APPLICATIONS OF SEMI-SUPERVISED MACHINE LEARNING TECHNIQUES IN THE SEARCH FOR NEW BOSONS FOCUSING ON DI-LEPTON FINAL STATES AT THE ATLAS EXPERIMENT

PRESENTED BY
BENJAMIN LIEBERMAN

INSTITUTE FOR COLLIDER PARTICLE PHYSICS
UNIVERSITY OF THE WITWATERSRAND

TOPIC HIGHLIGHTS

PRESENTATION OUTLINE

Madala Hypothesis

Di-Lepton Final State Dataset

Machine Learning Approaches in Physics

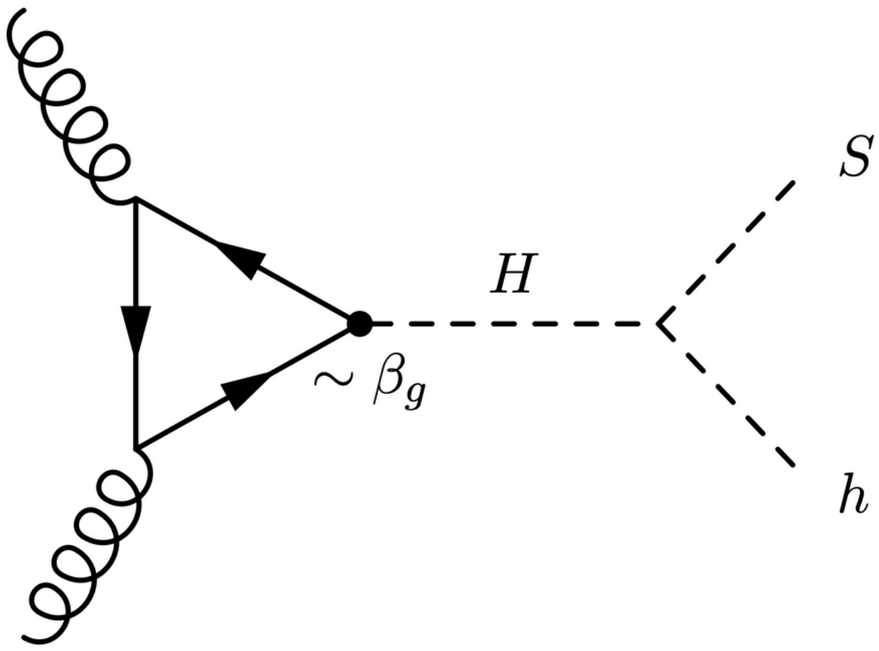
Input Variable Selection

TMVA Method Evaluation

Top Validation Region

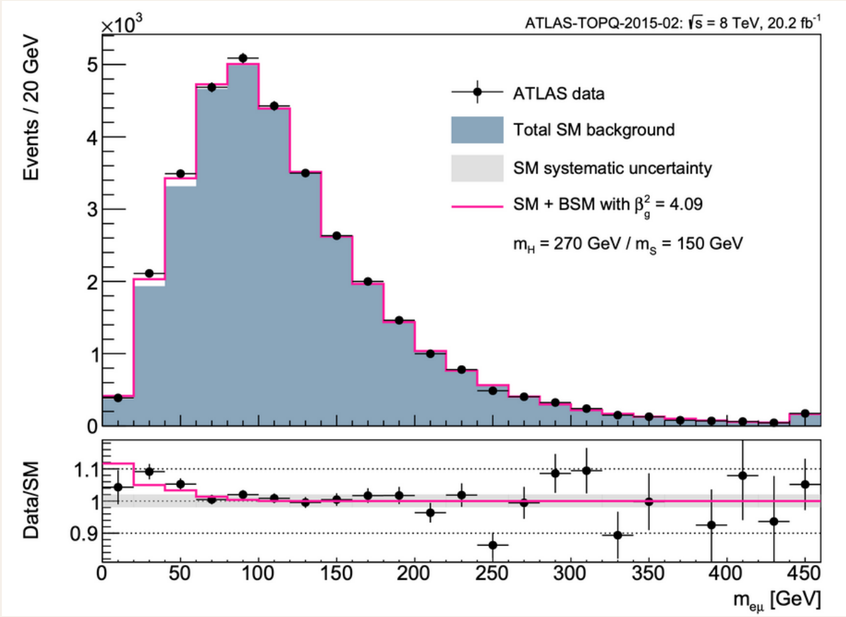
Conclusions

MADALA HYPOTHESIS

