Analysis of Frequentest Study Results in Quantifying Fake Signal Generated in the Training of Semi-Supervised DNN Classifiers

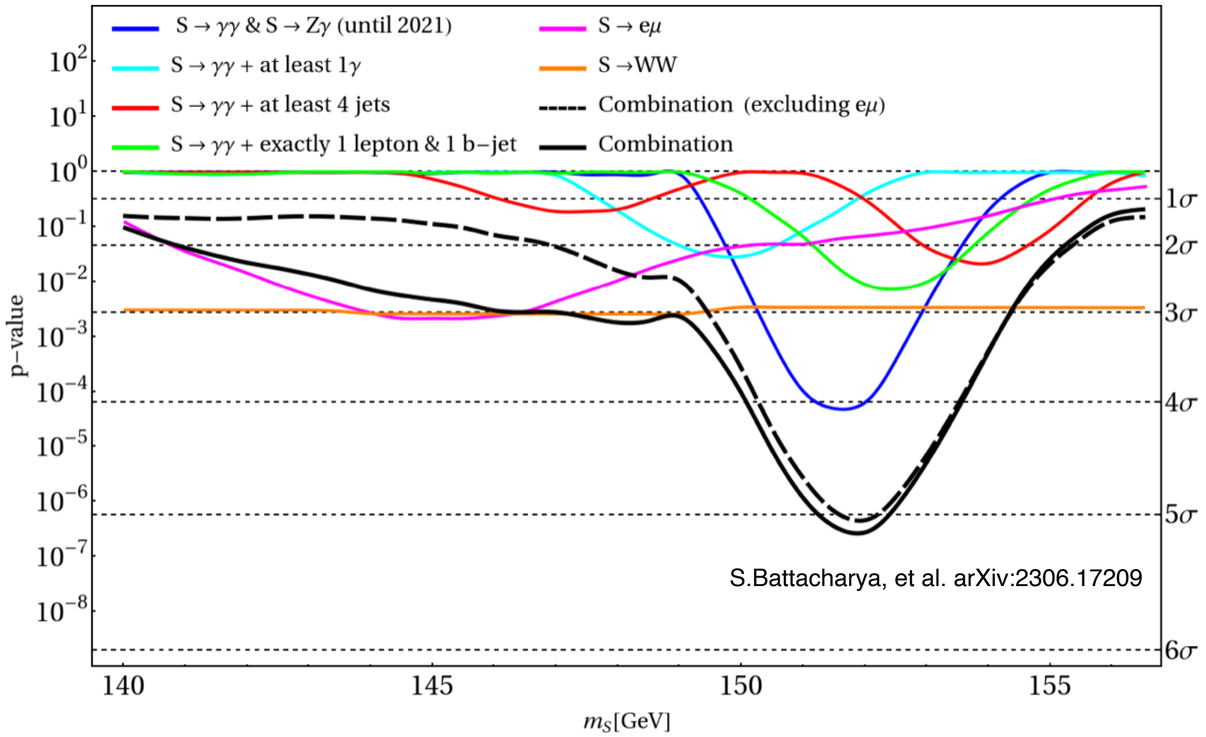
By **Benjamin Lieberman**,
Supervised by **Bruce Mellado**

INSTITUTE FOR COLLIDER PARTICLE PHYSICS
UNIVERSITY OF THE WITWATERSRAND

NRF National Research Foundation | iThemba LABS Laboratory for Accelerator Based Sciences

ATLAS EXPERIMENT

Introduction



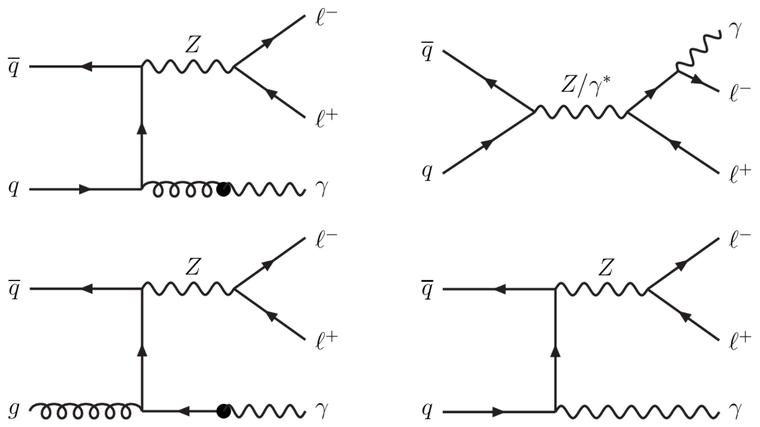
Context of Presentation:

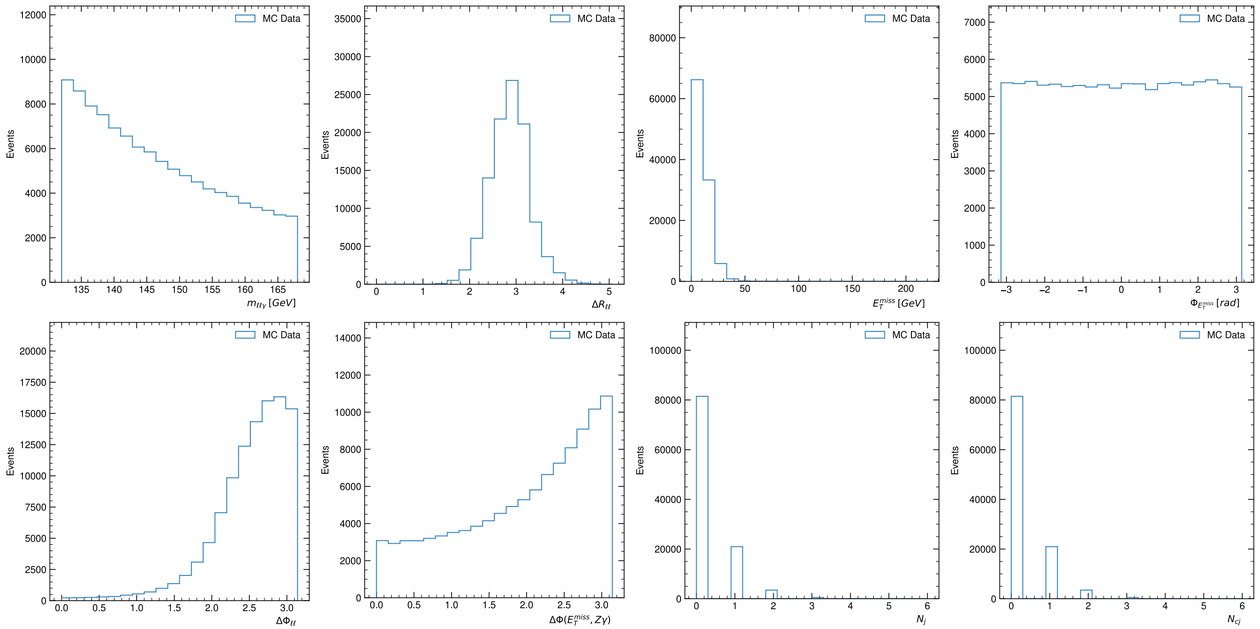
- Explore excess of potential heavy scalar, S, around ±150GeV
- BSM searches for Zy resonances
- Use weakly or semi-supervised machine learning classifier.
  - ➡Reduce model dependencies
- Expose internal error generated by using semi-supervised machine classifiers

Zγ Resonance Searches

- Model Used: 2HDM+S where S is a singlet scalar [Stefan von Buddenbrock *et al* 2017 *J. Phys.: Conf. Ser.* **889** 012020]
- Heavy scalar, H, decays predominantly into SS and Sh, where h is the SM Higgs boson.
- Model exposes, in multi-lepton anomalies and astro-physics anomalies when complimented by a Dark Matter Candidate.
- Certain models considered in this study, predicted the decay of the new heavy scalar to Zy final state.

