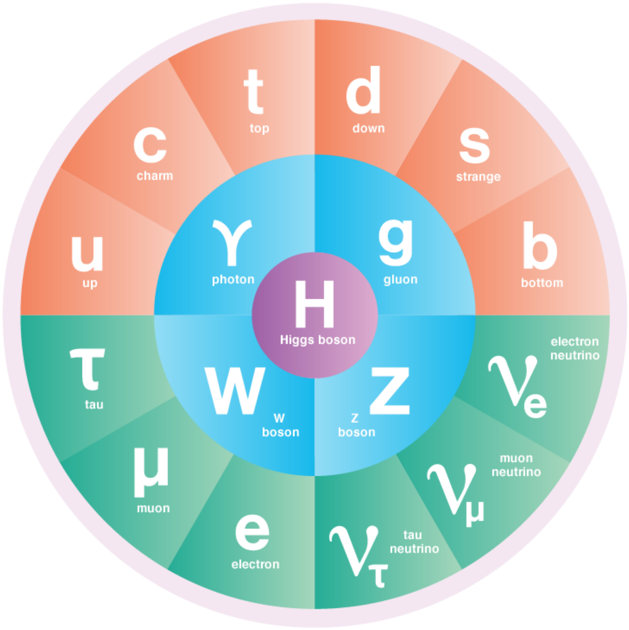
New Machine Learning based classifiers for the $t\bar{t}H(H \rightarrow b\bar{b})$ analysis

Christina Reissel

PhD Seminar 2019

The Standard Model



Higgs boson

- Higgs mechanism generates mass of fermions & weak gauge bosons
- coupling to fermions proportional to mass of fermion

Top quark

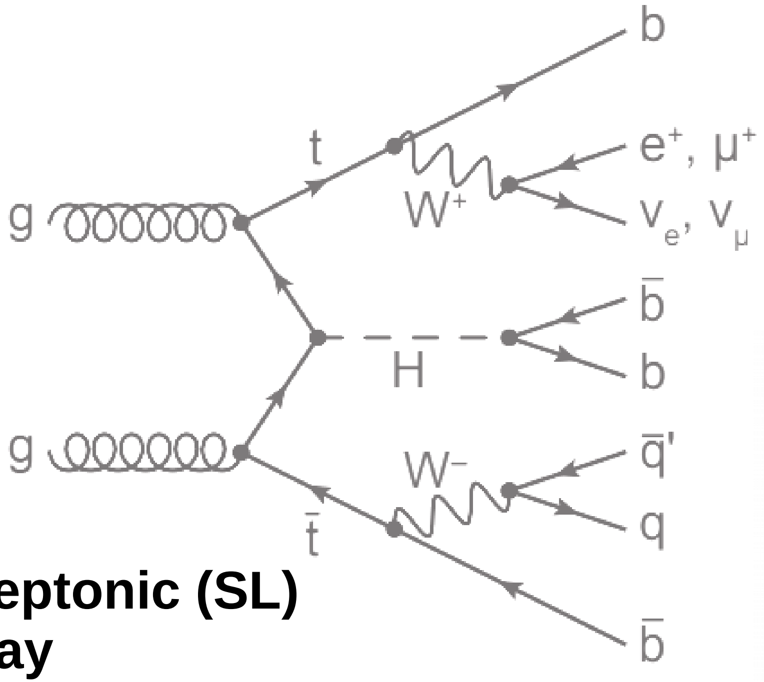
- spin ½ fermion
- heaviest elementary particle described by Standard model (SM)
  → strongest coupling to Higgs boson

