Search for single vector-like quarks at CMS with machine learning techniques

Francesco Carnevali
INFN & Università degli Studi di Napoli Federico II

21 December 2021

Vector-like quarks (VLQs)

Several extensions of the Standard Model (SM) predict the existence of VLQs

Properties of VLQs:

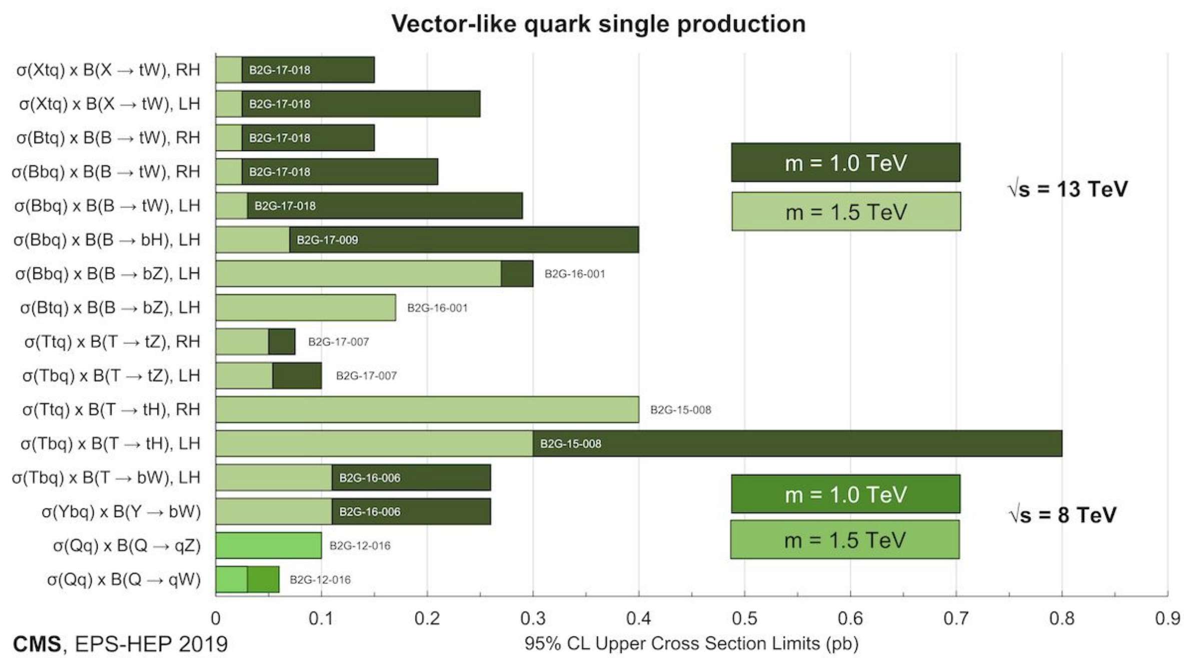
- Spin ½ fermions
- Left-handed and right-handed components behave in the same way under the SM symmetry group
- Vector current couplings to the weak gauge bosons
- Non-Yukawa coupling mass-terms for VLQs are allowed.

Type	Charge
X	+5/3
T	+2/3
B	-1/3
Y	-4/3

SU(2) Multiplets	
Singlets	T, B
Doublets	(T,B), (X,T), (B,Y)
Triplets	(X,T,B), (T,B,Y)

Search for VLQs at CMS

- Searches for vector-like quarks (VLQ) at the CMS experiment @ LHC
  - LHC Run II pp collision data, $\sqrt{s}$ =13 TeV , 137 fb^{-1}
- Single VLQs production:
  - T $\longrightarrow$ t Z/H/A*
- Different final states:
  - t $\longrightarrow$ b l ν
  - Z/H/A $\longrightarrow$ b b



*https://arxiv.org/abs/1907.05929

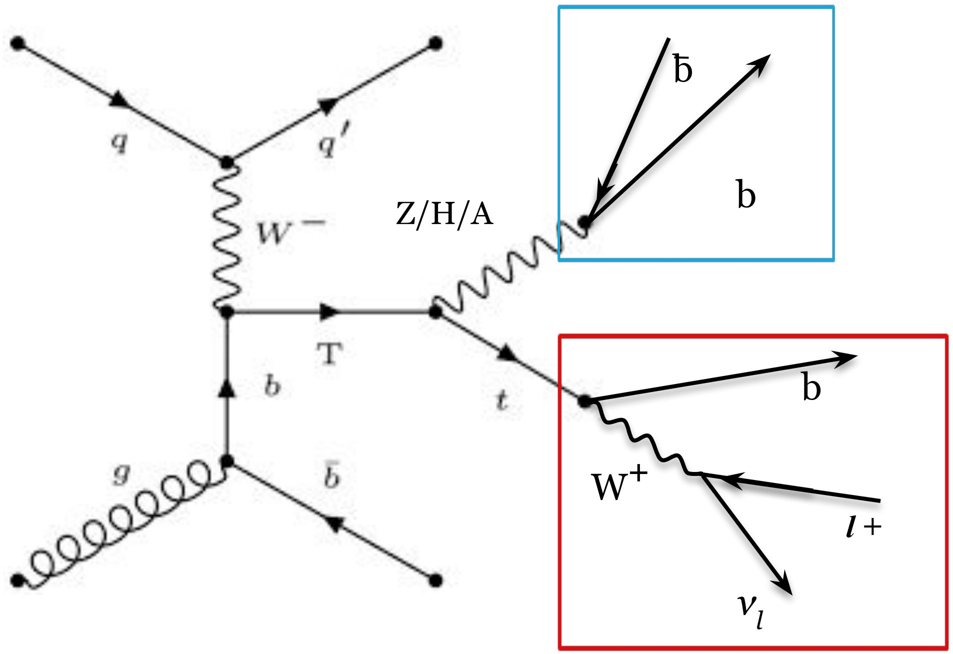
Francesco Carnevali 19/12/2021

Search for single VLQ T at CMS

- Signal: left-handed single VLQ T, in the t Z/H/A decay channel

bW ⟶ T
T ⟶ t Z/H/A

- Hadronic Z/H/A decay and in leptonic top quark decay
- Define new identification criteria for top quark reconstruction making use of ML techniques



Final state objects – top decay

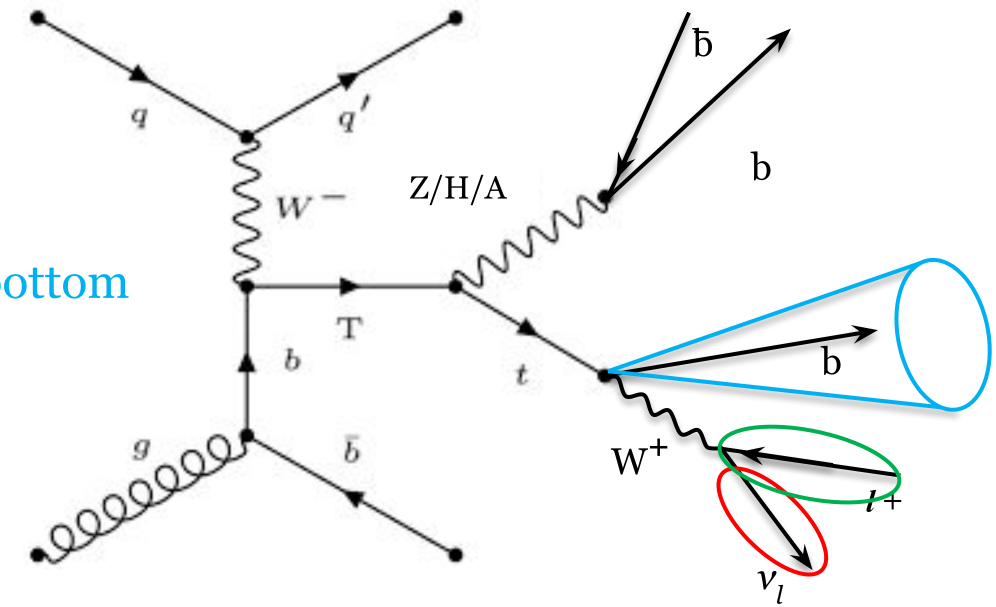
$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\varphi)^2}$

