2022/03/23

University of the Witwatersrand

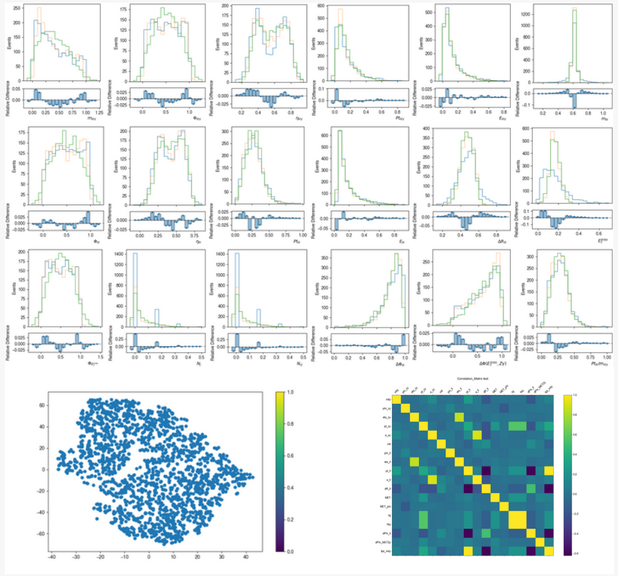
Supervisor

Prof Bruce Mellado

STUDENT

Finn Stevenson

The Use of a Variational Autoencoder in the Search for Resonances at the LHC



UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

First Pan-African Astro-Particle and Collider Physics Workshop

Table of Contents

		Page
I	Research Background & Motivation	3
II	Hypotheses	6
III	Methodology	9
IV	Research Results	10
V	Conclusion & Future Work	12

II Research Background & Motivation

The Standard Model (SM) of particle physics was completed by the discovery of the Higgs boson in 2012 by the ATLAS and CMS collaborations. However, the SM is not able to explain a number of phenomena and anomalies in the data. These discrepancies to the SM motivate the search for new bosons. In this reseacrh, searches for new bosons are completed by looking for Zgamma resonances in Zγ (pp →H →Zγ) fast simulation events.

This research is concentrated on two main processes related to the use of machine learning in the search:

- Part 1: Machine learning based Data Generation to complement traditional MC production mechanisms.
- Part 2: Machine learning based signal classification

II Research Background & Motivation

Data Generation Phase 1

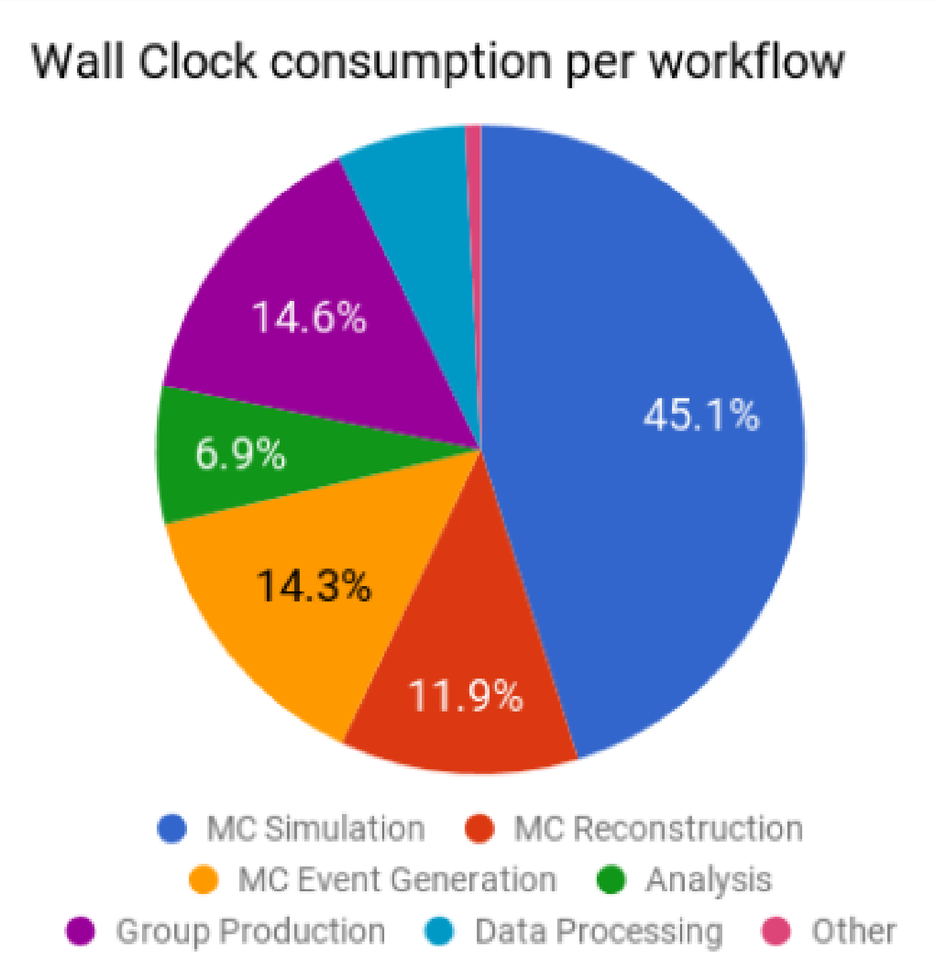
Data generation for particle physics data is a crucial process in the search for new bosons at the LHC.

