LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS
partículas e tecnologia

Autoencoders as a tool for unsupervised searches at the LHC.

Jhonathan Barrios | 9th September
Advisors:
Professor Nuno Castro and Miguel Romão, PhD
Summer LIP Internship Program 2022
University of Minho

Outline

Machine Learning: key concepts

Dataset

Autoencoders

1.
Machine Learning:

Key Concepts

> *Machine Learning is the science (and art) of programming computers, so they can learn from data [1].*

What is Machine Learning?

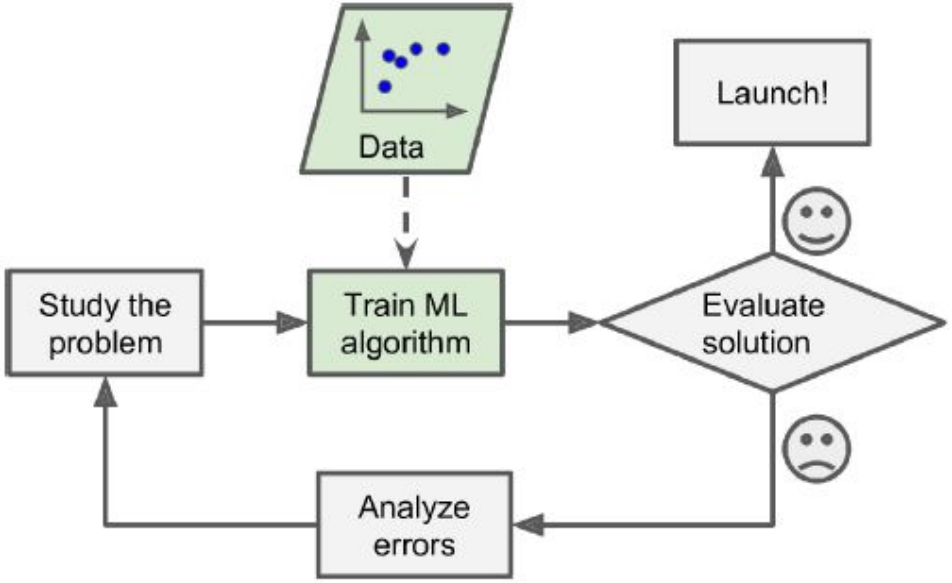
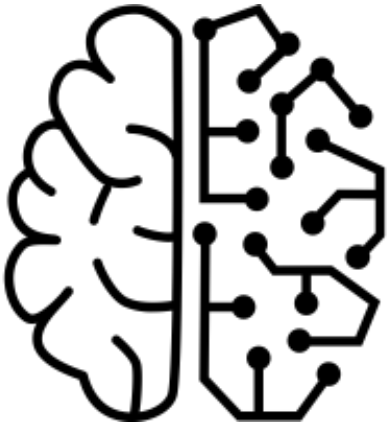


Figure 1: The Machine Learning approach [1].



[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

—Arthur Samuel, 1959

Supervised and Unsupervised Learning

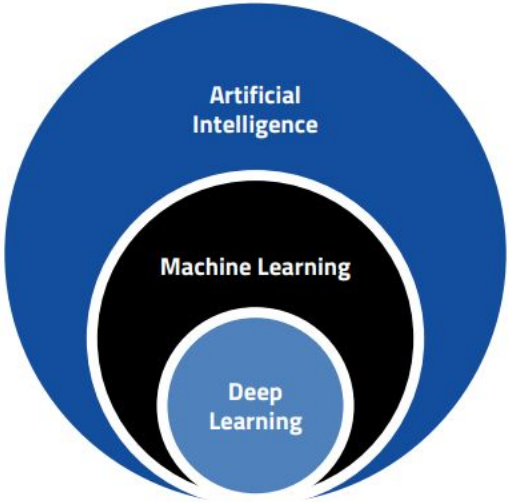
Supervised Learning
Supervised machine learning requires to be labelled input and output data during the training phase of the machine learning lifecycle.

Unsupervised Learning
Unsupervised machine learning is the training of models on raw and unlabelled training data.

