Optimizing the Machine Learning techniques used in the $H \rightarrow \tau\tau$ analysis at CMS

Speaker: Maryam Bayat Makou

DPG Physikerinnentagung 2020
Hamburg, November 2020

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Overview

- Physics of SM, LHC, CMS
- Higgs Boson
- Machine Learning
- $H \rightarrow \tau\tau$ Analysis
- Optimizing Methods in $H \rightarrow \tau\tau$
- Results of the Analysis
- Summary and Conclusion

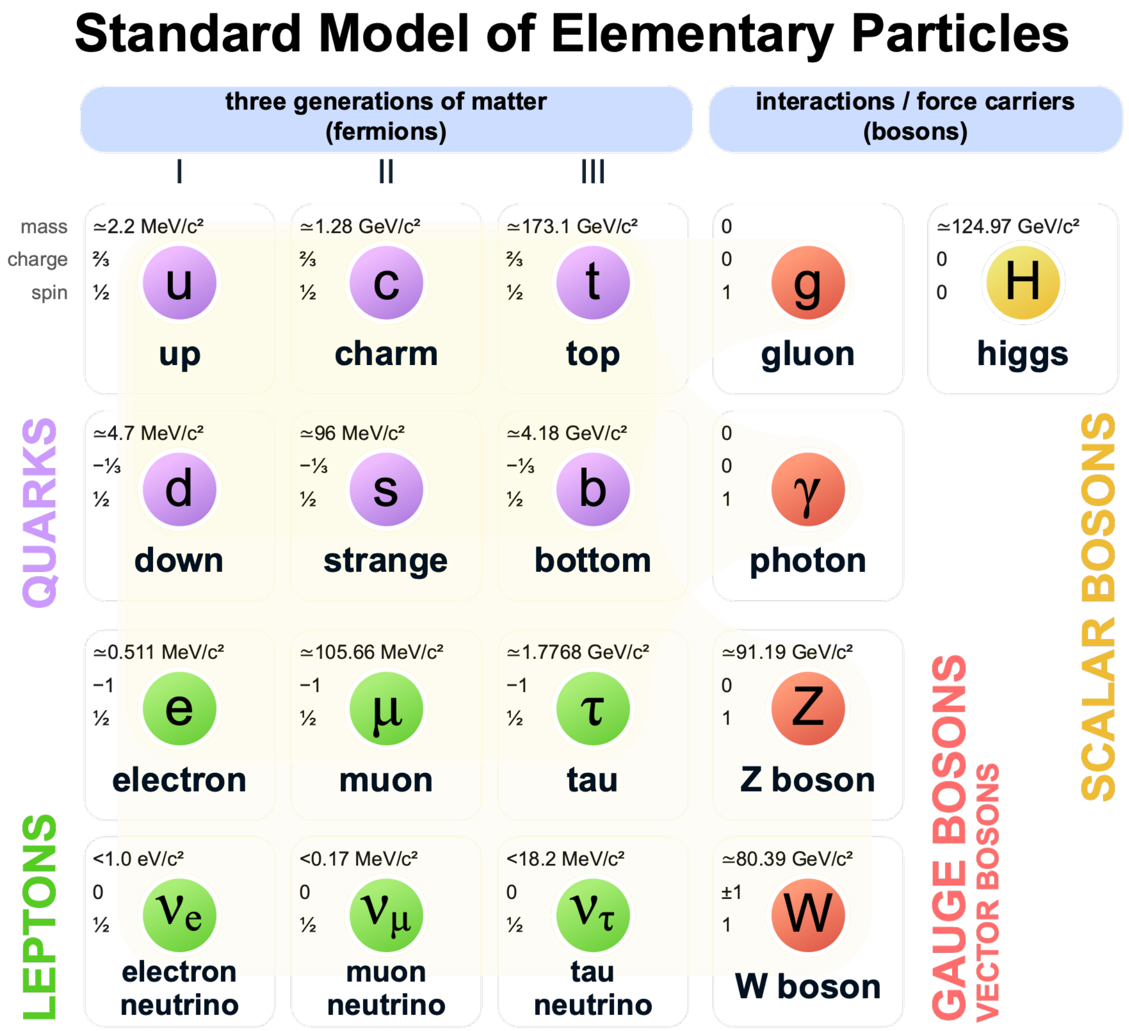
Standard Model of Elementary Particles

Exteremly Successful Theory:

- Precisely predicted several phenomena observed experimentally
- Describes 3 of the four fundamental forces and how particles interact among each other
- Explains how elementary particles acquire their mass, i.e. Brout-Englert-Higgs Mechanism

SM open Questions:

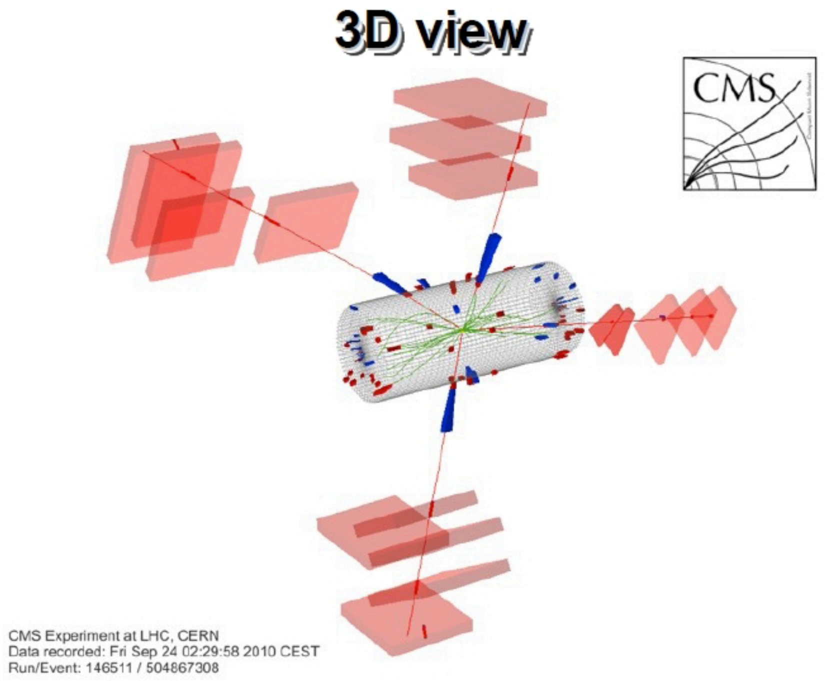
- Gravity is not described in Standard Model
- Several unexplained phenomena like Dark Matter, neutrino mass, etc…



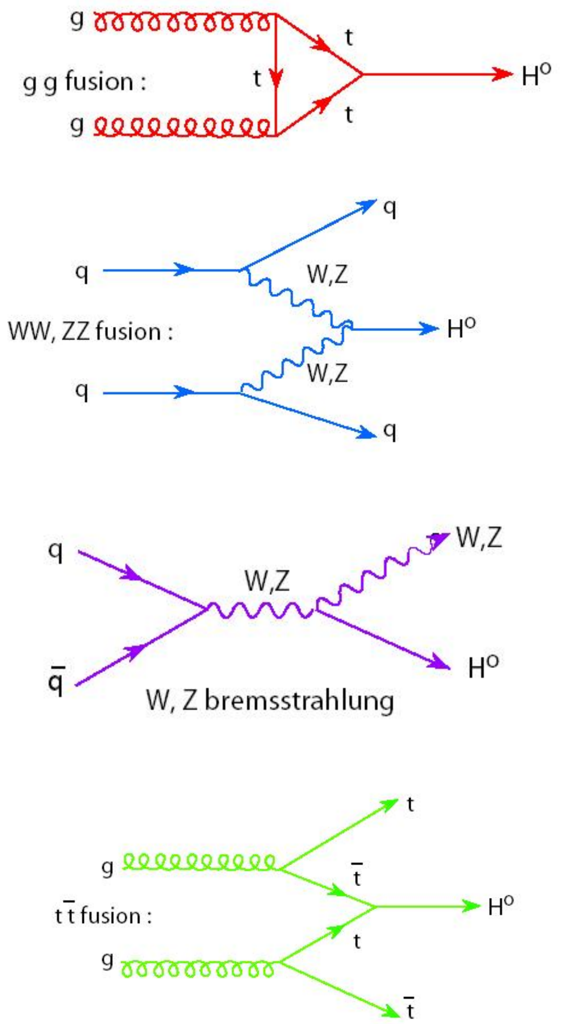
Higgs Boson

The only scalar elementary particle

- Gives mass to the gauge bosons W and Z (BEH mechanism 1964)
- Yukawa couplings to give mass to the fermions
- Discovery in July 2012 by ATLAS and CMS

