Advanced Analysis Techniques in the Search for Production of a Higgs boson in association with Top Quarks at CMS

Jason Slaunwhite
On behalf of
the CMS collaboration

Mass Hierarchy

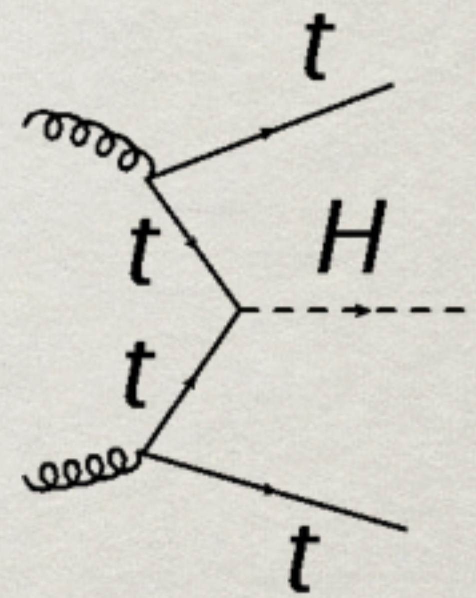
- One of the biggest questions remaining in the standard model:
  - Why do the electron and the top quark have such different masses?
- Top-Higgs coupling measurement is an important step in
  - Accessible via ttH production

Overview of this talk

- In this talk, we will see that TTH production is a challenging measurement because:
  - Signal production rate is small compared to backgrounds
  - Uncertainties are large
  - No single variable gives great discrimination
- We can overcome these issues using multivariate analysis techniques:
  - To identify the objects associated with ttH decay with high efficiency and purity
  - To distinguish ttH events from background

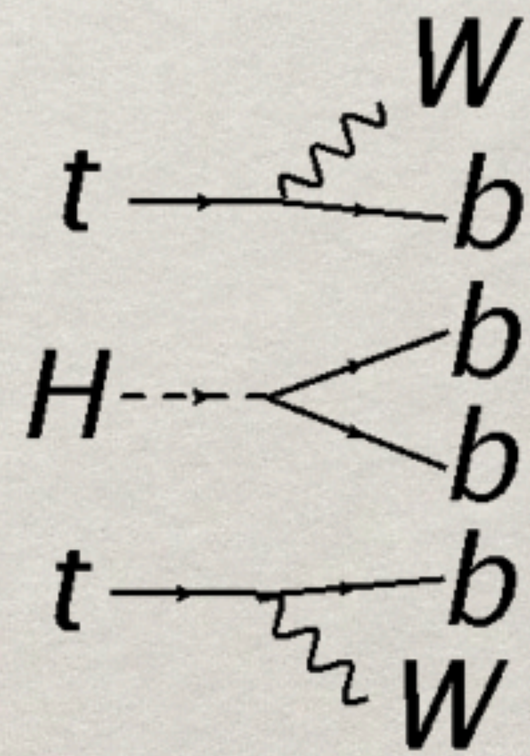
Signal Process

- Production: ttH
- Cross section: 130 fb at M=125 GeV and 8 TeV
- Focus on
  - H to bb (largest BR, 58%)
    - σ x BR(H to bb) = 75 fb
- Final state:
  - WWbbbb
- We require >=1 W to e,μ
  - 1 lepton and up to 6 jets.
    4 jets come from b-quarks.
  - 2 leptons and up to 4 jets.
    All 4 jets come from b-quarks.



Production

130 fb



Decay

Background Processes at 8 TeV

Compare to Signal
WWbbbb

- WWbbbb: **tt+bb**
  - ~2-4 pb
  - irreducible, ~24x larger than signal σ x BR(H to bb)
- WWbb+>=0jets: **tt+jets**
  - 234 pb
  - fewer jets/ fewer tags, ~3000x larger than signal
- **Single top, Dibson, W/Z+jets**
  - Many fewer jets and tags
- Classify events according to jets and tags

Object definitions

Electrons from W

Tight

- pT > 30 GeV
- eta < 2.5
- Tight Isolation
- **MVA ID**

Loose (main differences)

- pT > 15 GeV
- Loose Isolation

Muons from W

Tight

- pT > 30 GeV
- eta < 2.1
- Tight Isolation
- Tight ID

Loose (main differences)

- pT > 10 GeV
- Loose ID & Isolation

Jets from W, t, H

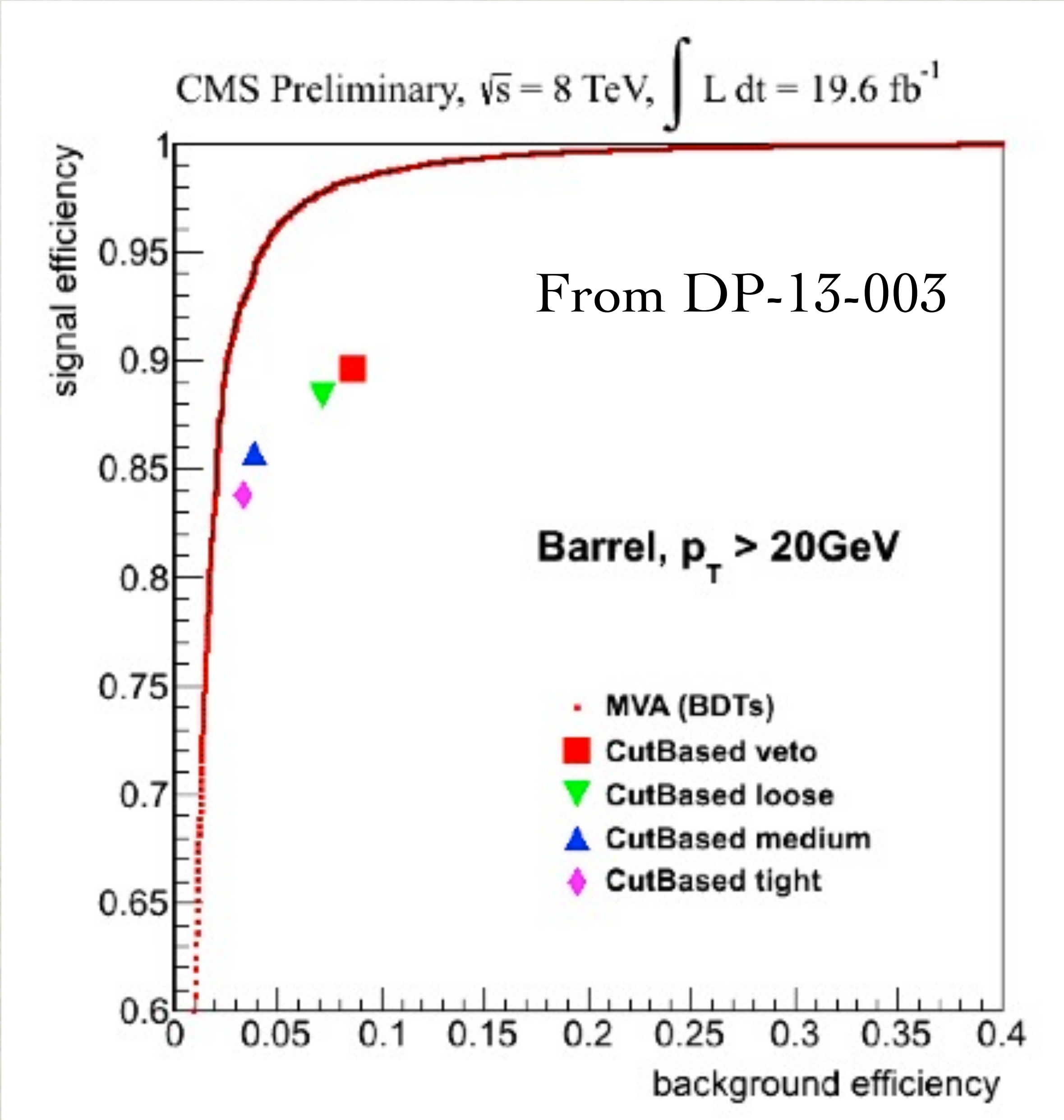
- Anti-kT size 0.5
- pT > 40 for jets 1,2,3
- pT > 30 each other jet
- Loose ID requirements