pp → H → Zγ→ (ℓ+ℓ–)γ





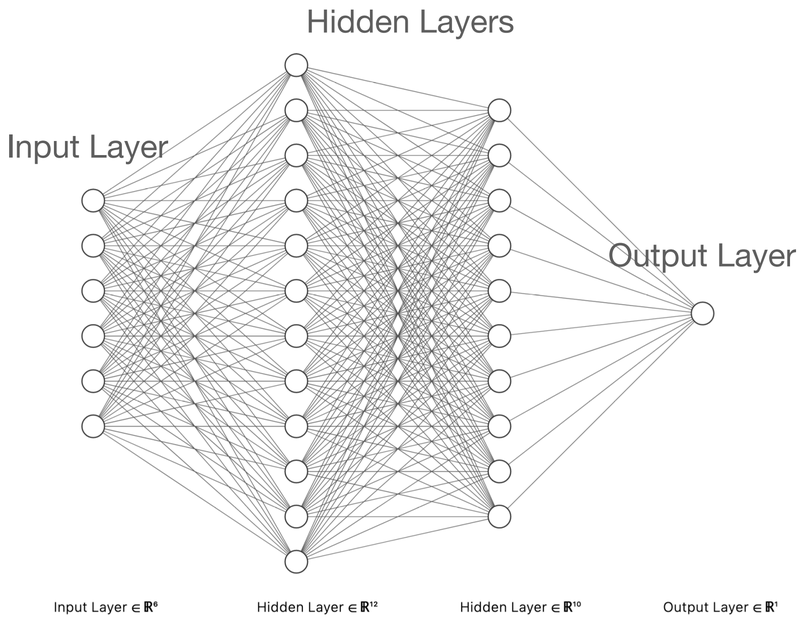
Fast Simulation Monte Carlo Zy Data generation:

Using Madgraph5 with NNPDF3.0 parton distribution function. Parton level generation is done using Pythia and detector level simulation is done using Delphes(v3)

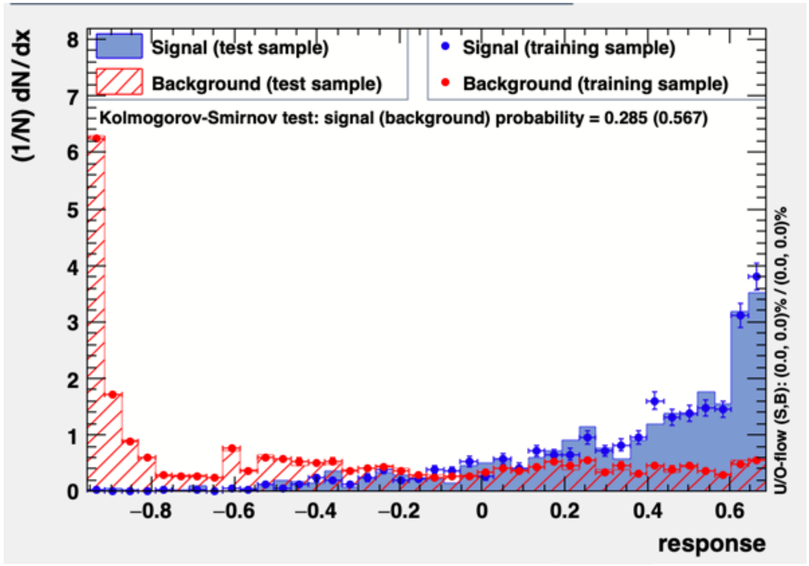
Introduction to Machine Learning

Classification and Anomaly Detection

Neural Networks

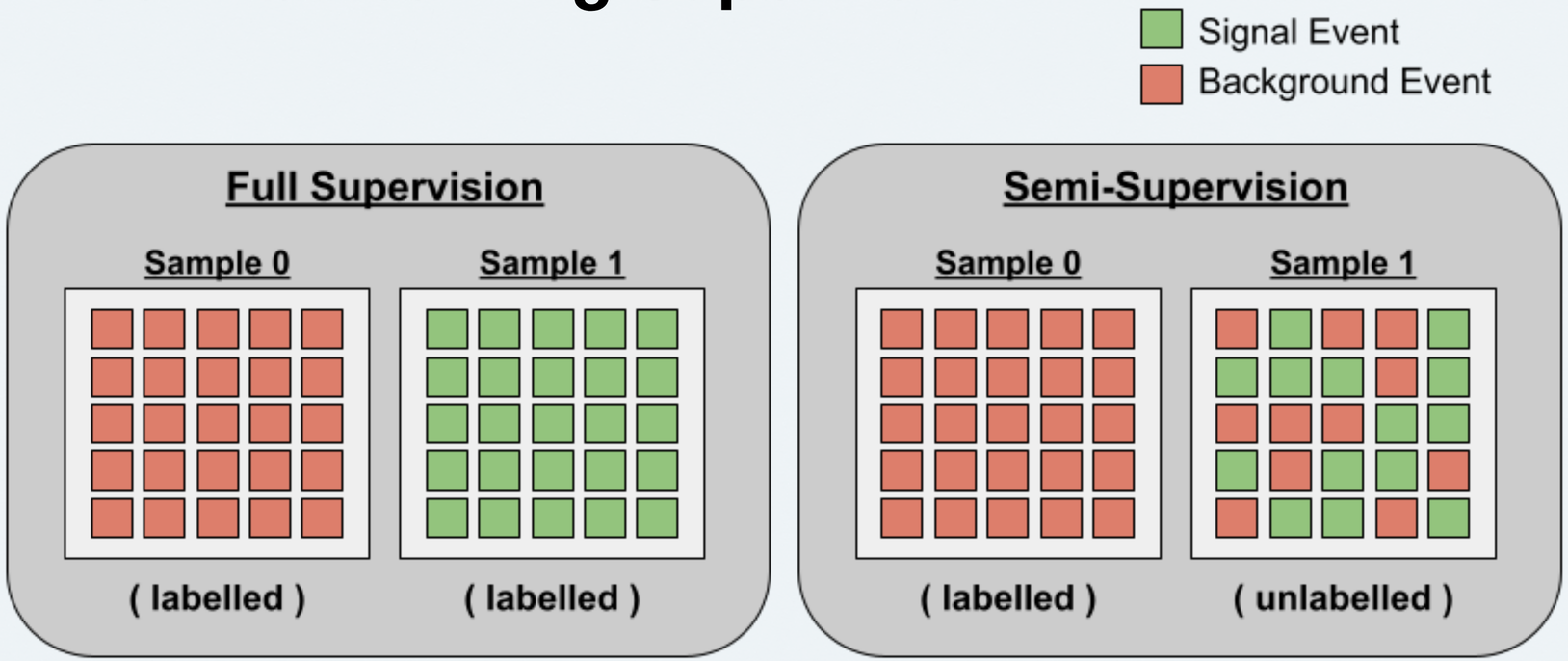


Response Distributions



Machine Learning Classifiers

Machine Learning Supervision



- Uses a fully labelled dataset
- Well defined “signal” and “background”
- Best Results

- Uses a partially labelled dataset
- Well defined “background”
- Good Results

Why would we use semi-supervision?

