Hadronic Top Tagging for 4-top Search in All-Hadronic and Single-Lepton Channels

Melissa Quinnan, Valentina Dutta, Huilin Qu, Loukas Gouskos
Owen Colegrove, Joseph Incandela

February 15, 2019

Introduction: 4-top Search

- We are interested in pursuing an analysis using the **full Run II dataset** for **all hadronic** and **single lepton** channels for 4-top search.
  - Goal: Make strides towards observing standard and non-standard model 4-top production
- **All Hadronic:** can be aided by top tagging and significant in combination with other channels
- **Single Lepton:** can add hadronic top tagging to legacy Run II analysis
- Happy to collaborate/combine results!

1) Introduction

2) Hadronic Top Taggers

3) Analysis Strategy

4) Feasibility Study

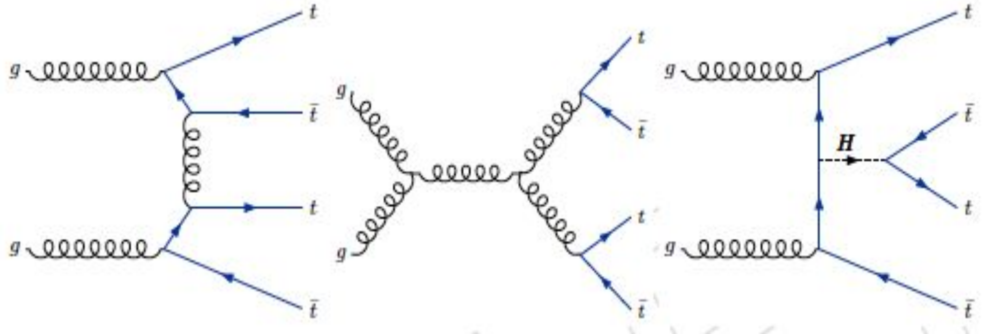
5) Background Estimation

6) Next Steps

Backup

Introduction: Physics of 4tops

- Rare Standard Model decay produced primarily by gluon fusion (can also happen via quark-antiquark annihilation) where gluons interact via a virtual boson and produce 2 tops and 2 anti-tops
- Abnormalities in tttt measurements can be indicative of BSM physics
- Signal includes leptons and met and/or jets, and b-quarks: very similar to expected backgrounds
- Hadronic top taggers can help here