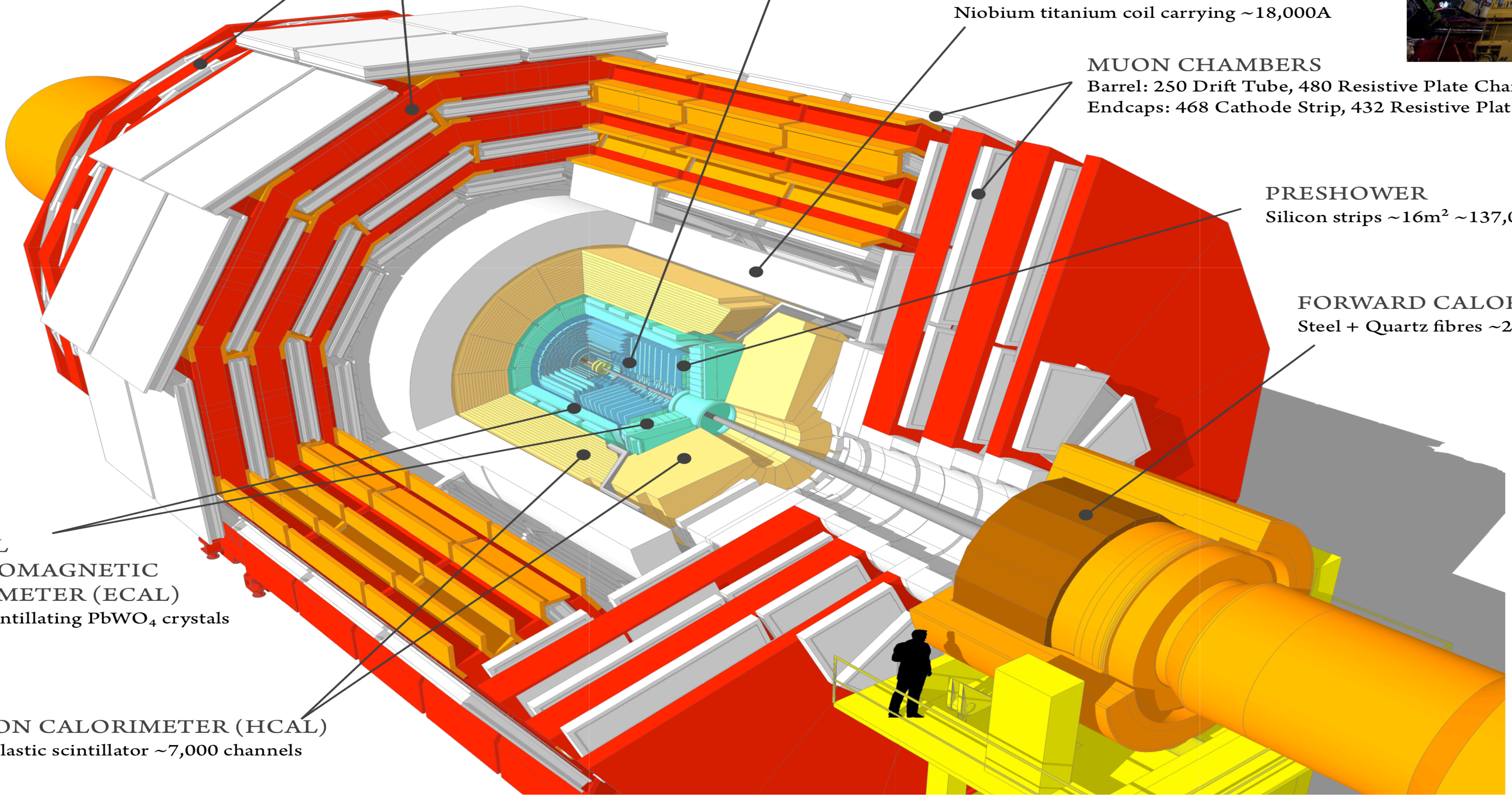


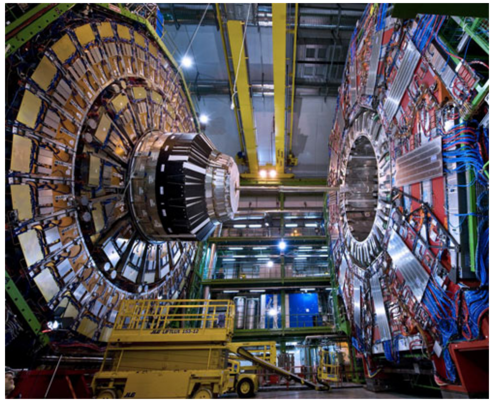
Higgs Production at LHC



Large Hadron Collider

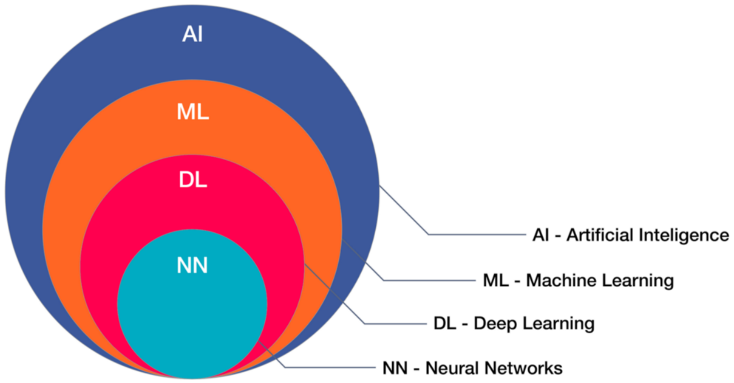
CMS Detector





Machine Learning

What is Machine Learning?



- The subfield of computer science that learns to behave intelligently based on data/experience

Using data for **answering questions**

Training — Predicting

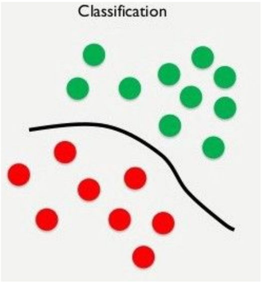
- Types of Machine Learning problems

Supervised

Learn through examples which we know the desired output.
Labeled data.

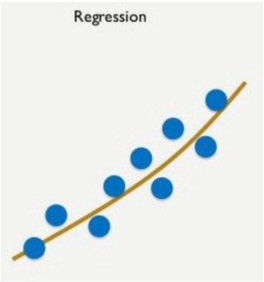
Classification

Output is a **discrete** variable (e.g., cat/dog)



Regression

Output is **continuous** e.g., price/temperature



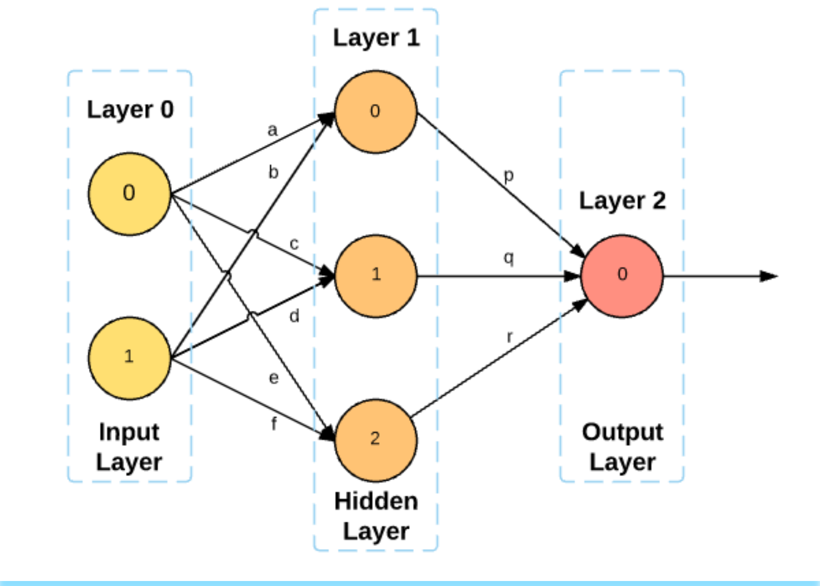
Unsupervised

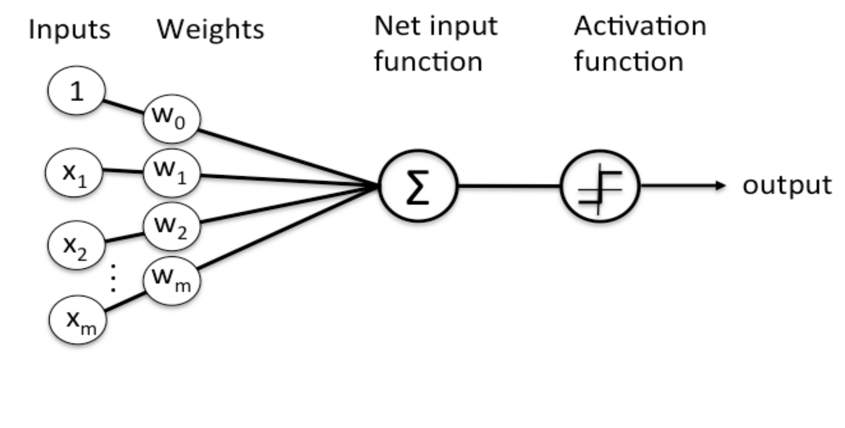
Unlabeled data

Machine Learning

Deep Neural Networks

- Neural networks (Deep Learning) is one of the algorithms used for the training
- Neural network can consist of several layers and nodes
- Given some input (x), the network calculates the output (y)
- Each input is multiplied with a weight and the output depends on the sum of inputs (and a bias term) transformed by activation function f
- Supervised learning means that we know the traget value y and we can quantify the difference between our model and the true value with the help of **loss function**
- Minimize the loss → Optimization of parameters with the help of **Gradient descent** method

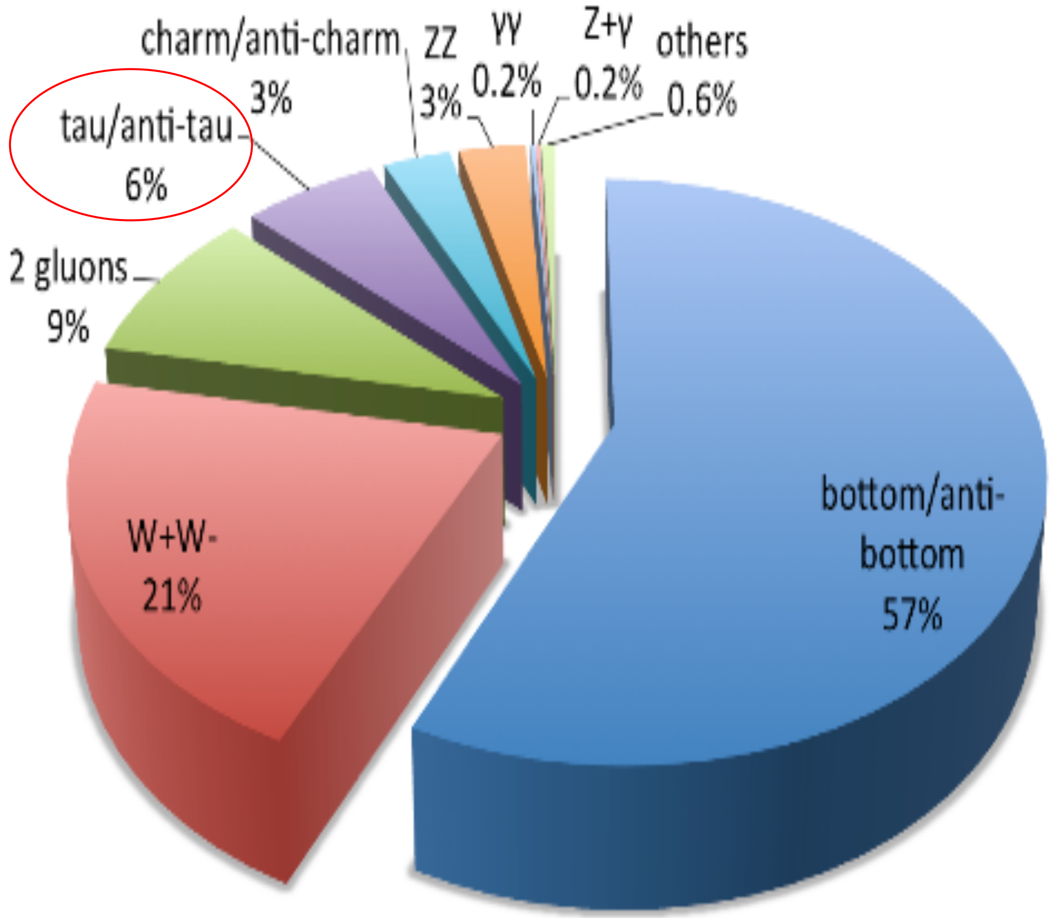




Schematic of Rosenblatt's perceptron.

