

Master's thesis

# Deployment of semi-unsupervised learning in the search for new physics at the LHC with the ATLAS detector

Benchmarking autoencoders for anomaly detection at the LHC with the ATLAS detector

**Sakarias Garcia de Presno Frette**

Computational Science: Physics  
60 ECTS study points

Department of Physics  
Faculty of Mathematics and Natural Sciences

6th April 2023





**Sakarias Garcia de Presno Frette**

**Deployment of semi-unsupervised learning  
in the search for new physics at the LHC  
with the ATLAS detector**

Benchmarking autoencoders for anomaly detection at the  
LHC with the ATLAS detector

Supervisors:  
Professor Farid Ould-Saada  
Dr. James Catmore



# **Abstract**



# Acknowledgments

To be blunt, this exercise is the equivalent of fishing in the dark, with a blindfold on, with a net of an arbitrary size, in an ocean trying to not catch fish, but something else that may or may not look like fish that might or might not be in the ocean.

---

*Sakarias Frette / Spring 2023*

- Veilledere
- William og Mikkel
- Foreldre
- Ernie Bernie
- NN



# Contents



# List of Figures



# List of Tables



# Introduction

Not only is the Universe stranger than we think, it is stranger than we can think.

---

*Werner Heisenberg*

So goes the quote by Werner Heisenberg, acclaimed for his work in quantum physics in the early to mid 20th century. The quote is a reminder and a statement of the fact that the world we live in is immensely strange, beautiful, complex and interesting. From the smallest atoms to the largest galaxy clusters, the universe is a place of wonder and mystery. It is perhaps easy to forget what science tries to do when studying nature. Its fundamental duty is to model the universe as well as possible given the tools available. These models get better over time, but unlike what scientific absolutists might think and argue for, we will never know the whole truth. One should always remember Heisenberg's quote, for it illustrates the very point that nature is indeed more strange than we can think, and science, with its predictive power is only an approximation, one can never truly know if the model we have of nature is a hyperfitted model or the actual instruction manual for nature itself.

As physicists we develop and extend, replaces and debunks models at the most fundamental level of nature, one of which is the Standard model. At higher energies, it is the most accurate, experimentally tested model to date, only rivalled by general relativity. From the 1930's to around 1973, there were huge leaps within the field of particle physics. However, besides experimentally verifying the existence of the Higgs boson, there have not been any major discoveries of the same order as before 1973. As the tools of science became better and better, more and more strange behavior was found around us, behavior that could not be explained by the Standard model. Attempts have been made to create theoretical frameworks that could include the strange behavior into the Standard Model, but all have yet to be experimentally verified. One could ask oneself why this is the case. Why is this new physics, whatever it may look like, so difficult to find? Is the framework wrong? Are the theories not well enough understood or wrong? Or could it be that the new physics are hiding within the data already collected, in some set of features, too subtle for the naked eye, but available through advanced data analysis tools? The answer is, perhaps not to anyone's shock, we have no earthly idea. In fact, they may all be true or false, to some proud physicist's dismay.

Luckily this story does not end here, as there are countless departments all over the world searching for this new physics. The ATLAS experiment at CERN is one organization that has taken upon itself this task, with my supervisors Professor Farid Ould-Saada and Dr. James Catmore working on searches for new physics at the University of Oslo. This thesis will take a somewhat different search approach than conventional analyses done at ATLAS. In a humble attempt, the assumption is made that if the new physics exists, it is too subtle for the naked eye to see, and that it is hiding in the data in some set of features. Further, it is assumed that by focusing on data we can label and that we know from experiment to exist, deviations from this would be of interest for more narrow searches. The hope is to filter out the events of interest, that with some certainty differs from the established theory, and then try to understand them better.

In the last decade, data analysis tools such as neural networks have become more and more available to the public, getting optimized and upgraded, extended and getting more computationally powerful every day. Google's Tensorflow[?] and Facebook's PyTorch[?] are two increasingly popular frameworks, with their own strengths and weaknesses that are used within industry, academia, and even within the particle physics community. Tensorflow was chosen for this thesis, and all of the neural networks are using this framework, together with several other third party softwares. In the epigraph below, the author of this thesis used ChatGPT to write a description of what is attempted in this thesis.

This endeavor resembles casting a line into the abyss, a murky world shrouded in darkness. With vision obscured and a net of unknown scale, the pursuit is not of fish, but of a nebulous entity that may or may not bear resemblance to its aquatic counterparts, if it exists in the briny deep at all.

---

*ChatGPT / Spring 2023*

## Or this?

In describing the work undertaken in this thesis, it feels appropriate to evoke imagery from the world of fishing. Imagine casting a line into the abyss, a murky world shrouded in darkness, where the only sounds that can be heard are the gentle lapping of waves and the occasional call of a seabird. The water is vast, and its depths are unknown. With vision obscured and a net of unknown scale, the pursuit is not of fish, but of a nebulous entity that may or may not bear resemblance to its aquatic counterparts, if it exists in the briny deep at all.

As a fisherman would use bait and tackle to catch their quarry, the tools used in this thesis are data analysis techniques such as neural networks. These are powerful tools that allow us to sift through vast amounts of information and identify patterns that might be invisible to the naked eye. By carefully selecting the data we work with and training our models to recognize specific features, we hope to identify anomalies that might point to the existence of new physics.

The world of particle physics is vast and complex, and there is much that we still do not understand. While the Standard Model has been successful in explaining a wide range of phenomena, it is known to be incomplete. There are many questions that it cannot answer, and many phenomena that it cannot explain. For example, there is currently no explanation for dark matter, which is believed to make up a significant portion of the matter in the universe. Similarly, there are many aspects of the behavior of particles at high energies that are not well understood.

The search for new physics is an ongoing endeavor, and there are many scientists and researchers around the world who are working to push the boundaries of our understanding. Organizations like CERN and the ATLAS experiment are at the forefront of this effort, using powerful tools and cutting-edge techniques to explore the mysteries of the universe.

In this thesis, we take a somewhat unconventional approach to the search for new physics. Rather than looking for large, dramatic signals that might indicate the presence of new particles or phenomena, we focus on more subtle features that might be hidden in the data. By carefully selecting the events we study and training our models to look for specific patterns, we hope to uncover anomalies that might point to the existence of new physics.

Of course, this is a challenging task, and there are many obstacles that must be overcome along the way. The data we work with is complex and noisy, and it can be difficult to separate signal from noise. Similarly, the tools we use to analyze the data are themselves complex, and it can be challenging to ensure that our models are working correctly and producing accurate results.

Despite these challenges, we remain optimistic about the potential of this approach. By focusing on subtle features and anomalies, we hope to identify signals that might be missed by other methods. If successful, this could lead to a deeper understanding of the universe and the fundamental laws that govern it.

In conclusion, the search for new physics is a daunting but exciting endeavor. By using powerful data analysis tools and carefully selecting the events we study, we hope to uncover new phenomena and deepen our understanding of the universe. While the road ahead may be long and challenging, we remain committed to the pursuit of knowledge and the exploration of the mysteries that lie beyond our current understanding.

## Outline of the Thesis

The master thesis is outlined in the following way. The first two chapters are dedicated to the necessary machine learning and standard model physics background required to understand the analysis done and tools used in the thesis. The third chapter goes through the implementation of the project, where the datasets comes from, the ATLAS architecture, the programming libraries, feature choice, and so on. Chapter four goes through the results from the implementation as well as discussion and interpretation of the results, the pros and cons of the implementation, aspects for future improvement, and other thoughts around the process. The final chapter is dedicated to the conclusion, where the findings are summarized.

# Chapter 1

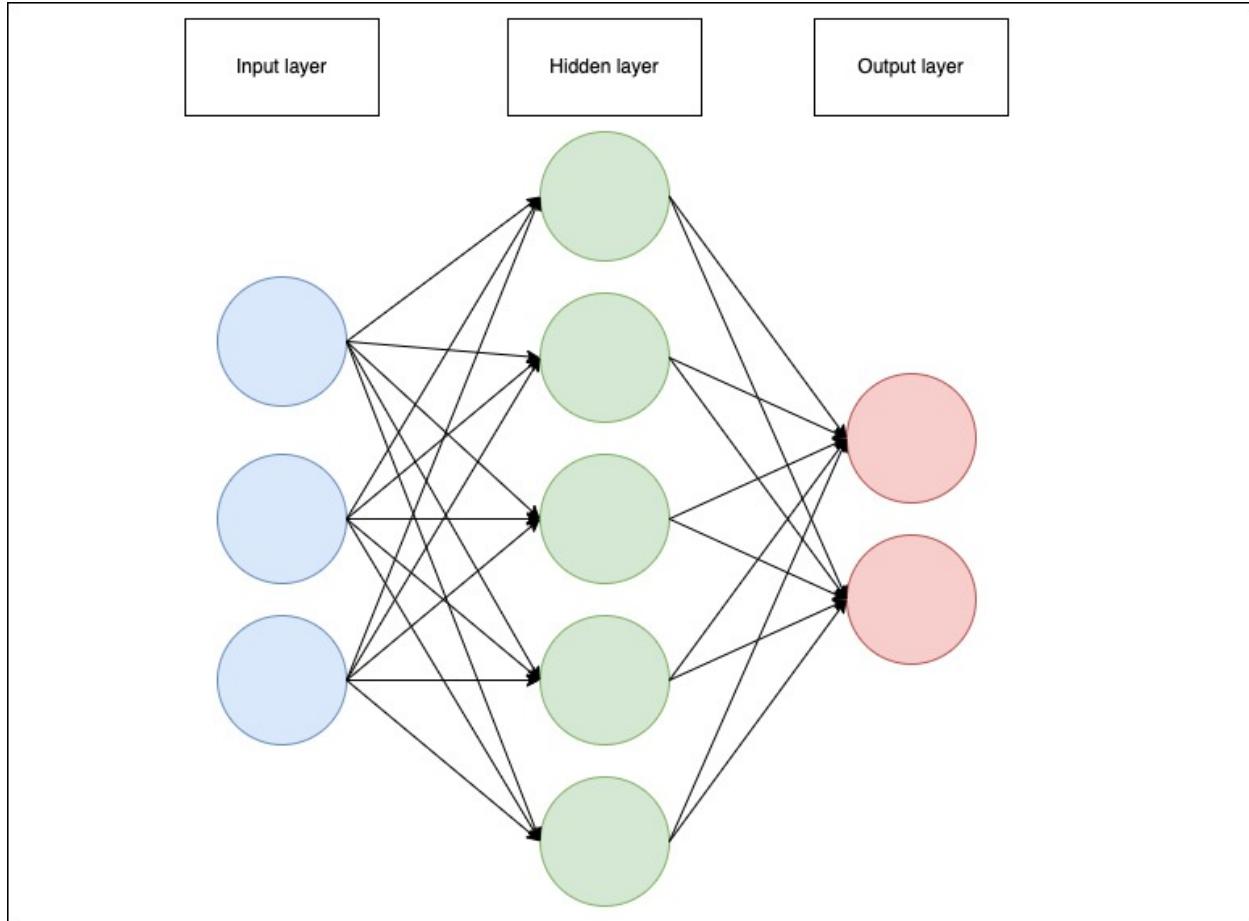
## Data Analysis

### 1.1 Anomaly detection

Anomaly detection is a versatile tool that finds application in a diverse range of scenarios, including fraud detection, anomalous sensor data analysis, and time series data. The primary objective of this tool is to identify data that deviates from a predetermined standard of normal behavior. The definition of this standard can vary from situation to situation, based on the context and the expected anomalous behavior. According to Chandola, Banerjee, and Kumar [?], anomalies can be classified into three categories: *point anomalies*, *contextual anomalies*, and *collective anomalies*. *Point anomalies* represent singular or few outliers from a larger group or context, and can occur in various situations. A notable example of a point anomaly is Michael Phelps, who is able to swim intensively for longer periods due to his body producing less lactic acid. *Contextual anomalies*, on the other hand, are determined based on the context of the anomaly and data, rather than as a whole. For instance, in the case of continuous gas flow data, a sudden change in flow on a Saturday, despite being within the range of Friday's flow, could be categorized as a contextual anomaly. The third type, *collective anomalies*, represents a group of anomalies that deviate from the expected behavior of the dataset. In particle physics experiments, collective anomalies are of particular interest as there are many sources of anomalous behavior, and only collective anomalies are worth investigating. Additionally, the noise generated by a large number of components in such experiments makes it essential to consider collective anomalies.

### 1.2 Neural Networks

In the field of machine learning, statistical algorithms are commonly employed for data analysis. Neural networks, a specific category of such algorithms, have experienced exponential growth in usage in both industry and academia over the past decade. These models find extensive use in a variety of applications, ranging from image analysis to weather prediction. The fundamental principle behind feed forward neural networks (FFNN) involves the data being fed forward through the network, with the end output evaluated and corrections then back propagated through the network to update the weights and biases. This training process is repeated until a certain threshold is met. figure ?? depicts a general layout of a neural network, wherein the input layer consists of one node per feature in the dataset. The number of hidden layers and nodes per layer can be fine-tuned, with the last hidden layer connected to the output layer. The latter is determined by the problem being addressed, and in the case of the binary classification problem depicted in figure ??, the nodes interact via tunable weights  $w$  and biases  $b$  that must be trained on the dataset prior to making predictions.



**Figure 1.1:** Simple neural network diagram drawn using Draw.io. Here the blue dots are the input layer, the green dots are a hidden layer, and the red dots are the output layer. The arrows shows the connections between the nodes.

In order to avoid confusion, we will adhere to table ?? for the notation used in the following sections.

## Gradient descent

Let us now consider a general n-dimensional problem, with parameters  $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_n\}$ . Our objective is to find the set of  $\boldsymbol{\theta}$  to minimize a cost function with respect to the data and target. One way to solve this problem is using ordinary least squares. For this approach, the optimal parameters  $\boldsymbol{\theta}_{opt}$  are derived from minimizing the cost function, as shown here:

$$\boldsymbol{\theta}_{opt} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{t},$$

where  $\mathbf{X}$  is the design matrix containing the data, and  $\mathbf{t}$  is the target vector. This however leads to a problem. Suppose the design matrix is sufficiently large, then the matrix inversion will get computaitonally expensive, or it might not even exist for a given  $\mathbf{X}$ . Thus, an alternative approch is to iteratively approximate the ideal parameters.

Suppose we we have a cost function  $C(\boldsymbol{\theta})$  for a given problem. We can approximate the minimum of the cost function by calculating the gradient  $\nabla_{\boldsymbol{\theta}} C$  with respect to  $\boldsymbol{\theta}$ . The negative of this gradient indicates the direction for the minimum of  $C$  when evaluating it in a specific point  $\boldsymbol{\theta}_i$  in the parameter space [? ]. This is formulized as follows

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \eta \nabla_{\boldsymbol{\theta}} C(\boldsymbol{\theta}_i), \quad (1.1)$$

where  $\eta$  is a step size, also called the learning rate. The choice of  $\eta$  is not a trivial case. It is one of several hyperparameters<sup>1</sup> that can be altered, and that highly depend on the given problem. With regards to the learning rate, there are only three situations to consider, shown in figure ??.

---

<sup>1</sup>Give reference to hyperparameters

**Table 1.1:** Notation

Matrices and vectors		
Notation	Description	Type
$X$	Design Matrix (input data).	$\mathbb{R}^{N \times \# \text{features}}$
$t$	Target values.	$\mathbb{R}^{N \times \# \text{categories}}$
$y$	Model output, the prediction from our network.	$\mathbb{R}^{N \times \# \text{categories}}$
$W^l$	The weight matrix associated with layer $l$ which handles the connections between layer $l - 1$ and $l$ .	$\mathbb{R}^{n_{l-1} \times n_l}$
$B^l$	The bias vector associated with layer $l$ which handles the biases for all nodes in layer $l$ .	$\mathbb{R}^{n_l \times 1}$
Elements		
$w_{ij}^l$	The weight connecting node $i$ in layer $l - 1$ to node $j$ in layer $l$ .	$\mathbb{R}$
$b_j^l$	Bias acting on node $j$ in layer $l$ .	$\mathbb{R}$
$z_j^l$	Node output before activation on node $j$ on layer $l$ .	$\mathbb{R}$
$a_j^l$	Activated node output on node $j$ on layer $l$ .	$\mathbb{R}$
Functions		
$C$	Cost function	
$\sigma^l$	Activation function associated with layer $l$ .	
Quantities		
$n_l$	The number of nodes in layer $l$ .	
$L$	Number of layers in total with $L - 2$ hidden layers.	
$N$	Total number of data points.	
All indexing starts from 1: $i, j, k, l = 1, 2, \dots$		

**Table 1.2:** Table containing notation used for deriving the mathematical formulas for the neural network [? ]



**Figure 1.2:** Figures showing different choices of learning rate for a given cost function, with respect to the tunable parameters. Source: [Jeremy Jordan](#), accessed 03.10.22.

Figure ?? visualizes the relation between the learning rate and the cost function. In the left most figure we note that the learning rate is too small. This leads to many iterations before you reach a minimum. In the right most figure we note that the learning rate is too high, and the result is that we get divergent behavior. Thus the goal is to find the optimal learning rate, shown in the middle figure.

A modified and preferred version of gradient descent is the so called stochastic gradient descent. Regular gradient descent can, for large datasets be quite slow, and is prone to getting stuck in a local minima. To circumvent this issue, mini batches are introduced.

## Feed forwarding

Inference (prediction) and training both use the same feed-forward algorithm. Lets then assume that we have generated a network. The network initializes the weights and biases usually with normal or uniformly distributed values, that can later be tuned. The procedure is to send the data through the network, weighting each connection according to the networks architecture, and produce an output. The procedure can be summarized in the following steps [? ]:

- The data is received by the input nodes in the network for each feature.
- Each input node weights the data value according to the connection of each node in the next layer.
- Every node in the hidden layers sums the weighted data values, and adds the bias associated to the given node, denoted as  $z$ .
- This value  $z$  is then sent through an activation function  $\sigma$ , which produces the output of the node, denoted as  $a = \sigma(z)$ .
- This process is repeated for each hidden layer, and it is important to note that the number of nodes in the hidden layers is not dependent on the number of features in the original dataset.
- The last hidden layer then sends the activated values to the output layer, where the number of nodes and choice of activation function depends on the problem to solve.

Mathematically this is expressed as follows:

$$z_j^l = \sum_{i=1}^{n_{l-1}} w_{ij}^l a_i^{l-1} + b_j^l, \quad a_j^l = \sigma^l(z_j^l), \quad (1.2)$$

where  $l$  is the layer index,  $j$  is the node index, and  $i$  is the index of the node in the previous layer, and  $l \neq 1$ , as it is not used on the input layer.

## Backpropagation

The way neural networks learn is conventionally by the use of the backpropagation algorithm, first proposed by Rumelhart et al[? ]. This is a bit misleading, as the backpropagation algorithm actually only refers to how to compute the gradient[? ]. The algorithm allows us to alter the weights and biases such that we get an ideal output. Assuming a cost function  $C$ , we can calculate the gradient  $\nabla_{w,b}C$ , and use this to back propagate the error correction. The gradient  $\nabla_{w,b}C$  is comprised of two derivatives<sup>2</sup>:

$$\nabla_{w,b}C = \left( \frac{\partial C}{\partial w_{i,j}^L}, \frac{\partial C}{\partial b_j^L} \right).$$

We have to use the chain rule to calculate the derivatives, and using that the last layer is  $l = L$ , we get the derivative with respect to the weights as

$$\frac{\partial C}{\partial w_{i,j}^L} = \frac{\partial C}{\partial a_j^L} \frac{\partial a_j^L}{\partial z_j^L} \frac{\partial z_j^L}{\partial w_{i,j}^L},$$

where

$$a_j^L = \sigma(z_j^L), \quad z_j^L = \sum_{i=1}^{n_L-1} w_{i,j}^L a_i^{L-1} + b_j^L.$$

This then gives us

$$\frac{\partial C}{\partial w_{i,j}^L} = \frac{\partial C}{\partial a_j^L} \sigma'(z_j^L) a_i^{L-1},$$

where we defined that

$$\sigma'(z_j^L) = \frac{\partial a_j^L}{\partial z_j^L}. \quad (1.3)$$

This derivative is very easy to calculate given a specific cost function and activation function. The derivative with respect to the bias is given as follows:

$$\frac{\partial C}{\partial b_k^L} = \frac{\partial C}{\partial a_j^L} \frac{\partial a_j^L}{\partial z_j^L} \frac{\partial z_j^L}{\partial b_j^L},$$

which gives us the final expression as

$$\frac{\partial C}{\partial b_k^L} = \frac{\partial C}{\partial a_j^L} \sigma'(z_j^L).$$

We will now introduce a new notation, a local gradient commonly called the "error". It reflects how the rate of change of the cost function depends on the  $j$ 'th node in the  $l$ 'th layer.

$$\delta_j^l \equiv \frac{\partial C}{\partial z_j^l}.$$

Using this we get the following expression:

$$\delta_j^L = \frac{\partial C}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L} \frac{\partial a_j^L}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L} \sigma'(z_j^L),$$

giving us the more compact forms of the derivatives with respect to the weights and biases:

$$\frac{\partial C}{\partial w_{i,j}^L} = \delta_j^L a_i^{L-1}, \quad \frac{\partial C}{\partial b_j^L} = \delta_j^L.$$

We can now let  $\boldsymbol{\delta}^l$  be the vector of all the errors in the  $l$ 'th layer, and  $\boldsymbol{\delta}^L$  be the vector of all the errors in the last layer. The error in the  $l$ 'th layer can then be expressed as a matrix equation for the last layer as follows:

$$\boldsymbol{\delta}^l = \nabla_a C \odot \frac{\partial \sigma}{\partial z^l}, \quad \nabla_a C = \left[ \frac{\partial C}{\partial a_1^L}, \frac{\partial C}{\partial a_2^L}, \dots, \frac{\partial C}{\partial a_{n_L}^L} \right]^T.$$

---

<sup>2</sup>NOTE TO SELF:: Note that this calculation is not a generalized algorithm for backpropagation, but rather for a special case using MSE.

Here  $\odot$  is the Hadamard product (element wise product). This local gradient can now be defined recursively for the  $j$ 'th node in a layer  $l$  as a function of the local error in the next layer:

$$\delta_j^l \equiv \frac{\partial C}{\partial z_j^l} = \sum_k \frac{\partial C}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial z_j^l} = \sum_k \frac{\partial z_k^{l+1}}{\partial z_j^l} \delta_k^{l+1}. \quad (1.4)$$

We also note that

$$z_k^{l+1} = \sum_{j=1}^{n_l} w_{j,k}^{l+1} a_j^l + b_k^{l+1} = \sum_{j=1}^{n_l} w_{j,k}^{l+1} \sigma(z_j^l) + b_k^{l+1},$$

thus the partial derivative is given as

$$\frac{\partial z_k^{l+1}}{\partial z_j^l} = w_{j,k}^{l+1} \sigma'(z_j^l), \quad (1.5)$$

using the substitution from equation ???. This allows us to substitute equation ?? into equation ?? to get the following expression:

$$\delta_j^l = \sum_k w_{j,k}^{l+1} \sigma'(z_j^l) \delta_k^{l+1}. \quad (1.6)$$

Using this we can derive a three step formula for the backpropagation algorithm:

- Compute the local gradient for the last layer,  $\delta^L$ .
- Recursively compute the local gradient for the remaining layers,  $\delta^l$  for  $l = L - 1, L - 2, \dots, 1$ .
- Update the weights and biases for all layers,  $l = 1, 2, \dots, L$ , given the learning rate  $\eta$  as shown below:

$$w_{i,j}^l \leftarrow w_{i,j}^l - \eta \delta_j^l a_i^{l-1},$$

$$b_j^l \leftarrow b_j^l - \eta \delta_j^l.$$

### 1.3 Autoencoders

Autoencoders are a subset of neural networks. Whereas a general neural network in principle can take any shape, autoencoders are more restrictive. This restrictiveness can in its most general sense be condensed into the following points:

- Same number of output categories as input categories
- A latent space with smaller dimensionality than the input/output layer

What we end up with are two funnel shaped parts linked together. The two funnels are called the encoder (left funnel) and decoder (right funnel) respectively. This architecture is not accidental, but rather designed with a very specific solution of problems in mind, reconstruction. A good example to illustrate this is image reconstruction, illustrated in figure ???. Suppose you have an image, and want to reconstruct it. By feeding the encoder an image, and comparing the decoder output to the actual image, the autoencoder can tune itself to recreate the images it trains on.



