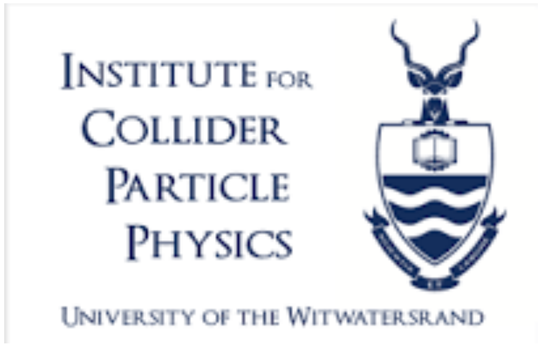
# Frequentist Approach to Quantify Fake Signals when using Semi-Supervised Machine Learning Classifiers

By **Benjamin Lieberman**

**Kruger 2022: Discovery Physics at the LHC**




# Overview

1. Introduction
2. Semi-Supervised DNN Classifier
3. Frequentest Study Methodology
4. Pseudo-Experiment Breakdown
5. Results
6. Data Generator/Sampler
7. WGAN as machine learning data generator

# Introduction

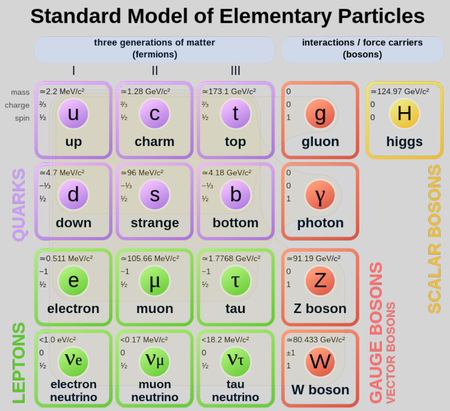
- SM is unable to explain various phenomena which explain substantial evidence, such as:

  ➡Dark Matter

  ➡The matter-anti-matter asymmetry

  ➡The origin of neutrino mass

- The search for new bosons is therefore motivated by these experimental discrepancies with the SM.




## Context of Presentation:

◉ Conduct BSM searches for Zγ resonances

◉ Use weakly or semi-supervised machine learning classifier.