Higgs decay to tau leptons

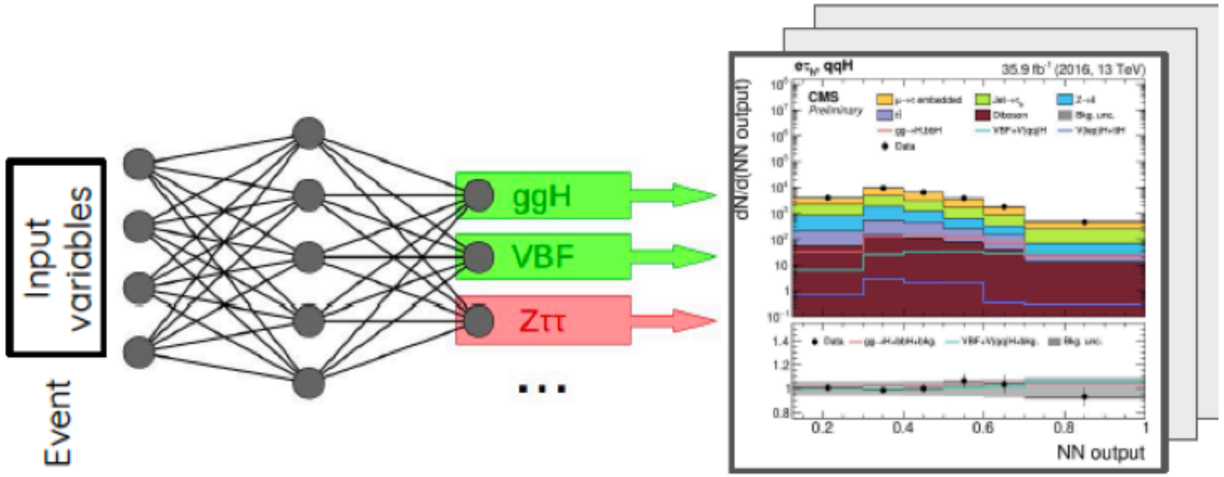
- Precise measurements of Higgs Yukawa couplings important test of SM
- H → $\tau\tau$ second largest branching ratio for Higgs to fermions, lower background than H → bb
- Higgs coupling to fermions was discovered in its decay to two tau leptons
- Observation of H → $\tau\tau$ at CMS only in 2017

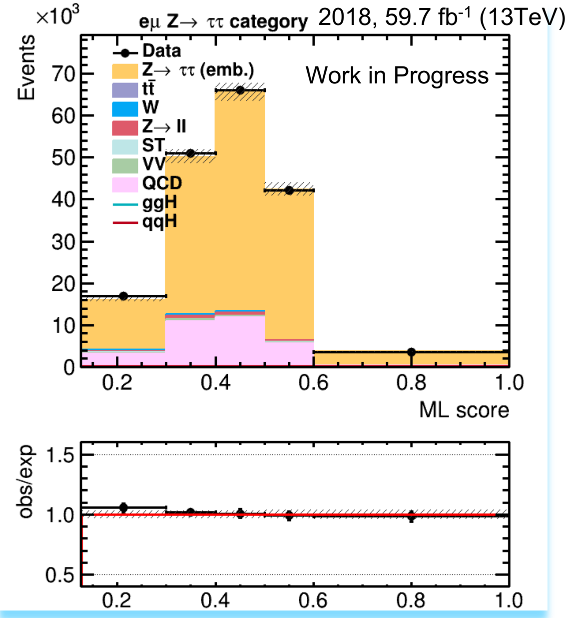


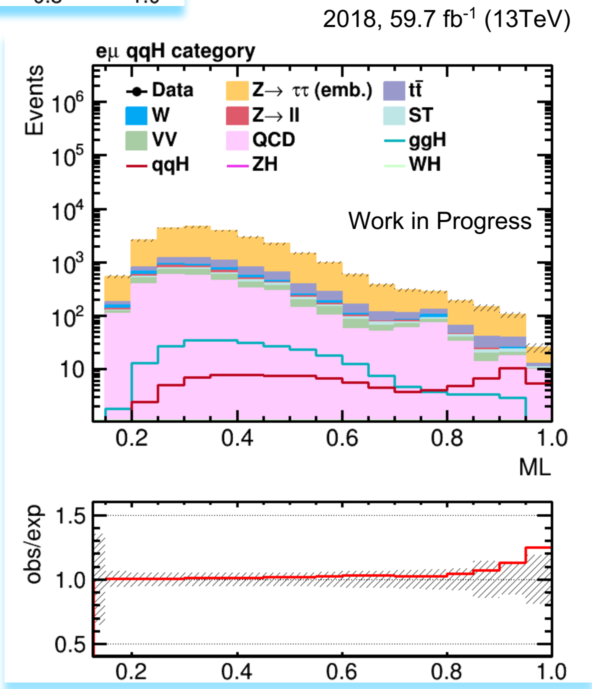
HTT Analysis

CMS Public Analysis Summary HIG-18-032

- 2016+2017 data, 77.4 fb^{-1} ($\sqrt{s}$ = 13 TeV)
- Four final state of $\tau\tau$ pair studied : **eμ, eτ_h, $\mu\tau_h$, $\tau_h\tau_h$**
- Multi-class NN , 2 signal classes (ggh and VBF) & several background classes (e.g. ZTT, ttbar, QCD, etc)
- Output : “probability” that event belongs to a certain class







Performance of NN

multi-class NN for event classification

- Performance of NN can be shown in **Confusion Matrix**

✓ **qqh : 74% of qqh true events are classified as qqh events**

✓ **ggh : only 20% of ggh true events are classified as ggh, 26% as ZTT!**

- Goal of Master thesis: NN performance improvement to better discriminate ggh from ztt.

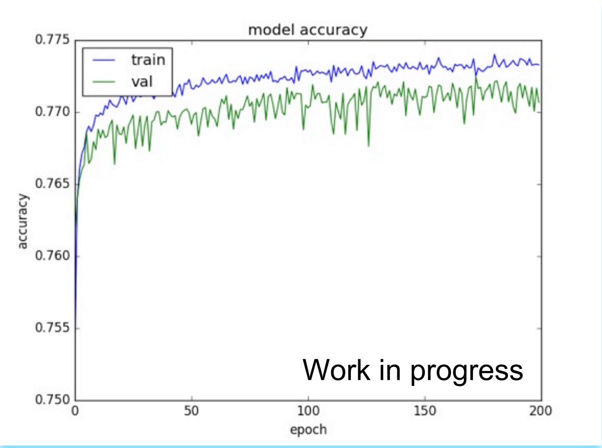
eμ (2017) CMS *Simulation Preliminary*

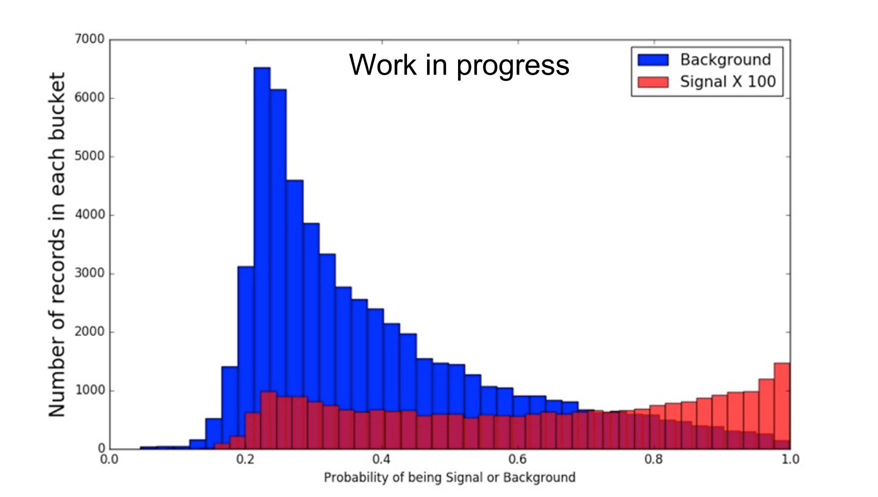
NN predicted event class \ True event class	ggH	qqH	ztt	qcd	tt	misc	db	st
ggH	0.20	0.05	0.12	0.06	0.01	0.11	0.08	0.03
qqH	0.26	0.74	0.13	0.06	0.16	0.09	0.07	0.17
ztt	0.26	0.03	0.52	0.24	0.00	0.16	0.07	0.01
qcd	0.07	0.03	0.11	0.45	0.03	0.18	0.05	0.04
tt	0.02	0.07	0.02	0.03	0.55	0.05	0.05	0.27
misc	0.07	0.02	0.05	0.11	0.02	0.24	0.07	0.05
db	0.08	0.02	0.03	0.04	0.04	0.10	0.46	0.12
st	0.03	0.04	0.01	0.02	0.19	0.06	0.14	0.30

