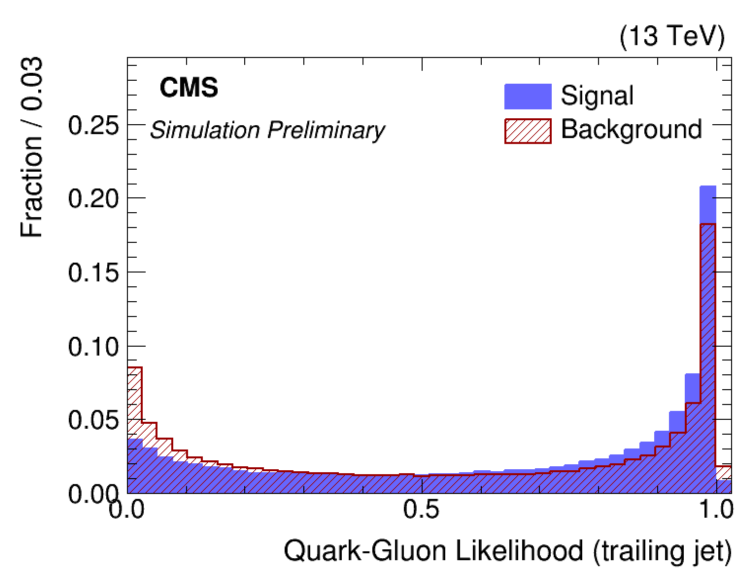






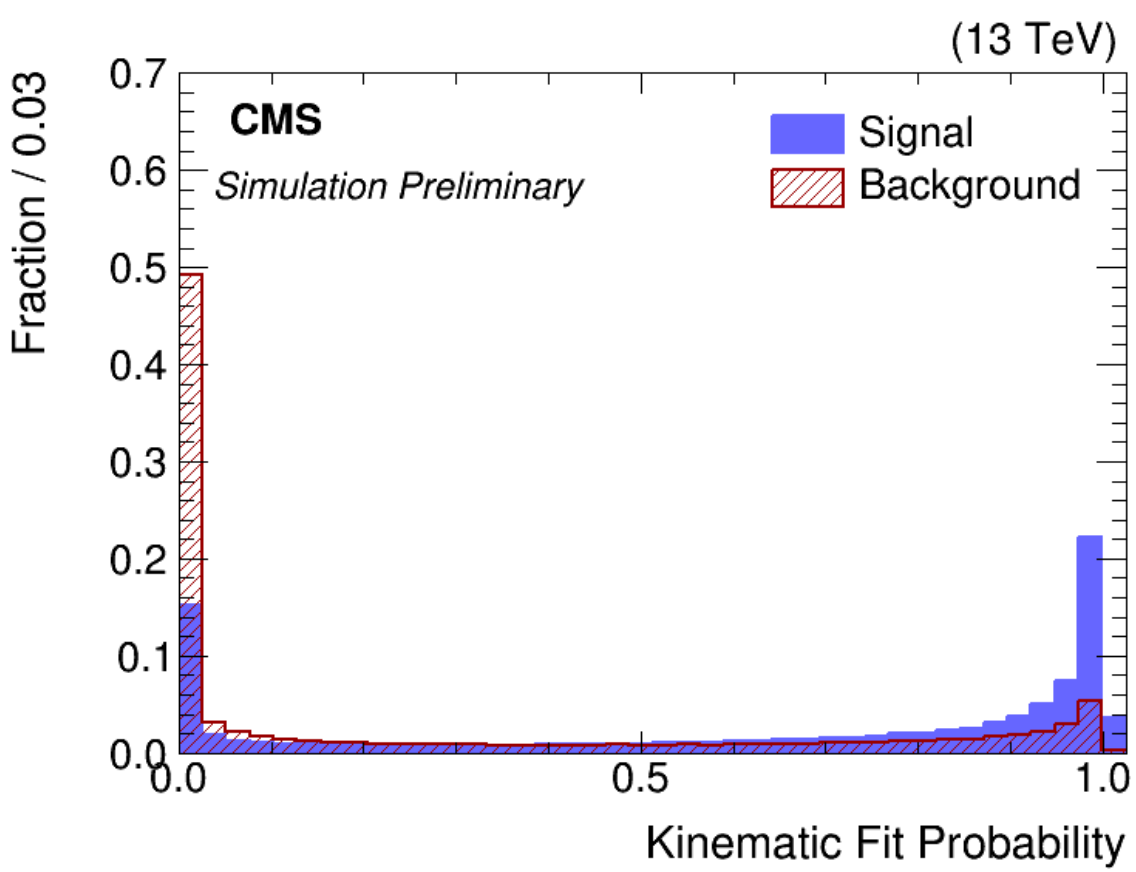
- **QGL:** quark-gluon likelihood discriminant
    - QGL of jets from W decay
    - Discriminant against combinatoric background from gluon-initiated jets