B-jets

- Pass all jet requirements
- **Combined Secondary Vertex (Medium operating point)**

Ele performance compare

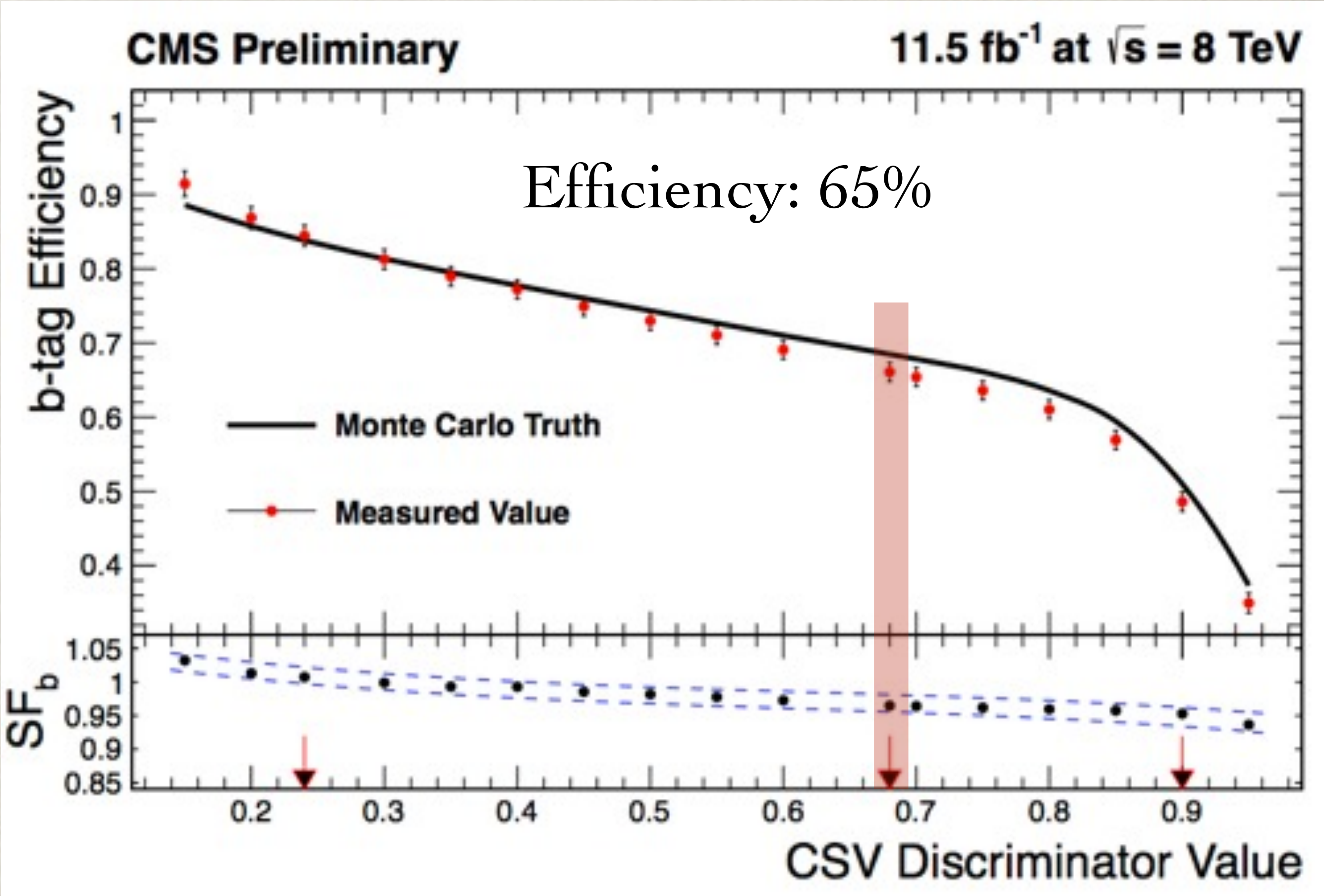
- MVA: Implemented with a Boosted decision tree
  - Trained for real vs fake electrons
- Ele MVA ID uses:
  - Tracking variables
  - Shower-shape variables
  - Geometric matching between track and calorimeter
  - Energy matching between track and calorimeter
- Has better efficiency for the same electron fake rejection



J. Slaunwhite

CSV tagger

- B-jets can be distinguished from other kinds of jets by looking for the decay of long-lived b-hadrons
  - Vertexing
  - Track impact parameter
- Combined Secondary Vertex (**CSV**) uses both
- Overcomes vertexing efficiency
- For the medium working point
  - Efficiency: 65% per jet
  - Fake rate: 1-1.5% per jet (tt+jets is 3000x larger than ttH)
  - For the same fake rate, a tagger using vertex-only information would have 55% efficiency



Fake Rate at this working point: 1-1.5%

Event categorization

- Background has fewer jets and tags, so classify events by num jets, and num tags
- Use all 9 categories in simultaneous fit

S/B Ratio - 1 tight lepton

	4jets	5jets	>=6jets
2tags	x	x	0.0031
3tags	0.0027	0.0063	0.011
>=4tags	0.028	0.037	0.040

Signal

8 TeV

S/B Ratio - 2 lepton

	2jets	>=3jets
2tags	0.0001	x
>=3tags	x	0.015

Signal

8 TeV

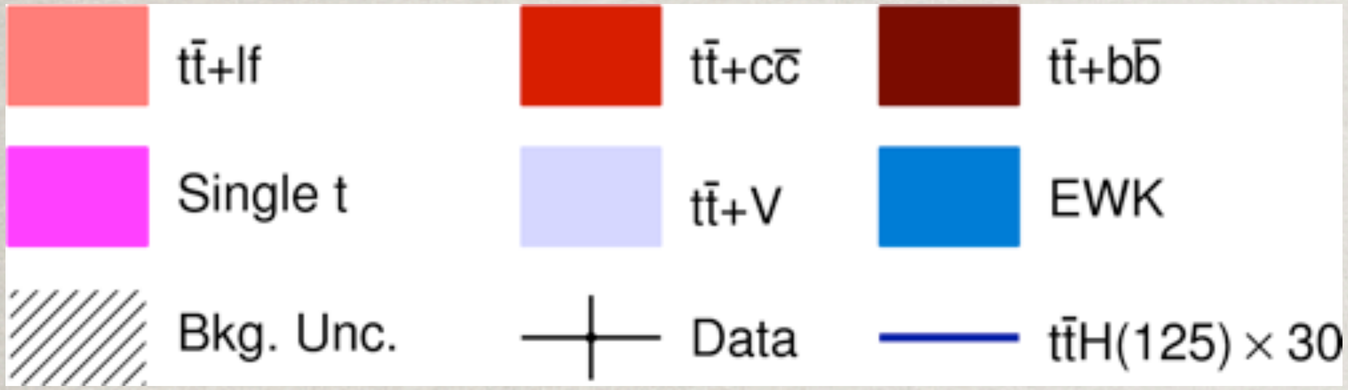
Uncertainties

- The uncertainties that have the greatest effect on the analysis are the ones that effect the number of jets/tags
  - Jet energy Scale, btag SF, mistag SF, madgraph scale
- The analysis is also sensitive to the amount of irreducible background
- Overall rate uncertainties in our prediction
- These are nuisance parameters in our fit

