Search for Vector-like T'(→ tH) in hadronic final states using Neural Networks

Jieun Choi

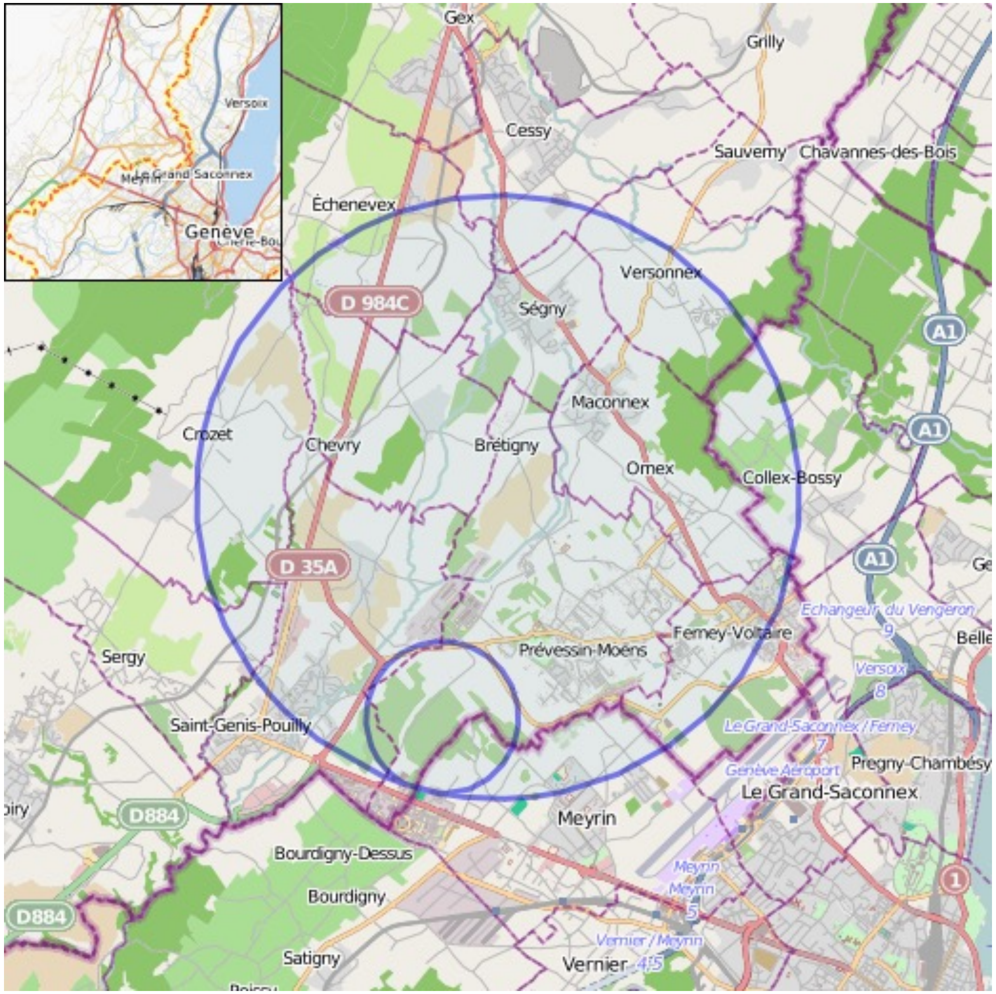
HYU (Seoul) / IP2I (Lyon)

April 25, 2023

ji.eun.choi@cern.ch

Introduction

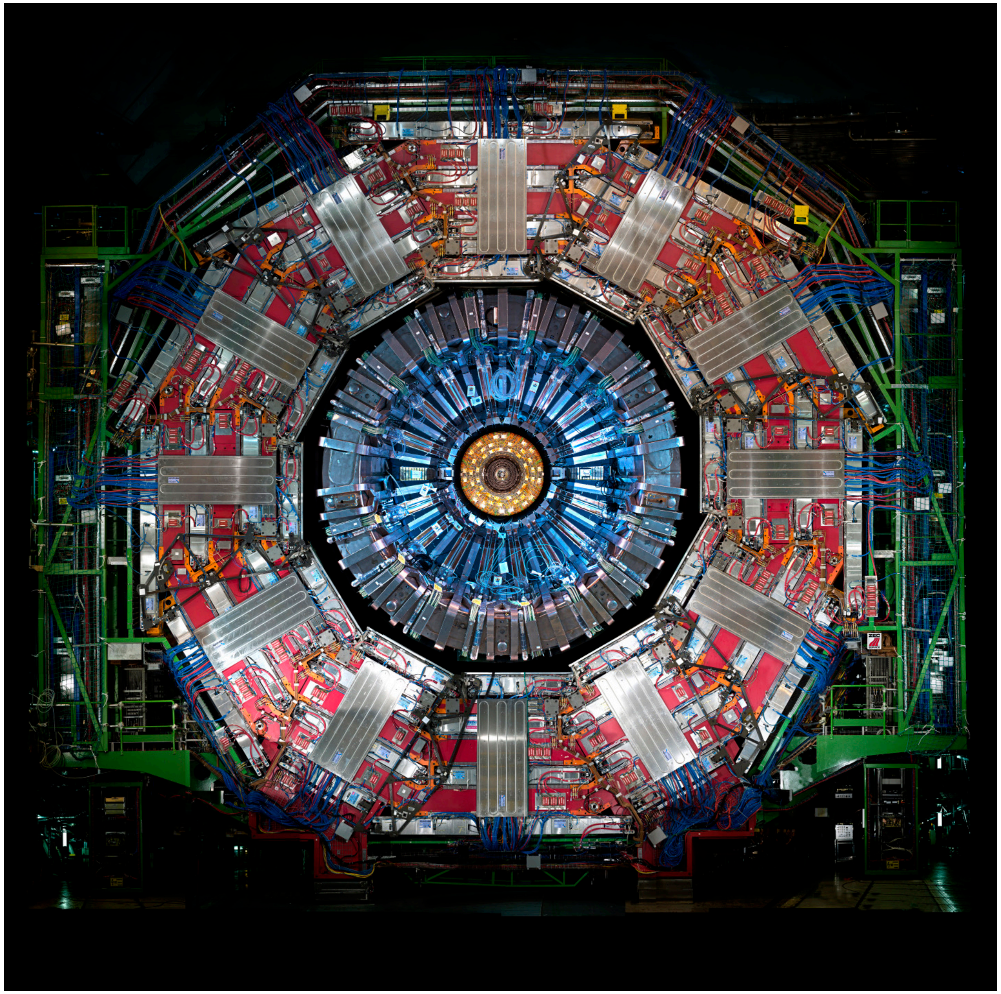
Large Hadron Collider



- The LHC is a particle accelerator that pushes protons to near the speed of light
- It consists of a 27 km ring of superconducting magnets with accelerating structures that boost the energy of the particles along the way
- It produces lots of particle physics phenomena from proton-proton collisions at the center of mass energy = 13 TeV

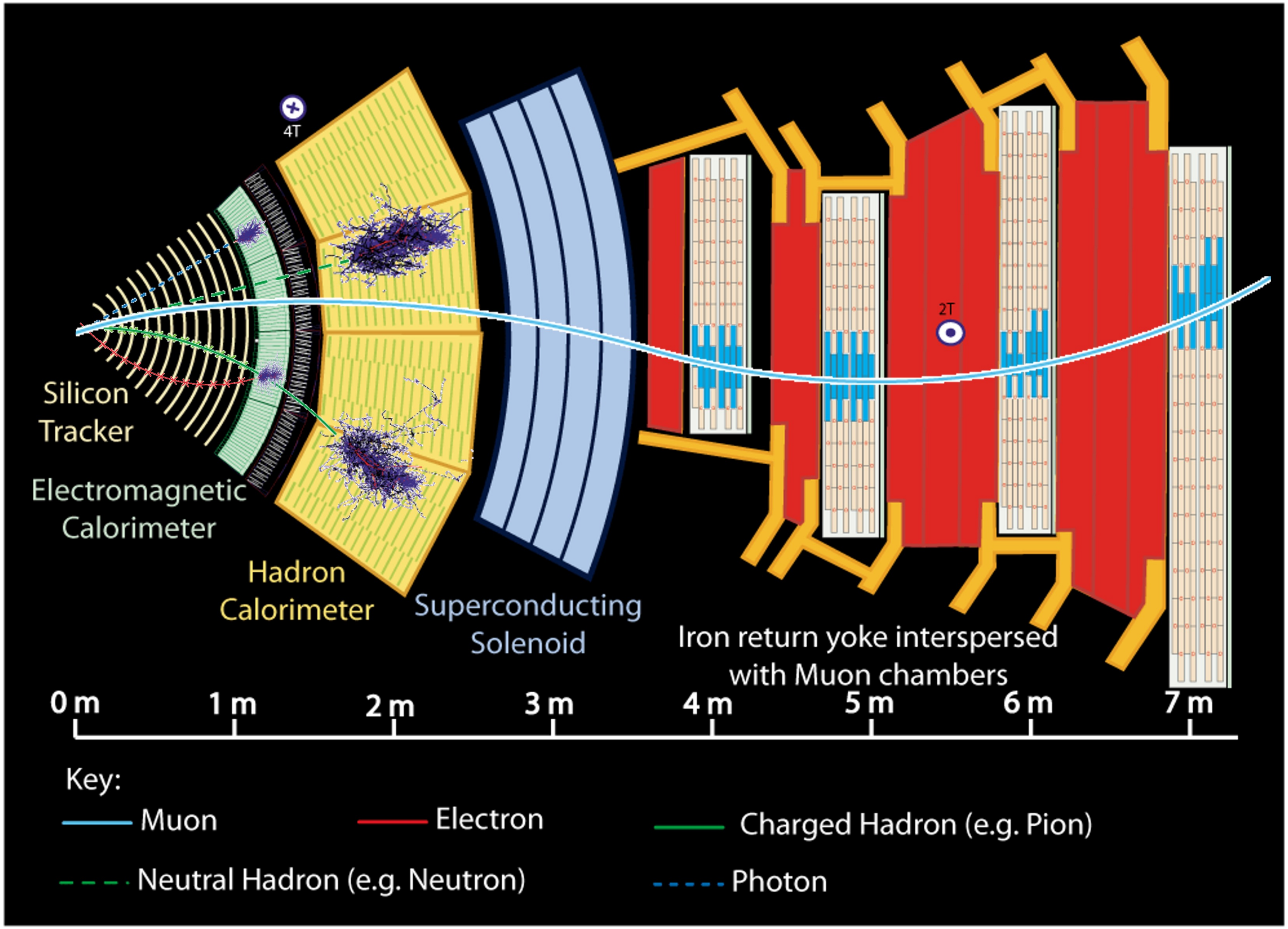
Introduction

Compact Muon Solenoid



- The CMS detector is located at one of the four collision points in LHC
- With 15 meters high and 21 meters long, CMS is “compact” for all detectors it contains
- It has the most powerful solenoid magnet ever made
- The discovery of Higgs boson at CMS and ATLAS detector in 2012 completed standard model
- However, some phenomena still exist that are not described by standard models

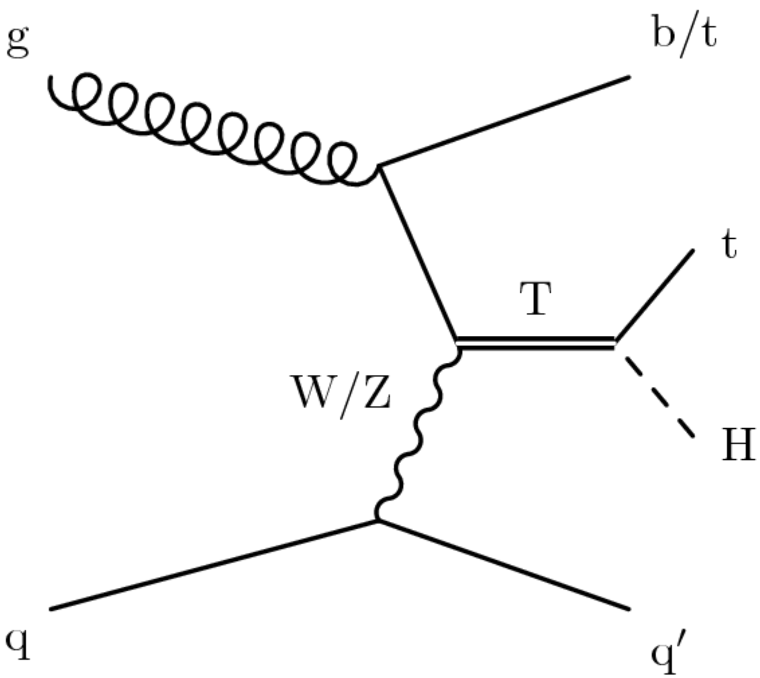
Introduction



Physics objects are reconstructed by information from various detectors

Motivation

Search for Vector Like Quark in hadronic final states



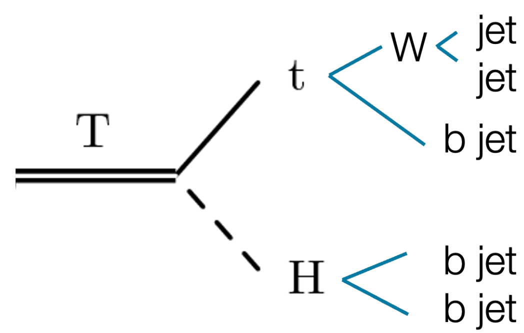
Why the Vector-Like Quarks?