4top production in the Standard Model through gluon fusion

1) Introduction

2) Hadronic Top Taggers

3) Analysis Strategy

4) Feasibility Study

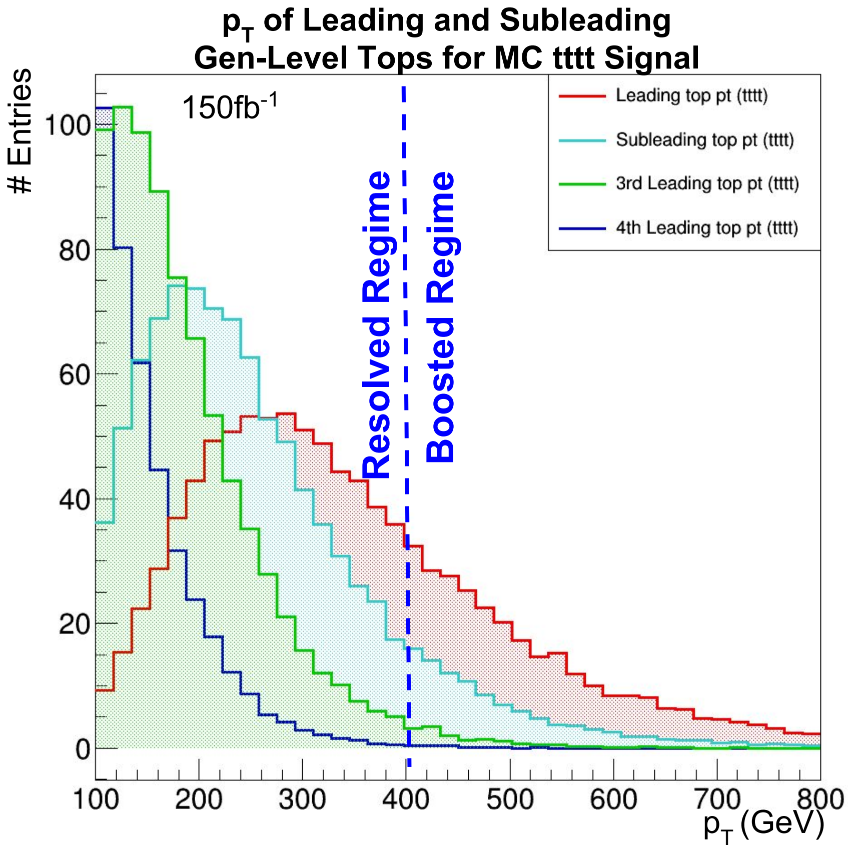
5) Background Estimation

6) Next Steps

Backup

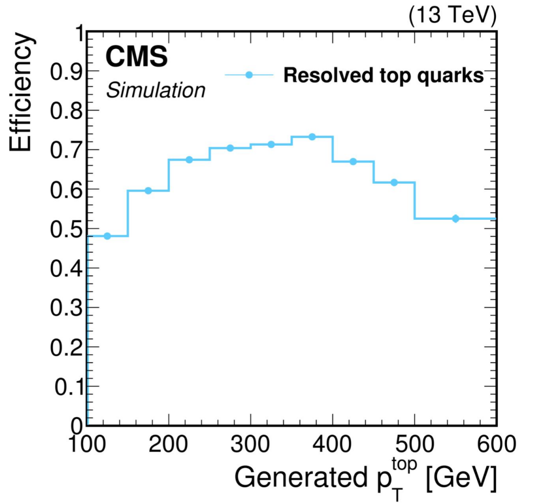
Introduction: Motivation

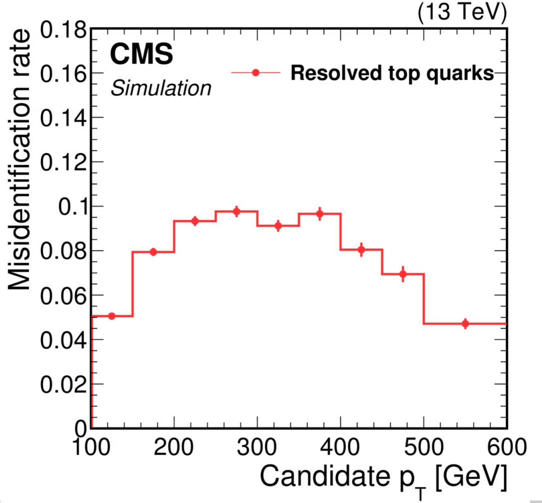
- Strategy: use hadronic top taggers to address 4-top single-lepton and all-hadronic channels
  - May help tackle difficult QCD background
  - Implement resolved top and boosted top tagger
  - Use number of boosted and resolved tops as discriminants



Resolved Top MVA

- Resolved top MVA from 2016 stop search (SUS-16-049)
- BDT to distinguish 3-jet (AK4) combinations from resolved tops against random jet combinations
    - trained using jet kinematics, jet flavor discriminants, and quark/gluon variables
- Reconstructs tops from 3-jet combinations of a b-jet and 2 jets from the W boson
- Up to 70% efficiency for ~10% mistag rate (medium working point)



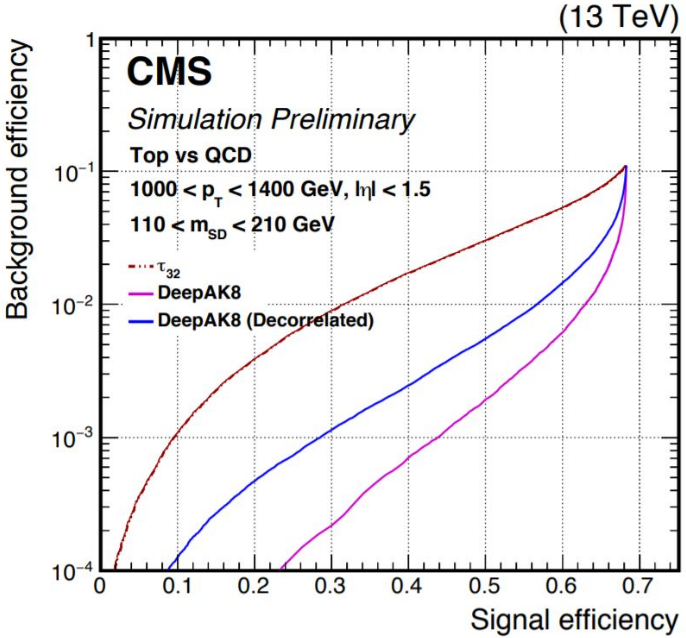


Top Figure : Efficiency in MC simulation to identify resolved top quark decays as a function of the p_T of the generated top quark.

Bottom Figure: Misidentification rate in MC simulation as a function of the p_T of resolved top quarks, in a sample dominated by the QCD multijet process.

DeepAK8 Boosted Top Tagging

- “DeepAK8” is a multi-class tagger for top, W, QCD or Higgs jets based on standard anti-kT R=0.8 (AK8) jets
- Deep neural network (DNN) with simultaneous low-level inputs:
  - Substructure info (PF candidates)
  - Flavor info (secondary vertices)
- Up to 70% signal efficiency for ~1% mistag rate (medium working point)
- CMS-DP-2017-049, DP-2017-049, AN-18-107



DeepAK8 boosted top tagging performance vs. QCD background efficiency. “Decorrelated” refers to correlation with jet mass/mass independence, and τ_{32} refers to the “traditional” CMS top tagging (ratio of n-subjettiness 3 and 2 jets)

Strategy: Selections & Key Variables

- Established baseline selections:

0-Lepton Baseline	**1-Lepton Baseline**
10+ jets, H_T>1000, no met, 3+ bjets, 1+ tagged resolved top	8+ jets, lepton pt>30, some met, 3+ bjets, 1+ tagged resolved top

- Identified set of possible discriminating variables:

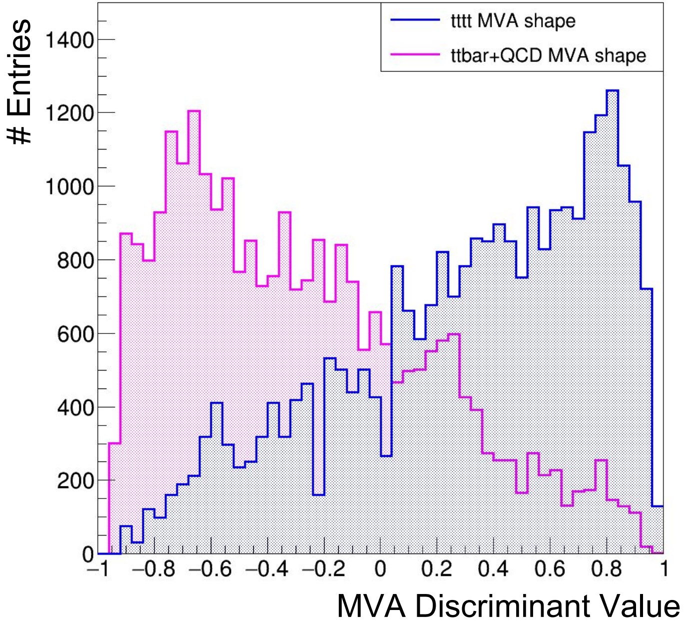
Kinematics:
  - Sum of transverse Momentum (H_T)
  - Missing transverse energy (met or E_T^{miss})
  - Number of jets (njets)
  - Number of b-jets (nbjets)
  - Sum of fat jet masses
  - Quark-gluon likelihood variables
  - Jet kinematics (pt, η, …)
  - Lepton kinematics (pt, η, …)