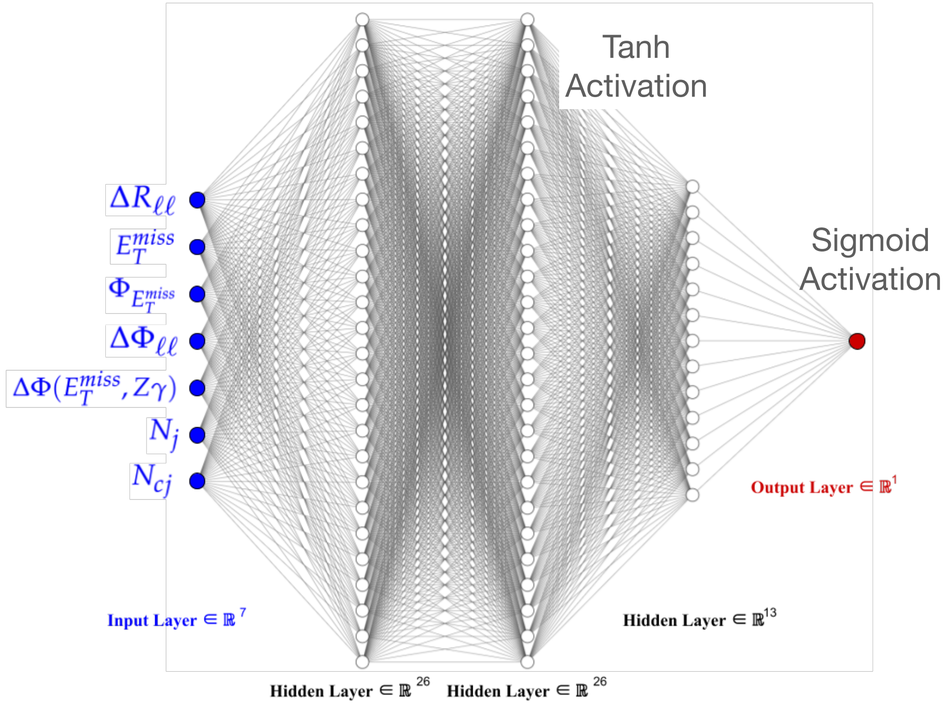
Reduce Model dependencies

Decreases biases caused by known physics. Reduces constraints placed on what “signal” must look like

Semi-Supervised DNN Classifier

Deep Neural Network Classifier

Model Architecture



During the training of neural networks, over-training/over-fitting can cause background events to be incorrectly classified as signals.

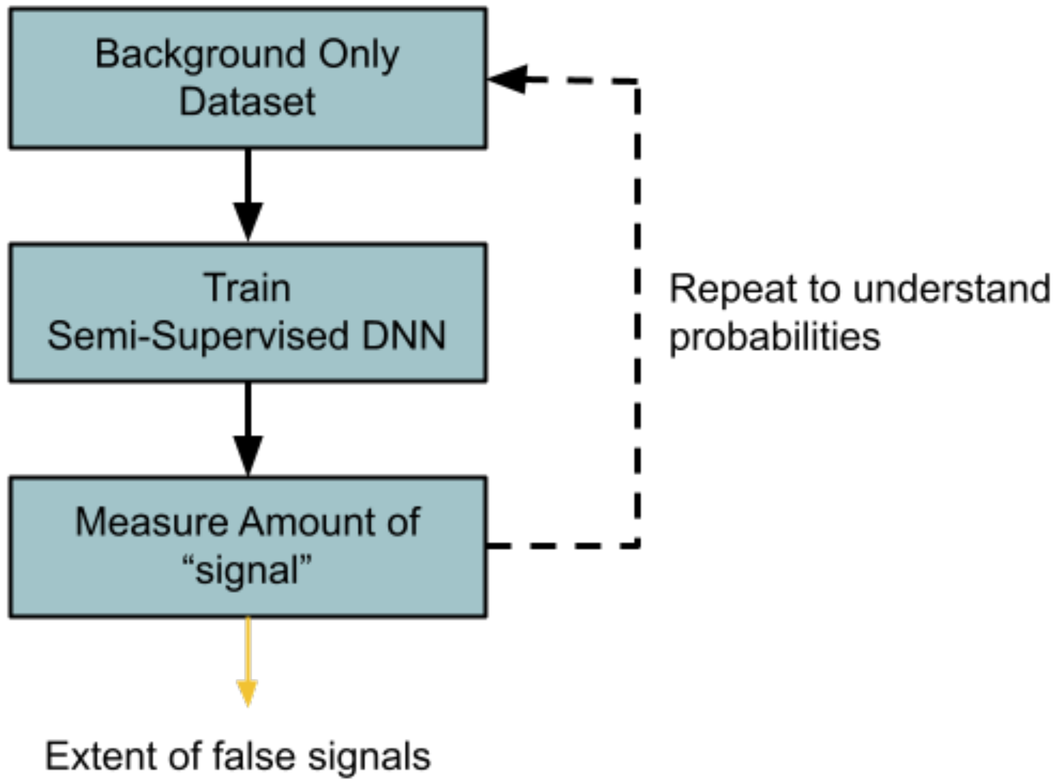
How often does the semi-supervised DNN model classify background processes as signal?

The optimised DNN hyper-parameters:

Learning Rate = $1*10^{-3}$
Batch size = 256
Optimiser = Adam

Error Measurement in Resonance Searches when using Semi-Supervised DNN Classifiers

Frequentest Approach: Pseudo Experiment



Why Frequentest Approach?

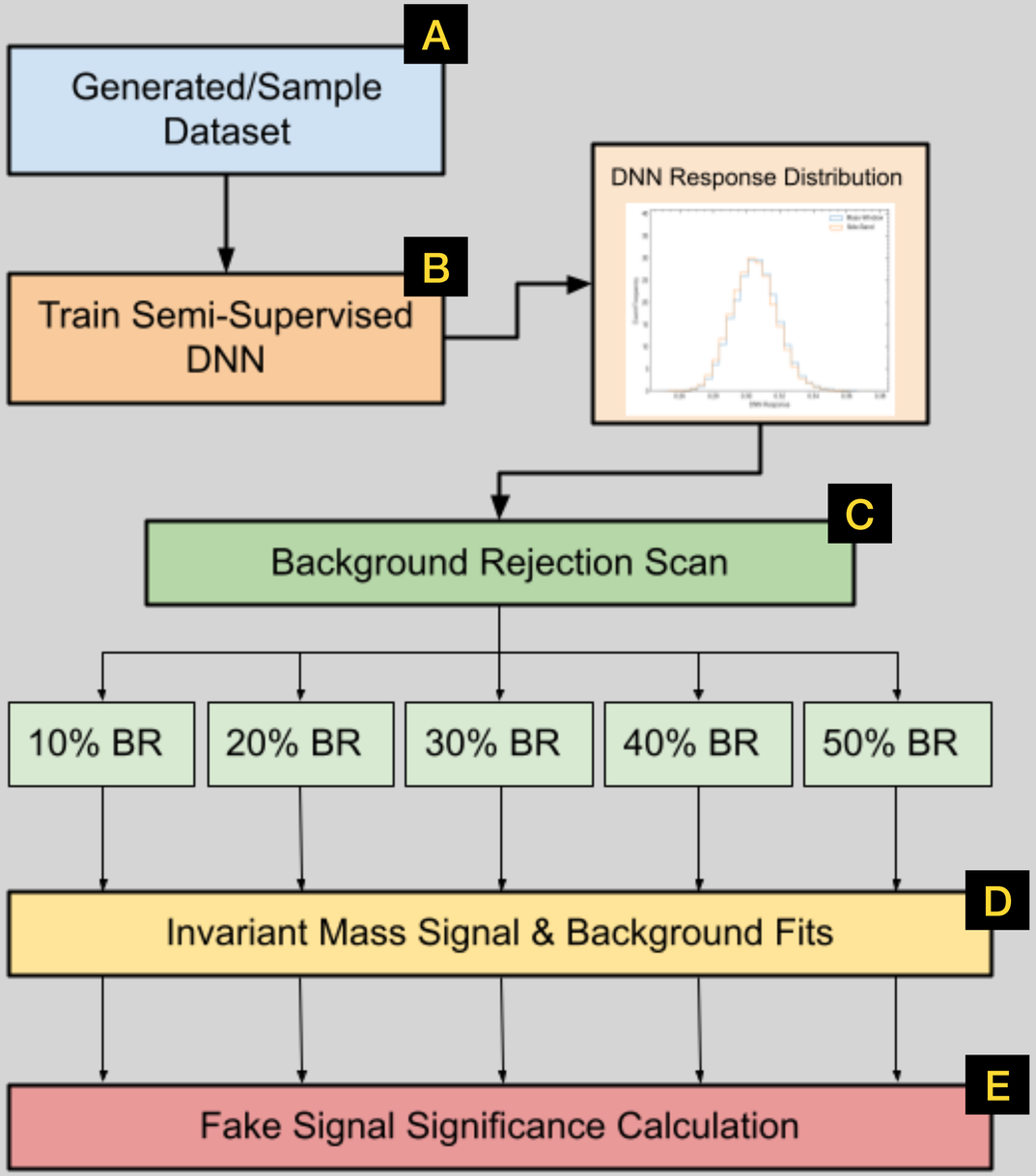
When conducting kinematic scans and/or resonance searches within a given mass range, the significance of observing a local excess of events, must consider the probability of observing the excess elsewhere within the range. This is known as the "**look elsewhere effect**".