- Evaluate many underlying models:
  - Stabilize the Higgs boson mass
  - Offers a potential solution to the hierarchy problem

  …

T’ decay in full hadronic final state

- T’ decaying into top and Higgs



- Main background:
  - ttbar in hadronic decay (tt → bbqqqq)
  - multi-jet event (QCD)

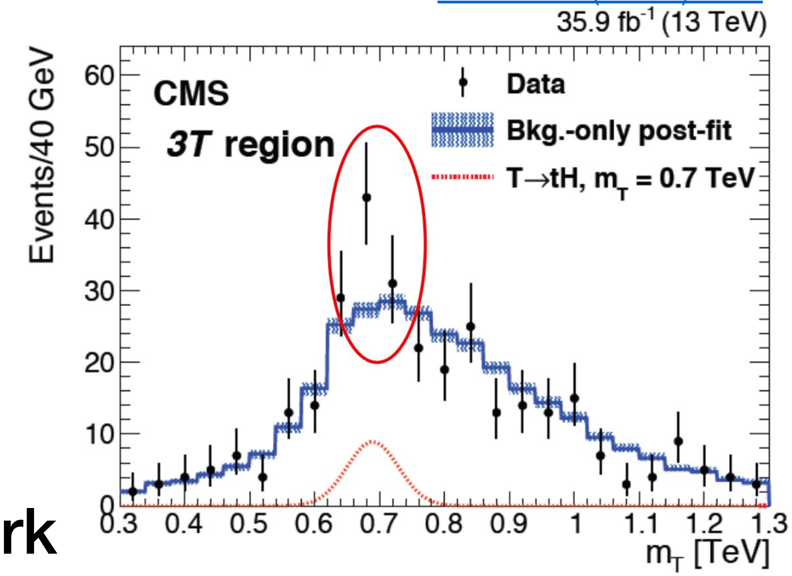
Motivation

Analysis using 2016 data in CMS

- Excess in T’ mass @ 680 GeV was observed
  - Using cut-Based method

→ Improve the significance with Neural Network !

JHEP 01 (2020) 036



cut-based method → Neural Network

- Cut-based method: Categorizing events with a certain “selection” criterion on a data

Vehicle dataset
Signal = Red bike

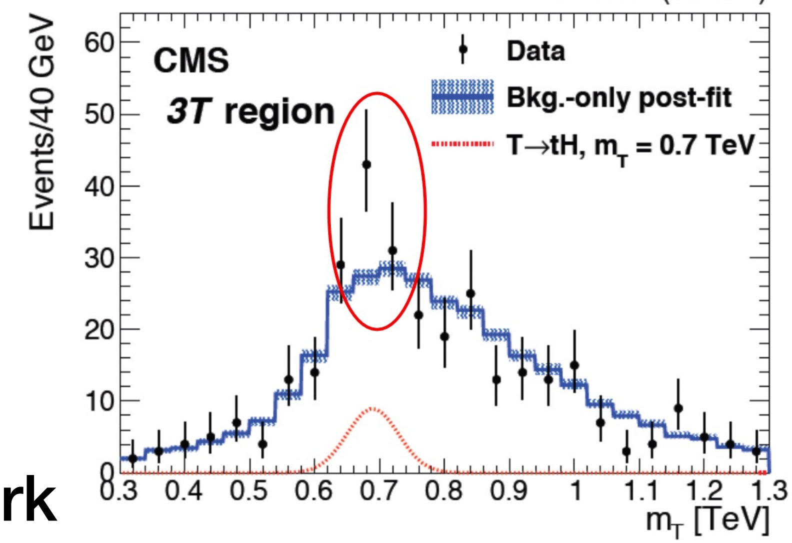
Motivation

Analysis using 2016 data in CMS

- Excess in T' mass @ 680 GeV was observed
  - Using cut-Based method

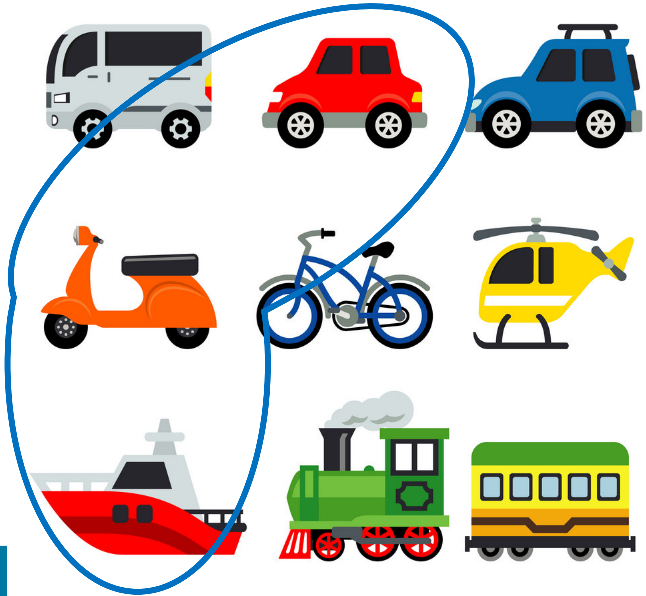
→ Improve the significance with Neural Network !

JHEP 01 (2020) 036



cut-based method → Neural Network

- Cut-based method: Categorizing events with a certain "selection" criterion on a data



Vehicle dataset
Signal = Red bike

Cut-Based method:

Color = Red

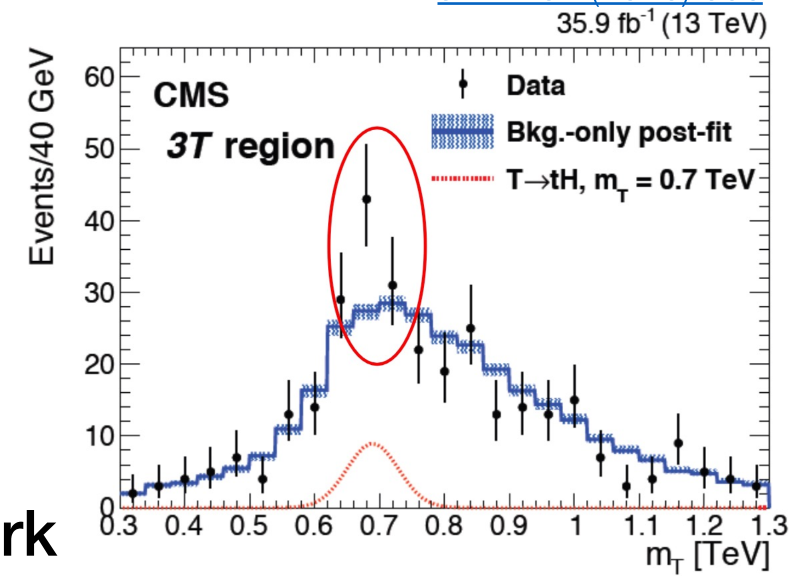
Motivation

Analysis using 2016 data in CMS

- Excess in T' mass @ 680 GeV was observed
  - Using cut-Based method

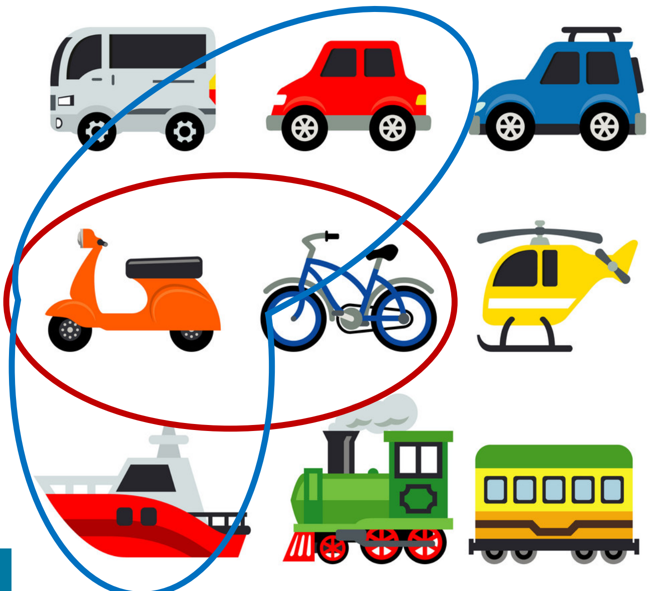
→ Improve the significance with Neural Network !

JHEP 01 (2020) 036



cut-based method → Neural Network

- Cut-based method: Categorizing events with a certain "selection" criterion on a data



Vehicle dataset
Signal = Red bike

Cut-Based method:

Color = Red

Number of wheels = 2

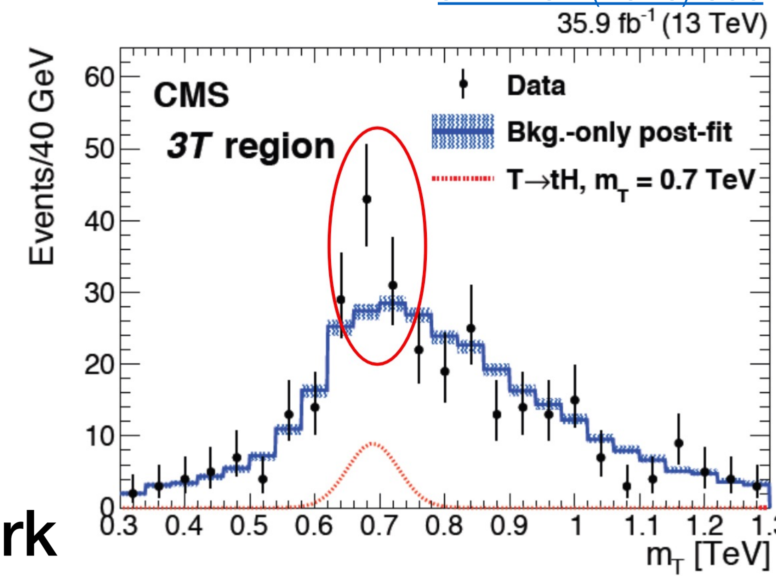
Motivation

Analysis using 2016 data in CMS

- Excess in T' mass @ 680 GeV was observed
    - Using cut-Based method

→ Improve the significance with Neural Network !

