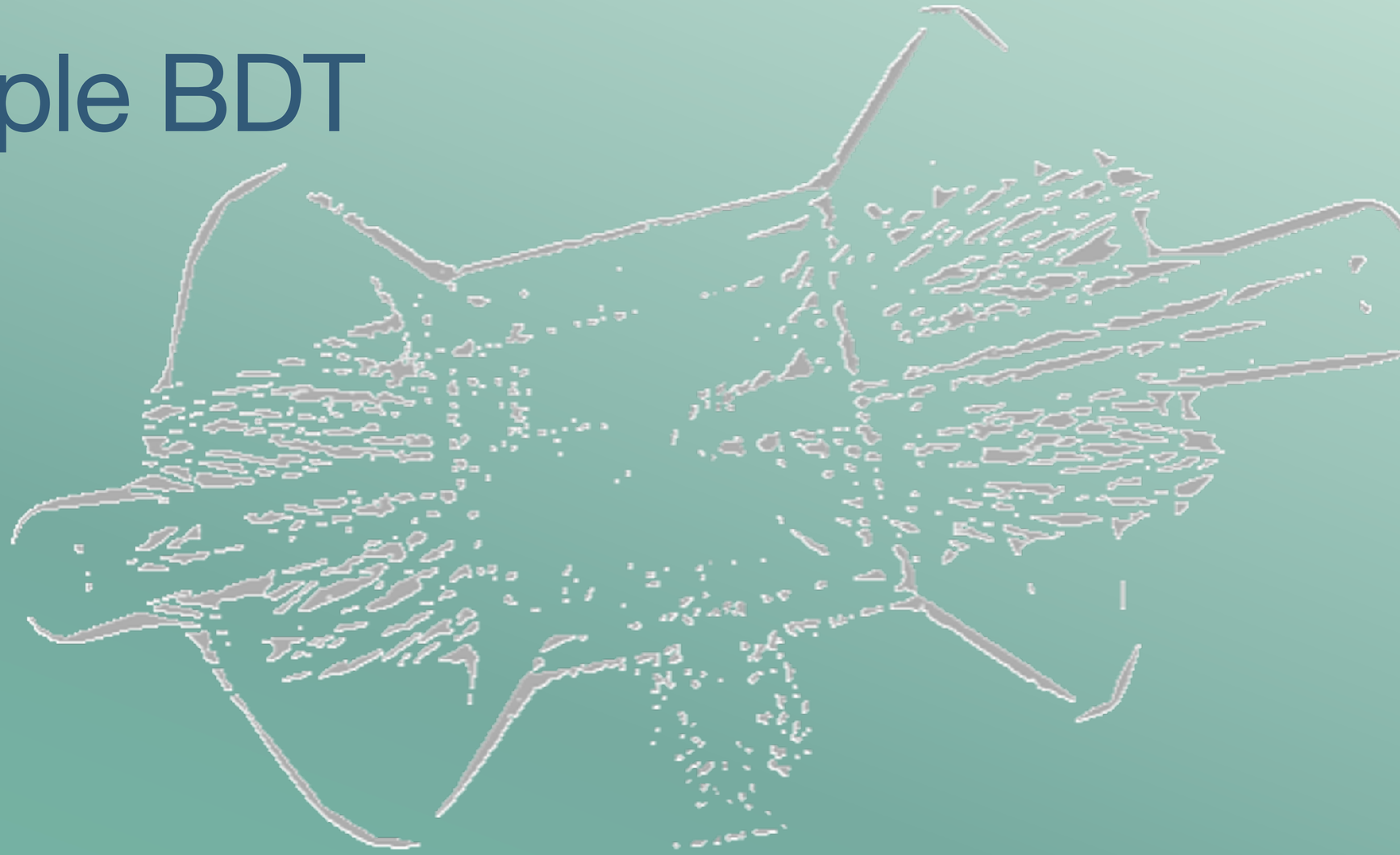
Vts with simple BDT

Nov 28, 2024

Xunwu Zuo



Previous results

Final selections

dilep cat

- missing energy > 80 GeV
- s-jet candidate energy > 45 GeV
- b-jet candidate energy > 25 GeV
- b-jet b-score > 0.9