Study Setup:

- Fixed mass around 150GeV
- Signal (mass-window) region: [144, 156] GeV
- Background (sideband) region: [132, 144) & (156, 168] GeV

Pseudo-Experiment

- A frequentest study consists of the repetition of a pseudo-experiment sufficient times to produce a statistically accurate distribution of results.
- In this study, each pseudo-experiment is used to measure the local signal significances resulting from the training of the semi-supervised DNN model.



A

Pseudo-Experiment: Data Sampling/ Generation

Data Sampling implemented:

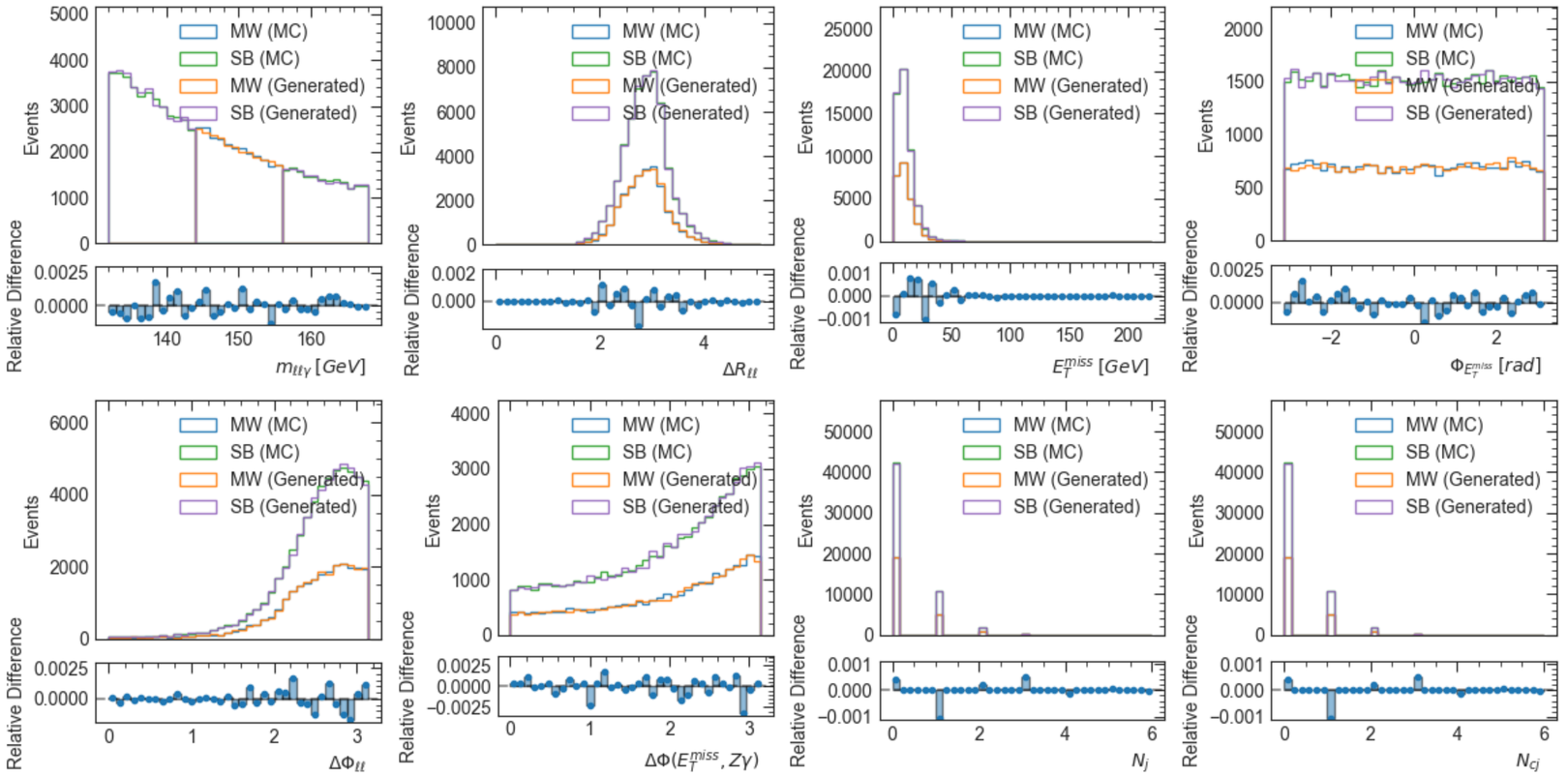
Kernal Density Estimation, **KDE**, method.

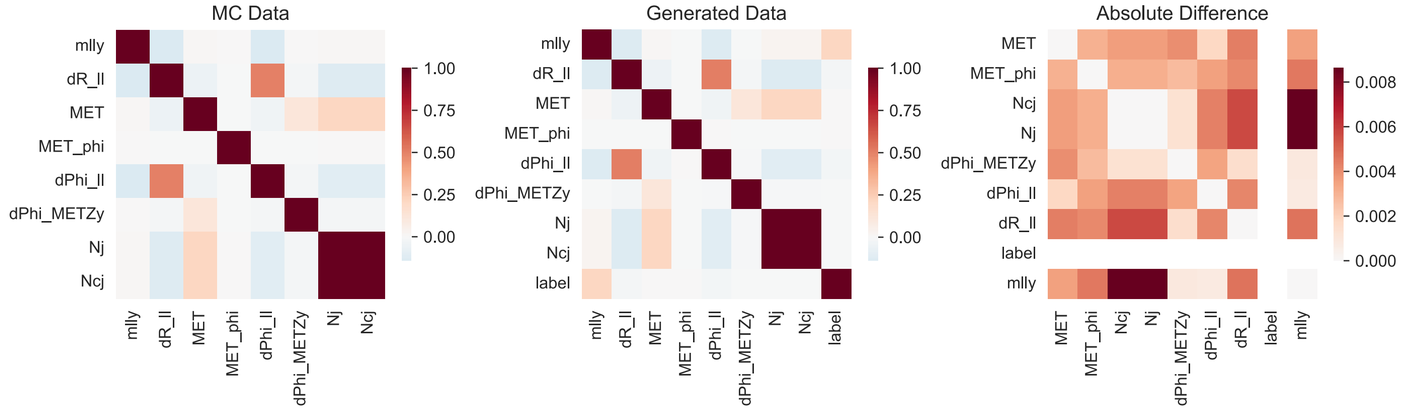
Excellent sampling method for synthesising events.

Data Sampling Considered:

Bootstrap, **KDE, VAE, GAN**.

Example of Generated training dataset:





B Pseudo-Experiment: DNN Training and Response Distribution

The semi-supervised DNN is trained on a generated/sampled Zy dataset.