**Figure 1.3:** Figure depicting a model for an autoencoder. Here the input  $\mathbf{x}$  is the original image,  $\mathbf{x}'$  is a reconstructed version of  $\mathbf{x}$ ,  $g_\phi$  is the encoder,  $f_\theta$  is the decoder, and  $\mathbf{z}$  is the latent space. Found 14.01.23 [here](#) [? ].

Mathematically this is represented as follows[? ]. Using the annotations of each component in figure ?? the decoded information is defined as follows

$$\mathbf{z} = \mathbf{g}_\phi(\mathbf{x}),$$

and the reconstruction given as

$$\mathbf{x}' = \mathbf{f}_\theta(\mathbf{g}_\phi(\mathbf{x})).$$

The parameters  $(\phi, \theta)$  are the tuneable parameters adjusted according to the loss function. In our case, the goal is reconstruction without copying, thus we can simply use mean squared error, given as

$$L_{AE}(\phi, \theta) = \frac{1}{N} \sum_{i=0}^{N-1} (\mathbf{x}^i - \mathbf{f}_\theta(\mathbf{g}_\phi(\mathbf{x}^i)))^2. \quad (1.7)$$

## 1.4 Variational autoencoders

Another popular method for reconstruction is the so called variational autoencoder. The work by Kingman and Welling [?] showed how one can use the variational bayesian approach for efficient approximate posterior inference, leading to the use of variational autoencoders. Here, contrary to regular autoencoders, the latent space is thought of as a distribution. In the context of reconstruction, this means that we want to create a latent space distribution based on the true distribution in the data, and use this latent space distribution to then generate data given that latent space.

### Variational Bayes

Let's assume some dataset  $X = \{\mathbf{x}^{(i)}\}_{i=1}^N$  with  $N$  samples for a variable  $\mathbf{x}$ , and also assume that the generation of the data comes from some random variable  $\mathbf{z}$ . We also assume that this variable  $\mathbf{z}$  is generated from a prior distribution  $p_\theta(\mathbf{z})$  and that the data is then generated from a conditional distribution  $p_\theta(\mathbf{x}|\mathbf{z})$ , and that both their respective probability density functions are sufficiently differentiable with respect to parameters  $\theta$  and  $\mathbf{z}$ . We then have that the true posterior distribution is given by  $p_\theta(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})/p_\theta(\mathbf{x})$ . The work done by Kingman and Welling [?] was motivated by the fact that  $p_\theta(\mathbf{z}|\mathbf{x})$  more often than not is intractable<sup>3</sup>, thus instead we will try to approximate the true posterior with a variational approximation  $p_\theta(\mathbf{z}|\mathbf{x}) \approx q_\phi(\mathbf{z}|\mathbf{x})$ . It can from here on out be useful to think of  $q_\phi(\mathbf{z}|\mathbf{x})$  as a probabilistic encoder, as it for a given data point  $\mathbf{x}$  will produce a distribution over possible values for  $\mathbf{z}$ , the latent space, that the datapoint could have been generated from. Similarly, it can be useful to think of the  $p_\theta(\mathbf{x}|\mathbf{z})$  as a probabilistic decoder, as it for a given  $\mathbf{z}$  will produce a distribution over possible values for  $\mathbf{x}$ .

<sup>3</sup>Intractability here refers to the inability to evaluate or differentiate a given distribution or function, or difficulty with solving a problem due to a large parameter space or finding the global optimum of a complex function.

### The variational bound

It can be shown that the marginal likelihood can be written as follows:

$$\log p_\theta(x^{(i)}) = D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z|x^{(i)})) + \mathcal{L}(\theta, \phi; x^{(i)}), \quad (1.8)$$

where the first term on the right hand side is the Kullback-Leibler (KL) divergence<sup>4</sup> and the second term is the evidence lower bound (ELBO) on the marginal likelihood. As the KL divergence is non-negative, the ELBO will always be equal to or less than the marginal likelihood, written below:

$$\log p_\theta(x^{(i)}) \geq \mathcal{L}(\theta, \phi; x^{(i)}) = \mathbb{E}_{q_\phi(z|x)}[-\log q_\phi(x|z) + \log p_\theta(x|z)]. \quad (1.9)$$

Rewriting we get that the ELBO is given as:

$$\mathcal{L}(\theta, \phi; x^{(i)}) = -D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z|x^{(i)})) + \mathbb{E}_{q_\phi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)], \quad (1.10)$$

which is also known as the loss function for variational inference. As a loss function, this needs to be differentiated with respect to  $(\phi, \theta)$ , but as shown by Kingman and Welling [?], the common method using Monte Carlo gradient estimator have high variance, and thus is not ideal.

### The Stochastic gradient variational bayes (SGVB) estimator

Kingman and Welling [?] instead proposes a practical estimator that both deals with effectiveness with large datasets, as well as the issue of intractability. This estimator is called the stochastic gradient variational bayes (SGVB) estimator, and first assumes an approximate posterior  $q_\phi(z|x)$ . Under certain conditions, as shown in subsection ??, we can reparameterize this approximate posterior to a random variable  $\tilde{z} \sim q_\theta(z|x)$ , by doing a differentiable transformation  $g_\phi(\epsilon|x)$ , such that

$$\tilde{z} = g_\phi(\epsilon|x), \quad \epsilon \sim p(\epsilon),$$

where  $\epsilon$  is a noise variable.

Using Monte Carlo estimates of expectations of a function  $f(z)$  with respect to  $q_\theta(z|x)$ , Kingman and Welling [?] shows that:

$$\mathbb{E}_{q_\theta(z|x^{(i)})}[f(z)] = \mathbb{E}_{p(\epsilon)}[f(g_\phi(\epsilon, x^{(i)}))] \approx \frac{1}{L} \sum_{l=1}^L f(g_\phi(\epsilon^{(l)}, x^{(i)})), \quad (1.11)$$

where  $\epsilon^{(l)} \sim p(\epsilon)$ . Using this technique, and assuming that the KL divergence  $D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z|x^{(i)}))$  can be integrated analytically, we only get one term in equation ?? that requires estimation by sampling:  $\mathbb{E}_{q_\phi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)]$ . What we end up with is a version of the SGVB estimator:  $\tilde{\mathcal{L}}^B(\theta, \phi; x^{(i)}) \simeq \mathcal{L}(\theta, \phi; x^{(i)})$ . This is given as:

$$\mathcal{L}(\theta, \phi; x^{(i)}) = -D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) + \frac{1}{L} \sum_{l=1}^L (\log p_\theta(x^{(i)}|z^{(i,l)})), \quad (1.12)$$

where  $z^{(i,l)} = g_\phi(\epsilon^{(i,l)}, x^{(i)})$  and  $\epsilon^{(i)} \sim p(\epsilon)$ . What we now have is a loss function that contains two terms. The first term, the KL divergence, acts as a regularizer, whereas the other term acts as a negative reconstruction error. In fact, we have here two objectives, we want to minimize the ELBO,  $\mathcal{L}(\theta, \phi; x^{(i)})$ , by minimizing the KL divergence and maximizing the expected log-likelihood.

### Reparameterization

If the latent space is assumed to be a continuous variable sampled from a conditional continuous distribution  $z \sim q_\phi(z|x)$ , then we can reparametrize  $z$  as a deterministic variable. This is very useful as we then can rewrite the expectation of the conditional distribution such that the Monte Carlo estimate of the expectation is differentiable with respect to the parameters  $\phi$ . The proof can be found in the article[?]. Now let's take the example of the univariate Gaussian distribution. First, let  $z \sim p(z|x) = \mathcal{N}(\mu, \sigma^2)$ . Then, a reparametrization of  $z$  can be  $z = \mu + \sigma\epsilon$ , where  $\epsilon \sim \mathcal{N}(0, 1)$ . Thus

$$\mathbb{E}_{\mathcal{N}(z;\mu,\sigma^2)}[f(z)] \simeq \frac{1}{L} \sum_{l=1}^L f(\mu + \sigma\epsilon^{(l)}).$$

---

<sup>4</sup>The KL divergence is a measure of how one probability distribution is different from a second, reference probability distribution. It is often used as a distance measure between two probability distributions.

## Variational autoencoders

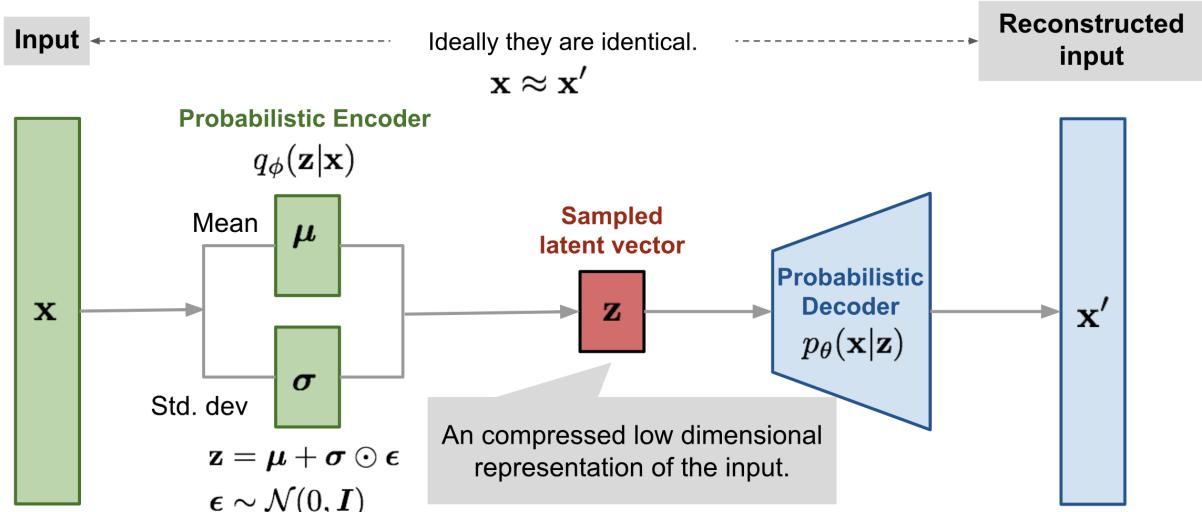
Now, this variational bayes can then be used to create a variational autoencoder. This contains two neural networks, the generative model  $p_\theta(x|z)$ , as well as the probabilistic encoder  $q_\phi(z|x)$ , which is used to approximate the posterior  $p_\theta(z|x)$ . We now let the prior over the latent space be a multivariate Gaussian, lacking parameters:  $p_\theta(z) = \mathcal{N}(z; 0, I)$ . We also have that  $p_\theta(x|z)$  is a multivariate Gaussian, with a distribution generated from  $z$ , whilst  $p_\theta(z|x)$  is in fact intractable. If we now assume that the true (intractable) posterior is a Gaussian with approximately diagonal covariance, we can let the variational approximate posterior be a multivariate Gaussian:

$$\log q_\theta(z|x^{(i)}) = \log \mathcal{N}(z; \mu^{(i)}, \sigma^{2(i)}I), \quad (1.13)$$

where the mean and standard deviation are given by the neural network. Now, we sample from the approximate posterior  $z^{(i,l)} \sim q_\theta(z|x^{(i)})$  where  $z^{(i,l)} = g_\phi(x^{(i)}, \epsilon^{(l)}) = \mu^{(i)} + \sigma^{(i)} \odot \epsilon^{(l)}$ ,  $\epsilon^{(l)} \sim \mathcal{N}(0, I)$  and  $\odot$  is the elementwise product. If both the prior  $p_\theta(z)$  and  $q_\theta(z|x)$  are Gaussian, the KL divergence can be computed analytically, and it can then be showed[?] that the variational approximate posterior is:

$$\mathcal{L}(\theta, \phi; x^{(i)}) \simeq \frac{1}{2} \sum_{j=1}^J (1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2) + \frac{1}{L} \sum_{l=1}^L \log p_\theta(x^{(i)}|z^{(i,l)}), \quad (1.14)$$

where  $z^{(i,l)} = \mu^{(i)} + \sigma^{(i)} \odot \epsilon^{(l)}$  and  $\epsilon^{(l)} \sim \mathcal{N}(0, I)$ .



**Figure 1.4:** Figure depicting a model for a variational autoencoder. Found 14.01.23 [here](#) [? ].

Figure ?? is a graphical representation of the variational autoencoder. The encoder is a neural network, which takes the input  $x$  and maps the input to a latent space by creating a Gaussian distribution. The decoder is a neural network, which takes the latent space  $z$  and outputs the parameters of a Gaussian distribution for the data. The loss function is given by equation ??.

## 1.5 Other tools and algorithms

### Adaptable moment estimation (ADAM) Optimizer

Stochastic gradient descent, though very useful, lack the ability to adapt to the feature space. One algorithm that address this issue is the ADAM optimizer[? ]. The Adaptable moment estimation (ADAM) algorithm uses stochastic gradient descent, but with an adaptive learning rate. This learning rate is adjusted by calculating estimates for the first and second moment<sup>5</sup>. Thus, a large gradient would indicate close proximity to a minimum in feature space, thus a lower learning rate would yield a more accurate result, whereas a small gradient would suggest far proximity to a local minimum, and thus a larger learning rate would increase the chance of approaching a minimum.

<sup>5</sup>In statistics the first moment is the expectation value for a distribution,  $E[X - \mu]$ . The second moment is the expectation value of the distribution squared, i.e the variance,  $E[(X - \mu)^2]$

## Hyperband

Hyperband is a tool for hyperparameter optimization[? ]. Hyperparameter optimization is of high importance in the search for ideal structures and architectures when using neural networks, as there is not a way to find an a priori setup for a given problem. Several algorithms are used, from random search, grid search, and bayesian optimization. Hyperband is an algorithm proposed by L. Li et al. It focuses on using successful halving[? ] but at the same time doing a grid search for how to allocate resources. Successive halving focuses on testing n configurations, and removing the bottom half, thus (hopefully) quickly converging to the ideal combination. However, it is not easy to a priori know the number of configurations n, and how much resources one needs, r, to quickly find the ideal set. This is where Hyperband comes in. In essence, it fetches and tries different combinations of r resources (time, data set subsampling or feature subsampling) and n configurations, to determine the ideal set of hyperparameters via successive halving, yielding 5x to 30x speedup compared to Bayesian optimization. One drawback for this algorithm is that one cannot guarantee that the configuration is optimal, but rather that it is good enough.

## Activation functions

Several activation functions are used in neural networks, and how one chooses the best combination for a given problem is not trivial. This leads often to the use of tuning. Below are a list of the functions used in this thesis, as well as their mathematical definitions.

1.  $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$
2.  $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
3.  $\text{ReLU}(x) = \max(0, x)$
4.  $\text{LeakyReLU}(x) = \max(\alpha x, x)$
5.  $\text{Softmax}(x) = \frac{e^x}{\sum_{i=1}^n e^x}$
6.  $\text{Linear}(x) = x$

## ROC curve

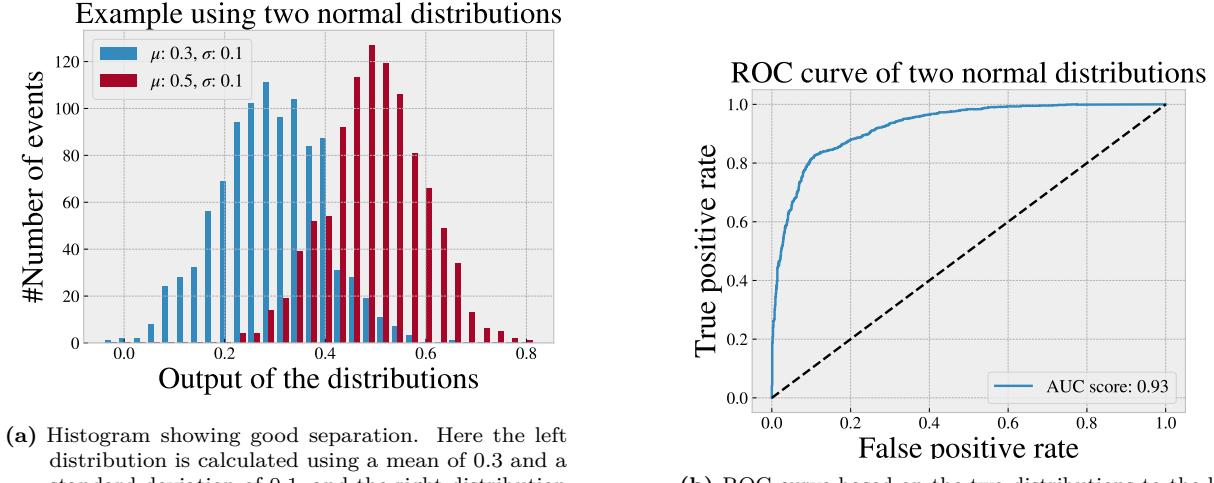
One way to measure the effectiveness of a prediction for a machine learning algorithm is with the use of a ROC curve. The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) which will be defined shortly. Suppose a classifier  $\mathcal{M}$  does a binary classification of positive and negative. If  $\mathcal{M}$  predicts a positive label, and the instance is indeed positive, we define this as a *true positive*. The same goes for negative prediction of a negative instance, which is defined as a *true negative*. Then there is the case where  $\mathcal{M}$  predicts a negative label when the instance is positive, defined as a *false negative*, and vice versa is then defined as a *false positive*[? ]. Now, the TPR is defined as the ratio of true positives to the total number of positive instances, i.e

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

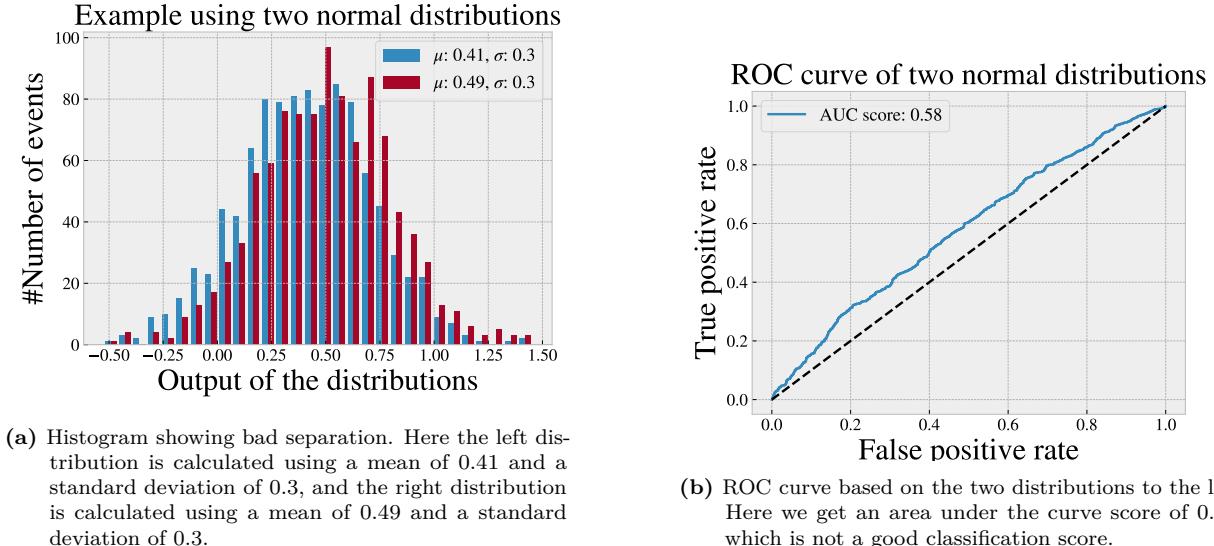
and the FPR is defined as the ratio of false positives to the total number of negative instances, i.e

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

Using this, we get a plot that tells us how well the classifier performs. As the goal usually is it to have a high TPR and a low FPR, the more "north west"[? ] of the plot the curve is, the better. Below are two examples of a ROC curve and the classifier output it is calculated from. For simplicity, the classifier output is in this case just two normal distributions with the mean and standard deviation listed in the legend of the distribution plots.



**Figure 1.5:** Histogram and ROC curve for two well separated distributions.



**Figure 1.6:** Histogram and ROC curve for two poorly separated distributions.

In figure ?? we have a histogram showing two distributions that are well separated, and the ROC curve that is calculated based on those distributions. If we compare this to figure ??, we see that the better separation of the two distributions, the better ROC curve. This is a useful tool to use with the autoencoder, as it provides valuable insight beyond just looking at the output distributions.

## Statistical significance

In the frequentist statistics, the Poisson distribution can be approximated with a Gaussian distribution in the limit of large number of events[? ]. The expression for the significance is given then as

$$Z = \frac{s}{\sqrt{b}}, \quad (1.15)$$

where  $s$  is the amount of signal, and  $b$  is the amount of background. For low statistics we have that the significance is given as

$$Z = \sqrt{2 \left[ (s+b) \ln \left( 1 + \frac{s}{b} \right) - s \right]}. \quad (1.16)$$

It can be shown that in the limit where  $s \ll b$ , the two expressions are approximately the same[? ].



## Chapter 2

# The Standard model and BSM physics



Figure 2.1: The standard model of elementary particles. Source [here](#). Accessed 07.10.22

## 2.1 Structure and composition of the Standard Model

This section will describe the standard model in a phenomenological way, as the mathematics and physical reasoning behind the theory is not of great importance to understand the work presented here, nor the results or discussion. For a more technical explanation, read (Pich, 2008)[?] for a well written paper containing some more standard model fundamentals, as well as summarizing the experimental status regarding the standard model. For more mathematical understanding of the standard model, Peskin and Schroeder's "An introduction to Quantum Field theory" (Peskin and Schroeder, 1995)[?] is highly recommended. Finally, see Thomson's "Modern Particle physics" (Thomson, 2013)[?] for a very comprehensive and up-to-date book that is easy to read and understand.

The standard model is to physicists what the periodic table is to chemists, and is to this day the most fundamental description of matter as we know it at the subatomic scale. It is comprised of two parent class particles, fermions and bosons, where fermions are comprised of quarks and leptons. The model contains 6 leptons, 6 quarks coming in 3 colors each, 4 gauge bosons mediating the electroweak interactions ( $\gamma, W^\pm, Z^0$ ), 8 gluons behind the strong interaction and one scalar Higgs explaining particle masses, all of which are shown in figure ??.

### Fermions

The fermions are the building blocks of matter, and contain two types of particles, leptons and quarks. the up and down quarks form protons and neutrons, which together with electrons forms the atoms around us. Fermions, unlike bosons, are spin half particles. The fermions are grouped into three so called families:

$$\begin{bmatrix} \nu_e & u \\ e^- & d' \end{bmatrix}, \quad \begin{bmatrix} \nu_\mu & c \\ \mu^- & s' \end{bmatrix}, \quad \begin{bmatrix} \nu_\tau & t \\ \tau^- & b' \end{bmatrix}$$

Note that the left column contains the leptons. whilst the right column contains the quarks. Within the left column, the subscripted  $\nu$  denotes what kind of neutrino that corresponds to the given lepton. Here, the first family consists of the electron, the electron neutrino, the up and down quarks. The second family consists of the muon and the muon neutrino, the charm and strange quarks. The third family consists of the tau and the tau neutrino, the top and bottom quarks. The masses of these particles increases for each particle in the matrix as the family number increases, i.e the muon is heavier than the electron, and the tau is heavier than the muon, and so on for the other charged fermions. It is not known whether or not the neutrinos follow this pattern as well due to their very low mass. Below is a table with specific properties of the fermions.

**Table 2.1:** Table showing properties of all the fermions, including name, symbol, antiparticle, spin, charge, generation and mass.