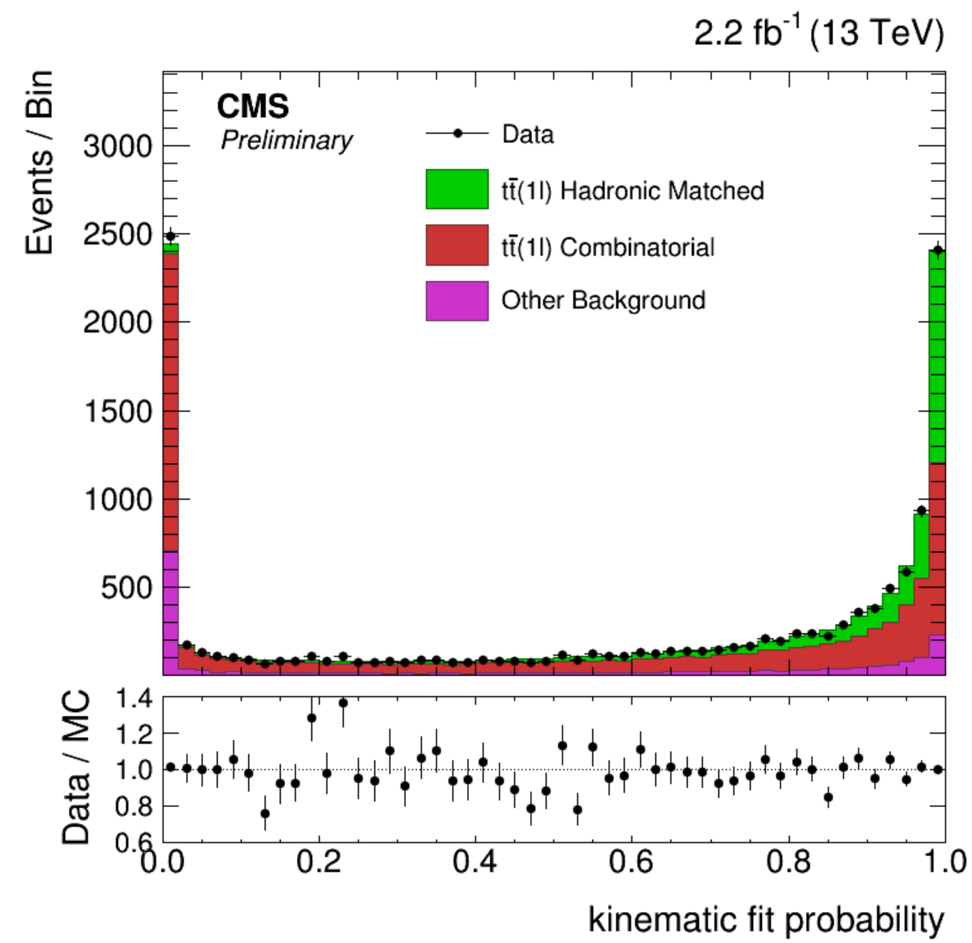




Inputs (2)

- Kinematic fit probability
  - jet triplet 4-vectors free to float within experimental uncertainties to best satisfy m_{top} and m_W constraints
  - Least squares method: probability translated from χ^2 of fit