[JHEP 01 (2020) 036]



cut-based method → Neural Network

- Cut-based method: Categorizing events with a certain "selection" criterion on a data

Vehicle dataset
Signal = Red bike

NN method:

Input: Color, Number of wheels
Signal / background label

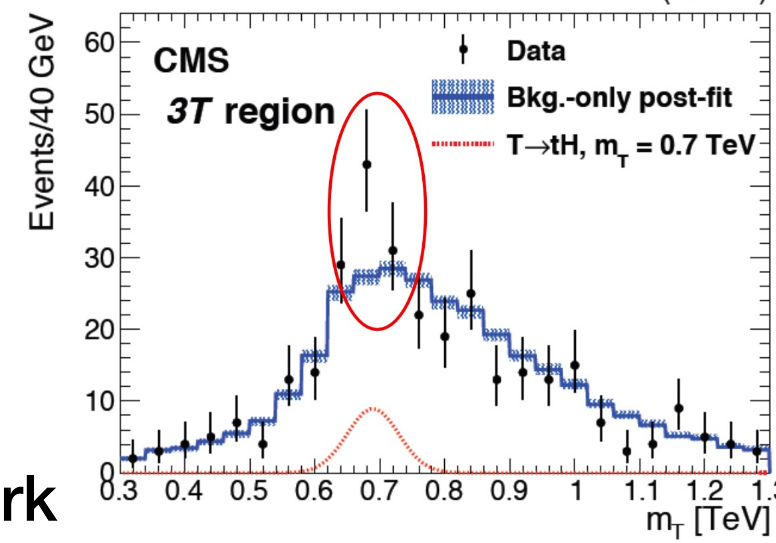
Motivation

JHEP 01 (2020) 036

Analysis using 2016 data in CMS

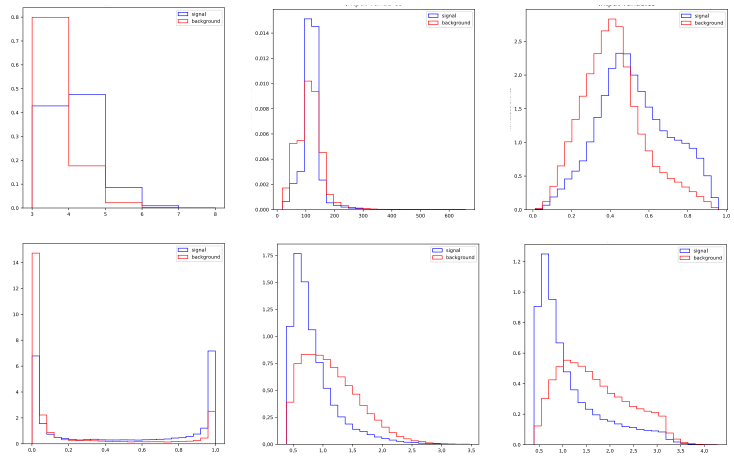
- Excess in T' mass @ 680 GeV was observed
    - Using cut-Based method

→ Improve the significance with Neural Network !



cut-based method → Neural Network

- Cut-based method: Categorizing events with a certain “selection” criterion on a data
- Selections optimized based on 47 kinematic observables for maximizing significance

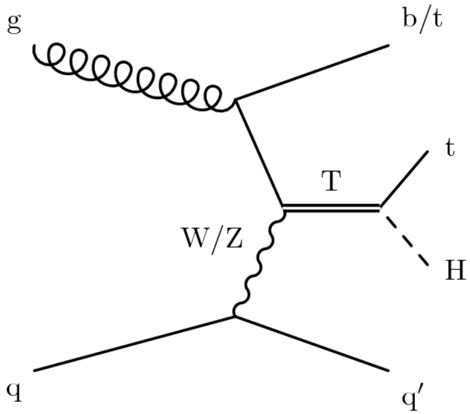


… Feed these information to neural network !

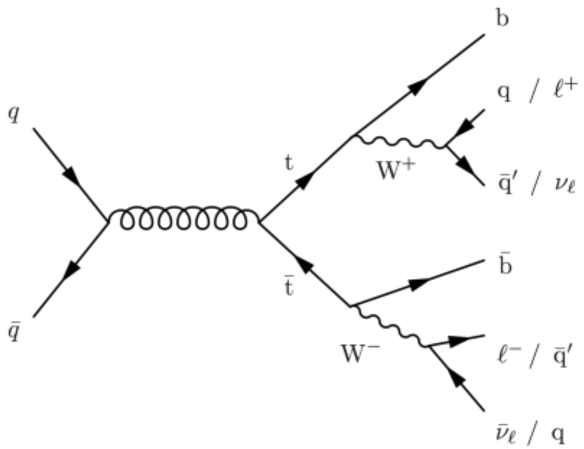
Strategy

Target Process

- Signal and Background Classification



Signal: T'→tH hadronic decay



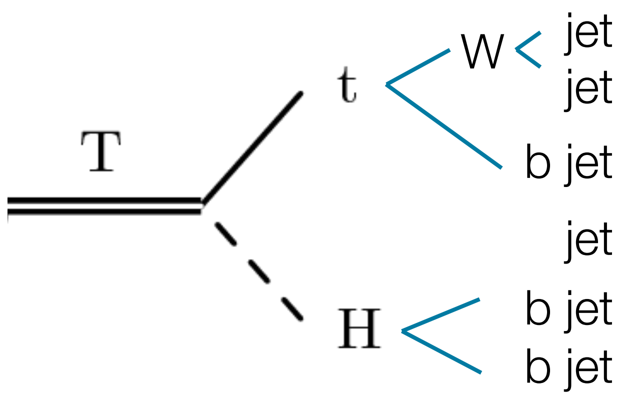
Background: ttbar hadronic decay

Benchmark

- Comparison between the cut-based method and Simple neural networks
- Sample analyzed: simulated CMS detector in ***2018***
  - Signal: Single T' → tH full hadronic M = 700 GeV
- Input features:
  - Low level feature: (b-tagged) jet information (energy, angular distribution…)
  - High level feature: selections used for the cut-Based method (angular differences between jets…)

Selections

- The cut-based method uses 7 selections (Cut 0 - 6) to maximize the significance
  - Most of selections are using variables from χ^2 reconstruction
  - To select jets reconstructing top/W/Higgs mass



$$\chi^2_{Z|Higgs} = \frac{(M_{Z|Higgs} - M_{bb})^2}{\sigma^2_{Z|Higgs}}$$

$$\chi^2_{w} = \frac{(M_W - M_{jj})^2}{\sigma^2_W}$$

$$\chi^2_{top} = \frac{(M_t - M_{bjj})^2}{\sigma^2_t}$$

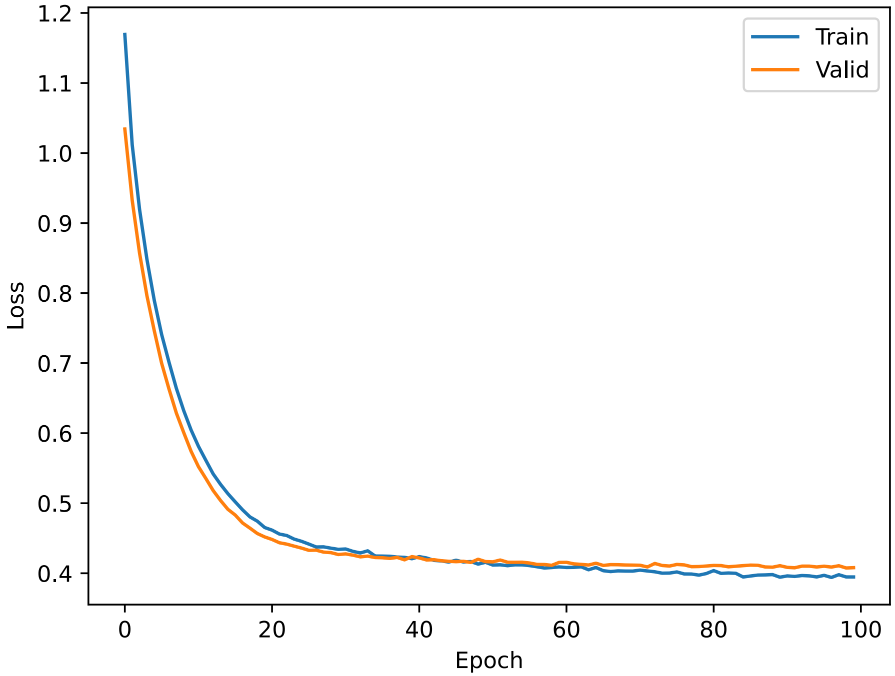
$$\chi^2 = \chi^2_{Z|Higgs} + \chi^2_w + \chi^2_{top}$$

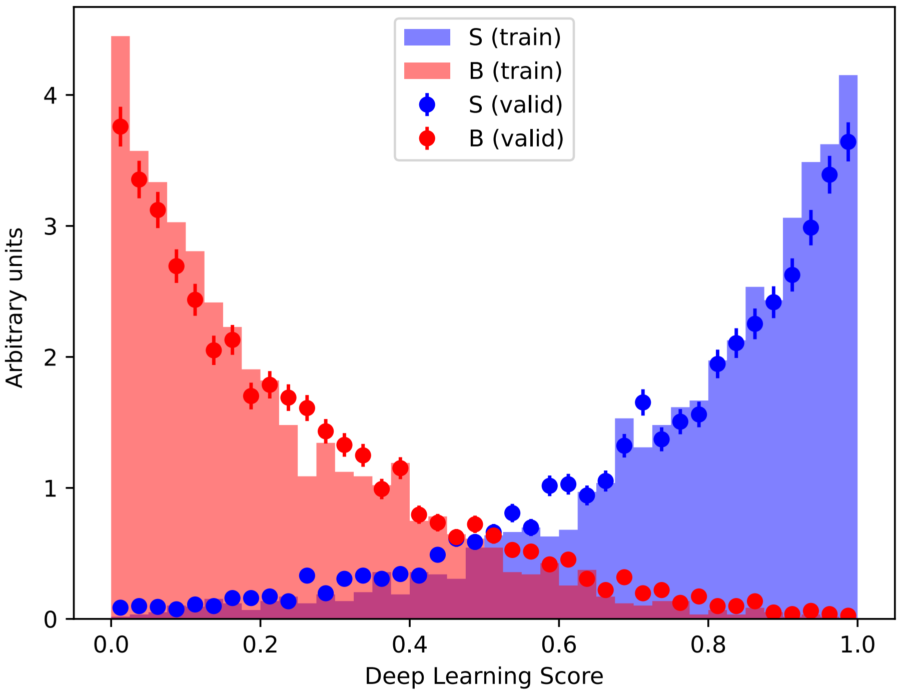
- Take the **baseline criteria (Cut 0)** from cut-based analysis
  - nJets >= 6, nbjets (mistag rate ~ 1%) >= 3
  - Jet_1 > 170, Jet_2 > 130, Jet_3 > 80 GeV
  - H_T > 500 GeV
  - Minimum χ^2 from mass reconstructions < 15
  - Invariant mass of second Top > 250 GeV
  - Invariant mass of Higgs from χ^2 reconstruction > 100 GeV

Used for Training

Used for Evaluation

Output score

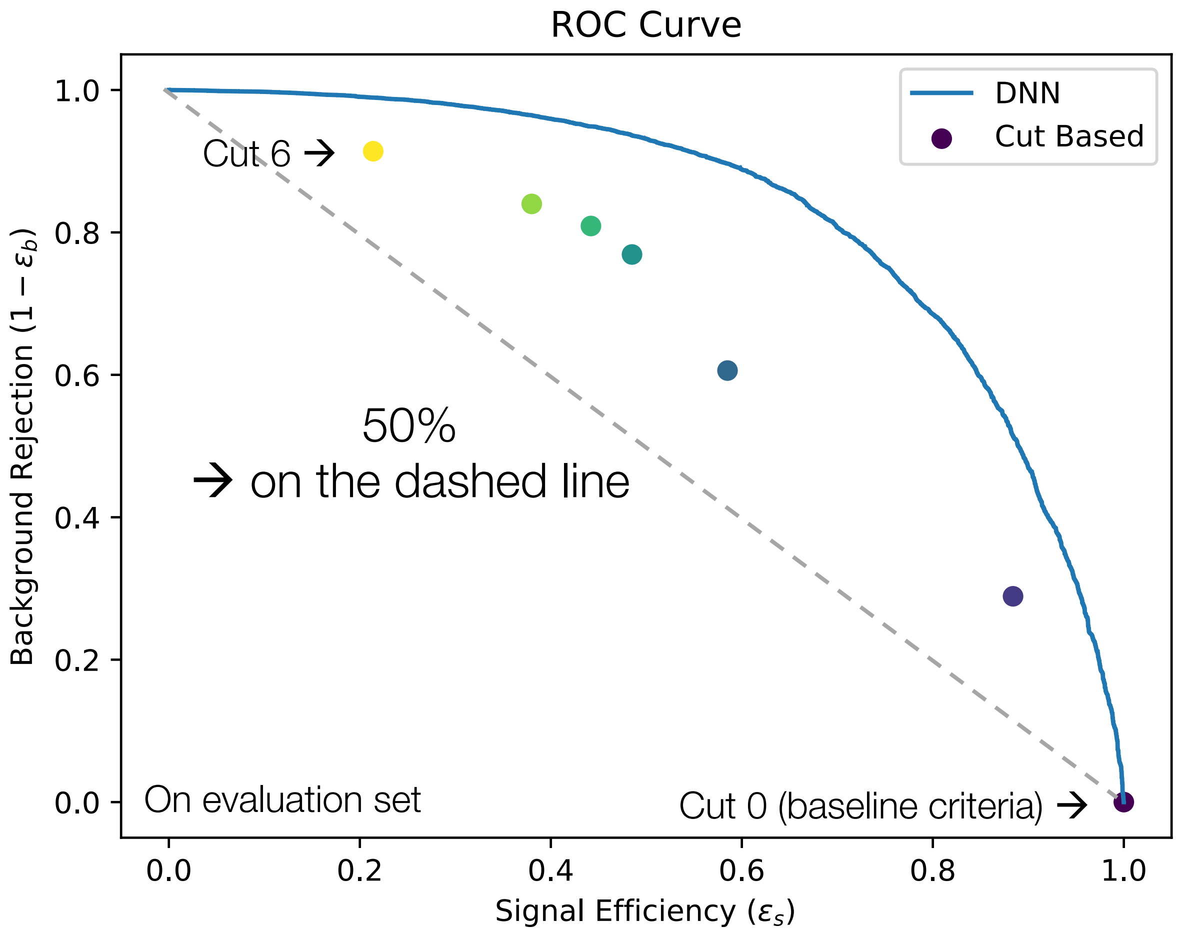




- Deep Neural Network (DNN) structure
  - Simple Dense Layer Network (3 Layer with 100 Nodes)
  - Input set: 80% for training, 20% for validation
- Overtraining check
  - Make sure if model is working not only on training set, but also on the other data
  - Check Loss curve + output distribution from Training and Validation set