Generation	Name	Symbol	Antiparticle	Spin	Charge	Mass (MeV/c <sup>2</sup> )
Quarks						
1	up	u	$\bar{u}$	1/2	2/3	$2.2^{+0.6}_{-0.4}$
	down	d	$\bar{d}$	1/2	-1/3	$4.6^{+0.5}_{-0.4}$
2	charm	c	$\bar{c}$	1/2	2/3	$1280 \pm 30$
	strange	s	$\bar{s}$	1/2	-1/3	$96^{+8}_{-4}$
3	top	t	$\bar{t}$	1/2	2/3	$172100 \pm 600$
	bottom	b	$\bar{b}$	1/2	-1/3	$4180^{+40}_{-30}$
Leptons						
1	electron	$e^-$	$\bar{e}^-$	1/2	-1	0.511
	electron neutrino	$\nu_e$	$\bar{\nu}_e$	1/2	0	< 0.0000022
2	muon	$\mu^-$	$\bar{\mu}^-$	1/2	-1	105.7
	muon neutrino	$\nu_\mu$	$\bar{\nu}_\mu$	1/2	0	< 0.170
3	tau	$\tau^-$	$\bar{\tau}$	1/2	-1	$1776.86 \pm 0.12$
	tau neutrino	$\nu_\tau$	$\bar{\nu}_\tau$	1/2	0	< 15.5

Another mystery regarding the neutrinos is in relation to anti particles. To each particle corresponds an antiparticle, i.e for  $l^- \rightarrow l^+$ ,  $q \rightarrow \bar{q}$ , where the anti particle and particle are different. With neutrinos however it is not known whether or not the neutrinos and antineutrinos are the same particle  $\nu = \bar{\nu}$  (Majorana neutrino), or if they are different  $\nu \neq \bar{\nu}$  (Dirac neutrino).

Quarks are fractional charge particles, with defined charge of either 2/3 or -1/3, as shown in table ???. They are the "main" building blocks of protons and neutrons, and are bound by the strong force, the

strongest of the four fundamental forces. The force mediator is the gluon. The other half of fermions are the leptons. They are split into the charged leptons (electrons, muons and taus), and the uncharged leptons (neutrinos). The charged leptons can interact via the electroweak force, where the Z, W bosons as well as the photon can be a mediator. Note that the neutrinos can only interact through the weak force.

## Bosons

Bosons are integer number spin particles, with spin 0, 1, 2, .... Within bosons there are so called elementary bosons, some of which are force carriers or mediators such as the  $W^\pm$ , Z and the photon. The Higgs boson is also an elementary scalar boson, but is not a force carrier. It provides masses for the fermions via a process called spontaneous symmetry breaking[? ]. Other bosons are so called composite bosons, mesons ( $q\bar{q}$ ) and baryons ( $qqq$ ) which are particles constructed by even or half integer spin. Bosons also have antiparticles, where  $(\gamma, g, H^0, Z^0)$  are equal to their respective antiparticle, and the  $W^- \rightarrow W^+$  are not equal.

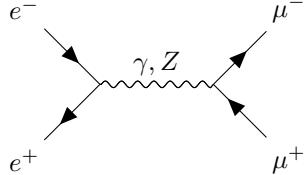
## Left and right handedness

Particles in the standard model are subject to a quantum mechanical property called chirality. Chirality is a property that describes a particle's ability to be superimposed on its mirror image. If a particle has chirality, it cannot be superimposed on its mirror image by any combination of translation, rotation, and reflection operations. [? ]

## Feynman diagrams

A graphical way to understand particle interactions are through so called Feynman diagrams. Feynman diagrams are drawn based on the Feynman rules for a given Lagrangian[? ][? ], and each component can be linked to a part in the Lagrangian for the system.

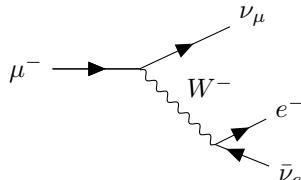
**Figure 2.2:** Feynman diagram of muon pair production from electron scattering. Here, both Z and  $\gamma$  can work as the propagator.



In figure ?? we have a Feynman diagram describing electron-positron annihilation into muon-antimuon pairs. In this thesis, all diagrams will be interpreted from left to right, i.e figure ???. The diagram contains all the components in the Lagrangian, and arrows, curly lines and so on all have its own meaning. A straight line with an arrow usually means a fermion, where the direction of the arrow tells if the particle (arrow towards the vertex) is an anti particle (arrow away from the vertex). There is often also a propagator between the left and right side of the Feynman diagram, and they depend on the processes we want to study. In the diagram above we have lepton scattering, thus we can both have the photon and the Z-boson as a propagator. This process is called a neutral current[? ], as the total charge coming out of the interaction is 0. As with neutral currents, we also have so called charged currents, where the sum of charge is not 0. Note that we only require charge conservation, thus there is nothing wrong with either having a neutral or a charged current, as long as charge conservation is preserved.

Feynman diagrams is not only used for visualizing scatterings, they can also visualize decays. An example is provided in figure ??.

**Figure 2.3:** Muon decay into an electron, an electron neutrino and a muon neutrino via the  $W^-$  boson. Read the graph from left to right.



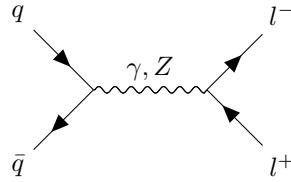
Here we have a decay of a muon into an electron and two neutrinos, through a charged current.

The examples above in figure ?? and ?? show interactions with the electroweak force, but along with the electroweak interactions, are also the quantum chromodynamics (QCD), responsible for interactions between quarks and gluons. A strange property of QCD, is that the coupling constant  $\alpha_S$ , unlike the  $\alpha_{EM}$  for electromagnetism, gets stronger as the energy decreases. This is because QCD (and weak interactions) are based on non-abelian groups[? ], thus to study such interactions, one needs to create collisions at very high energies.

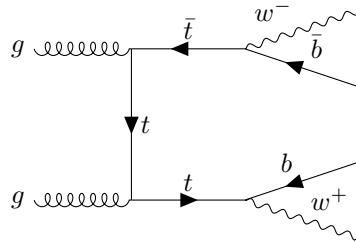
### Proton-proton collisions

Proton-proton collisions require high energy to collide. During an interaction, both the quarks and gluons in the protons colliding can interact in proton-proton collisions, as shown below:

**Figure 2.4:** Proton-proton collision with lepton pair production via the Z boson or photon. Read from left to right.



**Figure 2.5:** Proton-proton collision showing the  $t\bar{t}$  channel. Read from left to right



In figure ?? and figure ?? we have two examples of Feynman diagrams for possible interactions in proton-proton collisions. The first figure displays a lepton pair production via the Z boson or photon from quark-antiquark annihilation, and the second figure displays the  $t\bar{t}$  production via gluon-gluon fusion.

### Some limitations

All though the standard model has had great success comparing with experiments, there are still several problems not addressed by it. One example is gravity, the standard model as described above, does not and cannot incorporate gravity in a quantized way. There are models that try, without success so far, to address this problem, but they supplement the standard model, and does not derive it from it.

The problem that will be addressed in this thesis is a curious property of the weak interaction, namely that parity is broken. Parity as a mathematical operation is equivalent to the spatial inversion through the origin[? ]:

$$x \rightarrow -x. \quad (2.1)$$

In other words, parity can be thought of as left-right symmetry, or mirror symmetry. Breaking of parity is observed in weak currents, where the mediator of the charged currents,  $W^\pm$  only interacts with left handed fermions and right handed antifermions. In the standard model, neutrinos are assumed to be massless, and the righthanded neutrinos are sterile, i.e they do not interact in the standard model.

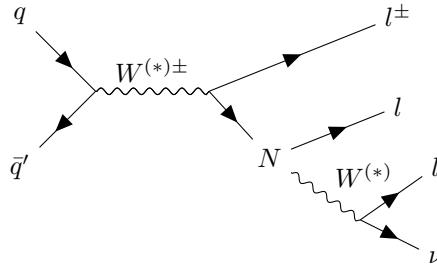
This asymmetry is strange, and hints towards new physics that perhaps can restore the parity breaking at much higher energies. Another note to make is that it has been experimentally verified that the neutrinos are massive[? ], with an upper limit on the mass for the anti electron neutrino of  $m_\nu < 0.8 eVc^2$  at 90% confidence level. This is somewhat problematic, as the tiny masses of the neutrinos are not predicted by the standard model.

## 2.2 BSM model physics

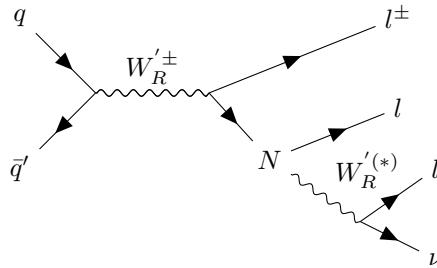
### Heavy neutrinos

Parity symmetry is suggested to be restored at high energies by the introduction of a right handed weak symmetry, leading to right handed weak charged bosons,  $W_R'^{\pm}$ . These mediators, as with the rest, decays faster than the detection ability at the LHC<sup>6</sup>, thus evidence of such a boson would come from the detection of the decay of a heavy neutrino. Heavy neutrinos can be produced in proton proton collisions through either the right handed  $W_R'^{\pm}$  bosons or the SM  $W^{\pm}$  bosons, under given conditions.

**Figure 2.6:** Proton-proton collision with heavy neutrino production via SM  $W^{\pm}$  boson into 3 lepton final state. Read the graph from left to right.



**Figure 2.7:** Proton-proton collision with heavy neutrino production via right-handed  $W_R^{\pm}$  boson into 3 lepton final state. Read the graph from left to right.



Note that the right most  $W$  and  $W_R$  bosons have an asterix as a superfix. This is because the bosons can be virtual. This means that if the mass of the heavy neutrino is less than the  $W$  boson, it cannot produce it, but it can produce a virtual one such that the decay still happens.

### Supersymmetry

Supersymmetry is another BSM theory that attempts to solve two other problems that the standard model has. First, the hierarchy problem . As the standard model is a perturbative theory, the Higgs mass increases at higher energies. The problem is that when you approach higher and higher energies, this mass goes to infinity, which is not physical. Supersymmetry solves this problem. The theory introduces a supersymmetric partner to each particle in the standard model. The result is that the contributions to the Higgs mass from fermions and bosons mainly cancel, thus fixing the hierarchy problem. Another problem we have with the standard model is that it does not have a field for dark matter. Some supersymmetry models have a dark matter candidate, thus the Supersymmetry search is quite large at CERN.

---

<sup>6</sup>Large Hadron Collider at CERN



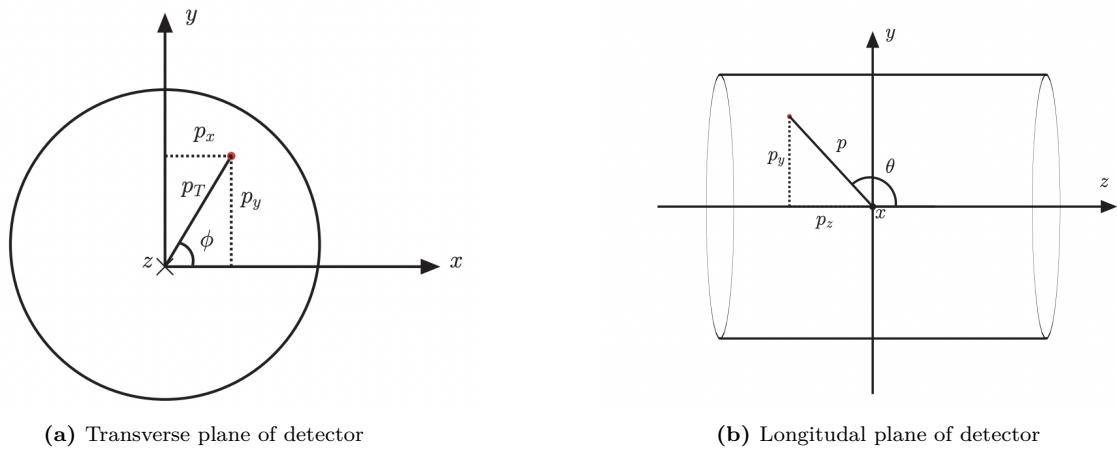
# Chapter 3

## Implementation

### 3.1 The ATLAS detector

#### Kinematics and detector geometry

Before one can analyse the data, it has to be collected and processed in a detector. The data used in this thesis are generated from proton-proton collisions in the ATLAS detector at the LHC. The ATLAS inner detector itself is a solenoid, and the kinematic variables are measured based on the following coordinate system. The z-axis is defined to go along the center axis of the solenoid, where as the y axis points upwards in the detector and the x axis radially outwards from the center axis. This allows for all transverse variables to be defined in the x-y plane[? ]. From this we construct the azimuthal and polar angles  $\phi$  and  $\theta$ , where the azimuthal angle  $\phi[? ]$  is the angle around the z-axis, and the polar angle  $\theta$  is the angle from the z-axis, as shown in figure ??.



**Figure 3.1:** Spherical coordinate definitions with the azimuthal and polar angles  $\phi$  and  $\theta$ . Here figure ?? of the transverse plane shows the z-axis into the paper, whereas figure ?? of the longitudinal plane shows the positive x-axis going out of the paper. Source [here](#). Accessed 16.03.23

In figure ?? the kinematic variables for proton-proton collisions are described by the energy, rest mass and momentum, given as  $E$ ,  $m$ , and  $\mathbf{p} = (p_x, p_y, p_z)$  respectively. As the particles move with very high energy, we will use the relativistic 4 momentum, given as  $\mathbf{P} = (E, \mathbf{p})$ . We also have that

$$\gamma = \frac{1}{\sqrt{1 - \beta^2}},$$

where  $\gamma$  is the Lorentz factor, and  $\beta = \frac{v}{c}$ , which gives us the following definitions for energy  $E = \gamma m$  and momentum  $\mathbf{p} = \beta \gamma m[? ]$ . From this we can derive the energy momentum formula:

$$\begin{aligned}
\mathbf{p}^2 &= \beta^2 \gamma^2 m^2 \\
\mathbf{p}^2 + m^2 &= m^2 (\beta^2 \gamma^2 + 1) \\
\mathbf{p}^2 + m^2 &= m^2 \gamma^2 \\
\mathbf{p}^2 + m^2 &= E^2
\end{aligned}$$

$$E = \sqrt{p^2 + m^2}. \quad (3.1)$$

It can be shown that the phase space of a particle is given by[? ]:

$$d\mathbf{p} = dp_x dp_y dp_z = p^2 dp d\Omega = dp_z p_T dp_T d\phi, \quad (3.2)$$

where  $p_z$  is the momentum along the beam direction,  $p_T$  is the projected momentum on the transverse plane, and  $\Omega$  is the solid angle. An analog to the relativistic longitudinal velocity is the rapidity  $y$ . To define this we have that the relativistic generalization of equation ?? is given by:

$$d^4 p \delta(E^2 - p^2 - m^2) = d\mathbf{p} \frac{1}{E} = p_T dp_T d\phi dy, \quad dy = \frac{dp_z}{E}.$$

Using the fact that  $p = \sqrt{p_T^2 + p_z^2}$  and equation ?? we can integrate  $dy$  to get the rapidity:

$$\begin{aligned}
\int dy &= \int \frac{dp_z}{\sqrt{p_T^2 + p_z^2 + m^2}} \\
y &= \cosh^{-1} \left( \frac{E}{\sqrt{p_T^2 + m^2}} \right)
\end{aligned} \quad (3.3)$$

For particles with little to no mass relative to the transverse momentum, we have that  $p_T^2 + m^2 \approx p_T^2$  where  $p_T = E \sin(\theta)$ , which gives us the following relations:

$$\begin{aligned}
\cosh(y) &= \frac{1}{\sin(\theta)}, \\
\sinh(y) &= \frac{1}{\tan(\theta)}, \\
\tanh(y) &= \cos(\theta).
\end{aligned}$$

which can be used to show that  $e^{-y} = \tan \frac{\theta}{2}$ . From this we define the pseudorapidity  $\eta$  as:

$$\eta = -\ln \left( \tan \frac{\theta}{2} \right), \quad (3.4)$$

which is, in the relativistic limit, the same as the rapidity  $y$ . A useful property of the pseudorapidity is that the phase space of a single particle is uniformly distributed for both  $\eta$  and  $\phi$ , making them good features to compare overlap between SM MC and ATLAS data[? ].

## Data collection



**Figure 3.2:** Figure describing how particles are detected at ATLAS, fetched from [ATLAS detector slice \(and particle visualisations\)](#), by Sascha Mehlhase [?].

The features used in this analysis are computed with or fetched from the features from the detector itself. Such features include the momentum, energy, angles etc, all of which are either directly measured or computed based on the measurements in the detector. In figure ?? a visualization shows how different particles move through the detector and how they are detected. For example, energy deposits are measured using calorimeters, and the different particles have calorimeters specially designed for them. Charged particles leaves tracks in the inner tracking device. Thus, electrons leaves tracks in the inner tracker and deposits energy in the electromagnetic calorimeter, the muons leaves tracks in the inner detector and deposits energy in the muon detector. The photons only deposits energy in the electromagnetic calorimeter. Hadrons like protons and neutrons deposits energy in the electromagnetic and hadronic calorimeters, and charged hadrons also leaves tracks in the inner tracker. This is shown in figure ??.

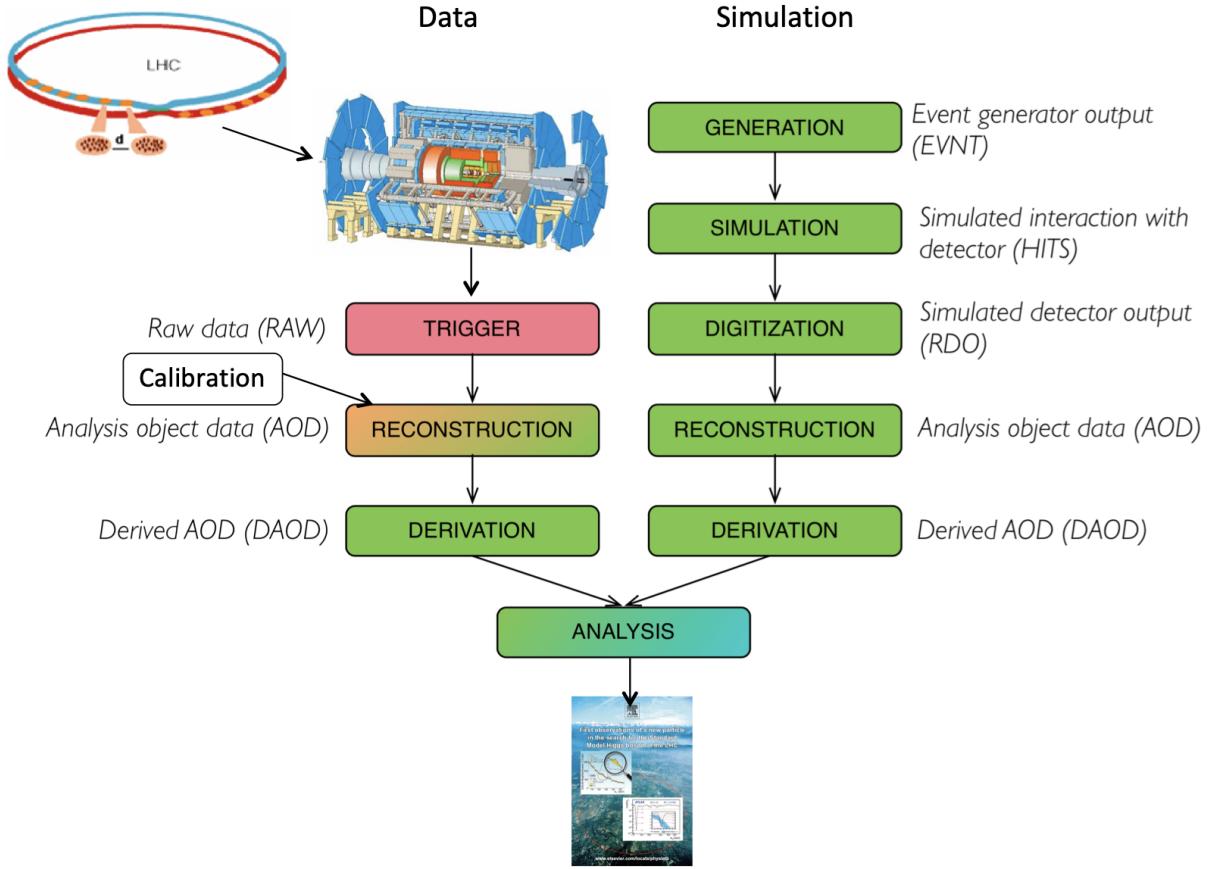
The ATLAS detector have a few thousand selection stages before the data is stored. In order to reach the highest intensity of collisions, the LHC accelerates packets of around  $10^{11}$  protons, and collides them at a rate of 25 nanoseconds, yeilding a collision rate of 40 MHz[?]. [?]

## Data preparation

During datagathering at the ATLAS detector, triggers on the hardware and software level select out the events that are of most interest. On the order of 99% or more of the recorded events are discarded, with about 1 in 40000 events being accepted, as the amount of recorded events simply as too high to realistically analyse. Also, a lot of the events are of no interest for new physics analysis anyways as they for example might have to little energy. Once the trigger selection is done, the data is reconstructed. This means that the objects in the recorded events are through software algorithms reconstructed into particles, jets, photons etc. The reconstruction is done based on the tracks and measurements in the detector, but it is not perfect, and can lead to fake leptons or jets. By fake it is meant that an object might look like a lepton but is in reality a jet or vice versa[?]. Once the reconstruction is done, further slimming of the n-tuples are done. Derivations are slimmings of the n-tuples where the selection of events are further reduced to match the needs of the different analysis groups.

The SM MC go through parts of the same process. These events are first generated and then run through the detector to simulate actual events. Once that is done, the events can be reconstructed and go through derivations just like the pp-collision data, to be used in analysis.

## Steps from Data Collection to Physics Results



**Figure 3.3:** Figure describing the steps to take for data collection at ATLAS, fetched from [Hybrid ATLAS Induction Day + Software Tutorial workshop](#), part [Computing and Data preparation](#), held by S.M Wang [?].

### Jets

Photons and electrons are detected in the electromagnetic calorimeter, and are easy to track and detect as they separate easily. Quarks, however, are bound by QCD and thus cannot be separated as individual particles. An illustration of how quarks and gluons materialise as jets during a proton-proton collision is shown below in figure ??.



**Figure 3.4:** Figure describing how quarks and gluons are treated in the detector, and thus why we name them jets, fetched from [the CMS webpage](#).

In a proton-proton collision, the quarks and gluons forms stable or unstable hadrons such that the color confinement<sup>7</sup> is upheld. These then decay to other stable hadrons that can be tracked, and these tracks are called jets. This is particularly difficult because one wants to isolate which hadrons came from the original quark in the Feynman diagram. Another point to make is that some quarks are of higher interest than others. For example, the b jet, coming from a b quark, is a good indicator for certain processes, thus identifying suchs particles is of huge interest.

## 3.2 ROOT

### N tuples

The main datastructure of ROOT is the so called N tuple structure. This datastructure contains each property for each type of particle in a given event, yeilding a ragged structure.

	$jetP_T$	$jetPhi$	$lepP_T$	$lepPhi$	Rowlength
0	[120.2, 57]	[1.2, 0.5]	[223.3, 57.5, 9.7]	[0.545, 0.2, -0.3]	10
1	[, ]	[, ]	[121.343, 89.323]	[0.886, -0.855]	4
2	[86.112]	[86.112]	[57.75, 34.5]	[0.33, 0.255]	6

### RDataFrame

[? ]

RDataFrame's main purpose is to make reading and handling of root files easier, especially in relation to modern machine learning tools and their respective frameworks and environments. This is done by creating a dataframe like structure of the root n-tuples, and then lazily<sup>8</sup> apply contraints to the data. Using PyROOT, RDataFrame can be accessed in Python, as the functionality is wrapped around a C++ class. Below is an example of how to create a RDataFrame object, apply a cut and then create a column for later use. Here, good leptons are defined first, denoted as "ele\_SG" and "muo\_SG". A cut is then applied where we require that the number of good leptons is always 3. Finally, a column is created where the combination of type of leptons in the 3 lepton system is stored, as well as creating a histogram containing the results for that given channel<sup>9</sup> k. Notice here that if the variable already exist as a column in the dataframe, arithmetic and logic can be applied directly using those columns to create new one. More complicated variables, such as the flavor combination for the leptons, or the invariant mass of two particles must be found or calculated using C++ functions. An example of such a function is shown below the python code listing.