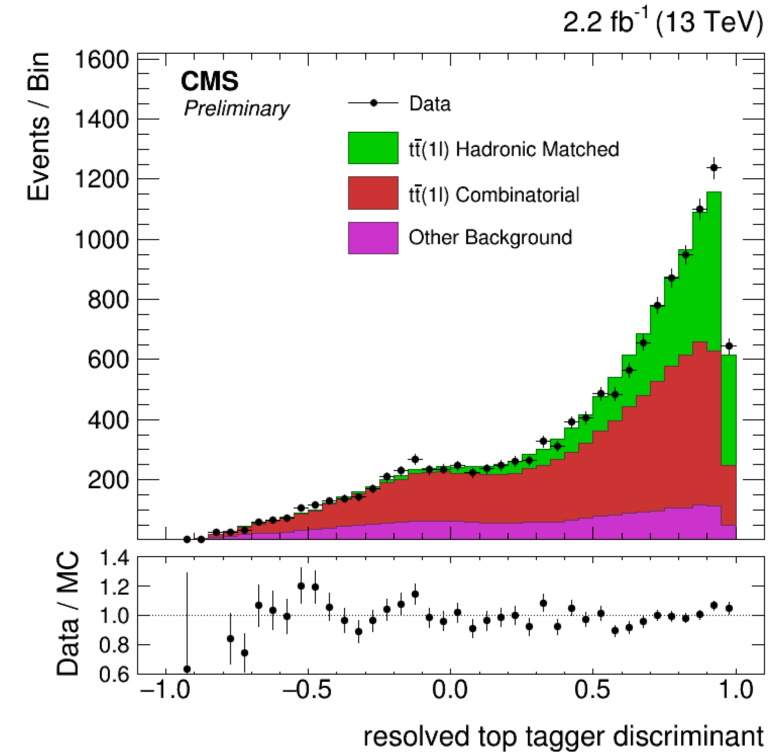


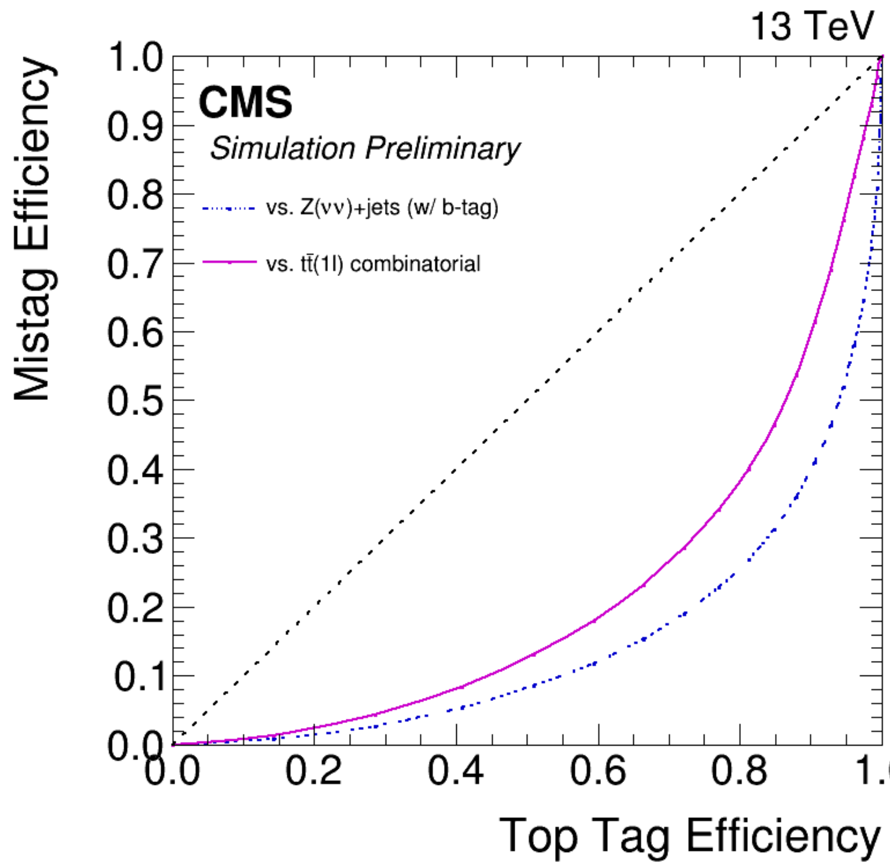
- ✦ Most powerful input variable to RTT training
- ✦ 2015 data well described by simulation

Expected Performance

- Observe good level of agreement between 2015 CMS data and simulation for RTT discriminant
  - 3 efficiencies in data & MC: signal, tt(1ℓ) combinatorial background, non-tt(1ℓ) background
  - measured for working point **RTT > 0**

Split by signal and backgrounds





- Expected performance in simulation
  - signal combination is a matched jet triplet from hadronic top decay in semileptonic tt sample
  - background is highest RTT scoring combination containing at least one jet not matched to a hadronic top decay quark

Characterization Procedure (1)

- Efficiencies measured using 2 step procedure:

  (1) **non-tt(1ℓ) bkg** efficiency measured in **dileptonic tt** control region

  (2) **signal** and **tt(1ℓ) combinatorial bkg** efficiencies measured in **semileptonic tt** sample

Characterization Procedure (2)