Performance

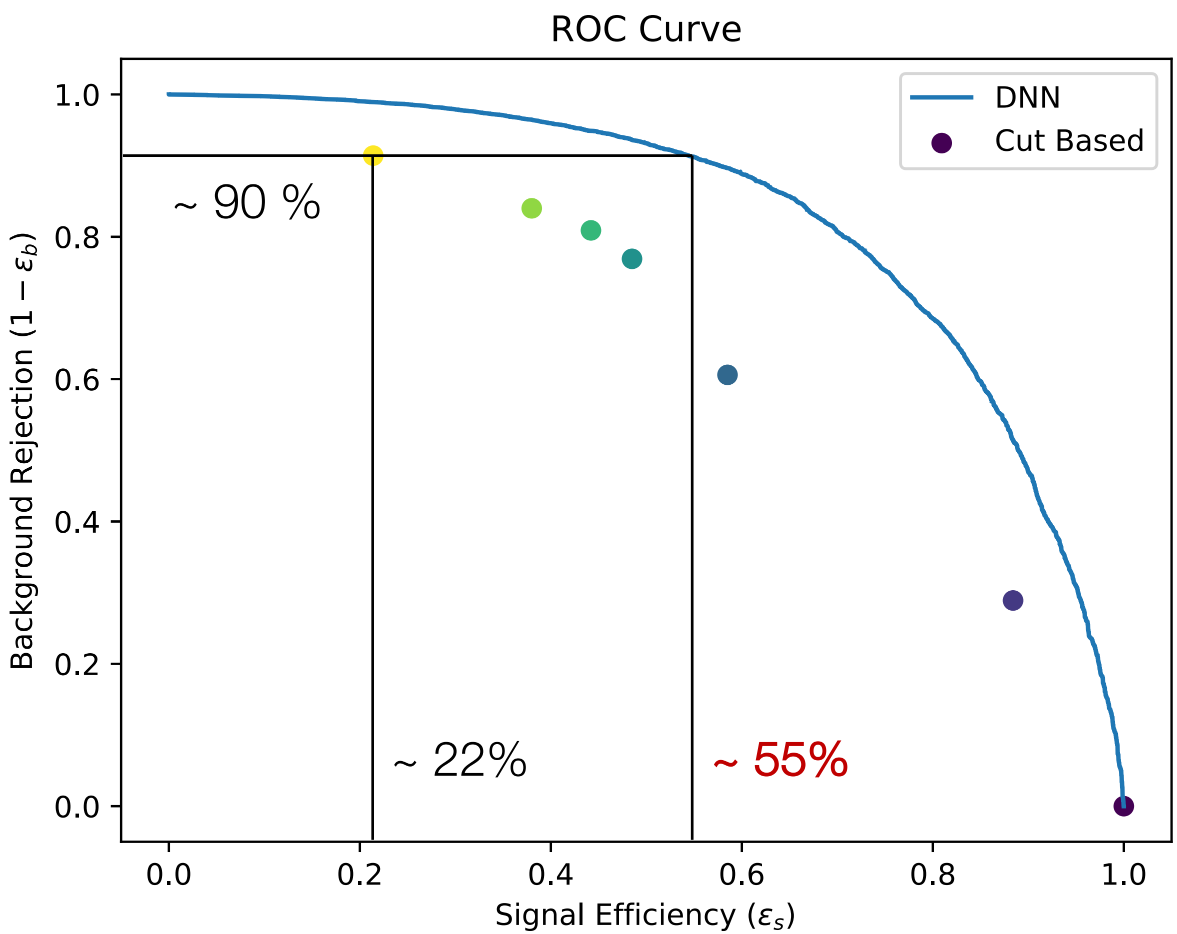
Receiver Operating Characteristic curve



NN works better than cut-based method even without any optimization!

Performance

Receiver Operating Characteristic curve

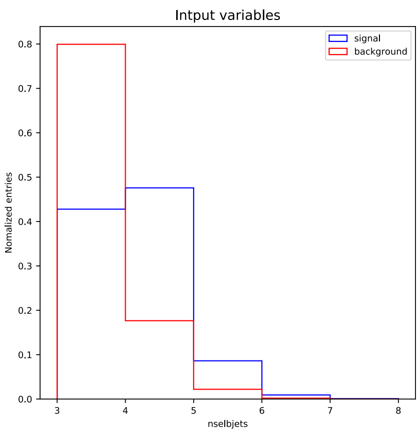


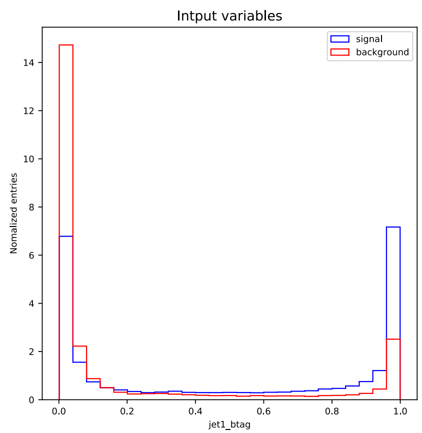
NN works better than cut-based method even without any optimization!

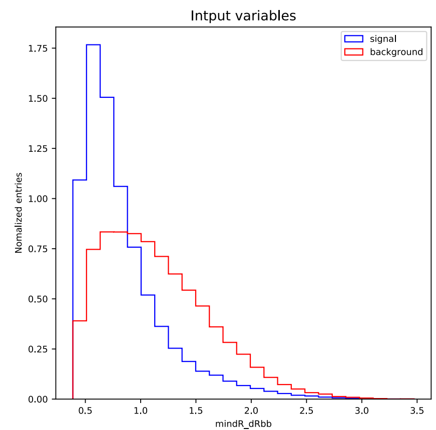
Input feature study

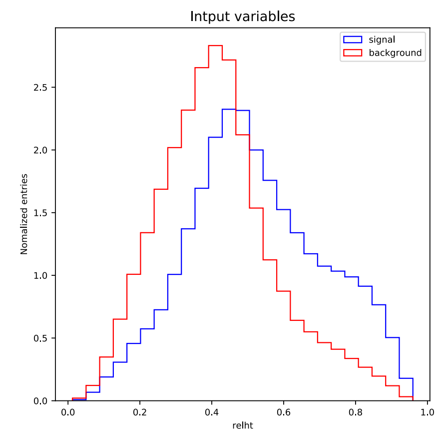
Study on input features

- Initial question:
  - How should we determine which physics observables are “more important”?
  - Which input feature has the largest impact on the NN output node?









Which are “good” and “bad” observables?

Methodology

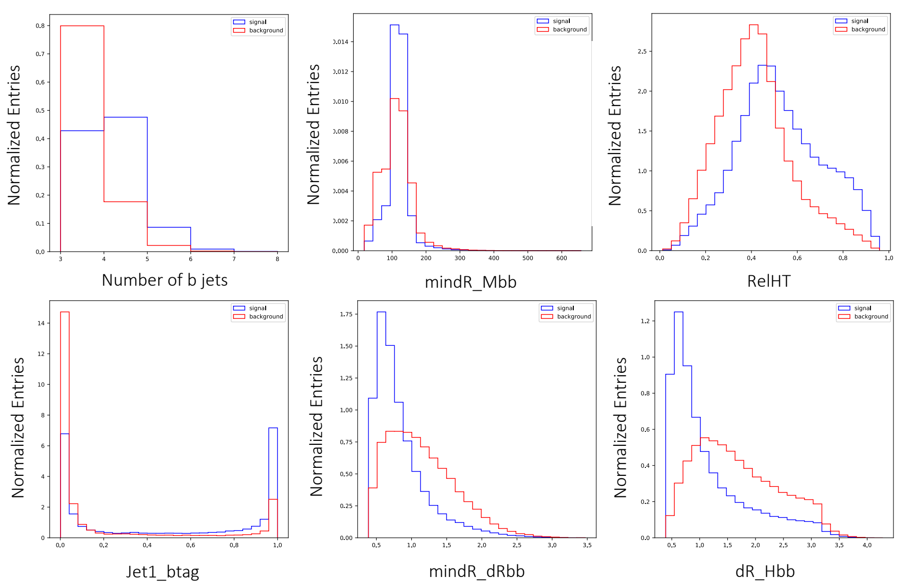
- Pearson Correlation Coefficient
  - Measure linear correlations
- Talyor expansion of the output function at the minima (model)*
  - Calculate gradients of output(node) w.r.t. inputs(event)
  - Extract average gradient for each input features

*arXiv:1803.08782

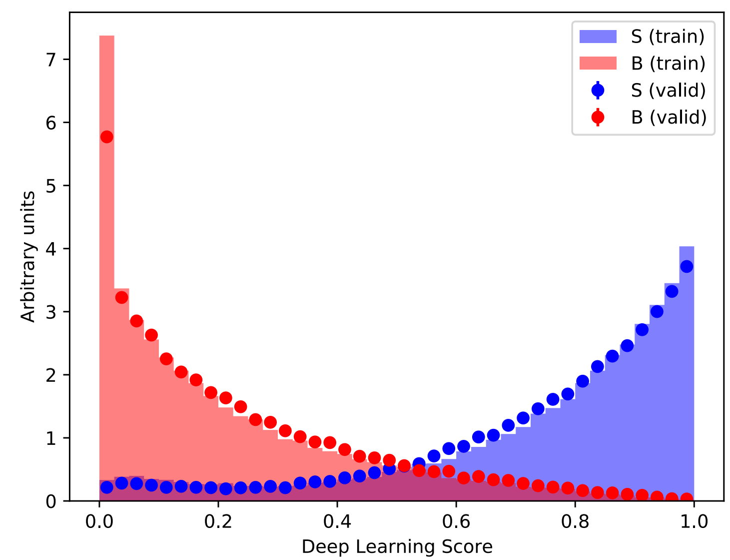
Input feature study

Input Features

- Add all 47 kinematic observables studied from the cut-based method
  - ((6)+15) Information of (2) 5 leading (b) jets (η, energy, b disc. value)
  - (2) Number of (b-tagged) jets
  - (2) ΔR, invariant mass of two b jets having minimum ΔR
  - (5) Min, Max χ^2 value from H, Top, W reconstruction
  - (6) Invariant mass of 5 leading jets, χ^2 candidates (H, (second) Top, W, W+H)
  - (2) H_T, Relative H_T between Top and Higgs
  - (4) ΔR between (b) jets from χ^2 candidates (T', H, W)
  - (5) Ratio of invariant mass, p_T between χ^2 candidates (T' (second) Top, H, W)