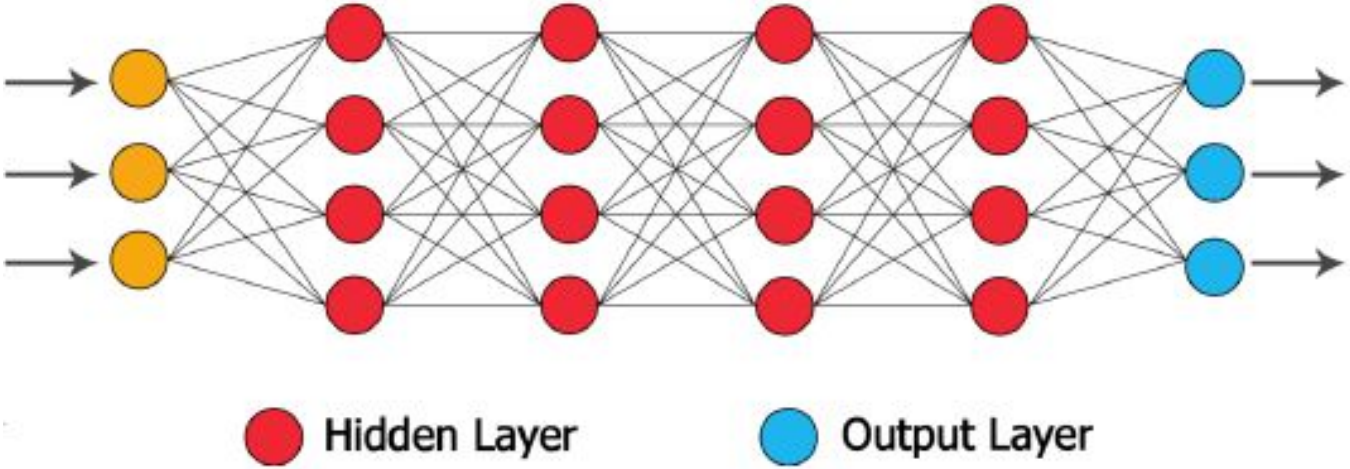
The main difference between supervised vs unsupervised learning is the need for labelled training data. Supervised machine learning relies on labelled input and output training data, whereas unsupervised learning processes unlabelled or raw data.

Deep Learning

Deep Learning is a subclass of Machine Learning algorithms that train Neural Networks to perform tasks



Deep Learning



- Differentiable models that can be trained with Stochastic Gradient Descent
- Unmatched representational power and are capable of feature abstraction: deeper layers abstract more complex relations.
- Extremely versatile and can take in data of many different shapes and formats.
- All state-of-the-art Machine Learning applications are based on Deep Learning and implement Neural Networks.

2.
Dataset

Dataset

https://zenodo.org/record/5126747#.YxnTCyHMKBw

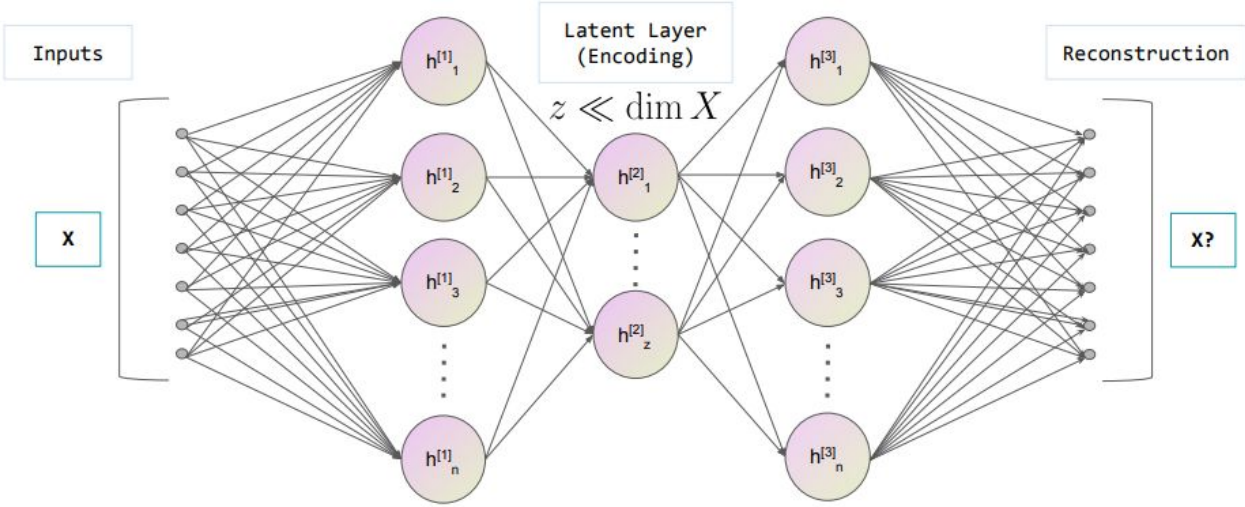
Dataset

This data-set comprises simulated events of pp collisions at 13 TeV with 2 leptons + 1 bottom jet signal state, with HT > 500 GeV. It includes the following samples