---

<sup>7</sup>Add link to a source or explanation for this.

<sup>8</sup>In this context lazily means that the functions and or cuts are done first after all have been registered, see [ROOT guidelines](#) for more.

<sup>9</sup>A channel here refers to a certain decay channel. The standard model has several, and some look more alike than others. One example is the Higgs decay channel, with possible decays such as two photons, W bosons or Z bosons.

```

1 import ROOT as R
2
3 R.EnableImplicitMT(200)
4 R.gROOT.ProcessLine(".L helperFunctions.cxx+")
5 R.gSystem.AddDynamicPath(str(dynamic_path))
6 R.gInterpreter.Declare(
7     '#include "helperFunctions.h"'
8 ) # Header with the definition of the myFilter function
9 R.gSystem.Load("helperFunctions_cxx.so") # Library with the myFilter function
10
11 df_mc = getDataFrames(mypath_mc)
12 df_data = getDataFrames(mypath_data)
13 df = {**df_mc, **df_data}
14
15 for k in df.keys():
16
17     # Signal leptons
18     df[k] = df[k].Define(
19         "ele_SG",
20         "ele_BL && lepIsoLoose_VarRad && lepTight && (lepDOSig <= 5 && lepDOSig >= -5)",
21     )
22     df[k] = df[k].Define(
23         "muo_SG",
24         "muo_BL && lepIsoLoose_VarRad && (lepDOSig <= 3 && lepDOSig >= -3)",
25     )
26     df[k] = df[k].Define("isGoodLep", "ele_SG || muo_SG")
27     df[k] = df[k].Define(
28         "nlep_SG", "ROOT::VecOps::Sum(ele_SG)+ROOT::VecOps::Sum(muo_SG)"
29     )
30
31     df[k] = df[k].Filter("nlep_SG == 3", "3 SG leptons")
32
33     # Define flavor combination based on
34     df[k] = df[k].Define("flcomp", "flavourComp3L(lepFlavor[ele_SG || muo_SG])")
35     histo[f"flcomp_{k}"] = df[k].Histo1D(
36         (
37             f"h_flcomp_{k}",
38             f"len(fldic.keys())",
39             len(fldic.keys()),
40             0,
41             len(fldic.keys()),
42         ),
43         "flcomp",
44         "wgt_SG",
45     )
46

```

In the code listing above we see an example of how RDataframe can be used for event selection. Line 1-9 are settings for ROOT, how many threads to use in the parallelization, extra helper functions written in C++ with .h and .so files an the path to the folder. The next three lines creates a dictionary containing the ROOT RDataFrames to do event selection on. These are categorized by channel name. The loop then does event selection for each channel sample, defining new variables in the RDataFrame, applying filters, and creating histograms. Some variables are constructed using variables already in the ROOT files, such as energy and mass and so on, which through custom C++ functions can be added. In the codelisting below we see a custom C++ function which is used in this thesis.

```

1 double getM(VecF_t &pt_i, VecF_t &eta_i, VecF_t &phi_i, VecF_t &e_i,
2             VecF_t &pt_j, VecF_t &eta_j, VecF_t &phi_j, VecF_t &e_j,
3             int i, int j)
4 {
5     /* Gets the invariant mass between two particles, be it jets or leptons */
6
7     const auto size_i = int(pt_i.size());
8     const auto size_j = int(pt_j.size());
9
10    if (size_i == 0 || size_j == 0){return 0.;}
11    if (i > size_i-1){return 0.;}
12    if (j > size_j-1){return 0.;}
13
14    TLorentzVector p1;
15    TLorentzVector p2;
16
17    p1.SetPtEtaPhiM(pt_i[i], eta_i[i], phi_i[i], e_i[i]);
18    p2.SetPtEtaPhiM(pt_j[j], eta_j[j], phi_j[j], e_j[j]);
19
20    double inv_mass = (p1 + p2).M();
21
22    return inv_mass;
23 }

```

The C++ function listed above creates Lorentzvectors for two particles, and then constructs the invariant mass based on the parameters sent in. this function will be used on all the leptons in a given event, and in the case that one particle or both do not exist, the C++ function will return 0 as the invariant mass.

```

1 import pandas as pd
2
3 cols = df.keys()
4
5 for k in cols:
6
7     print(f"Transforming {k}.ROOT to numpy")
8     numpy = df[k].AsNumpy(DATAFRAME_COLS)
9     print(f"Numpy conversion done for {k}.ROOT")
10    df1 = pd.DataFrame(data=numpy)
11    print(f"Transformation done")
12
13
14    df1.to_hdf(
15        PATH_TO_STORE + f"/{k}_3lep_df_forML_bkg_signal_fromRDF.hdf5", "mini"
16    )

```

Once eventselection is done, the features have been chosen and histograms have been drawn, the Rdataframe can be converted to a pandas dataframe, which is a very popular choice for data structure when doing data analysis in python. This is done through an intermediary step of converting the RDataframe to a numpy filestructure, which then can be converted to a pandas[?] ] dataframe, or some other framwork. Here the new pandas dataframe is stored as hdf5[?] files to be used later, as the hdf5 format has a very good compression ratio, and is very fast to read and write.

### 3.3 Background samples

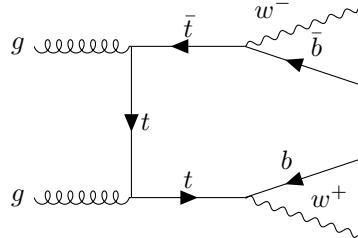
To look for anomalies in the three lepton final state we need to train on background MonteCarlo samples with that specific final state. This means in a sense that we want the autoencoder to learn what is expected from the Standard model in terms of this final state. The 2 and 3 lepton +  $e_T^{miss}$  background MonteCarlo contains the following channels:

**Table 3.1:** SM MC channels for both the 2 and 3 lepton +  $e_T^{miss}$  final state background.

3 lepton + $e_T^{miss}$	2 lepton + $e_T^{miss}$
Channel names:	
Wjets	Wjets
ttbar	ttbar
Singletop	Singletop
ZeeJets	ZeeJets
ZmmJets	ZmmJets
ZttJets	ZttJets
Higgs	Diboson
Triboson	
TopOther	
Diboson2L	
Diboson3L	
Diboson4L	

Below are three Feynman diagrams, two of which are both represented in the 2 and 3 lepton +  $e_T^{miss}$  background MonteCarlo shown in table ???. The selected ones are likely Feynman diagrams for  $t\bar{t}$ , Higgs and Zeejets channels.

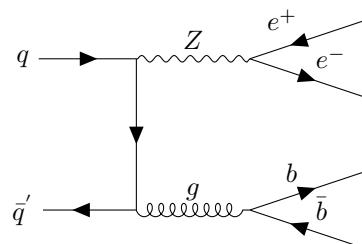
**Figure 3.5:** Proton-proton collision showing the  $t\bar{t}$  channel. Here the w bosons decay leptonically and one or more jets can be misreconstructed as fake leptons by the detector.



**Figure 3.6:** Proton-proton collision showing the Higgs channel. Here the Z bosons decay leptonically.



**Figure 3.7:** Proton-proton collision showing the Zeejets channel. Here one of the Z bosons decay leptonically and the gluon decays hadronically.



### 3.4 The dataset features

#### The Rapidity-Mass matrix (RMM)

Most of the features in the analysis are elements in the so called Rapidity-Mass matrix (RMM) inspired by the work of Chekanov [? ].

The RMM is a convenient structure to create a feature space for the dataset. It contains information of various reconstructed objects and their combinations about mass, rapidity, momenta and missing transverse energy, all of which are useful in searches for new physics[? ] in HEP. One example of an analysis that have used some of the features from the RMM is demonstrated in [? ]. The main reason however for using this structure is the systematic layout and automated featurespace, that maintains low to no correlation between the cells in the matrix, as this is ideal when using neural networks.

Its composition is determined as a square matrix of  $1 + \sum_{i=1}^T N_i$  columns and rows, where T is the total number of objects (i.e jets, electrons etc.), and  $N_i$  is the multiplicity of a given object. In the case of the same number of a given object for all objects, we can denote the RMM matrix as a TmNn matrix, where m is the multiplicity of T, and n is the number of particle per type. Thus there is already room for evaluation, as the combination of number of objects and the number of each object type highly affects the analysis as well as computational resources. Each cell in the matrix contains information about either single or two particle properties. This could in principle be generalized to three particle properties, which would make a three dimensional RMM. The scope of thesis will for simplicity only cover the two dimensional case. An example is shown in matrix ??.

$$\begin{pmatrix} e_T^{miss} & m_T(j_1) & m_T(j_2) & m_T(e_1) & m_T(e_2) \\ h_L(j_1) & e_T(j_1) & m(j_1, j_2) & m(j_1, e_1) & m(j_1, e_2) \\ h_L(j_2) & h(j_2, j_1) & \delta e_T(j_2) & m(j_2, e_1) & m(j_2, e_2) \\ h_L(e_1) & h(e_1, j_1) & h(e_1, j_2) & e_T(e_1) & m(e_1, e_2) \\ h_L(e_2) & h(e_2, j_1) & h(e_2, j_2) & h(e_2, e_1) & \delta e_T(e_2) \end{pmatrix} \quad (3.5)$$

In matrix ?? we have the RMM matrix for a T2N2 system, in other words we have two types of particles, jets <sup>10</sup> and electrons, where each type has two particles. The matrix itself is partitioned into three parts. The diagonal represents energy properties, the upper triangular represents mass properties, and the lower triangular represents longitudinal properties related to rapidity. The diagonal has three different properties,  $e_T^{miss}$ ,  $e_T$  and  $\delta e_T$ .  $e_T^{miss}$  is placed in the (0, 0) position in the matrix. It accounts for the missing transverse energy for the system, which is of high interest for this analysis due to the search for heavy neutrinos.  $e_T$  is the transverse energy defined as

$$e_T = \sqrt{m^2 + p_T^2} \quad (3.6)$$

but for light particles such as electrons, this can be approximated to  $e_T \approx p_T$ .  $\delta e_T$  is the transverse energy imbalance. It is defined as

$$\delta e_T = \frac{E_T(i_n - 1) - E_T(i_n)}{E_T(i_n - 1) + E_T(i_n)}, \quad n = 2, \dots, N. \quad (3.7)$$

The first column in the RMM matrix, with the exception of the first element, is related to the longitudinal property of the given particle. It is defined as

$$h_L(i_n) = C(\cosh(y) - 1),$$

where C is a constant to ensure that the average  $h_L(i_n)$  values do not deviate too much from the ranges of the invariant masses of the transverse masses, found to be 0.15, as it ensures that rapidity ranges in the range [-2.5, 2.5] produces  $h_L(i_n)$  values in the [0, 1] interval[? ]. y is the rapidity of the particle, and  $i_n$  is the particle number. On the lower triangle we have the longitudinal properties of the combinations of particles. Similar to  $h_L(i_n)$ , this property is defined as

$$h(i_n, j_k) = C(\cosh(\Delta y) - 1),$$

where  $\Delta y = y_{i_n} - y_{j_k}$  is the rapidity difference for particle  $i_n$  and  $j_k$ .

---

<sup>10</sup>Jets here can both be b- or ljets. Ljet is defined as jets with jetdl1r < 0.665, where as bjet77 is defined as jets with jetdl1r >= 2.195, where jetdl1r is a machine learning output from a network trained to distinguish b- and ljets.

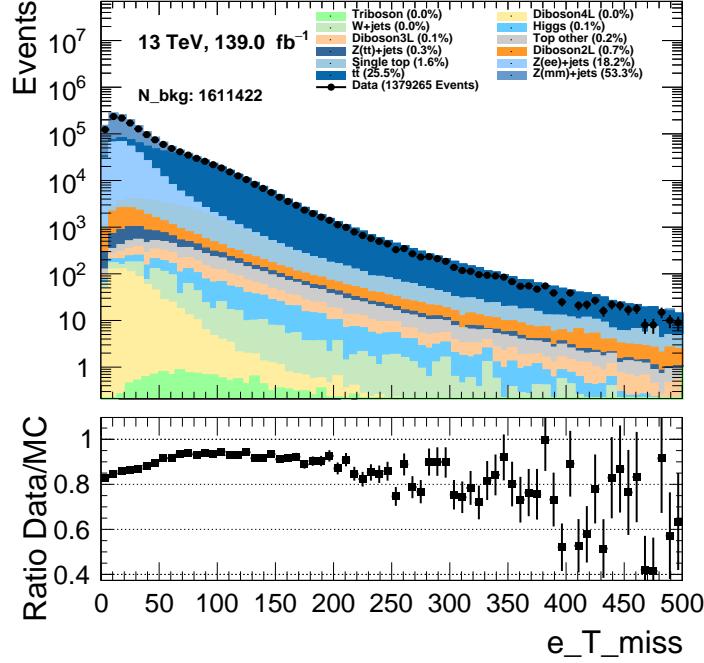
## Tabular and sparse data

A consequence of using the RMM structure is that the data and Monte Carlo are sparse. This is due to the fact that the RMM allows for the variety of final states of the reconstructed events, i.e that one event has two ljets, zero bjets, one electron and two muons, whereas another event can have 4 ljet, 3 bjets and three electrons. This means that the RMM matrix for each event will have a different size, and for neural networks this is a problem. To solve this problem, Chekanov simply pads the missing values with 0s[? ].

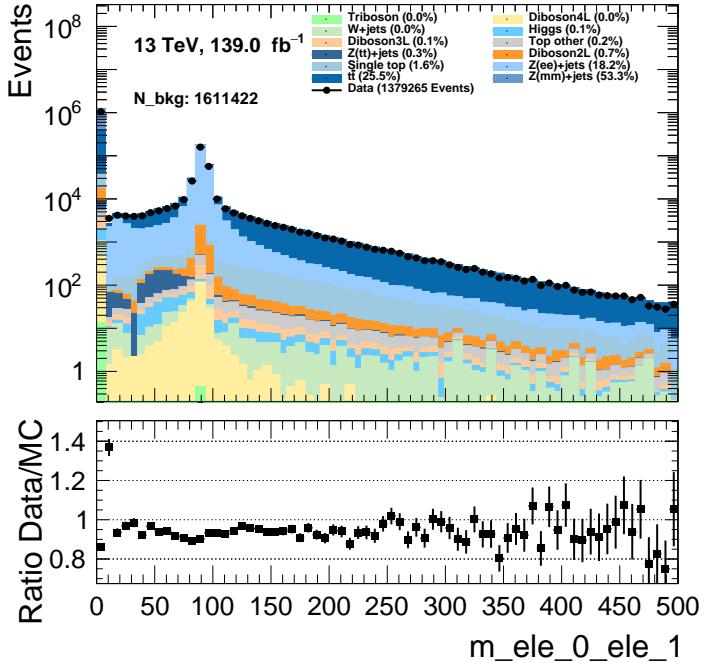
## MonteCarlo and data comparison

Before we can start the analysis, we need to compare the MonteCarlo and data. This is done to ensure that the measured features used are reconstructed well by the MonteCarlo training samples we use. As described by R. Stuart Geiger et al. [? ], the concept of "Garbage in, garbage out" is of key importance in computer science, and indeed important in high energy physics. To ensure that the models we train actually learn physical processes, the training set must represent the physics "status quo". If the training samples do not match the physical reality, we regard it, in the context of high energy physics, as garbage in, which will in turn give garbage out. The Monte Carlo standard model simulations are indeed very good, but they are numerical approximations, and can sometimes be off. Thus, every feature that will be used for training have to be checked before being used. This is done by comparing the distributions of the features in the MonteCarlo and ATLAS data. MonteCarlo simulations are based on the actual theory itself, and comparisons with data taken from ATLAS and other detectors alike are necessary to prove that the standard model is a good model.

Now, if we compare all SM MonteCarlo and ATLAS data, we would usually expect there to be a good overlap. To ensure that the standard model MonteCarlo actually represents the physics, we create signal and background regions to optimize for a signal and or background. If we can create a background region where we believe with very high certainty that only standard model processes can occur, and we get a good match, we usually conclude that the MonteCarlo is good enough. Now, for this thesis, simply comparing all ATLAS data to all standard model MonteCarlo is enough, as this data batch has been analysed by the ATLAS collaboration for multiple years without finding any new physics, concluding that if the signals are there, they are too small for so called visual cuts. Traditional searches have only excluded some models, which is why machine learning is getting more popular. The hope is that the signal, whatever it might be, can be revealed with clever feature engineering and smart machine learning algorithms. Particle physics differs here from more day to day machine learning as the target data is unlabeled.



(a) Missing transverse energy for the three lepton final state in GeV. The histogram contains the entire Run 2 dataset.



(b) Invariant mass for the first and second electron in GeV. The histogram contains the entire Run 2 dataset.

**Figure 3.8:** Comparison of the MonteCarlo and data for the three lepton +  $e_T^{miss}$  final state with the features  $e_T^{miss}$  and flavor composition.

In figure ?? two features have been selected to visualize the comparison between Monte Carlo and ATLAS data,  $e_T^{miss}$  and  $m(ele_0, ele_1)$  in the 3 lepton +  $e_T^{miss}$  dataset. We see that both  $e_T^{miss}$  and  $m(ele_0, ele_1)$  satisfy a good ratio between Monte Carlo and ATLAS data, thus we can safely move forward with the analysis. All

features were checked, and can be found in the Github repository for this thesis at [Figures/Histo\\_var\\_check<sup>11</sup>](#) under the 3lep folder.



(a) Missing transverse energy for the three lepton final state in GeV. The histogram contains the entire Run 2 dataset. Note here the lack of good overlap from about 200-500 GeV.



(b) Invariant mass for the first and second electron in GeV. The histogram contains the entire Run 2 dataset. Note the sharp reduction of events from 10-70 GeV.

**Figure 3.9:** Comparison of the MonteCarlo and data for the 2 lepton +  $e_T^{miss}$  final state with the features  $e_T^{miss}$  and flavor composition.

<sup>11</sup>Full link: [https://github.com/Gadangadang/MasterThesis/tree/main/Figures/histo\\_var\\_check/LEP3](https://github.com/Gadangadang/MasterThesis/tree/main/Figures/histo_var_check/LEP3)

The same checks were done for the 2 lepton +  $e_T^{miss}$  dataset. In figure ?? we see that the features  $e_T^{miss}$  and  $m(e_{l0}, e_{l1})$  do not satisfy the same ratio between SM MC and ATLAS data, as with the 3 lepton +  $e_T^{miss}$  case. It is suspected that the trigger matching and triggers are not implemented properly, but neither the author or the supervisors have found a solution to the issue, and due to time constraints, chose to move forward with the analysis regardless. Thus, for all results regarding the 2 lepton +  $e_T^{miss}$  dataset should be interpreted with this in mind. One attempt can be shown in figure ?? where a cut on the invariant mass above 70 GeV of at least the two leptons with the highest energy, to try to accommodate the trigger issue. It did not seem to work. All features were checked, and can be found in the Github repository for this thesis at [Figures/Histo\\_var\\_check<sup>12</sup>](#) under the 2lep folder.

### Triggers

The following triggers where implemented for the 2 lepton +  $e_T^{miss}$  and 3 lepton +  $e_T^{miss}$  dataset are written below:

**Table 3.2:** Triggers used in the 2015 MonteCarlo and ATLAS data samples for the 2 lepton 0  $e_T^{miss}$  dataset.

	Name	Description
2015	<i>HLT_2e15_lhvloose_nod0_L12EM13VH</i>	
	<i>HLT_2e12_lhloose_L12EM10VH</i>	
	<i>HLT_2mu10</i>	
	<i>HLT_mu18_mu8noL1</i>	
	<i>HLT_e17_lhloose_mu14</i>	
	<i>HLT_e7_lhmedium_mu24</i>	

**Table 3.3:** Triggers used in the 2016 MonteCarlo and ATLAS data samples for the 2 lepton 0  $e_T^{miss}$  dataset.

	Name	Description
2016	<i>HLT_2e15_lhvloose_nod0_L12EM13VH</i>	
	<i>HLT_2e17_lhvloose_nod0</i>	
	<i>HLT_2mu10</i>	
	<i>HLT_2mu14</i>	
	<i>HLT_mu20_mu8noL1</i>	
	<i>HLT_mu22_mu8noL1</i>	
	<i>HLT_e17_lhloose_nod0_mu14</i>	
	<i>HLT_e24_lhmedium_nod0_L1EM20VHI_mu8noL1</i>	
	<i>HLT_e7_lhmedium_nod0_mu24</i>	

**Table 3.4:** Triggers used in the 2017 MonteCarlo and ATLAS data samples for the 2 lepton 0  $e_T^{miss}$  dataset.

	Name	Description
2017	<i>HLT_2e17_lhvloose_nod0_L12EM15VHI</i>	
	<i>HLT_2e24_lhvloose_nod0</i>	
	<i>HLT_2mu14</i>	
	<i>HLT_mu22_mu8noL1</i>	
	<i>HLT_e17_lhloose_nod0_mu14</i>	
	<i>HLT_e26_lhmedium_nod0_mu8noL1</i>	
	<i>HLT_e7_lhmedium_nod0_mu24</i>	

<sup>12</sup>Full link: [https://github.com/Gadangadang/MasterThesis/tree/main/main/Tables/histo\\_var\\_check/LEP2](https://github.com/Gadangadang/MasterThesis/tree/main/main/Tables/histo_var_check/LEP2)

**Table 3.5:** Triggers used in the 2018 MonteCarlo and ATLAS data samples for the 2 lepton  $0 e_T^{miss}$  dataset.

	Name	Description
2018	<i>HLT_2e17_lhvloose_nod0_L12EM15VHI</i>	
	<i>HLT_2e24_lhvloose_nod0</i>	
	<i>HLT_2mu14</i>	
	<i>HLT_mu22_mu8noL1</i>	
	<i>HLT_e17_lhloose_nod0_mu14</i>	
	<i>HLT_e26_lhmedium_nod0_mu8noL1</i>	
	<i>HLT_e7_lhmedium_nod0_mu24</i>	

In tables ??, ??, ??, ?? we have the triggers used for the different data generations used for the 2 lepton +  $e_T^{miss}$ , i.e the data taken from the 4 years of data collection in Run 2. Each trigger corresponds a certain detection in the calorimeters and reconstruction. For example, *HLT\_2mu10* means that at least 2 muons with a  $p_T$  of atleast 10 GeV is detected.

For the 3 lepton +  $e_T^{miss}$  final state, the trigger system did not implement well in RDataframe, thus a simple cut of requiring at least two leptons with a  $p_T$  above 20 GeV was implemented. Note that further triggers would refine the dataset even more.

## 3.5 Code implementation

### Machine learning implementation

The machine learning analysis was written with Keras[?] using the Tensorflow api[?]. The machine learning structure was written using a functional structure<sup>13</sup>. In practise, this model could just as well have been written as a Sequential model<sup>14</sup>, but at a cost of flexibility and lack of potential non-linear structure in the architecture. The code consists of one general class for the autoencoder, where the different testing cases are different classes inheriting from the parent class.