Top-Tagger Discriminants:
  - Number of tagged resolved tops
  - Number of tagged boosted tops

Strategy: All-Hadronic Approach

- In all-hadronic case we cut on an MVA before binning in discriminating variables:
    1. Train an MVA on kinematic variables
    2. Cut on MVA discriminant (or shape) to maximize signal/$\sqrt{\Sigma(\text{background})}$
    3. Discriminate by binning in kinematic variables in categories of resolved or boosted tops.
- Cleaned resolved against boosted tops
- Can later refine training and use MVA shape analysis rather than binning in variable distributions

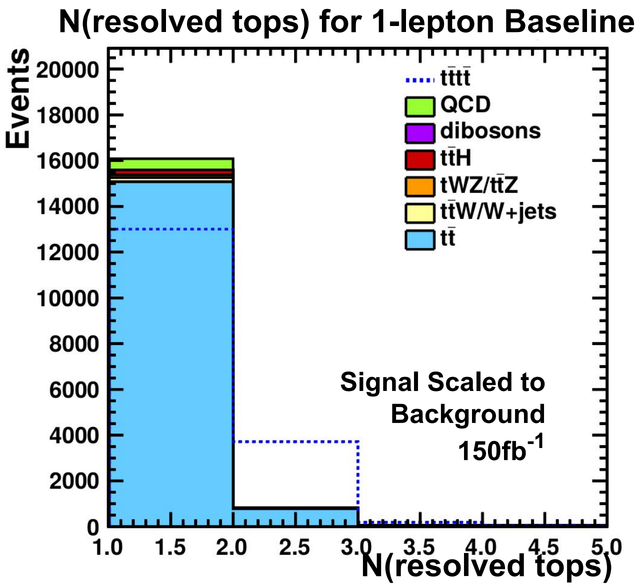
Shapes of tttt Signal and ttbar+QCD background MVA Discriminants

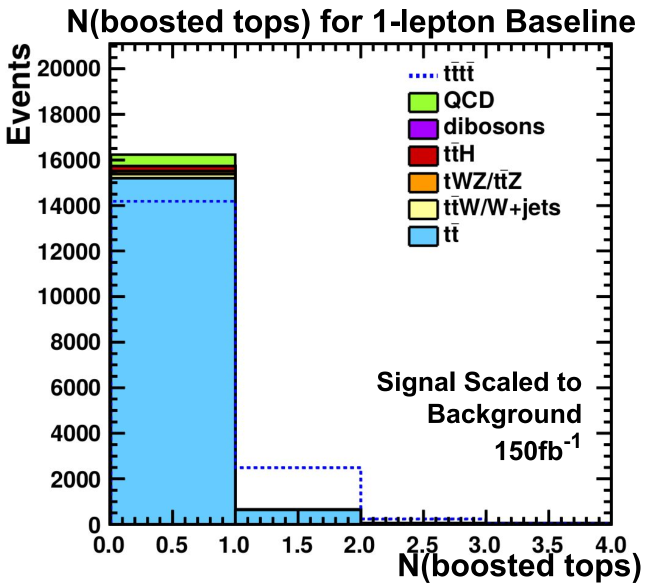


Example of MVA shape for training inputting H_T, Σ(fat jet mass), and other kinematic variables. 0-Lepton baseline. Signal scaled to background, 150fb^{-1}

Strategy: Single-Lepton Approach

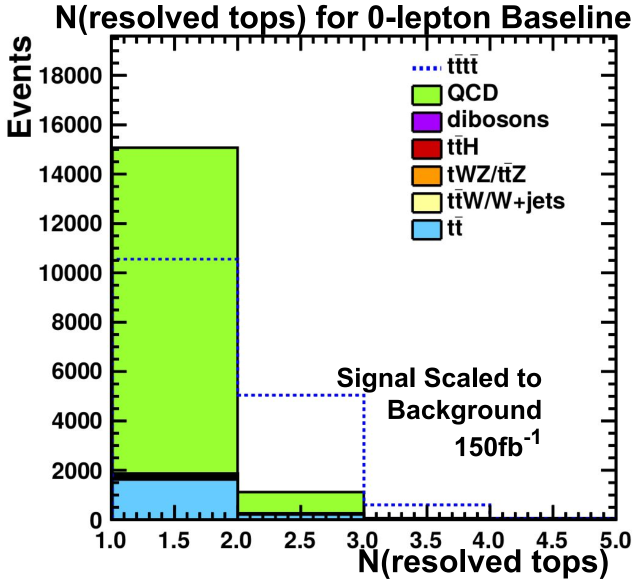
- Used a simple cut-based approach
- Discriminate based on tagged number of resolved and boosted tops and bin in kinematic variables
- Can gain from binning in number of tops

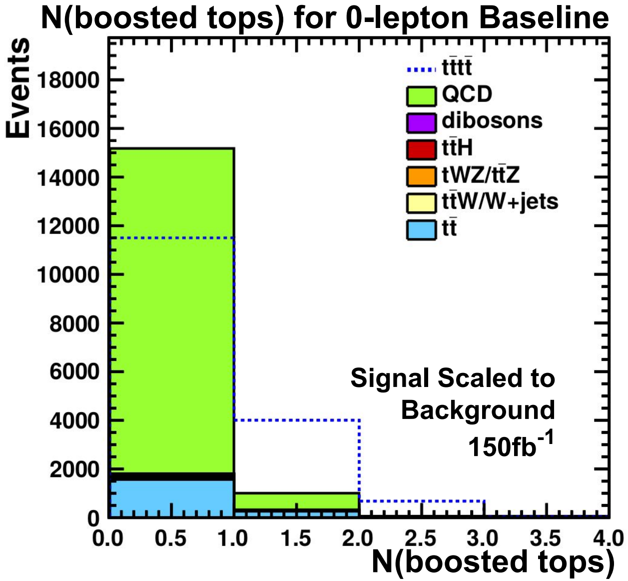




All-Hadronic Distributions

- Pre-MVA cut (0-lepton baseline) distributions for number of tagged resolved and boosted tops
- QCD is a significant background
- Can gain from binning in number of tops