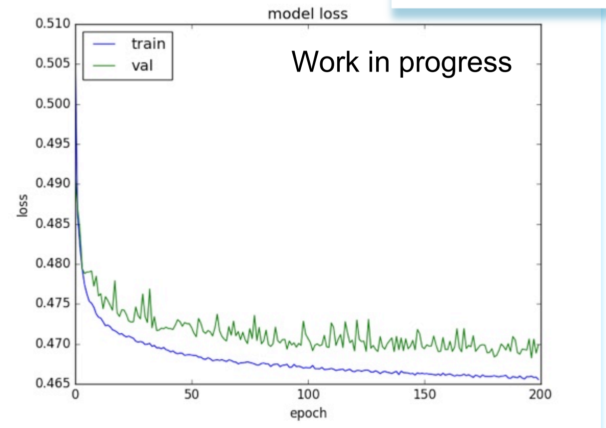
CMS PAS HIG-18-032

Signal(ggh) / bkg (ZTT) separation

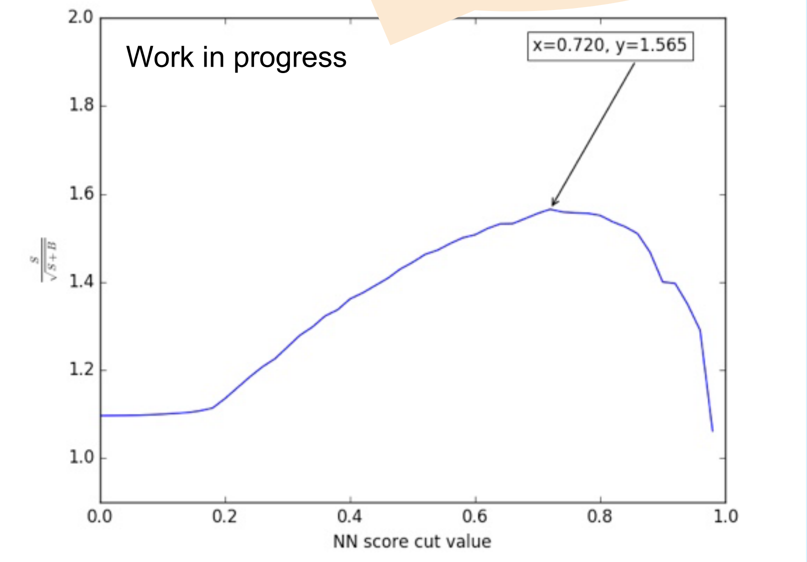
14 input variables :momenta of e, μ and m_{vis} ,etc → Preprocessing of input data → Feed forward NN to classify ggh / ZTT → 2 hidden layers, 100 nodes each → Test splitting: 30%, Validation splitting:25 %







Using the $S/\sqrt{S+B}$ to observe the NN Performance

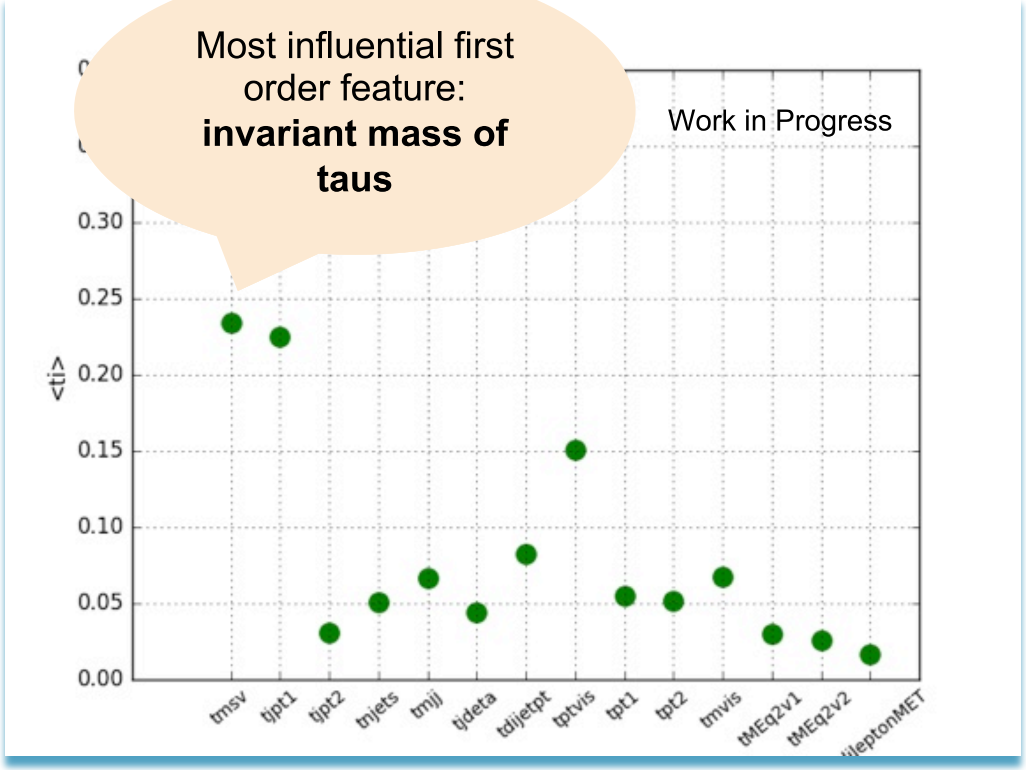


3.Taylor expansion of NN output wrt input

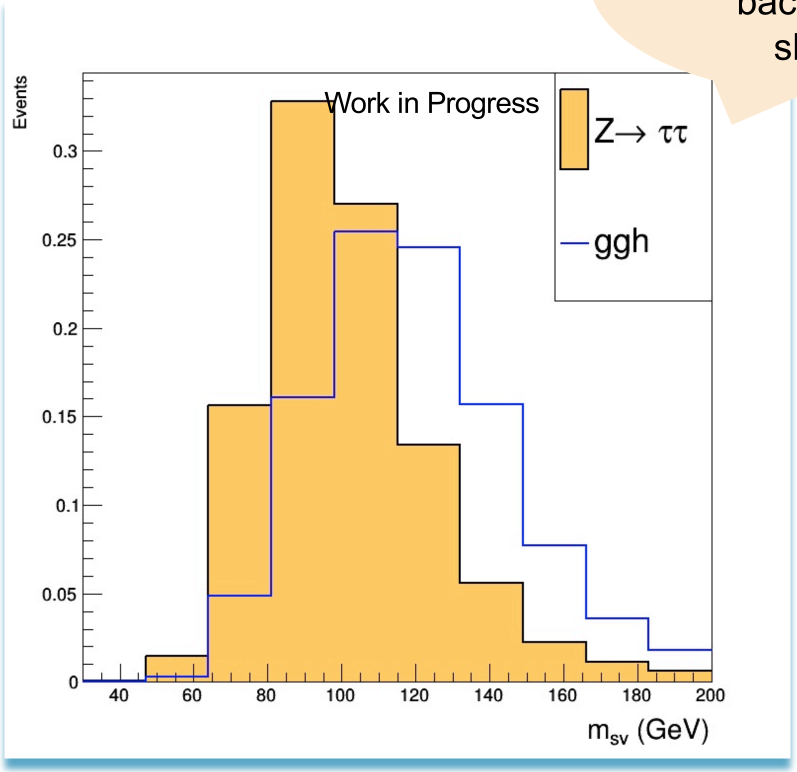
Based on KIT publications : https://arxiv.org/pdf/1803.08782.pdf

- Taylor expansion of the full NN function up to second order, which allows to connect the input space directly to the output
- Taylor coefficients identify those characteristics of the input space that have a large influence on the NN output
- First order features: captures the influence of single input elements on the NN output

Most influential first order feature: **invariant mass of taus**



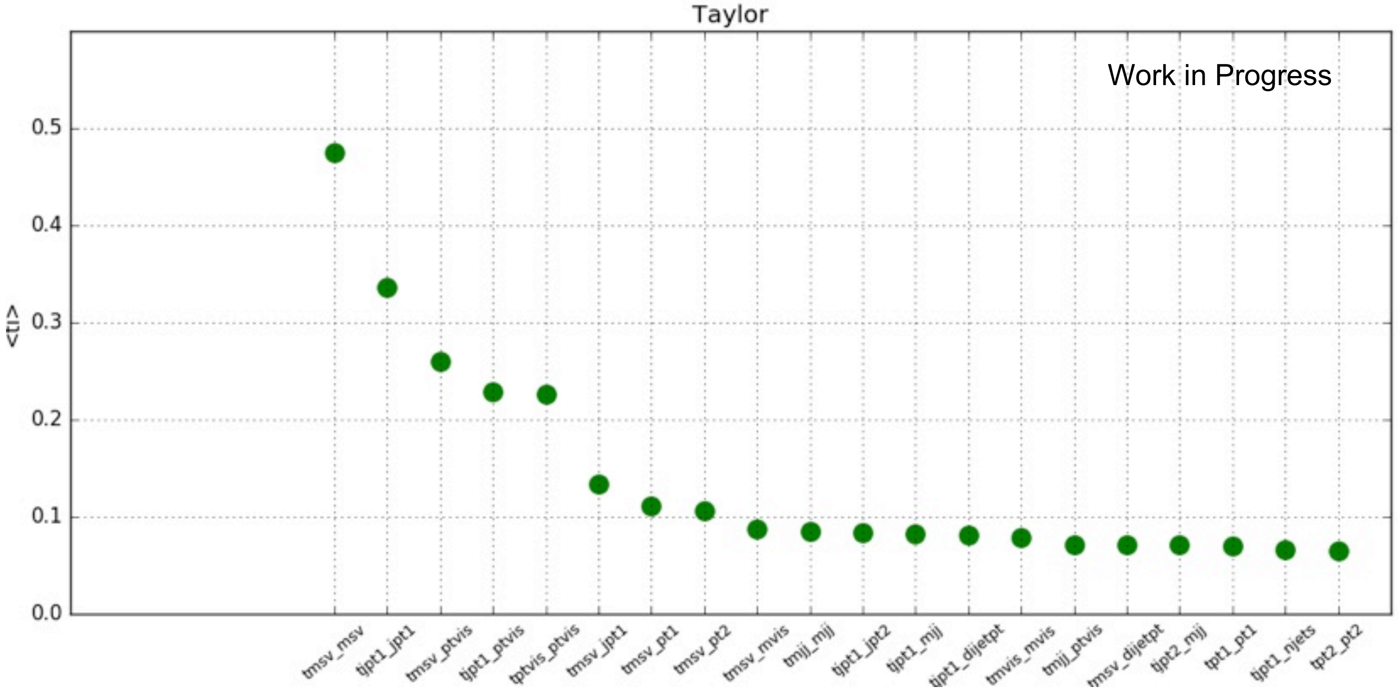
Difference of signal and background shapes



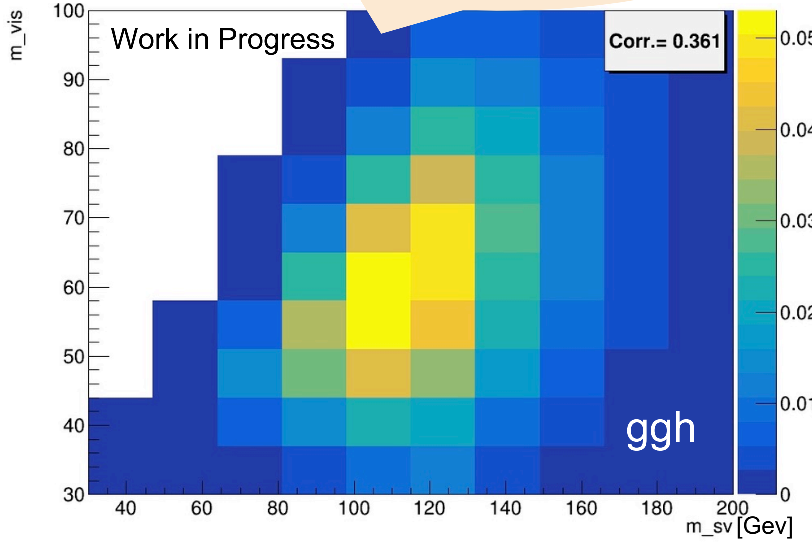
3.Taylor expansion of NN output wrt input