### Construction of a neural network in Tensorflow

Using the functional structure, a general neural network in the Tensorflow API can be constructed as shown below.

```

1 import tensorflow as tf
2
3
4 inputs = tf.keras.layers.Input(shape=data_shape, name="input")
5
6 # First hidden layer
7 First_layer = tf.keras.layers.Dense(
8     units=30,
9     activation="relu"
10 )(inputs)
11
12 # Second hidden layer
13 Second_layer = tf.keras.layers.Dense(
14     units=45,
15     activation="relu"
16 )(First_layer)
17
18 # Output layer
19 output_layer = tf.keras.layers.Dense(
20     units=1,
21     activation="sigmoid"
22 )(Second_layer)
23
24
25 # Model definition
26 nn_model = tf.keras.Model(inputs, output_layer, name="nn_model")
27
28 hp_learning_rate = 0.0015
29 optimizer = tf.keras.optimizers.Adam(hp_learning_rate)
30 nn_model.compile(loss="mse", optimizer=optimizer, metrics=["mse"])

```

The neural network here contains one input layer, two hidden layers, and an output layer. The choice of nodes and activation functions are arbitrary here as the use case has not been defined. Note that this is exactly the same as the previous example, but using the sequential structure.

```

1 import tensorflow as tf
2
3 nn_model = tf.keras.Sequential(
4     [
5         tf.keras.layers.Dense(30, activation="relu", input_shape=data_shape),
6         tf.keras.layers.Dense(45, activation="relu"),
7         tf.keras.layers.Dense(1, activation="sigmoid"),
8     ]
9 )
10
11 hp_learning_rate = 0.0015
12 optimizer = tf.keras.optimizers.Adam(hp_learning_rate)
13 nn_model.compile(loss="mse", optimizer=optimizer, metrics=["mse"])

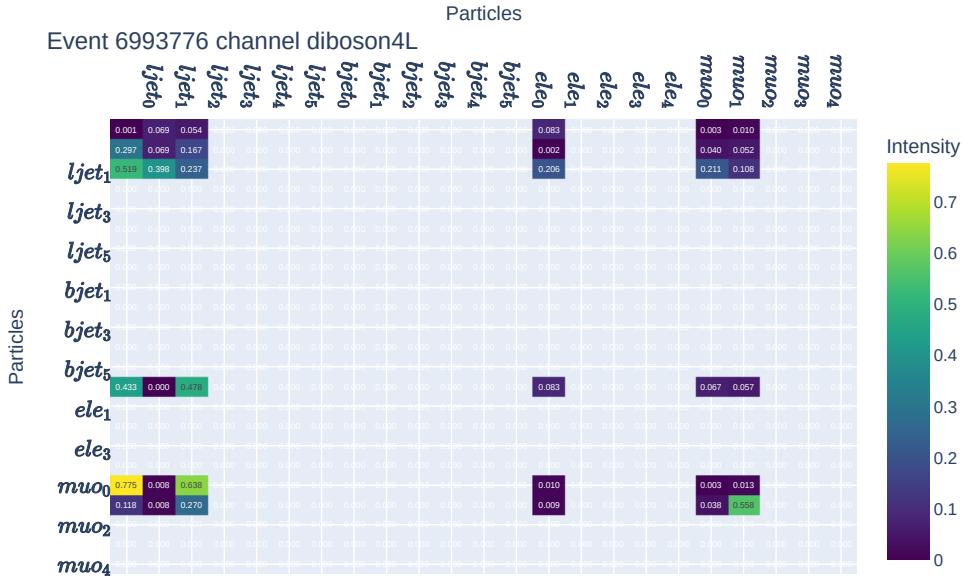
```

<sup>13</sup>Functional structure uses a function call for layers, i.e for layers a,b, then b(a) will connect the two layers, and equals a sequential link a → b. This allows for more flexible structures. More on the functional api can be found [here](#).

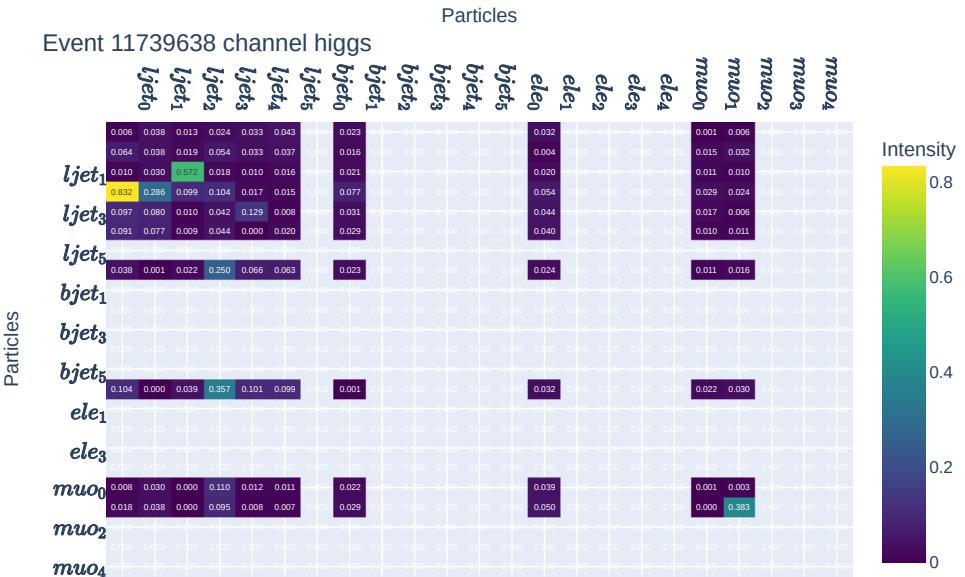
<sup>14</sup>Sequential structure adds layers in sequence, i.e for layers a, b, c we have that a → b → c, with a strict structure. This allows for more organized code. More on sequential models can be found [here](#).

**Data handling python side****Implementation of the RMM matrix**

An example of the RMM matrices used in this thesis is shown in figure ?? below:



(a) RMM matrix for event number 6993776 from the MonteCarlo diboson4L sample. Each feature is scaled based on a fit for that feature for all events in the training set ( $\approx 80\%$  of total MC). This sample contains two ljets, one electron and two muons.



(b) RMM matrix for event number 11739638 from the MonteCarlo Higgs sample. Each feature is scaled based on a fit for that feature for all events in the training set ( $\approx 80\%$  of total MC). This sample contains five ljets, one bjet, one electron and two muons.

**Figure 3.10:** Note here that the y axis for the RMM's lack every other label, due to lack of space in the y axis of the plot. If looked more closely upon, one can see that each figure have all RMM cells, just that the labels, which are identical to the x axis label, only show every other. Note also here that the labels only tell which particle are used for that row/column, i.e in figure ?? in row 0 column 3 we have the invariant mass of the first and second ljet. The RMM plots were created using Plotly[? ].

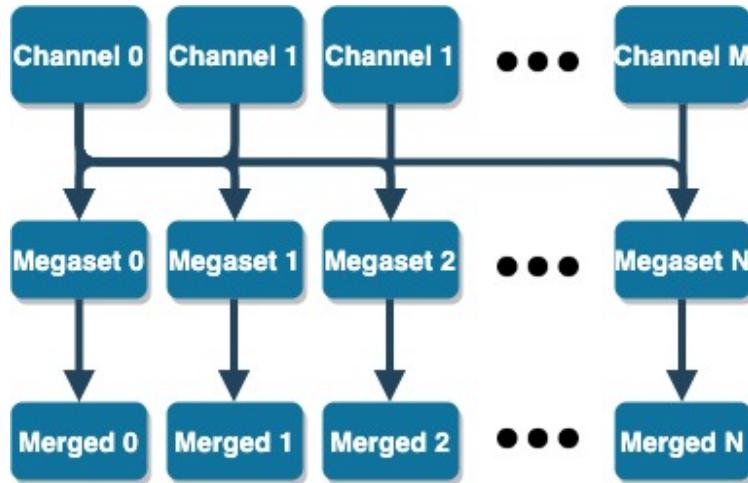
In figure ?? we see two RMM matrices created from two different channels in the MonteCarlo samples. This RMM is of type T4N5<sup>15</sup>. For easier interpretability, the gray area corresponds to a missing value, leading to so called "islands" in the RMM matrix.

## Setup for 3 lepton dataset

The 3 lepton dataset is about 96 Giga bytes of data when implementing the RMM structure for 6 b- and ljets, 5 electrons and 5 muons. The dataset is converted from a ROOT RDataframe to a Pandas dataframe<sup>[?]</sup> for further preprocessing. Having added the channel column in the dataset<sup>16</sup>, the channel categories, weights, missing transverse energy and trilepton mass<sup>17</sup> as well as the RMM structure are divided into a training and validation/test set in an 80-20 split. This was done using the `".fit_transform()"` and `".transform()"` functions from the Scikit-learn library<sup>[?]</sup>. The training and validation/test set are then converted to numpy arrays<sup>[?]</sup> for faster loading and easier indexing, and saved as `".npy"` files. This allows for faster reuse of the arrays.

### Setup via iterative training for 2 lepton dataset

The two lepton dataset contains about 1.5 - 2 Terra bytes of data when implementing the RMM structure for 6 b- and ljets, 5 electrons and 5 muons, as for the 3 lepton dataset. This is too much to hold in memory at the same time, thus it had to be split into several smaller datasets, called megasets. Below is a figure visualizing the structure used.



**Figure 3.11:** Megaset structure for the 2lep dataset. This figure generalises to M channels, and N megasets. The more you increase the number of megasets, the smaller each megaset will be in bytesize, but in order to keep the SM MC distribution, it is not recommended to make too small sets.

In figure ?? we see a generalized dividing structure. In the case of this thesis, each channel was divided into 10 equal parts, and stored in their respective folder. Pandas is not built for very large datasets, not running in a parallelized way. To handle this, the library Polars<sup>18</sup> [?] was used instead. When all channels were split, a merging was done combining all the channels in a given megaset to a separate dataset. The selection of events from each channel was done randomly, which is important, as we want to the best of our ability keep the distribution signature of the entire dataset in each megaset. If not, the model will be biased towards those datasets with the most events. Once each of the megasets were merged, the training could begin in an iterative fashion. Because Tensorflow is statically compiled, you cannot call the fit function over and over again. Instead, the weights trained based on one megaset is stored and reloaded into a new model, thus the weights are still trained on the entire set, but in a batch like manner.

<sup>15</sup>T4 → 4 particle types: bjets, ljets, electrons and muons. N5 → 5 particles per particle type. Note here that we have 5 particles only for the leptons, and 6 particles for each of the types of jets.

<sup>16</sup>This column would in a fully supervised setting be used as a target vector, but in this thesis it will only be used for legends in histograms and to index out certain channels in the validation and training set.

<sup>17</sup>Invariant mass of three leptons. This is assured to exist from event selection in the 3 lepton dataset.

<sup>18</sup>Polars uses all available cores on the system and has excellent memory handling capability, see [Polars User Guide](#)

```

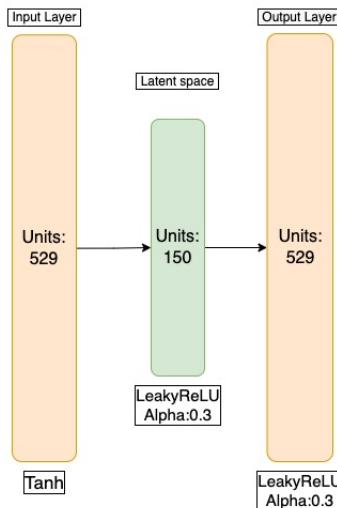
1 for megaset in range(totmegasets):
2
3     #* Load model
4
5     autoencoder = getModel()
6
7     if megaset != 0:
8         autoencoder.load_weights('./checkpoints/Megabatch_checkpoint')
9
10
11    #* Run Training
12    with tf.device("/GPU:0"):
13
14        tf.config.optimizer.set_jit("autoclustering")
15
16        autoencoder.fit(
17            xtrain,
18            xtrain,
19            epochs=epochs,
20            batch_size=b_size,
21            validation_data=(xval, xval),
22            sample_weight=x_train_weights,
23        )
24
25
26    AE_model.save_weights('./checkpoints/Megabatch_checkpoint')

```

## 3.6 The chosen neural network architectures

### The regular Autoencoder

Below are the two models used for the regular autoencoder.



**Figure 3.12:** Small autoencoder architecture.

In figure ?? we have the small autoencoder. It consists of an input and output layer of 529 nodes, with one latent space layer of 150 nodes. The activation functions for the input, latent space and output are the Tanh and LeakyReLU with  $\alpha = 0.3$  respectively,

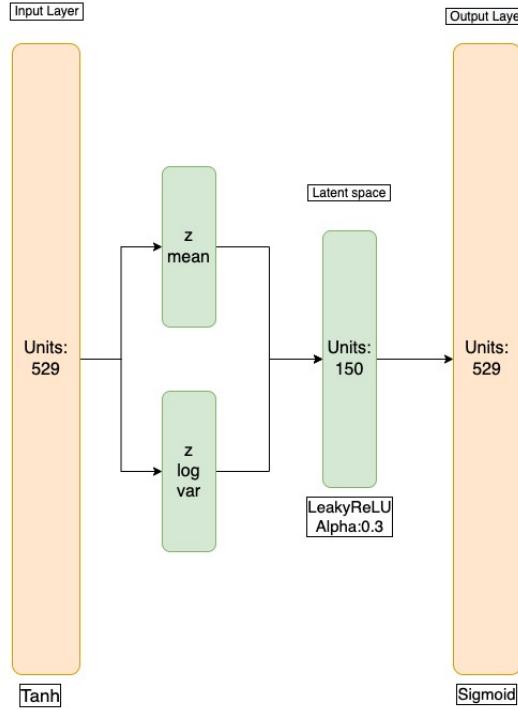


**Figure 3.13:** Large autoencoder architecture.

In figure ?? we have the large autoencoder. It consists of an input and output layer of 529 nodes, with three hidden layers of 450, 300 and 200 nodes respectively in the encoder and three hidden layers of 200, 300 and 450 respectively. The activation functions for the input and output layers are the Tanh and LeakyReLU with  $\alpha = 0.3$  respectively. The hidden layers in the encoder have the activation functions Tanh, LeakyReLU with  $\alpha = 0.3$  and ReLU respectively. The hidden layers in the decoder have the activation functions ReLU, LeakyReLU with  $\alpha = 0.3$  and Tanh respectively. The latent space has 150 nodes, with the LeakyReLU activation function with  $\alpha = 0.3$ .

### The variational Autoencoder

Below are the two models used for the variational autoencoder.



**Figure 3.14:** Small variational autoencoder architecture.

In figure ?? we have the small variational autoencoder. It consists of an input and output layer of 529 nodes, with one latent space layer of 150 nodes sampling from a mean and variance layer of same size. The activation functions for the input and output are the Tanh and LeakyReLU with  $\alpha = 0.3$  respectively.



**Figure 3.15:** Large variational autoencoder architecture.

In figure ?? we have the large autoencoder. It consists of an input and output layer of 529 nodes, with three hidden layers of 400, 300 and 200 nodes respectively in the encoder and three hidden layers of 200, 300 and 400 respectively. The activation functions for the input and ouput layers are the ReLU and Sigmoid respectively. The hidden layers in the encoder have the activation functions Tanh, ReLU, LeakyReLU with  $\alpha = 0.3$  respectively. The hidden layers in the decoder have the activation functions LeakyReLU with  $\alpha = 0.3$ , ReLU and Tanh respectively.

### 3.7 The search strategy

The strategy used to look for anomalies in the three lepton + missing energy final state is presented as follows. First, the MonteCarlo and ATLAS data for training and inference are constructed with the RMM from section ?? as features. After scaling and splitting, 80% of the MC will be used for training the neural network, and the remaining 20% will be used for inference. To separate the anomalies, the reconstruction error of the input data will create a distribution. The same will then be done for a test signal, and eventually the ATLAS data. The ATLAS data is completely unlabeled which makes validation difficult. However, the test signals are labeled, thus we can analyze the reconstruction error distribution of those signals as a weak<sup>19</sup> validation of performance when later running on ATLAS data.

Standard analyses creates what is called a signal region. The signal region is a region in the feature space where the signal is maximized. We use this region to calculate the significance of a result in the search, which is really the only metric that is of use. The statistical uncertainty and noise is proportional to the amount of SM MC, thus with lower amounts of SM MC, the better the significance will be. Using the reconstruction error, the autoencoder can create its own signal region, namely the areas of high reconstruction error. A cut here will then be used to separate the anomalies from the Standard model in for example missing transverse energy, or other features of interest.

To avoid bias by the author, some of the test signals should contain some signal samples that the creator of the model has not seen before, to ensure no changes have been made to the network to adjust for that signal. ROC curves will then be used to evaluate the binary classification ability of the autoencoder.

## Initial testings

Before testing the AE and VAE on signal samples, it was of interest to test the sensitivity of the two models on alterations on the Monte Carlo. This can be thought of as initial testing.

### Channel removing

As the goal of the autoencoder is to reconstruction data is has looked on, one idea was to remove one of the channels in the standard model. The idea was that some of the channels differs enough in the final states they produce and thus the RMMs for the events in the given selection. All channels were tested on as signal, but one can expect some to have more similar results than others.

### Altering transverse momentum

Another idea for anomaly detection testing with the MonteCarlo was to alter the transverse momentum of some of the particles. Random events were selected and had the transverse energy changed, in accordance with equation ???. The hope is that especially events with above 5 time increase in transverse momentum should be picked up. Note that by changing the transverse momentum, the change in transverse energy also changes. From equation ?? we have that a scale change  $k$  in  $p_T$  yields the following new relation:

$$\delta e_T^k = \frac{k E_T(i_n - 1) - E_T(i_n)}{k E_T(i_n - 1) + E_T(i_n)}, n = 2, \dots, N. \quad (3.8)$$

Thus, both the transverse energy and the change in transverse energy is changed for this test.

### Feature shuffling

Another idea was to shuffle the features of the events. This is done by randomly selecting a feature and swapping it with another feature. This will create fake and unphysical events, which should be picked up by the autoencoder.

### Dummy data

The last idea was to create dummy data. This is done by selecting both a percentage of the rows and columns, and swapping them, making the data unphysical.

---

<sup>19</sup>Weak means here that although the test signals can provide valuable insight in how the autoencoder can separate anomalies, we must be cautious about generalizing the evaluation value of those results. There are many contributing factors that affect the reconstruction error separation, which will be such as similarities with the Standard model, large amounts of missing energy, high energy particles and more, discussed in chapter ??.

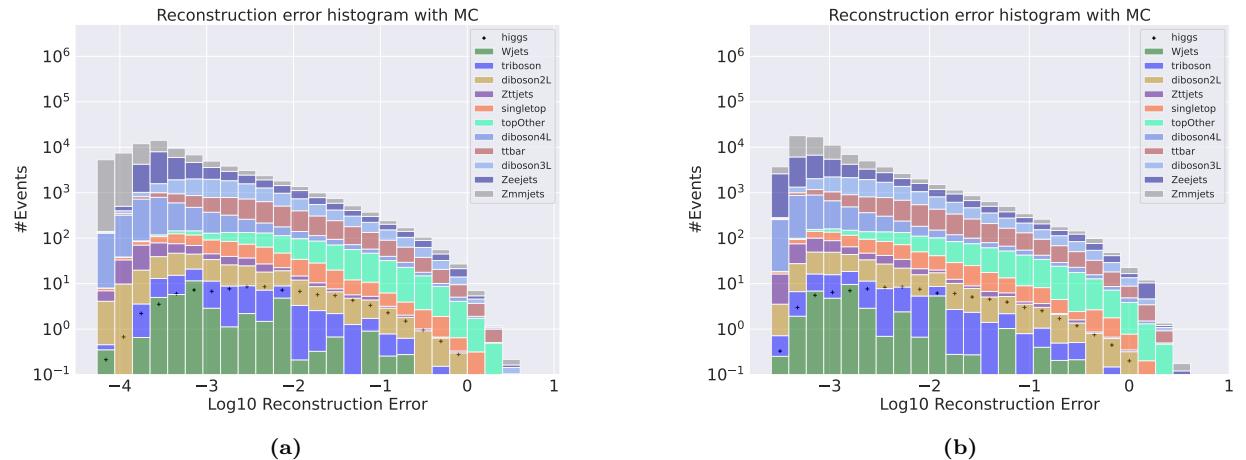
# Chapter 4

## Results and Discussion

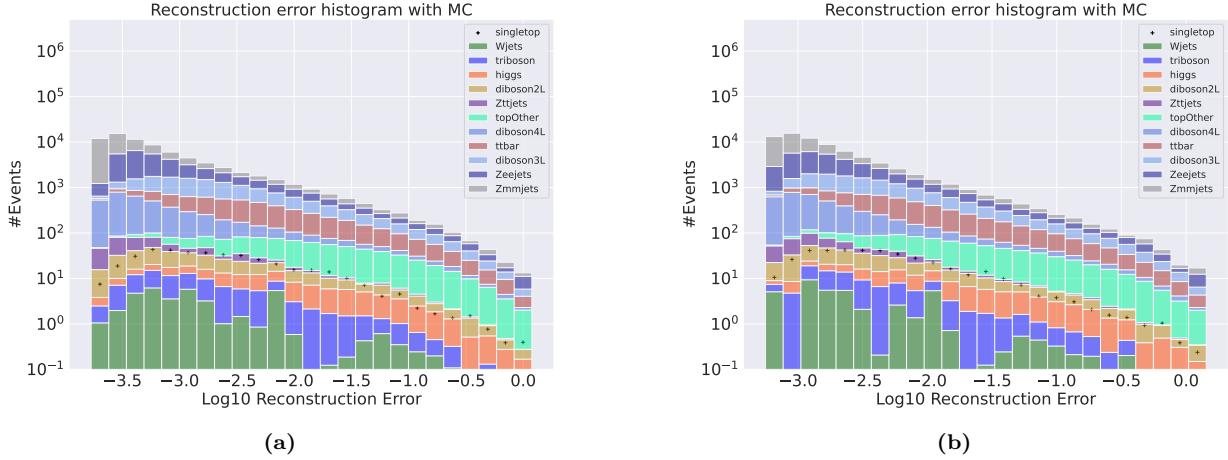
### 4.1 Non signal testing of the regular and variational Autoencoder

#### Channel removing

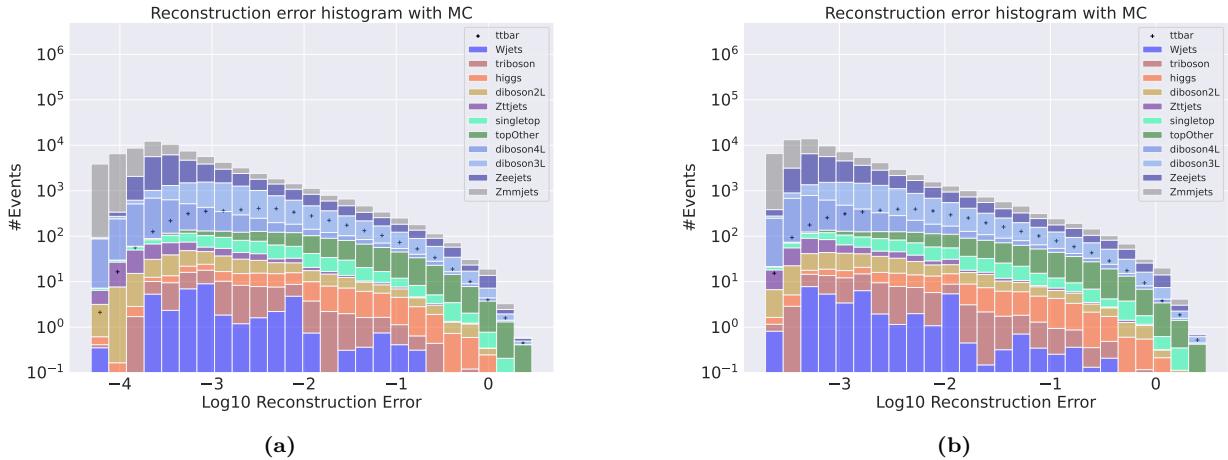
Both the large and small regular and variational autoencoder produced results, and are shown below. The Higgs, singletop and ttbar channels have been selected here, as they are from a physics stand point the channels that looks most unlike the other channels.



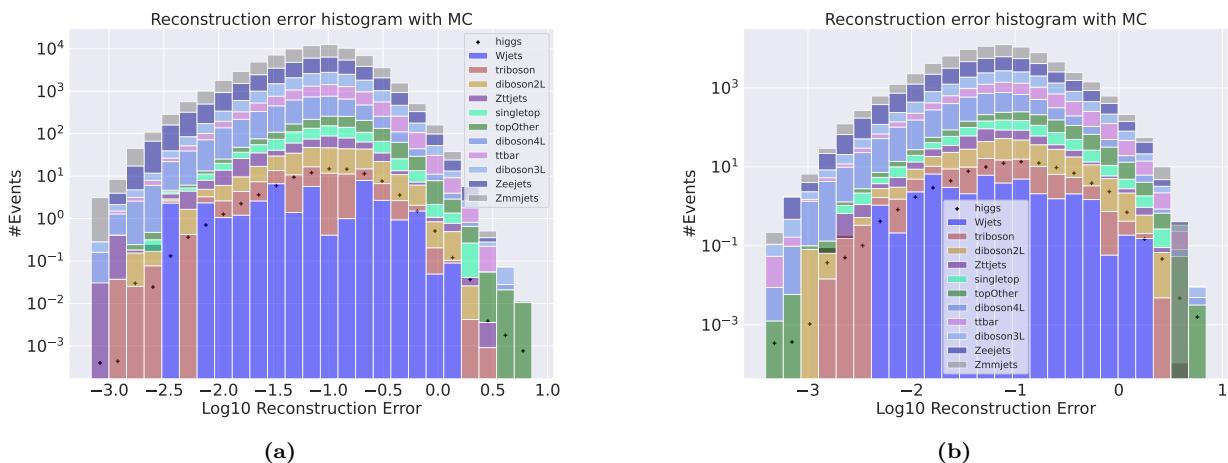
**Figure 4.1:** Reconstruction error on validation SM MC from the small (left) and large (right) Autoencoders. Here the higgs channel has been removed from training and is used as signal. No significant difference in distributions are found.