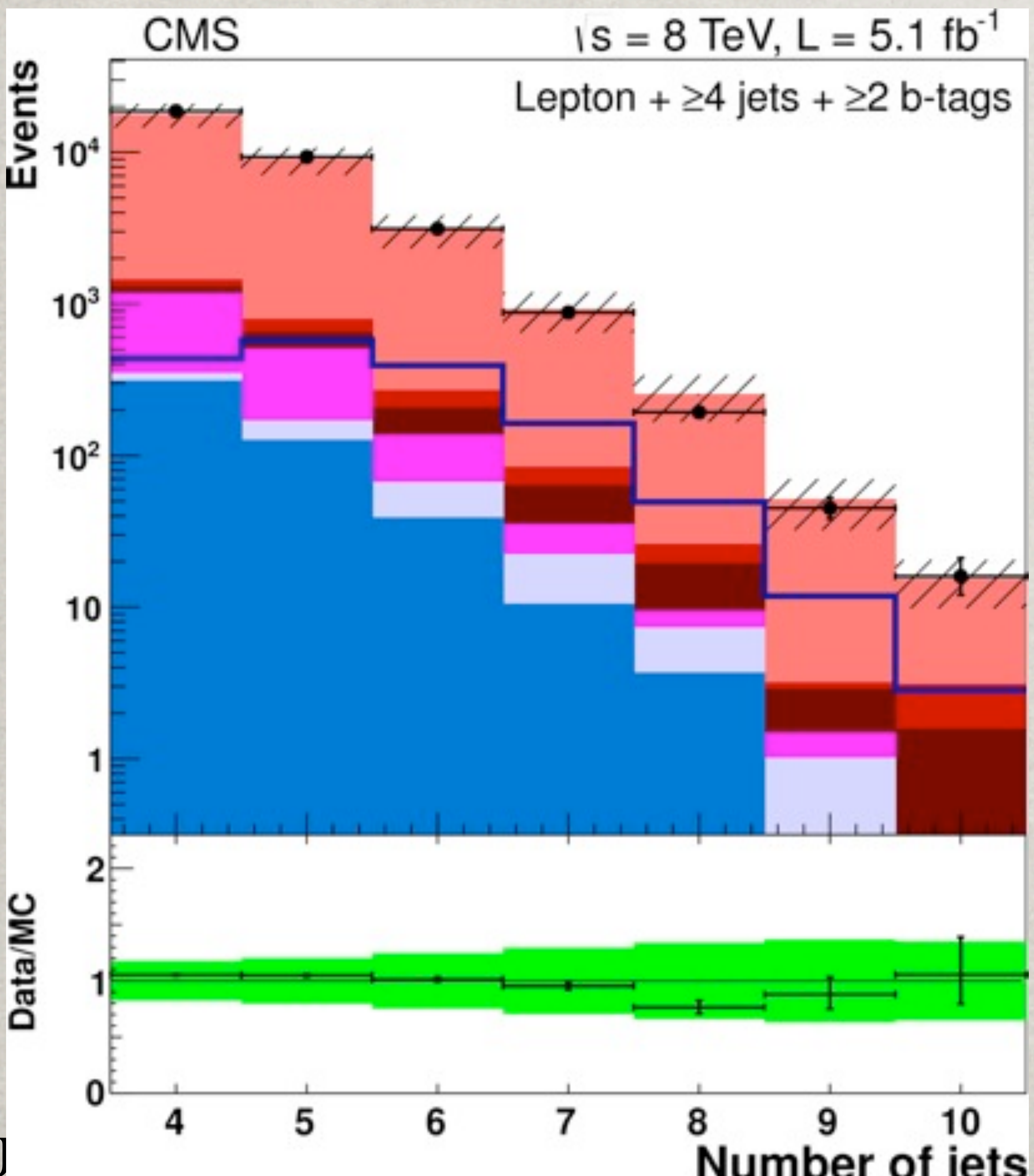
Uncertainty	Max Rate Impact
Jet Energy Scale	60%
tt+bb ONLY (theory)	50% (only tt+bb)
Btag SF	34%
Mistag SF	24%
Madgraph Scale	20%
Theory xsecs, Lumi, lepton efficiencies, etc	~15%