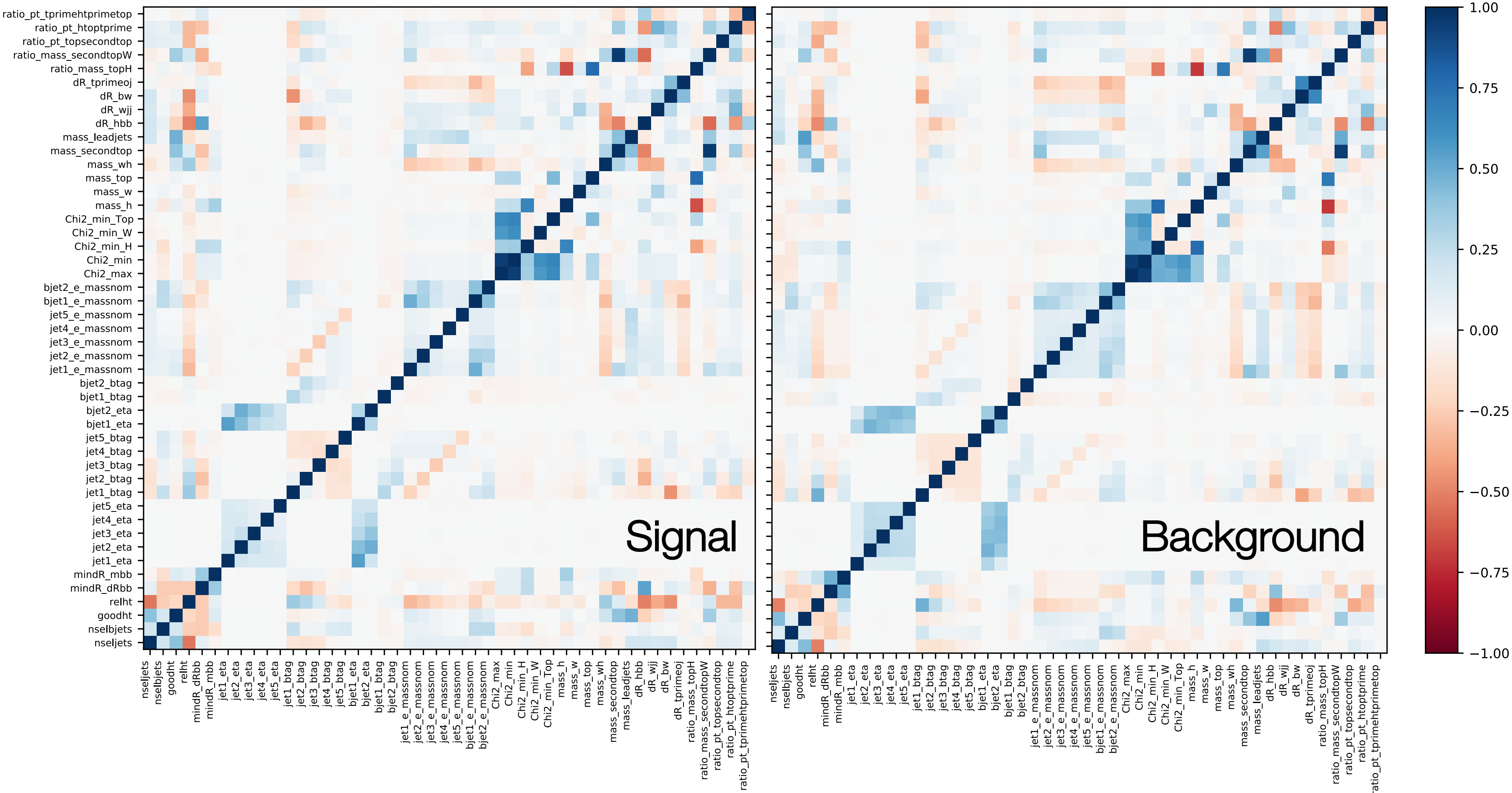


→



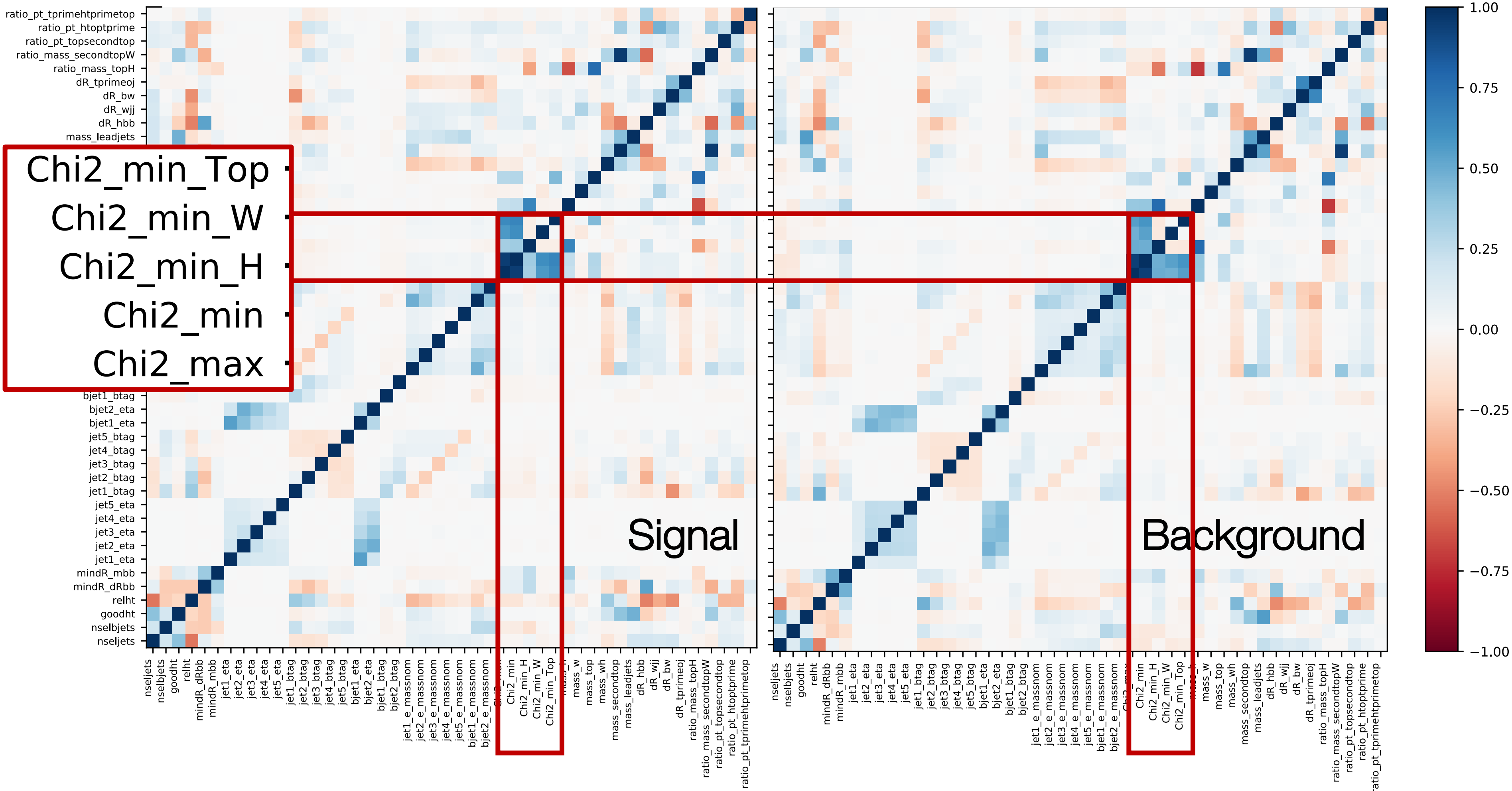
Input feature study

Correlation Matrices



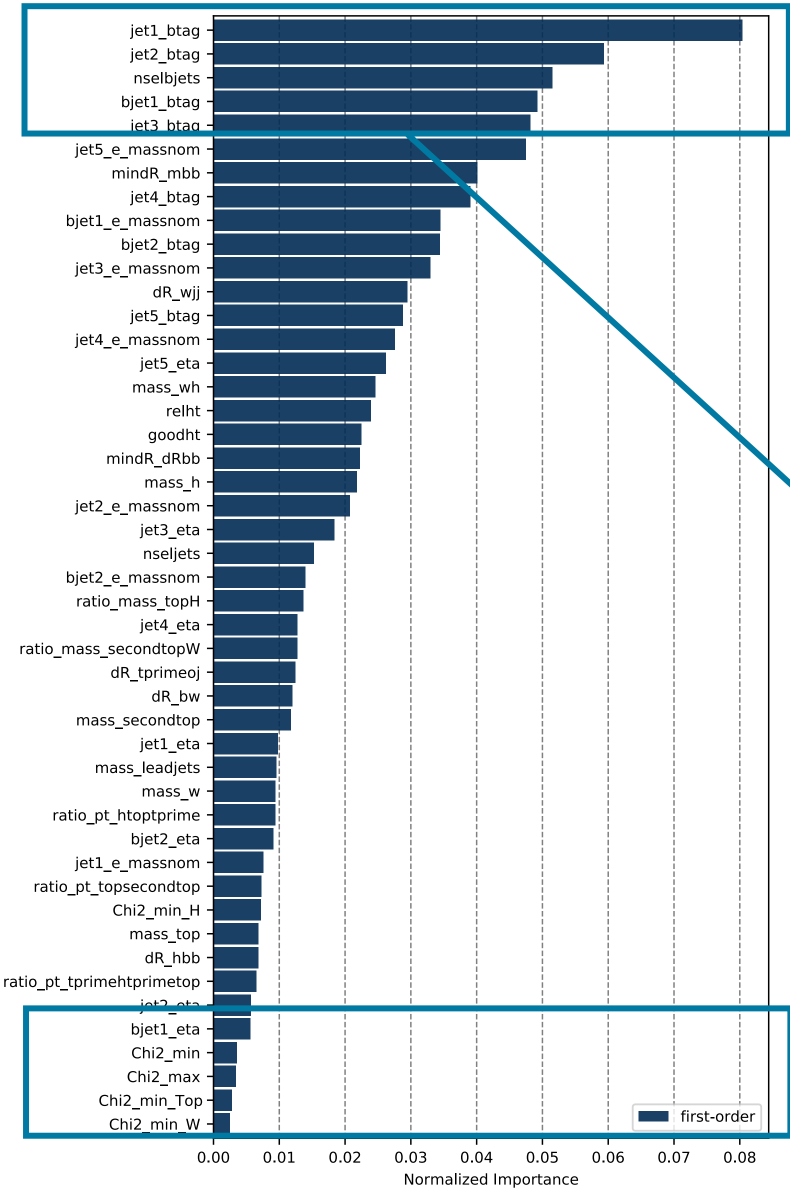
Input feature study

Correlation Matrices



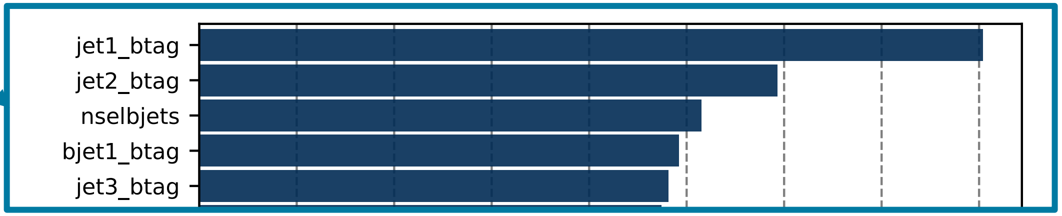
High correlations on χ^2 reconstruction observed

Input feature study



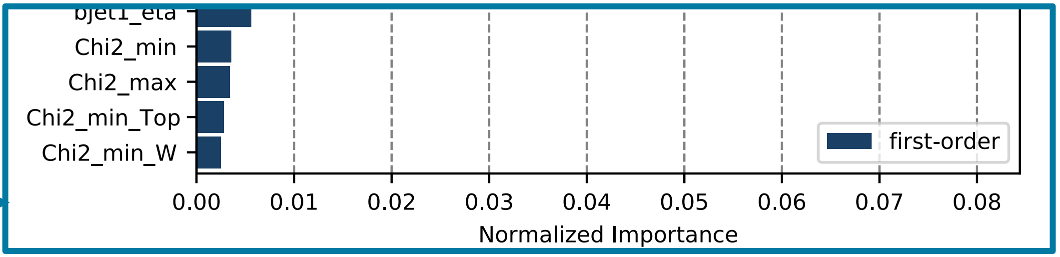
Feature importance

- Talyor expansion of gradient
  - Gradient: slope of the loss function w.r.t. model parameters
  - To measure how much a feature effect on a model