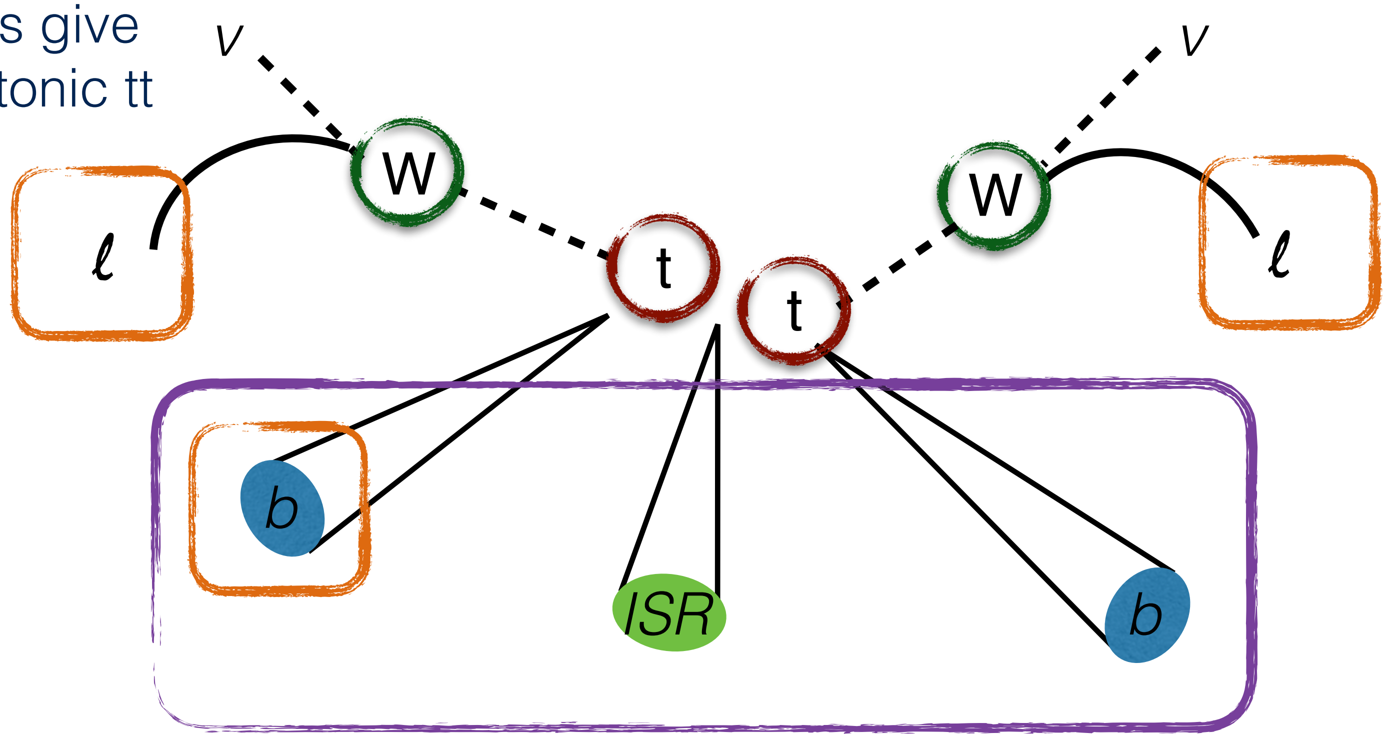
- Efficiencies measured using 2 step procedure:

(1) non-tt(1ℓ) bkg efficiency measured in dileptonic tt control region

- **provides sample with similar composition as Z(νν)+jets for fake rate measurement**

Control region selection

- object requirements give high purity of dileptonic tt



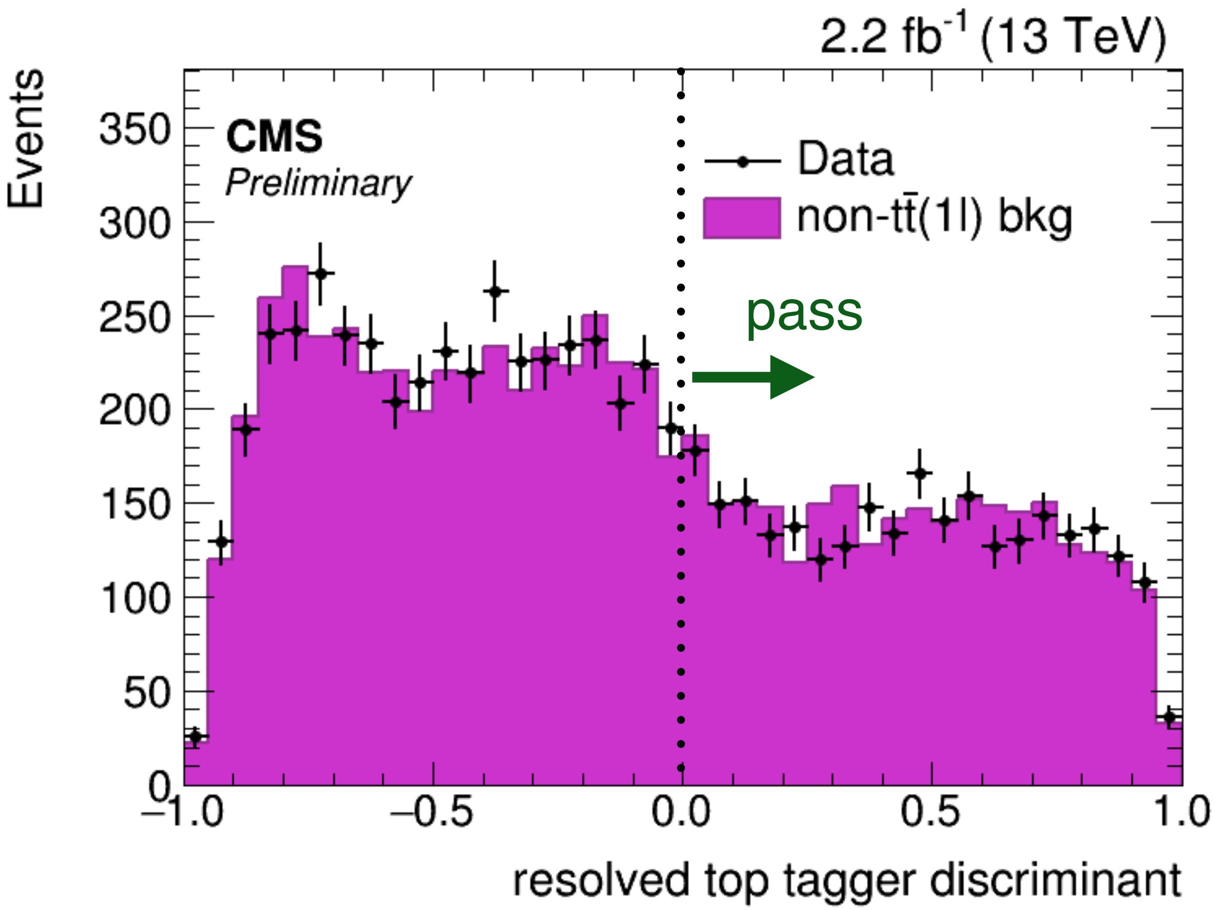
Jet triplet for non-tt(1ℓ) efficiency measurement