Direct measurement of Top-Higgs Yukawa coupling crucial test of the SM

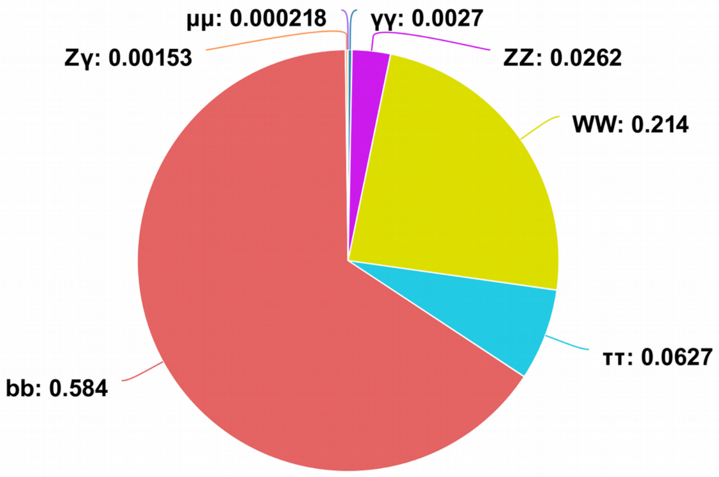
The $t\bar{t}H(b\bar{b})$ process

$t\bar{t}$H production
direct measurement of Top-Higgs coupling

ggF	48.6 pb
VBF	3.78 pb
WH	1.37 pb
ZH	0.88 pb
$t\bar{t}$H	0.50 pb

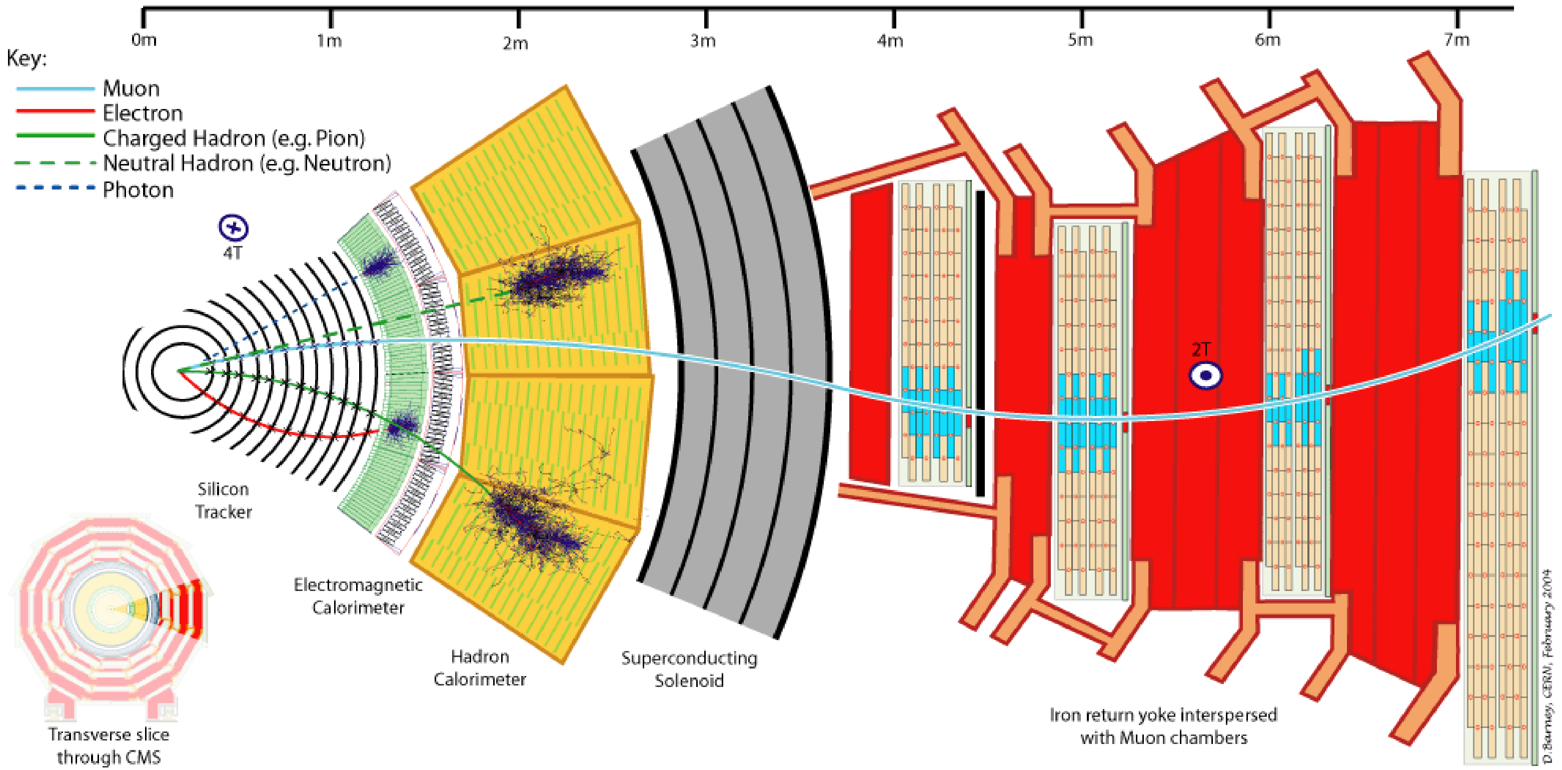


H → $b\bar{b}$
largest branching ratio
→ reasonable rates



semi-leptonic (SL) W decay
suppression of QCD background
→ $t\bar{t}$+jets dominant background

The CMS Detector



all sub-detectors of CMS experiment needed for $t\bar{t}H(b\bar{b})$ measurement

Analysis Strategy

Data

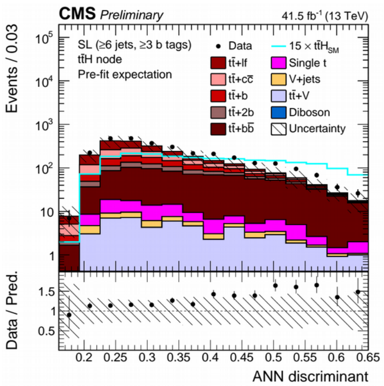
MC Simulation

Reconstruction

Filtering
- objects
  (e.g. muon pt)
- event
  (e.g. number of jets)

Classifier Calculation
(variable for optimal signal-to-background discrimination)