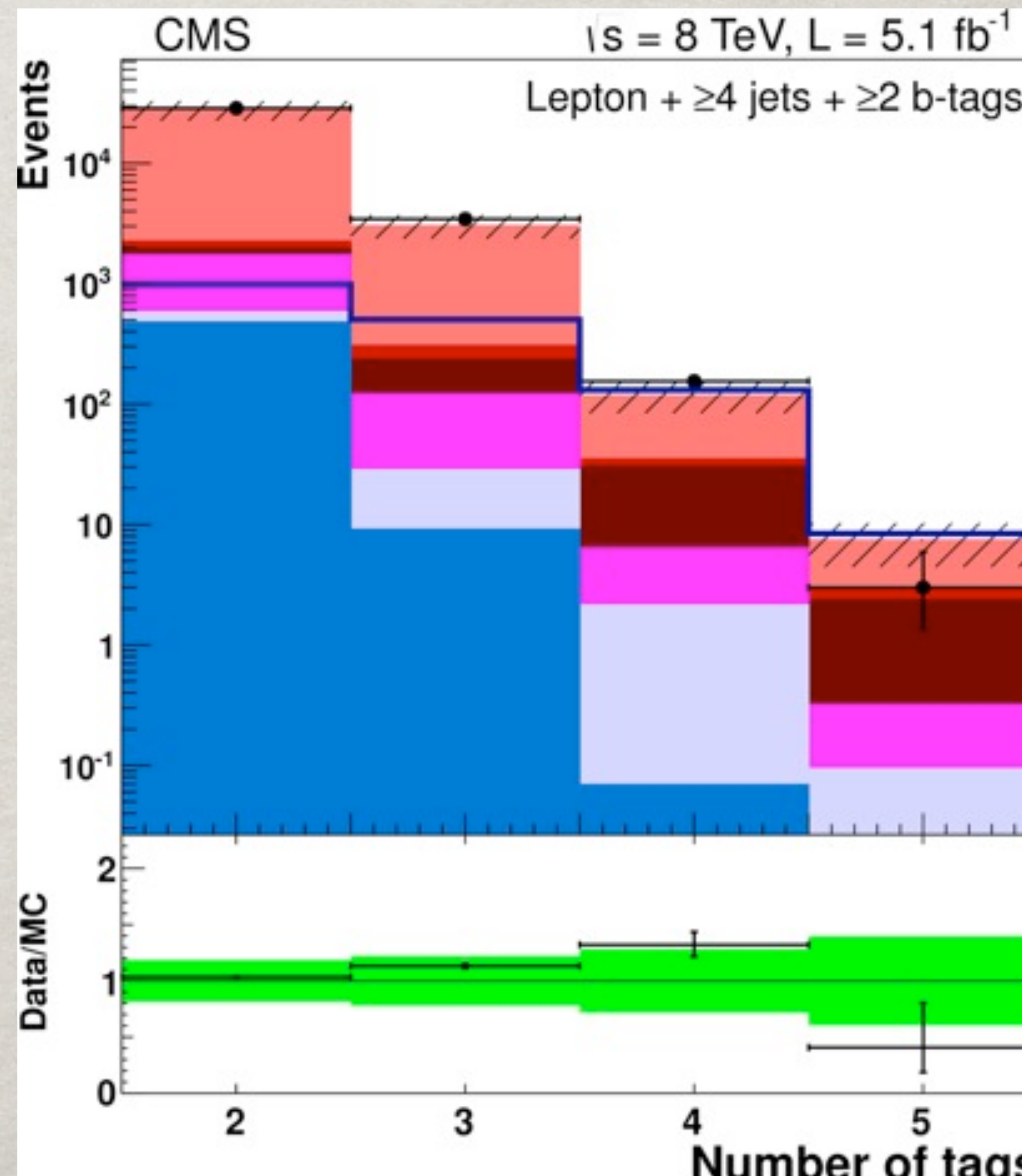
Signal size: ~ 4% of background

YIELD SUMMARY: 1 LEPTON EVENTS



Yields agree overall
Majority of background is
tt+light 65% - 90% of all events



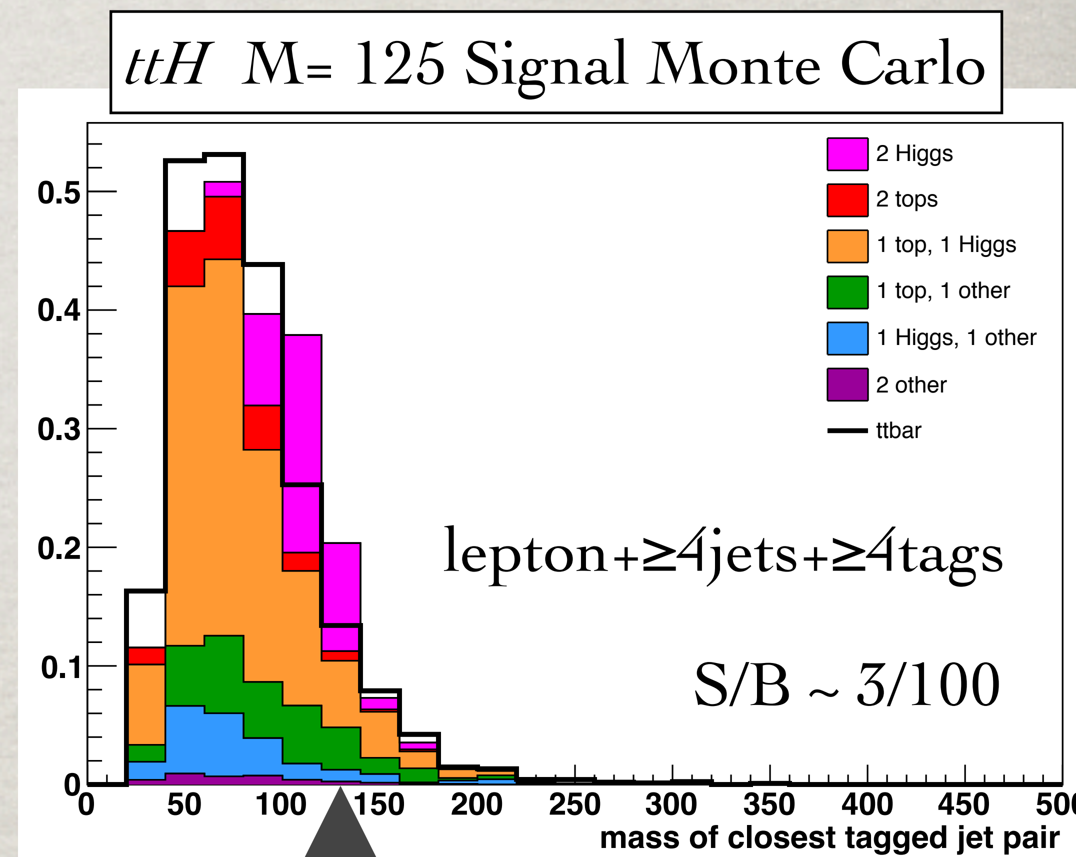


Signal Extraction Strategy

- Yield in >=6jets >=4tags
  - 2.5 Signal on background of 63 +/- 21
  - Counting experiment will not be very sensitive
- Improve sensitivity by simultaneously fitting discriminating distributions in all categories
  - Treat uncertainties as nuisance parameters in the fit
- Start by establishing a baseline using one kinematic variable in each category
- Then measure impact of combining multiple variables with an MVA technique

Higgs Mass in TTH

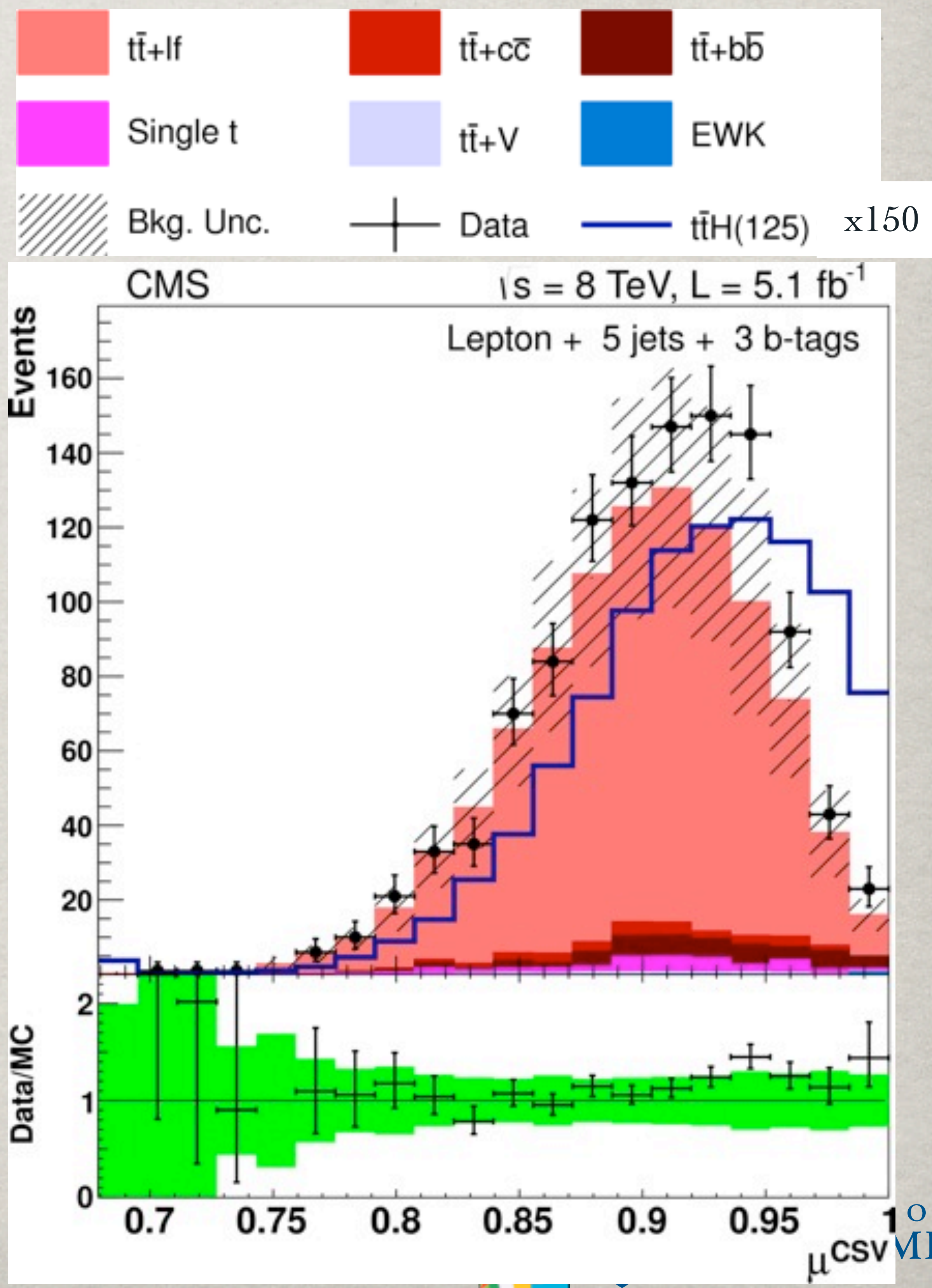
- Initially expect the Higgs mass resonance to provide distinguishing power
  - This is where discovery modes H to ZZ and H to γγ get their power
- For ttH, mass is not so powerful
  - Helps somewhat in 6 jets 4 tags, but it is not the most sensitive
- Reasons:
  - b-jet energy resolution worse than photon/e/μ energy resolution
  - Combinatorics of b-jets in final state can wash out resonance