Characterization Procedure (3)

- Efficiencies measured using 2 step procedure:

(1) non-tt(1ℓ) bkg efficiency measured in dileptonic tt control region

- **provides sample with similar composition as Z(vv)+jets for fake rate measurement**



- Count the fraction of events passing RTT discriminant > 0 in data & MC to derive scale factor for non-tt(1ℓ) bkg

Characterization Procedure (4)

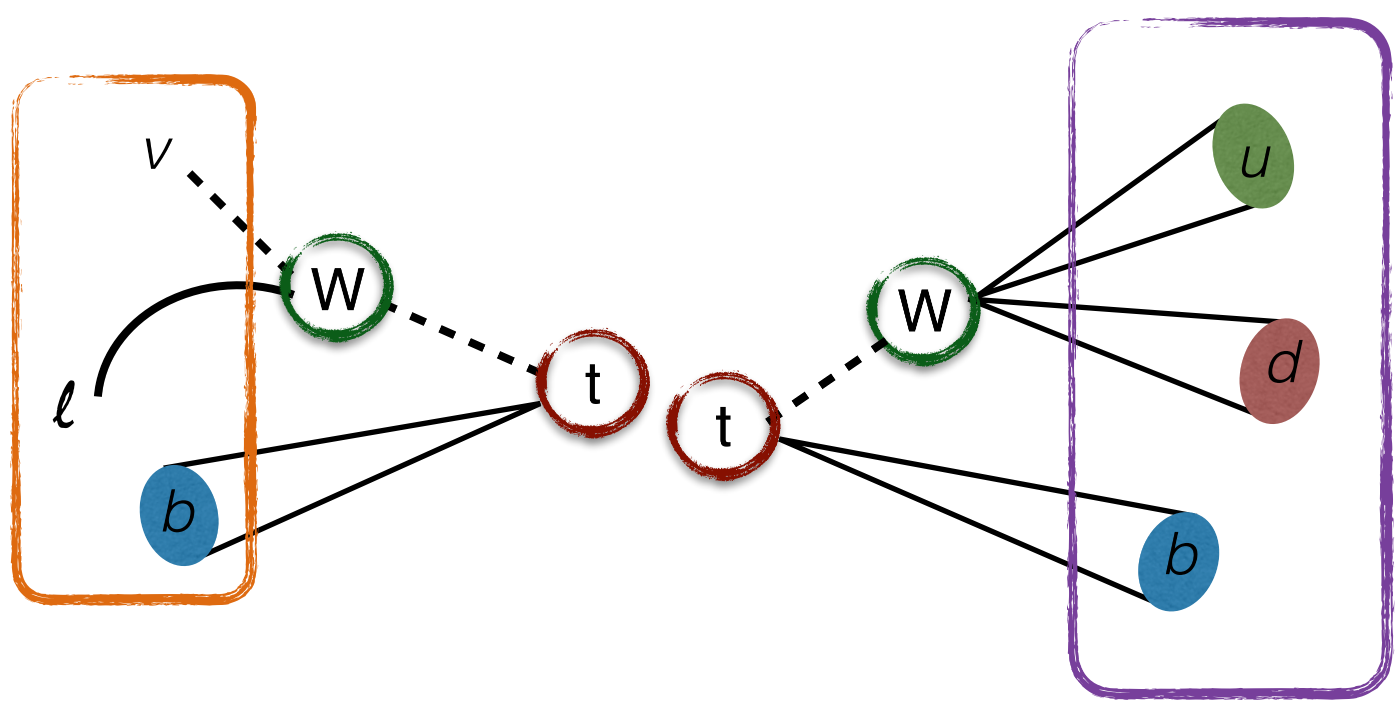
- Efficiencies measured using 2 step procedure:

(1) non-tt(1ℓ) bkg efficiency term measured in dileptonic tt sample

(2) signal and tt(1ℓ) combinatorial bkg efficiency terms measured in semileptonic tt

- **"tag" leptonic top decay, look at "probes"**

$E_T^{miss} > 40$ GeV



"Tag"