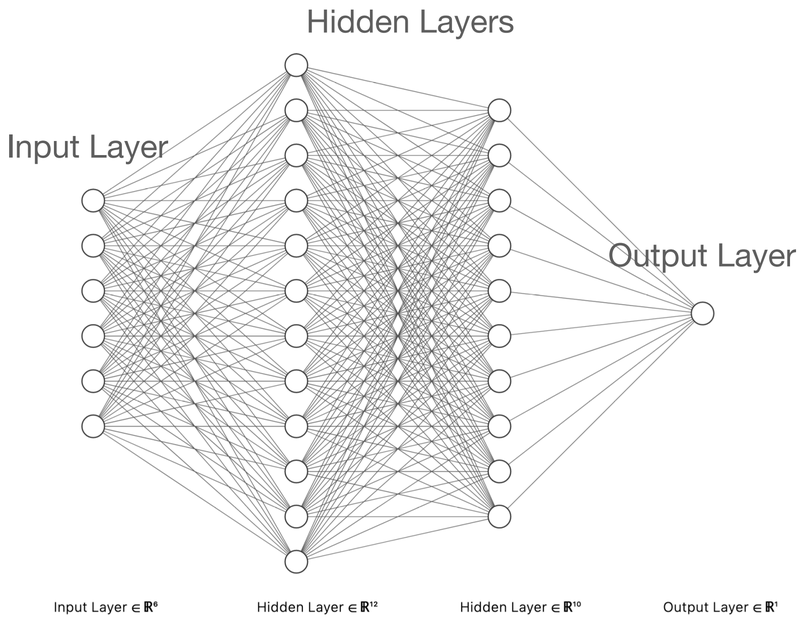
➡Reduce model dependencies

◉ Expose internal error generated by using semi-supervised machine classifiers

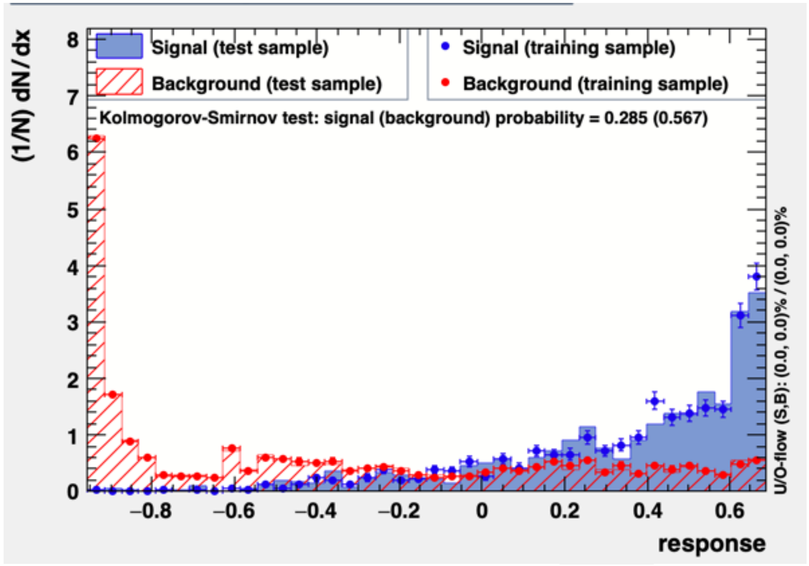
# Introduction to Machine Learning

## Classification and Anomaly Detection

### Neural Networks

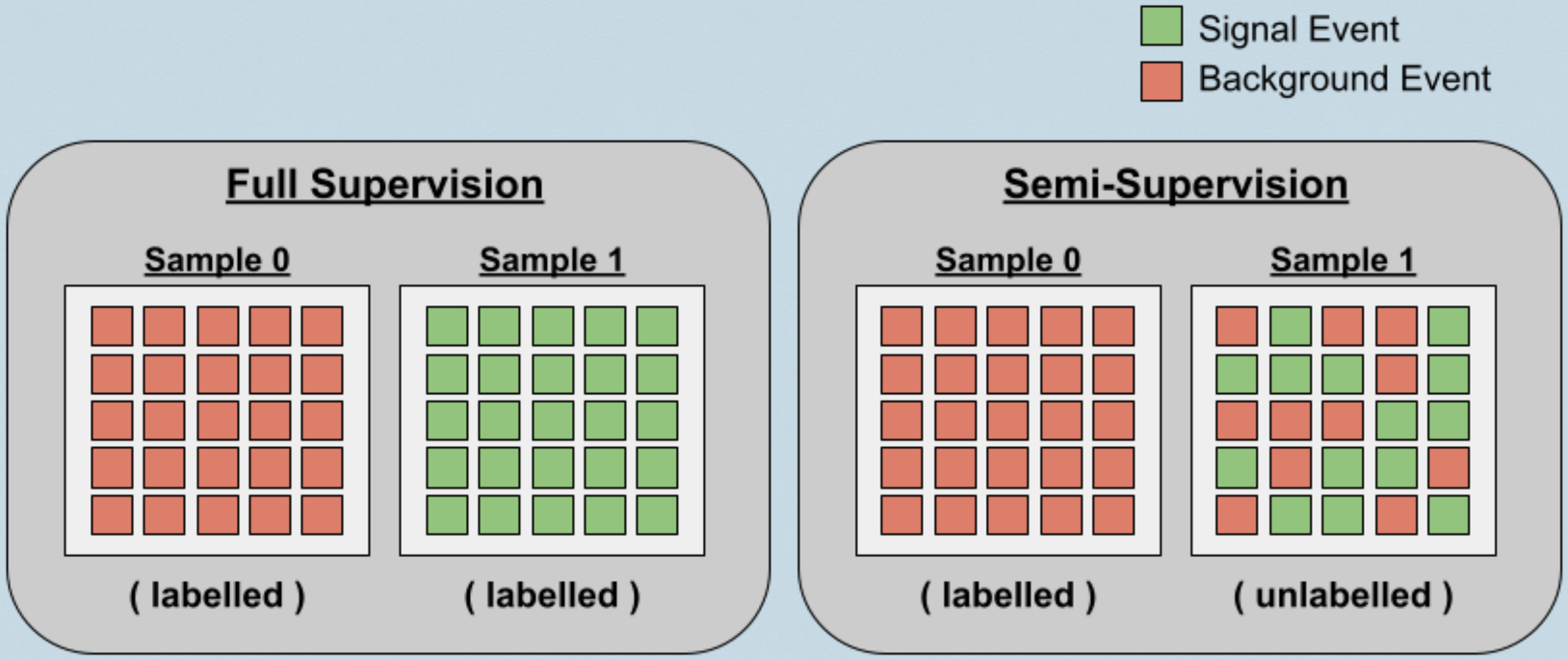



### Response Distributions



# Machine Learning Classifiers

## Machine Learning Semi-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?**

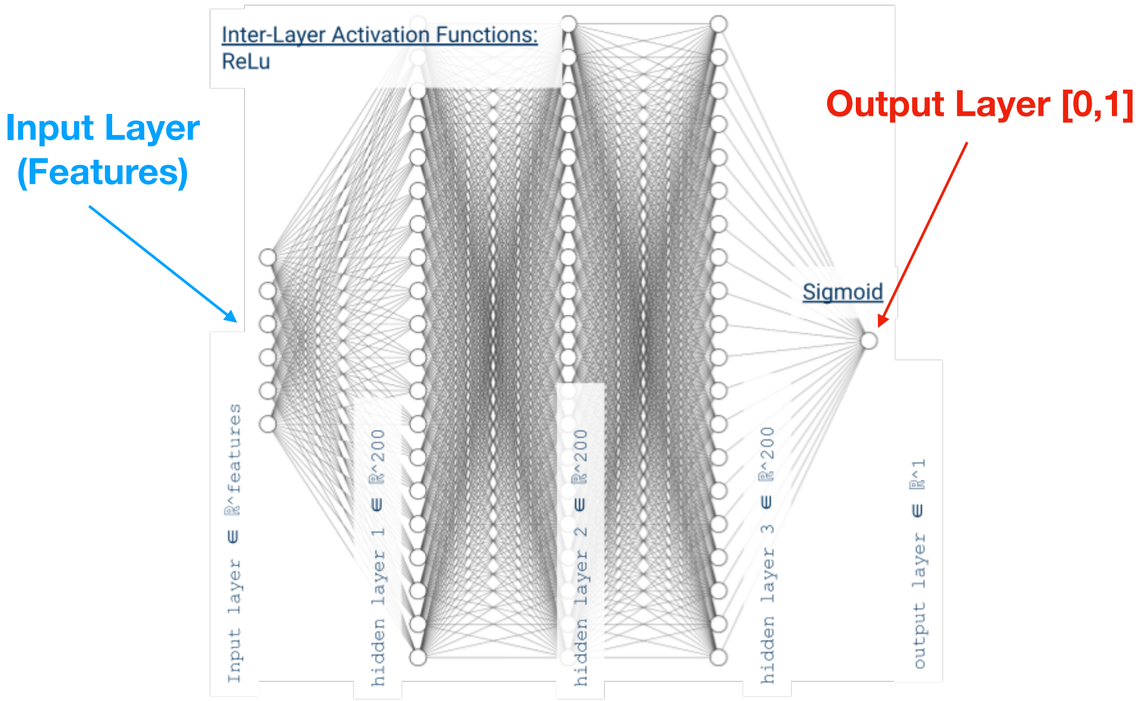
**Reduce Model dependencies**

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




**The optimised DNN hyper-parameters:**

Learning Rate = $1*10^{-3}$

Batch size = 256

Optimiser = Adam

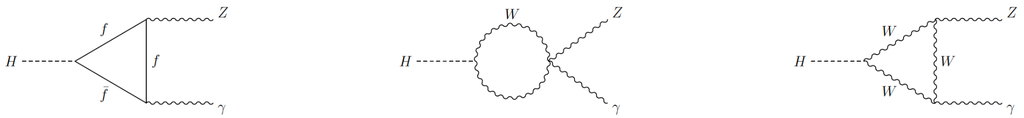
During the training of neural networks, over-training/over-fitting can cause background events to be incorrectly classified as signals.

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

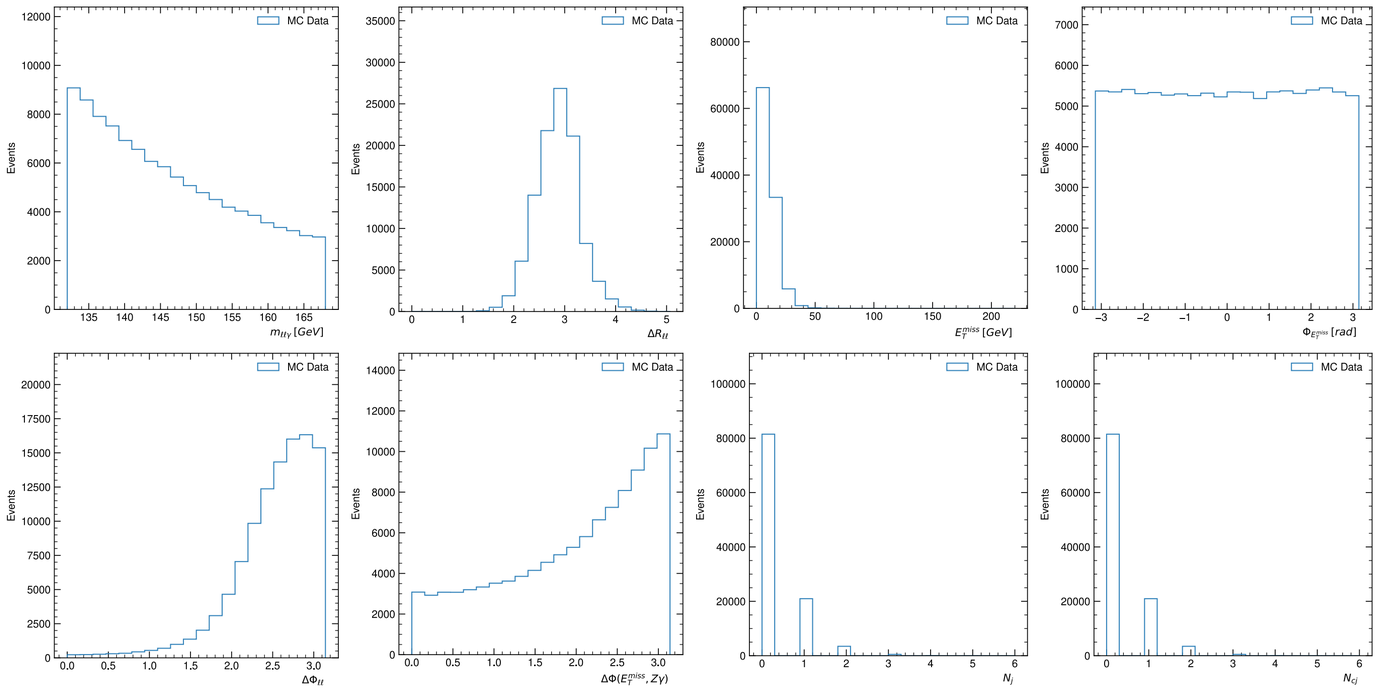
# Zγ Resonance Searches

- 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  Zγ final state.

## $pp \rightarrow H \rightarrow Z\gamma \rightarrow (\ell^+\ell^-)\gamma$

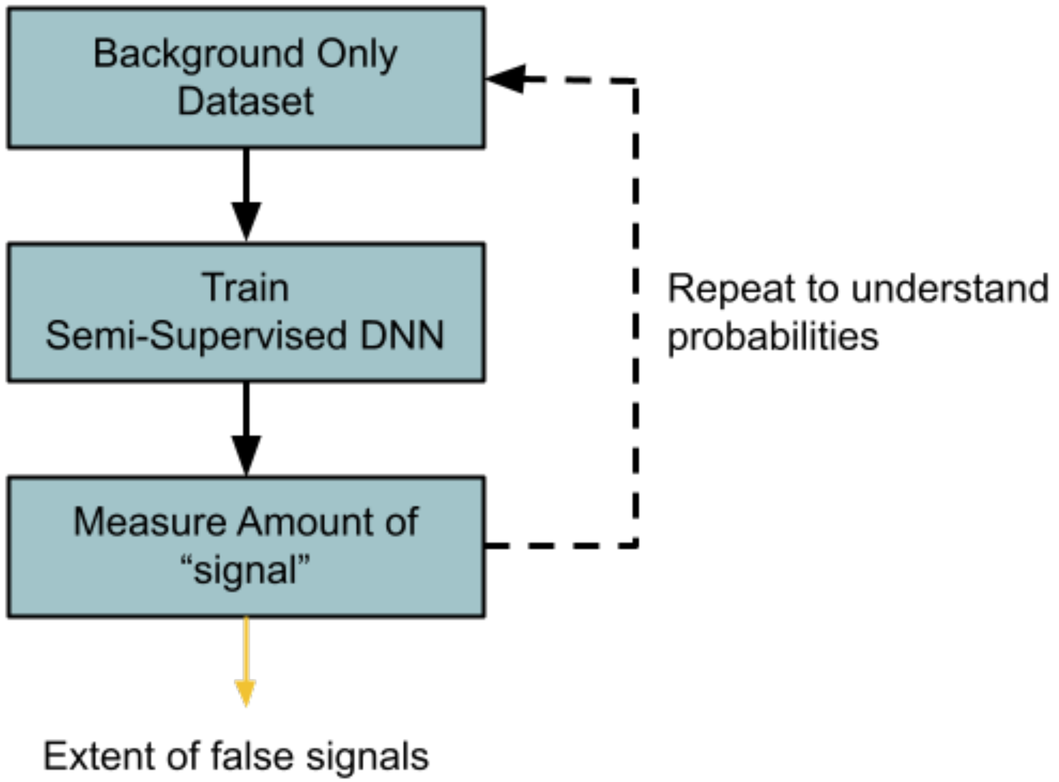



**Fast Simulation Monte Carlo Zγ Data generated:**
Using Madgraph5 with NNPDF3.0 parton distribution function. Parton level generation is done using Pythia and detector level simulation is done using Delphes(v3)