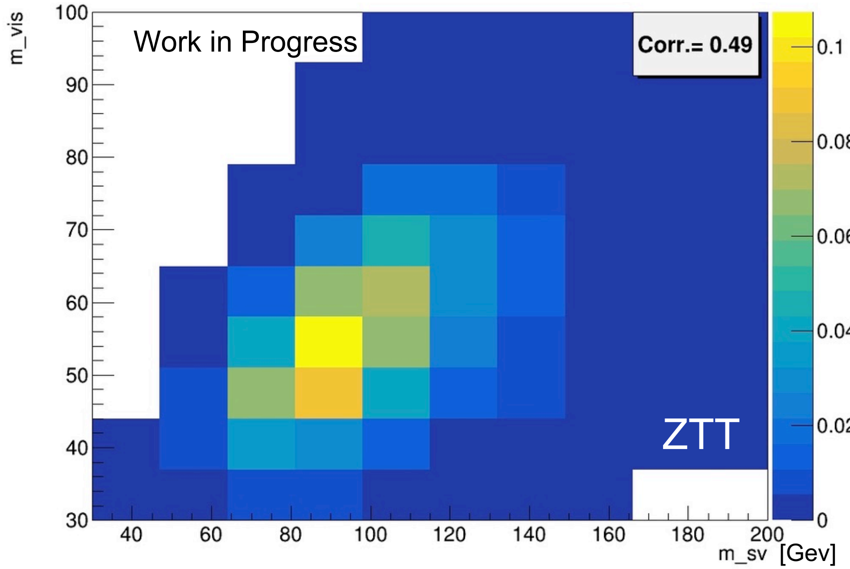
second oder coefficients

- Second order coefficients: captures the influence of pair-wise or auto-correlations among the input elements.
- Plot of second order Taylor coefficients showes only the first 20 maximum value coefficients



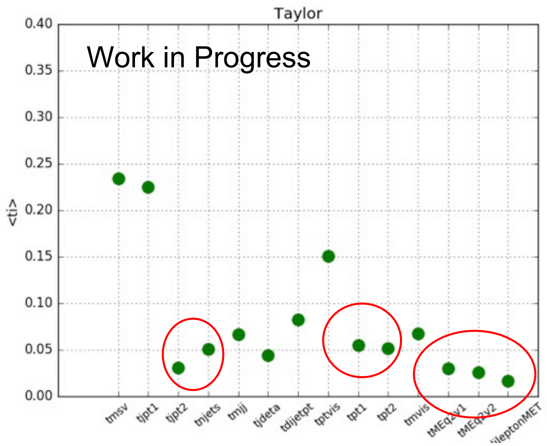
Variable correlations of $e\mu$ channel



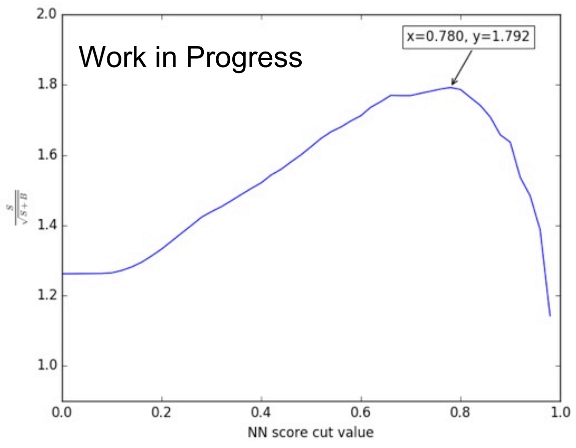


Optimizing input variables

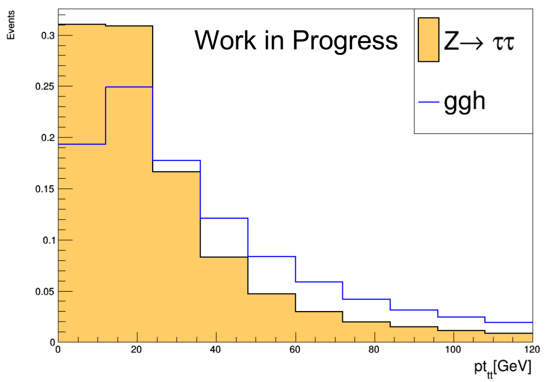
Omitting the low value Taylor coefficients → Observing the result without these coefficients → Finding 14 new variables → Train and observe the result with new variables



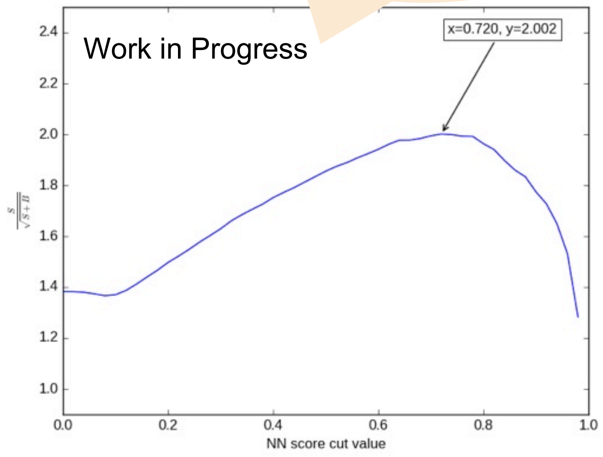
Dimensional reduction from 14 to 6



New variables selection based on discriminating power of their correlation distributions in signal vs background



- **NN improvement has been observed**
- $^{s}/_{\sqrt{s+B}}$ **= 2.002**



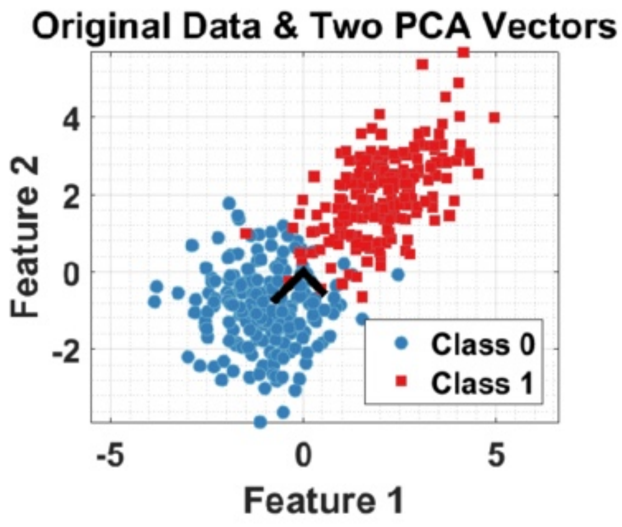
What is PCA

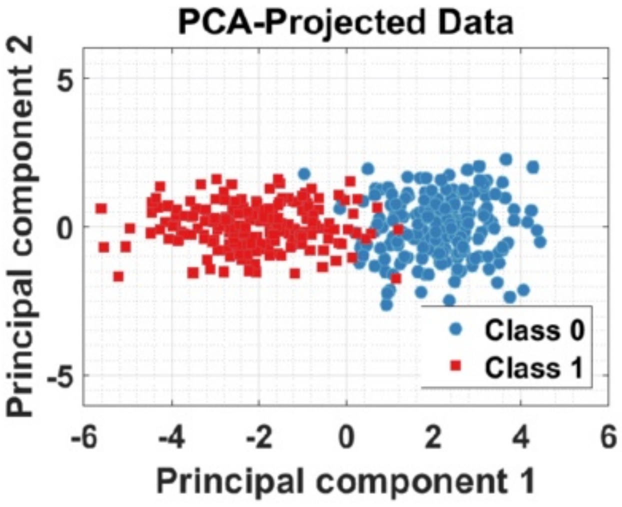
Principal Component Analysis

- Set of a data as a linear combination of an **orthonormal** set of vectors which define a new coordinate system

PCA procedure

- Find eigenvalues and eigenvectors
- Decide which are significant
- Form a new coordinate system defined by the significant eigenvectors
- Map data to the new space





why PCA?

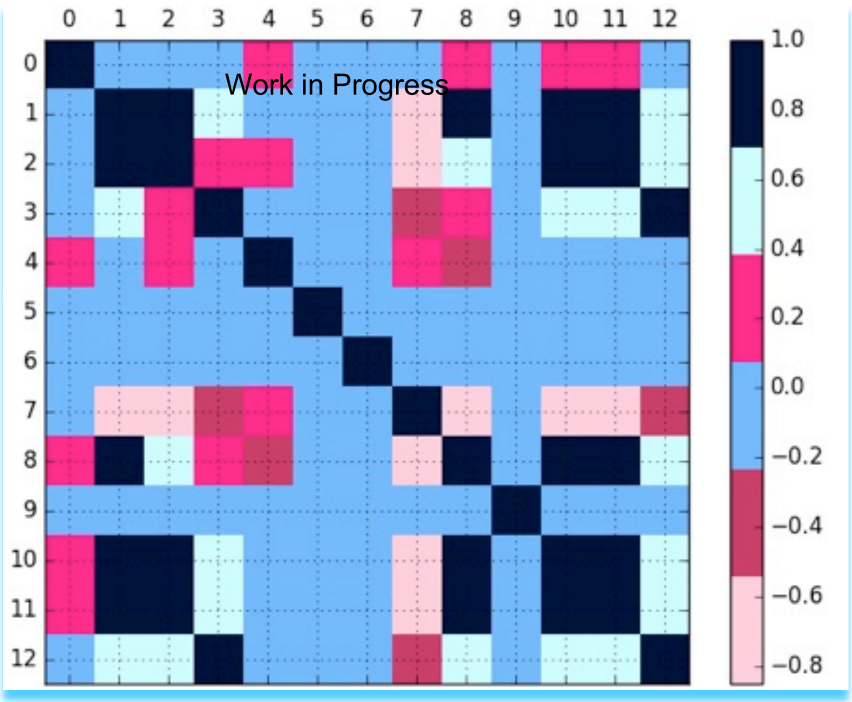
- PCA as a dimension reduction technique
- PCA as a tool for input variables ***decorrelation*** and ***performance improvement***