*_btag: likelihood that the jet is from b quark

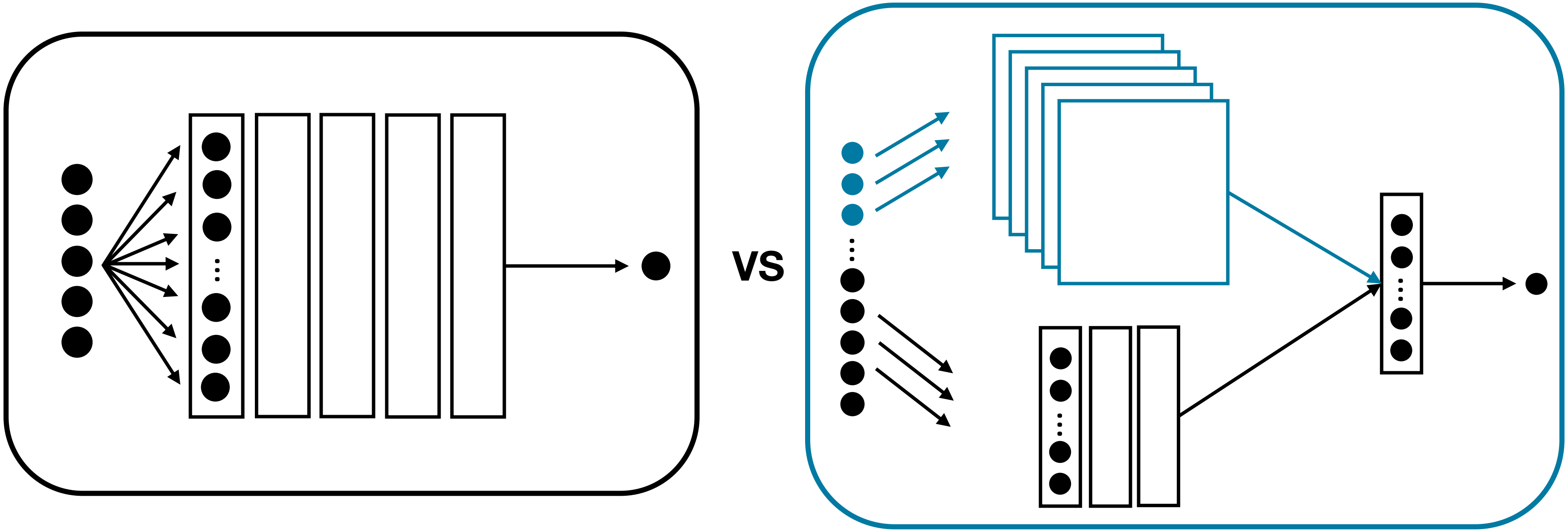
- χ^2 values from reconstruction have lower importance compared to jet information



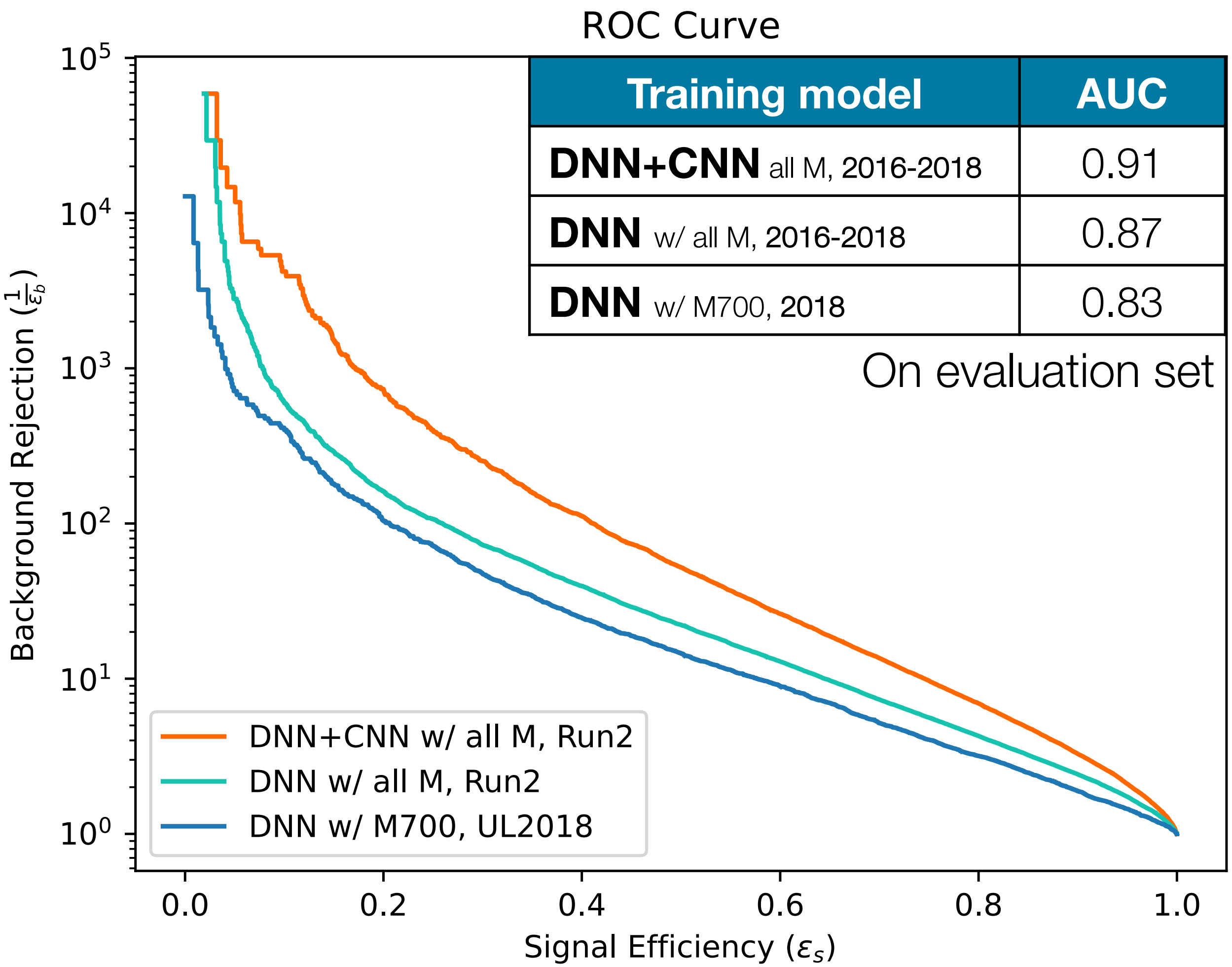
Strategy

Improving Neural Networks

- More statistics, deeper structures!
    - Trained on: With all mass variation (M = 600-1200 GeV) + full MC statistics
- Deeper DNN
    - All information without ordering
- DNN + Convolutional NN (CNN)
    - DNN: Event level information
    - CNN: Jet information (4 vector, b disc. value of 5 leading jets)



Performance



Training model	AUC
DNN+CNN all M, 2016-2018	0.91
DNN w/ all M, 2016-2018	0.87
DNN w/ M700, 2018	0.83

On evaluation set

Using DNN+CNN with the full Run2 (2016-2018) datasets and all mass variations has the highest performance in all criteria

Conclusion

- Analysis on search for T’ in hadronic final state is ongoing, while excess on 2016 CMS data was observed.
- Compared two methods: cut-Based vs NN to gain significance.
- The NN method showed better performance than the cut-Based method.
- Adding more input datasets and deeper model could improve the performance of the NN method.
- This model could help increase the significance in the search for the vector like T’ in hadronic final states with Run 3 (2022-) data
- To do list:
  - Look into the Run2 data
  - Optimize hyperparameter: input variables, number of layers, nodes, optimizers, techniques..
  - Try different architecture: Graph Neural Network, parameterized Neural Network…
  - Continue working on Run 3 data

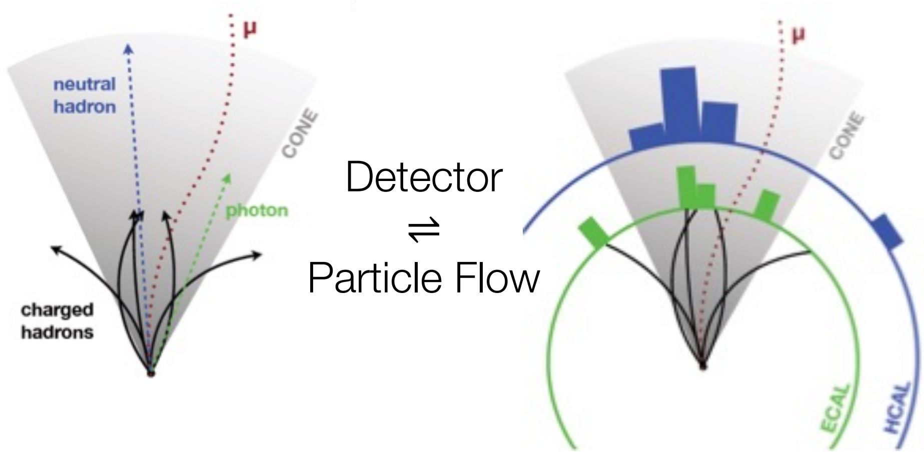
BACKUP

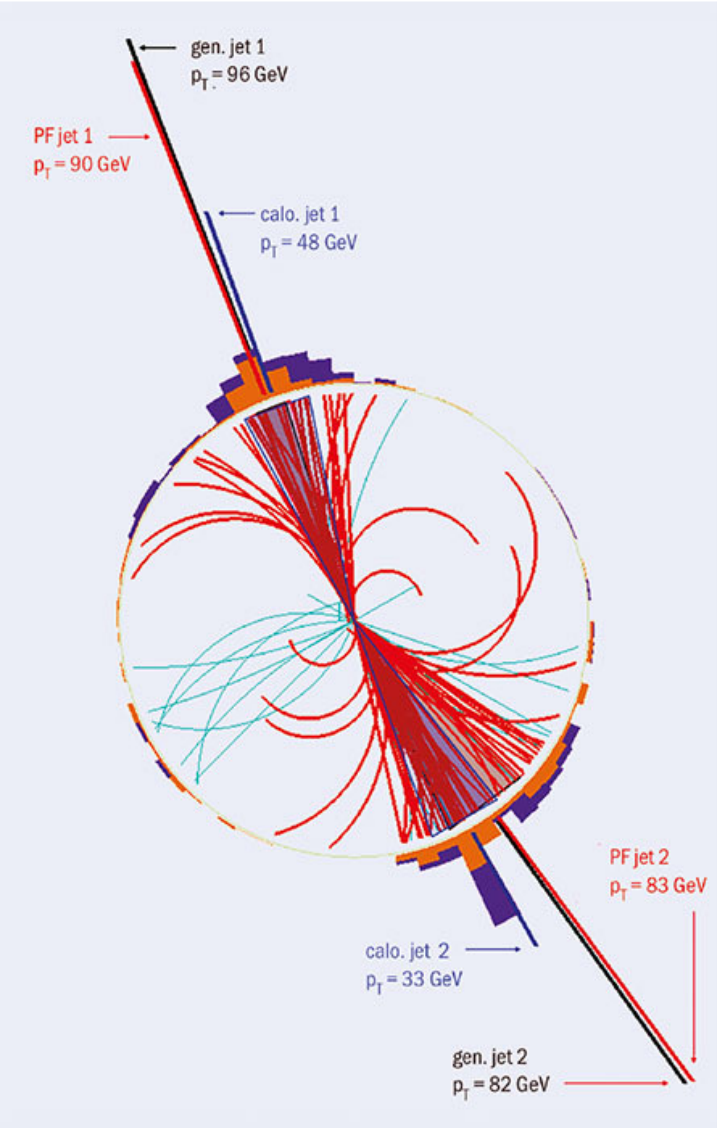
Particle Flow

How we reconstruct jet in CMS

- Calorimeter based approach
- Jet-Plus-Track approach: Calorimeter jet + tracks

- Particle Flow approach
  - Reconstruct each particle individually in the event based on information from all sub-detectors
  - Jet composition:
    ~ 65% charged hadrons
    ~ 25% photons
    ~ 10% neutral hadrons





https://indico.in2p3.fr/event/14415/#2-pflow-in-cms by Florian Beaudette in Particle Flow Day @ LLR

Physics observables in cut-based method

Basic Selection Criteria
Trigger and p_T,η and $n_b^{DeepCSVT} \geq 3$
$j^1_{p_T} > 170$ GeV/c, $j^2_{p_T} > 130$ GeV/c, $j^3_{p_T} > 80$ GeV/c and $H_T > 500$ GeV/c
$\chi^2 < 15$
2nd Top Mass> 250 GeV/c^2
Higgs Mass> 100 GeV/c^2