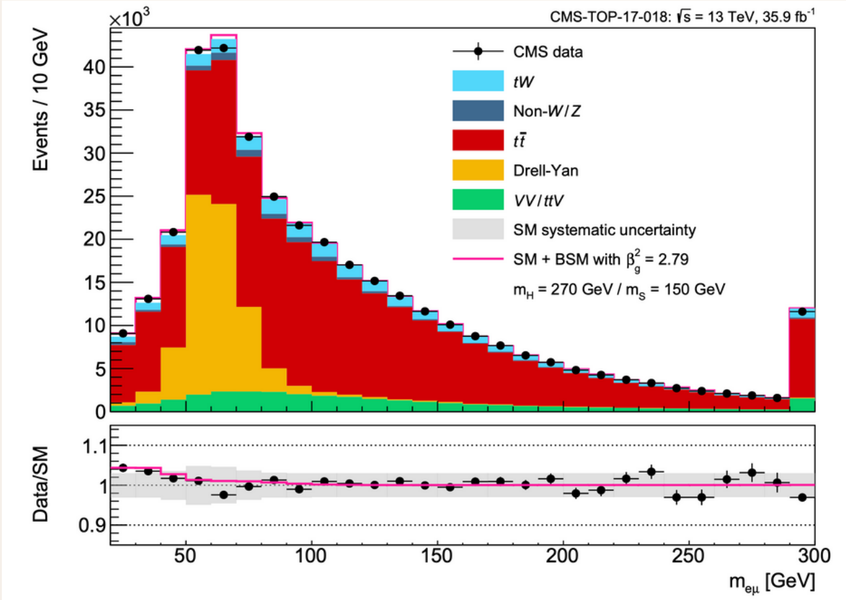


Observed discrepancies between data and standard model at low mll regions

TOP ANALYSIS MASS GRAPHS

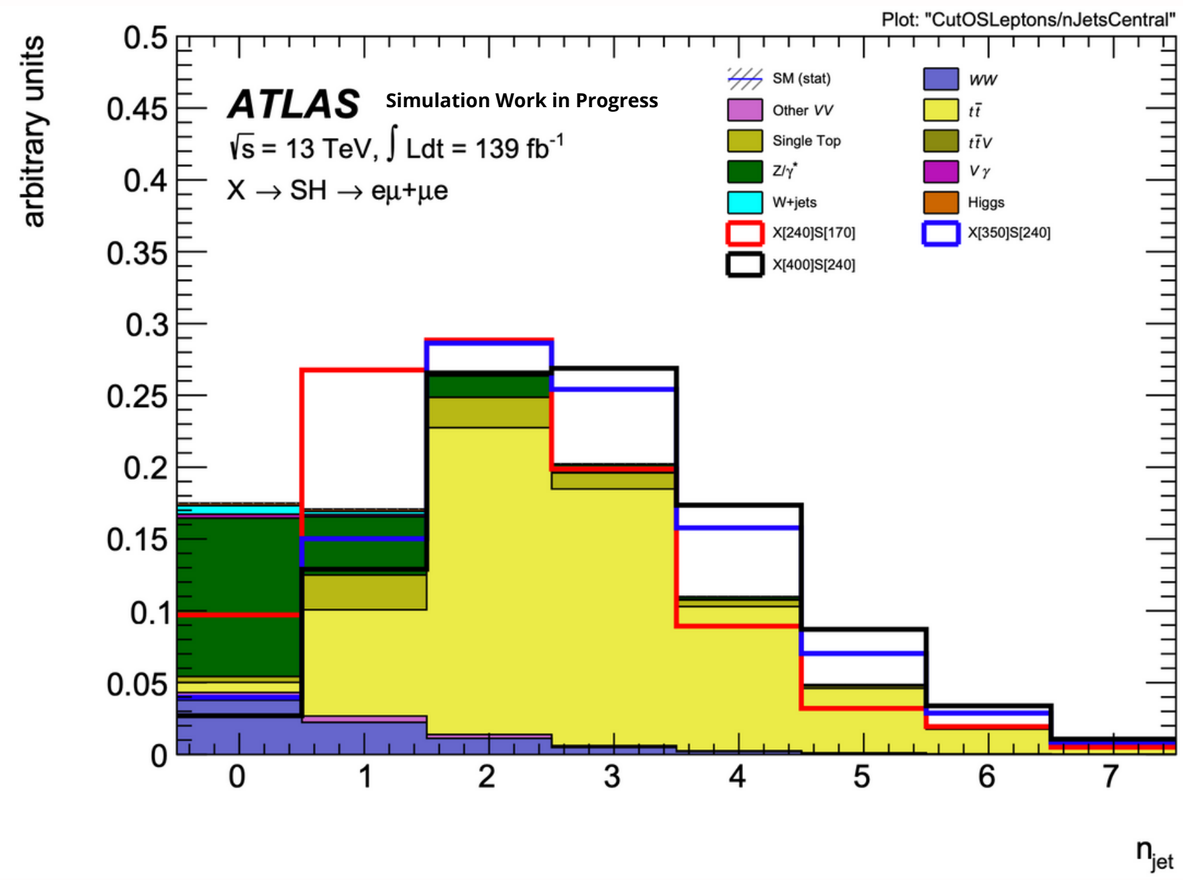


Run 1



Run 2

DI-LEPTON FINAL STATE MONTE CARLO DATASET



Di-Lepton Dataset:

- Selection: 1 electron + 1 muon with total electric charge = 0
- pt(leading) > 27GeV, pt(subleading) > 15GeV
- We apply the common trigger selection, jet cleaning and goodrun list to the dataset.

Di-lepton final state signals:

X[240]->S[170]H,
X[350]->S[240]H,
X[400]->S[240]H

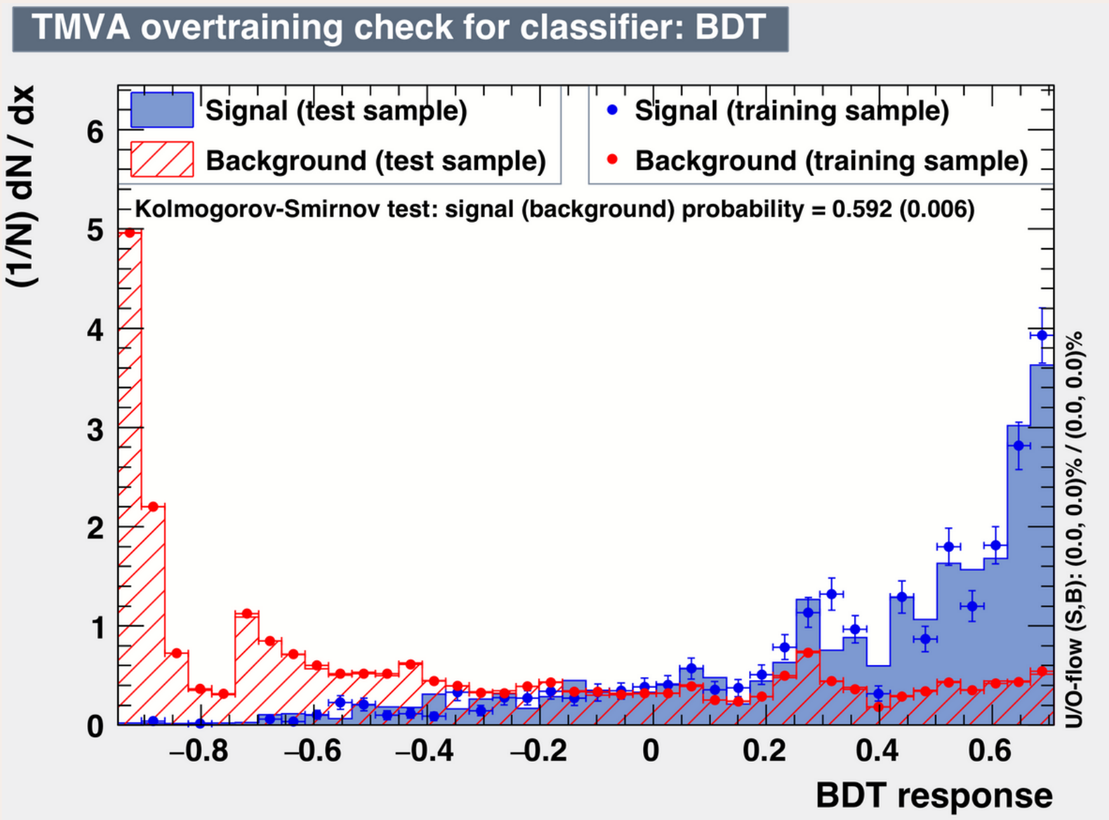
Background processes:

HIGGS,
WW,
OTHER VV,
Vγ,
TTBAR,
SINGLE TOP,
Z/γ*,
W+jets MC

MACHINE LEARNING APPROACH IN PARTICLE PHYSICS

Determine trends and/or features in a dataset to extract a signal process from a magnitude of background processes.

FULL SUPERVISION CLASSIFICATION EXAMPLE:
SEPARATION OF SIGNAL AND BACKGROUND DATASETS



FULL SUPERVISION VS WEAK SUPERVISION

FULL SUPERVISION

TRAINING DATASETS

Sample 1: Labelled Background dataset
Sample 2: Labelled Signal dataset

CHARACTERISTICS

Excellent classification of signal from background based on well defined physics in training datasets. Results however are biased to characteristics of given training set.

WEAK SUPERVISION

TRAINING DATASETS

Sample 1: Labelled Background dataset
Sample 2: Unlabelled Signal + Background mixed dataset

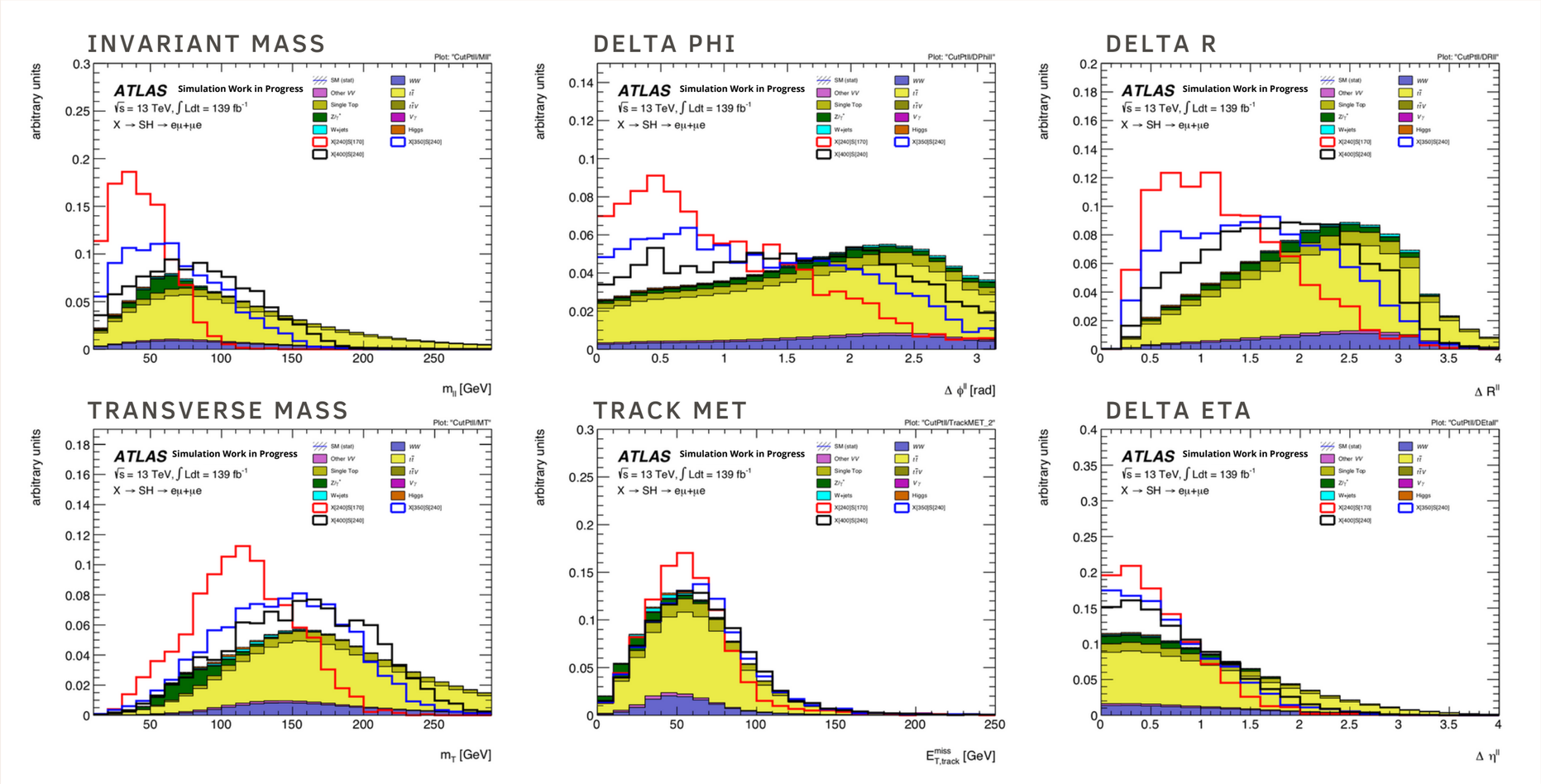
CHARACTERISTICS

Classification of signal from background with well guided training samples.
Provide classification of datasets that are not as well defined without limiting results by currently understood physics.