Overview on input data correlation before and after PCA fit

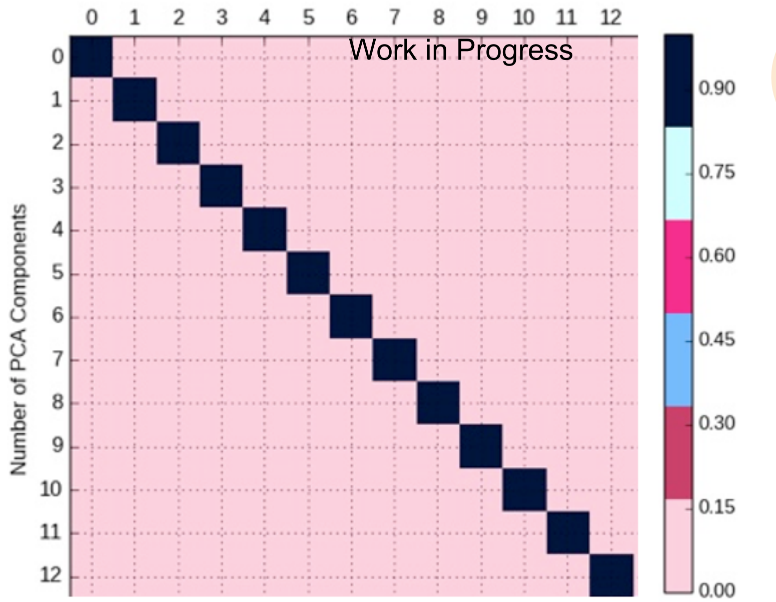
the principal components are perpendicular to one another, therefore they are statistically linearly independent of one another

List of input variables:
0: m_sv,
1: jpt_1,
2: ptvis,
3: jdeta,
4: mvis,
5: d0_1,
6: d0_2,
7: dr_tt,
8: met,
9: eta_1,
10: pt_sv,
11: pt_tt,
12: pt_ttjj

Correlation Matrix Plot of 13 input variables



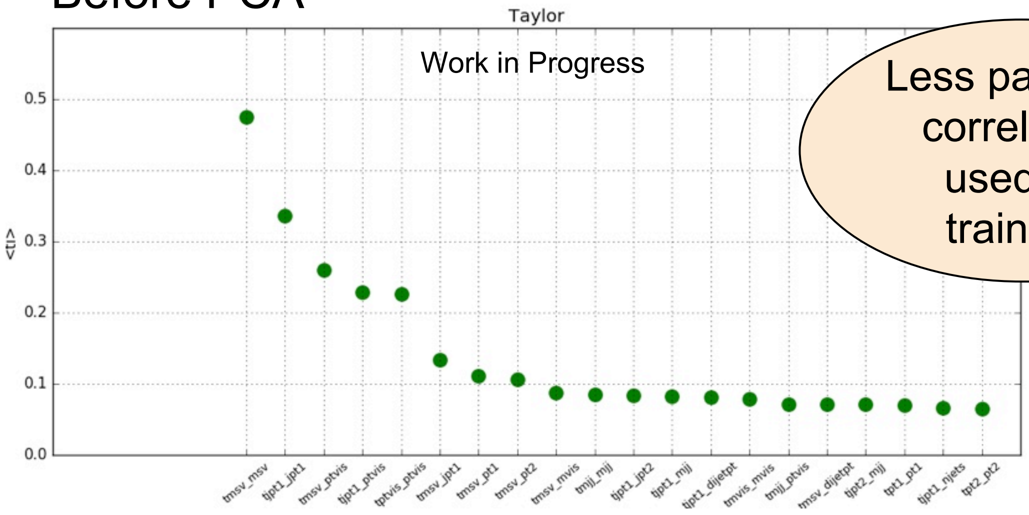
Correlation Matrix Plot after PCA fit

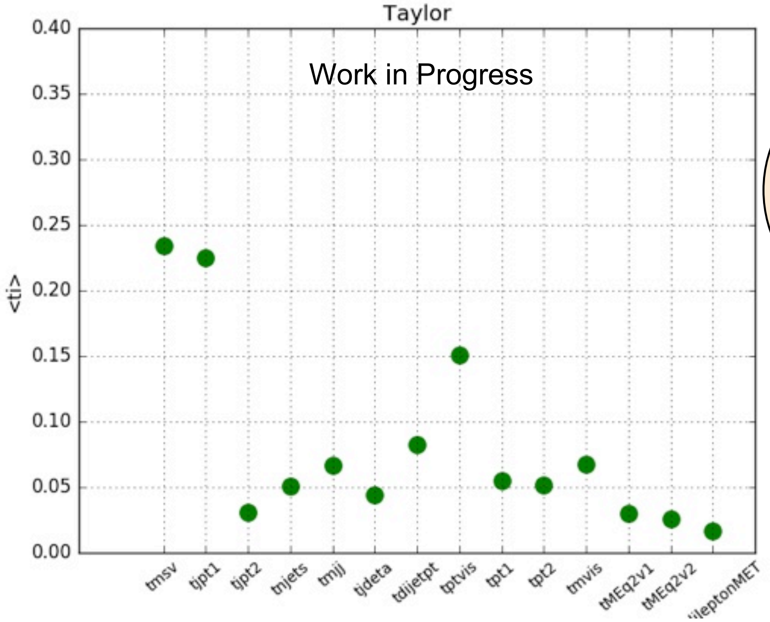


Conclusion: as expected input variables are decorrelated

Taylor coefficients before and after PCA application

Before PCA

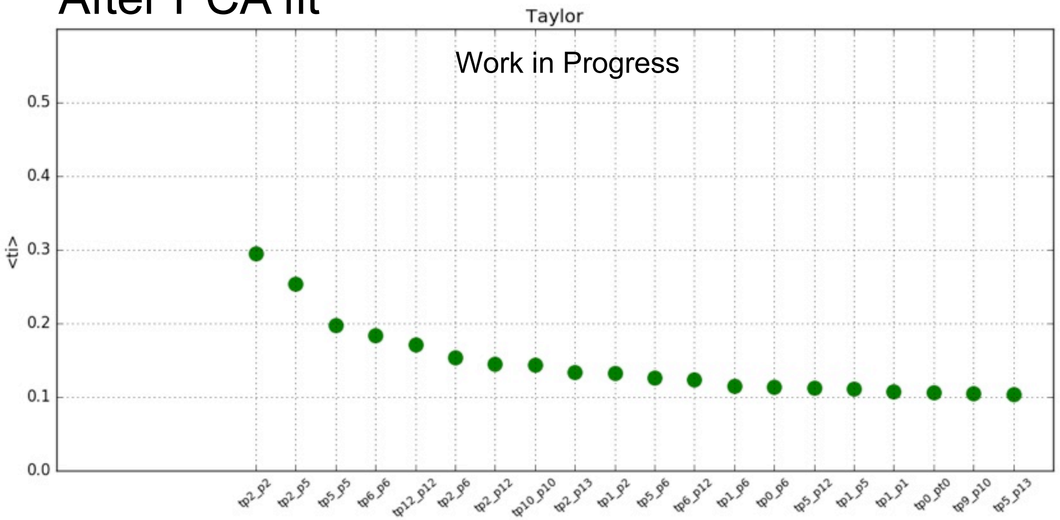


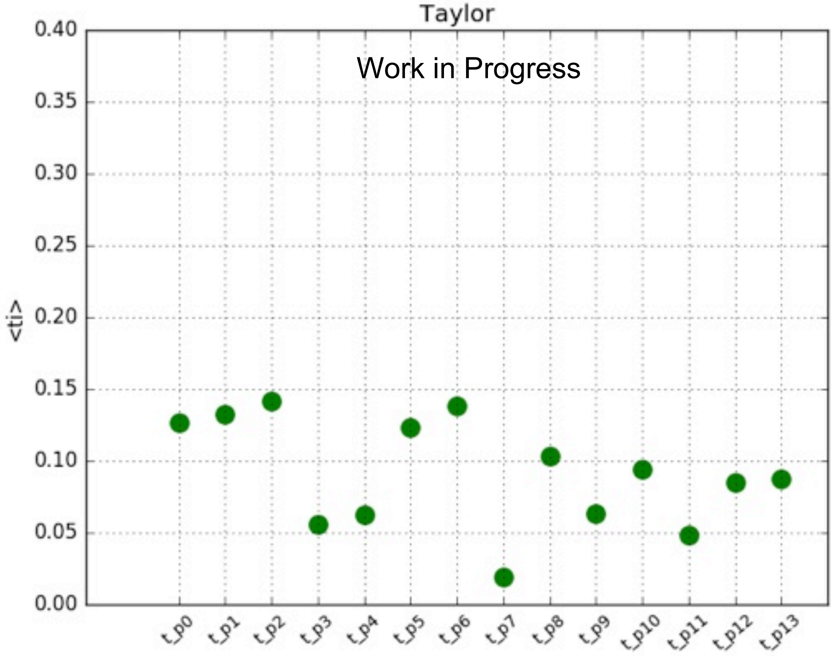


Less pair-wise correlation used for training.

Variables have similar influence in the training.