# Quantification of False Signals Generated in the Training of 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".

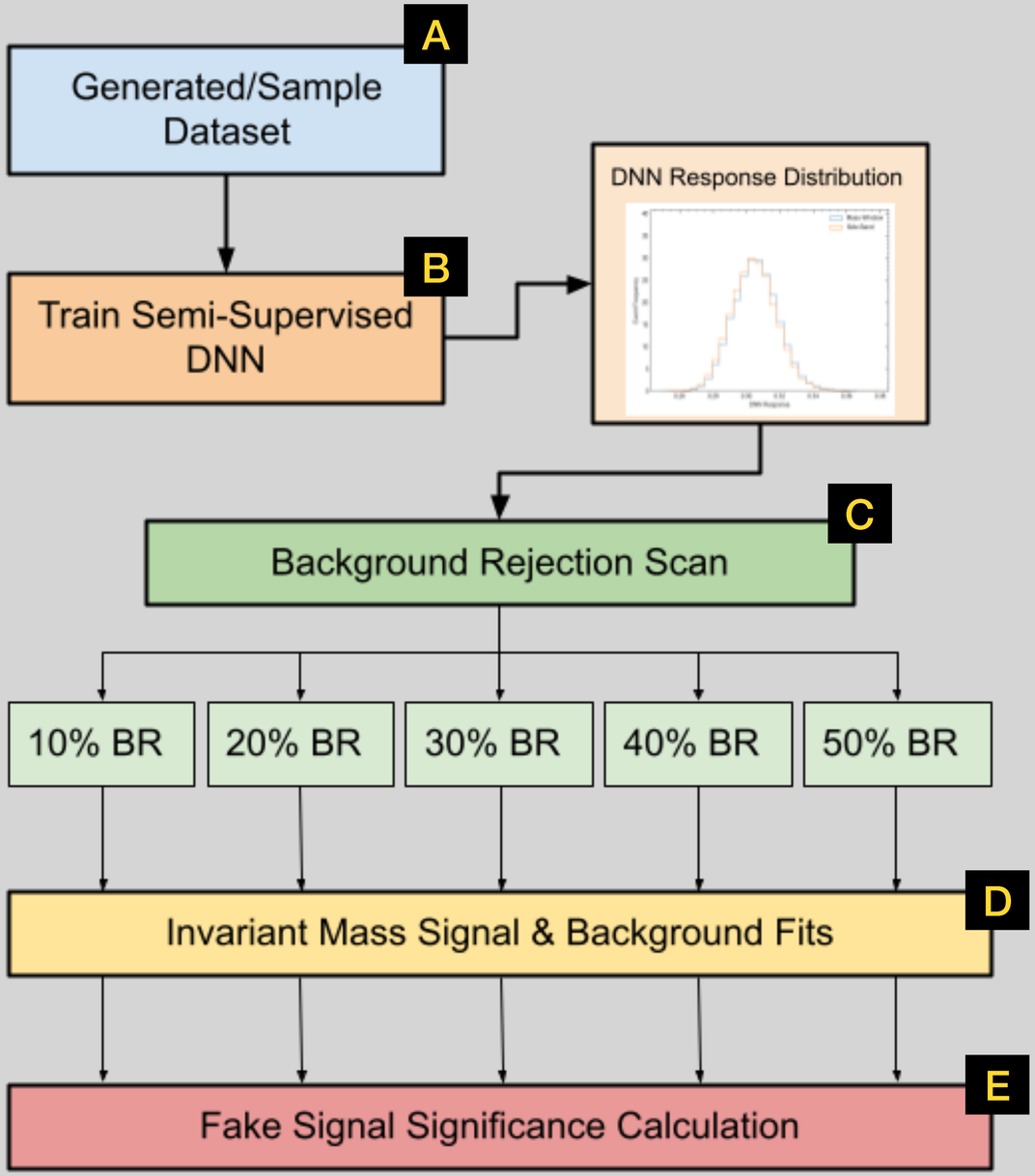
### Study Setup:

Fixed Mass:

- Center of mass = 150GeV
- Mass-window region = [144, 156] GeV
- 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

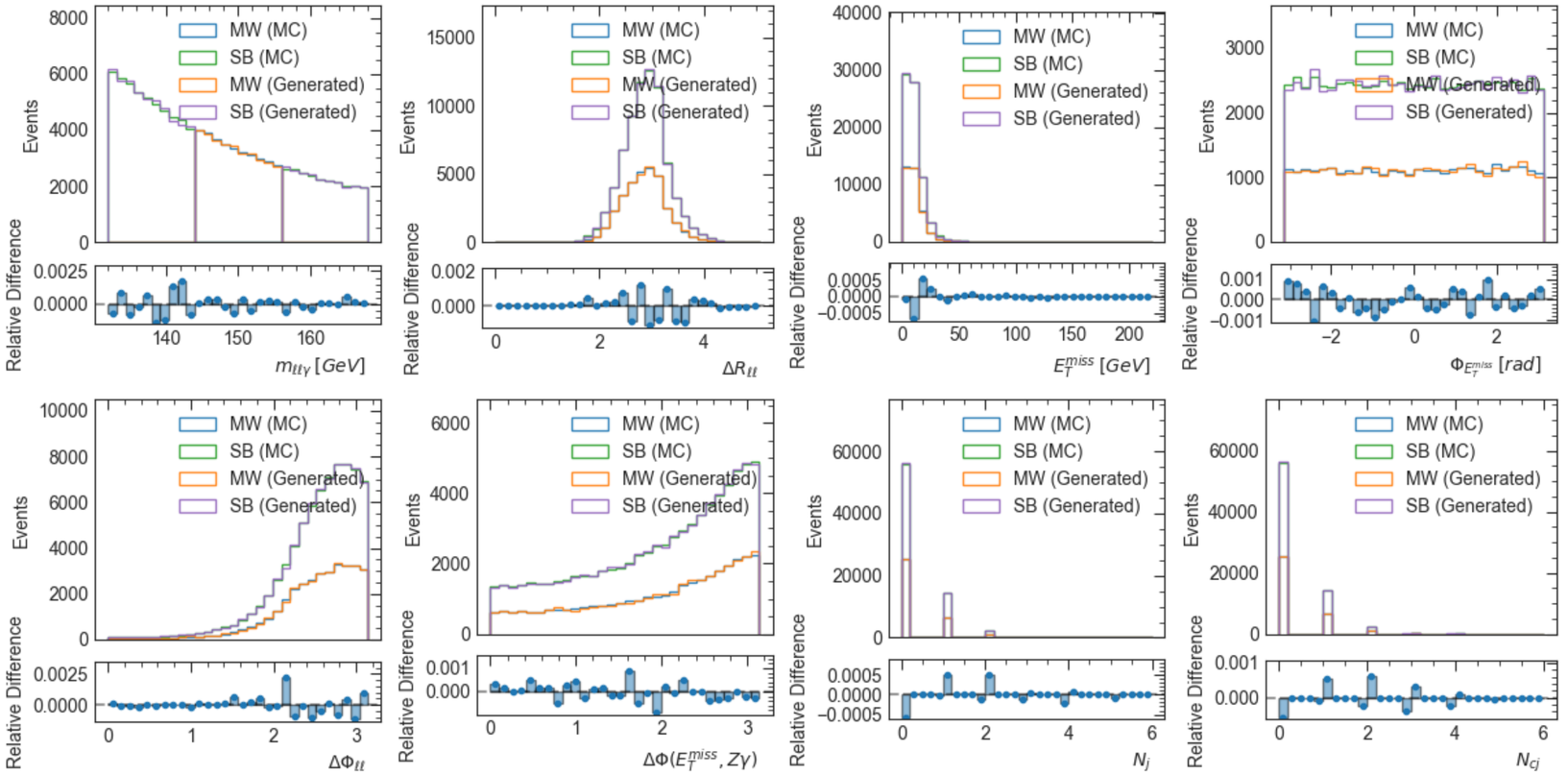
# Pseudo-Experiment: Data Sampling/ Generation

