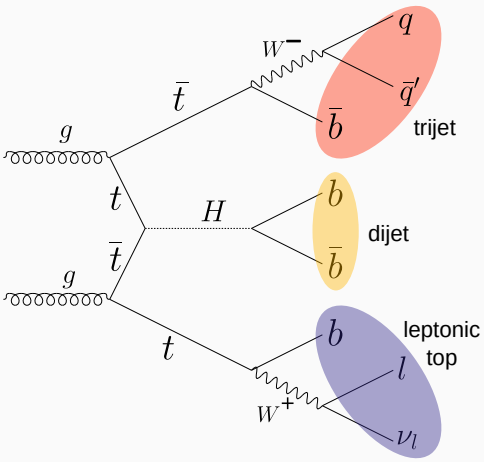
Attention-based reconstruction for $t\bar{t}H(b\bar{b})$ in CMS

L. Benato[1], A. Calandri[2], M. Donega[2], A. Gomez Espinosa[2], G. Kasieczka[1], T. Lösche[1], M. Meinhard[2], C. Reissel[2], D. Ruini[2], R. Wallny[2]

"Physics at the Terascale" Annual Meeting, 26.11.2019

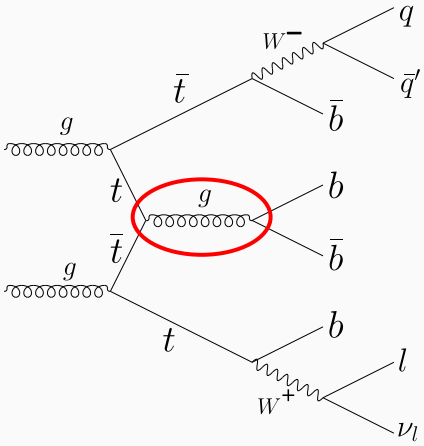
[1]IExp, Universität Hamburg; [2]IPA, ETH Zürich

Introduction - $t\bar{t}H$ motivation



- SM: fermion masses accounted for by Yukawa interactions between Higgs and fermion field
- top Yukawa coupling essential probe of SM (expected: $Y_t \approx 1$)
- coupling can be observed in Higgs production in association with top-quarks
- consider single-lepton channel

Introduction - $t\bar{t}H(b\bar{b})$ challenges

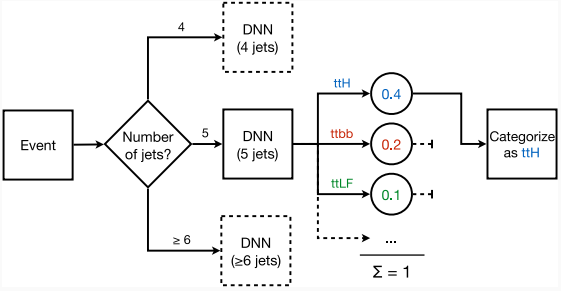


Problems:

- high number of jets in final state
- irreducible background ($t\bar{t}$ + $b\bar{b}$)
  $\rightarrow$ cross section 8 times higher than signal
- often final state not completely reconstructed

$\Rightarrow$ need to handle combinatorial assignment of jets to improve discrimination

Current state



- current state-of-the-art for SL channel: ANN output used as final discriminant [CMS-PAS-HIG-18-030]
- inputs: basic object kinematics and high-level variables (e.g. Matrix-Element-Method (MEM) discriminant)
- problem: MEM discriminant time consuming in computation (up to 10 min. per event on single CPU)
- solution: develop new DNN architecture with competitive discrimination power using simple kinematics as input

COBRA - attention based DNN

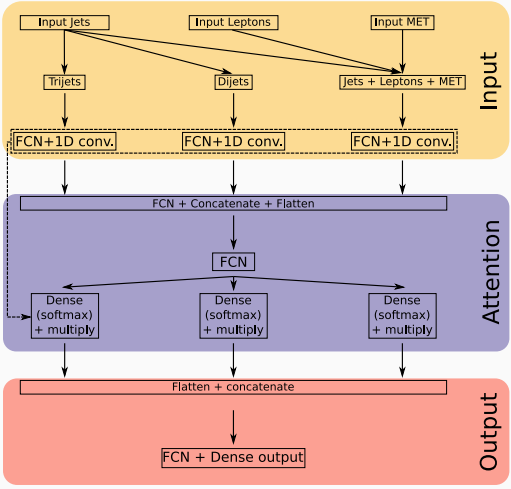
A woman is throwing a frisbee in a park.