MC simulation at the LHC -> large amount of the total CPU hours of the ATLAS experiment (see graph right).

Using trained deep learning models to produce valid MC fast simulation data instead of MC production mechanisms -> Can help alleviate the some of the required CPU hours and free up for other tasks

The luminosity of the detectors at the LHC is increasing continuously, this will increase the need for MC simulated data or deep learning-based generated MC simulated data.

In this research, Variational Autoencoders VAEs are assessed as a deep learning-based event production mechanism.



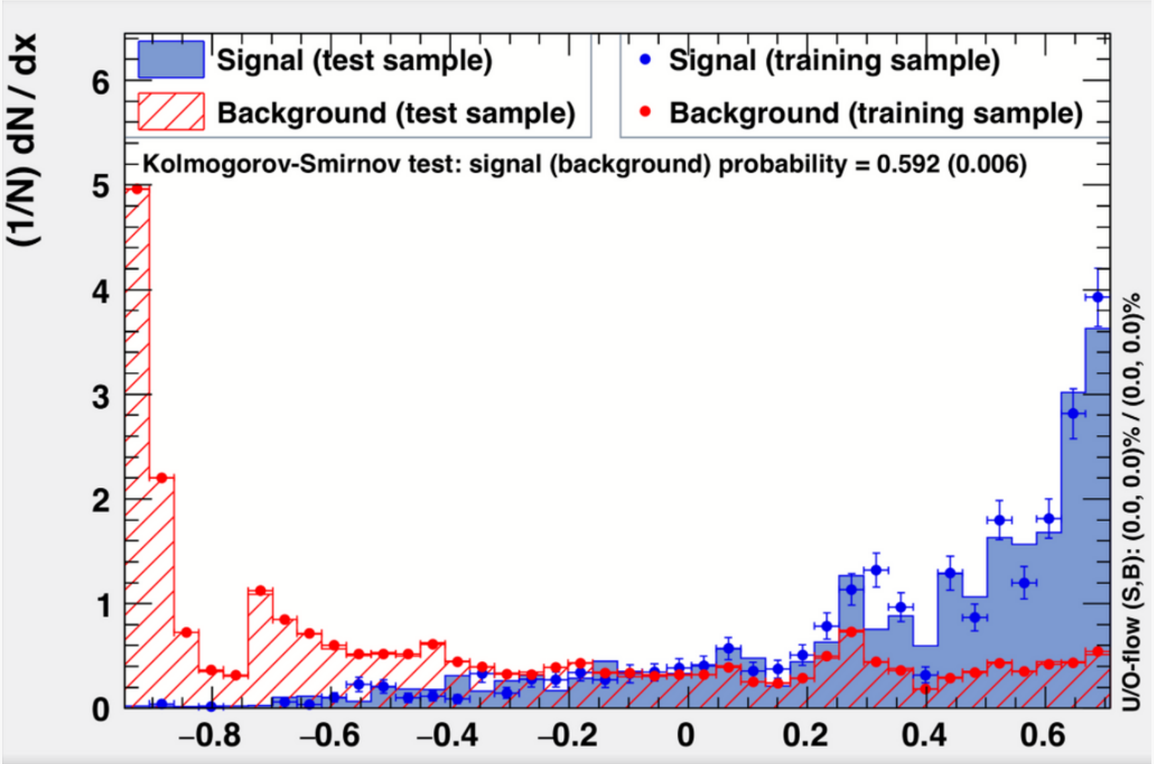
II Research Background & Motivation

Signal Classification Phase 2

Signal classification -> crucial process at the LHC in the search for new bosons. Machine learning techniques are great at this, especially with extremely large datasets!