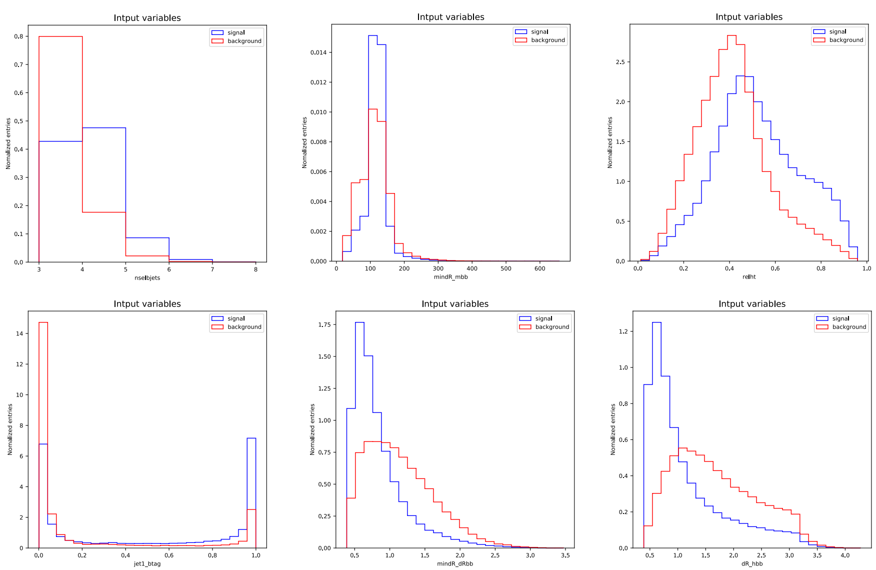
Label	Cuts
Cut 0	Basic selection
Cut 1	Relative $H_T > 0.4$
Cut 2	Max(χ^2) < 3
Cut 3	$\Delta R(b_{Higgs}, b_{Higgs}) < 1.1$
Cut 4	$\chi^2_{Higgs} < 1.5$
Cut 5	$\Delta R(j_W, j_W) < 1.75$
Cut 6	$\Delta R(b_{Top}, W) < 1.2$

Criteria	Quick description
P_T of each jets	Signal should have harder P_T than QCD
p_T (T′)	
Nb Good Jets	QCD could have larger jet multiplicity
χ^2	Signal peaks at 0
χ^2_{Higgs}	Signal peaks at 0, background is larger
χ^2_{Top}	Signal peaks at 0, QCD is larger
χ^2_W	Signal peaks at 0, QCD is larger
Max(χ^2)	Maximum (χ^2_{Higgs},χ^2_{Top},χ^2_W)
M($Higgs_{cand}$)	Invariante mass of Higgs candidate
M(top_{cand})	Invariante mass of Top candidate
M(W_{cand})	Invariante mass of W candidate
M(W_{cand}+$Higgs_{cand}$)	Invariante mass of sum of Higgs and W candidate [4 jet mass]
M(6 Jets)	Invariante mass of the 6 selected jets
2nd Top Mass	Invariante mass of Higgs candidate and 6^{th} jet
$\frac{M_{top}-M_{Higgs}}{M_{top}+M_{Higgs}}$	Ratio of invariante masses
$\frac{M^{2nd}_{top}+M^{2nd}_{W}}{M_{Higgs}}$	Ratio of invariante masses
$\frac{M(W+H)}{M(Top+H+6^{th}Jet)}$	Ratio of invariante masses
H_T	
Relative H_T	$\frac{p_T(H_{cand})+p_T(top_{cand})}{H_T}$
New Relative H_T	$\frac{p_T(H_{cand})+p_T(top_{cand})+p_T(6^{th}Jet)}{H_T}$
$\Delta R(T', 6^{th}Jet)$	LO signal tends to give back to back results
$\Delta R(b_{Higgs}, b_{Higgs})$	Separation between the two jets making the Higgs candidate
$\Delta R(j_W, j_W)$	Separation between the two jets making the W candidate
$\Delta R(Higgs, Top)$	Separation between the Higgs and Top candidates
$\Delta\eta(W, H)$	Eta separation between the Higgs and W candidates
$\Delta R(b_{Top}, W)$	Separation between b-jet making the top candidate and the W candidate
$\Delta R \times \Delta R$	Product of all ΔR in the event. All ΔR are computed to be peaking at 0
Max(ΔR)	Maximum of all ΔR in the event. All ΔR are computed to be peaking at 0
$\Delta\phi(Higgs, Top)$	Phi separation between the Higgs and Top candidates
$\frac{P_T^{2ndtop}-P_T^{top}}{P_T^{2ndtop}}$	Ratio of P_T candidates
$\frac{p_T(H_{cand})-p_T(top_{cand}}{T'}$	Ratio of P_T candidates
$\frac{T'}{p_T(H_{cand})} - \frac{T'}{p_T(top_{cand})}$	Ratio of P_T candidates

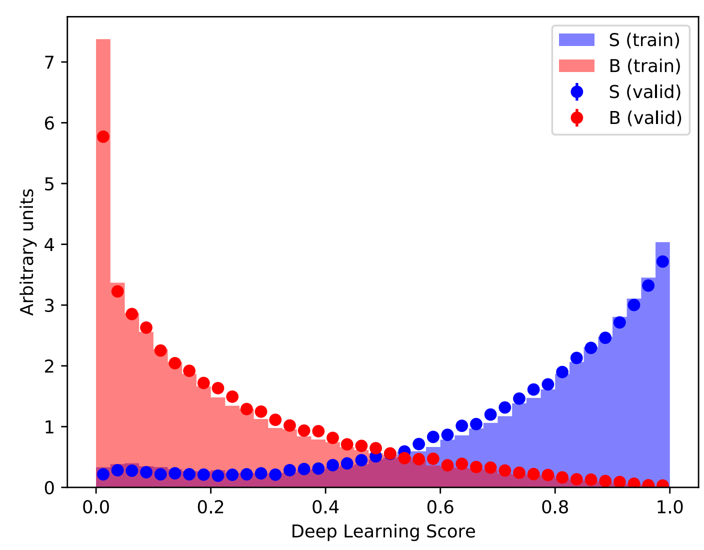
Input feature study

Input Features

- All kinematic observables studied in the cut-based method
- 47 input variables
  - ((6)+15) Information of (2) 5 leading (b) jets (η, energy, b disc. value)
  - (2) Number of (b-tagged) jets
  - (2) ΔR, invariant mass of two b jets having minimum ΔR
  - (5) Min, Max χ^2 value from H, Top, W reconstruction
  - (6) Invariant mass of 5 leading jets, χ^2 candidates (H, (second) Top, W, W+H)
  - (2) H_T, Relative H_T between Top and Higgs
  - (4) ΔR between (b) jets from χ^2 candidates (T', H, W)
  - (5) Ratio of invariant mass, p_T between χ^2 candidates (T' (second) Top, H, W)



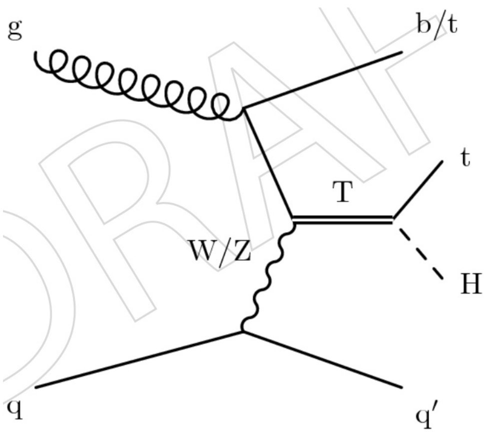
→



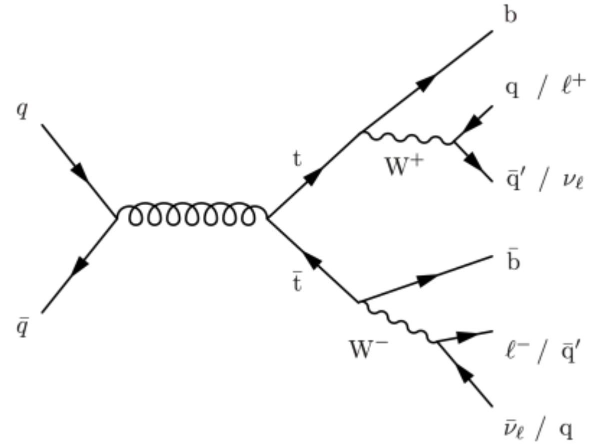
How to avoid bias in NN

Target Process

- Signal and Background Classification



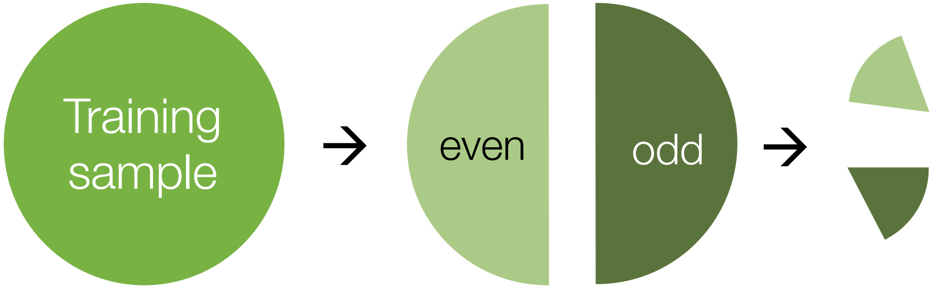
Tprime



ttbar hadronic decay

Strategy

- Compare ROC curves with cutBased (signal efficiency vs background rejection)
  - Evaluate NN at the level of Cut 0 for the pair comparison



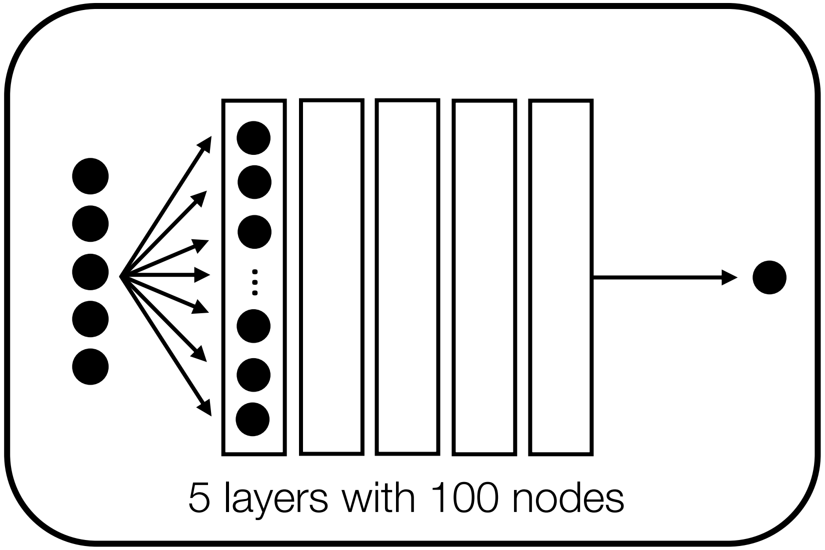
: After Cut 0 → for Evaluation

: After DNN Cut → for Training

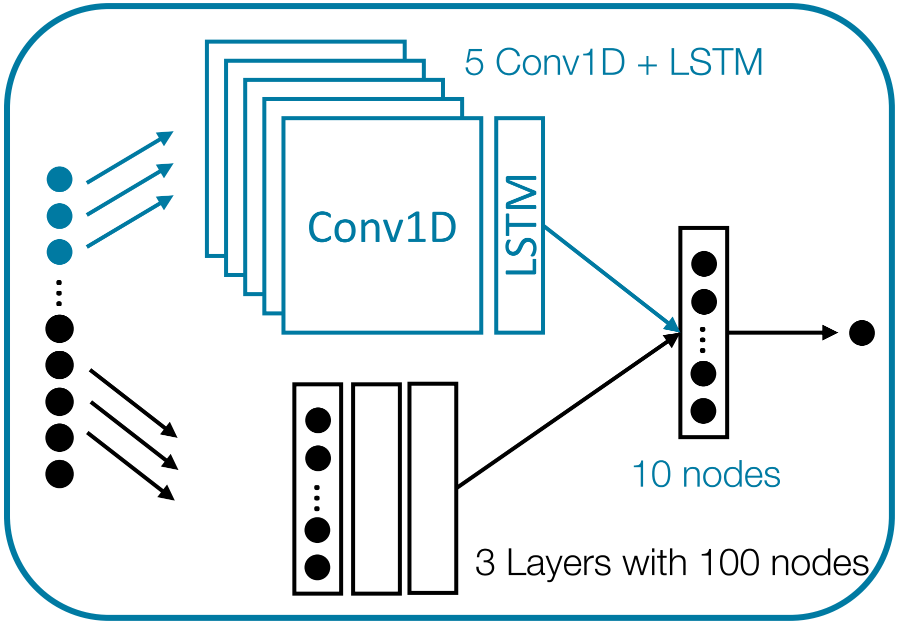
Improving Neural Networks

- Deeper DNN
  - 5 Layers with 100 nodes
- DNN + CNN
  - DNN: 3 Layers with 100 nodes
  - CNN: 5 Conv1D + LSTM (Long Short-Term Memory)
    - Conv1D: 150, 100, 100, 25, 25 filters
    - LSTM: 10 nodes