DILEPTON KINEMATIC INPUT VARIABLE SELECTION

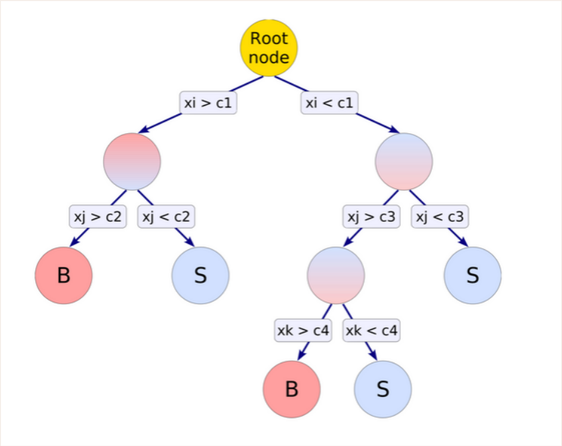
BASIC

- Two leptons with different flavour and opposite sign
- At most 1b-tagged jet
- Ptll > 30 GeV



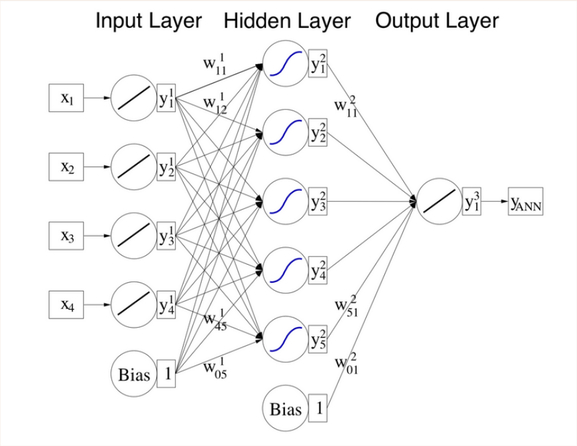
TMVA METHOD ANALYSIS

BOOSTED DECISION TREES (BDT)



- Binary tree structured classifier.
- Repeat binary yes/no decisions for a single variable at a time until the criterion is fulfilled.

MULTILAYER PERCEPTRON (MLP)



- Basic Neural Network.
- Single Hidden layer.
- Always feed farward.

DEEP NEURAL NETWORK (DNN)

$$W_{i,j}^k \rightarrow W_{i,j}^k - \alpha \frac{\partial J(\mathbf{x}_b, \mathbf{y}_b)}{\partial W_{i,j}^k}$$

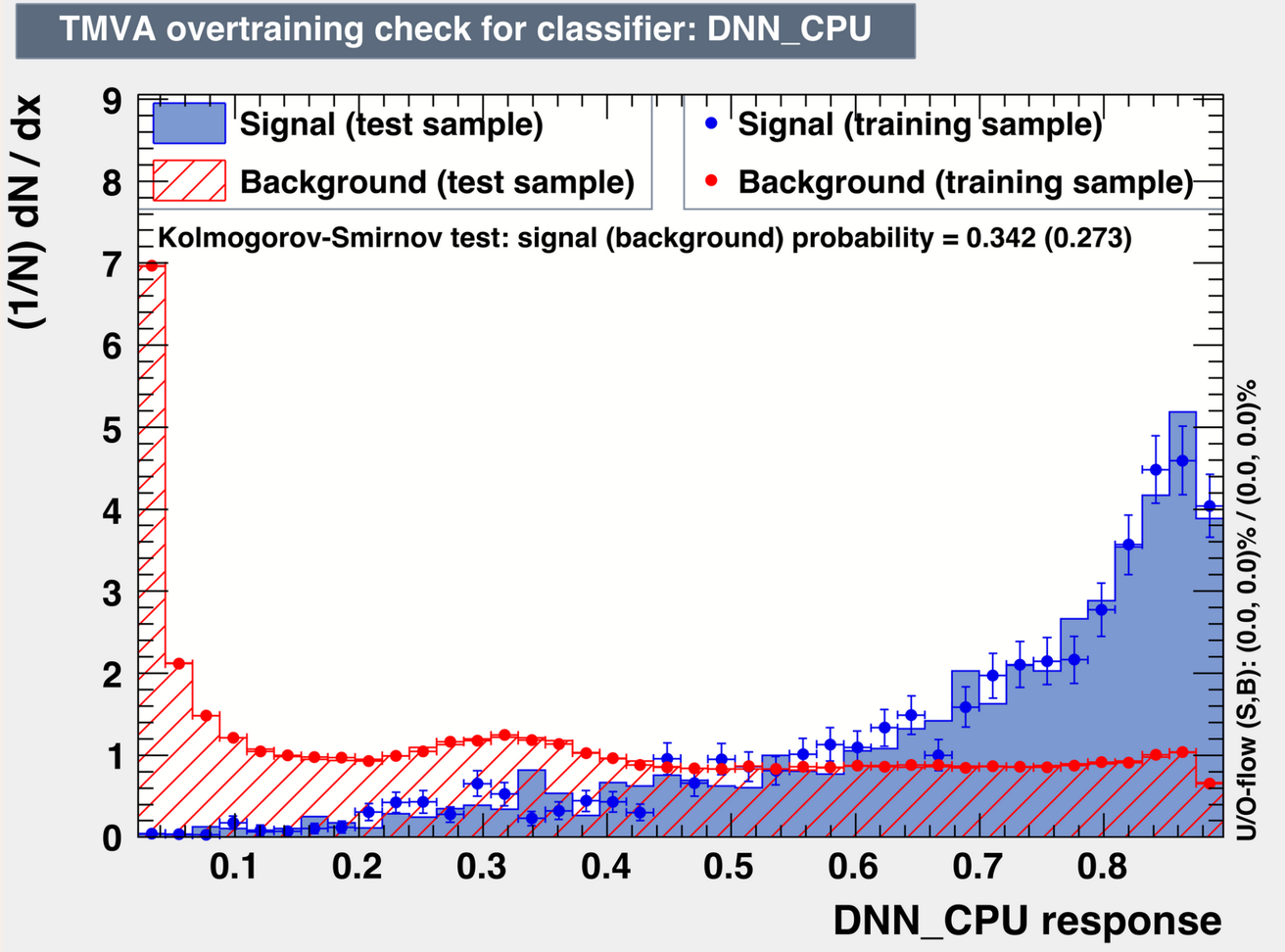
$$\theta_i^k \rightarrow \theta_i^k - \alpha \frac{\partial J(\mathbf{x}_b, \mathbf{y}_b)}{\partial \theta_i^k}$$