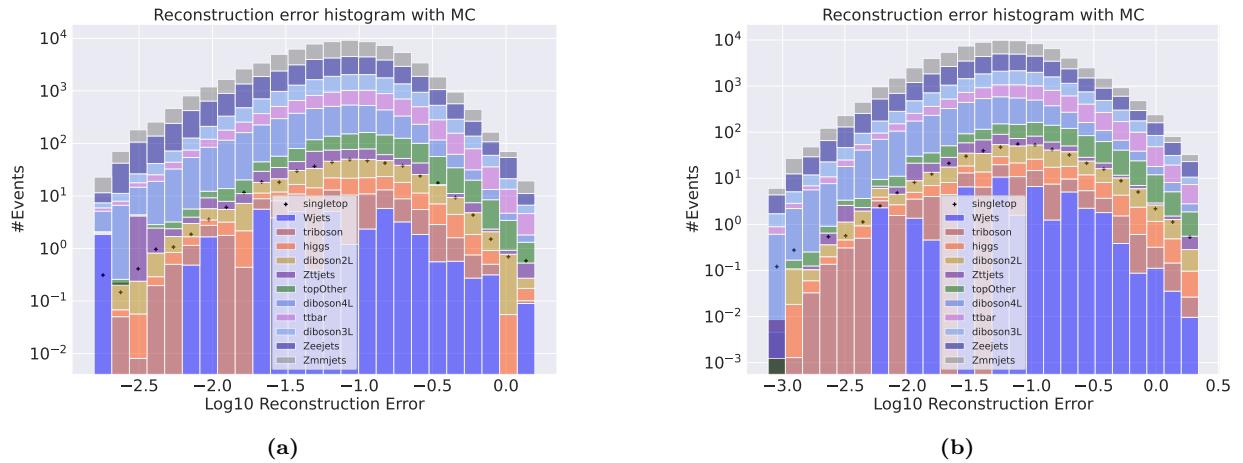
**Figure 4.2:** Reconstruction error on validation SM MC from the small (left) and large (right) Autoencoders. Here the singletop channel has been removed from training and is used as signal. No significant difference in distributions are found.



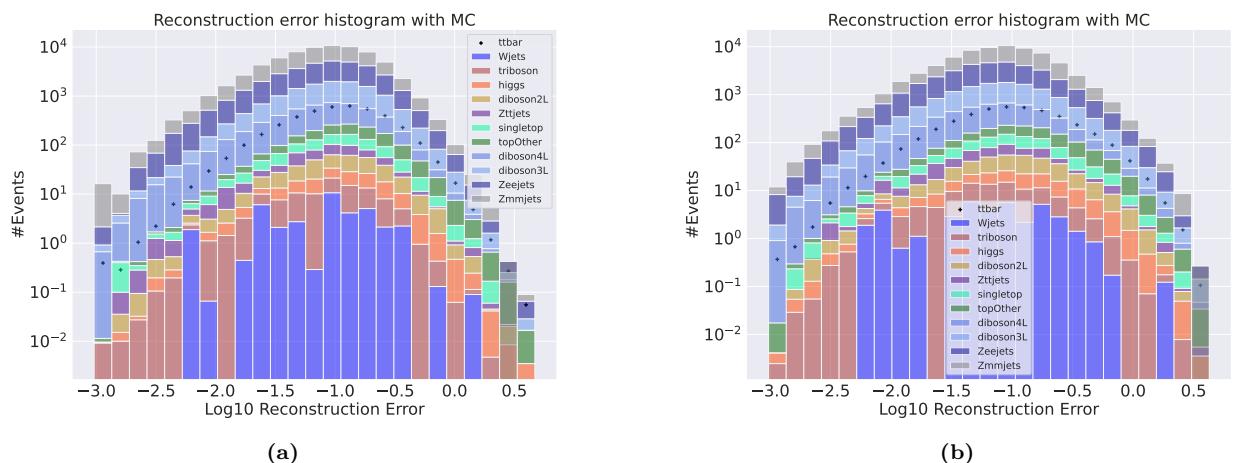
**Figure 4.3:** Reconstruction error on validation SM MC from the small (left) and large (right) Autoencoder. Here the ttbar channel has been removed from training and is used as signal. No significant difference in distributions are found.



**Figure 4.4:** Reconstruction error on validation SM MC from the small (left) and large (right) variational Autoencoder. Here the higgs channel has been removed from training and is used as signal. No significant difference in distributions are found.



**Figure 4.5:** Reconstruction error on validation SM MC from the small (left) and large (right) variational Autoencoder. Here the singletop channel has been removed from training and is used as signal. No significant difference in distributions are found.

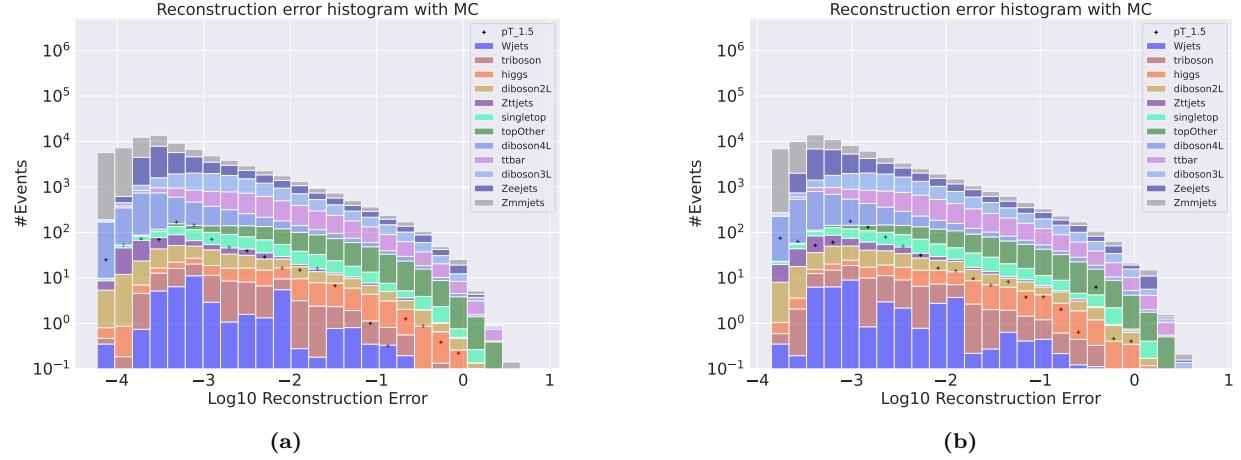


**Figure 4.6:** Reconstruction error on validation SM MC from the small (left) and large (right) variational Autoencoder. Here the ttbar channel has been removed from training and is used as signal. No significant difference in distributions are found.

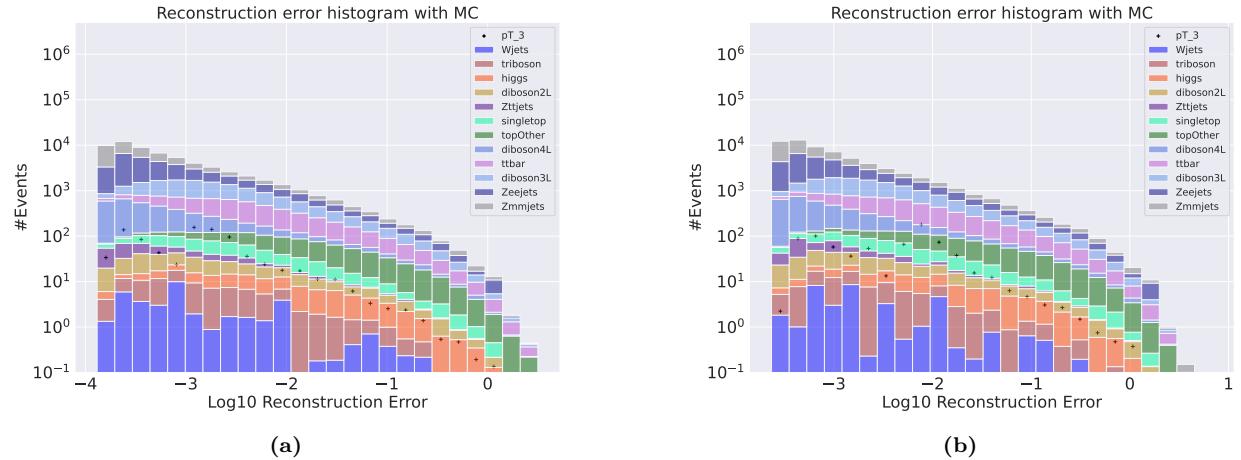
In figures ?? to ?? we have the reconstruction error distributions for the SM MC using Higgs, singletop and ttbar as signal samples for the small (to the left) and the big (to the right) regular and variational autoencoder. It is clear from figures that the reconstruction error distributions for the removed SM channel and the remaining SM MC are very similar. The same behavior is also shown for the other channels, which can be found in appendix 2. All though the two models struggled to get a separation of reconstruction error distributions, the regular autoencoder seems to have a more shifted pattern of reconstruction to lower values than the variational autoencoder. In fact, the regular autoencoder’s reconstruction error distribution peak is about 3 order of magnitude lower than the variational autoencoder’s error distribution peak. There could be several reasons for this, one of which is that the variational autoencoder requires more input data to approximate the distribution, as well as the fact that the balance between the KL divergence and MSE loss can be difficult to handle[? ]. This can lead to poor performance of the network. It should be noted that all though the three channels selected here, as well as the rest in appendix 2, do not differ that much from one another, and the hopes were not high that the networks would be able to separate the two distributions created. However it does provide us with a baseline, as well as insight for what to expect if we were to test on signals that looks a lot like some channels.

## Altering of transverse momentum

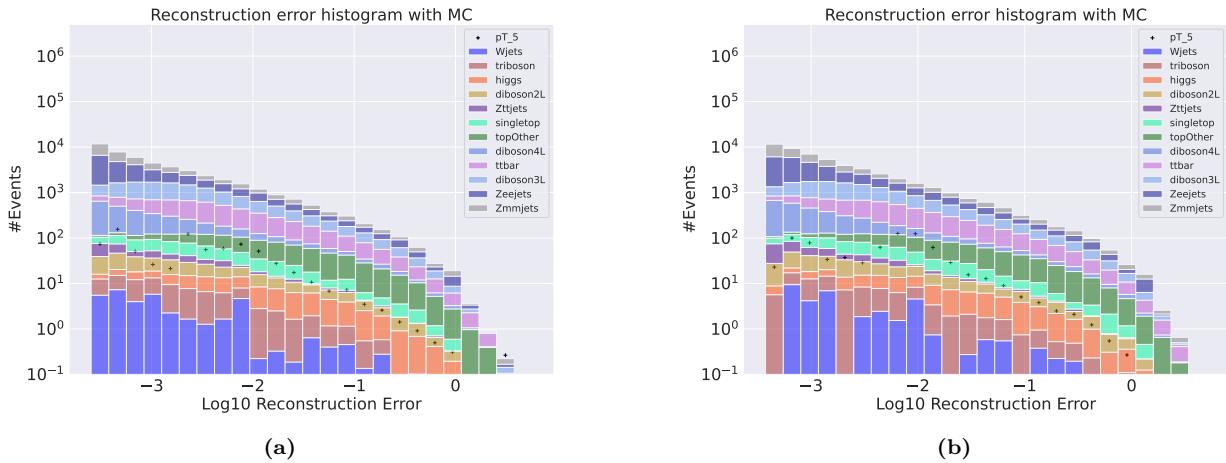
Altering the transverse momentum of some particles would in the extreme be anomalous and the hypothesis was that some of those trends would be picked up by the autoencoder. Several scales for the transverse momentum were used, with the following results.



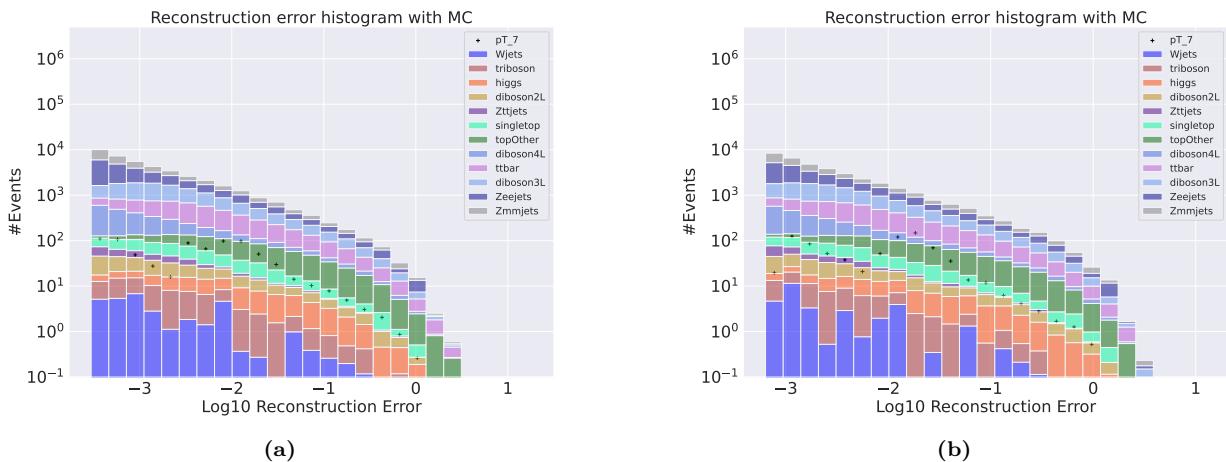
**Figure 4.7:** Reconstruction error on validation SM MC from the small (left) and large (right) regular Autoencoder. Here the signal is a subsample of the validation set where the transverse momentum of the first electron and the first muon has been increased with a scale of 1.5. The change of transverse energy has thusly also been changed according to the scaling of transverse momentum. No significant difference in distributions are found.



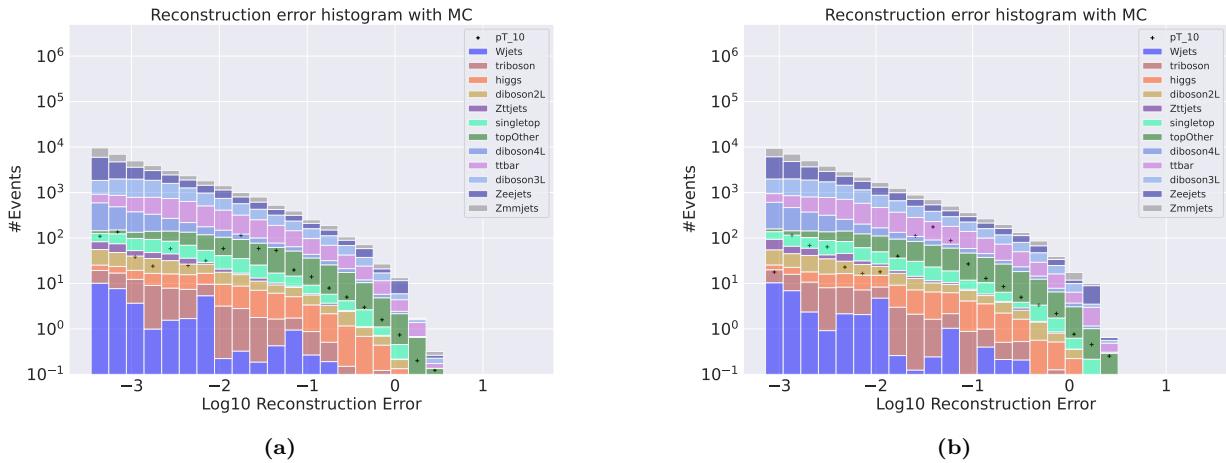
**Figure 4.8:** Reconstruction error on validation SM MC from the small (left) and large (right) regular Autoencoder. Here the signal is a subsample of the validation set where the transverse momentum of the first electron and the first muon has been increased with a scale of 3. The change of transverse energy has thusly also been changed according to the scaling of transverse momentum. No significant difference in distributions are found.



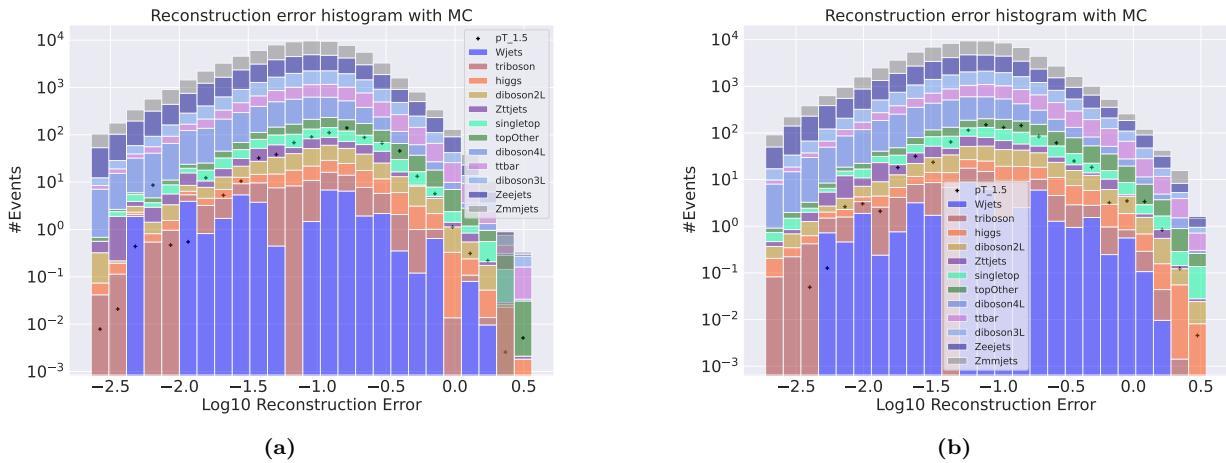
**Figure 4.9:** Reconstruction error on validation SM MC from the small (left) and large (right) regular Autoencoder. Here the signal is a subsample of the validation set where the transverse momentum of the first electron and the first muon has been increased with a scale of 5. The change of transverse energy has thusly also been changed according to the scaling of transverse momentum. No significant difference in distributions are found.



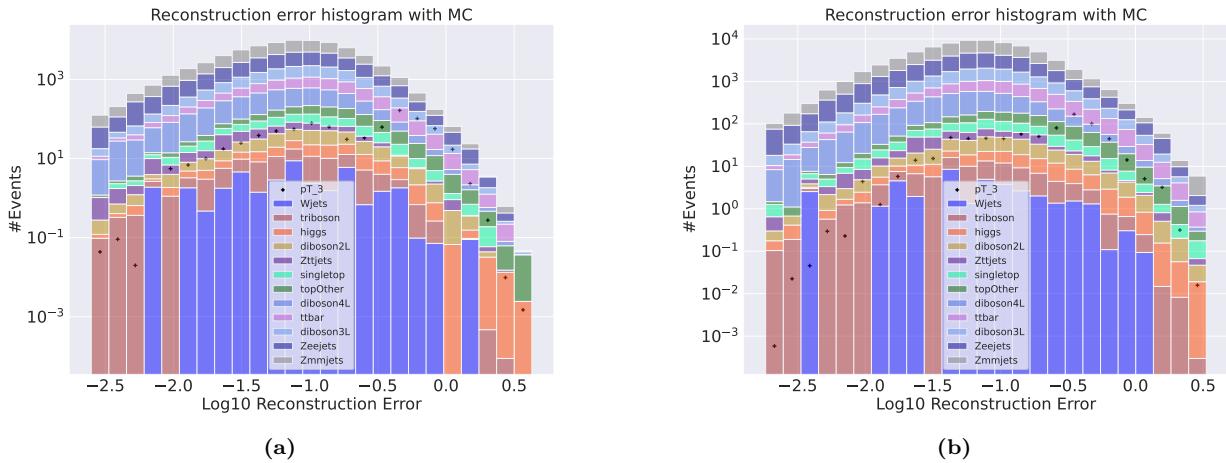
**Figure 4.10:** Reconstruction error on validation SM MC from the small (left) and large (right) regular Autoencoder. Here the signal is a subsample of the validation set where the transverse momentum of the first electron and the first muon has been increased with a scale of 7. The change of transverse energy has thusly also been changed according to the scaling of transverse momentum. No significant difference in distributions are found.



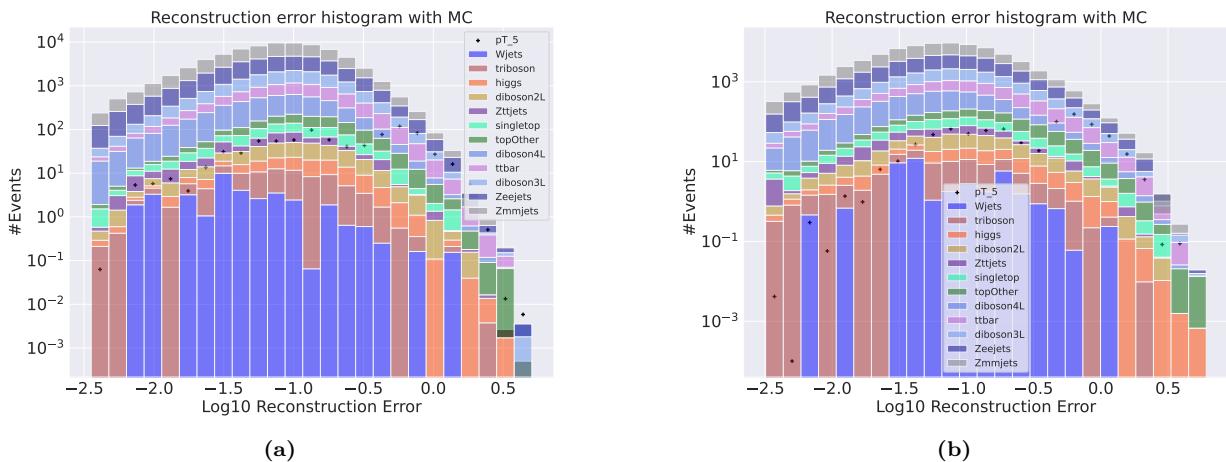
**Figure 4.11:** Reconstruction error on validation SM MC from the small (left) and large (right) regular Autoencoder. Here the signal is a subsample of the validation set where the transverse momentum of the first electron and the first muon has been increased with a scale of 10. The change of transverse energy has thusly also been changed according to the scaling of transverse momentum. No significant difference in distributions are found.