Leptonic top quark decay:

1. >= 1 Lepton
2. Met assigned to neutrino production
3. >=1 JetAK4 (ΔR=0.4), from bottom quark hadronization

At least 1 Top quark candidate:

1. JetAK4 Pt>30 GeV
2. Lepton Pt>10 GeV



Top Quark Reconstruction

Top quark candidates category depends on ΔR(j,l)

$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\varphi)^2}$

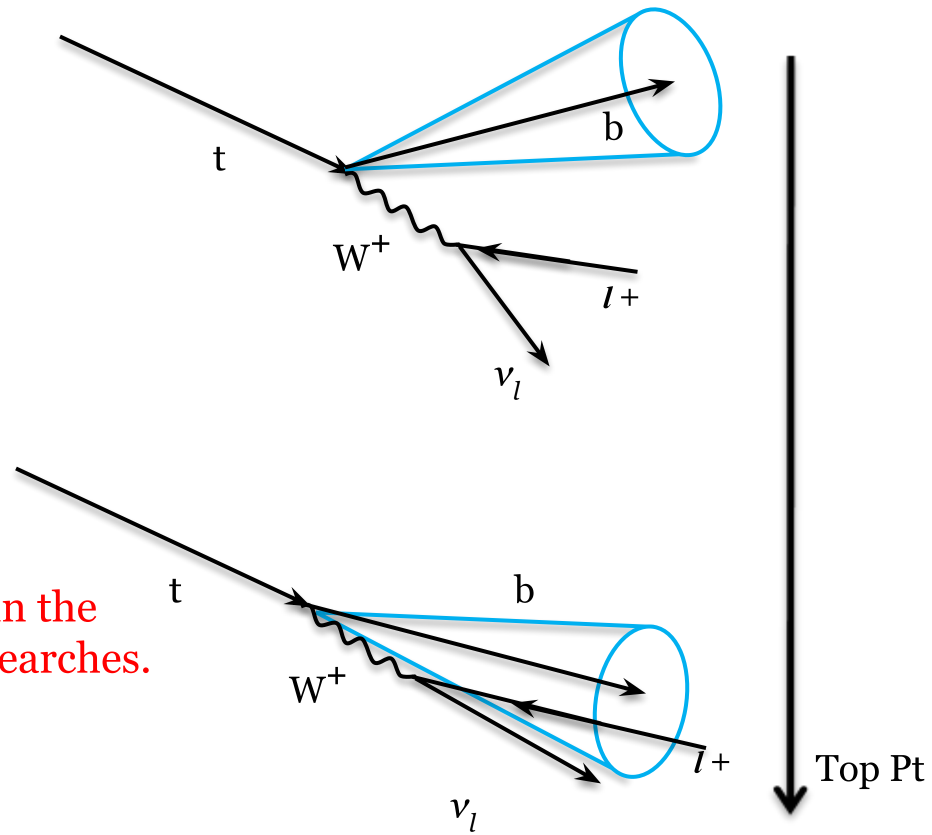
If 0.4<ΔR<2, the top quark reconstructed is classified as:

Top Resolved.

If ΔR<0.4, the top quark reconstructed is classified as:

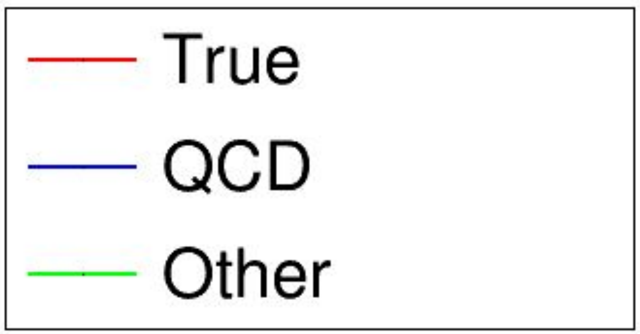
Top Merged. → Not considered in the previous VLQs searches.

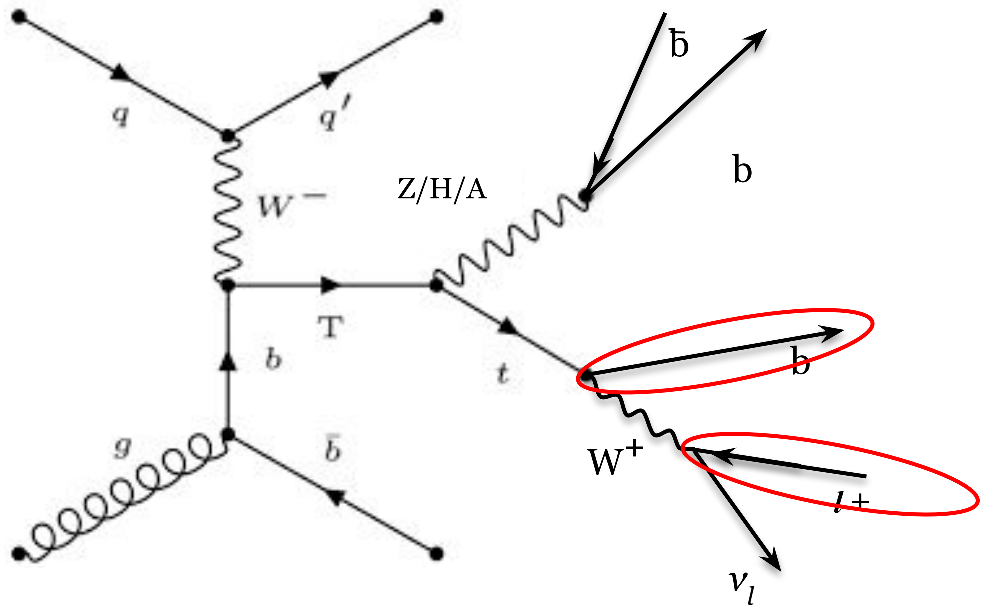
It's expected the "Top Merged" configuration if the quark top is boosted.