- Convolutional Neural Network with multiple hidden layers.
- Weights term is updated for each layer to account for internal biases..

MACHINE LEARNING OVERTRAINING CHECKS

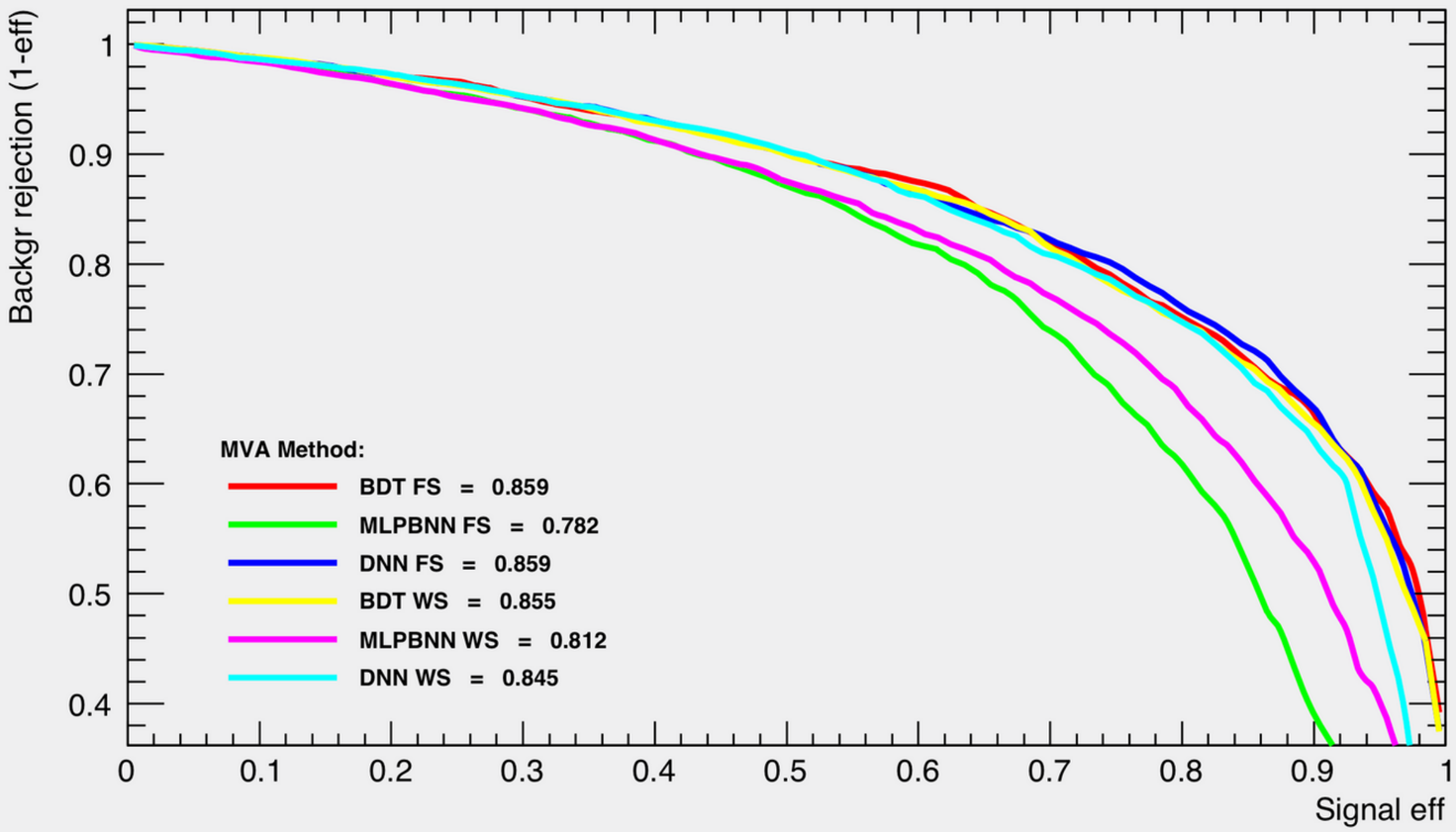
KOLMOGOROV-SMIRNOV TEST

- Quantifies the ability of the algorithm to not be biased to the training dataset, due to learning to many parameters of that training data.
- Provides a good indication of how effectively the model can be applied to a new dataset.
- KS Score: DNN > MLP > BDT