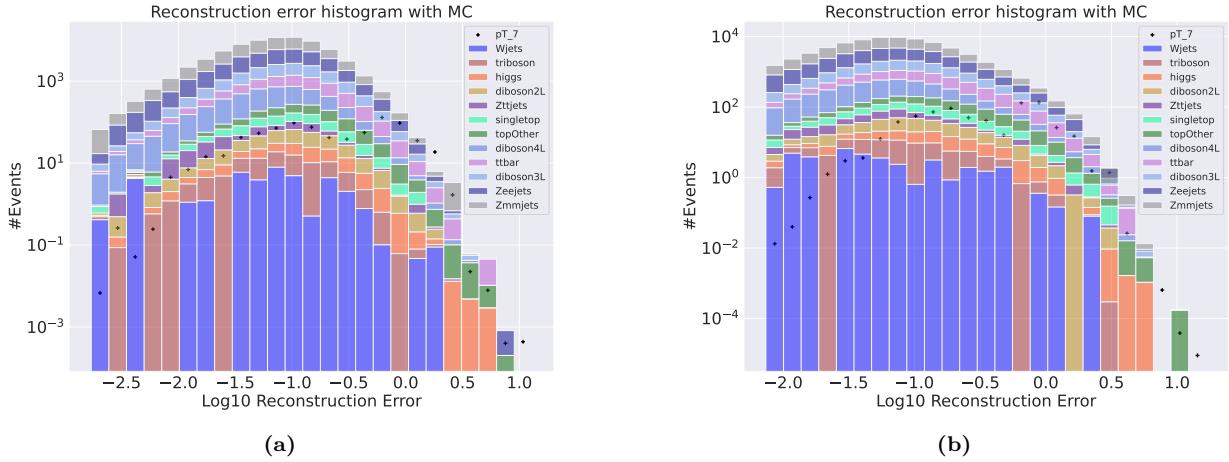
**Figure 4.12:** Reconstruction error on validation SM MC from the small (left) and large (right) variational Autoencoder. Here the signal is a subsample of the validation set where the transverse momentum of the first electron and the first muon has been increased with a scale of 1.5. The change of transverse energy has thusly also been changed according to the scaling of transverse momentum. No significant difference in distributions are found.



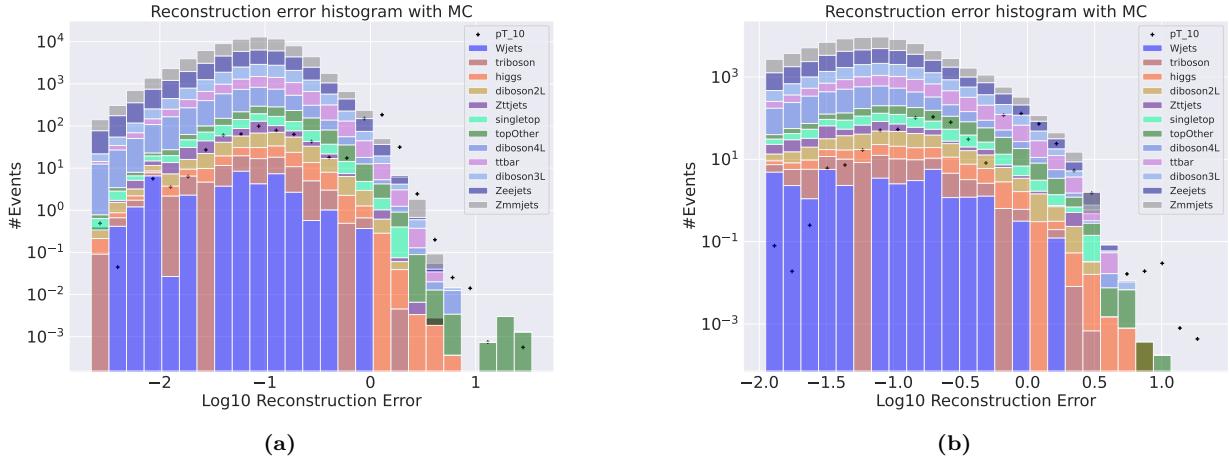
**Figure 4.13:** Reconstruction error on validation SM MC from the small (left) and large (right) variational Autoencoder. Here the signal is a subsample of the validation set where the transverse momentum of the first electron and the first muon has been increased with a scale of 3. The change of transverse energy has thusly also been changed according to the scaling of transverse momentum. No significant difference in distributions are found.



**Figure 4.14:** Reconstruction error on validation SM MC from the small (left) and large (right) variational Autoencoder. Here the signal is a subsample of the validation set where the transverse momentum of the first electron and the first muon has been increased with a scale of 5. The change of transverse energy has thusly also been changed according to the scaling of transverse momentum. No significant difference in distributions are found.



**Figure 4.15:** Reconstruction error on validation SM MC from the small (left) and large (right) variational Autoencoder. Here the signal is a subsample of the validation set where the transverse momentum of the first electron and the first muon has been increased with a scale of 7. The change of transverse energy has thusly also been changed according to the scaling of transverse momentum. No significant difference in distributions are found.



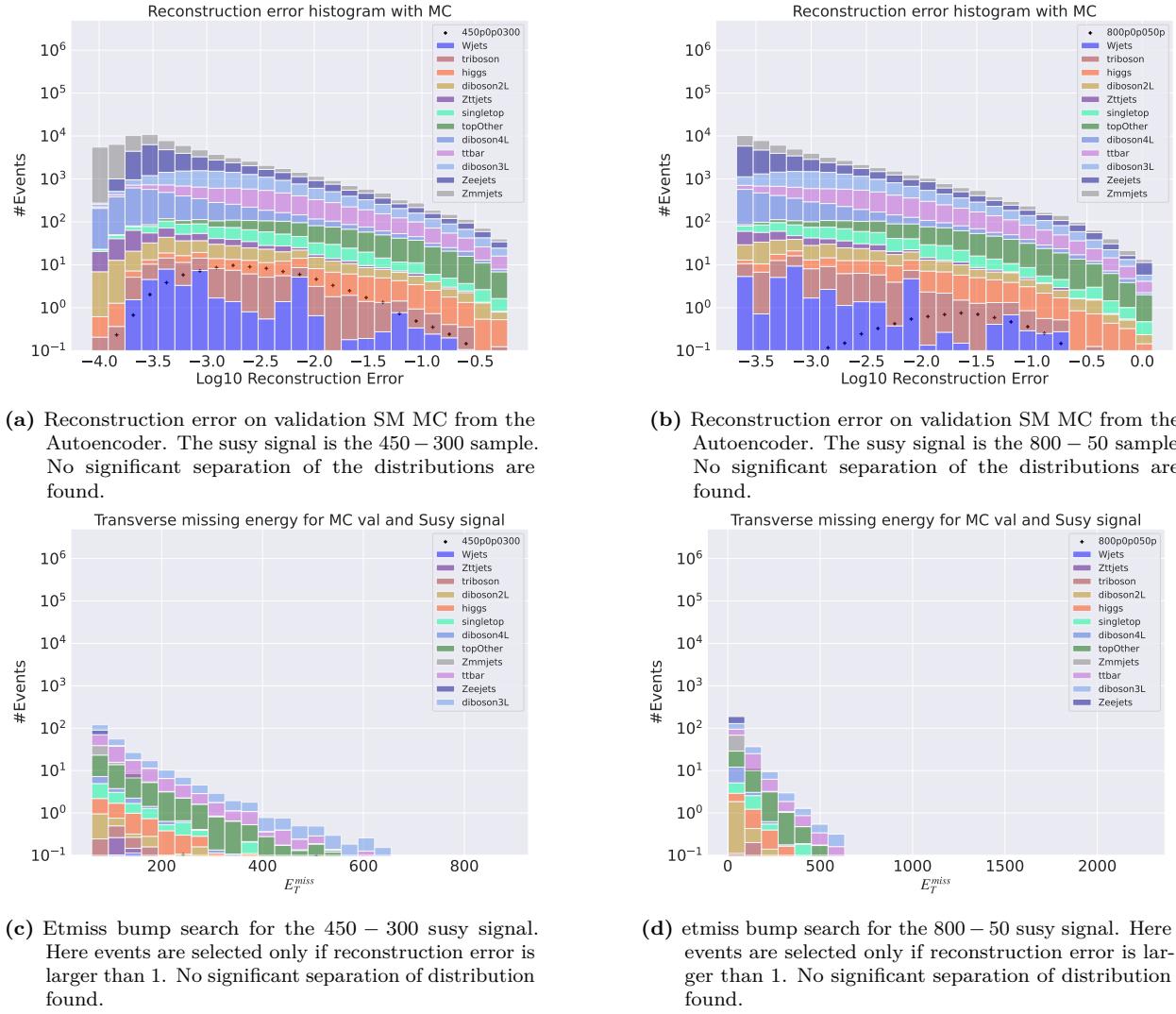
**Figure 4.16:** Reconstruction error on validation SM MC from the small (left) and large (right) variational Autoencoder. Here the signal is a subsample of the validation set where the transverse momentum of the first electron and the first muon has been increased with a scale of 10. The change of transverse energy has thusly also been changed according to the scaling of transverse momentum. No significant difference in distributions are found.

The first thing to notice here is that, as with the channel removal, the distribution shape of the output from the regular autoencoder compared to the variational autoencoder is different. The shape from the regular autoencoder is clearly shifted to the left end, with low reconstruction error compared to the shape from the variational autoencoder, where the peak of the distribution is around 1. From the  $p_T$  altering test we also see here that the variational autoencoder falls short to the regular one. It is also interesting to observe here that the high  $p_T$  events are picked up as anomalous from both autoencoders, with both shallow and deep structure. In the regular autoencoder it is arguably even separated distributions. This is also a good sign as it indicates that the network has learned which ranges the  $p_T$  should be in for the 3 lepton plus missing energy final state. The most extreme case is as expected in the case where the  $p_T$  is multiplied by 10, and the  $\delta e_T^{10}$ , there is a clear separation in reconstruction error distributions for the signal and background for the deep regular autoencoder, and a slightly more subtle separation of reconstruction error distributions for the shallow regular autoencoder. In the variational autoencoder, the separation is not as clear, but the signal is still picked up as anomalous. The peaks are separated, but not to the same degree.

## 4.2 Dummy signal testing of the regular and variational Autoencoder

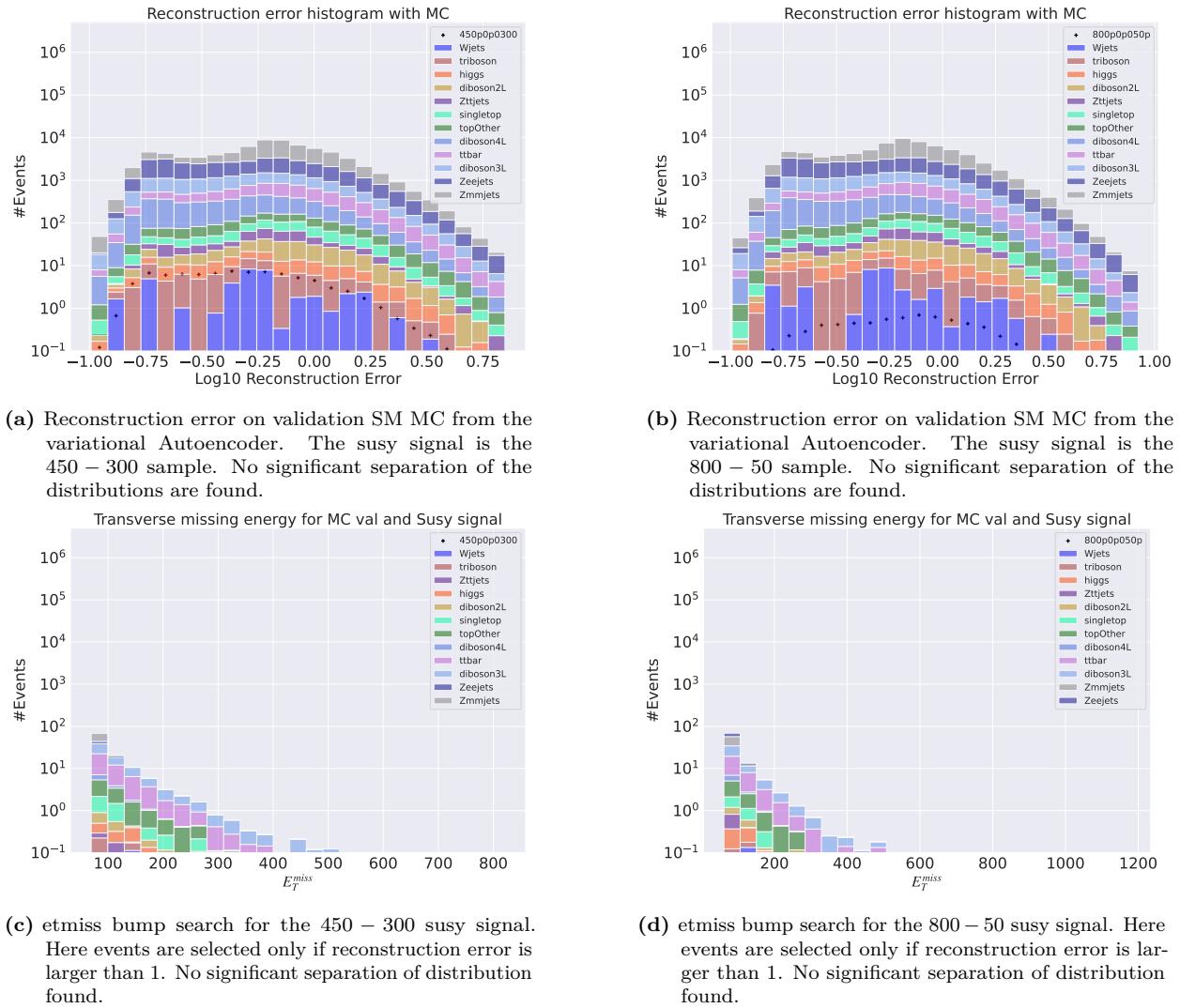
Another important part of benchmarking the algorithm is to test it on a signal sample. This is the closest test we can do before we no longer can alter the algorithm, as that would be supervised learning. Results from both the regular and the variational autoencoder are shown below.

### Autoencoder



**Figure 4.17:**  $e_T^{miss}$  bump search on the 450 – 300 and 800 – 50 susy signals. The reconstruction error cut is set to 1.

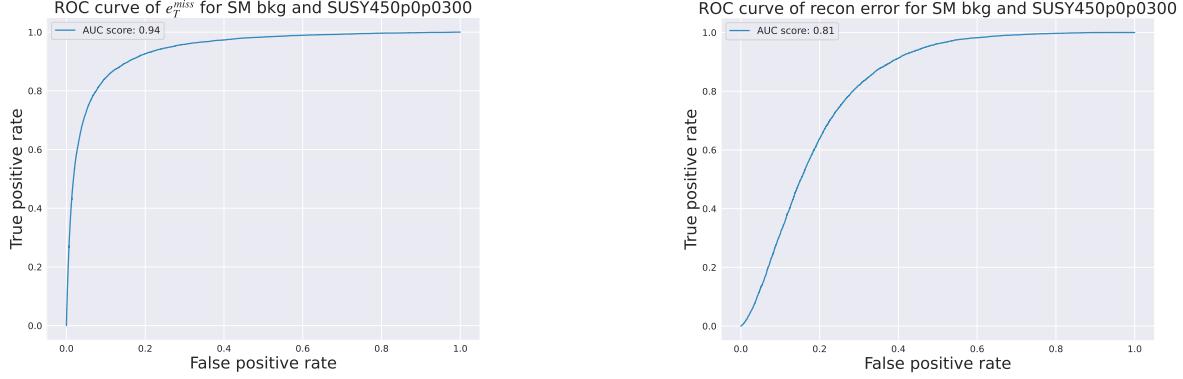
## Variational Autoencoder



**Figure 4.18:** Reconstruction error and etmiss bump search on the 450 – 300 and 800 – 50 susy signals. The reconstruction error cut is set to 1.

## ROC curves on $e_T^{miss}$ and reconstruction error

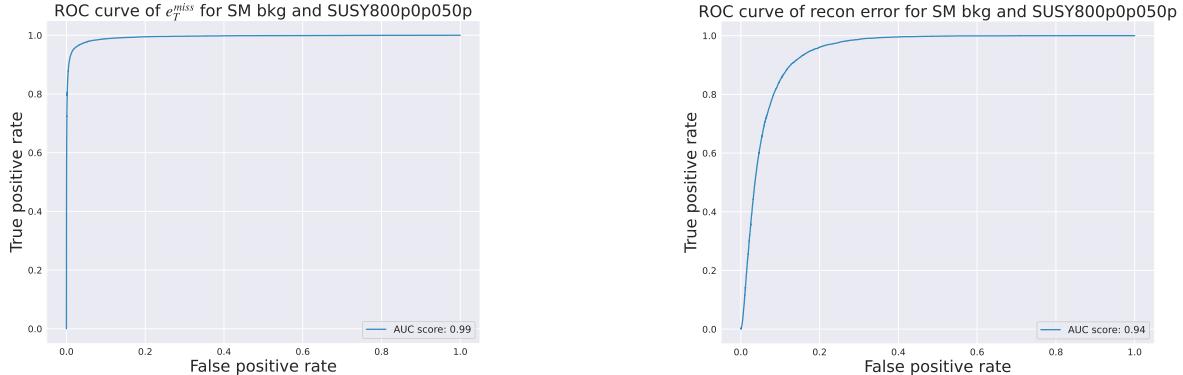
Now based on the reconstruction error on the SUSY signals, especially the one in figure ??, it is tempting to declare that the autoencoder has done a good job reconstructing the standard model and a poor job on the SUSY signal, resulting in a separation of distributions. However, there is one thing we need to remember. The SUSY signals have very high  $e_T^{miss}$ , so much so that one can separate them by eye on that distribution alone. This is presented in the figure below:



(a) ROC curve for  $e_T^{miss}$  of both SM MC and the SUSY 450 – 300 signal. We see here that based on that feature alone, you can separate the distributions with relative ease, with an AUC of about 0.94.

(b) ROC curve for reconstruction error of both SM MC and the SUSY 450 – 300 signal. We see here that there is relatively good separation, with a small AUC of about 0.81.

**Figure 4.19:** ROC curve for both  $e_T^{miss}$  and reconstruction error with the SUSY 450 – 300 signal.



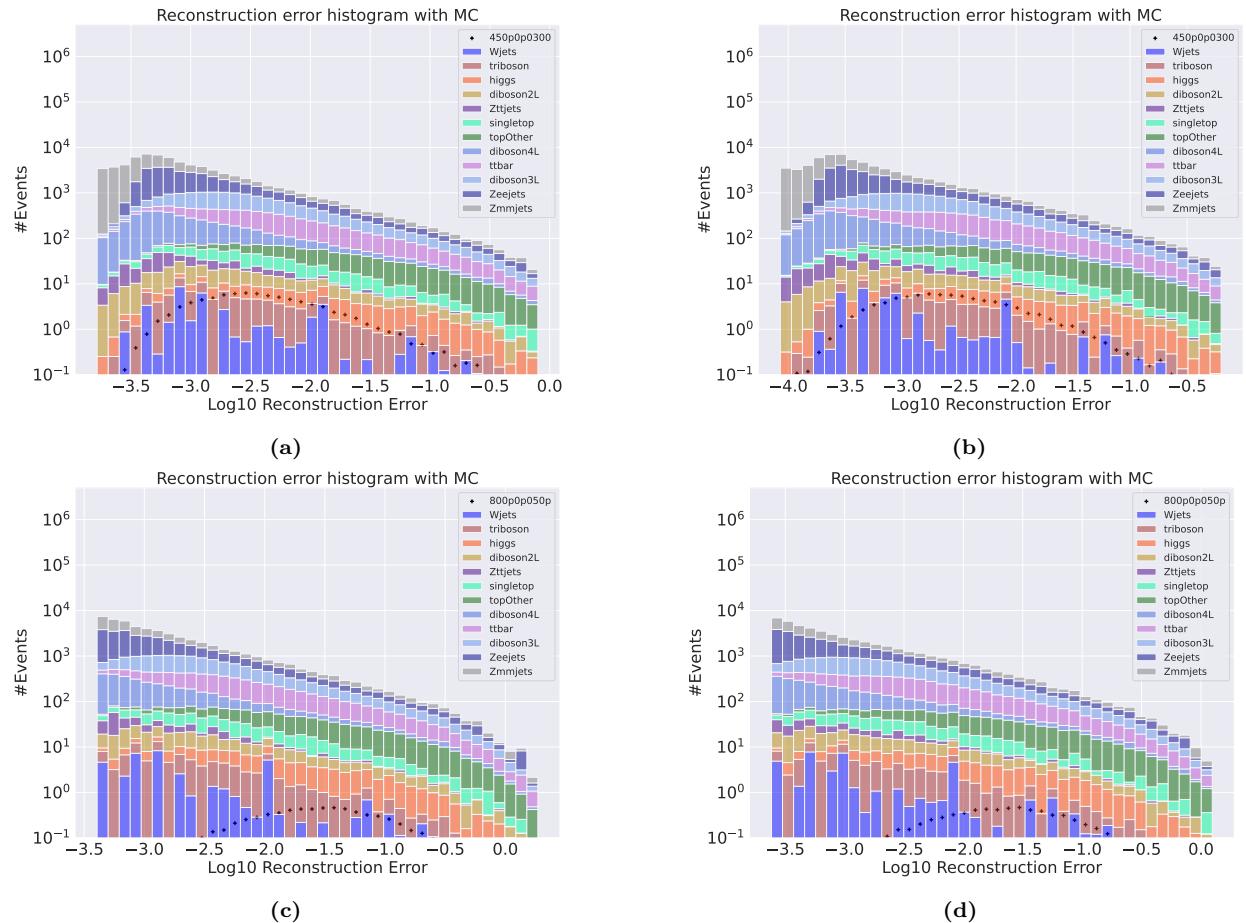
(a) ROC curve for  $e_T^{miss}$  of both SM MC and the SUSY 800 – 50 signal. We see here that based on that feature alone, you can separate the distributions with relative ease, with an AUC of about 0.99.

(b) ROC curve for reconstruction error of both SM MC and the SUSY 800 – 50 signal. We see here that there is a very good separation between the distributions , with an AUC of about 0.94.

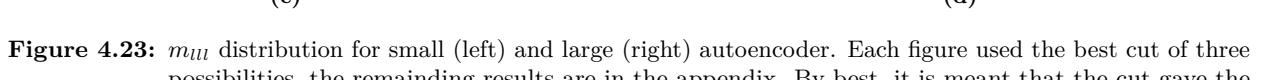
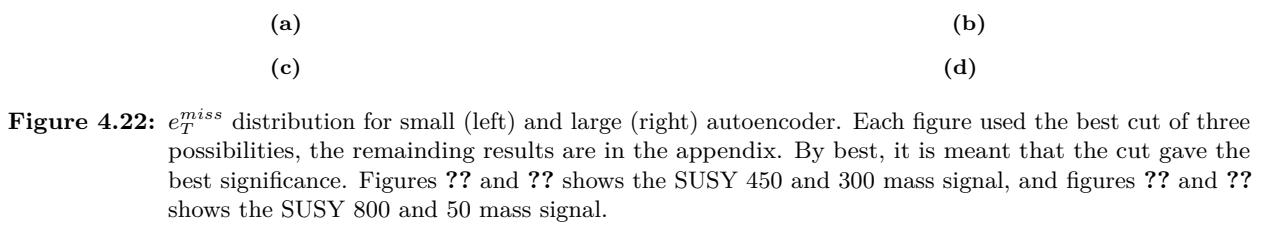
**Figure 4.20**

From figure ?? and ?? we see that even though the reconstruction error is a good discriminator for the SM MC and the SUSY signals,  $e_T^{miss}$  provides better separation. This is because the SUSY signals have very high  $e_T^{miss}$ , so much so that one can separate them by eye on that distribution alone. The interpretation of this is that the autoencoder has shown it can be used for the purpose of separating the SM MC from the SUSY signals, but not to the level of using just  $e_T^{miss}$ . The same behavior is shown for the small and large variational autoencoder in the appendix, aswell as for the small and large variational autoencoder.

### 4.3 3 lepton training for bump search testing



**Figure 4.21:** Reconstruction error distribution for the small (left) and large (right) regular autoencoder, using the 3 lepton +  $e_T^{miss}$  dataset as training and test set. The signals used are 3 lepton +  $e_T^{miss}$  finalstate SUSY signals. Figures ?? and ?? shows the SUSY 450 and 300 mass signal, and figures ?? and ?? shows the SUSY 800 and 50 mass signal.