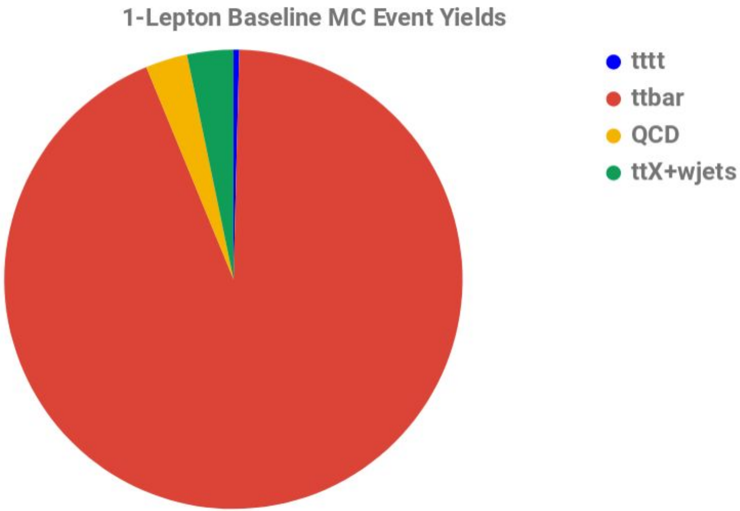




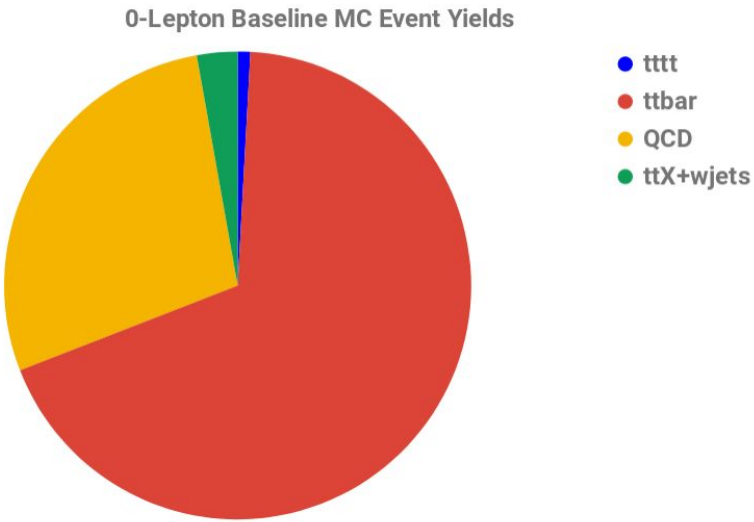
Feasibility Study: MC Event Yields

1-Lepton Baseline
8+ jets, lepton pt>30, some met, 3+ bjets, 1+ tagged resolved top

0-Lepton Baseline + MVA Cut
10+ jets, H_T>1000, no met, 3+ bjets, 1+ tagged resolved top



1-Lep baseline MC event yields for 16987 total events and 150fb^{-1}



0-Lep baseline MC event yields with cut on MVA discriminant corresponding to 50% signal efficiency for 2532 total events and 150fb^{-1}

Feasibility Study: Search Regions

- **MC-based feasibility study to test search strategy**
- Identified and tested search region (SR) bins based on discriminating variables to optimize expected limits. SRs include baseline selections.
- **Resolved/Boosted Top SR Bins**:

Resolved Top SR Bins
- N(resolved tops) = 1
- N(resolved tops) ≥2

Boosted Top SR Bins
- N(boosted tops) = 1
- N(boosted tops) ≥ 2

Combined Top SR Bins
- N(resolved tops) = 1 and N(boosted tops) = 0
- N(resolved tops) = 1 and N(boosted tops) = 1
- N(resolved tops) = 1 and N(boosted tops) ≥ 2
- N(resolved tops) = 2 and N(boosted tops) = 0
- N(resolved tops) = 2 and N(boosted tops) ≥ 1

- **Kinematic SR Bins** (2-3 bins per top bin):
  - H_T
  - N(jets)

Feasibility Study: Predicted Limits

Expected Limits/σ for SM

Predicted for Full Run II Integrated Luminosity (150fb^{-1}):

	Predicted Limits	Predicted % Improvement
All-Hadronic Channel	4.0	n/a
Single-Lepton Channel	3.0	~30%*
Combined	2.6	n/a

- From 2016 Monte Carlo based feasibility study
- *Predicted improvement with respect to TOP-17-019 scaled to full Run II
- Limits include statistical uncertainties and 20% uncorrelated systematics

Background Estimation Strategy

- **General strategy for CRs:** predict background contributions in search regions (SRs) using orthogonal control regions (CRs) for each background and corresponding scale factors
- Tentative ideas (Still refining strategy):
  - In general, CRs are the same as SRs with the listed exceptions

Single-Lepton Triggered CRs

- **ttbar 1-lepton CR:** 1 lepton, 2+ bjets, 3+ jets, isolated lepton, small ht
- **ttbar 2-lepton CR:** 2 leptons, 1+ bjets
- **ttX (Z,W,H) 2-lepton CR:** 2 leptons, dilepton pair close to invariant mass of the boson, simultaneous fit
- **ttX (Z,W,H) 3+ lepton CR:** 3 (or more) leptons required, simultaneous fit