- object requirements give high purity of semileptonic tt

"Probe"

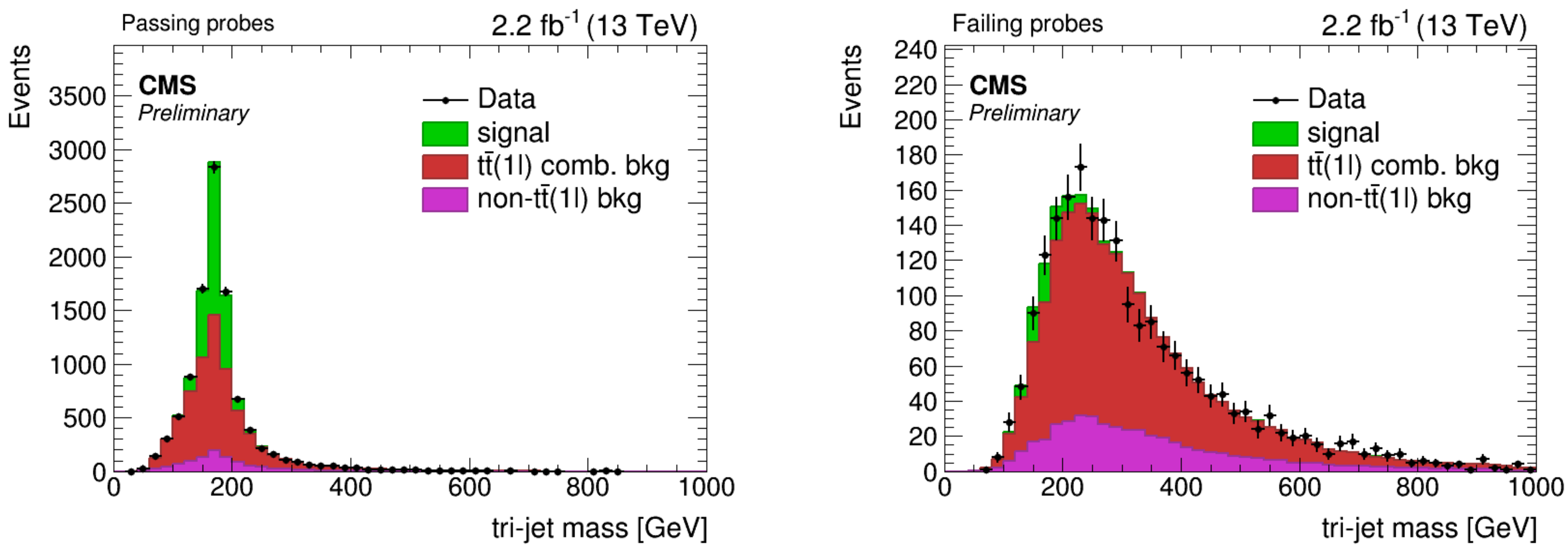
- highest scoring jet triplet in the event

Characterization Procedure (5)

- Efficiencies measured using 2 step procedure:

(1) non-tt(1ℓ) bkg efficiency term measured in dileptonic tt sample

(2) signal and tt(1ℓ) combinatorial bkg efficiency terms measured in semileptonic tt



- MC mass templates constructed from tt(1ℓ) signal, tt(1ℓ) background and non-tt(1ℓ) background events
  - non-tt(1ℓ) bkg efficiency constrained from step 1
- Simultaneously fit reconstructed top mass (tri-jet mass) in orthogonal passing and failing "probe" samples to extract efficiencies

Evaluating Systematic Uncertainties

- **Jet Energy Scale and Resolution:** efficiency measurement repeated where MC mass templates are convoluted with Gaussians
  - μ floats within [-2,2] GeV and σ floats within [0,4] GeV