A little girl sitting on a bed with a teddy bear.

arXiv:1502.03044 [cs.LG]

- COBRA = COmBinatoRics based deep Attention network, developed and implemented at ETH Zurich
- attention originally developed by Google (arXiv:1706.03762 [cs.CL])
- allows network to focus on subset of inputs (e.g. certain parts of image)
- can detect specific features
- extensively used in image recognition and natural language processing
- idea for $t\bar{t}H$: focus on jet-combinations to get a handle on combinatorics

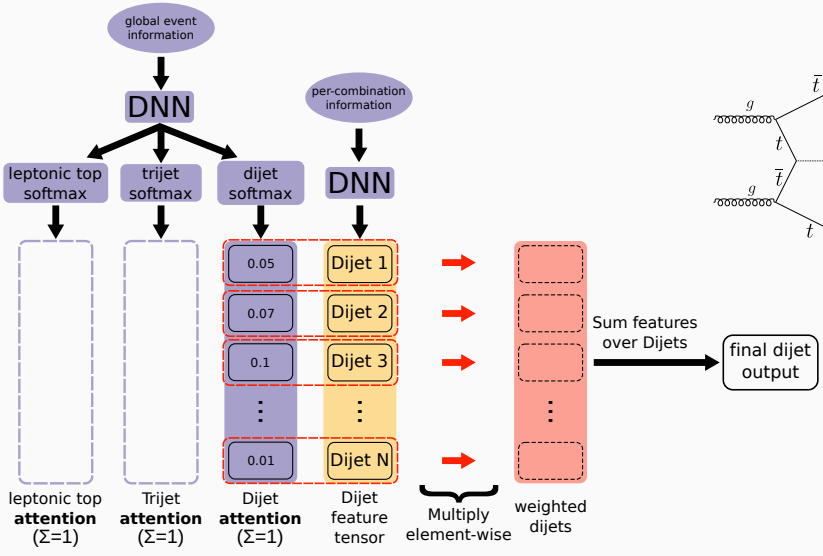
COBRA architecture & inputs

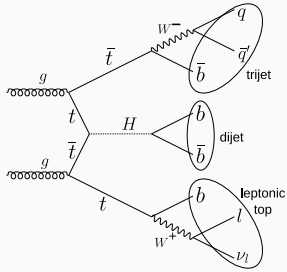


input variables (Monte Carlo simulation):

- jets [8 floats/jet]:
  - η, ϕ, Energy, p_x, p_y, p_z, p_T, b-tag
- leptons [7 floats]:
  - η, ϕ, Energy, p_x, p_y, p_z, p_T
- MET [5 floats]:
  - ϕ, $\sum E_T$, p_x, p_y, p_T
- no prior categorization based on jet or b-tag multiplicity
- always one lepton and ten jets (zero-padded if less) per event
  $[8 \cdot 10 + 7 + 5 = 92$ floats total$]$