H_T-Triggered CRs

- **QCD 0-lepton CR:** 0 leptons, 0 bjets, low met
- Look into extrapolating from 0 to some number of tagged tops using the measured fake rate for top taggers

Simulation-Based Predictions

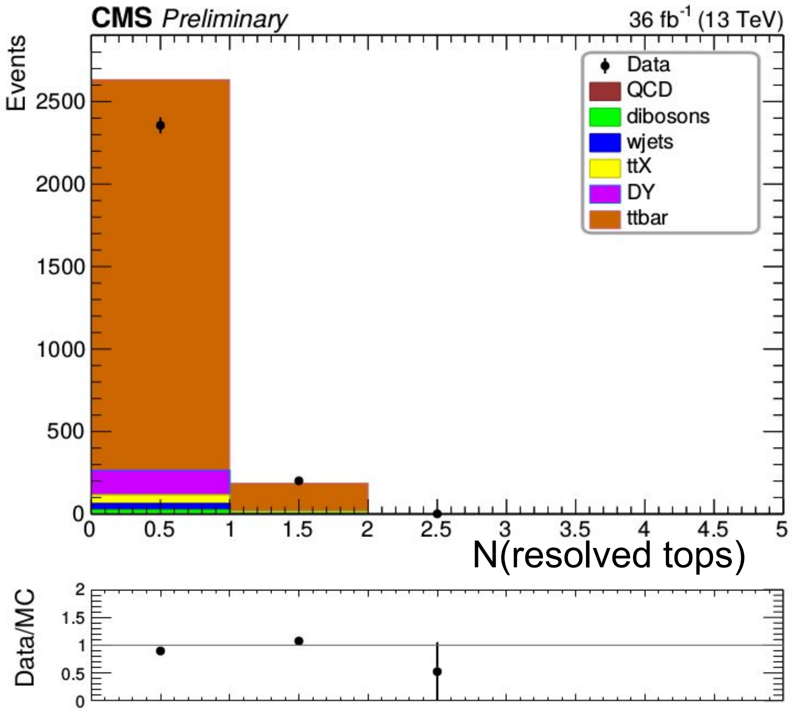
(znunu, dibosons, small contributions…)

- Use Monte Carlo

Control Regions: Di-Lepton

- **Started establishing control regions for each channel and validating against 2016 data**

N(resolved tops) in Di-Electron CR



- Example of di-lepton control region (first look)
- MC vs. 2016 single-electron triggered data
- 2lep ee CR (targeting ttbar or ttX): 2 electrons with e1 p_T>40 GeV and e2 p_T>25 GeV, H_T>500 GeV, N(jets)>2, E_T^{miss}>100GeV , N(bjets)>1
- No scale factors added yet to MC (pileup and b-tagging corrections and lepton scale factors included)
- Statistical uncertainties shown

1) Introduction

2) Hadronic Top Taggers

3) Analysis Strategy

4) Feasibility Study

5) Background Estimation

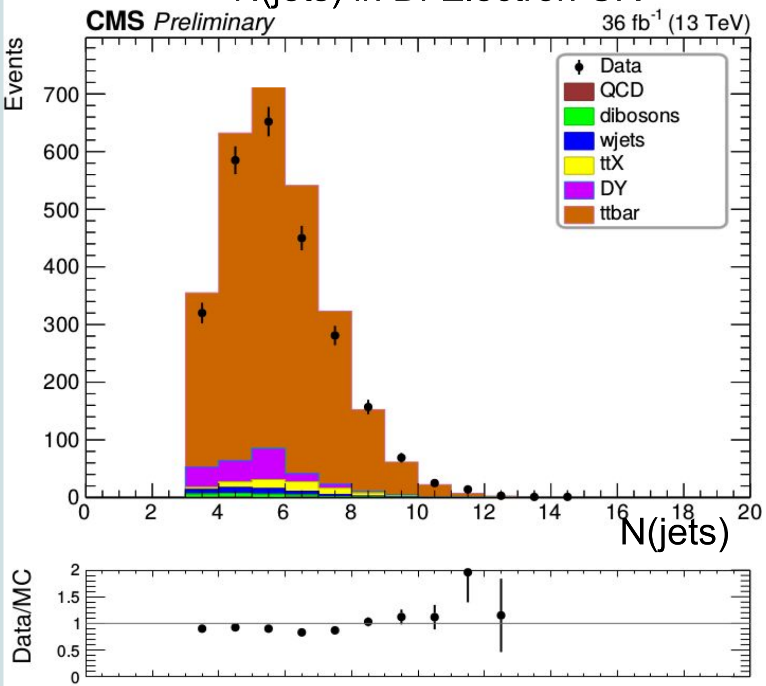
6) Next Steps

Backup

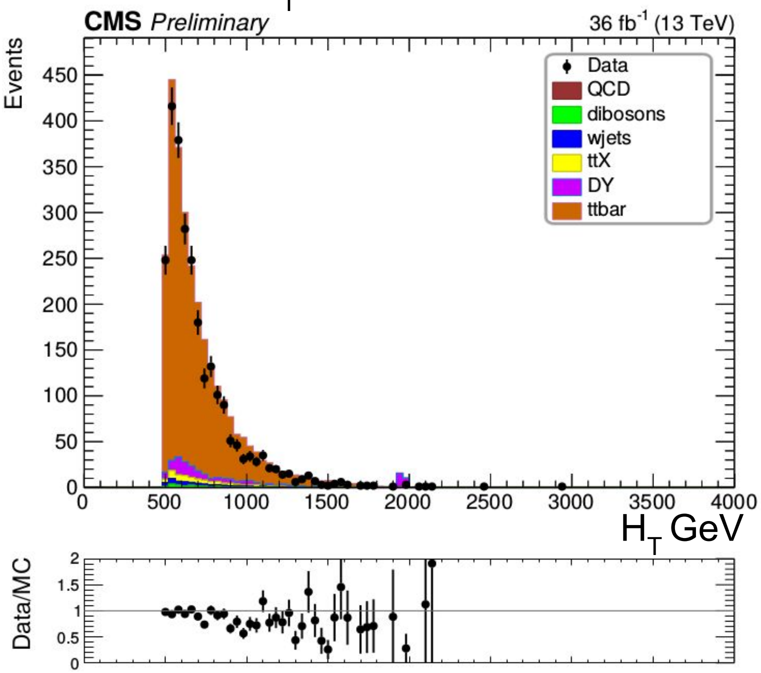
Control Regions: Di-Lepton

- **Started establishing control regions for each channel and validating against 2016 data**