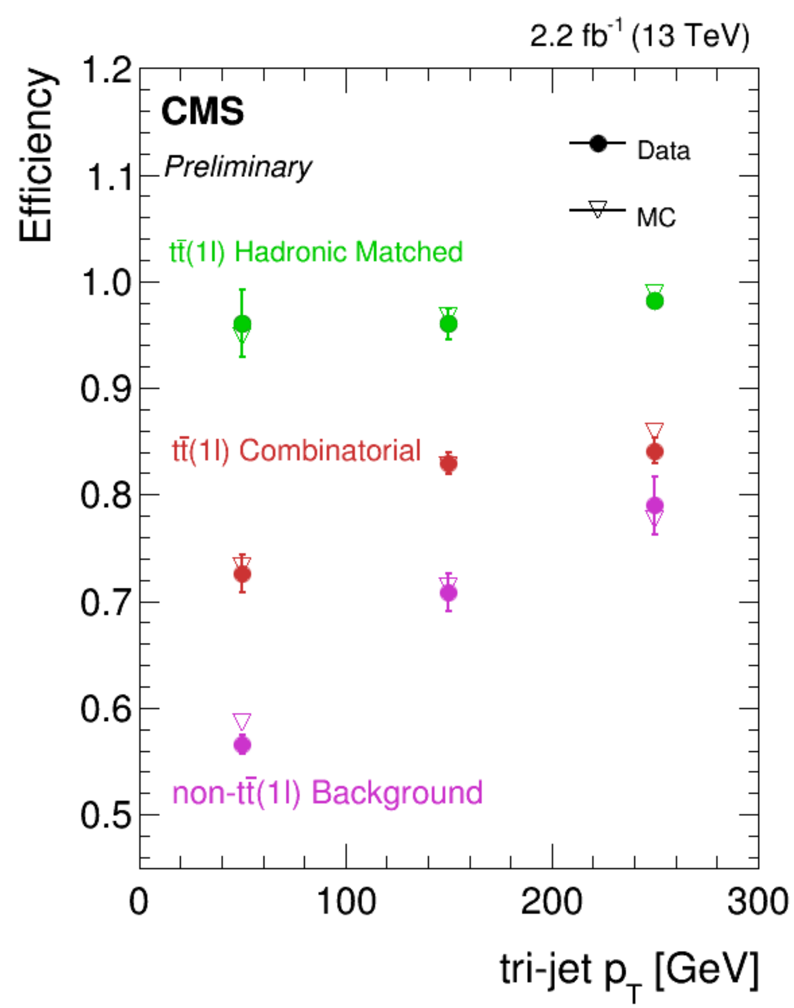
- **Showering Scheme:** TnP procedure repeated with MC mass templates from `HERWIG++` showered tt sample (nominally showered with `PYTHIA8`)

Sources of uncertainties

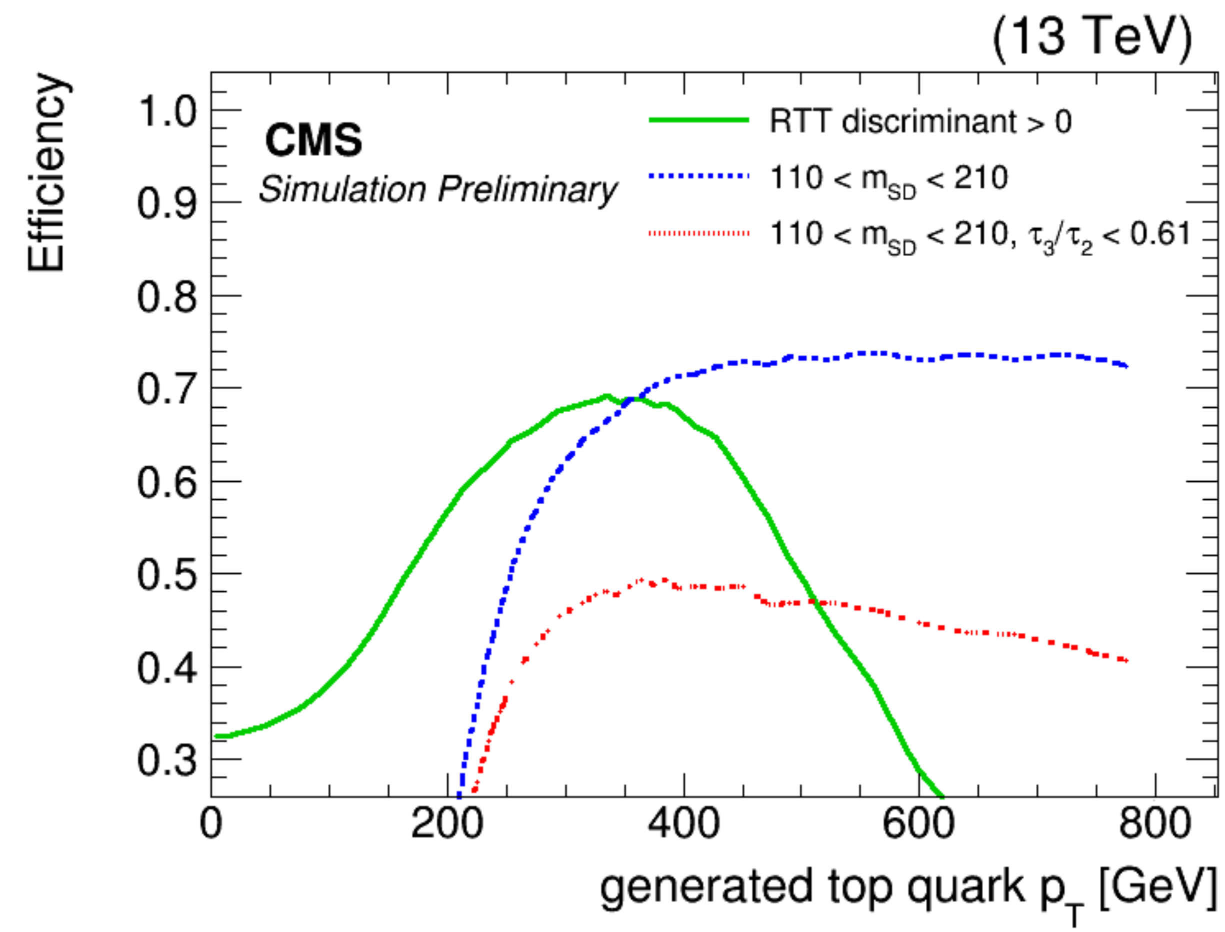
Parameter	Statistical	JEC/JER	Showering Scheme	Total
ϵ_{sig}	0.01	0.01	0.03	0.03
$\epsilon_{t\bar{t}(1l)comb}$	0.03	0.03	0.04	0.06
$\epsilon_{nont\bar{t}(1l)}$	0.01	0.03	0.01	0.03