- Standard-Model background (bkg), generated at leading order includes the sub-samples Z+Jets, ttbar, WW, WZ, and ZZ.
- Vector-like T-quarks with masses 1.0, 1.2, 1.4 TeV (hq1000, hq1200, hq14000) pair produced either through the Standard-Model gluon (wohg) or through a BSM 3TeV heavy gluon (hg3000)
- tZ production through a Flavour Changing Neutral Current (fcnc) vertex.

3.
Autoencoders

Autoencoders

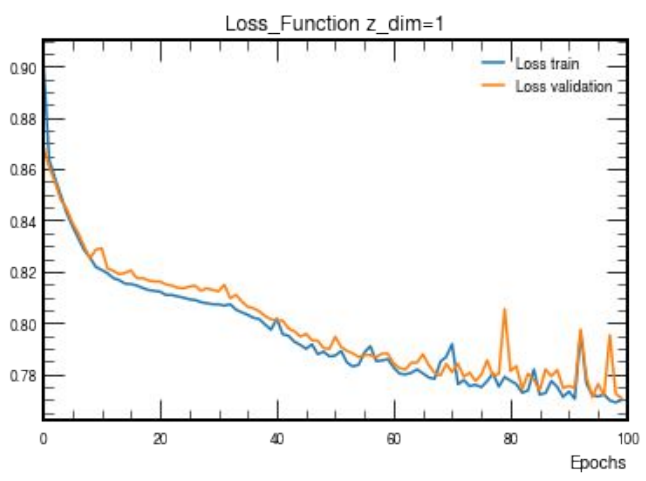


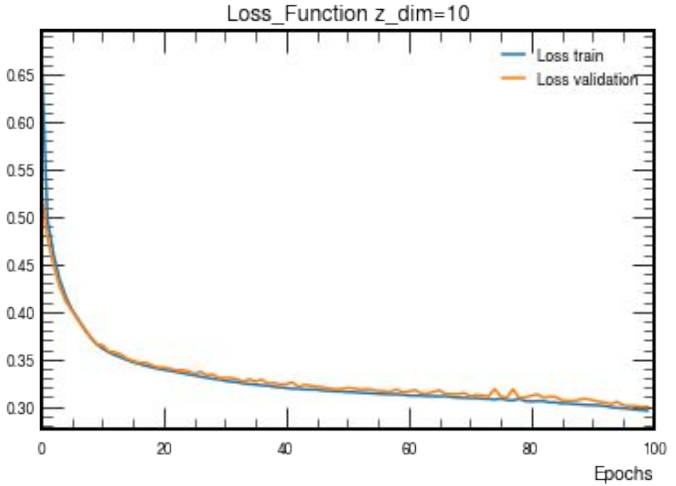
- The Network is trained by minimizing the reconstruction error:

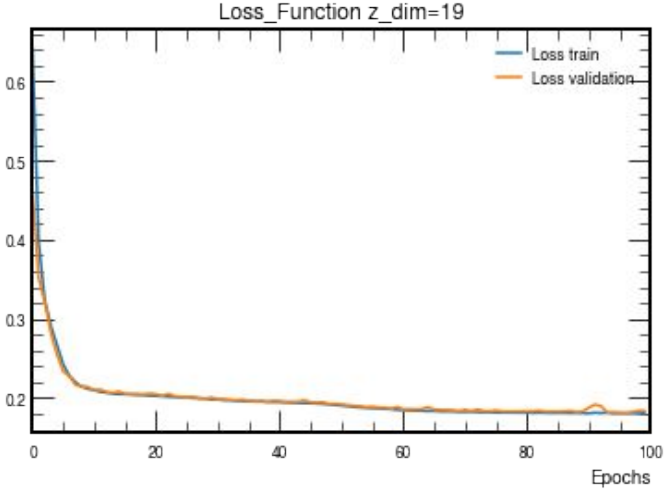
$$L = \frac{1}{N}\sum_{i=1}^{N} |x_i - \mathrm{AE}(x_i)|^2$$