semilep_cs cat

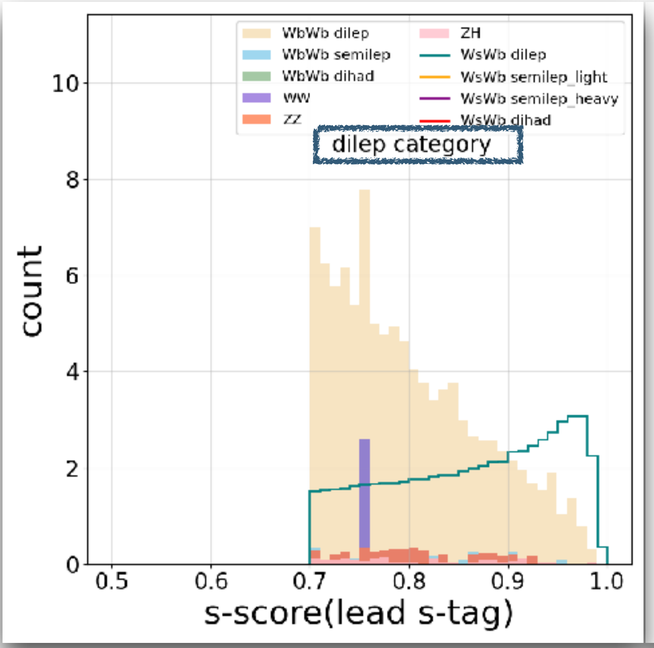
- missing energy > 30 GeV
- Can reconstruct 1 $t \to W(cs)b$ or $t \to W(cs)s$ decay
  - c-score and s-score > 0.5
  - $60 < m_W^{cs} < 80$ GeV
  - $140 < m_{top}^{bcs/scs} < 175$ GeV

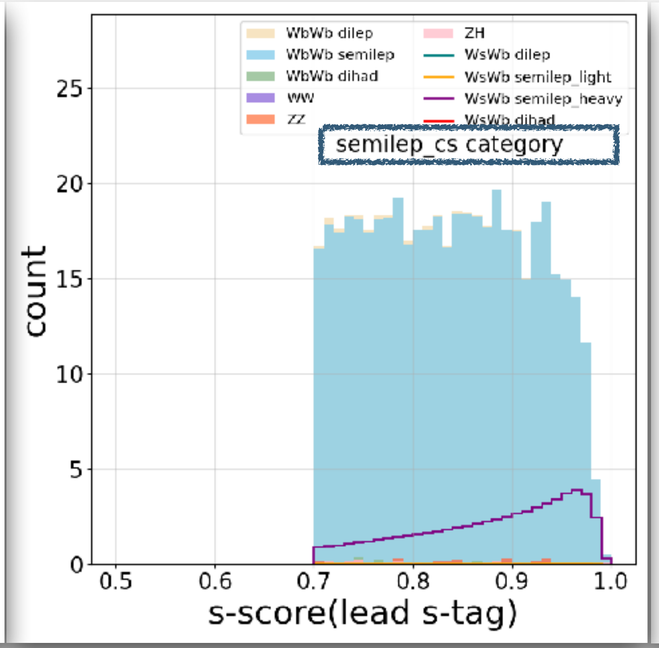
semilep_ud cat

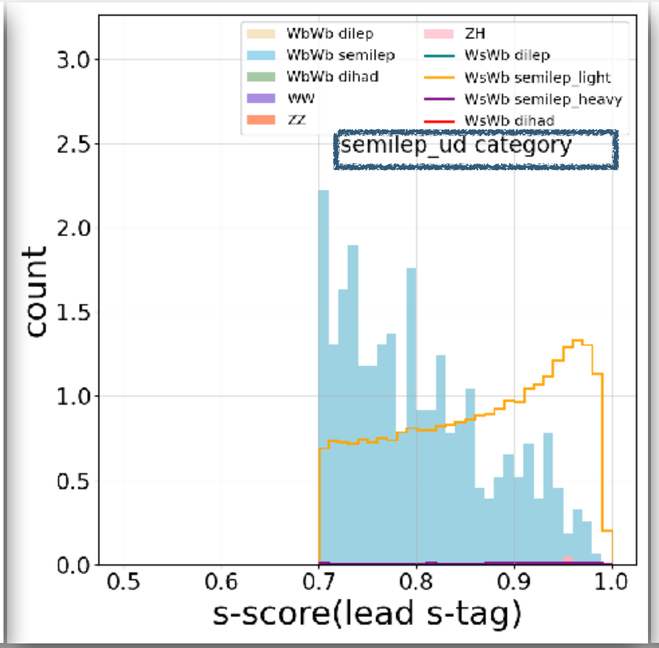
- missing energy > 30 GeV
- Can reconstruct 1 $t \to W(ud)b$ or $t \to W(ud)s$ decay
- s-jet candidate energy > 60 GeV
- s-jet candidate energy > 45 GeV
- b-jet b-score > 0.9