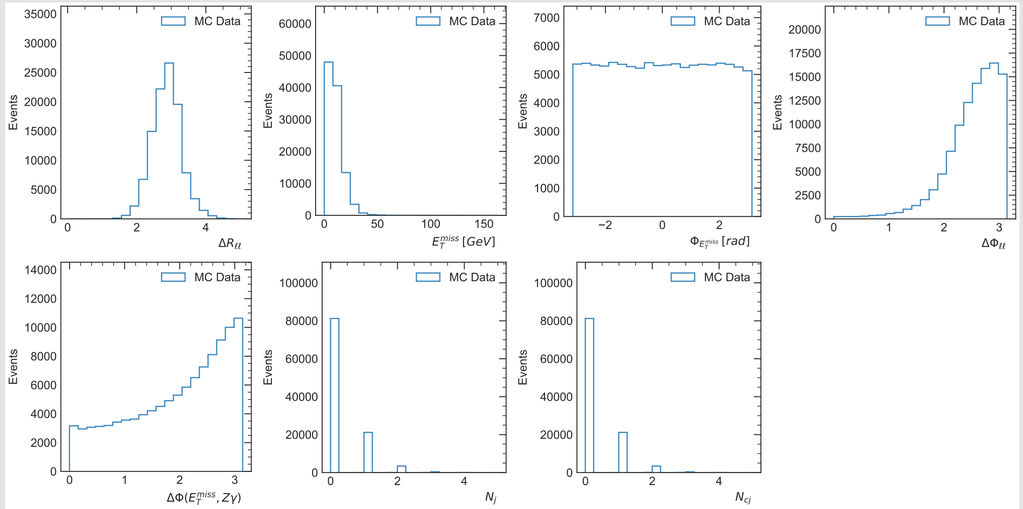
Sample 0 (background / side-band region):

$(132 \leq m_{\ell\ell\gamma} < 144)$ and $(156 < m_{\ell\ell\gamma} \leq 168)$

Sample 1 (Signal / mass-window region):

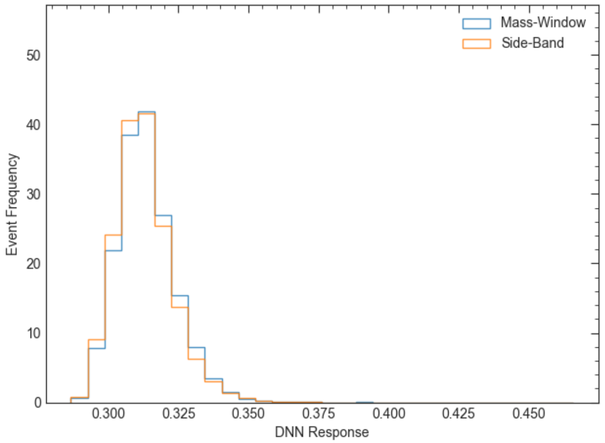
$(144 \leq m_{\ell\ell\gamma} \leq 156)$

Zy Training Dataset Feature Distributions

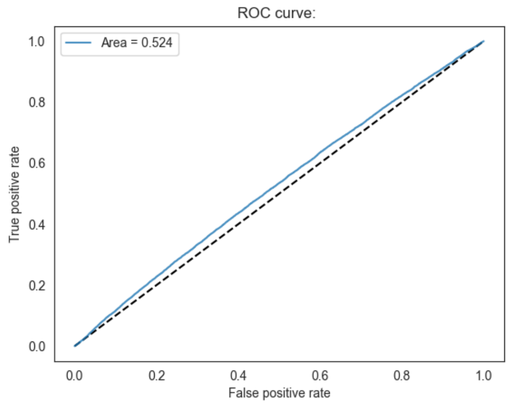


DNN Outputs

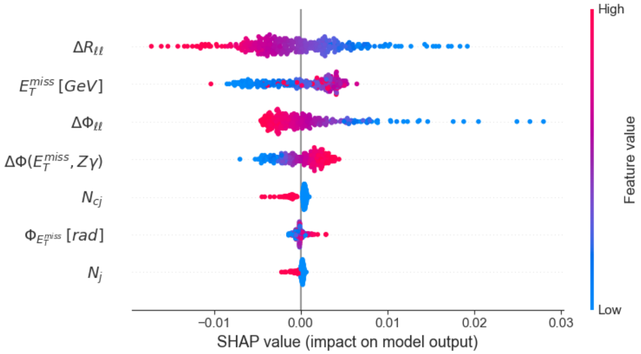
Response Distribution



ROC Curve

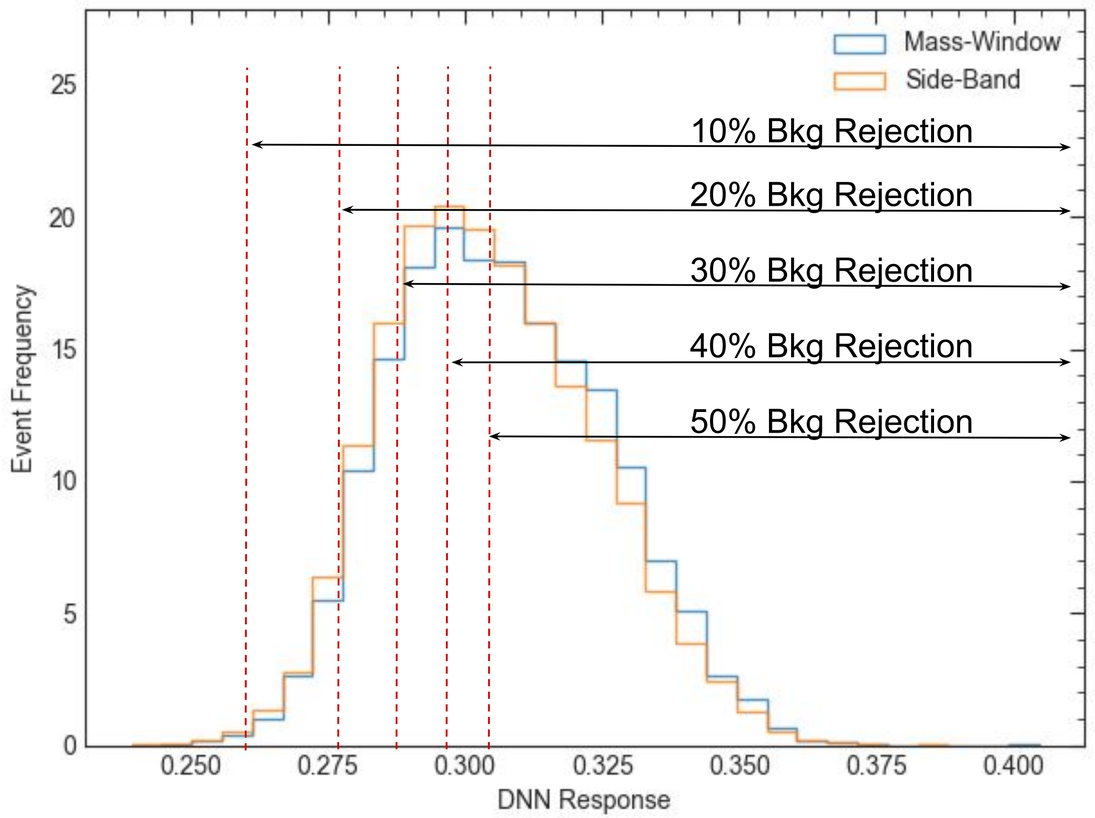


SHAP Feature Ranking



C Pseudo-Experiment: Background Rejection Scan

1. Scan response distribution extracting batches of events.
2. Each batch excludes percentages of events considered background (closer to zero).
3. Each batch of events is mapped to their corresponding invariant mass, mlly.
4. Each batches invariant mass distribution can therefore be used to extract a local significance