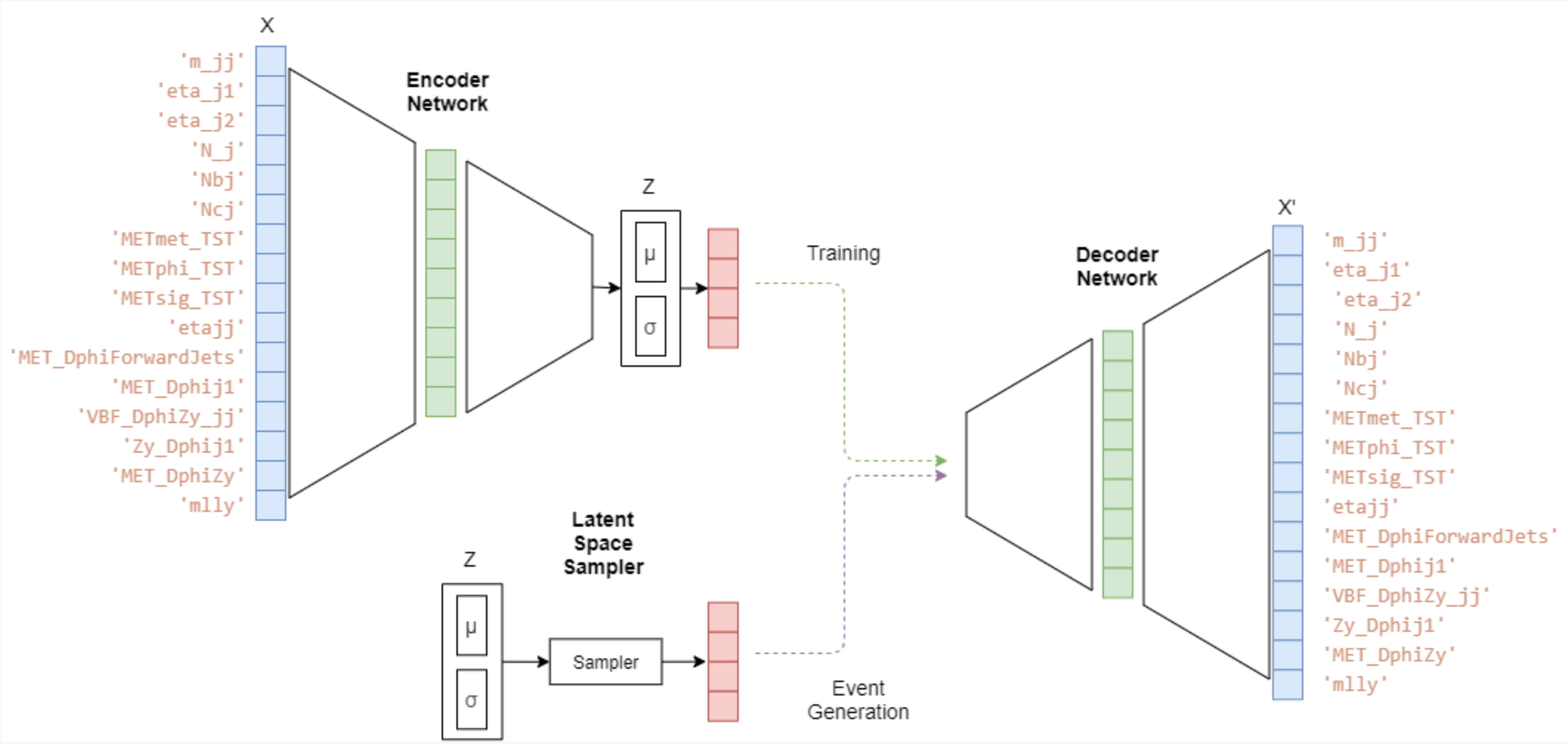
A VAE can be trained as both a **Signal Classification** and a **Data Generation** model in one.

VAEs can be used to search for resonances in data by training on a specific selection of background events, and then during testing background samples with some injected signal events cant be fed through the VAE and **the reconstruction loss metric can be used to identify whether there are any signal events present in the test sample.**



I Hypotheses

A well trained VAE can be used in the search for new boson at the LHC simultaneously for data generation and signal classification.



Dataset Exploration & Kinematic Variable Selection

Kinematic variables selected from the Zγ (pp →H →Zγ) MC generated dataset where used to test the hypothesis.

Model Developement

A Variational Autoencoder base model was selected and developed as the base model. VAEs are tried and tested generative and classification arhcitectures.

Optimisation

The base model requires extensive hyper-parameter optimisation to achieve results that are favourable fo the search.

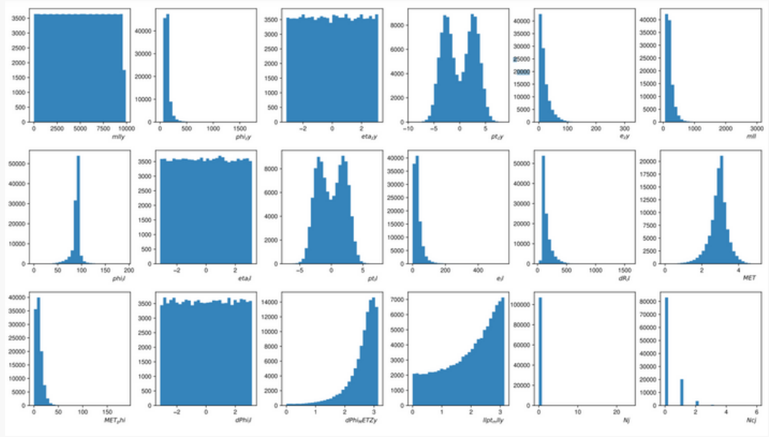
Evaluation

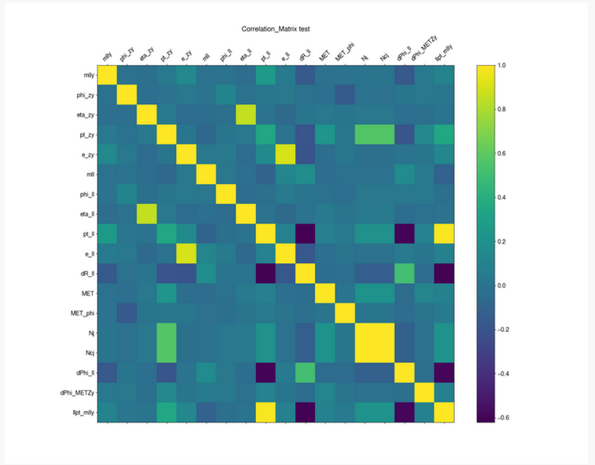
A number of evaluation metrics have been selected to evaluate the models generative and classification capabilities.