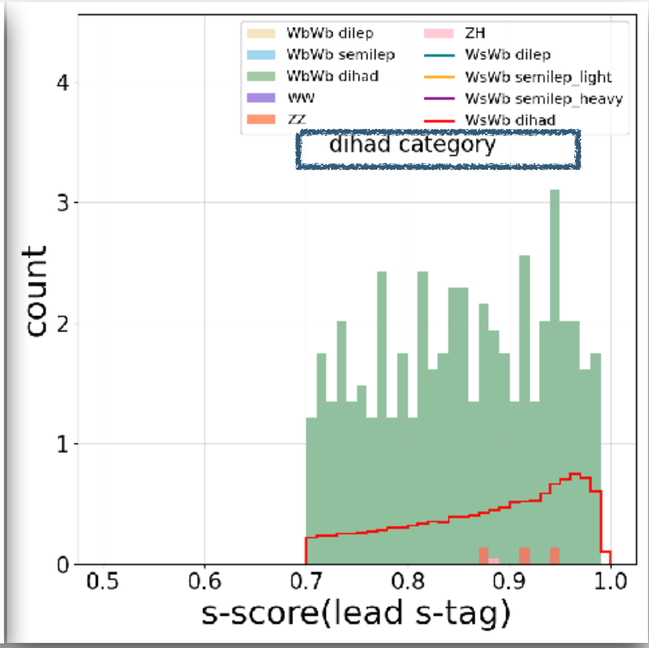
dihad cat

- missing energy < 20 GeV
- s-jet candidate energy > 60 GeV
- b-jet candidate energy > 40 GeV
- b-jet b-score > 0.9
- Can reconstruct 2 top candidates









Expected precision

- Preliminary fit with HiggsCombine
  - Very basic datacards, no uncertainty included
  - For significance,

    combine -M Significance datacard.txt -t -1 —expectedSignal=1
  - For signal strength,

    text2workspace datacard.txt -o ws.root

    combine -M FitDiagnostics -t -1 —expectedSignal=1 ws.root

category	dilep	semilep_cs	semilep_ud	dihad	combined
significance	5.83	1.13	4.78	1.49	7.77
precision	+29%/-25%	+97%/-88%	+42%/-33%	+177%/-99%	+22%/-20%

Event selection efficiency

Signal event efficiency

- Baseline selection: correct number of leptons and jets, exactly 1 jet with b-score > 0.5, at least 1 jet with s-score > 0.5, N(s-jet) - N(c-jet) = 1
- Final selection: given in slide 3

sig mode		dilep	semilep_light	semilep_heavy	dihad
Expected yield		644.4	1337.6	1337.6	2760.3
Total raw events		3862800	7459200	7792200	15051600
categories					
dilep	base	1.57×10^{-1}	0.0	0.0	0.0
	final	8.98×10^{-2}	0.0	0.0	0.0
semilep_light	base	1.86×10^{-3}	1.56×10^{-1}	1.51×10^{-2}	3.59×10^{-6}
	final	2.59×10^{-7}	2.01×10^{-2}	2.21×10^{-4}	0.0
semilep_heavy	base	1.50×10^{-4}	1.85×10^{-3}	6.87×10^{-2}	5.39×10^{-6}
	final	1.73×10^{-5}	9.92×10^{-4}	4.61×10^{-2}	0.0
dihad	base	2.88×10^{-6}	2.75×10^{-3}	1.43×10^{-3}	1.08×10^{-1}
	final	0.0	0.0	0.0	4.61×10^{-3}

bkg event efficiency

- Baseline selection: correct number of leptons and jets, exactly 1 jet with b-score > 0.5, at least 1 jet with s-score > 0.5, N(s-jet) - N(c-jet) = 1
- Final selection: given in slide 3

bkg mode		dilep	semilep	dihad	WW	ZZ	ZH
Expected yield		0.201×10^6	0.836×10^6	0.863×10^6	26791250	1607000	79262.93103
Total raw events		6400000	12800000	6400000	11754213	11470944	9800000
categories							
dilep	base	2.68×10^{-3}	3.52×10^{-6}	0	1.70×10^{-6}	2.96×-4	$1.11 \times^{-3}$
	final	4.70×10^{-4}	5.47×10^{-7}	0.0	8.51×10^{-8}	1.92×10^{-6}	1.32×10^{-5}
semilep_light	base	1.10×10^{-3}	5.03×10^{-3}	1.13×10^{-5}	1.59×10^{-5}	3.34×10^{-5}	2.82×10^{-4}
	final	0.00	3.12×10^{-5}	0.00	0.00	0.00	1.02×10^{-7}
semilep_heavy	base	2.20×10^{-4}	1.88×10^{-3}	4.22×10^{-6}	4.85×10^{-6}	1.39×10^{-5}	7.37×10^{-5}
	final	1.28×10^{-5}	5.79×10^{-4}	4.69×10^{-7}	0.00	7.85×10^{-7}	7.04×10^{-6}
dihad	base	4.38×10^{-6}	9.51×10^{-4}	1.04×10^{-2}	1.00×10^{-4}	1.26×10^{-4}	1.03×10^{-3}
	final	0.00	0.00	6.02×10^{-5}	0.00	2.62×10^{-7}	1.02×10^{-7}