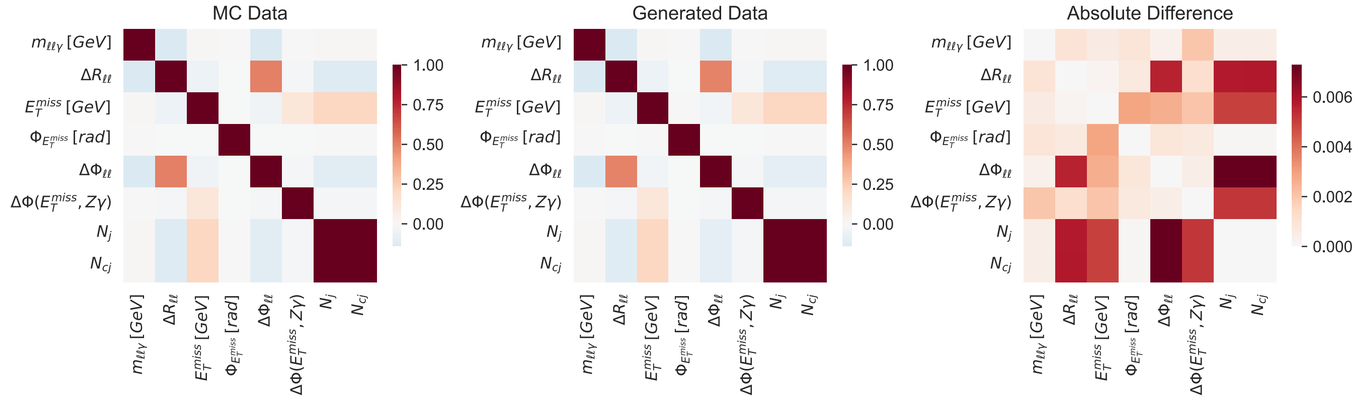
## Data Sampling using:

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

Excellent sampling method for synthesising events.

## Example of Generated training dataset using:



# B Pseudo-Experiment: DNN Training and Response Distribution

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

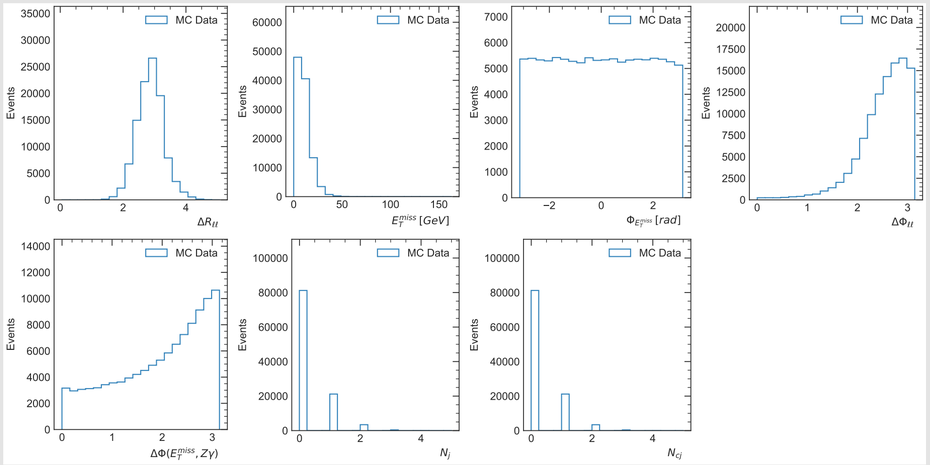
**Sample 0** (background / side-band region):

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

**Sample 1** (Signal / mass-window region):

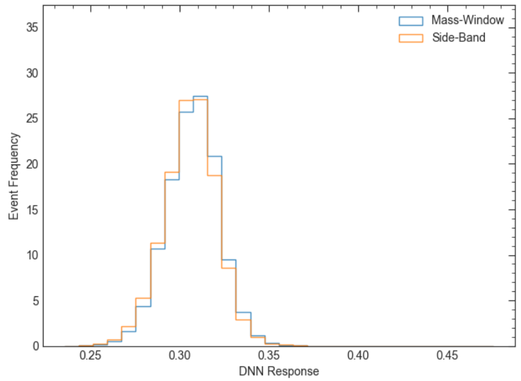
$(144 \leq m_{\ell\ell\gamma} \leq 156)$

## Zy Training Dataset Feature Distributions

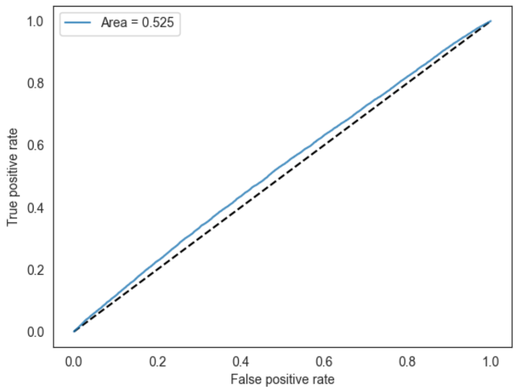



## DNN Outputs

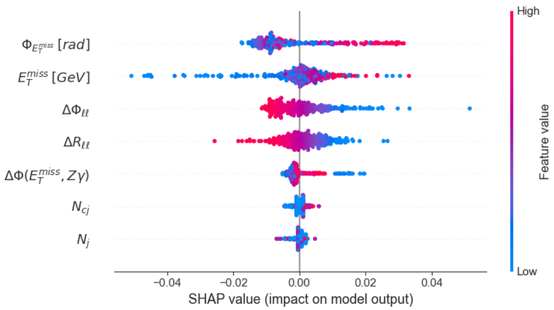
### Response Distribution




### ROC Curve

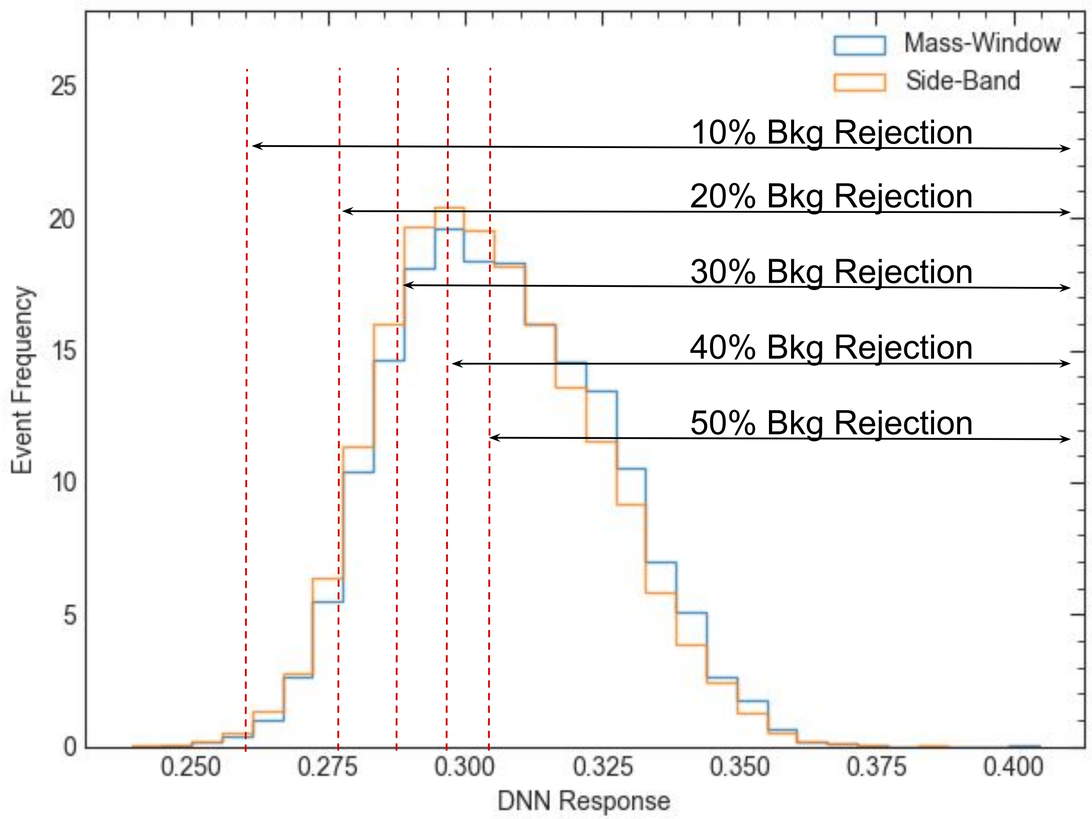



### SHAP Feature Ranking



# 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.
4. Each batches invariant mass distribution can therefore be used to extract a local significance