V Methodology

Dataset Description & Kinematic Variable Selection

The MC events corresponding to the Zy sample in this analysis have been generated using Madgraph5 with the NNPDF3.0 parton distribution functions. Here we have utilized the Standard Model (SM) of particle physics for which the UFO model files required by the Madgraph is obtained from FeynRules. The parton level generation if followed by the parton showering and hadronization by Pythia and then the detector level simulation is performed using Delphes(v3). The jets at this level has been constructed using Fastjet which involves the anti-KT jet algorithm with PT > 20 GeV and radius R = 0.5. While generating the sample we decayed the Z boson to leptons. We also have applied some baseline cuts on the leptons and photons at the Madgraph level to enhance the statistics.





Kinematic variable selection:

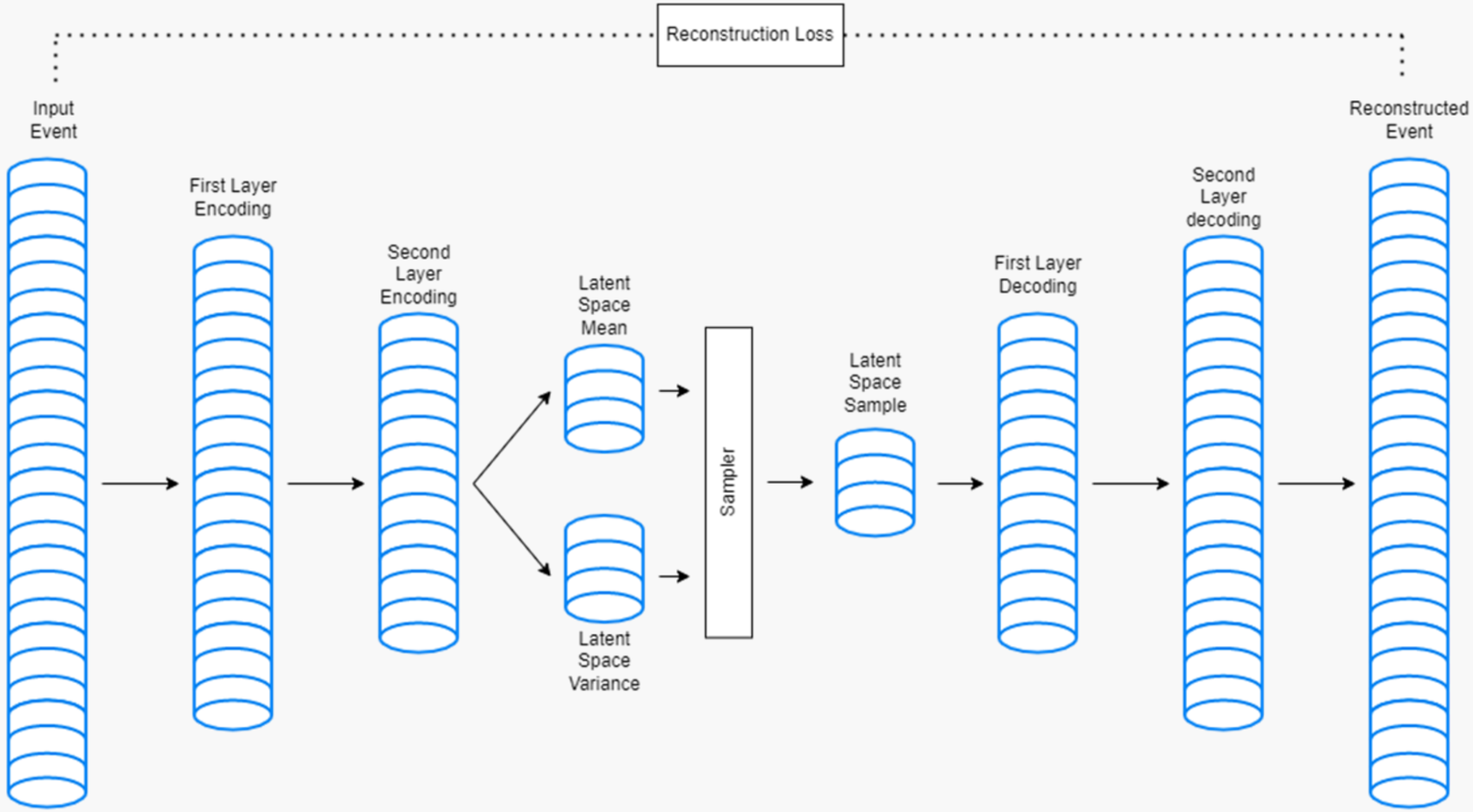
['mlly', 'phi_zy', 'eta_zy', 'pt_zy', 'e_zy', 'mll', 'phi_ll', 'eta_ll', 'pt_ll', 'e_ll', 'dR_ll', 'MET', 'MET_phi', 'Nj', 'Ncj', 'dPhi_ll', 'dPhi_METZy', 'llpt_mlly']

V Methodology

Model Development

During training, the VAE learns to reconstruct the training events that are fed forward through the network. Weights are learnt using backpropagation and optimisation of the loss function.