FULL SUPERVISION VS WEAK SUPERVISION COMPARISON FOR TMVA METHODS

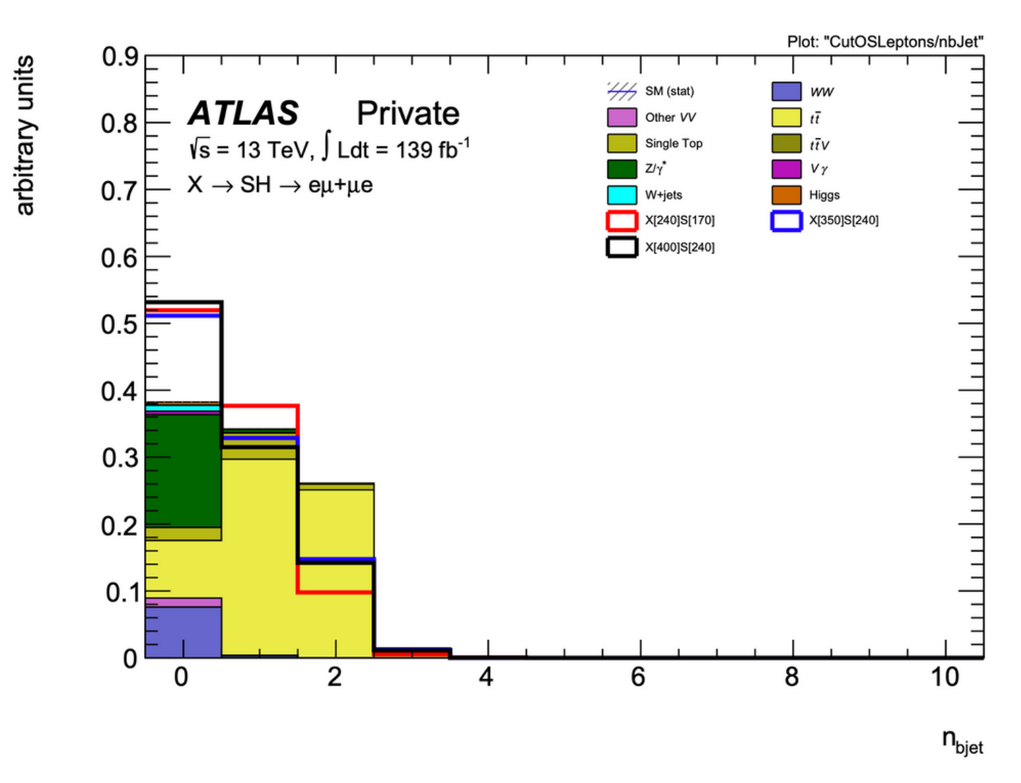
ROC CURVE COMPARISON OF TMVA METHODS FOR FS AND WS



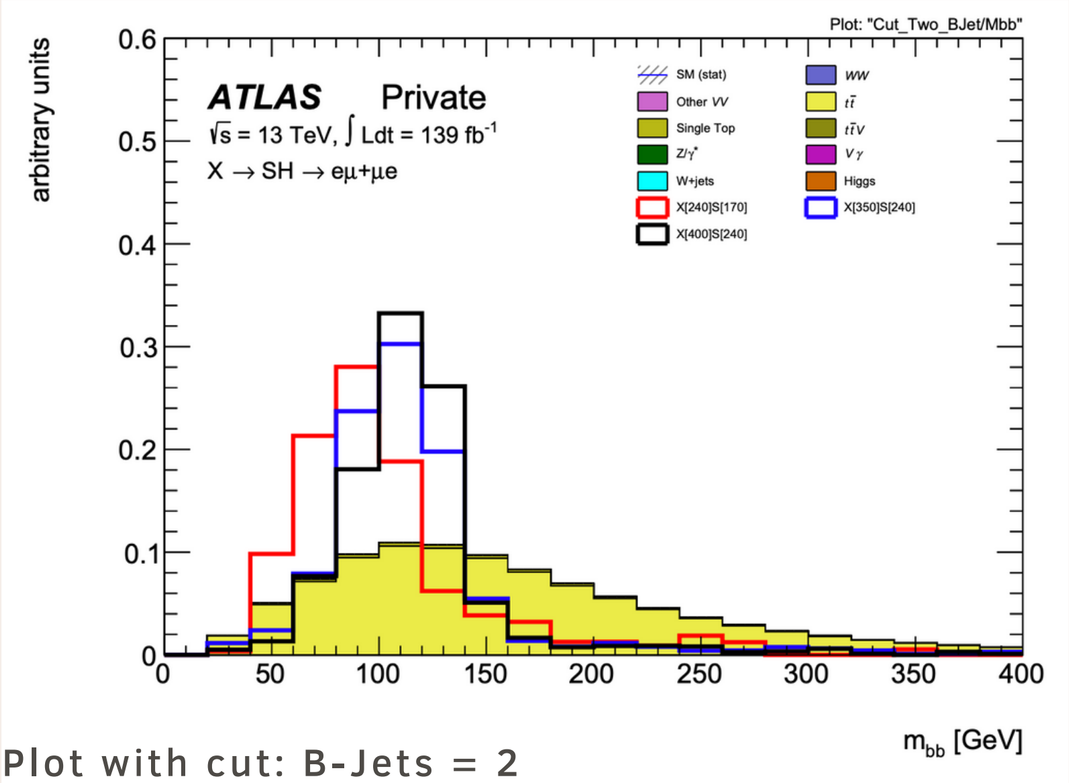
- Both BDT and DNN show excellent signal classification (±86%) for both FS and WS.
- BDT, MLP and DNN all show excellent correlation between FS and WS results. This shows that for the dilepton dataset weak supervision is a viable method.