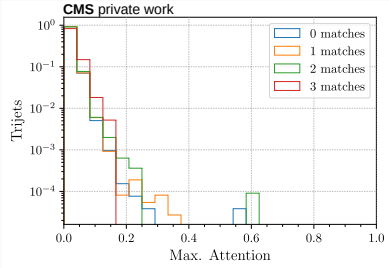
Attention mechanism



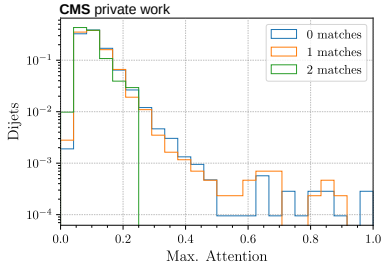


Attention output

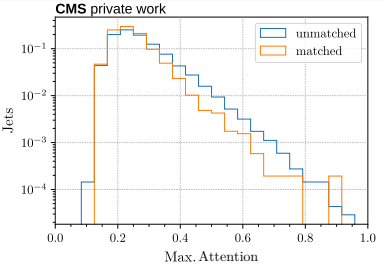
Trijet attention



Dijet attention

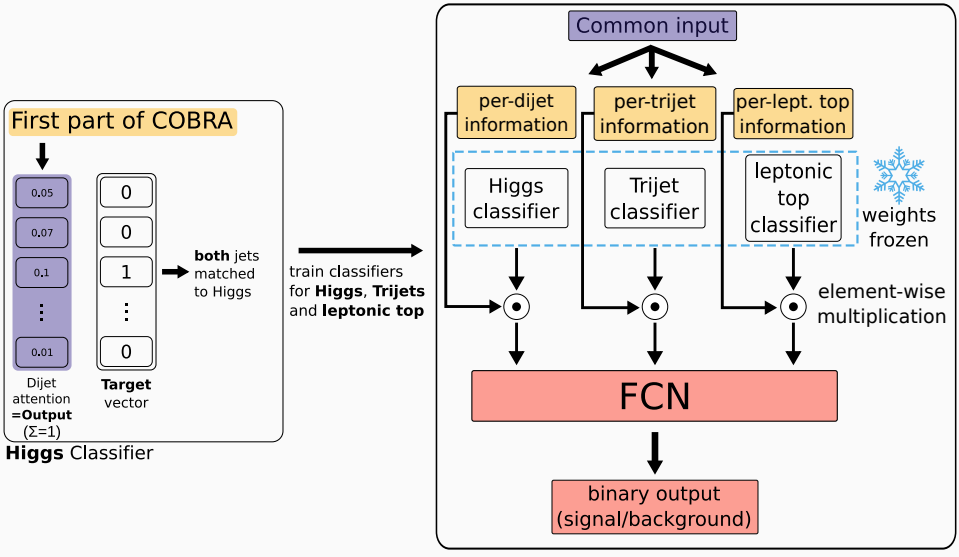


Leptonic jet att.



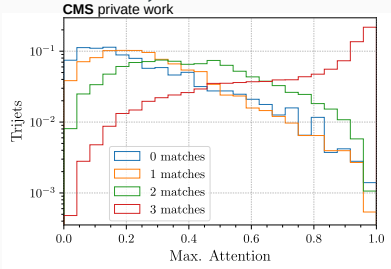
- completely matched sample:
  - 3 jets matched to hadronically decaying top
  - 2 jets matched to Higgs
  - 1 jet matched to leptonically decaying top
- expectation: Higher attention weights for higher number of matched jets → not the case

Classifier architecture

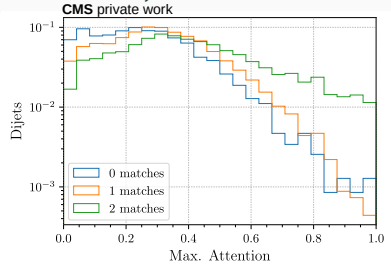


Classifier output

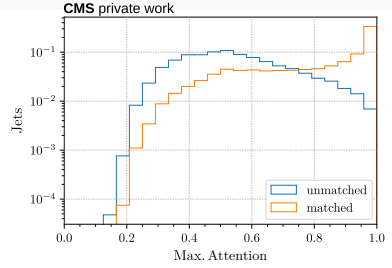
Trijet attention



Dijet attention



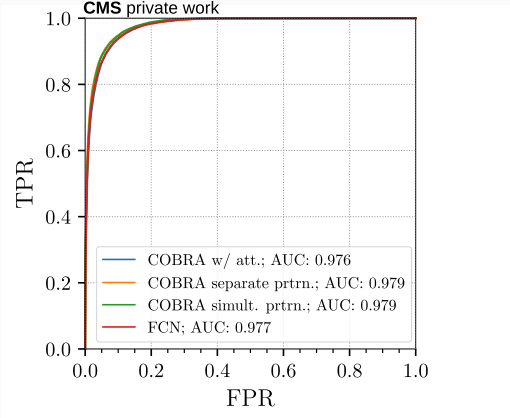
Leptonic jet att.