No mass peak visible on top of combinatoric background

Performance with Best Variable

- For most categories, the average CSV value for tagged jets is the best discriminant
  - Helps reject largest background: tt + light flavor
- Fit best single variable in each category and extract upper limit on xsec
  - **6.6x SM expectation**
  - "If cross section was more than 6.6 times what we expect, then we would have seen it with this measurement"



ANN Design and Training

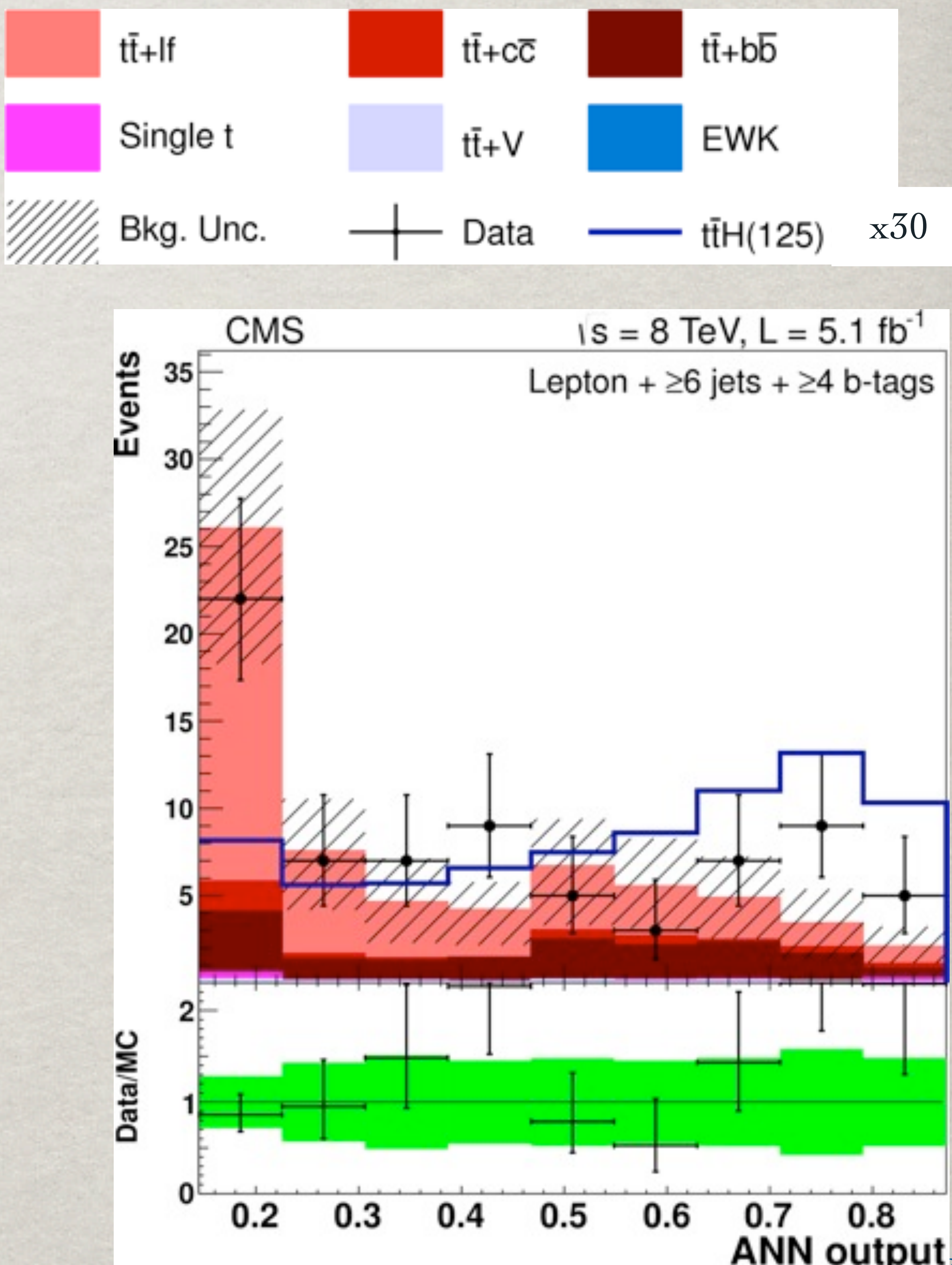
- We use Artificial Neural Networks (ANN) to combine multiple variables into a single discriminant
  - Multi-layer perceptron as implemented in ROOT and TMVA
- Create one ANN per category with own set of input variables
- Structure: N inputs, 2 hidden layers, one output
  - Hidden layer 1: N nodes
  - Hidden layer 2: N-1 nodes
- Training
  - 50% Signal = ttH, M(H)=120
  - 50% Background = tt
  - Reserved testing sample for overtraining check

Categories of variables
Kinematics of objects, single and composite
Kinematics of jet pairs
Event shape
Btag CSV discriminant