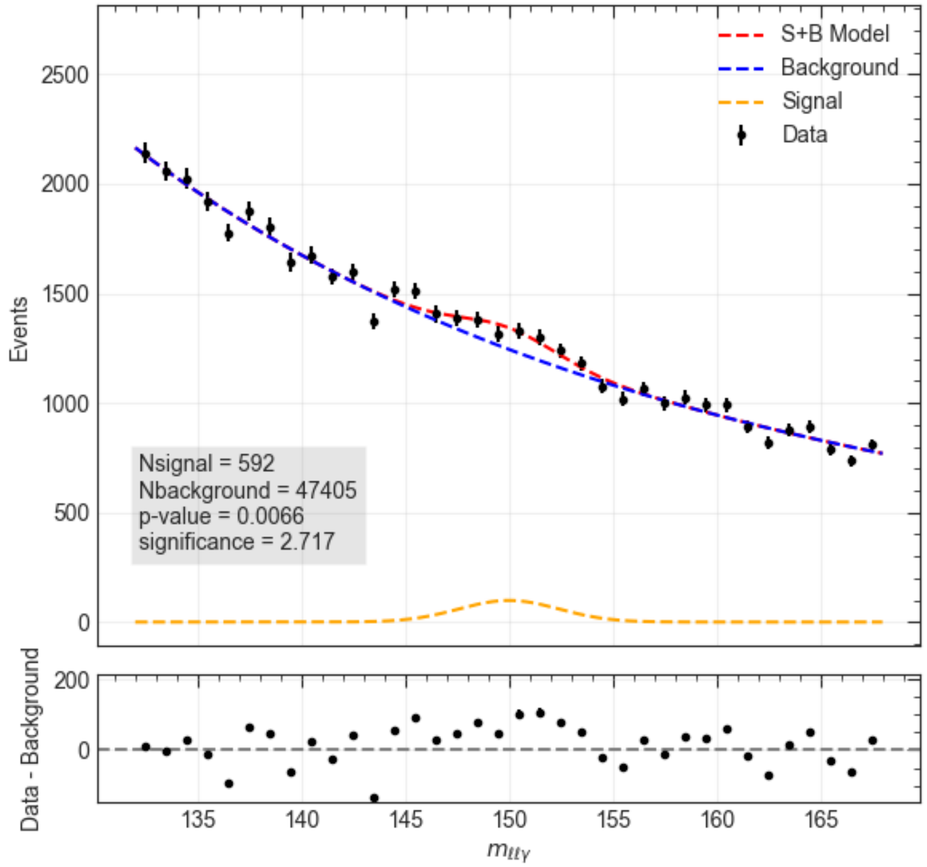
D Pseudo-Experiment: Invariant Mass Background Fits

Background and Signal Fitting

1. Background Function, f(x):
   - A mixed poly-log-exponential function of first order $f(x) = (1-x)^{c_0} * x^{c_1 + c_2 * \log(x)}$

2. Signal Function:
   - The signal function is a gaussian function
   - Mean, μ, fixed at mass = 150GeV
   - Sigma. σ, is the resolution = 2.4

3. Signal + Background Function, g(x):
   - The background fit parameters are fixed
   - Measures gaussian size



E Pseudo-Experiment: Fake Signal Significance Calculation

Parameter of interest: number of signal events

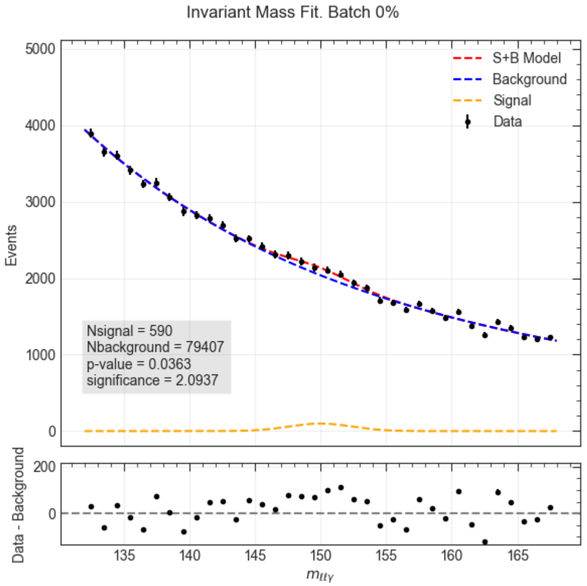
Observable: Invariant Mass, mlly

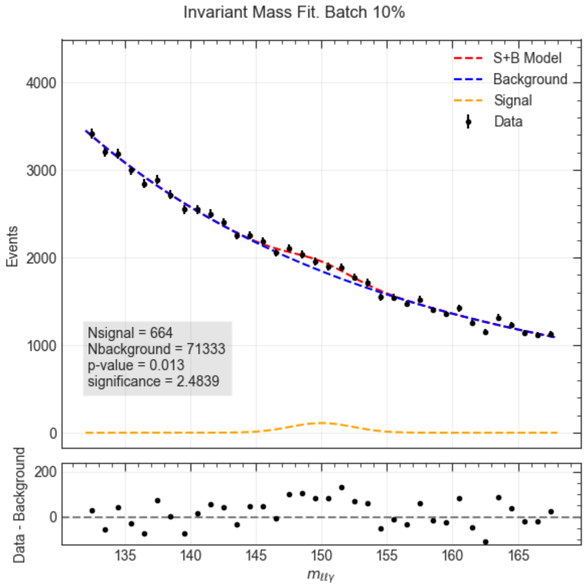
Null (background) Hypothesis: no signal events will be found in signal region

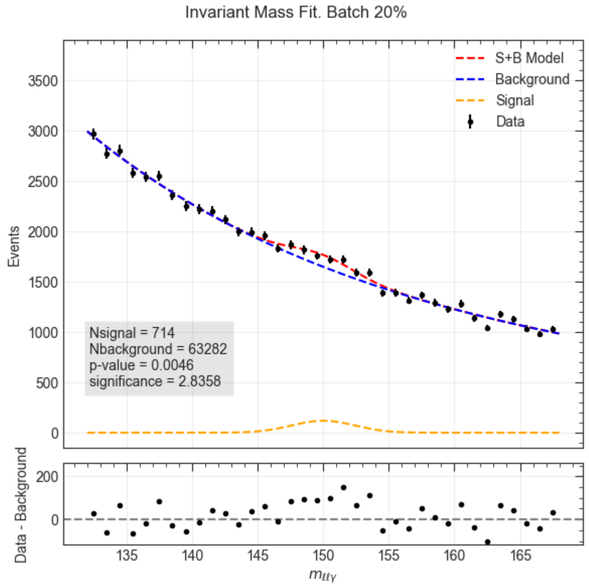
1. Signal and Background Fits by minimising Negative Log Likelihood
2. Integrate over background function to get number of background events, $\mathbf{N_b}$
3. Integrate over signal function to get number of signal events, $\mathbf{N_s}$
4. Calculate significance, Z, using the formula:

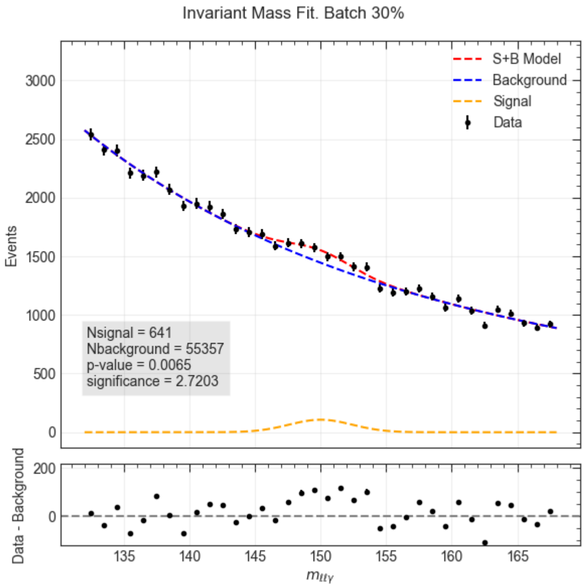
$$Z = \sqrt{2 \cdot ((N_s + N_b) \cdot \log(1 + \frac{N_s}{N_b}) - N_s)}$$