Application of Corrections & Calculation of Uncertainties
(e.g. different efficiencies in simulation & data, cross section calculation)



central

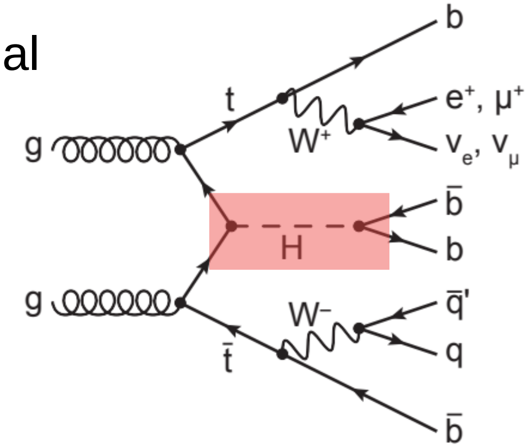
analysis specific

→ development of new analysis framework based on arrays suitable for GPU usage

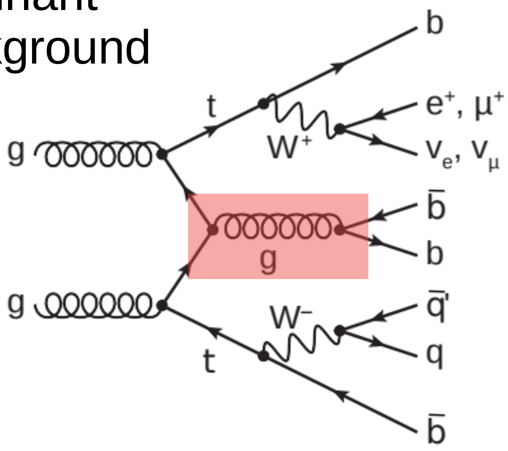
→ speed up by several orders of magnitude

Challenges

signal



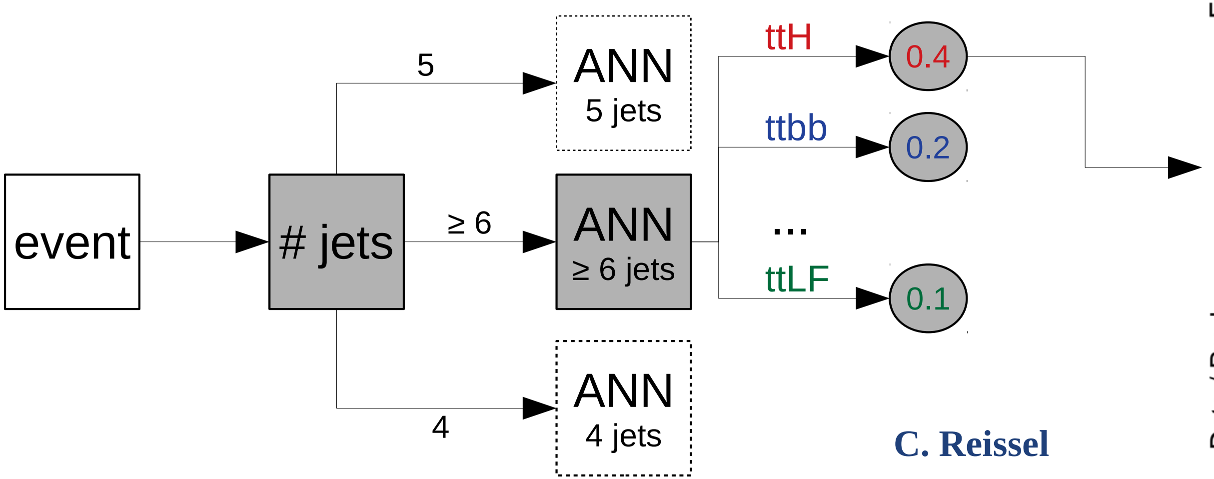
dominant background

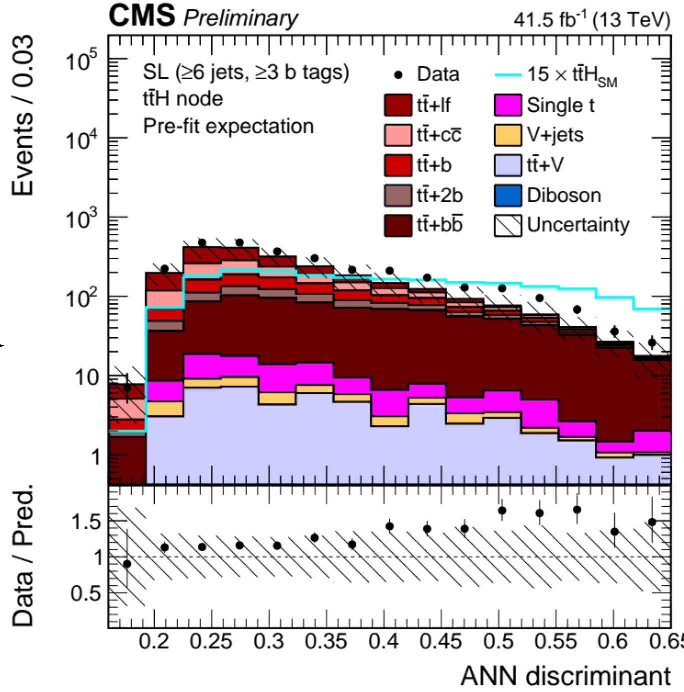


- Irreducible $t\bar{t}+b\bar{b}$ background with final state being indistinguishable from signal
- Background simulation comes with large theory modelling uncertainties
- large number of jets in final state: assignment of jets to partons not trivial → **combinatorial background**
- information lost: significant amount of events cannot be fully reconstructed because jets are out of acceptance