DI-LEPTON TOP VALIDATION REGION ANALYSIS

NUMBER B-JETS ≥ 2



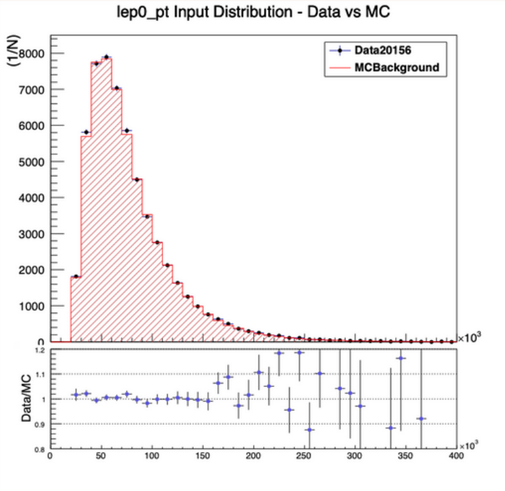
Mbb > 150 GeV

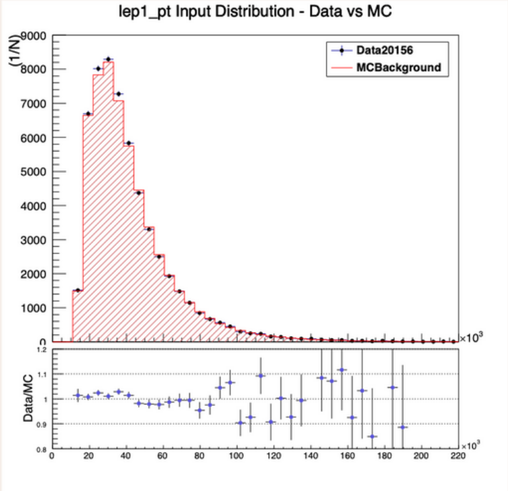


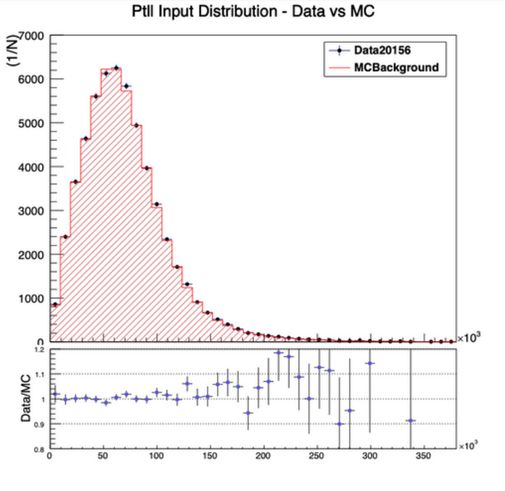
Plot with cut: B-Jets = 2

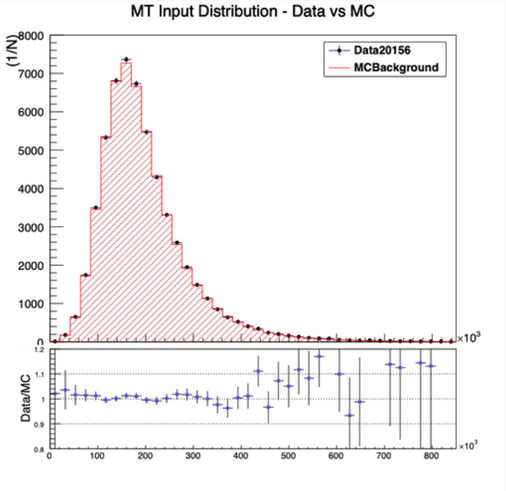
TOP VALIDATION REGION INPUT VARIABLES

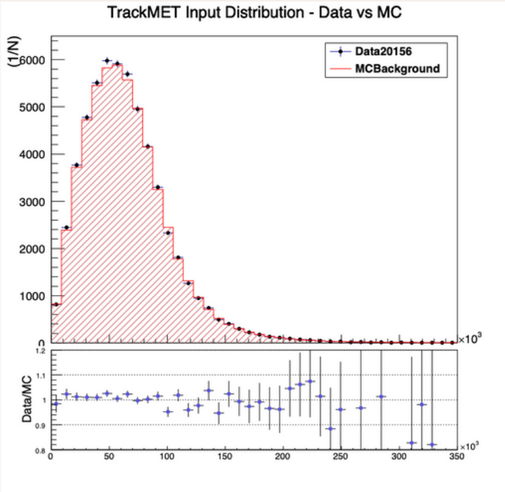
ATLAS 2015/16 DATA VS MONTE CARLO DATASET