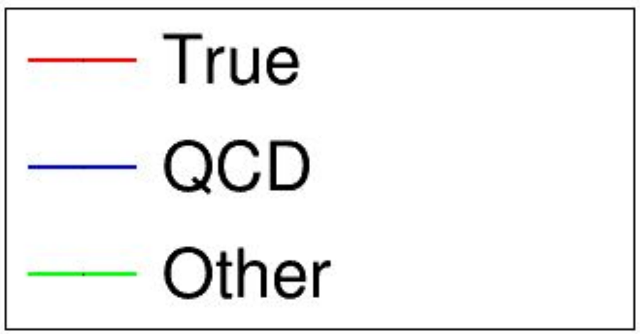
Francesco Carnevali 19/12/2021

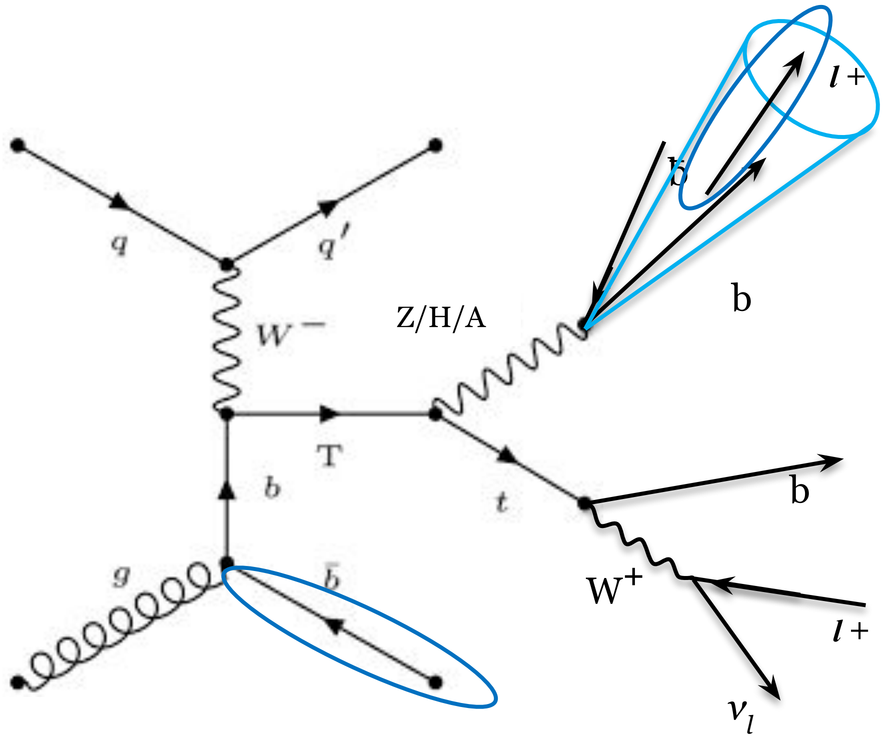
Top Quark Classification - True

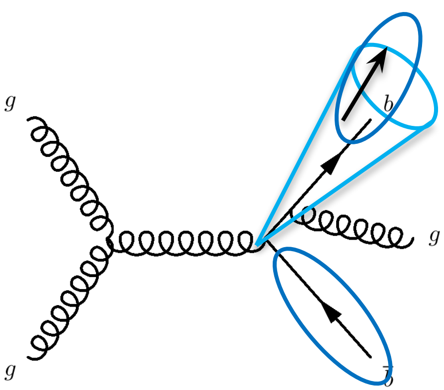




Top Quark Classification - QCDlike

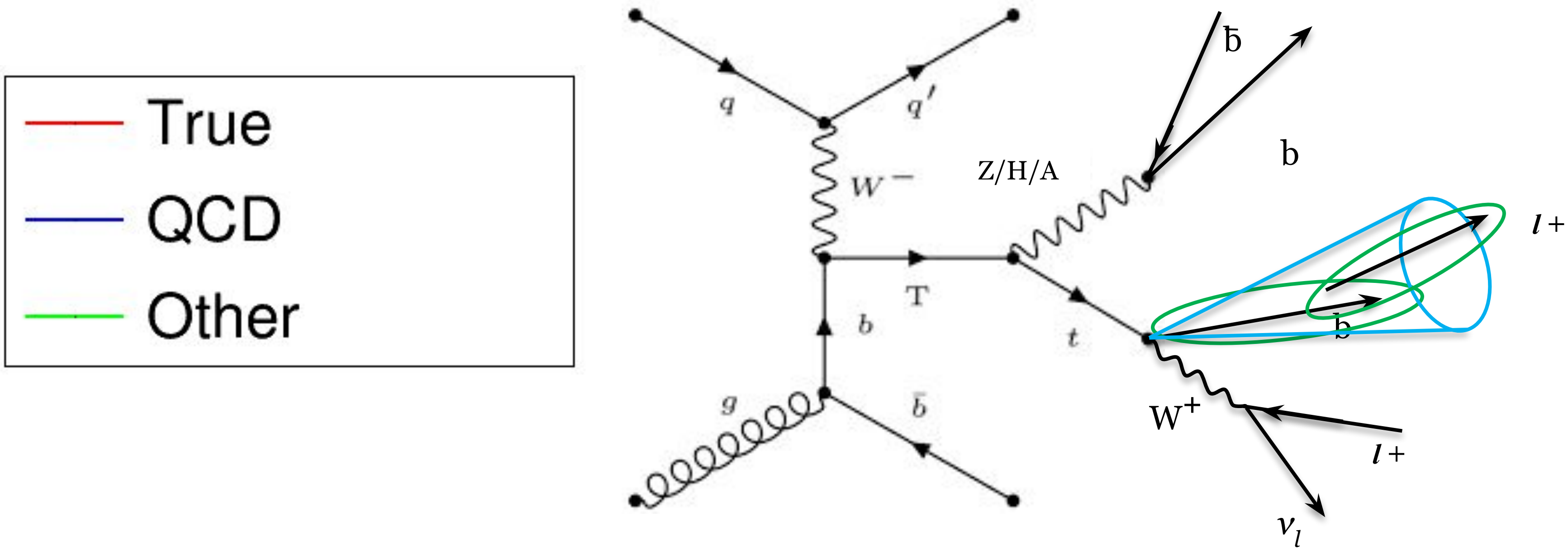






QCD MultiJet process

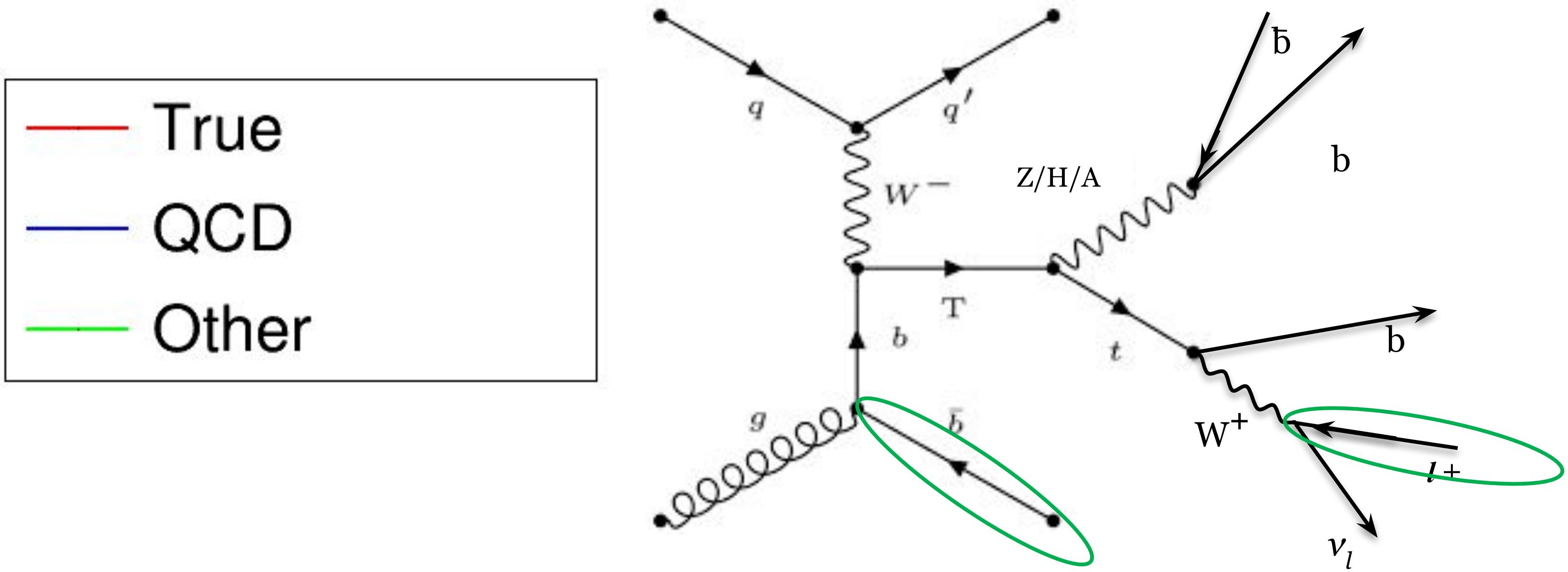
Top Quark Classification - Other



Top Merged: $b \rightarrow ..+\pi \rightarrow ...+\mu\nu$

Top Resolved: $Z/\gamma \rightarrow \mu^{+}\mu^{-}$, $b \rightarrow ..+\pi \rightarrow ...+\mu\nu$

Top Quark Classification - Other



Multiclass BDT Training

Use of machine learning algorithm (XGBoost) to improve top quark selection:

- Top category: Resolved or Merged
- Lepton: Muon or Electron
- Top p_T bin: Low [0,500) or High [500,Inf)