A few things to improve

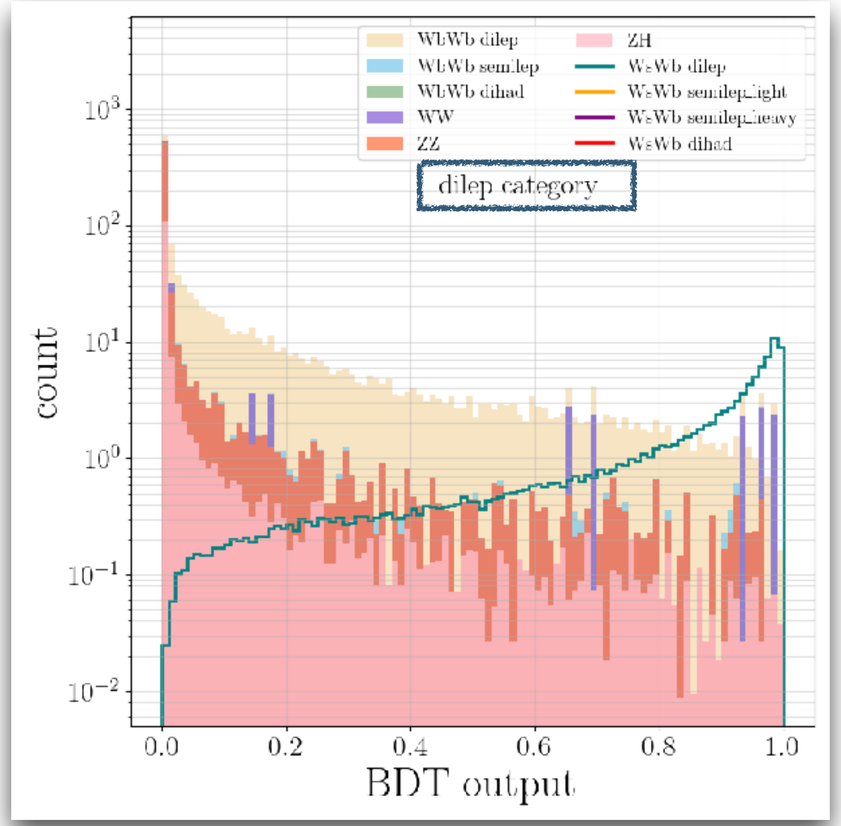
- Optimize config for jet clustering
- Proper full reconstruction of hadronic top decays
- Use an MVA discriminator rather than s-tag score for signal extraction
  - especially in semilep_heavy category