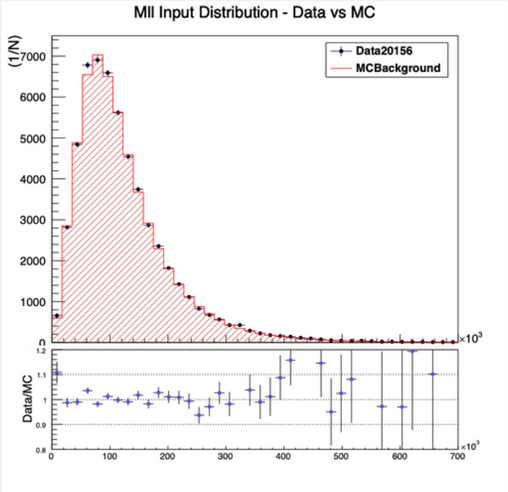


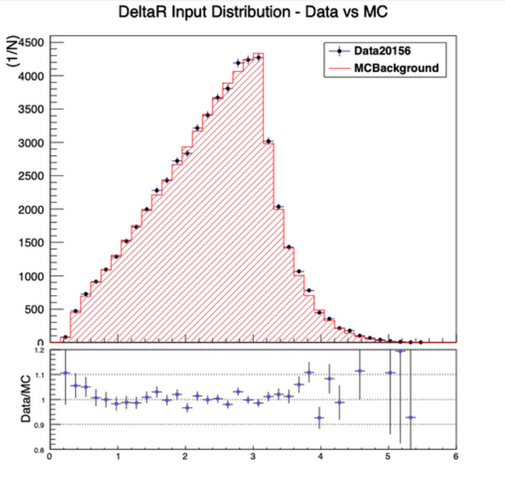


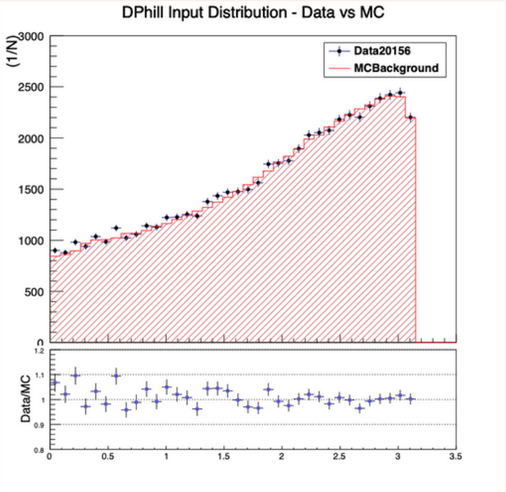










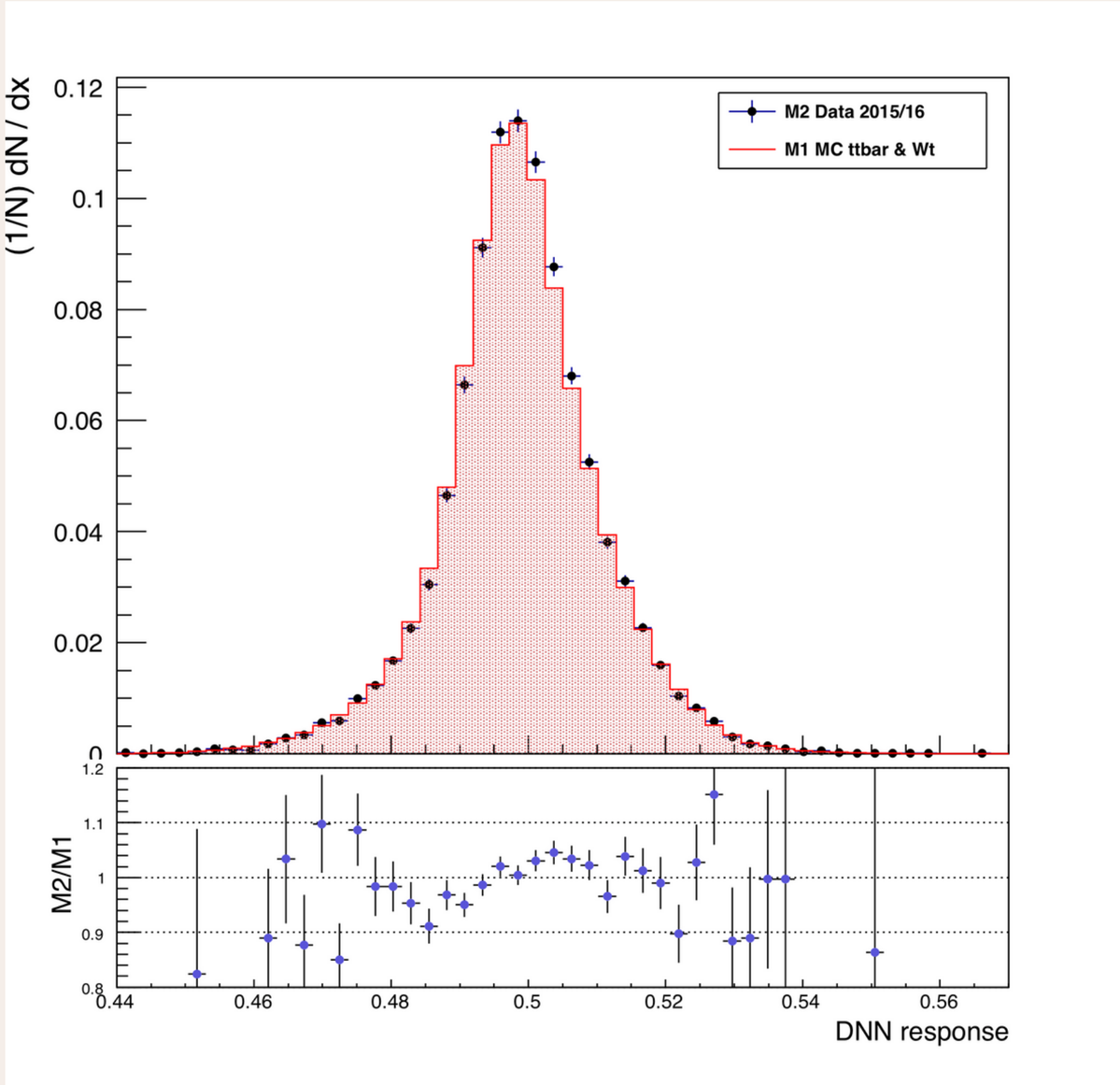


DNN OUTPUT DISTRIBUTION EVALUATION

OUTPUT DISTRIBUTION BIASIS

1. Possible mis-modeling on TMVA method
2. New Physics

TOP VALIDATION REGION OUTPUT DISTRIBUTION: 2015/16 DATA VS MONTE CARLO



CONCLUSIONS

WEAK SUPERVISION

Weak supervision is shown to be a viable solution for classification of datasets that aren't as well defined.

MACHINE LEARNING METHOD FOR SPECIFIC DATASET

DNNs provide minimal overtraining and internal bias as compared to BDT and MLP methods for dilepton dataset.

FUTURE

DNN bias determined in the top validation region needs to be further examined in order to evaluate the source of biasis.
Algorithm can then be applied to signal region of dilepton dataset to search for potential new physics.

REFERENCES

- Von Buddenbrock, Stefan et al. “The Emergence of Multi-Lepton Anomalies at the LHC and Their Compatibility with New Physics at the EW Scale.” Journal of High Energy Physics 2019.10 (2019): n. pag. Crossref. Web.

- Hoecker, A., Speckmayer, P., Stelzer, J., Therhaag, J., von Toerne, E., Voss, H., Backes, M., Carli, T., Cohen, O., Christov, A. and Dannheim, D., 2007. TMVA-toolkit for multivariate data analysis. arXiv preprint physics/0703039.