Example ANN: One lepton 6 jets and 4 tags

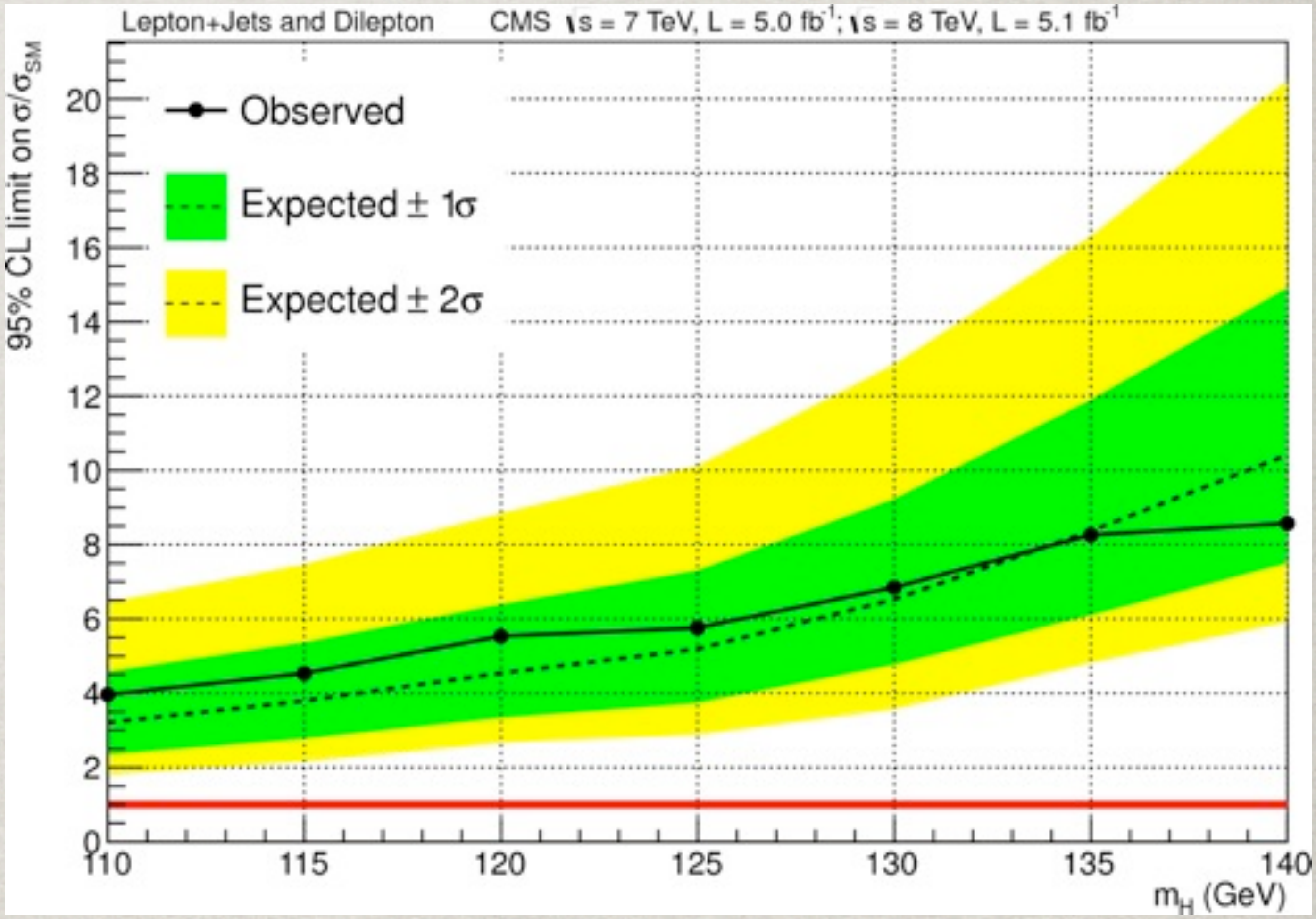
11 input variables in total

Variable	Category
Mass (lep, MET, Jets)	Kin. of composite obj
Mass (j,j) closest jets	Jet pairs
Mass (j,j) best	Jet pairs
Average ΔR(tag, tag)	Jet pairs
Minimum ΔR(lep, jet)	Shape
Sphericity	Shape
H2	Shape
H3	Shape
Average CSV*	**Btag***
2nd-highest CSV	Btag
lowest CSV	Btag



Limit Results

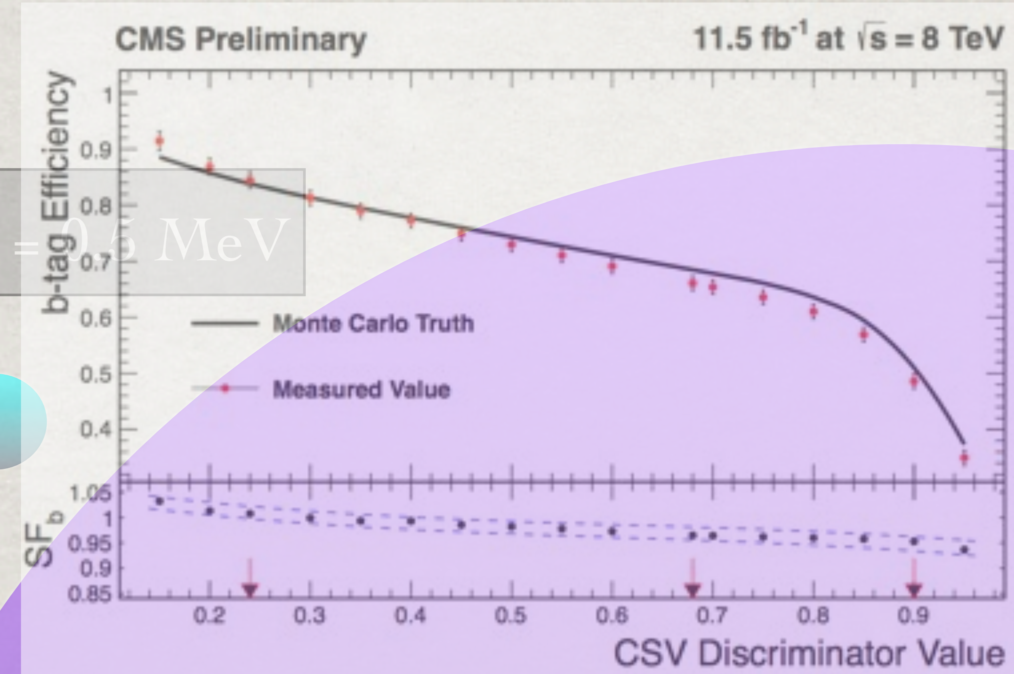
- Fit NN output distribution simultaneously in all 9 categories to extract overall limit
- **5.2xSM expectation at M=125**
- 27% improvement over single variable
- Equivalent to increasing data collected by 60%
  - Effectively 3/fb additional in this dataset
  - Effectively 12/fb on full dataset
  - Worth half a year of data taking