The VAE loss function consists of 2 terms.
-The reconstruction loss (which can be seen in the diagram)
-The KL divergence



VAE loss function:

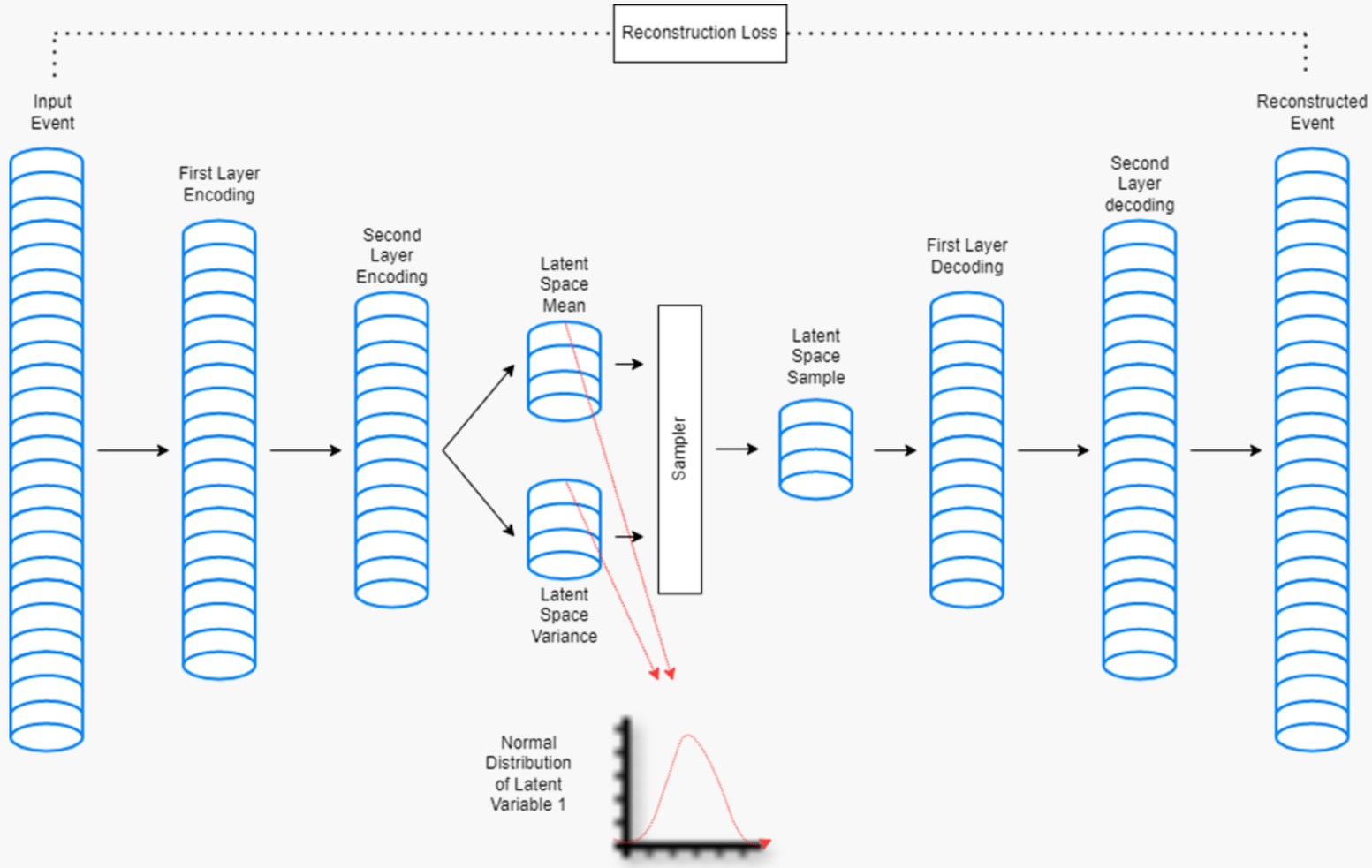
$$VAE_{loss} = reconstruction_{loss} + variational_{beta} * KLdivergence_{loss}$$

V Methodology

Model Development

Latent space variables (mean and variance) for each latent variable normal distribution are learnt through variational inference (as a result of the addition of the KL divergence to the loss function).

These learned latent space distributions can then be used to generate new events.