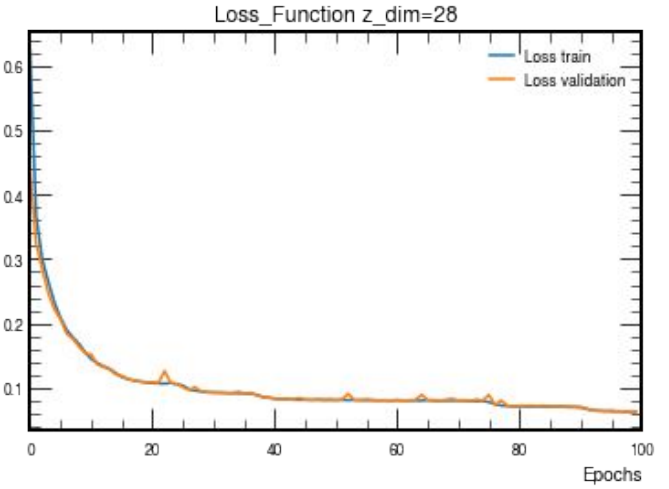
- In principle, events that are easier to reconstruct are the most common.
- Reconstruction error of an event can be a measure of how rare it is, BSM events should have higher reconstruction error.

Loss Function

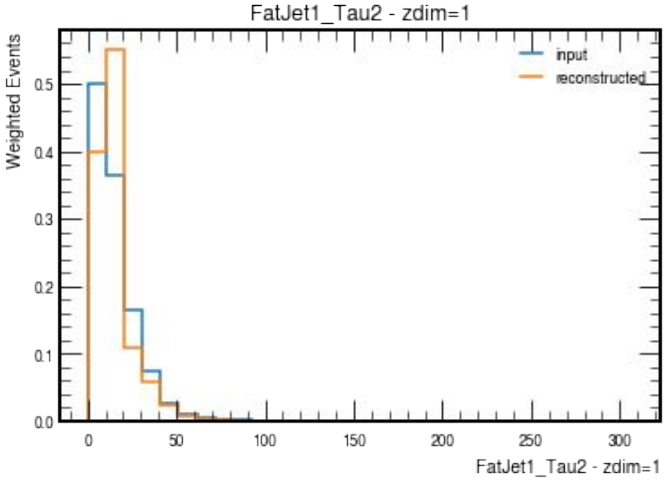


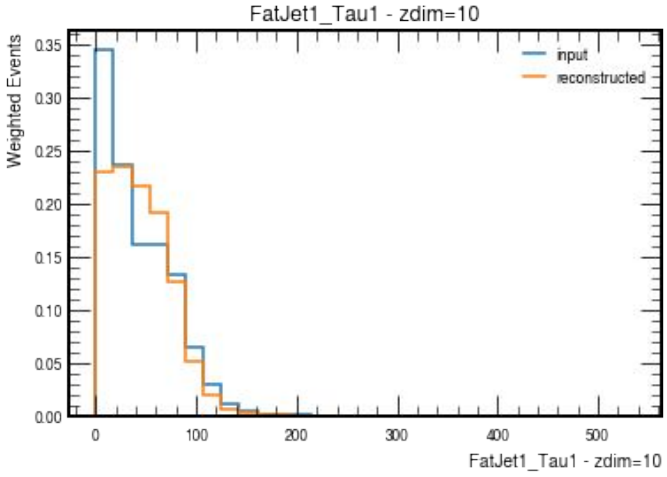


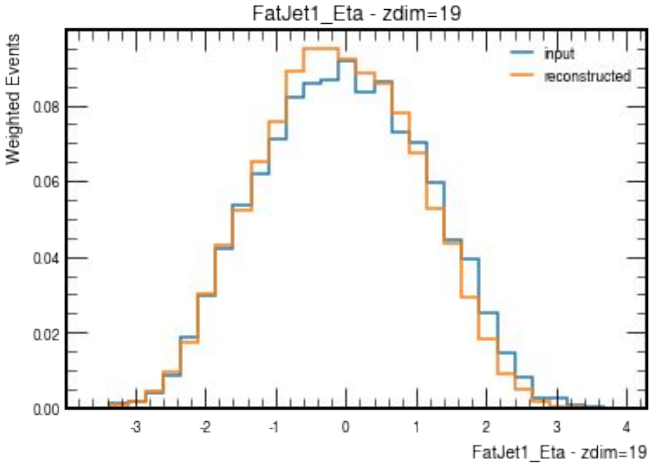


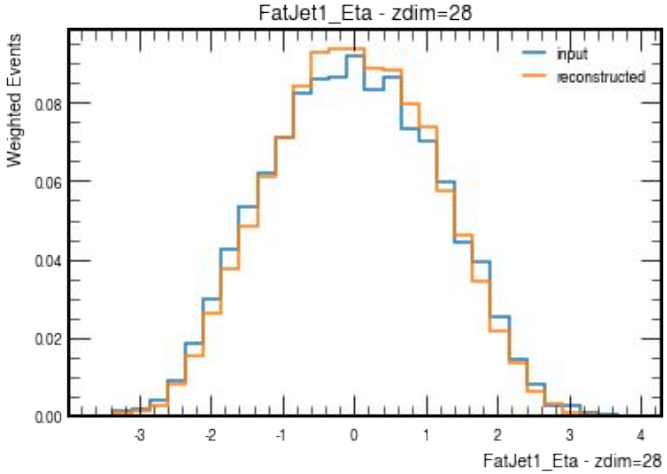


FatJet1_tau2

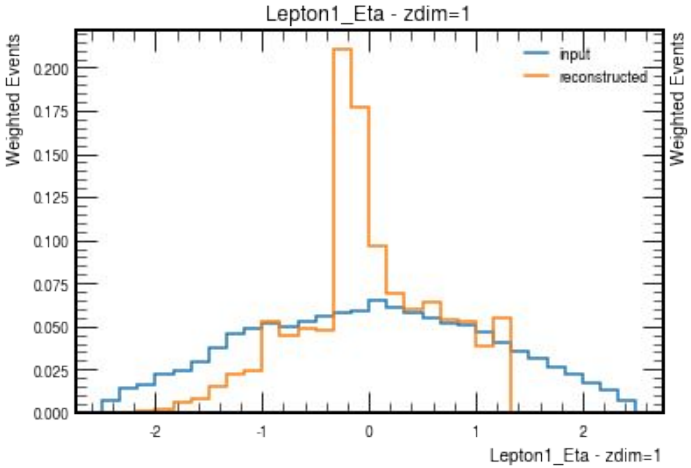


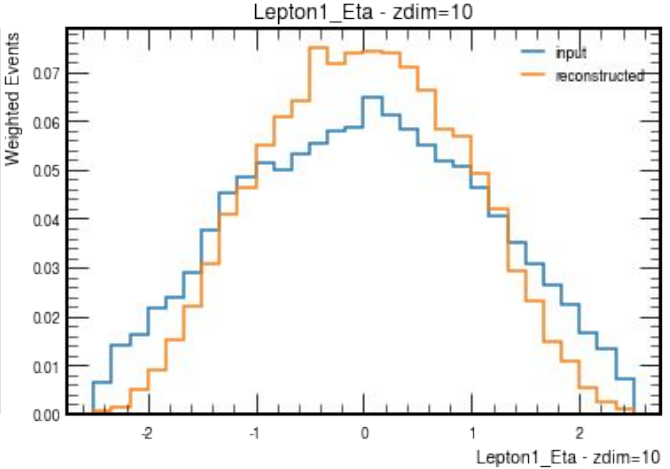