# D Pseudo-Experiment: Invariant Mass Background Fits

## Fitting

1. Data (Background) is fit with exponential, f(x)

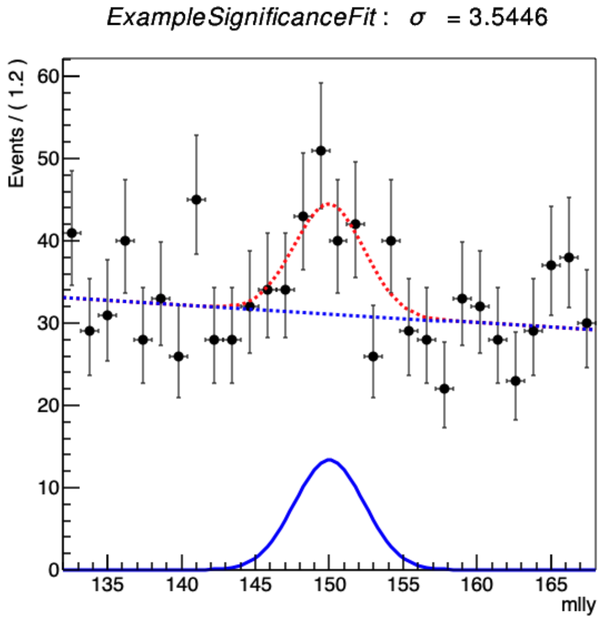
2. Background + Signal is fit with exponential + gaussian, g(x)
   - Exponential component of g(x) uses fixed parameters from f(x)
   - Mean, μ, is the centre of mass = 150GeV
   - Sigma. σ, is the resolution = 2.4

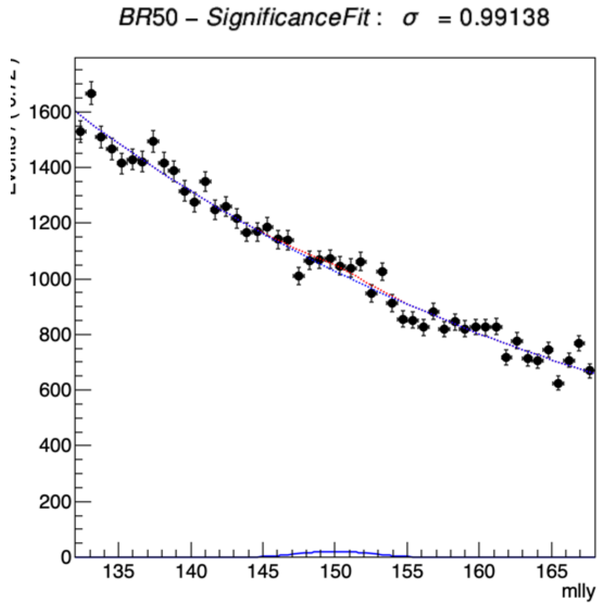
**Side-Band (Background) Fit:**

$$f(x) = n_0 * e^{ax+bx^2}$$

**Mass-Window (Signal) Fit:**

$$g(x) = n_0 * e^{ax+bx^2} + n_1 * e^{\frac{(x-\mu)^2}{2\sigma}},$$



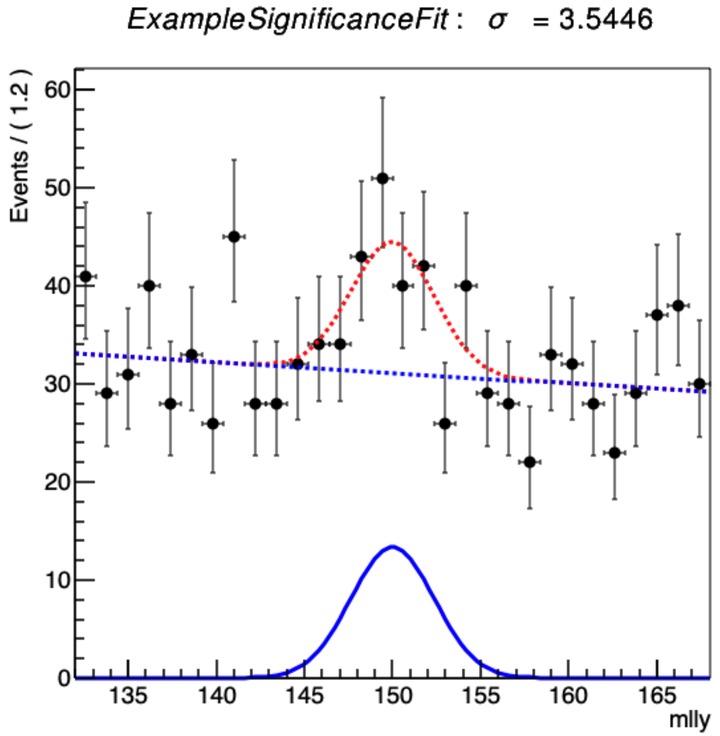
# Pseudo-Experiment: Fake Signal Significance Calculation

- Using Signal and background fits, the signal significance is calculated using **ROOT**, CERN's library designed for particle physics analysis.
- The **AsymptoticCalculator** used, performs hypothesis tests using the asymptotic formula for the **profile likelihood**, and uses the **Asimov data set** to compute expected significances or limits.

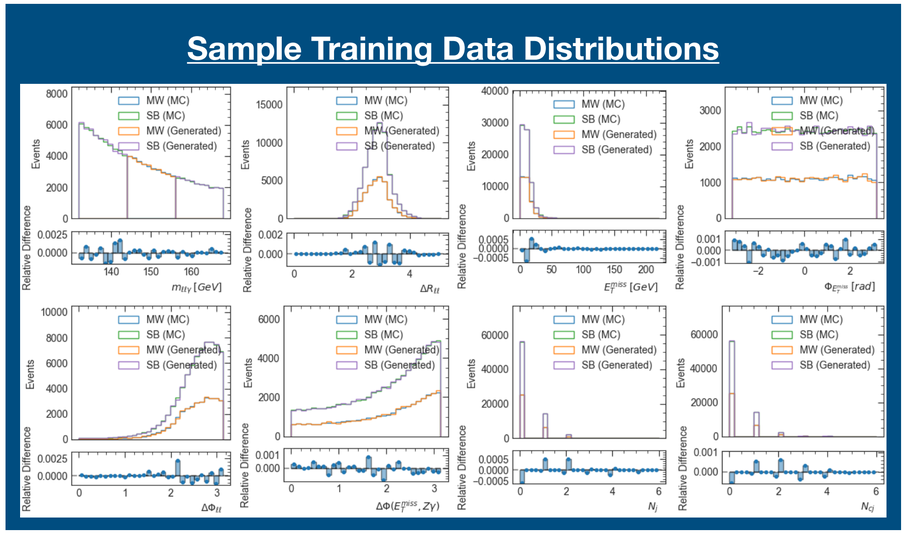
**Parameter of interest**: number of signal events

**Observable**: Invariant Mass, mlly

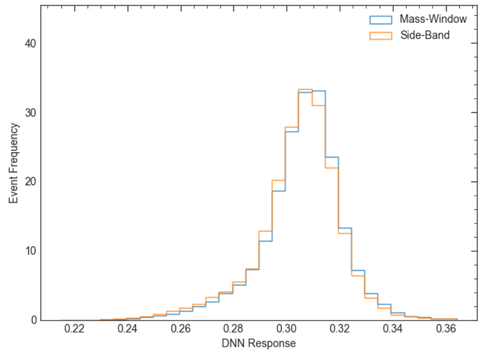
**Null (background) Hypothesis**: no signal events will be found in signal region