Current Strategy

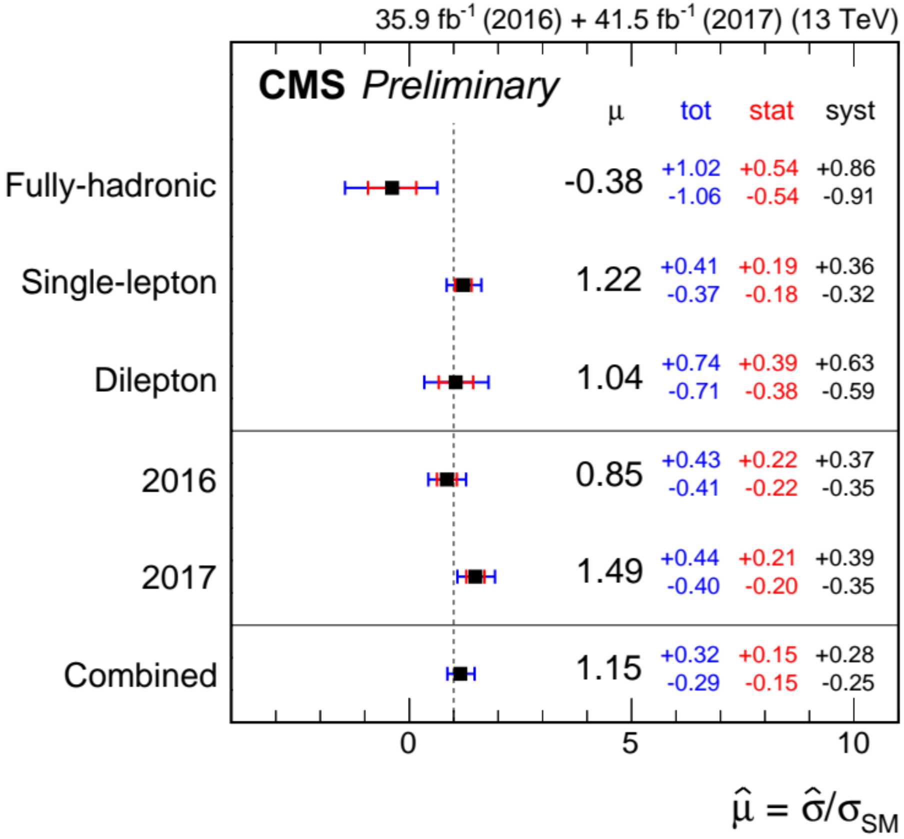
- events are classified according to number of jets, number of btagged jets and output node of ANN (artificial neural network) → results in 3x6 = 18 categories
- inputs to ANN: kinematics of leptons, jets and missing transverse energy, but also high-level variables (e.g. event shape or Matrix-Element Method discriminant)
- ANN discriminant output used also as final discriminant in fit





C. Reissel

Current Status

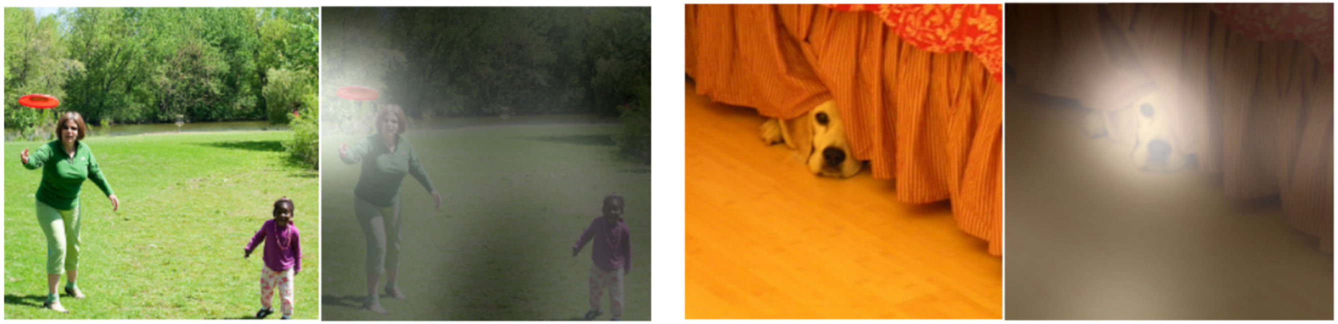


- simultaneous binned likelihood fit in all analysis categories optimizing the signal strength modifier $\mu = \sigma/\sigma_{SM}$ (σ: cross section)
- results compatible with SM expectation of μ=1
- evidence for $t\bar{t}H(b\bar{b})$ found (3.9σ), aiming for discovery (5σ)
- full $t\bar{t}H$ analysis combining Higgs boson decays in different channels observed with 5σ significance

Going beyond

- Run III & beyond: amount of data taken by CMS experiment will increase
- Important high-level variables are time consuming in computation (e.g. Matrix-Element variable takes up to 15min on single CPU per event)
- development of neural networks aiming at modeling the full event kinematics using mainly low level variables
- Model built on:
  - 4-momenta of all reconstructed objects (jet, lepton, MET)
  - for jets: b tagging information
- **solving combinatorics with the help of modern machine learning algorithms**