After PCA fit



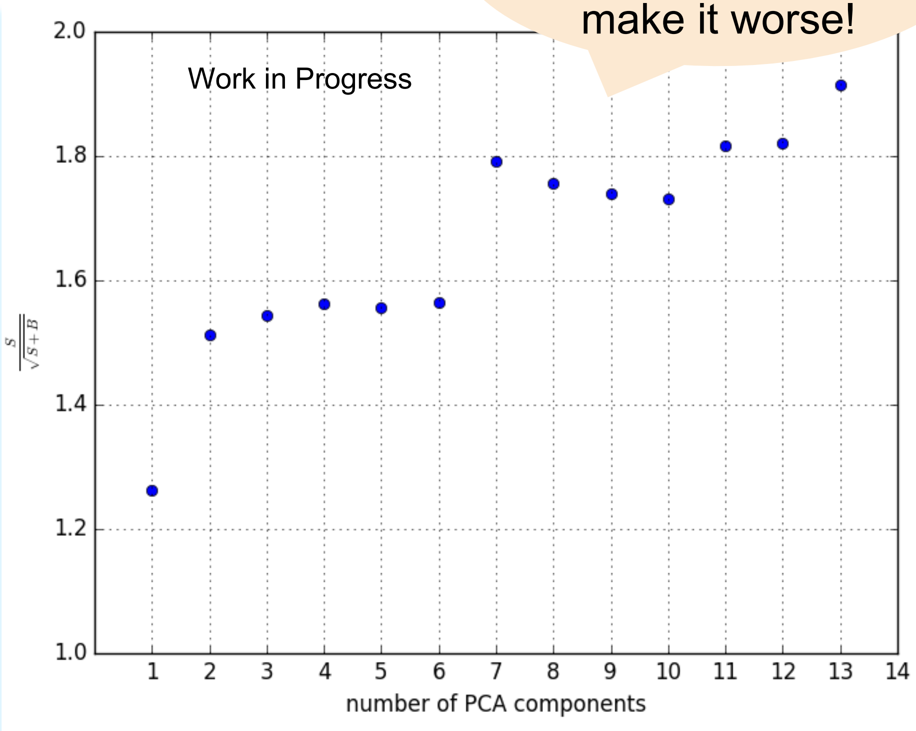


Apply PCA on Signal/Background classification

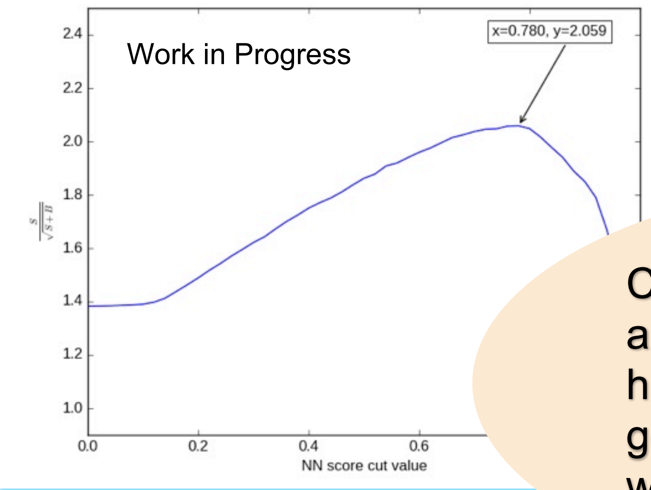
2. PCA for performance Improvement

- Monitoring the s/$\sqrt{s+b}$ with new input variables set after PCA fit:

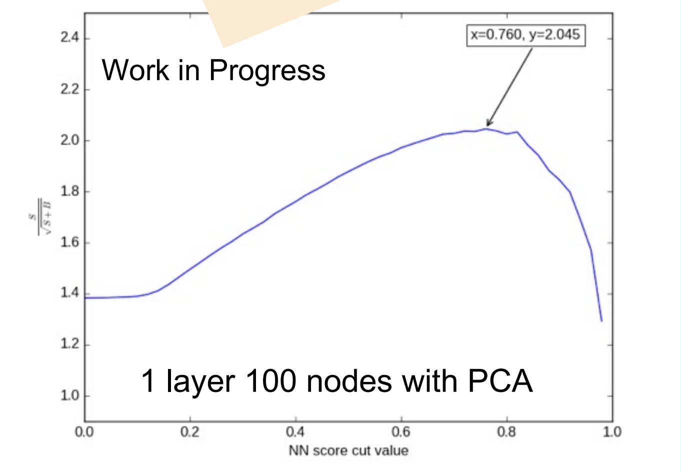
PCA neither improves the performance nor make it worse!



2 layers each 100 nodes with PCA



Conclusion: PCA helps to train an easier Network (e.g: less hidden layers, less nodes) and get the same results as before with a complex network !



Summary and Conclusion

- Results on classification of ggh/ZTT have been shown and NN performance has been optimized.
- The **ML optimization** may improve the measurements on the Higgs boson properties in its decay to two tau leptons
- With inclusion of PCA the complexity of the NN can be reduced, this has several advantages:

  1. reduces the training time without sacrificing accuracy
  2. the number of parameters in the NN is reduced, making the NN less prone to overtraining

- Promising perspective on NN applications to Higgs searches with leptonic decay of taus.

Thank you.

BACK UP

List of the input variables

- p_T (first tau cand)
- p_T (second tau cand.)
- Transverse mass (di-lepton,MET)
- p_T leading jet and subleading jet
- Number of jets
- Mass of dijet system
- $\Delta\eta$ (2 leading jets)
- p_T (2 leading jets)
- SVFit di tau mass
- Visible di-tau p_T
- MELA energy transfer quantities

Identifier

- Pt_1
- Pt_2
- mTdileptonMET
- jpt_1 and jpt_2
- njets
- Mjj
- Jdeta
- Dijetpt
- M_sv
- M_vis
- Pt_vis
- ME_q2v1
- ME_q2v2

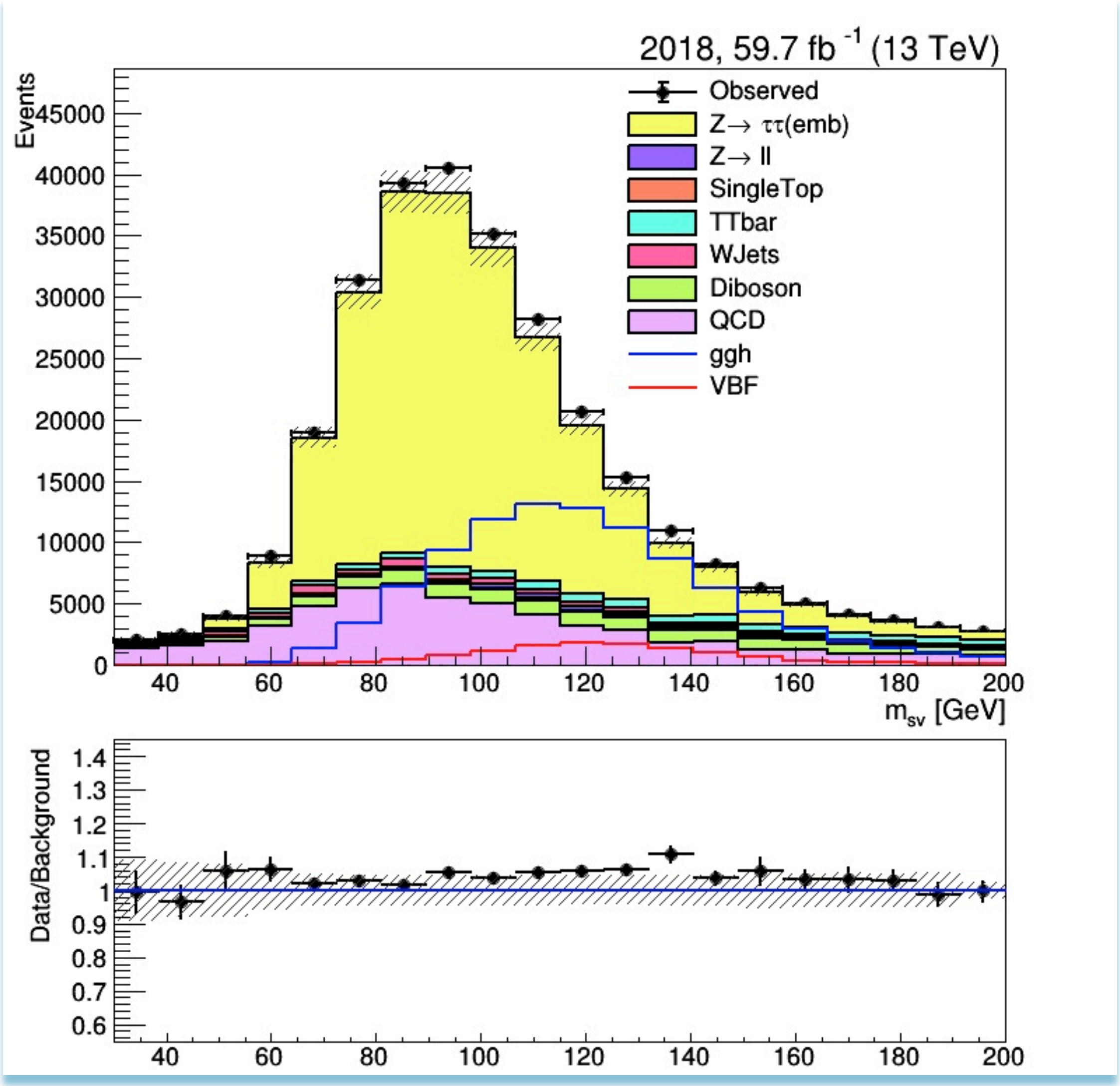
1.Data/MC control Plots

CMS 2018 Data , Integrated luminosity 59.7 fb^{-1}

Work in Progress

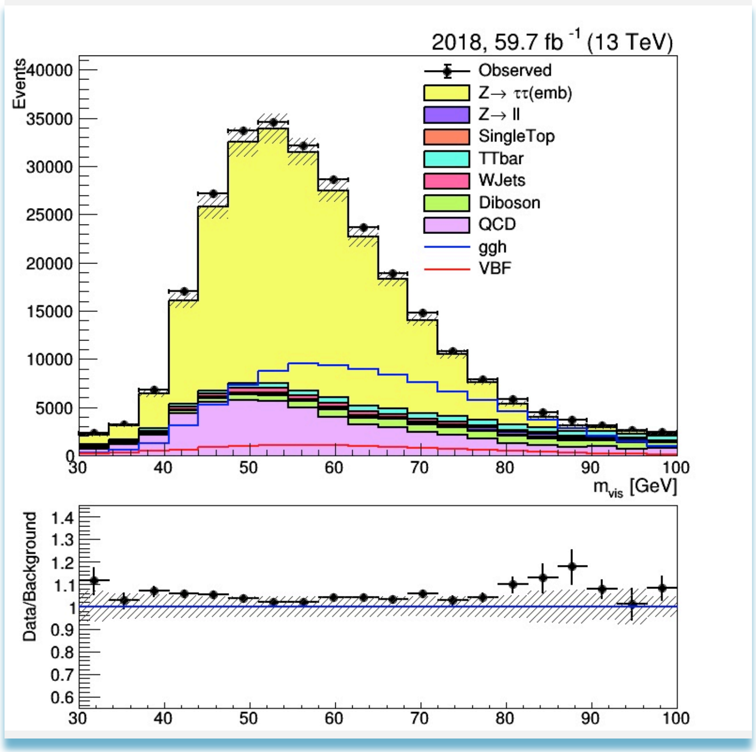
- Kinematics and event selection:
  - $e\mu$ pair selection
  - Isolation variable $I_{rel}^{e(\mu)} <$ 0.15 (0.2)
  - B-tag veto
  - $m_T^{e\mu} = \sqrt{2P_T^e P_T^\mu\,(1-\cos\Delta\phi^{e\mu})} <$ 60 GeV