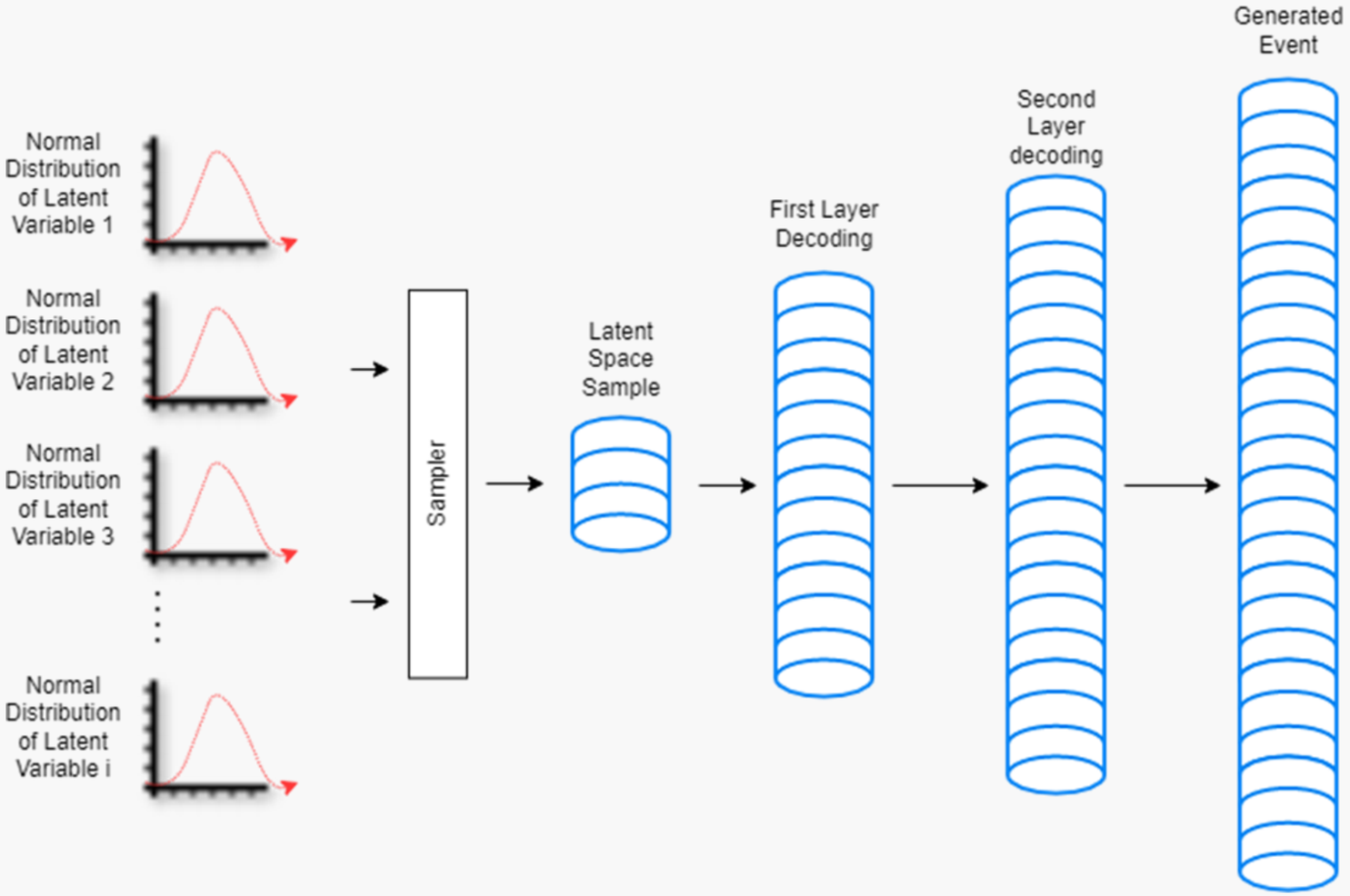
$$VAE_{loss} = reconstruction_{loss} + variational_{beta} * KLdivergence_{loss}$$

V Methodology

Model Development

The diagram on the right shows how the latent space learned distributions can be sampled from and the sample can be fed forward through the decoder in order to generate a new event.

This can be done using random sampling from the latent space to generate many events.



$$VAE_{loss} = reconstruction_{loss} + variational_{beta} * KLdivergence_{loss}$$

Model Development

A look at some of the important components of the VAE architecture and loss function.

The loss function:

$$VAE_{loss} = reconstruction_{loss} + variational_{beta} * KLdivergence_{loss}$$

The VAE loss function is made up of two terms. The reconstruction loss is responsible for minimising the difference between an input event and a reconstructed event, during training.
The KL divergence loss term is a regularization term and is responsible for generating a latent space that is appropriate for generative purposes. Variational inference is used to create normal distributions for each latent space variable.