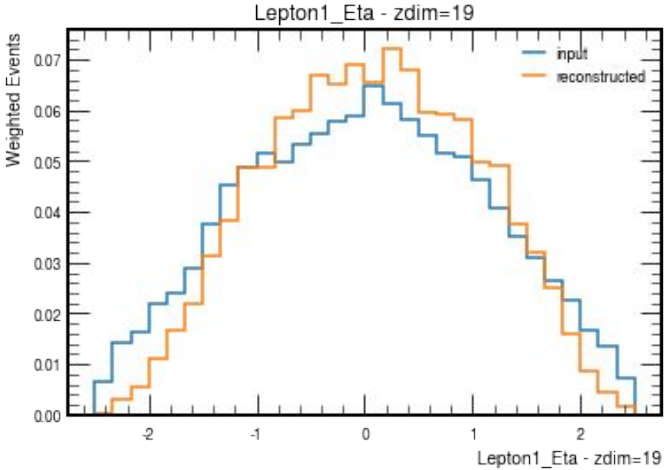


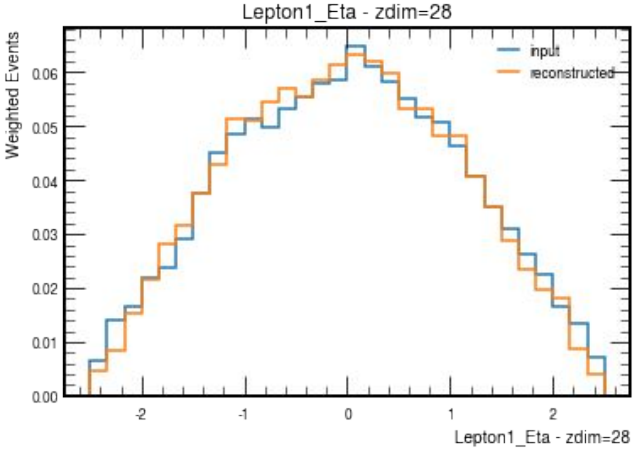


Lepton1_eta

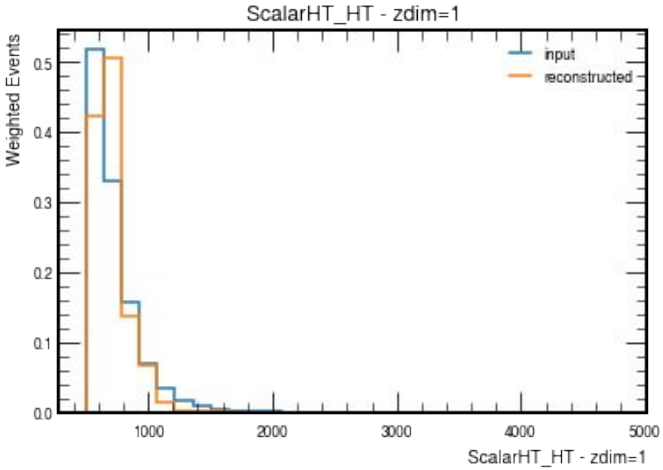

Lepton1_Eta - zdim=1
Weighted Events
input
reconstructed
Lepton1_Eta - zdim=1

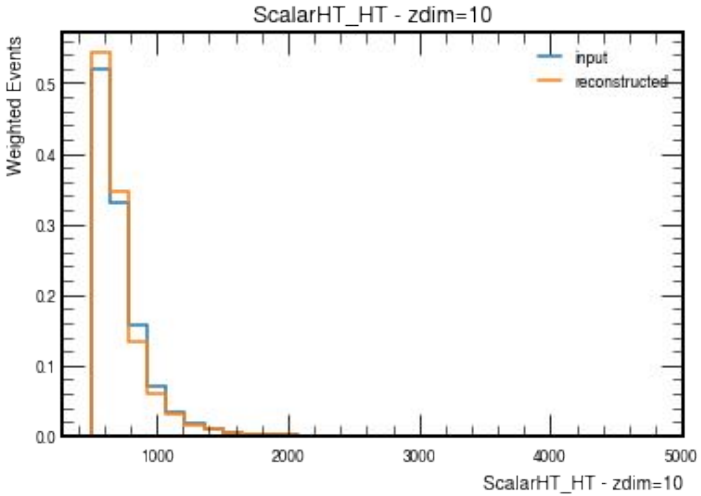

Lepton1_Eta - zdim=10
Weighted Events
input
reconstructed
Lepton1_Eta - zdim=10

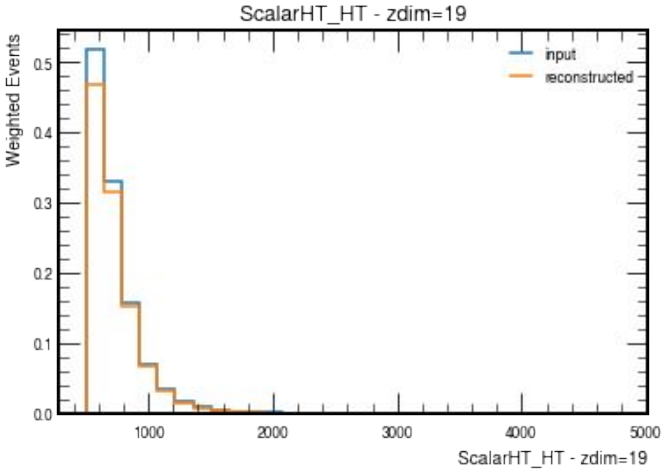

Lepton1_Eta - zdim=19
Weighted Events
input
reconstructed
Lepton1_Eta - zdim=19

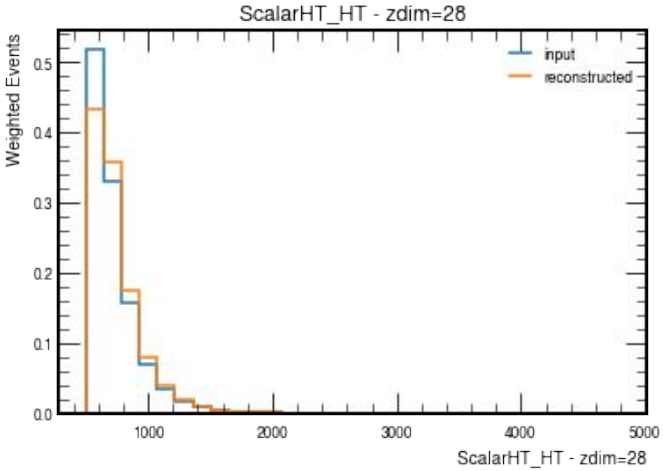

Lepton1_Eta - zdim=28
Weighted Events
input
reconstructed