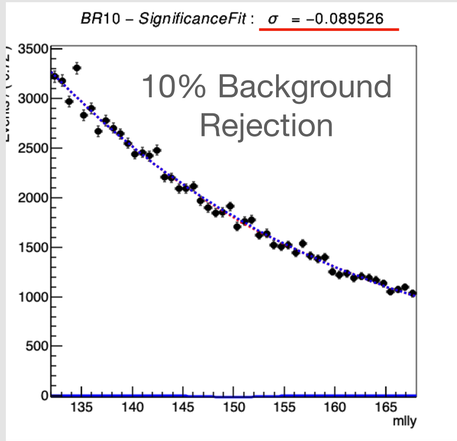


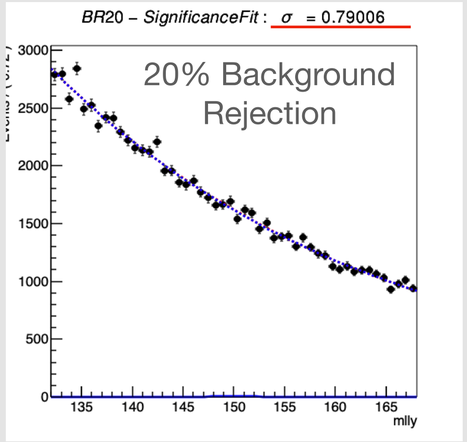
# Example Pseudo-Experiment

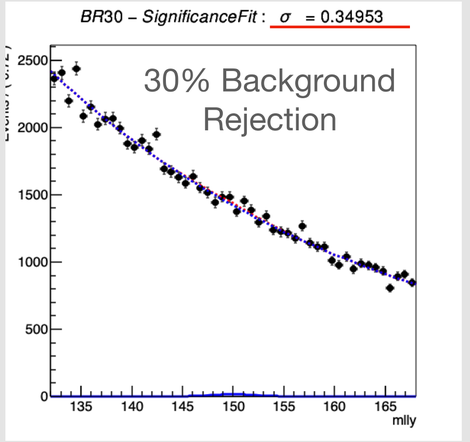


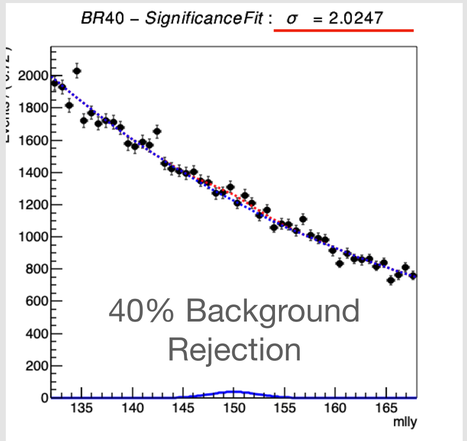

Train Semi-Supervised DNN

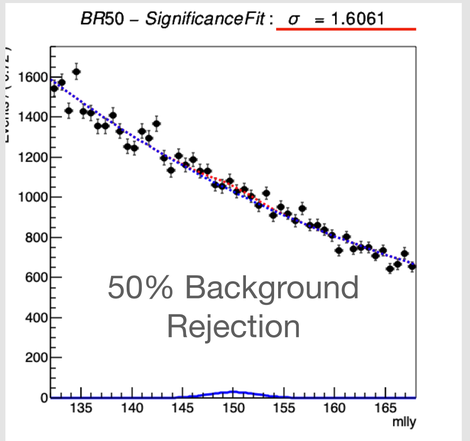


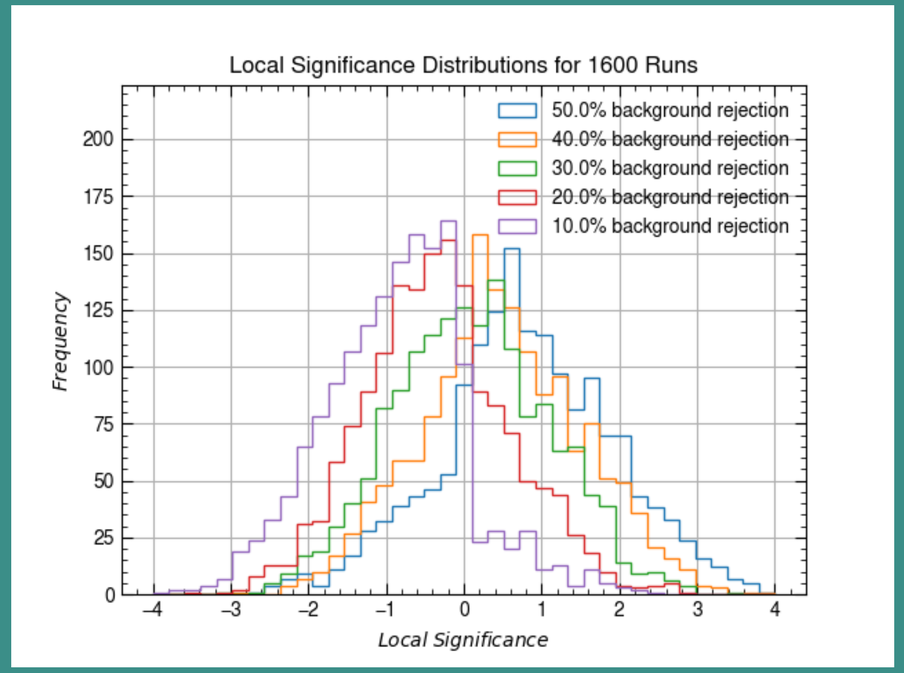



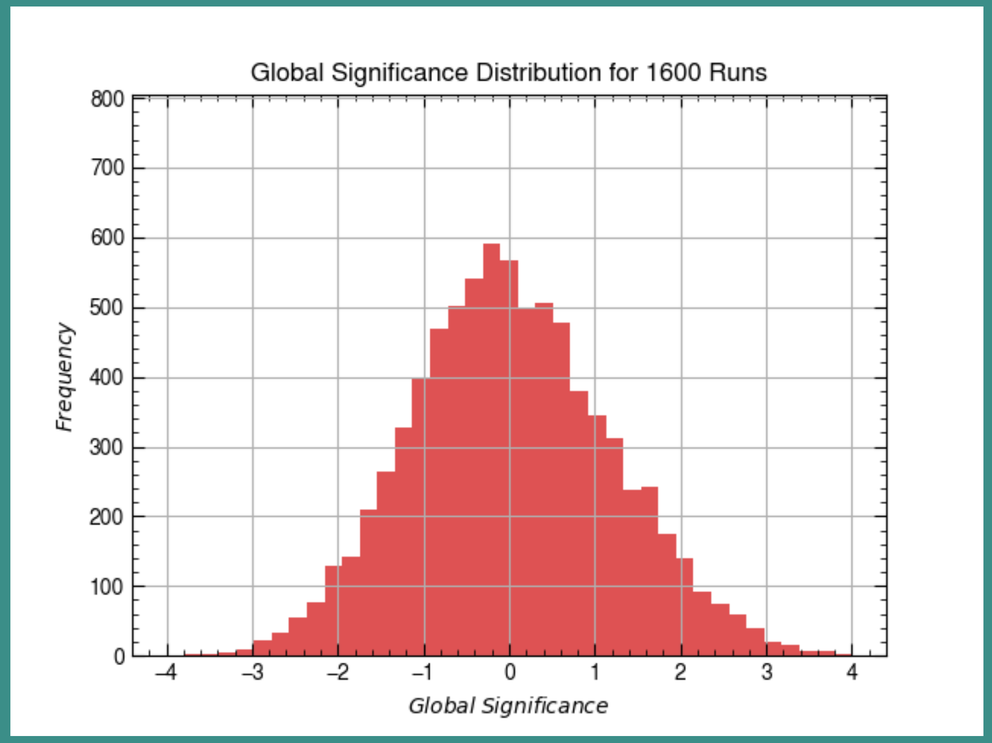



# Frequentest Study Initial Results




| | 1σ | 2σ | 2.5σ | 3σ | 3.5σ |
|---|---|---|---|---|---|
| **Local Runs** | 3154 | 721 | 267 | 71 | 15 |
| **Pvalue [%]** | 39.51 | 9.03 | 3.34 | 0.89 | 0.19 |