- Background estimation Methods:
  - Embedded samples for events with two genuine taus
  - QCD events estimated from SS region in Data
  - Other Background events: MC

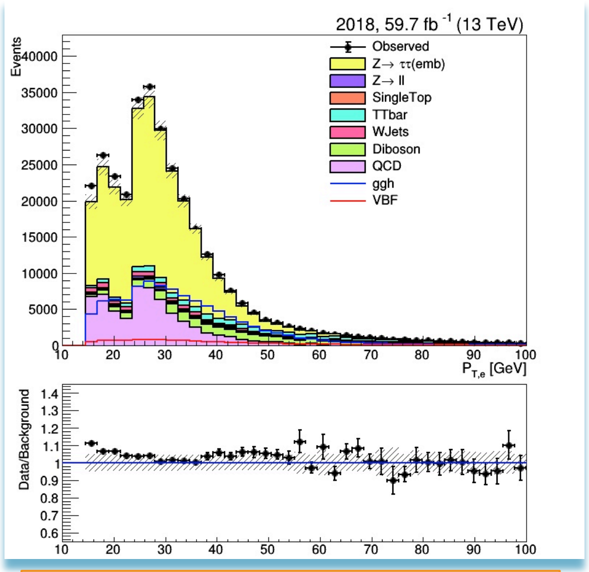


Result : Data/MC Control plots

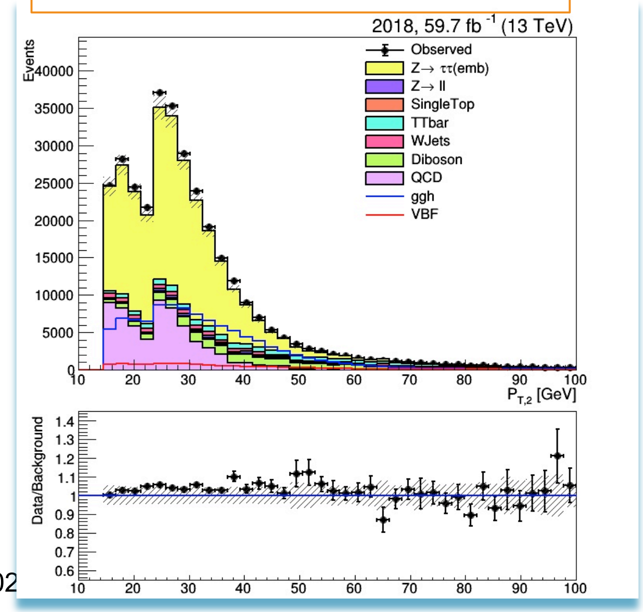
Visible mass of di-taus



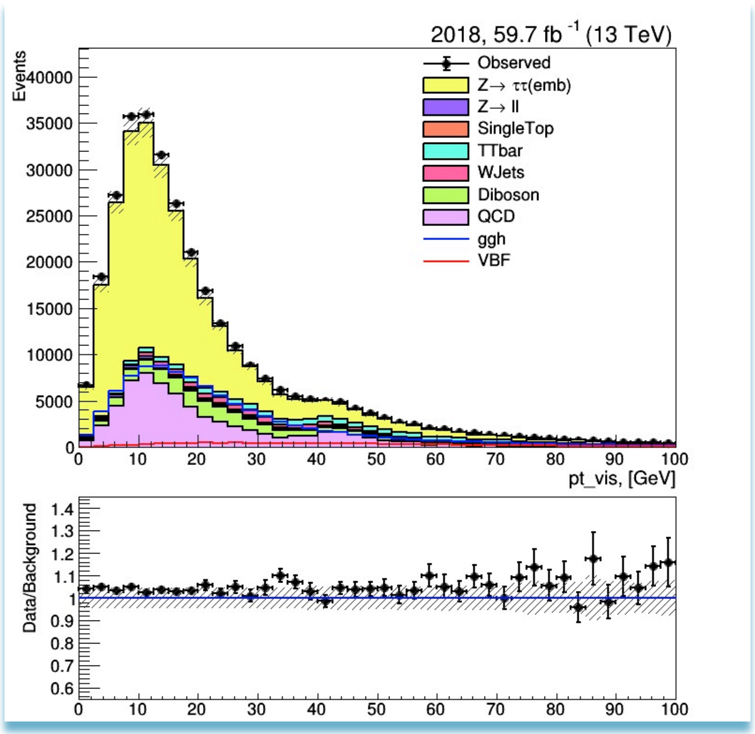
Transverse momenta of electron



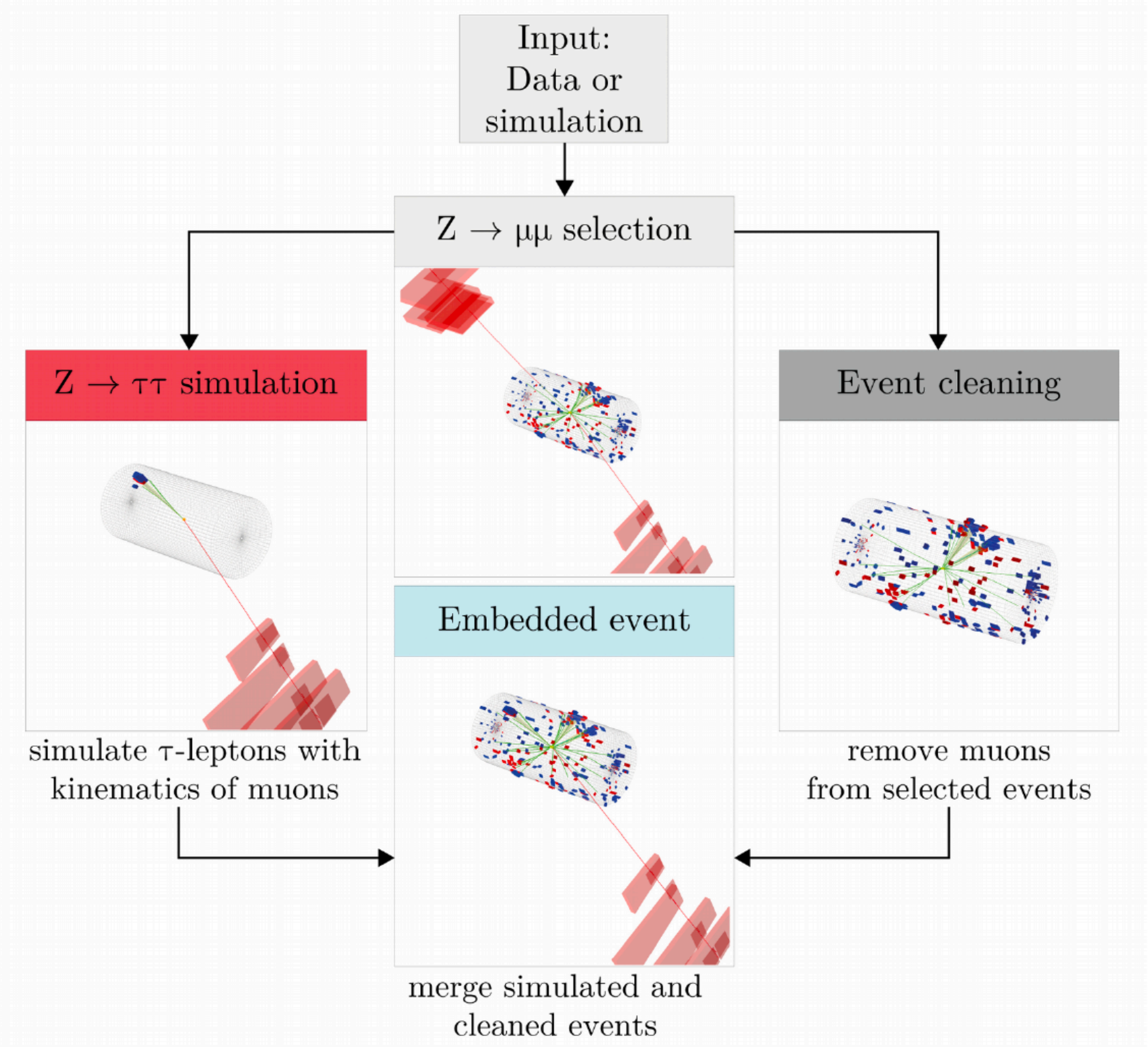
Transverse momenta of muon



Visible transverse momenta of taus



Data-driven bkg estimation: embedding

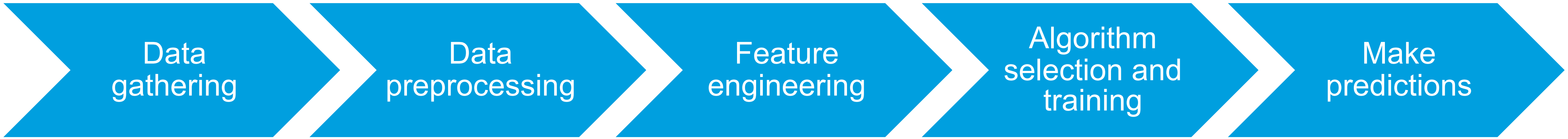


- Used to estimate $Z/\gamma^* \rightarrow \tau\tau$, and other processes with two genuine τ leptons
- It is derived based on the principle of lepton universality
- It combines $Z \rightarrow \mu\mu$ collected events with $Z \rightarrow \tau\tau$, tt-bar and VV simulation
- Increases the effective statistic for these processes and reduces number of systematics

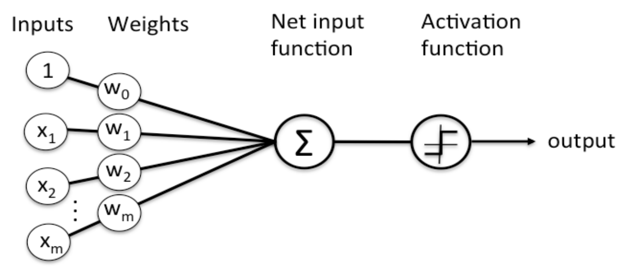
Introduction to Machine Learning

Deep Neural Networks

- Steps to solve a Machine Learning problem



- Neural Networks (Deep Learning) is one of the Algorithms used for the training
- Other Algorithm choices: Random Forests, Decision Tree and etc.
- First model of artificial neural networks propsed on 1943
- Neural network can consist of several layers and nodes
- Given some input (x), the Network calculates the output (y), using a set of weights (w)



Schematic of Rosenblatt's perceptron.