Attention mechanism

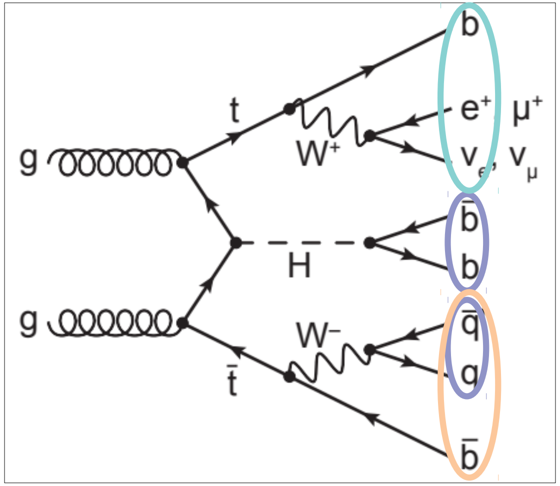


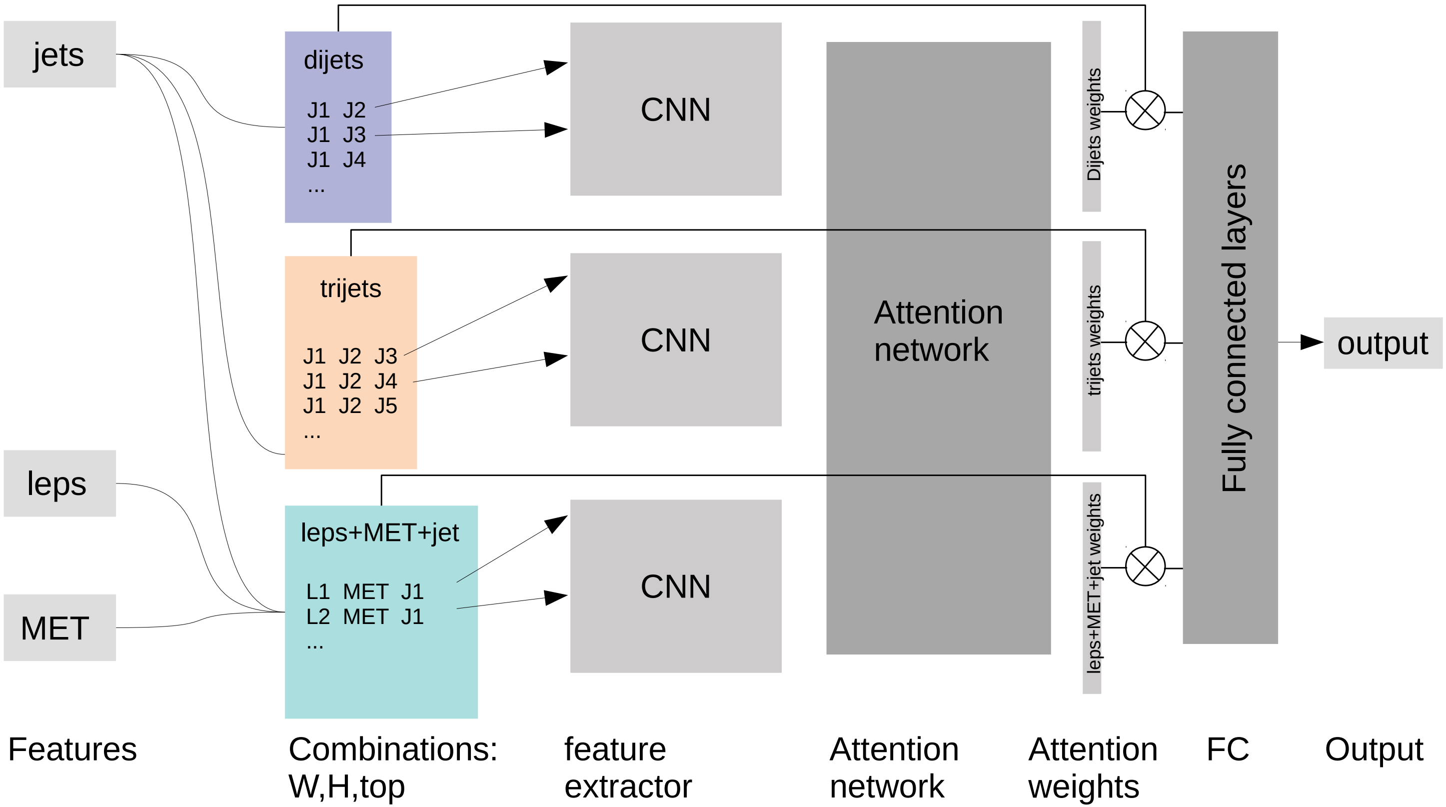
A woman is throwing a frisbee in a park. A dog is standing on a hardwood floor.

K.Xu et. al: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

- origin in image recognition and translation
- focus on special region of input phase space
- interpretation as a vector of importance weights
- application for $t\bar{t}H(b\bar{b})$: solve combinations by making a Deep Neural Network (DNN) learn how important a jet combination is in order to distinguish signal and background