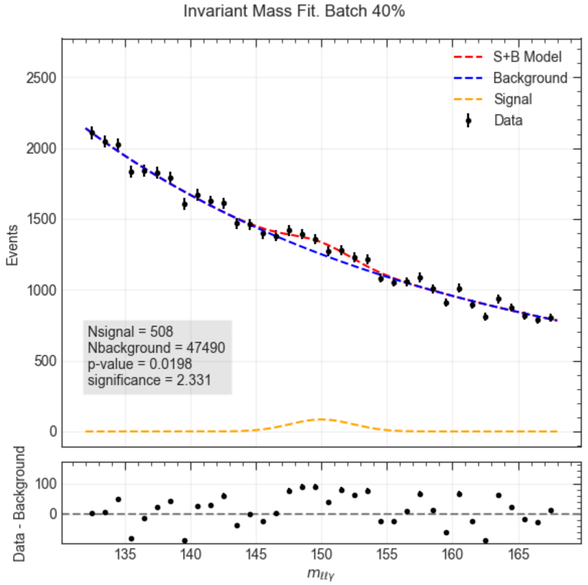
Example Pseudo-Experiment Results

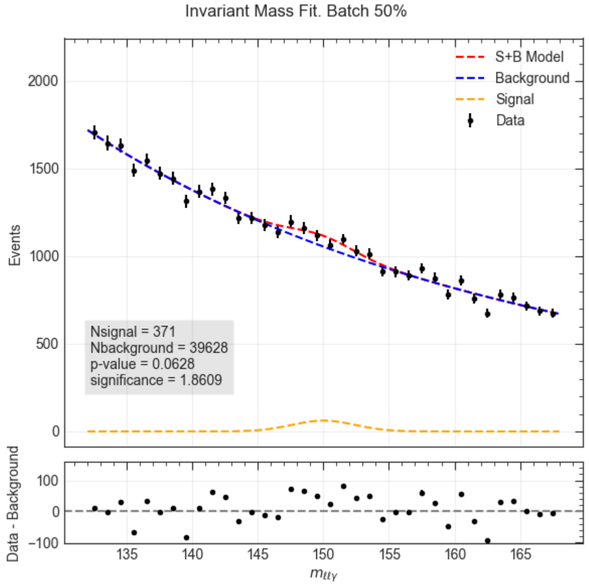






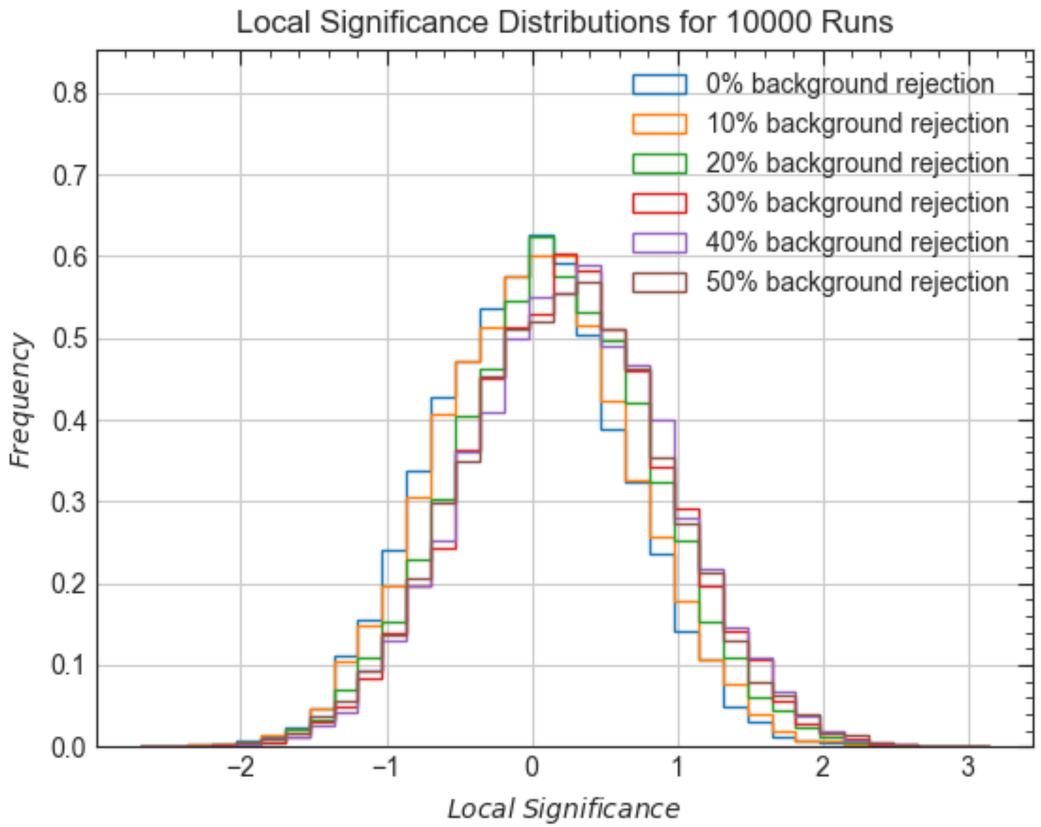


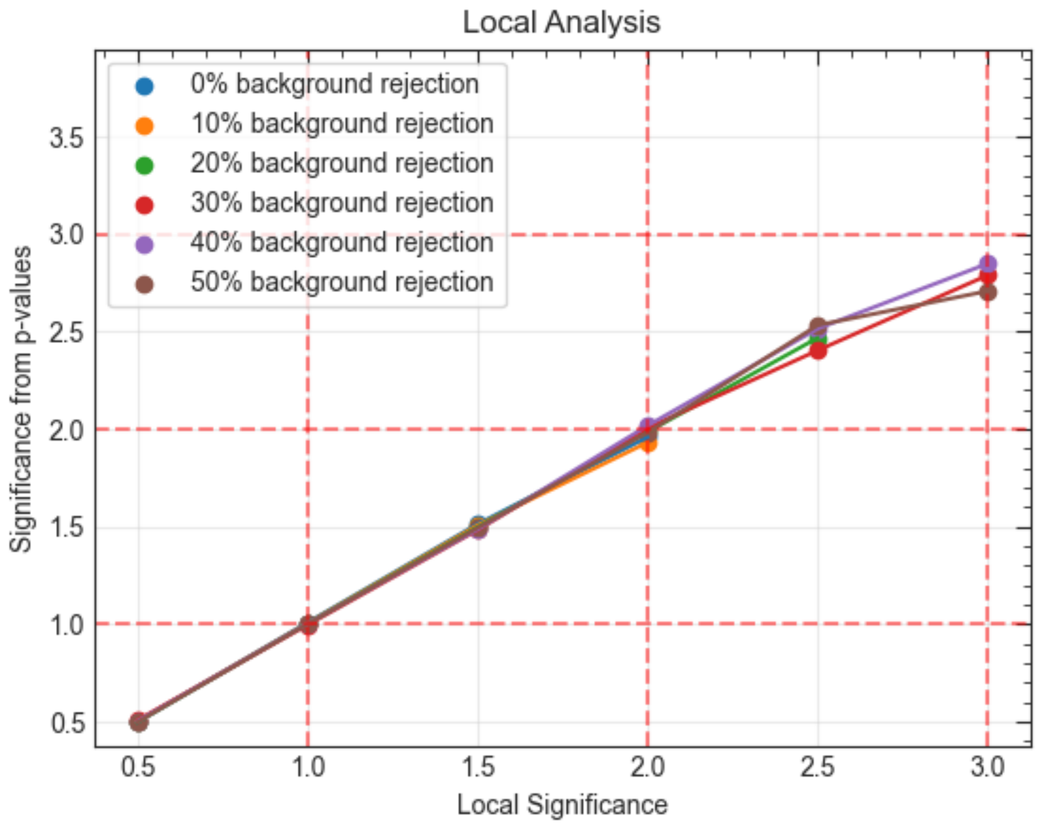




Frequentest Study Results - Local

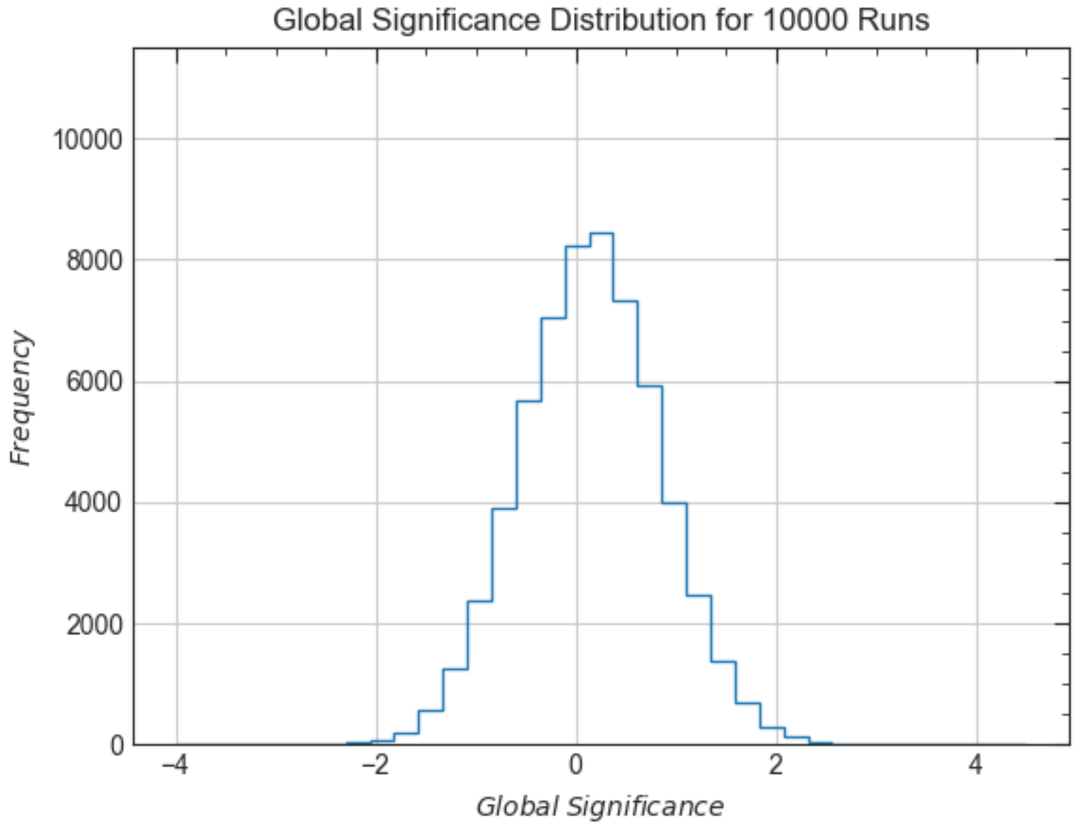
Local significance results for 10,000 pseudo-experiments:

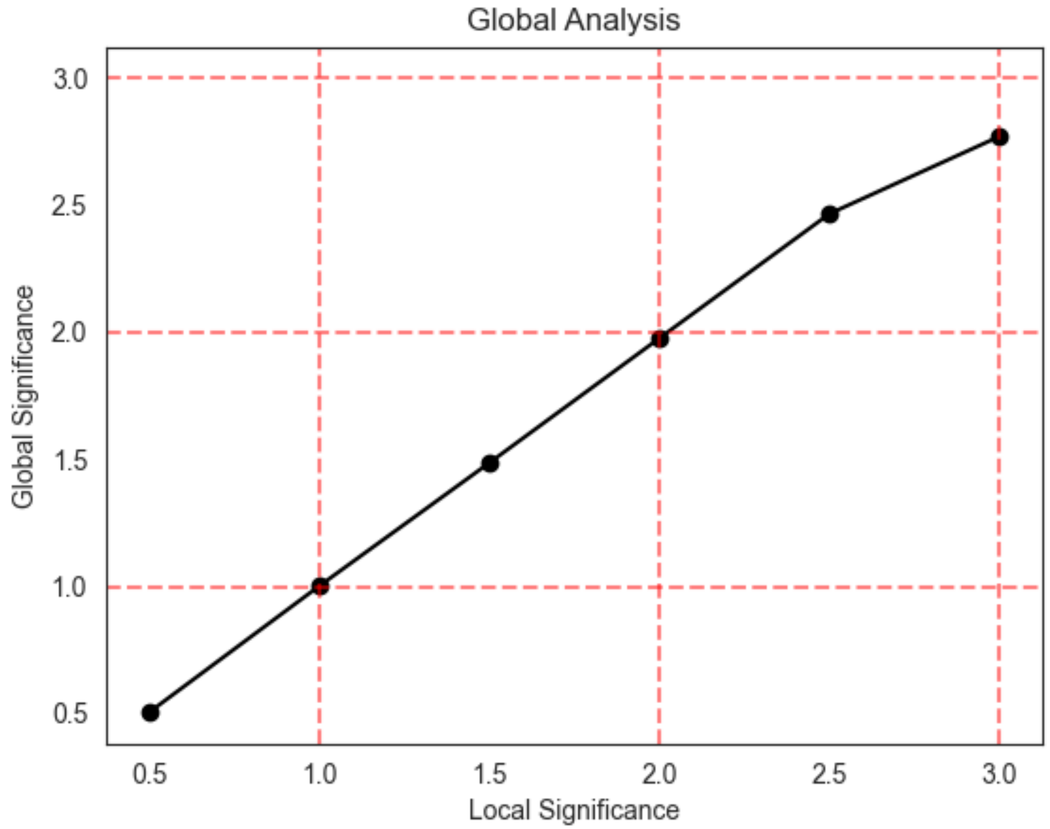




Frequentest Study Results - Global

Global significance results for 10,000 pseudo-experiments:





Summary

- **<u>Semi-Supervised DNN classifiers:</u>**

    1. Show good potential for use in BSM searches and can be used to **reduce model dependencies**.

    2. Extent of “fake” signal manifested in training can be evaluated using frequentest study

    3. Consider “**look elsewhere effect**” on significance measurements

- **<u>Frequentest Study Results:</u>**

    4. Using 10,000 pseudo-experiments it can be seen that the current results correspond to normal distribution.

    5. In order to complete the study (for 3-Sigma) more iterations are needed

Thank You

References

[1] von Buddenbrock S, Chakrabarty N, Cornell A S, Kar D, Kumar M, Mandal T, Mellado B, Mukhopadhyaya B and Reed R G 2015