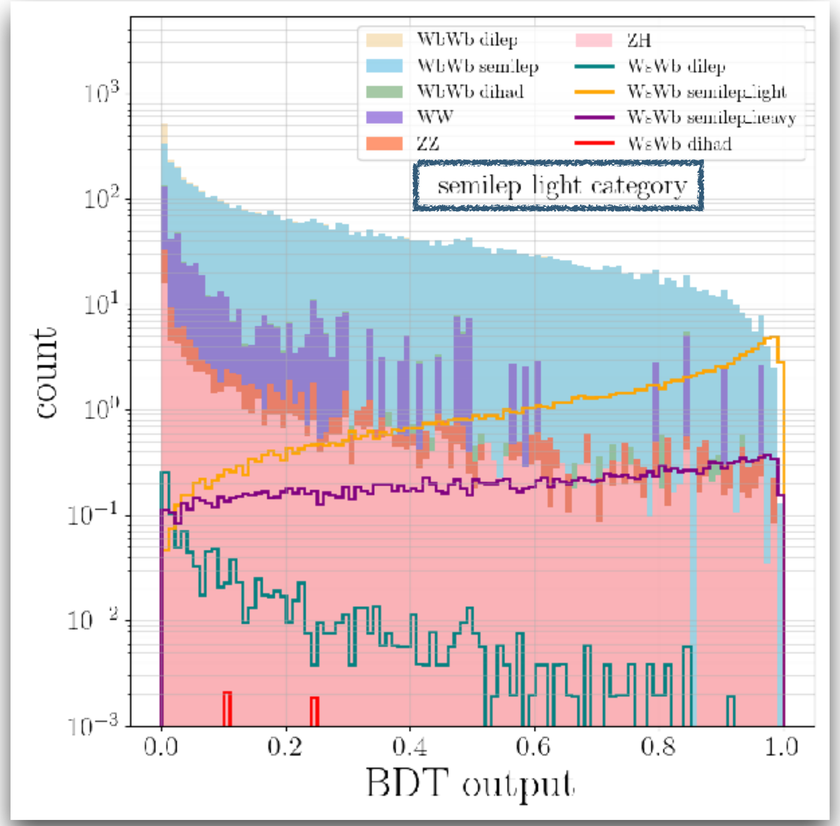
BDT training

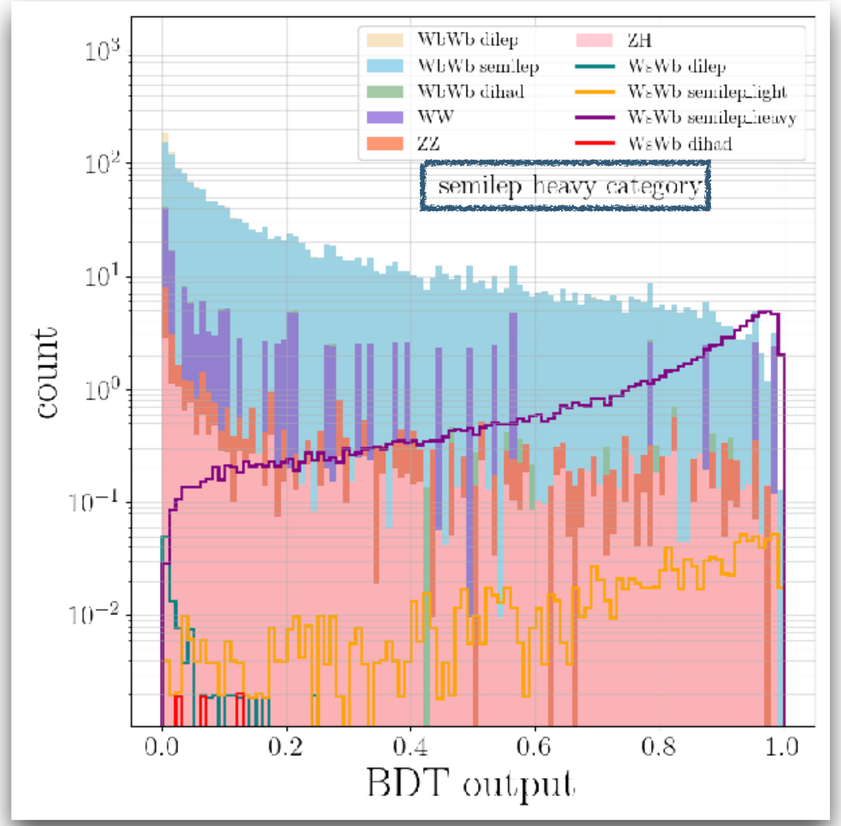
- For signal extraction
  - Hopefully also integrate information of hadronic top decay
  - Currently using 200 trees, depth of 3. Bigger BDTs lead to overtraining
  - Turned out not enough to “figure out” hadronic top decay combinatorics

- BDT vars
  - https://xzuo.web.cern.ch/FCC/topVts/BDT/dilep_vars/
  - https://xzuo.web.cern.ch/FCC/topVts/BDT/semilep_heavy_vars/
- Performance and overtraining test
  - https://xzuo.web.cern.ch/FCC/topVts/BDT/

BDT output for fit







Fit result

- Cut-based result for comparison

category	dilep	semilep_cs	semilep_ud	dihad	combined
significance	5.83	1.13	4.78	1.49	7.77
precision	+29%/-25%	+97%/-88%	+42%/-33%	+177%/-99%	+22%/-20%

- BDT-based result

category	dilep	semilep_light	semilep_heavy	combined
significance	8.83	4.09	2.69	10.5
uncertainty	$^{+20\%}_{-18\%}$	$^{+35\%}_{-30\%}$	$^{+50\%}_{-41\%}$	$^{+16\%}_{-15\%}$

Further ideas

- Currently have O(200k) events for each training, BDT not enough to explore complex correlations
  - Need better algorithm to extract hadronic top reco

- Systematically optimize jet clustering algorithm to recover signal efficiency