# Pseudo-Experiment: Data Sampling/Generation

## Problem:

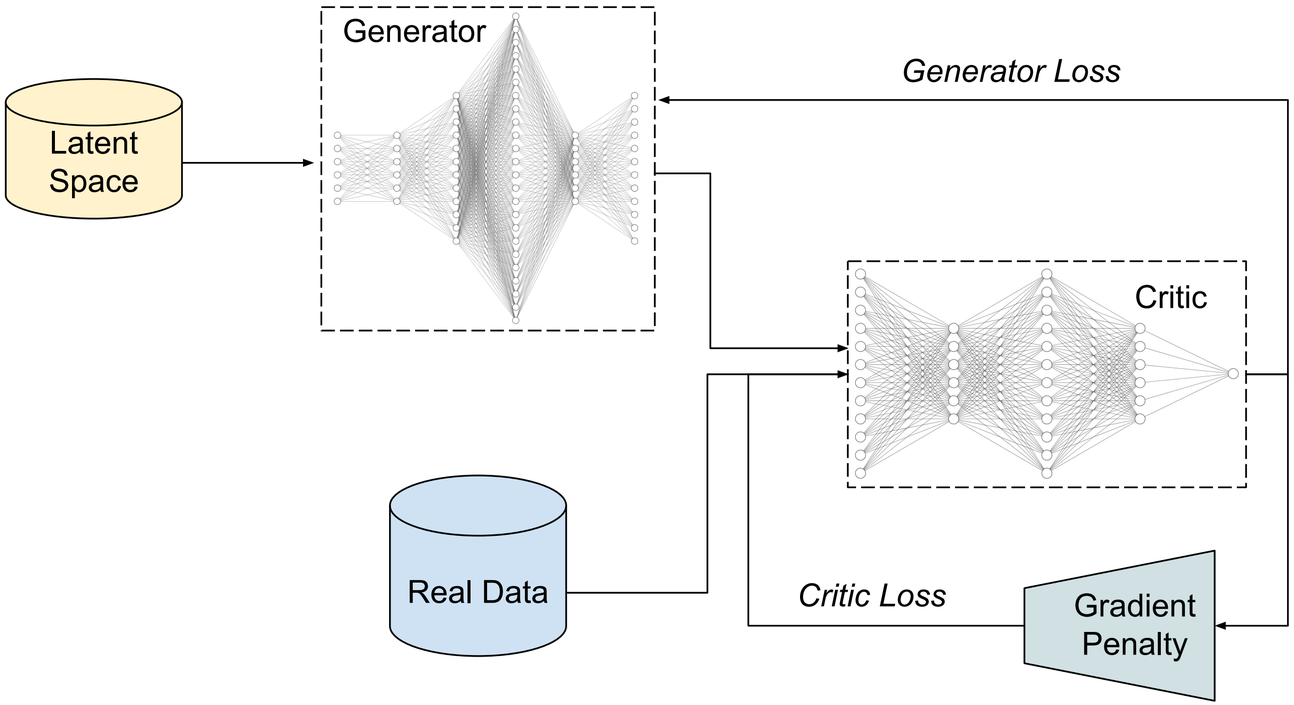
- For each pseudo-experiment a **statistically independent dataset** is needed.
- **±200,000 events** is ideal for the training and evaluation of the DNN for each pseudo-experiment.
- In order to complete the frequentest study, the pseudo-experiment must be run more than **50,000** times.
- Monte Carlo event generation of sufficient events is computationally excessive and will take take excessive time

Therefore the study requires **±10x10$^9$ events**

## Solutions:

1. **Event Sampling**: Bootstrap or other event sampling methods can enable batches of events to be sampled while maintaining statistics.
2. **Machine Learning Generators**: GANs, VAE and other machine learning data generators can learn to generate statistically accurate events at scale

# Data Generator: Wasserstein Generative Adversarial Network



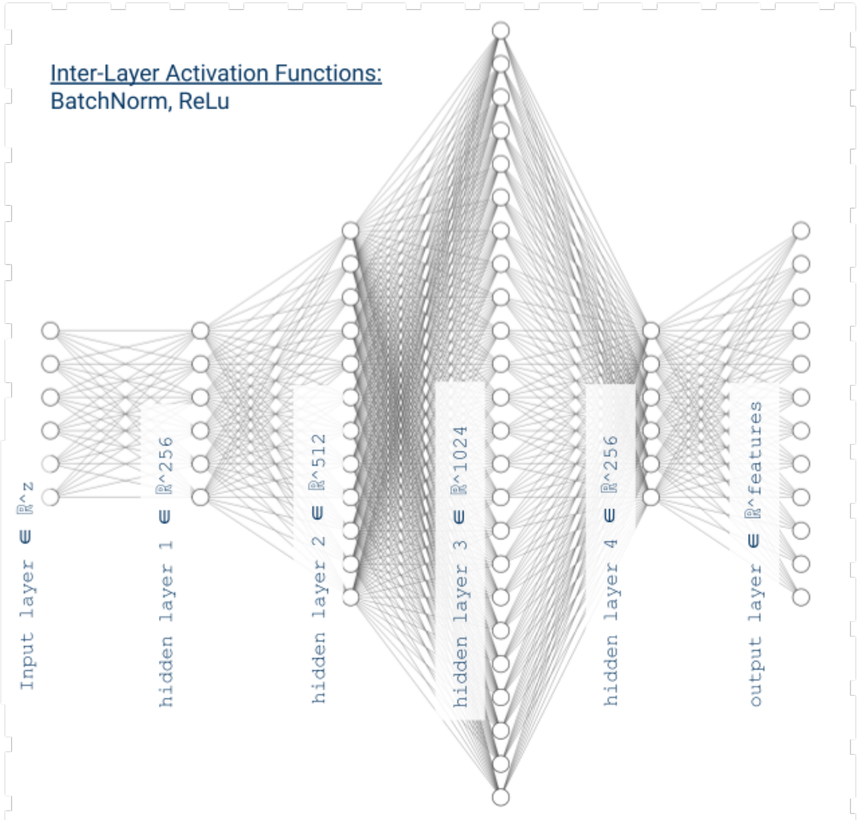

$$L_g = \min \mathop{\mathbb{E}}_{\tilde{x} \sim \mathbb{P}_g} [C(\tilde{x})]$$

$$L_c = \max \mathop{\mathbb{E}}_{x \sim \mathbb{P}_r} [C(x)] - \mathop{\mathbb{E}}_{\tilde{x} \sim \mathbb{P}_g} [C(\tilde{x})] + \lambda * GP$$

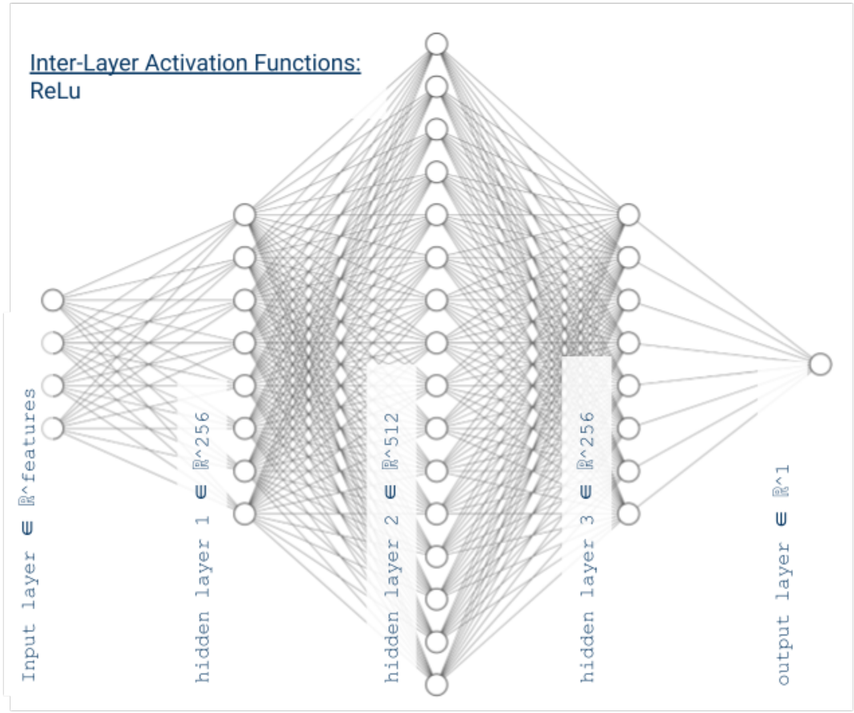
$$GP = \mathop{\mathbb{E}}_{\tilde{x} \sim \mathbb{P}_{\tilde{x}}} [(||\nabla_{\tilde{x}} C(\tilde{x})||_2 - 1)^2]$$