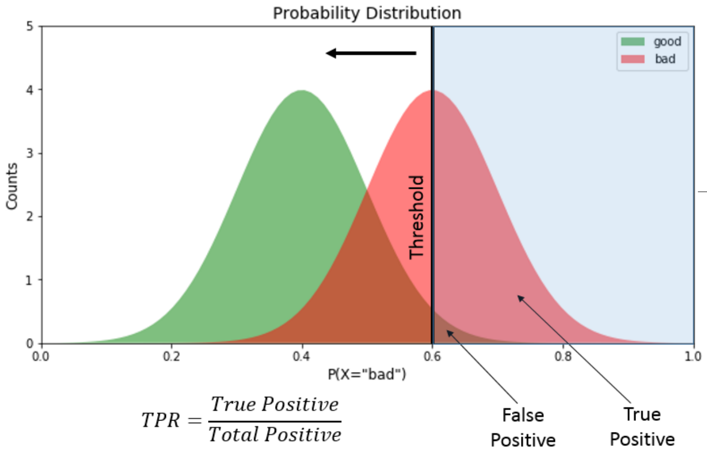
COBRA architecture





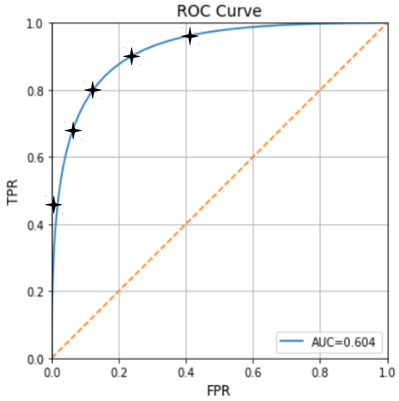
ROC curves

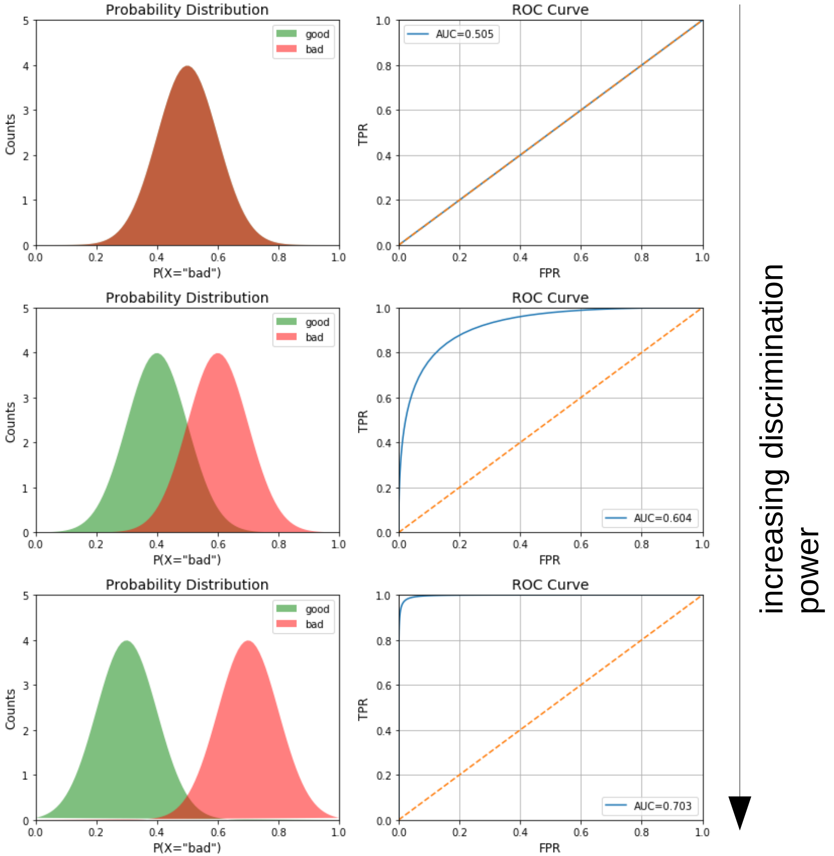
- Receiver operating characteristic (ROC) curves visualize discrimination power of discriminant
- Overall comparison often done with Area Under Curve (AUC)



$$TPR = \frac{True\ Positive}{Total\ Positive}$$

$$FPR = \frac{False\ Positive}{Total\ Negative}$$





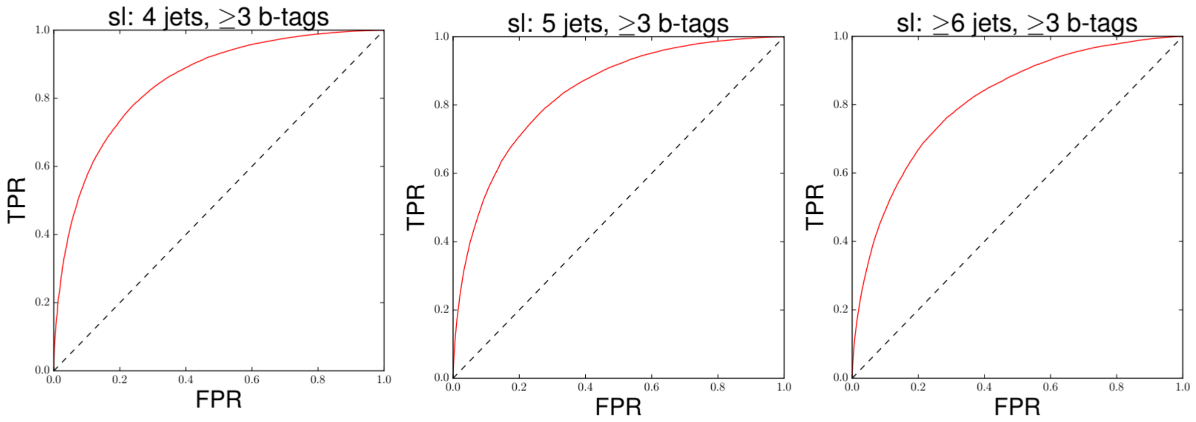
Results

AUC

	(4 jets, $\geq$ 3 b-tags)	(5 jets, $\geq$ 3 b-tags)	($\geq$ 6 jets, $\geq$ 3 b-tags)
Current CMS ✶	0.684	0.694	0.708
COBRA ✶	0.849	0.837	0.811