[2] von Buddenbrock S, Chakrabarty N, Cornell A S, Kar D, Kumar M, Mandal T, Mellado B, Mukhopadhyaya B, Reed R G and Ruan X 2016 Eur. Phys. J.

[3] Crivellin A, Fang Y, Fischer O, Kumar A, Kumar M, Malwa E, Mellado B, Rapheeha N, Ruan X and Sha Q 2021

[4] von Buddenbrock S, Cornell A S, Fadol A, Kumar M, Mellado B and Ruan X 2018 J. Phys. G 45 115003

[5] Hernandez Y, Kumar M, Cornell A S, Dahbi S E, Fang Y, Lieberman B, Mellado B, Monnakgotla K, Ruan X and Xin S 2021 Eur. Phys. J. C 81 365

[6] Beck G, Kumar M, Malwa E, Mellado B and Temo R 2021 (Preprint 2102.10596)

[7] Sabatta D, Cornell A S, Goyal A, Kumar M, Mellado B and Ruan X 2020 Chin. Phys. C 44 063103

[8] Abi B et al. (Muon g-2) 2021 Phys. Rev. Lett. 126 141801 (Preprint 2104.03281)

[9] S.Battacharya, et al. arXiv:2306.17209

[10] A. Crivellin, et al. arXiv:2109.02650

Additional Slides

Invariant Mass Background Fits

Exponential

$$f(x) = c_0 * e^{-c_1 * x}$$

Quadratic Exponential

$$f(x) = c_0 * e^{-c_1 * x + c_2 * x^2}$$

Polynomial

$$f(x) = c_0 + c_1 * x + C_2 * x^2$$

Mixed Poly-Log-Exponential

$$f(x) = (1 - x)^{c_0} * x^{c_1 + c_2 * \log(x)}$$