Lepton1_Eta - zdim=28

ScalarHT


ScalarHT_HT - zdim=1
Weighted Events
input
reconstructed
ScalarHT_HT - zdim=1


ScalarHT_HT - zdim=10
Weighted Events
input
reconstructed
ScalarHT_HT - zdim=10

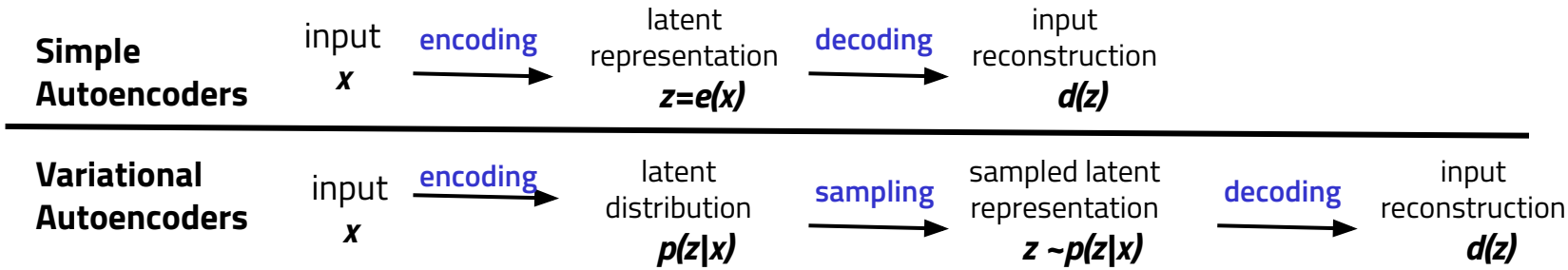

ScalarHT_HT - zdim=19
Weighted Events
input
reconstructed
ScalarHT_HT - zdim=19


ScalarHT_HT - zdim=28
Weighted Events
input
reconstructed
ScalarHT_HT - zdim=28

VAE

A variational autoencoder can be defined as being an autoencoder whose training is regularized to avoid overfitting and ensure that the latent space has good properties that enable generative process.

Instead of encoding an input as a single point, we encode it as a distribution over the latent space.



Overarching Conclusions

Here are some general conclusions of the work carried out:

- An unsupervised architecture was implemented.
- The autoencoder can be optimized using an automatic hyperparameter optimization software framework → OPTUNA.
- Variational autoencoders (VAE) can be implemented for a better reconstruction of the data.

Thanks!

jhonathanbarrios21@gmail.com