RTT Efficiency

Efficiencies in data/MC



Signal efficiencies in MC



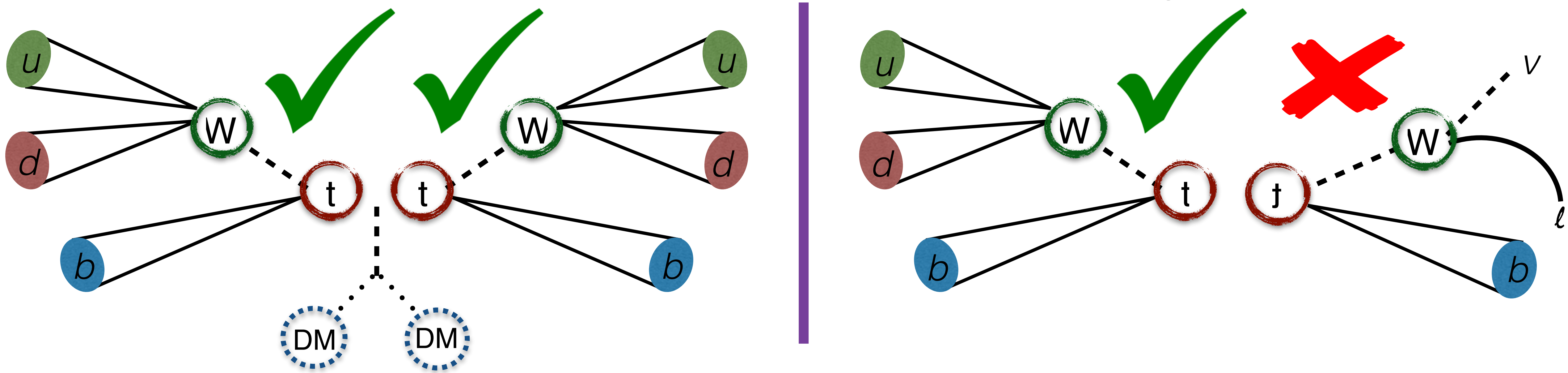
- CA R=1.5 fat jet, soft drop mass
- CA R=1.5 fat jet, soft drop mass, N-subjettiness

- Observe good agreement between efficiencies in data and simulation
- RTT covers top p_T range more suitable for tt+DM than typical boosted techniques
  - e.g. comparison with top tagging used in monotop+DM search (EXO-16-017)

Application: RTT in Action

- Employed as a means of categorization in E_T^{miss} shape analysis search for top pairs produced in association with dark matter with 2015 CMS data
  - hadronic channel categorized into events with 2 top tags and < 2 top tags

tt(had) + DM   tt semileptonic bkg



- Enables gains of up to ~30% over non-categorized strategy
  - 2 tag category allows for better rejection of tt(1ℓ)
  - < 2 tag category: $N_{jets} > 3$, $N_{bjets} > 1$
    - recover signal acceptance for soft p_T jets or merged jets in boosted tops

- Envision categorization of semileptonic channel in next iteration

Conclusions

- Developed a novel MVA-based top tagging tool
  - Performs well against backgrounds similar in composition to targeted application sample
- Efficiency measured in data and simulation
  - 2015 data is well described by simulation
  - Major systematics assessed: JES/JER and showering scheme
- Enables significant gains in tt+DM search
- Potential for wide application in top-philic BSM searches
  - Complimentary to taggers developed for boosted tops
  - Extension to other searches underway

Back Up

Tag-and-Probe Details

- Selection for dilepton sample used to derive $\mathbf{SF}^{\text{non-tt}(1\ell)\text{ bkg}}$
  - two leptons passing stringent ID with p_T>30 and |η|<2.4
  - at least three jets with p_T>30 and |η|<2.4 (at least one is b-tagged)

- Selection for single muon sample used to derive $\mathbf{SF}^{\text{signal}}$ and $\mathbf{SF}^{\text{tt}(1\ell)\text{ combinatorial bkg}}$
  - exactly one muon passing stringent ID with p_T>30 and |η|<2.4
  - E_T^{miss} > 40
  - at least four jets with p_T > 30 and |η|<2.4 (at least two b-tagged)