- combinations with higher number of jets matched have higher attention
- accuracies (separately trained classifiers):
  trijet classifier 64.35%; Higgs classifier 38.7%; leptonic top classifier 71.86%

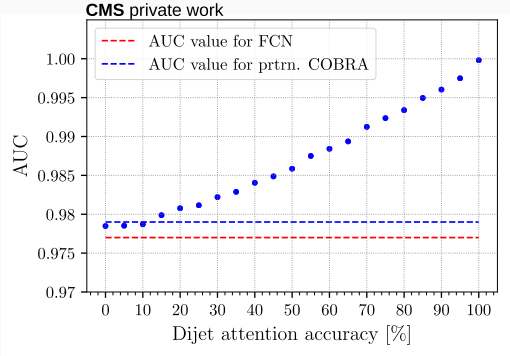
Classifier performance

optimal conditions
(completely matched sample)



- use of pre-trained classifier does not enhance performance when trained on completely matched sample
- accuracy of 38.7% of dijet classifier probably too low to cause performance increase
- provide network with ”optimal” dijet attention:
  - no jet matched $\rightarrow 0$
  - 1 jet matched $\rightarrow 1$
  - 2 jets matched $\rightarrow 2$
  - normalise scores such that the sum is one

Attention accuracy analysis



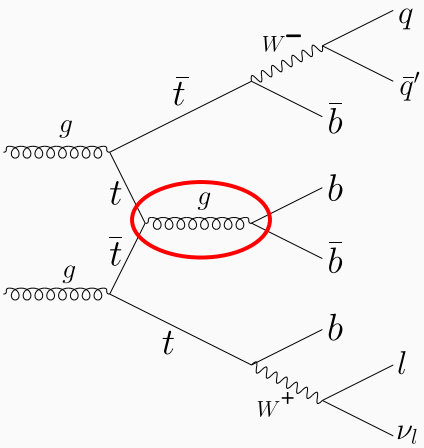
- AUC increases with provided attention accuracy
- AUC of pretrained COBRA lower than value at ≈39%
- further research in enhancing the ability of COBRA to find the Higgs necessary

Summary

- $t\bar{t}H$ important for better understanding mass generation in SM
- major challenge: irreducible background ($t\bar{t}$ + $b\bar{b}$) 8 times higher than signal
- we investigated the use of attention to improve combinatorial assignment of jets in $t\bar{t}H$
- when used for signal/background classification, attention distribution independent of number of matched jets
- if trained as classifier $\rightarrow$ attention increases for higher number of matched jets
- proof-of-principle: performance increases with attention accuracy
- pretrained classifiers show similar perfromance as FCN $\rightarrow$ accuracy of attention network needs to be improved

Backup

Background



- $t\bar{t}$ + b-jets: extra b-jets from (overlapping) b-hadrons
- $t\bar{t}$ + $c\bar{c}$: at least one extra charm jet from one or more overlapping hadrons
- $t\bar{t}$ + light flavour (LF): events which do not fit in other categories
- minor backgrounds: $t\bar{t}$-production in association with massive vector bosons or jets, production of W, Z, or γ with jets, diboson processes and single top quark production

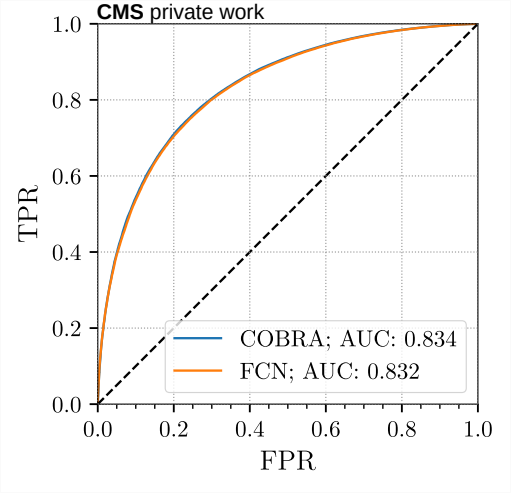
MEM method

- Problem: difficult to discriminate between $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ + $b\bar{b}$
  $\Rightarrow$ Matrix-Element-Method (MEM)
- uses both experimental and theoretical information by evaluating scattering amplitudes
- for every event, two associated weights are computed according to:

$$P(\mathbf{x}|\alpha) \propto \frac{1}{\sigma_\alpha} \int d\Phi(\mathbf{y}) \quad |M_\alpha|^2(\mathbf{y}) \quad W(\mathbf{x},\mathbf{y})$$

- α can stand for signal or background-only hypothesis
- $|M_\alpha|^2(\mathbf{y})$ is LO matrix element, $W(\mathbf{x},\mathbf{y})$ transfer function simulating detector response with $\mathbf{x}$ being measured and $\mathbf{y}$ the true parameters
- discriminant is likelihood ratio: $M_i = \frac{P(\mathbf{x}_i|sig)}{P(\mathbf{x}_i|sig)+k\cdot P(\mathbf{x}_i|bkg)}$; k =optimised parameter

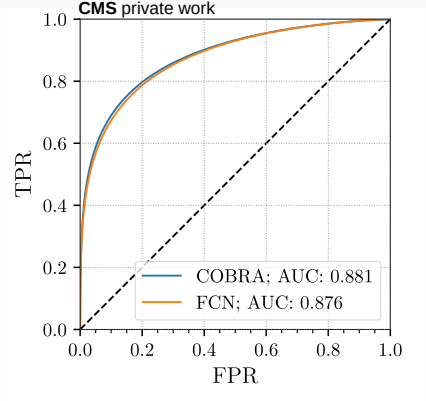
Performance comparison



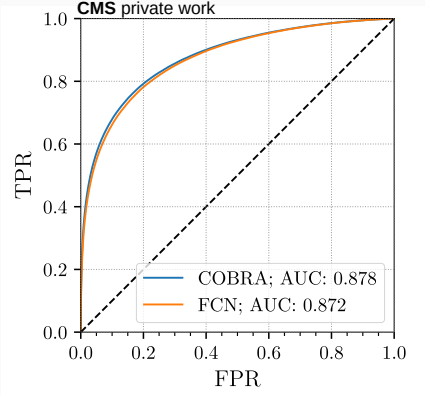
- sample: ≥ 4 jets, ≥ 3 btags
- no significant difference in performance visible
- other architectures (e.g. with pretrained classifiers) show no performance increase $\rightarrow$ rate at which correct Higgs combination is found still too low

Comparison of b-tagging algorithms

Full sample deepCSV



Full sample CSV v2



- AUC performance decrease for CSV v2
- no significant change in the difference of performance between FCN and COBRA

Performance comparison ($\geq$ 4 jets, $\geq$ 3 b-tags)

DNN model	COBRA w/ att.	COBRA w/o att.	FCN comb.	simple FCN	Current CMS
AUC	0.80	0.80	0.79	0.80	N/A
Upper stat. uncertainty	0.194	0.206	0.203	0.194	0.22
Lower stat. uncertainty	0.194	0.206	0.203	0.194	0.21

Object selection and samples

Object selection	SL
Application of all MET filters	✓
Electrons: tight ID, isolation cuts	$p_T > 30$ GeV, $\|\eta\| < 2.1$
Muons: tight ID, isolation cuts	$p_T > 26$ GeV, $\|\eta\| < 2.1$
Jets: AK4PFchs,	$p_T > 30$ GeV, $\|\eta\| < 2.4$
Loose ID with lepton removal and PU jet ID	
b-tagging via deepCSV	
MET	> 20 GeV

Samples:

- process: ttH(bb) SL; tag: 12Apr2018
- process: ttjets SL; tag: 12Apr2018_new_pmx
- number of events used in training $\approx$ 12.5 Mio.
- no further event selection used in training