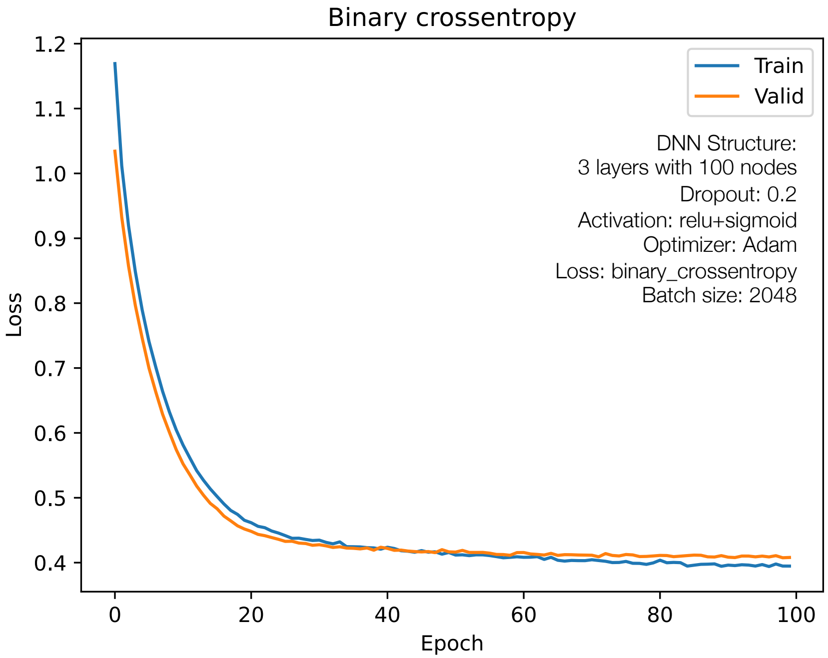
NN Structure
Keras tensorflow backend
Batch Normalization applied
Batch size = 1024
Activation: ReLU
Optimizer: Adam

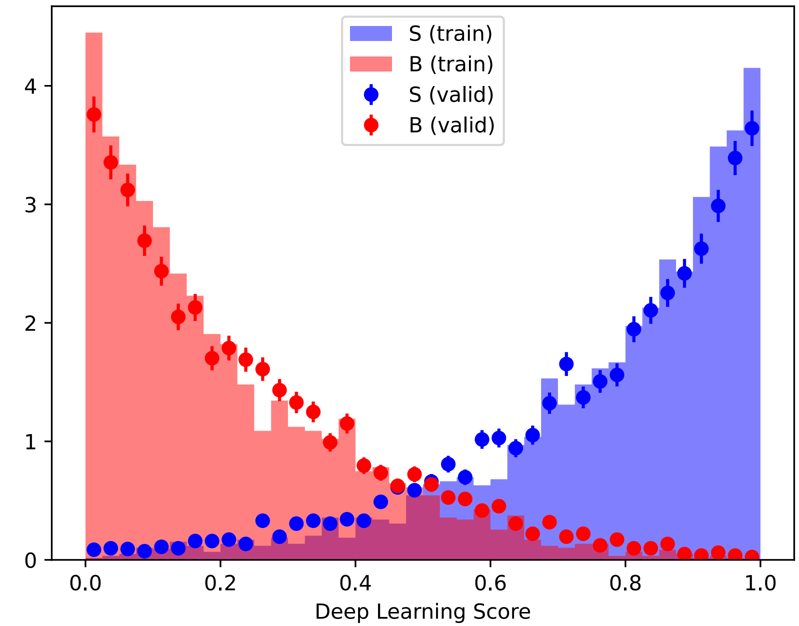


vs



Overtraining check

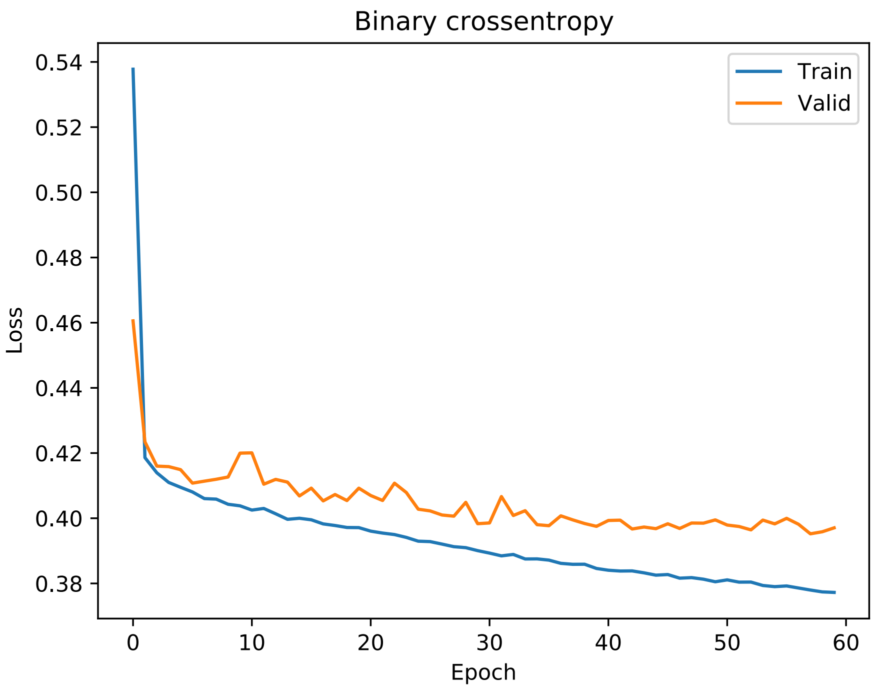


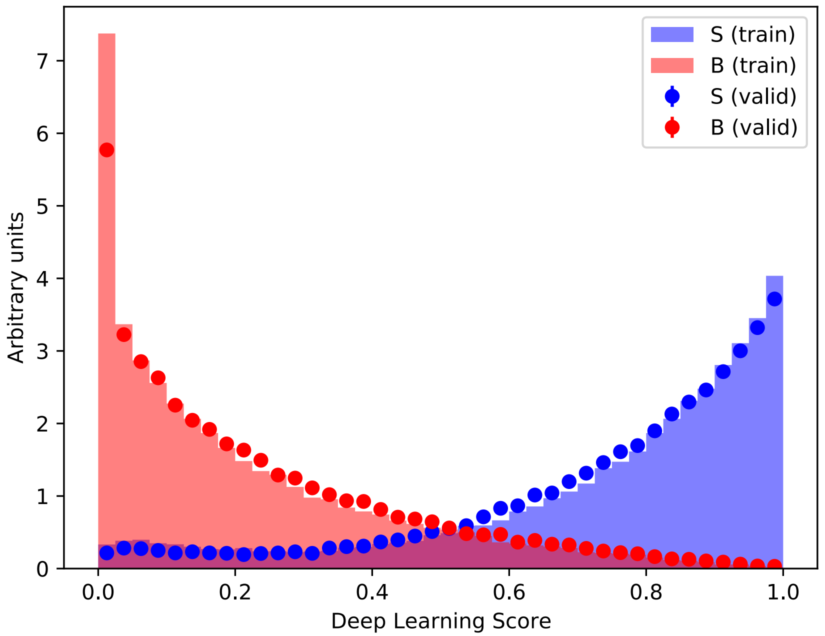


Detail

- Trained in CC server (Training time: 10ms/epoch)
- Input set : Half of TprimeBToTH_M-700 after selection (odd numbered event, 23210 entries)
    - 80 % for training, 20 % for validation
    - Keep even numbered event for evaluation: to avoid bias (using the same event) for performance estimation
    - Epoch: 100 → Validation Loss / Acc are stable, does not diverge yet

Training with more statistics: DNN

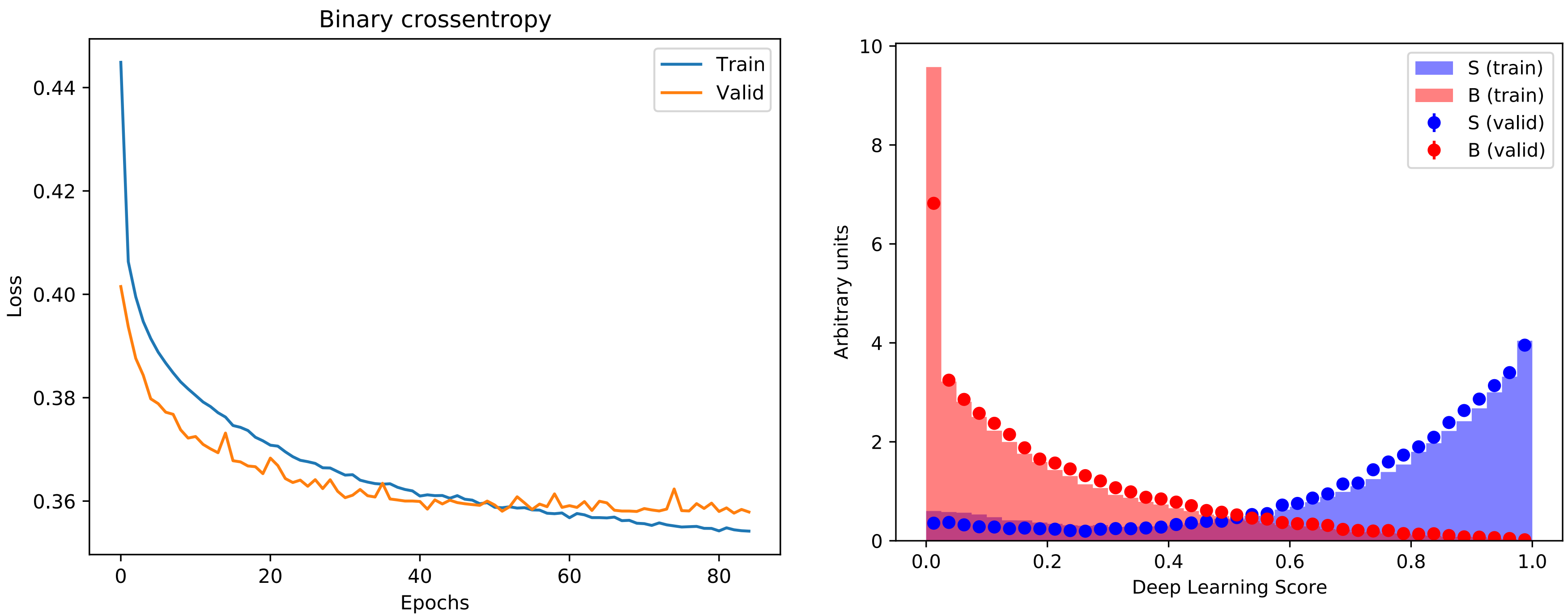




Strategy

- Trained in CC server (Training time: 1s 8ms/epoch)
- Do the same with full Run2 (2016-2018 datasets) and different mass range
- Train on signal samples M=600~1200 GeV (181724 entries (M700 entries * 7)) + the same amount of statistics from TTToHadronic

Training with more statistics: DNN + CNN



Strategy

- Do the same with full Run2 (2016-2018 datasets) and different mass range
- Train on signal samples M=600~1200 GeV (181724 entries (M700 entries * 7)) + the same amount of statistics from TTToHadronic

Feature importance

Motivation

- Initial question: What are the input features with the largest impact on the NN output nodes?
- Extract average gradient for each input features
- Will be able to ”see” how much each variable “effects” on training model
- arXiv:1803.08782

Methodology

- Talyor expansion of the output function at the minima (model)
- Tensorflow.GradientTape()
  - Allow us to record the history of operations applied to target input features
  - Calculate gradients of output(node) w.r.t. inputs(event)
- 1^{st} order: Physical location of feature/marginal distributions – weight w_i for x_i
- 2^{nd} order: Curvature of NN output function - correlations across two features: Gradient of each element of the source w.r.t target – weight w_ij for x_i * x_j

$$\langle t_\alpha \rangle = \frac{1}{N} \sum_{k=1}^{N} \left| t_\alpha \left(\left\{ x_j^{(k)} \right\} \right) \right|$$

N : Sample size

t_α : Taylor coefficient labeled by α

- $\langle t_\alpha \rangle$ is the arithmetic mean of $|t_\alpha|$, evaluated on the whole input space that is sampled by the test data set.

- Introduce nomenclature of *generalized features* of the input feature space:

$\alpha = x_1, x_2, \ldots$ 1. order feature of input space (~ 1. order derivative)

$\alpha = x_1 x_1, x_1 x_2, \ldots$ 2. order feature of input space (~ 2. order derivative)

$\alpha = x_1 x_1 x_1, x_1 x_1 x_2, \ldots$ 3. order feature of input space (~ 3. order derivative)

$\vdots$ $\vdots$