} 8 Training Categories

Lepton preselection requirements: standard Id, Isolation, and PV distances.

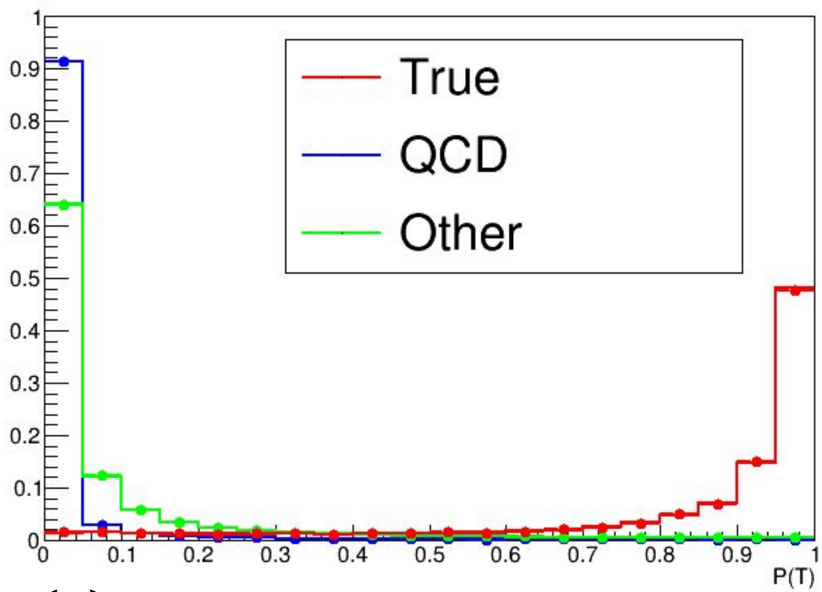
The BDT algorithm provides three different scores for each candidate:

- True
- QCD like (both jet and lepton don't match with MC truth)
- Other (1 object matches)

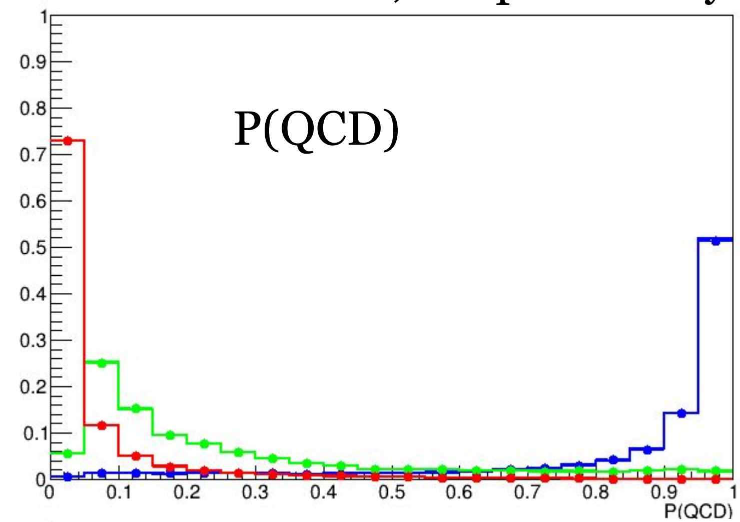
Multiclass BDT Training

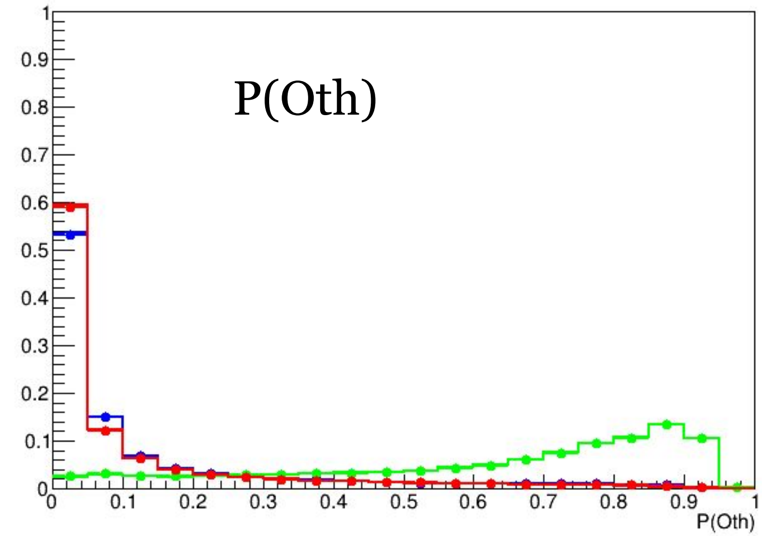
The BDT algorithm provides three different scores for each candidate, the probability of belonging to each of the three groups:

$$P(T) + P(QCD) + P(Oth) = 1$$



P(T)





Top Candidates: Resolved Low Pt (μ+Jet)

Multiclass BDT Training

The BDT algorithm provides three different scores for each candidate, the probability of belonging to each of the three groups:

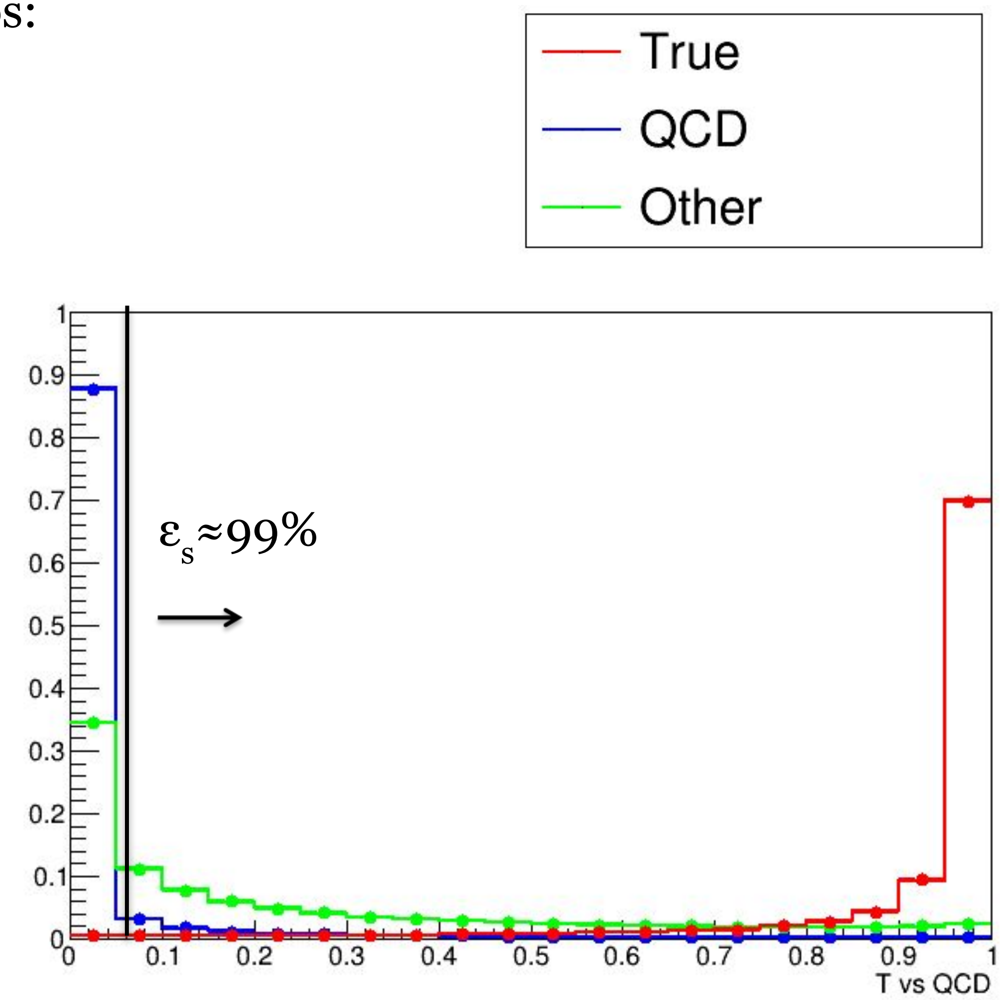
$$P(T)+P(QCD)+P(Oth)=1$$

Top taggers:

$$TvsQCD=\frac{P(T)}{1-P(Oth)}=\frac{P(T)}{P(T)+P(QCD)}$$

Preselection cut:

$\varepsilon_s \approx 99\%$



Multiclass BDT Training

The BDT algorithm provides three different scores for each candidate, the probability of belonging to each of the three groups:

$$P(T)+P(QCD)+P(Oth)=1$$

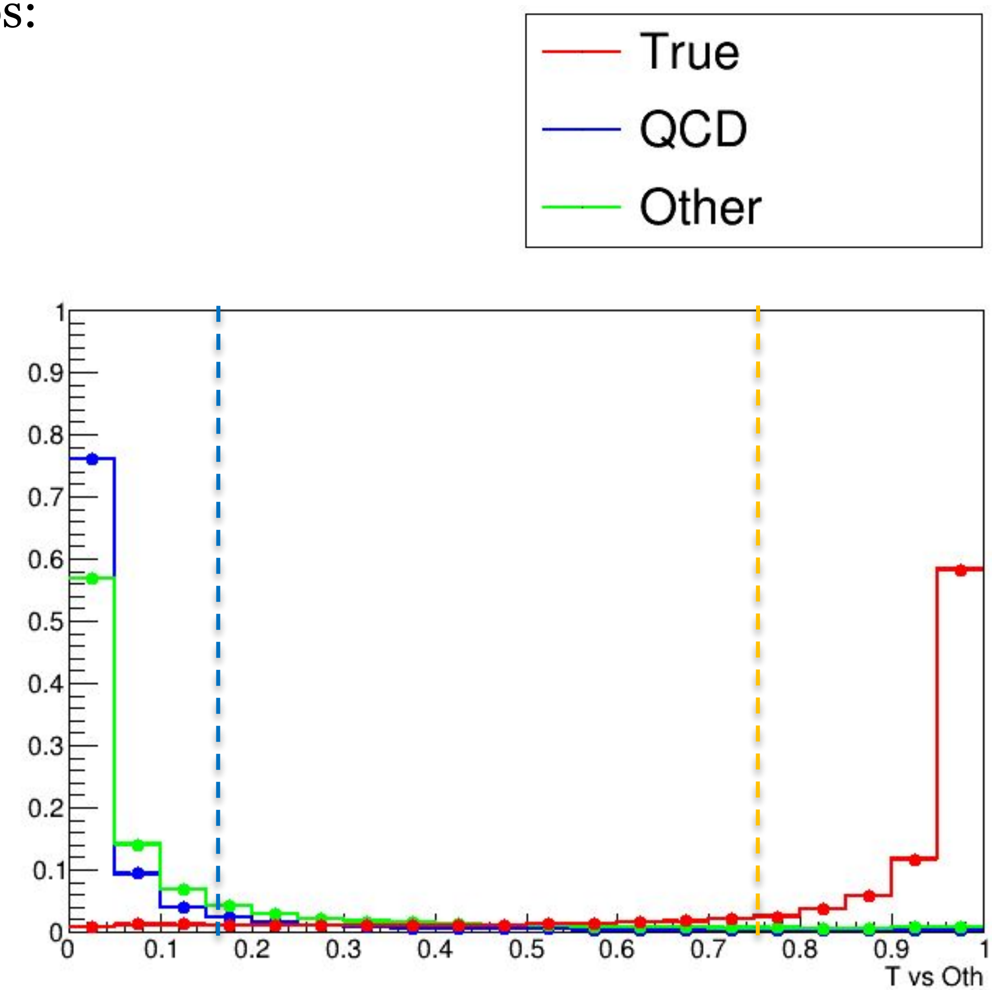
Top taggers:

$$TvsOth=\frac{P(T)}{1-P(QCD)}=\frac{P(T)}{P(T)+P(Oth)}$$

Selection cut:

WP Loose (misId Oth≈10%)