N(jets) in Di-Electron CR



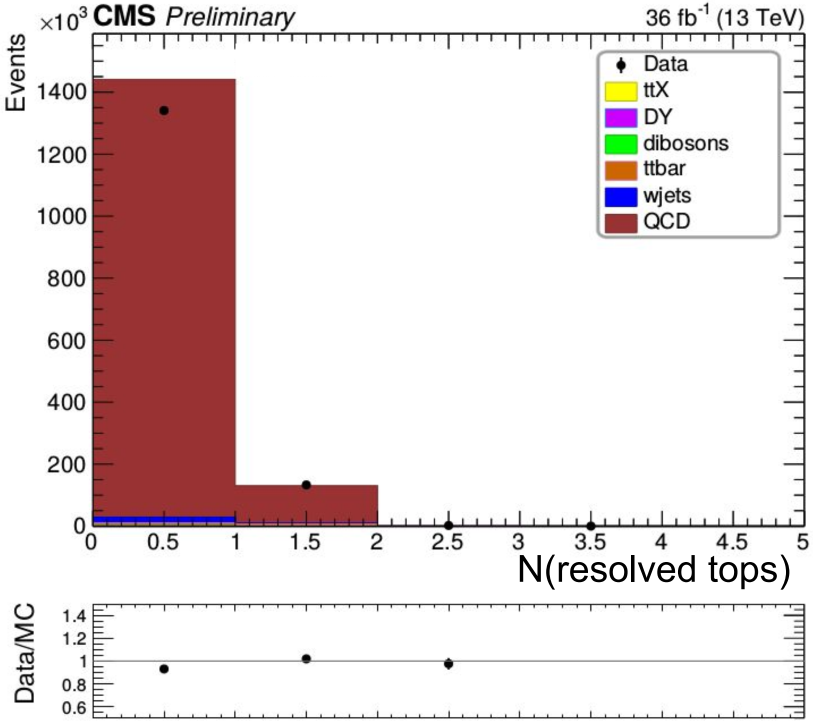
H_T in Di-Electron CR



Control Regions: 0-Lepton

- **Started establishing control regions for each channel and validating against 2016 ht data**

N(resolved tops) in 0-Lepton QCD CR

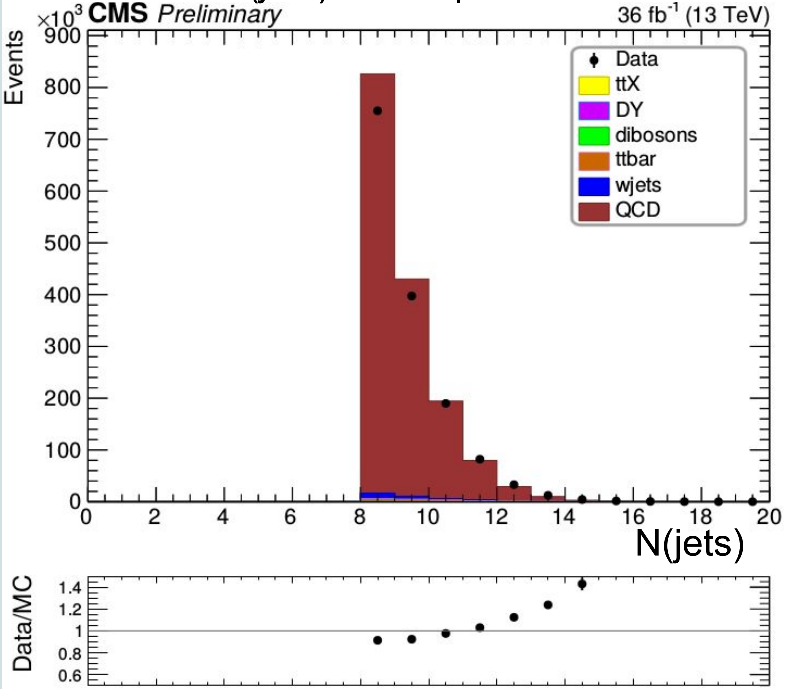


- Example of all hadronic control region (first look)
- MC vs. 2016 H_T triggered data
- 0-Lepton CR (targeting QCD): 0 leptons, N(bjets)=0, N(jets)>7, H_T>1000 GeV, E_T^{miss}<500 GeV
- No scale factors added yet to Monte Carlo (pileup and b-tagging corrections and lepton scale factors included)
- Error bars show statistical uncertainties