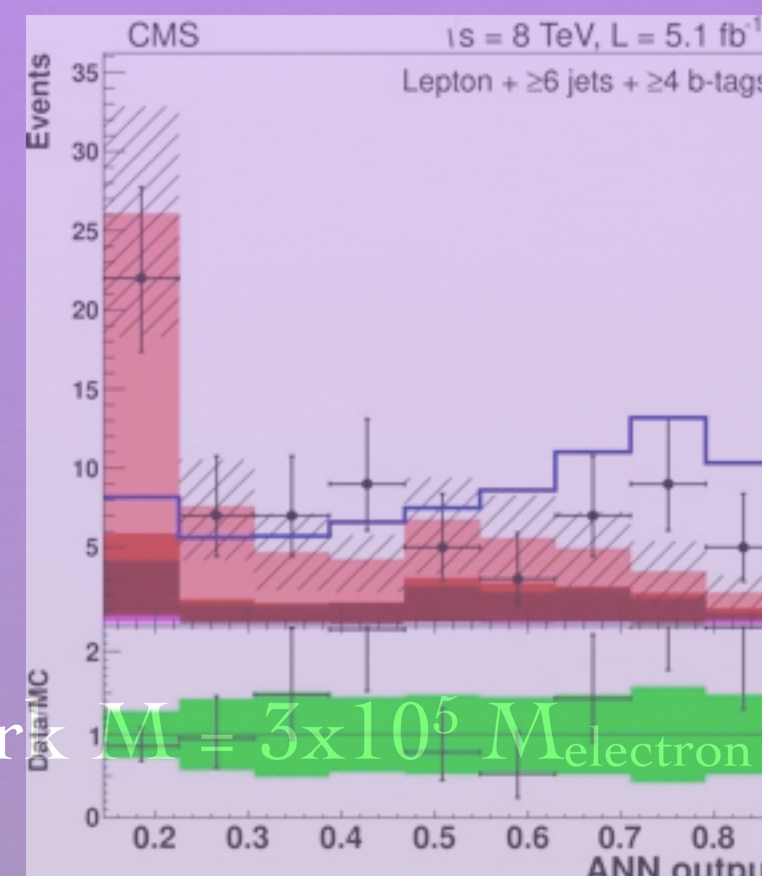


Expected @ 125: 5.2xSM
Observed @ 125: 5.8xSM

Summary

- Mass hierarchy is a compelling problem that can be explored through ttH
- Challenging: ttH cross section is small compare to the backgrounds, the uncertainties are large, and the mass resonance is not especially powerful
- Multivariate techniques help us overcome some these challenges by optimizing:
  - Object identification (b-tags, electrons)
  - Signal discrimination
- The optimizations help us get more performance out of the data we collected





Top Quark M = 3×10^5 $M_{electron}$

BACKUPS

BTAG PERFORMANCE

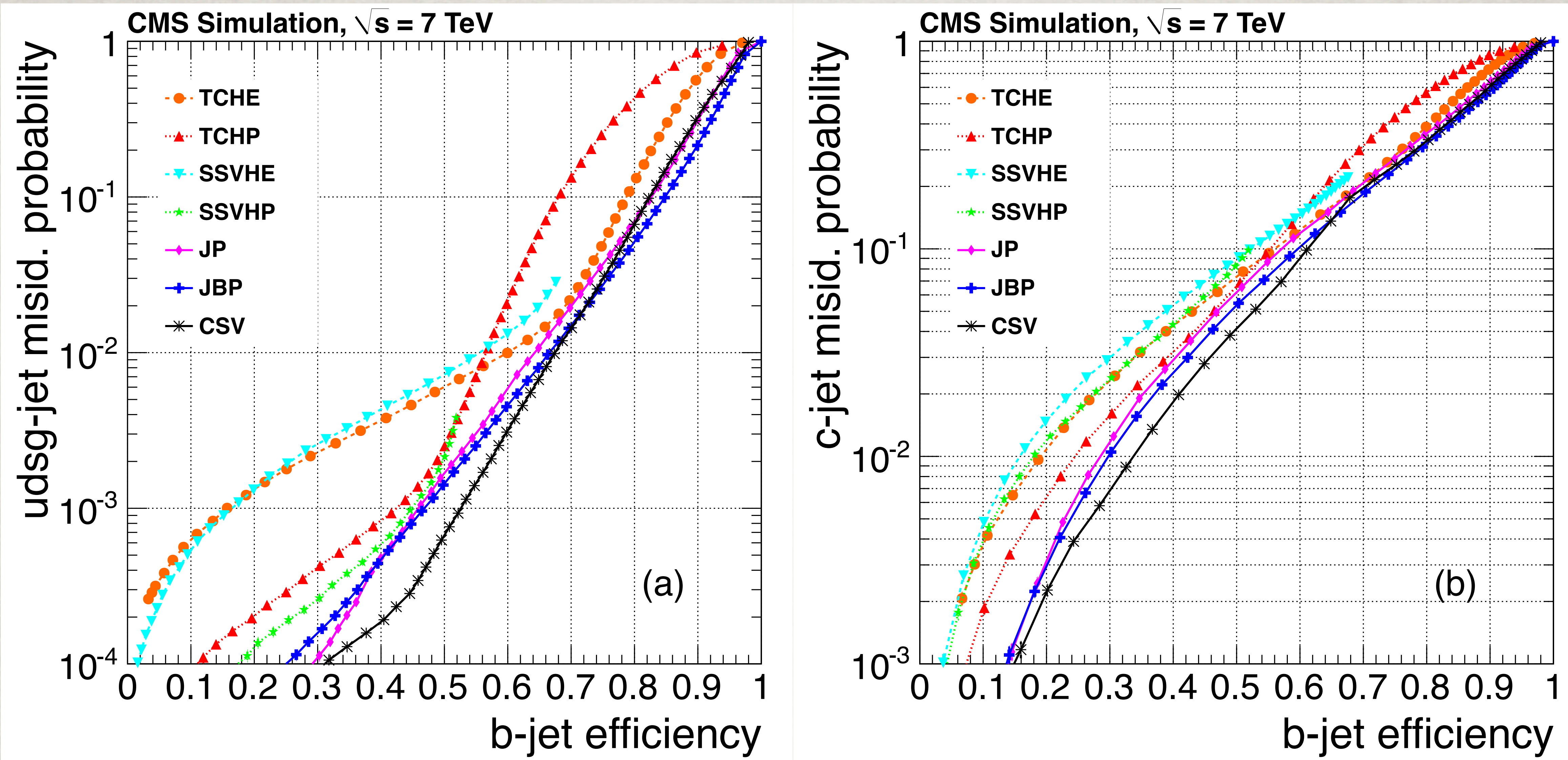
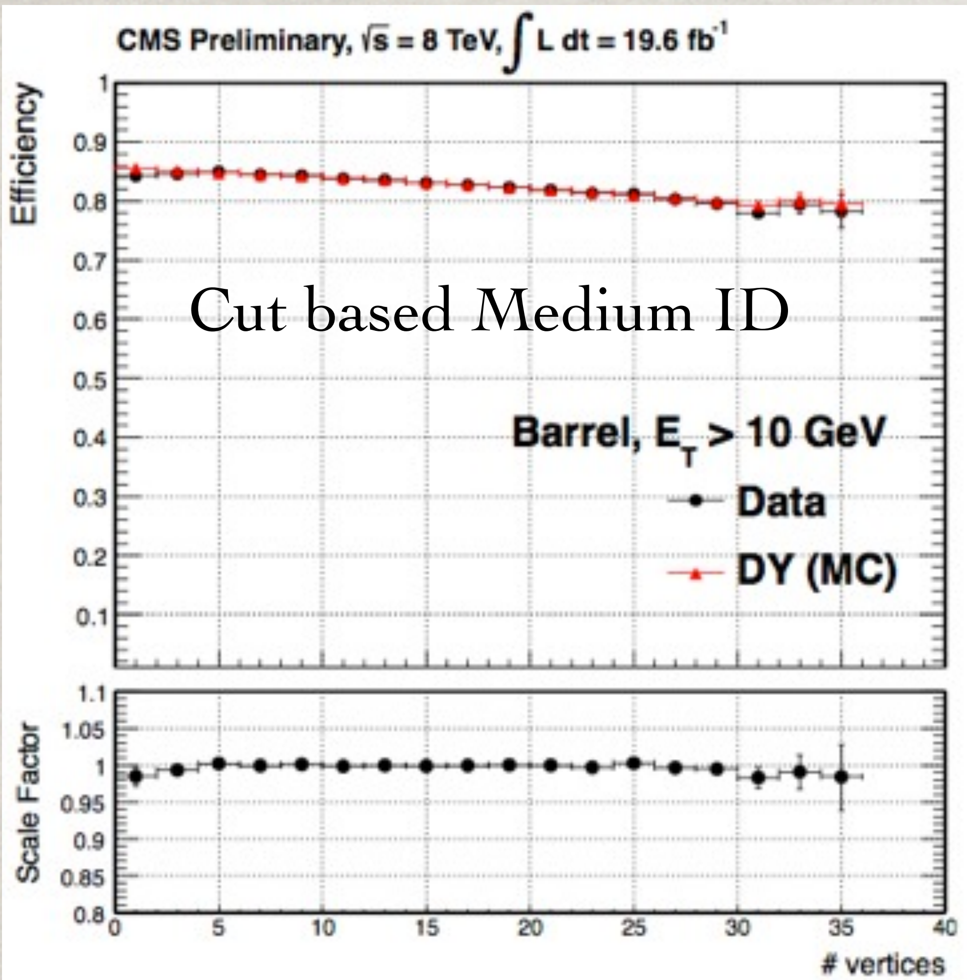


Figure 6: Performance curves obtained from simulation for the algorithms described in the text. (a) light-parton- and (b) c-jet misidentification probabilities as a function of the b-jet efficiency.