# Data Generator: Wasserstein Generative Adversarial Network

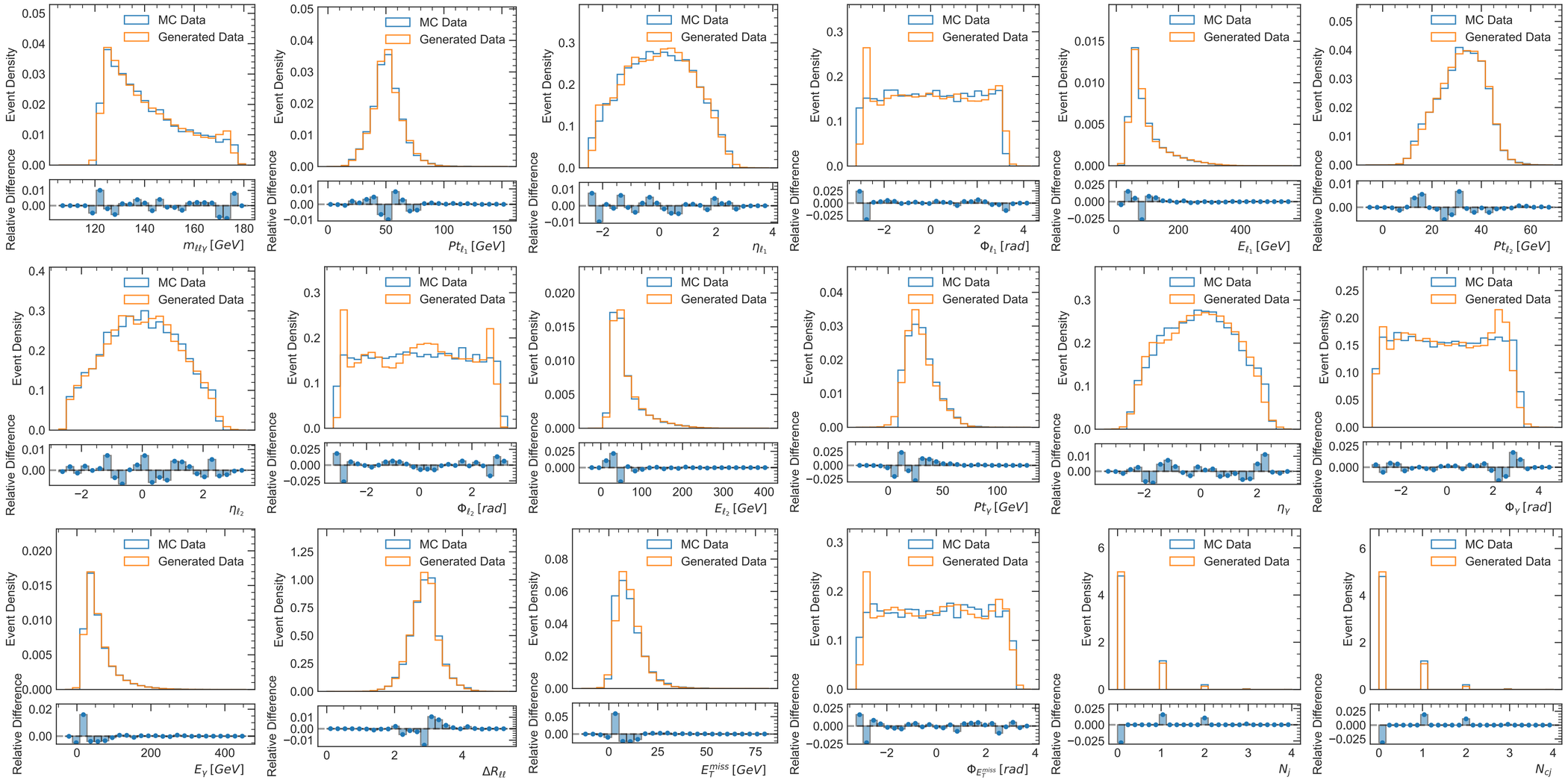
## Generator Model



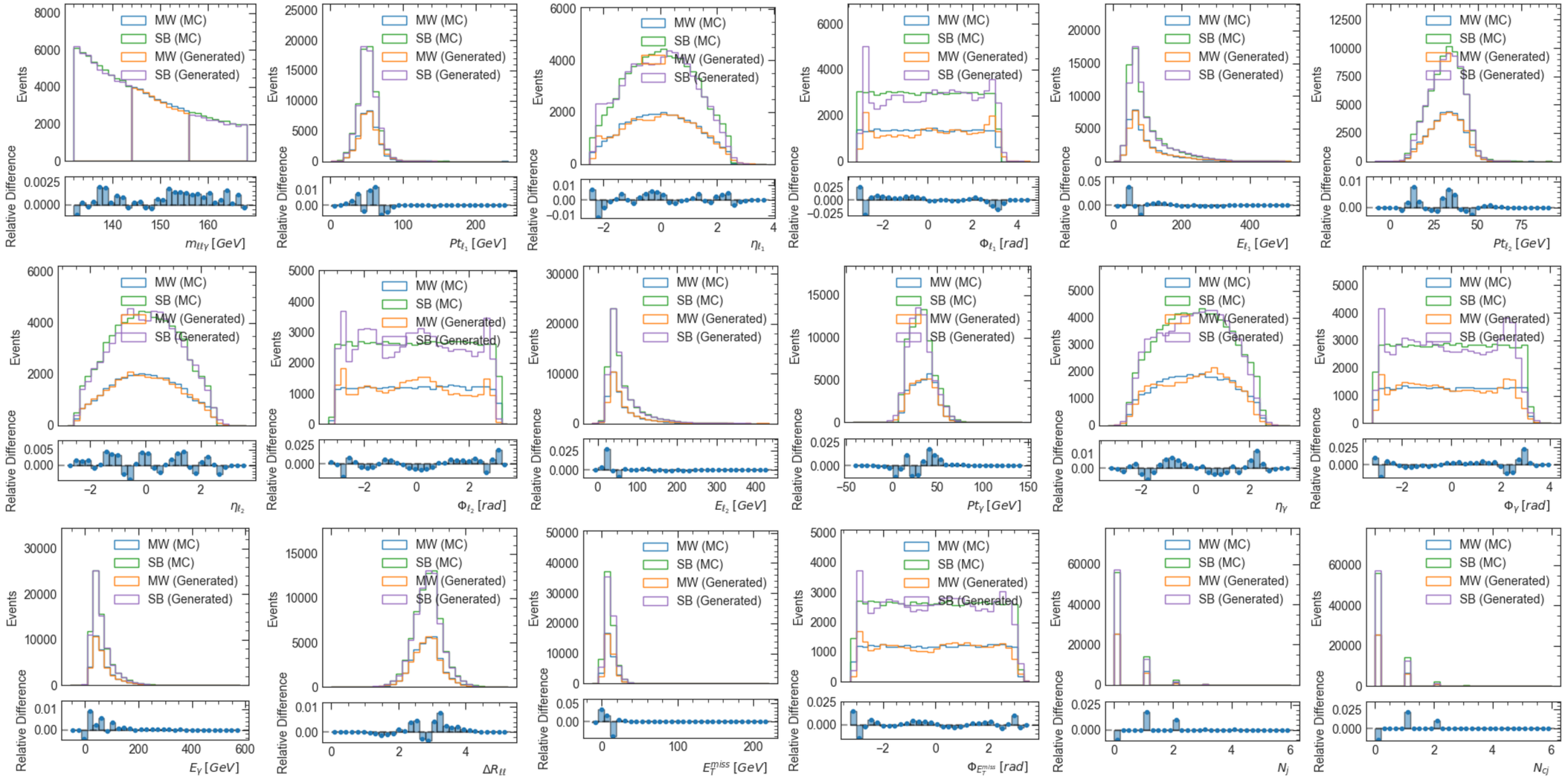

## Critic Model



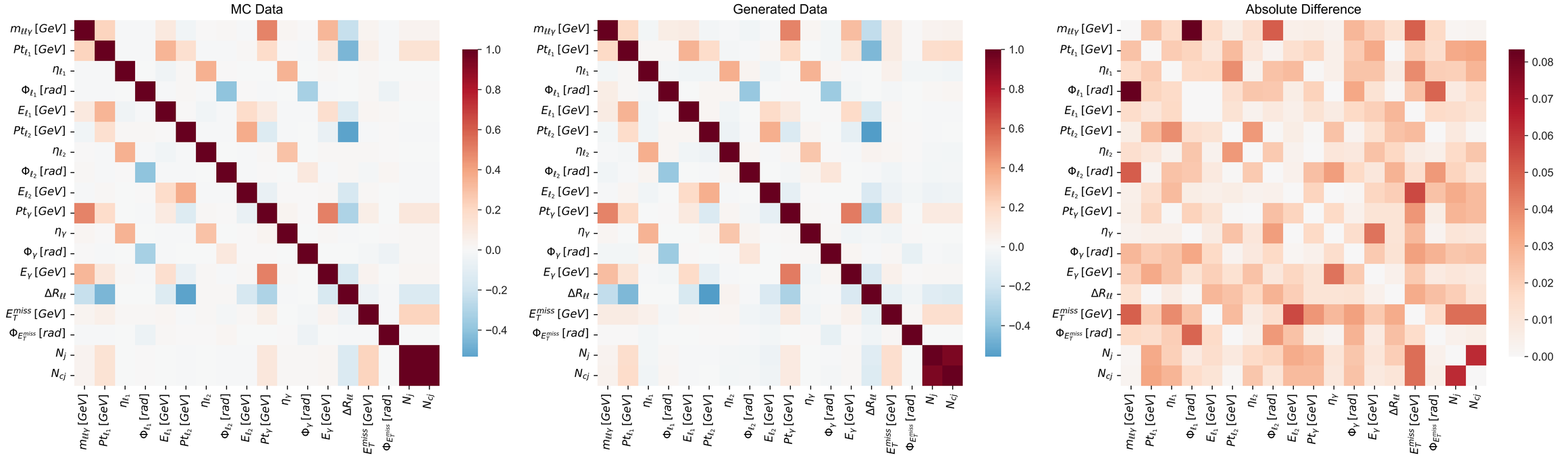
# Data Generator: Wasserstein Generative Adversarial Network
# Generated Data Feature Distributions



# Data Generator: Wasserstein Generative Adversarial Network
# Generated Data Feature Distributions



# Data Generator: Wasserstein Generative Adversarial Network
# Generated Data Feature Correlation



# 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)