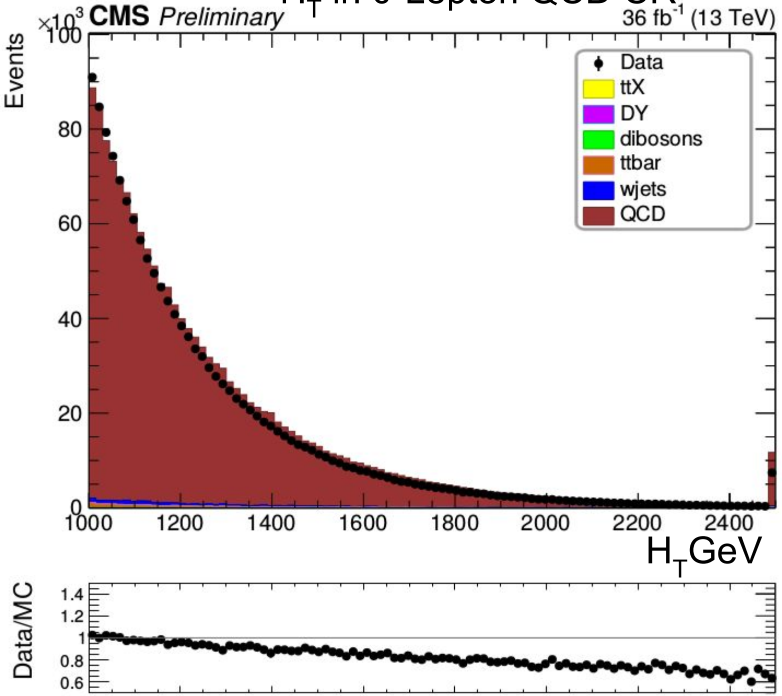
Control Regions: 0-Lepton

- **Started establishing control regions for each channel and validating against 2016 ht data**

N(jets) in 0-Lepton QCD CR



H_T in 0-Lepton QCD CR



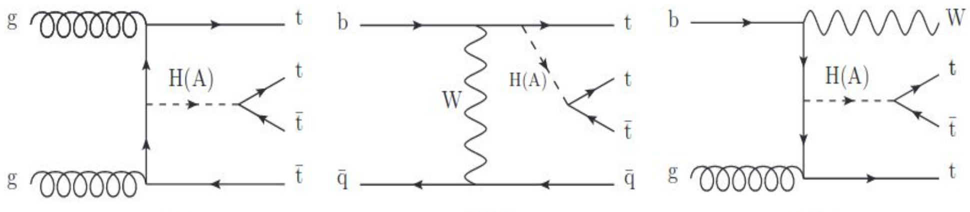
Next Steps

- Continuing to refine background estimation strategies
    - Look into 2017+2018 data
- We look forward to feedback and some discussion about how to move forward!

Thank You!

Backup

4-tops Beyond the Standard Model



(a) tt̄H (b) tHq (c) tHW

Top-associated Heavy Scalar Production Modes

Many BSM theories predict an enhancement of the 4-top cross section:

- SUSY gluino pair production
- Production of heavy scalar or psuedo-scalar boson in association with ttbar in 2 Higgs Doublet Models (2HDM)
  - Assume 2 doublets of Higgs bosons
  - Attractive source of CP violation and matter/antimatter asymmetry
  - Enhanced tttt cross section signature

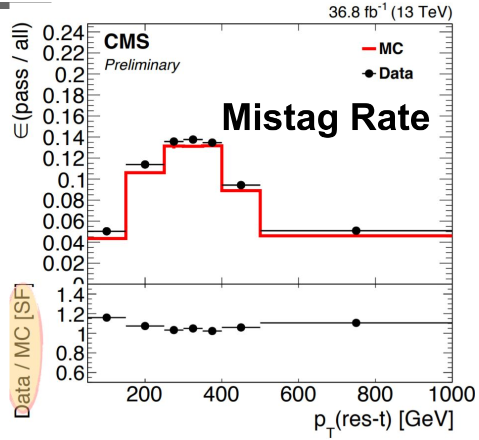
Res-Top MVA Training

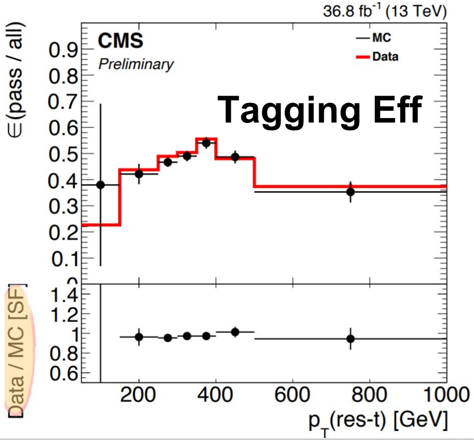
CMS PAS SUS-16-049

- **Resolved-Top Jet Candidates:** consider collection of 3 separate jets (distance parameter 0.4)
  - Contains one “b” jet with highest CSV discriminant value and 2 “W” jets (within 40GeV of W mass)
  - cleaned with a distance R>0.8 with respect to boosted Ws and boosted tops
  - Combined 3-jet system within 80GeV of top mass
- **BDT Training:**
  - Trained using simulated ttbar sample
  - Inputs:
    - Jet Kinematics: masses, angular separations, kinematics between jets
    - Quark-gluon discriminants
    - B-tagging discriminant
    - Charm-to-light discriminant (CMS-PAS-BTV-16-001)
- In high-pT regime a drop in tagging efficiency is due to identification instead of boosted tops (top quark decay products contained in R=0.8 instead and not resolved into 3 separate jets)

Res-Top MVA Scale Factors

- **Scale Factors extracted and used to correct Monte Carlo in background estimation**
- **Mistagging Rate:**
  - Used QCD dominated sample
  - HT>1 TeV, N(b-jets)>=1, no leptons
- **Tagging Efficiency:**
  - Used ttbar(1L) enhanced sample
  - 1μ, N(b-jets)>=1, MET>50 GeV, Δφ(μ,b-jet)<1, Δφ(μ, top candidate)>2, background subtracted using mistag scale factor
- Systematics were propagated to background estimation
- CMS-SUS-16-049