Electron MVA

From DP-13-003



From DP-13-003

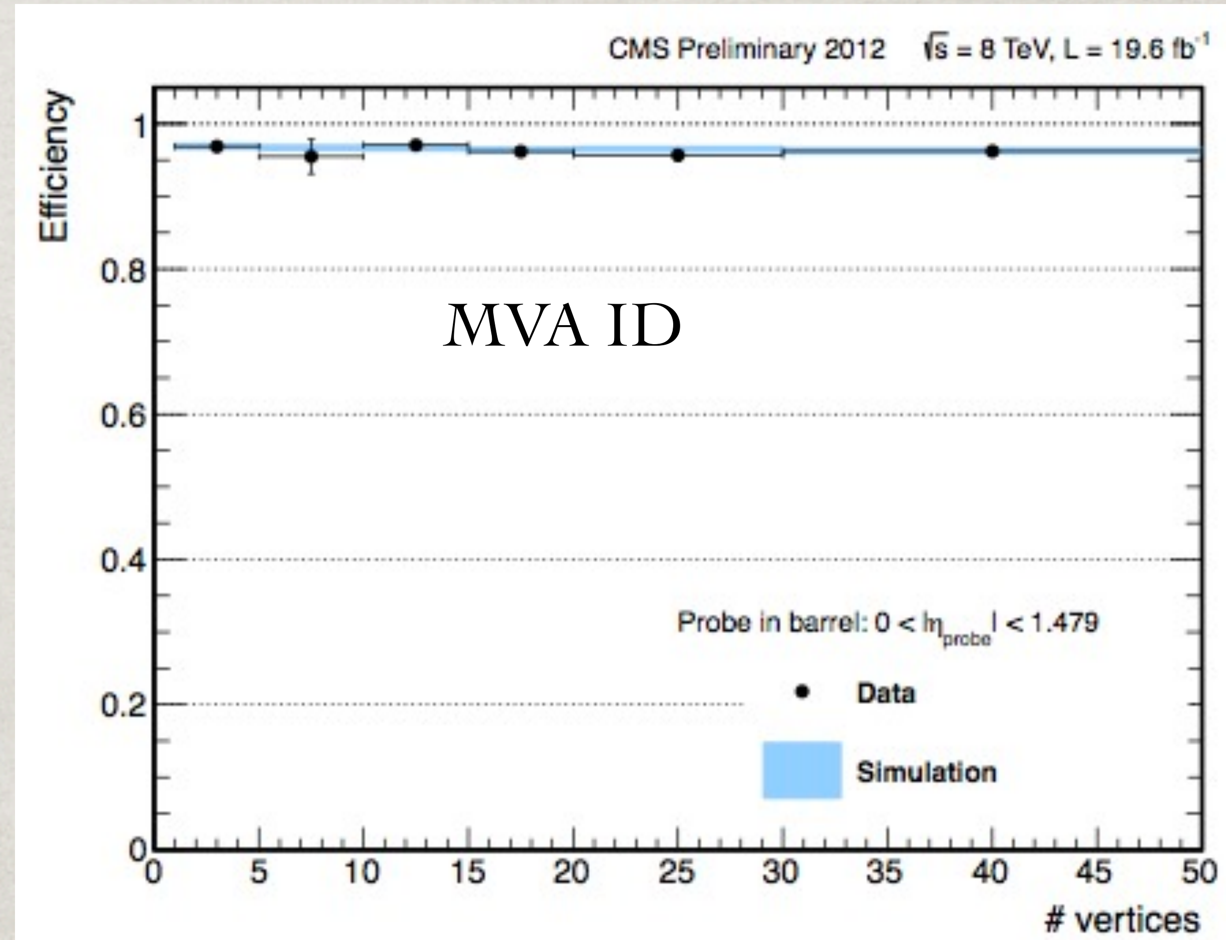


Table 4: The ANN inputs for the nine jet-tag categories in the 8 TeV $\mathrm{t\bar{t}H}$ analysis in the lepton+jets and dilepton channels. The choice of inputs is optimized for each category. Definitions of the variables are given in the text. The best input variable for each jet-tag category is denoted by ★.

	Lepton+Jets							Dilepton	
Jets	$\geq$6	4	5	$\geq$6	4	5	$\geq$6	2	$\geq$3
Tags	2	3	3	3	4	$\geq$4	$\geq$4	2	$\geq$3
Jet 1 p_{T}		✓	✓		✓			★	✓
Jet 2 p_{T}		✓	✓						
Jet 3 p_{T}	✓	✓	✓			✓			
Jet 4 p_{T}	✓	✓	✓			✓			
N_{jets}									✓
$p_{\mathrm{T}}(\ell, E_{\mathrm{T}}^{\text{miss}}, \text{jets})$		★	✓		✓	✓		✓	✓
$M(\ell, E_{\mathrm{T}}^{\text{miss}}, \text{jets})$	✓	✓		✓	✓		✓		
Average $M((\mathrm{j}_m^{\text{untag}}, \mathrm{j}_n^{\text{untag}}))$	✓			✓					
$M((\mathrm{j}_m^{\text{tag}}, \mathrm{j}_n^{\text{tag}})_{\text{closest}})$							✓		
$M((\mathrm{j}_m^{\text{tag}}, \mathrm{j}_n^{\text{tag}})_{\text{best}})$							✓		
Average $\Delta R(\mathrm{j}_m^{\text{tag}}, \mathrm{j}_n^{\text{tag}})$				✓	✓	✓	✓		
Minimum $\Delta R(\mathrm{j}_m^{\text{tag}}, \mathrm{j}_n^{\text{tag}})$			✓					✓	✓
$\Delta R(\ell, \mathrm{j}_{\text{closest}})$						✓	✓	✓	✓
Sphericity	✓			✓			✓		
Aplanarity	✓				✓				
H_0	✓								
H_1	✓				✓				
H_2				✓			✓		
H_3	★			✓			✓		
μ^{CSV}	✓	✓	★	★	★	★	★	✓	★
$(\sigma_n^{\text{CSV}})^2$		✓	✓	✓	✓	✓			
Highest CSV value						✓			
2^{nd}-highest CSV value		✓	✓	✓	✓	✓	✓		
Lowest CSV value		✓	✓	✓	✓	✓	✓		

Significance

$$\langle S^2 \rangle = \frac{1}{2} \int \frac{\left(\hat{y}_S(y) - \hat{y}_B(y)\right)^2}{\hat{y}_S(y) + \hat{y}_B(y)} \mathrm{d}y,$$