✶ multi classifier

✶ binary classifier



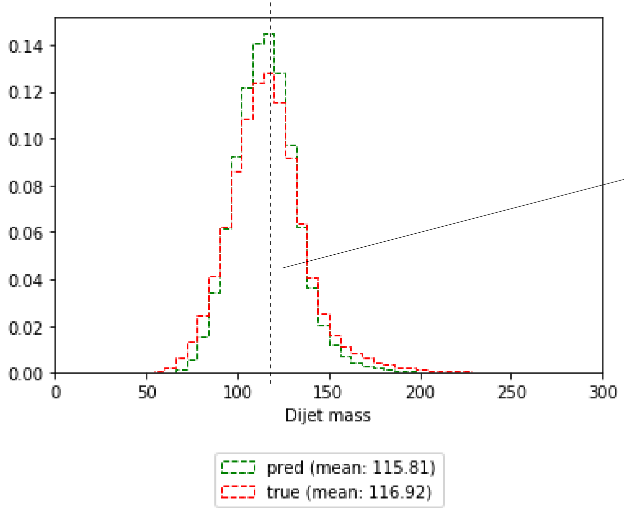
- discrimination power of COBRA with 2017 CMS simulation samples promising
- further analysis of DNN architecture needed

Ongoing work

- attention weights of COBRA can give us insights in what the DNN learns

  → unravel the black box of the DNN

- use attention information in order to reconstruct Higgs candidate or Top candidate



mass of dijet pair linked to Higgs boson peaks at approx. m_{Higgs}

- approach can be transferred to events in which Top quark system decays di-leptonic or fully hadronic

Summary

- $t\bar{t}H(b\bar{b})$ process is important test of SM
  (direct access to the Top-Higgs Yukawa coupling)
- Challenging analysis due to irreducible $t\bar{t}+b\bar{b}$ background
  & large number of jets in final state (combinatorial background)
- Current results by CMS in agreement with SM prediction, aiming for discovery
- high-level variables computationally expensive
  → development of new COBRA DNN based only on basic object kinematics
- Fast evaluation of DNN with newly developed framework able to run on GPUs
- further studies ongoing in order to understand & use COBRA

Backup

Matrix-Element Method

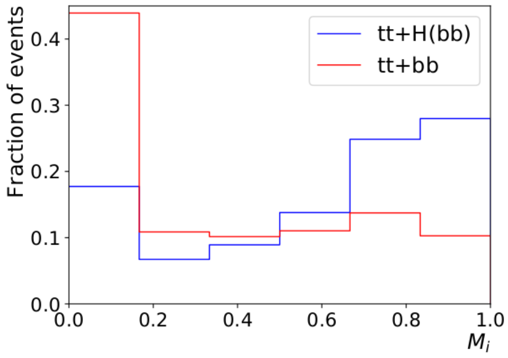
- maximal use of experimental and theoretical information via Matrix element re-weighting method
- association of a weight

$$P(\mathbf{x}|\alpha) \propto \underbrace{\frac{1}{\sigma_\alpha}}_{\text{cross section}} \int \underbrace{d\Phi(\mathbf{y})}_{\text{phase space measure}} \underbrace{|M_\alpha|^2(\mathbf{y})}_{\text{LO matrix element}} \underbrace{W(\mathbf{x},\mathbf{y})}_{\text{transfer function}}$$

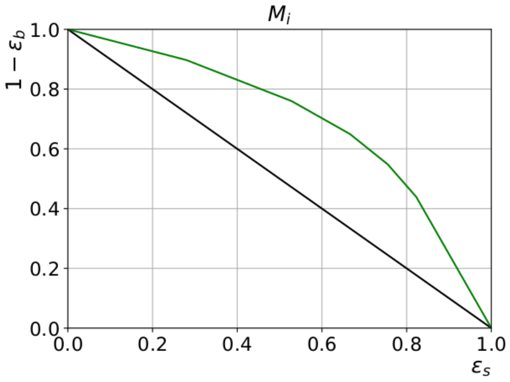
- construction of MEM-based variable

$$M_i = \frac{P(\mathbf{x}_i|S)}{P(\mathbf{x}_i|S) + k \cdot P(\mathbf{x}_i|B)}$$