Res-Top MVA Systematics

Source	$t\bar{t}$/W+jets	$Z \to \nu\nu$	QCD	Rare	Signal
Resolved t-tagging					
Generator	<1%	-	-	<1%	<3%
Parton Showering	1–12%	-	-	1–16%	1–31%
Remaining sources	1–18%	1–17%	1–17%	1–16%	1–20%

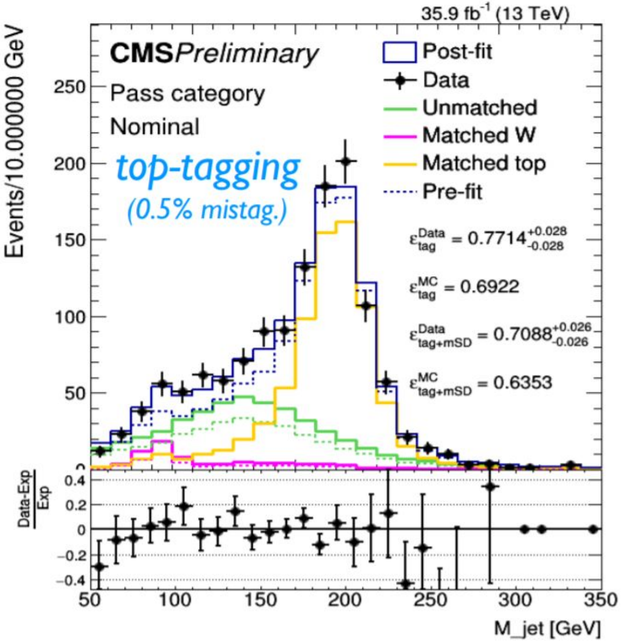
- From CMS-SUS-16-049:
  Search for direct production of supersymmetric partners of the top quark in the all-jets final state in proton-proton collisions at 13 TeV

DeepAK8 Training Inputs

- Particle/Substructure Inputs:
  - Up to 100 PF candidates (Number chosen to include all candidates for ≥ 90% of the events)
  - Sorted in descending pT order
  - Uses basic kinematics, Puppi weights, etc
  - Uses properties (quality, covariance, displacement, etc.) of the associated tracks for the charged particle
- Secondary Vertices (SVs)/ Flavour Inputs:
  - Up to 5 SVs (inside jet cone) (Number chosen to include all candidates for ≥ 90% of the events)
  - Sorted in descending SIP2D order
  - Uses SV kinematics and properties (quality, displacement, etc.)

DeepAK8 Scale Factors

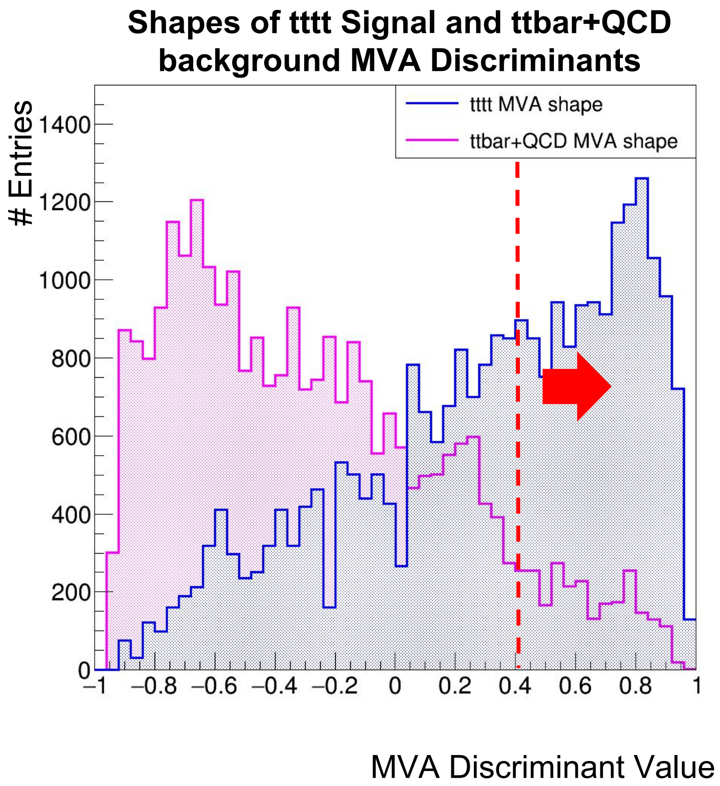
- Performance for top tagging tested in data with a ttbar dominated sample:
  - 1 tight muon (pT>45 GeV, |η|<2.1), MET>50 GeV, N(jets)(AK4)≥2, N(b-jets)(tight)≥1
  - select highest pT AK8 jet opposite to the muon as the candidate
  - define three mass templates: top-matched, W-matched and unmatched
  - simultaneously fitting the "pass" and "fail" categories to extract SFs for the tagging efficiency



Data/MC = 1.12 ± 0.04 *

Strategy: All-Hadronic MVA

- **MVA:** TMVA
- **Signal:** tttt
- **Background:** QCD and ttbar (inclusive)
- **Baseline:** 0 leptons, at least 1 resolved top, ht >=1000, at least 10 njets and 3 bjets
- **Training:** met, qglsum (sum of quark-gluon likelihoods), nsdw (number boosted Ws), metovsqrtht (met/sqrt(ht), nbjets, ht, and sumfatjetmass (or puppijet), number of ResTop MVA restops, and number of DNN boosted tops
- Cut on my 4tmva signal efficiency that maximized signal yield/sqrt(bkgd yield) (ie. 30%, 45%...)
  - This ended up being a cut of 50%, or 0.42 on the MVA discriminator



Strategy: Expected Systematics

- For now, we assumed 20% systematics
- We expect systematics consistent with data-driven control regions and the top taggers used
  - Statistics will of course be a source of uncertainty
  - Systematics from resolved top MVA and DeepAK8
  - Applying normal scale factors, jet id, pileup, lepton scale factors etc.

DeepAK8 ROC Curve