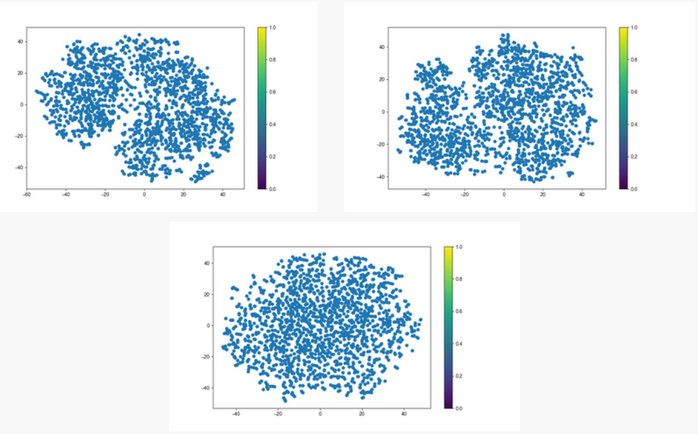
Variational beta:

The Variational-beta parameter in the loss function is used to weight the importance of the KL-divergence loss term against that of the reconstruction loss term.

This is useful to optimise for different batch_sizes architectural complexity and learning rates.

The latent space:

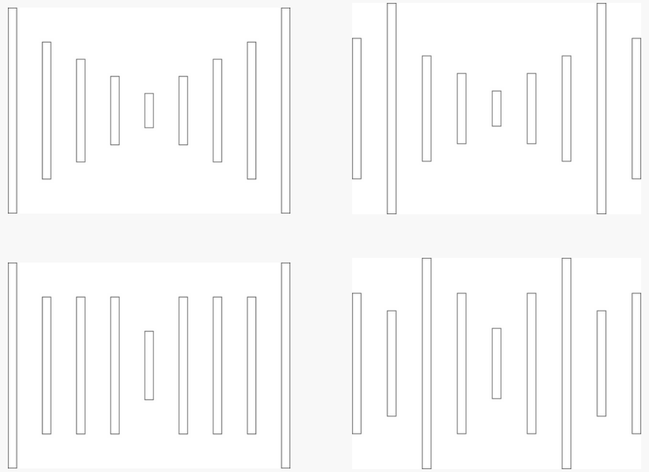
The latent space of a Variational Autoencoder is made up of a normal distribution for each latent space variable. Visualisations of this Latent space can be done using PCA and tSNE.



Architectural Considerations

Different architecture shapes need to be assessed during hyper-parameter optimisation to see which one is more appropriate for event generation.

The most traditional shape can be seen in the top left. Whilst the least frequently used shape can be seen bottom right.



Model Development - Hyperparameter Optimisation

Hyper-parameters can be optimised in order to optimise generation accuracy. The following hyper-parameters were selected for hyper-parameter optimisation:

-Training batch_size [real]:

-Asyemetry [truth]:

-Learning rate [float]:

-Activation function [str]:

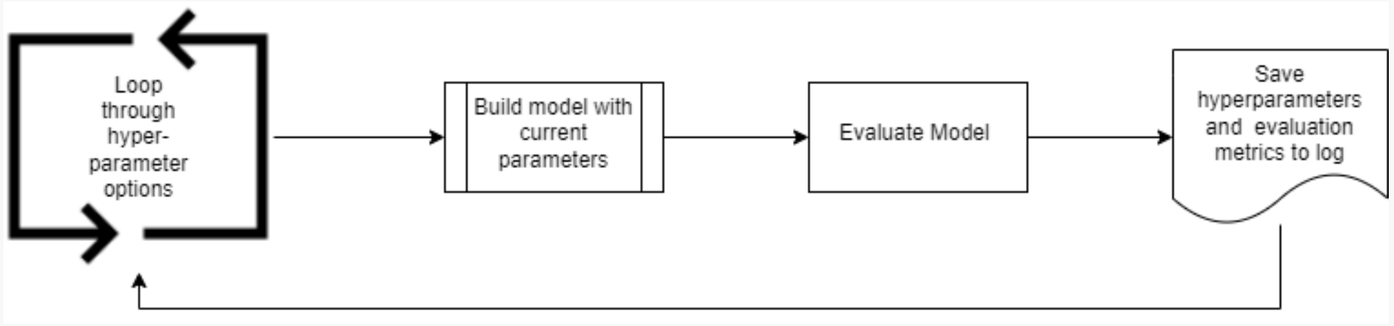
-Latent dimension size [real]:

-Number of hidden layers [real]:

-Number of nodes in hidden layer 1:

-Number of nodes in hidden layer 2:

-Number of nodes in hidden layer 3:



Evaluation Metrics

Evaluation metrics are required in both the training of the VAE and for the final evaluation of the generated results.

During Training Evaluation Metrics:

-Simply the loss and the loss components:

- reconstruction loss
- KL divergence
- VAE loss

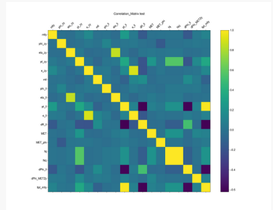
Data Generation Evaluation Metrics:

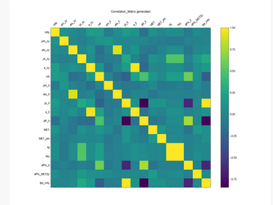
- KL Divergence (between input events data set and generated)
- Pearsons correlation coefficient

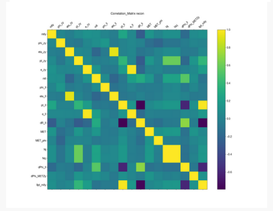
Plot based evaluation:

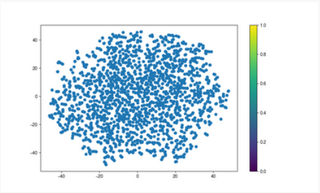
- Distribution Plots
- Correlation Plots
- Latent Space tSNE plots

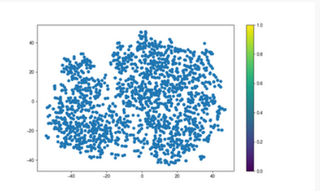
V Results

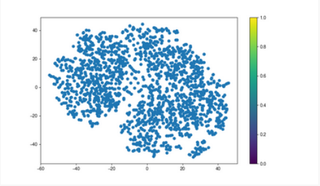


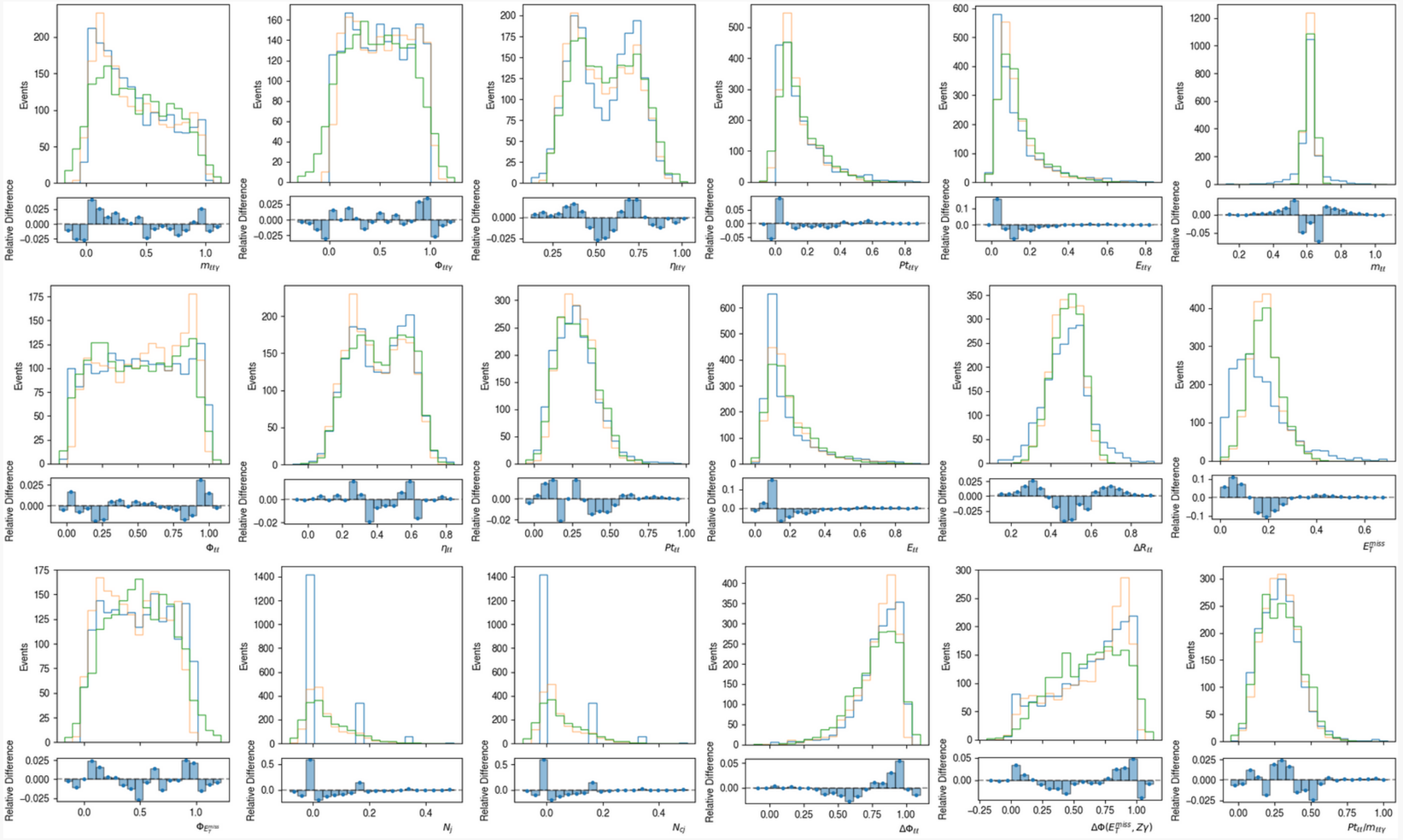












Preliminary results for phase 1 of the research (Data Generation) show promise. Further evaluation will be completed to obtain even better results before continuing to phase 2.

I Conclusion

The VAE architecture is a valuable resource in the search for new bosons at the LHC. This research has produced decent results for the first phase (data generation).

Future Work:

In the second phase of the research, the classification capabilities will be assessed whilst still assuring sufficient data generation capability during optimisation of classification.