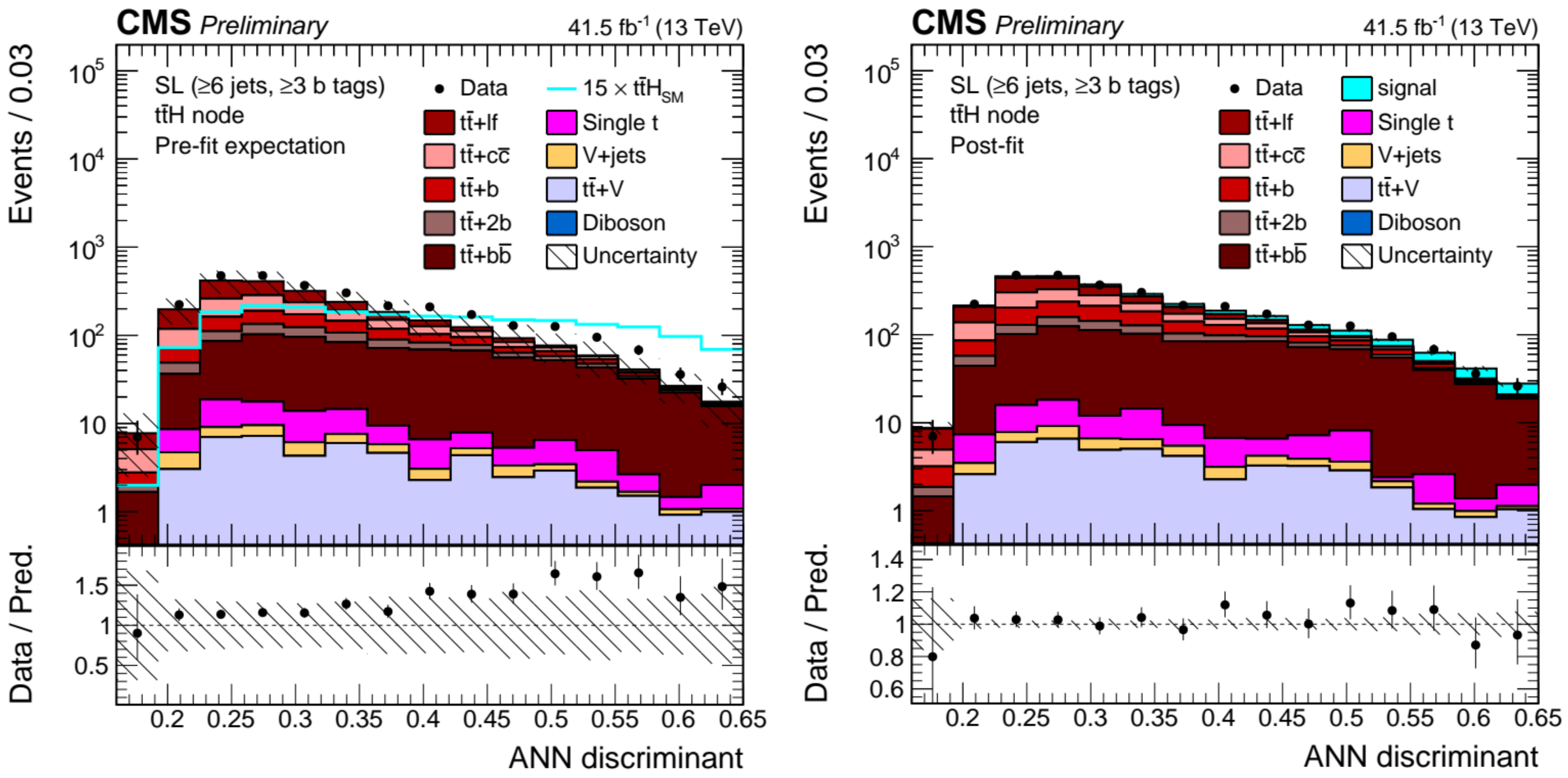
- not very sensitive to generator choice



sl: ≥ 6 jets, ≥ 4 b-tags

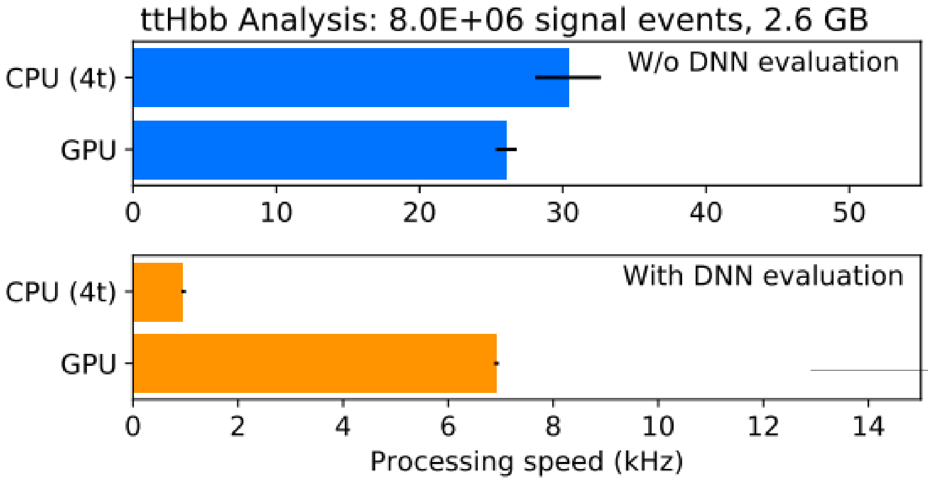


Post-fit ANN discriminant



hepaccelerate framework

- simplified, fast analysis software based on arrays
- able to run whole analysis on GPUs
  → perfectly suitable for modern machine learning (ML) applications
- fast iterations of analysis → easy optimization of ML classifiers



DNN (COBRA): approx. 745.000 pre-trained weights based on 92 inputs

GPU speed up by factor of 7 with respect to CPU