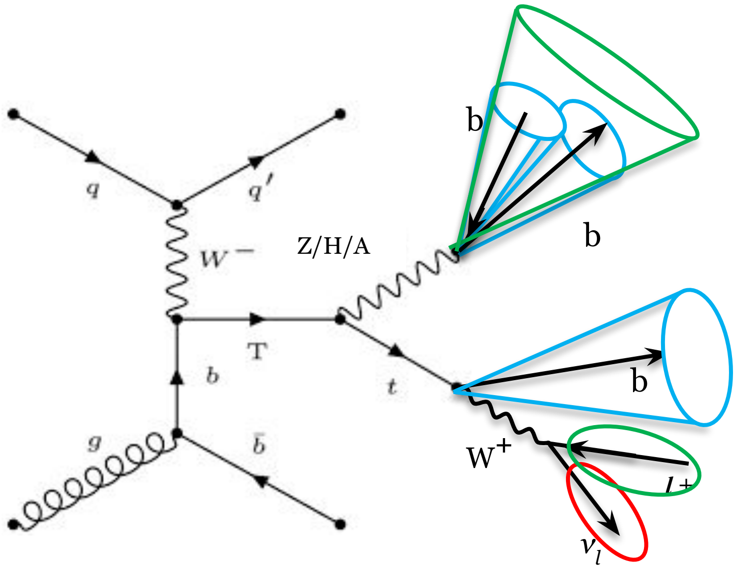
WP Tight (misId Oth≈1%)



Physics Background

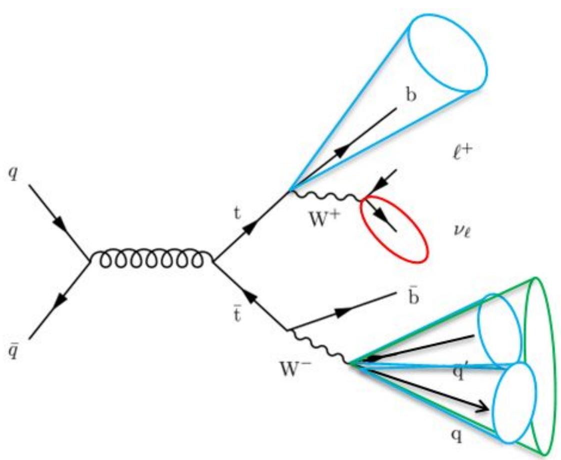
The signal region is characterized by:

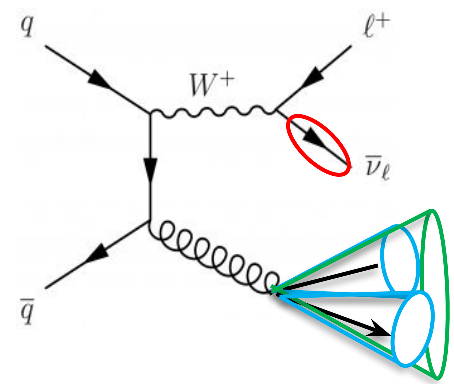
- 1 *tight* Top,
- 1 *tight* H/Z/A-tagged jet.



The main SM backgrounds:

- $t\bar{t}$,
- W+Jets,
- QCD MultiJet.





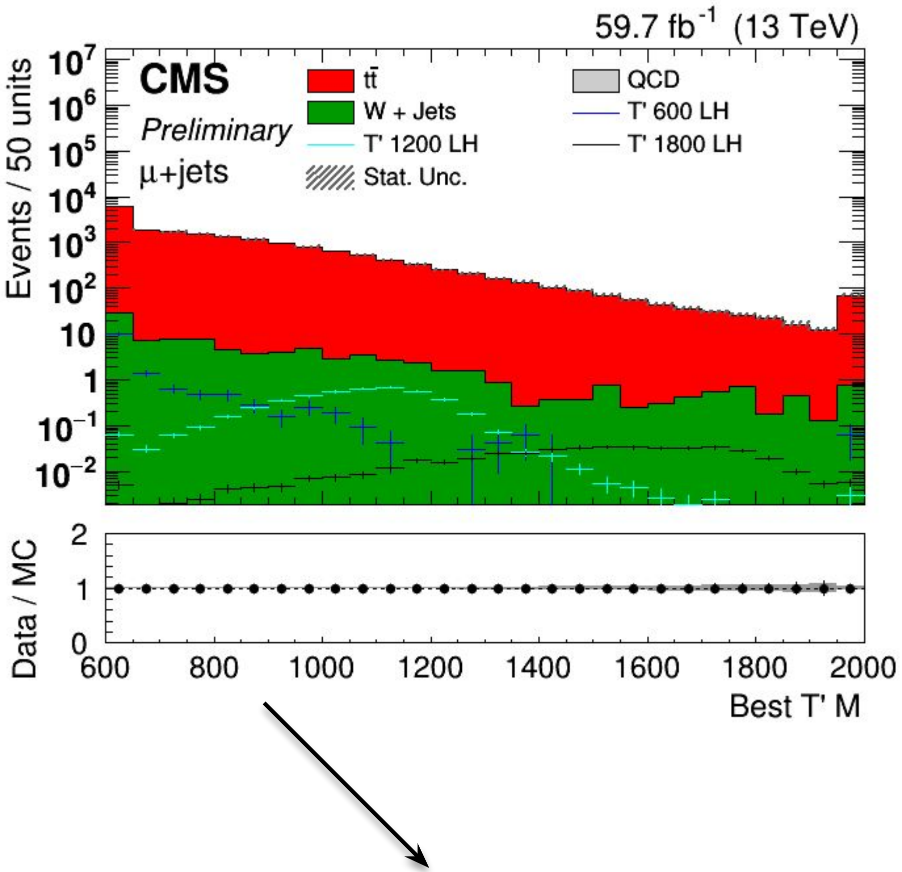
Conclusions and Next Steps

Conclusions:

- Search for single VLQ T production, focusing on the decay T→t Z/H/A → blν bb
- ML techniques are used as identification criteria for the leptonic top quark decay

Next Steps:

- Extend the analysis to all Run II Data
- Identify discriminating variables to improve signal to background discrimination
- Set T production cross section upper limits



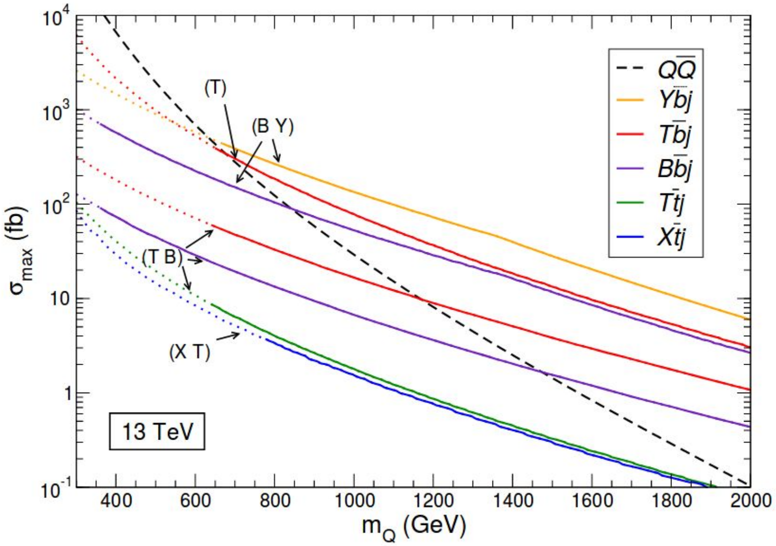
Signal Region
RunII 2018

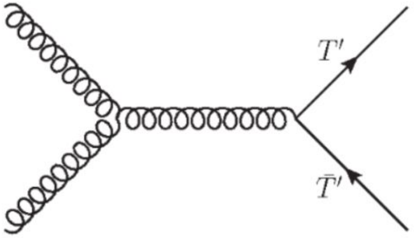
BACKUP

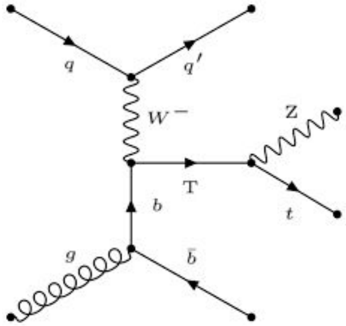
VLQ production and decay

VLQs could be produced both singly and in pair:

- Pair production
  - Strong interaction processes
  - Model independent cross section, suppressed for large VLQ mass
- Single production
  - Electroweak processes
  - Cross section depending on VLQ mass and coupling to SM particles
  - Models foresee preferential mixing with 3rd generation SM quarks







Type	Decay channel
X	tW
T	tZ,tH,bW
B	bZ,bH,tW
Y	bW

BR depending on VLQ mass and model

Final state objects – Z/H/A decay

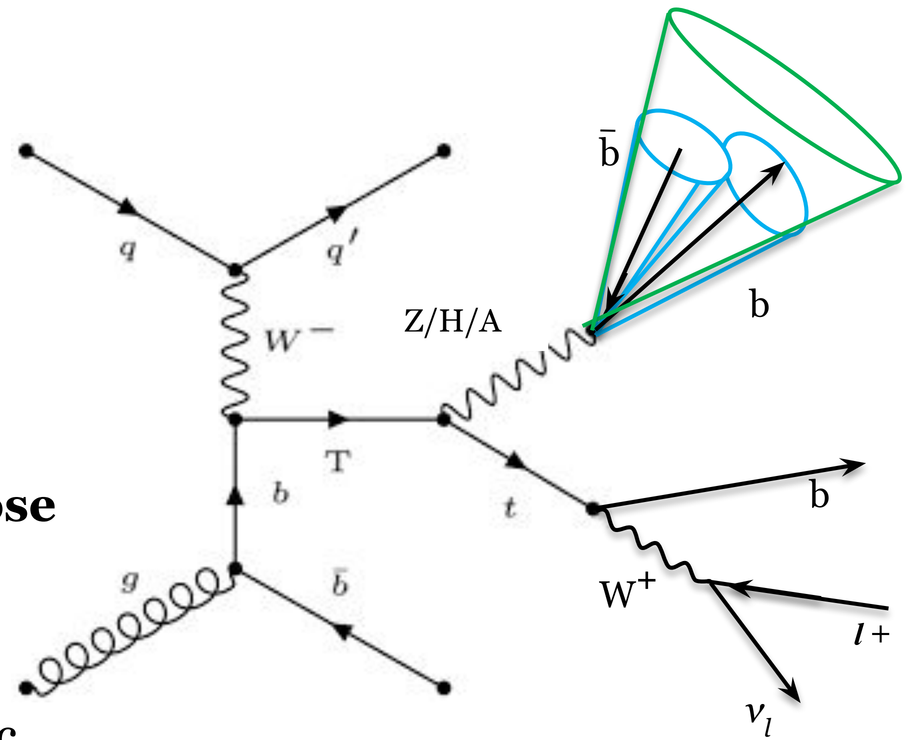
Z/H/A hadronic decay:

- >=1 JetAK8:
  - 60 Gev ≤ Mass < 140 GeV

or

  - **Z/H/A-> bb Tagger Loose WP (misId 10%),**

Mass-decorrelated tagger, Dynamic Graph Convolutional Neural Network

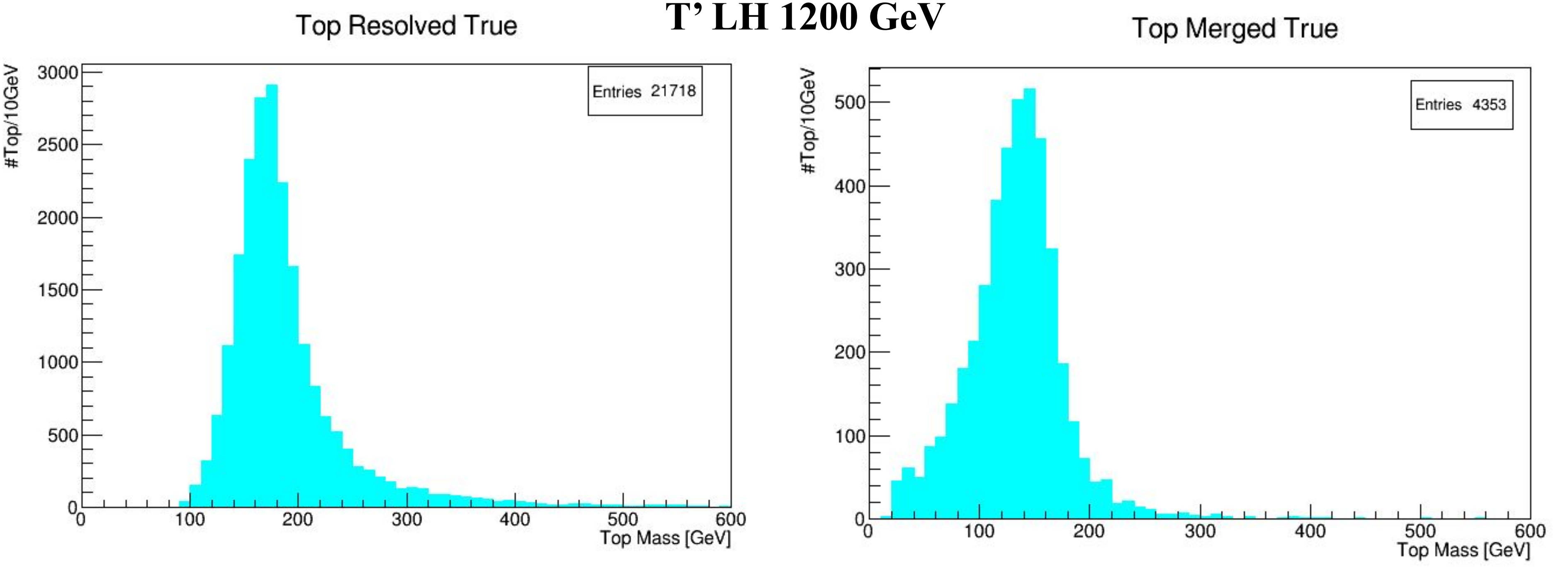


Top Quark Reconstruction

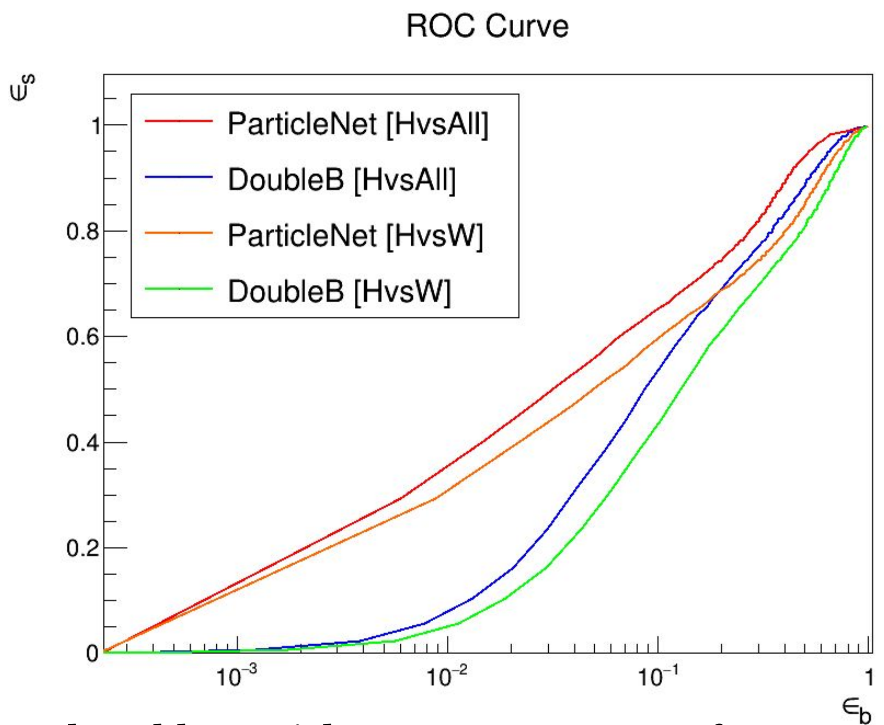
Mass distribution of the reconstructed top quark, with the lepton, Met and Jet.

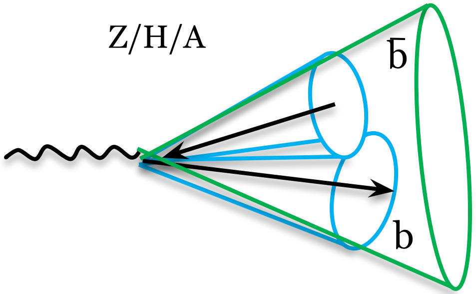
The neutrino momentum along the beam axis was estimated by imposing
$\sqrt{s}$ (l,MET)= 80,4 GeV (W Mass).

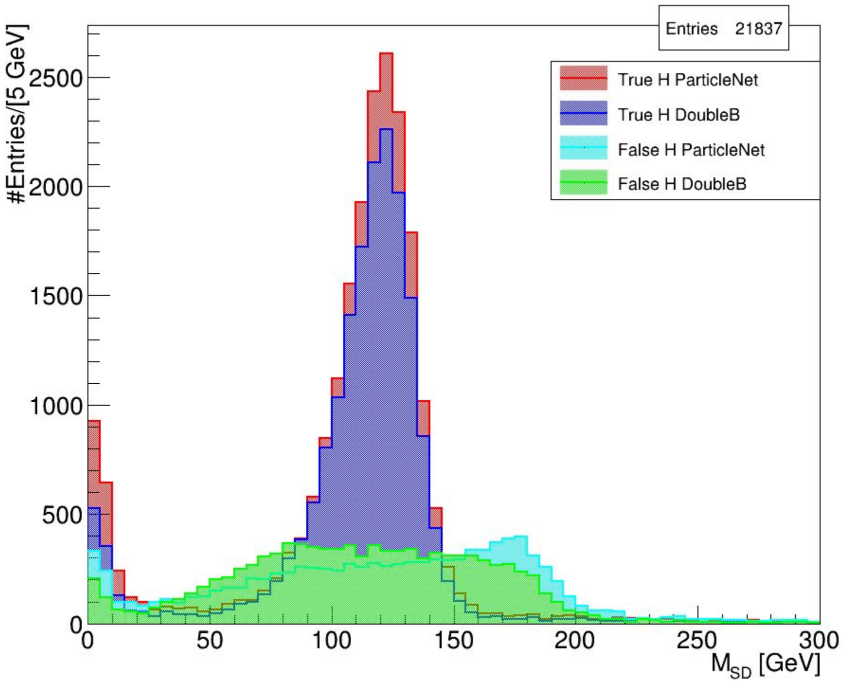
T' LH 1200 GeV



Xbb ParticleNet Tagger





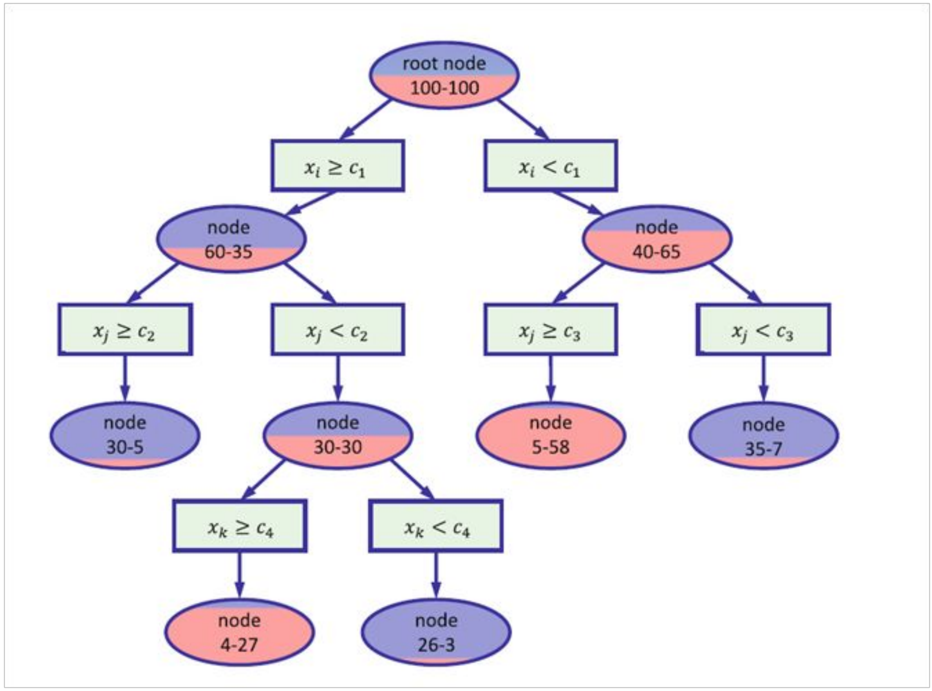


The Xbb ParticleNet MD tagger performs a better selection than the Double B tagger, both in the H vs All, and H vs W.

Studies has been done also on the Z->bb selection, and the tagger performs better than the DeepAk8 tagger.

BDT

BDT algorithm is used to improve the top quark reconstruction



- Input variables (top, jet, lepton):

$x_{1,\dots,n}$;

- Each tree has a weight ω;

- Final output is the weighted average of each output:

$$y(\vec{x}) = \sum_{k=1}^{N_{trees}} w_k C^k(\vec{x})$$

- The Multiclass BDT provides k different scores, each of them is the probability of belonging to the i-th category.

Multiclass BDT Selection

The BDT training has been improved, using the same variables, but with a large training dataset.

In each event only Top Candidates with a score greater than a threshold, in order to reject the 90% background, are selected (WP90).

Only events with at least 1 Top WP90 are selected.

Two differente Top Selected Collections:

- Top WP90, Loose
- Top WP99, Tight

μ/e	Resolved	Merged
Low p_T	2	4
High p_T	3	1