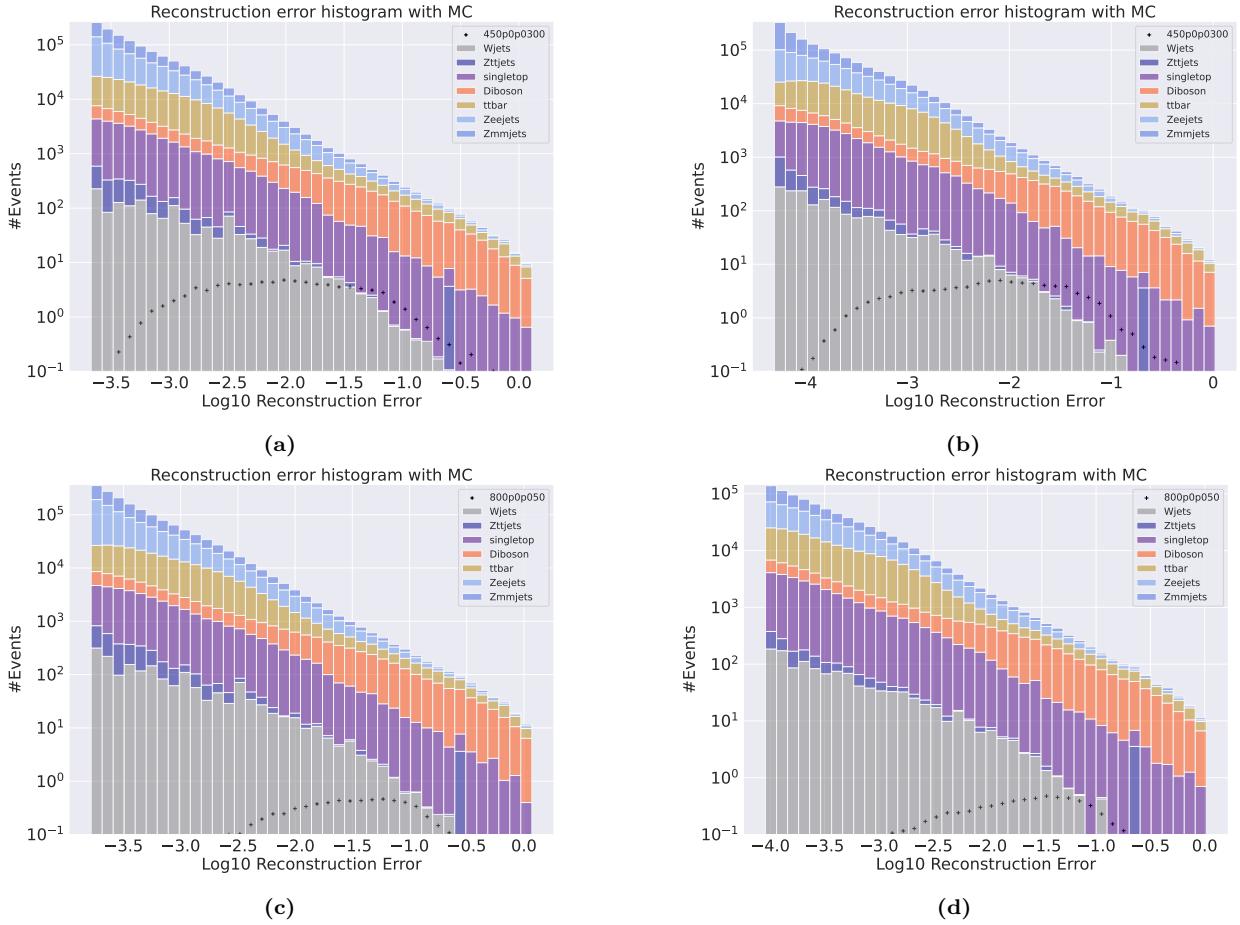
**Figure 4.23:**  $m_{lll}$  distribution for small (left) and large (right) autoencoder. Each figure used the best cut of three possibilities, the remaining results are in the appendix. By best, it is meant that the cut gave the best significance. Figures ?? and ?? shows the SUSY 450 and 300 mass signal, and figures ?? and ?? shows the SUSY 800 and 50 mass signal.

**Table 4.1:** Significance table for the regular autoencoder models. For each signal for both the small and large regular autoencoder, the significance is calculated with equation small (left column) and equation large (right column). The reconstruction error cuts used where [] for the SUSY 450p300 signal and [] for the SUSY 800p50 signal for the large autoencoder. The reconstruction error cuts used for the small autoencoder where [] for the SUSY 450p300 signal and [] for the SUSY 800p50 signal.

Regular Autoencoder					
Small		Large			
450p300	800p50	450p300	800p50		
Pre reconstruction error cut significance					
Post reconstruction error cut significance					

## 4.4 2 lepton training for bump search testing

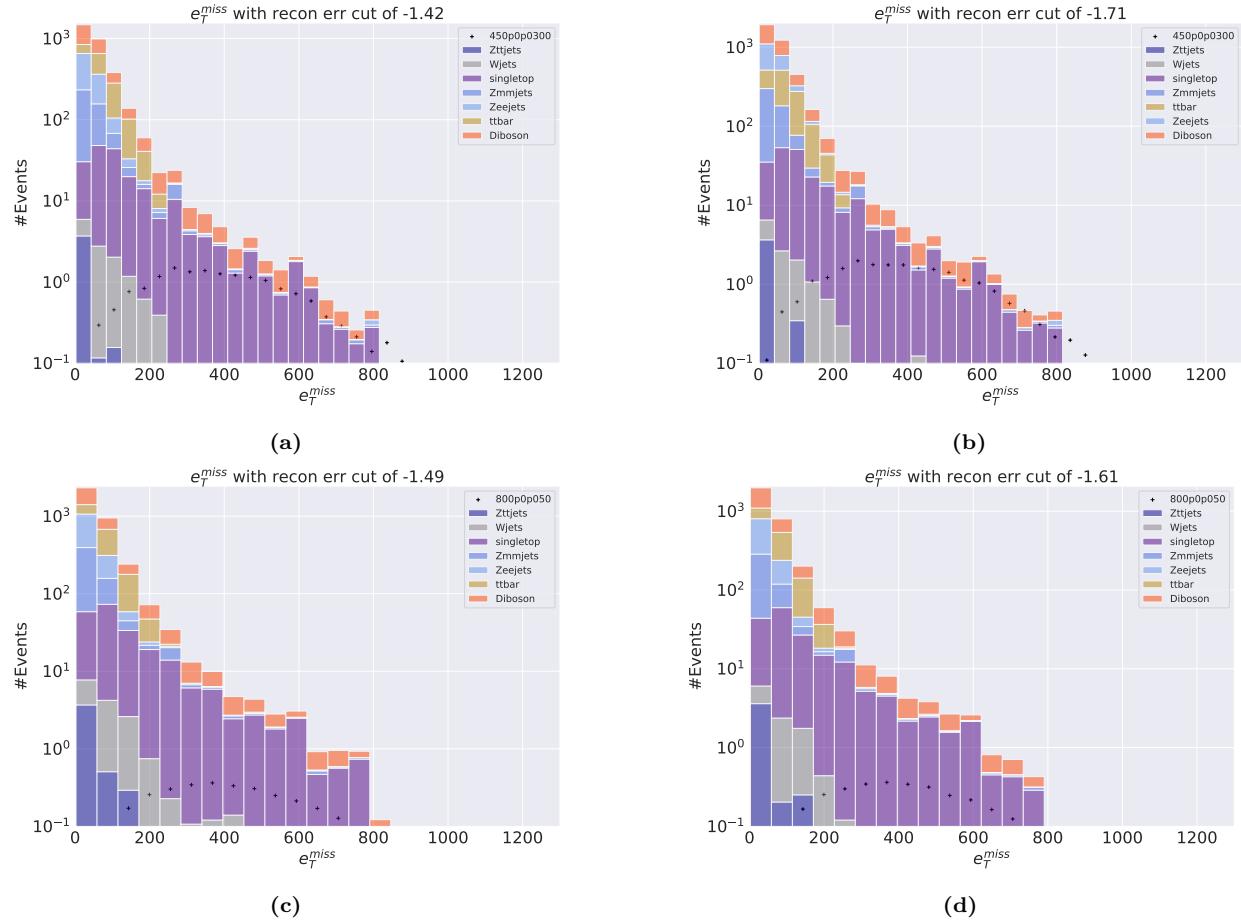
The 3 lepton +  $e_T^{miss}$  dataset has less events, and thus allows for less training of the neural networks. Thus, the 2 lepton +  $e_T^{miss}$  dataset was tried instead. The event selection was done choosing atleast 2 leptons, meaning that the RMM signatures of some of the events will look similar to the RMM signatures of the 3 lepton +  $e_T^{miss}$  dataset. A consequence of this is that the signal samples for the 3 lepton case would be a good start point for testing. The signal region for the regular autoencoder models were created by calculating the median  $m_{err}$  of the reconstruction error. Then, 3 cuts were made, starting at  $m_{err} + im_{err}/5$  for  $i = 1, 2, 3$ . As the shape of the background reconstruction error has a hill like shape, the median is a good place to start to remove alot of the background. This is however just a guess for an optimal signal region, as the true signal is unknown, and the method has to be as unbiased as possible. Below are some of the results from training on the 2 lepton case, using the same two SUSY signals as test cases.



**Figure 4.24:** Reconstruction error distribution for the small (left) and large (right) regular autoencoder, using the 2 lepton +  $e_T^{miss}$  dataset as training and test set. The signals used are the same SUSY signals from the 3 lepton case. Figures ?? and ?? shows the SUSY 450 and 300 mass signal, and figures ?? and ?? shows the SUSY 800 and 50 mass signal.

In figure ?? we have the reconstruction error distributions for the 2 lepton +  $e_T^{miss}$  dataset for the regular autoencoder models. There is a general trend here for both the small and large regular autoencoder, which is that the distribution is highly shifted to the lower end of the reconstruction error range. This is a continuing trend following the 3 lepton +  $e_T^{miss}$  shown in figure ???. This indicates that as we increase the statistics, in other words the amount of background events, the ability of the autoencoder to learn the internal structure increases. It is also interesting to observe here that as the reconstruction error increases, the amount of each sample in a given bin changes a lot. As expected, Zmmjets and Zeejets along with ttbar are the events with the highest statistics in the 2 lepton +  $e_T^{miss}$  dataset, thus it should be easier to learn to better reconstruct those events. However, note the amount of Diboson in the higher end of the reconstruction error histograms, as well as in the  $e_T^{miss}$  post reconstruction error cut distributions. One explanation for this could be the sample discrepancy explained in section ???. It would appear that too much of the diboson samples is cut, leading to not enough samples for it to appropriately learn the RMM structure for that channel, leading to the high reconstruction error values for some of those events. Note also here that the reconstruction threshold where this effect is getting more noticeable around  $10^{-2}$  and larger, where the events are on the order of a 1000 or less per bin. Thus it is not a lot of events, but still enough to note.

Using the signal region definition from above we set cuts on the reconstruction error and then calculate the significance of the signal. The figures with the best significance are shown below:



**Figure 4.25:**  $e_T^{\text{miss}}$  distribution for small (left) and large (right) autoencoder. Each figure used the best cut of three possibilities, the remaining results are in the appendix. By best, it is meant that the cut gave the best significance. Figures ?? and ?? shows the SUSY 450 and 300 mass signal, and figures ?? and ?? shows the SUSY 800 and 50 mass signal.

In figure ?? we have the  $e_T^{\text{miss}}$  distributions for the best cuts for the regular autoencoder models. The cuts then gives rise to the significance for each signal region. Now, one indication that the method could work is if the models can improve the significance from just looking at the  $e_T^{\text{miss}}$  distributions of the background and the signal in mind. The significances corresponding to each reconstruction error cuts are all listed below. Note here that for  $e_T^{\text{miss}}$ , no physics based cuts have been used, but there are certain cuts that can increase the significance, so the results in the following table should be thought carefully of.

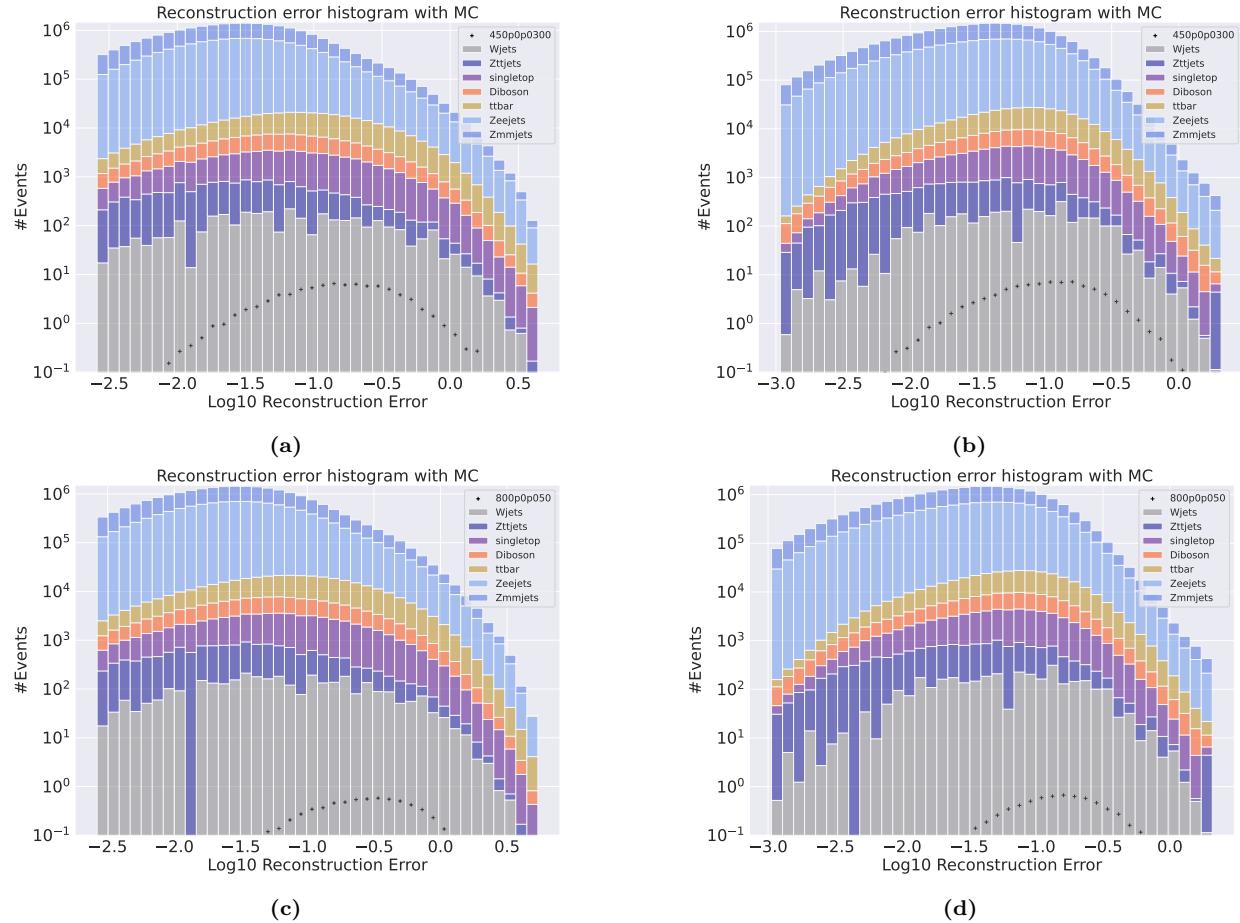
**Table 4.2:** Significance table for the regular autoencoder models. For each signal for both the small and large regular autoencoder, the significance is calculated with equation small (left column) and equation large (right column). The reconstruction error cuts used where  $[-1.42, -1.07, -0.71]$  for the SUSY 450p300 signal and  $[-1.49, -1.11, -0.74]$  for the SUSY 800p50 signal for the large autoencoder. The reconstruction error cuts used for the small autoencoder where  $[-1.71, -1.29, -0.86]$  for the SUSY 450p300 signal and  $[-1.61, -1.21, -0.81]$  for the SUSY 800p50 signal.

Regular Autoencoder								
Small				Large				
450p300		800p50		450p300		800p50		
Pre reconstruction error cut significance								
0.35	0.02	<i>NaN</i>	0.002	0.35	0.02	<i>NaN</i>	0.002	
Post reconstruction error cut significance								
0.35	0.35	0.06	0.06	0.29	0.29	0.05	0.05	
0.20	0.2	0.04	0.04	0.14	0.14	0.04	0.04	
0.06	0.06	0.02	0.02	0.06	0.06	0.02	0.02	

In table ?? we have the small data and large data significance on  $e_T^{miss}$  for both the small and large regular autoencoder models, calculated for both the SUSY 450p300 and 800p50 models. The table compares the pre reconstruction error significance with post reconstruction error significance. Note that the significance have been rounded up to two decimals, and the actual values with 17 decimal precision can be found in the *significance\_MODELTYPE\_MODELSIZE.txt* in the codebase folder in github: [MasterThesis/tree/main/CodeBase/Analysis](#). MODELTYPE refers to the type of autoencoder, either AE for regular autoencoder, or VAE for variational autoencoder. MODELSIZE refers to the size of the autoencoder, either small for the shallow model, or big for the deep model.

First, it should be noted that the significance for the 800p50 signal model is a lot smaller than for the 450p300 signal model, even though the separation shown in figures ?? and ?? would suggest otherwise. All though the peak in both signal models are fairly separated from the peak of the SM MC, the SUSY 800p50 signal model is shifted a lot more. This is consistant with expectation, as that specific signal model has a lot of missing transverse energy, compared to even the other SUSY signal model. Now, The reason for the low significance is most likely the fact that the signal sample contains low statistics, in other words, the weights are just a lot lower, compared to the other signal sample. We also have to cases where the significance is *NaN*. This only happens in the case we use equation ??, and is either because the square root has a negative term inside it, or because the logarithm has a negative term inside it.

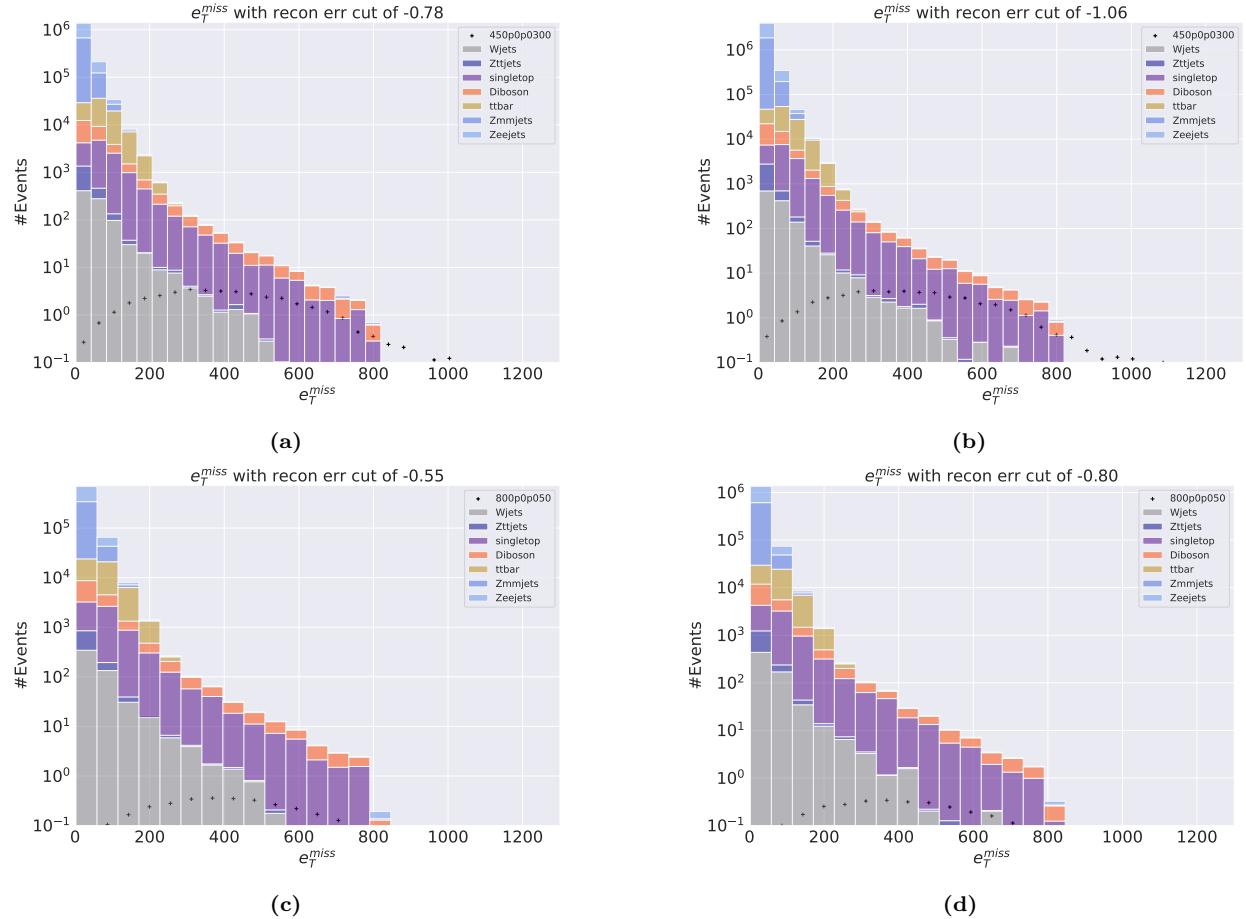
From figure ?? below we have the variational autoencoder output for the 2 lepton case.



**Figure 4.26:** Reconstruction error distribution for the small (left) and large (right) regular autoencoder, using the 2 lepton +  $e_T^{miss}$  dataset as training and test set. The signals used are the same SUSY signals from the 3 lepton case. Figures ?? and ?? shows the SUSY 450 and 300 mass signal, and figures ?? and ?? shows the SUSY 800 and 50 mass signal.

In figure ?? we have the reconstruction error distributions for both SUSY signals for the small and large variational autoencoder. Here we observe that the peak of the distributions for the SM MC in all four cases are somewhat centered in the middle of the reconstruction error range, which differs from the steep slope we saw in figure ???. Interestingly, we see here that the deepness of the neural network here plays a role, which is different from the regular autoencoder output, where both the small and large autoencoder made a steep slope shape of the SM MC reconstruction error distribution. The peak of the distribution here is slightly

shifted to the left for the shallow autoencoder model, and slightly shifted to the right of the center with the deep autoencoder model. One possible reason for this somewhat Gaussian like distribution could be that the variational autoencoder, via the reparametrization trick from section ??, samples from a Gaussian distribution that has yet to be trained on enough data to produce a good enough error distribution. It could also be that the batch size is too large, and that the model has to train on smaller batches to get a better result.



**Figure 4.27:**  $e_T^{\text{miss}}$  distribution for small (left) and large (right) autoencoder. Each figure used the best cut of three possibilities, the remaining results are in the appendix. By best, it is meant that the cut gave the best significance. Figures ?? and ?? shows the SUSY 450 and 300 mass signal, and figures ?? and ?? shows the SUSY 800 and 50 mass signal.

In figure ?? we have the  $e_T^{\text{miss}}$  distribution for the best cut for each signal. By best, it is meant the cut that gave the best significance. We see that the cuts are somewhat similar to the regular autoencoder, but with two key differences.

First, because the peaks of the distributions from figure ?? are so close, the cuts allowed for more background events in the signal region. Here, as with the regular autoencoder output,  $m_{\text{err}}$  was used, but was not a good discriminator for the background events. One could perhaps make a new strategy for setting the cuts in the event one has a "fat gaussian" shape. Still, because we set cuts based on reconstruction error to minimize the background in the signal region, if one does not have a slope like shape, the results will be poor in comparison.

Secondly, the background that remains are slightly different from the signal region from the regular autoencoder. In the lower energy range there is a large excess of Zeejets, Zmmjets and ttbar events that have a high reconstruction error, which is not the case for the variational autoencoder, dominated by diboson events in all bins. In the higher energy range, diboson are largest contributer to the background, but the number of bins are exceptionally smaller than the Zmmjets/Zeejets/ttbar events, by a magnitude of 3 at the most.

**Table 4.3:** Significance table for the variational autoencoder models. For each signal for both the small and large regular autoencoder, the significance is calculated with equation small (left column) and equation large (right column). The reconstruction error cuts used where  $[-0.78, -0.58, -0.39]$  for the SUSY 450p300 signal and  $[-0.74, -0.55, -0.37]$  for the SUSY 800p50 signal for the large autoencoder. The reconstruction error cuts used for the small autoencoder where  $[-1.06, -0.79, -0.53]$  for the SUSY 450p300 signal and  $[-1.06, -0.79, -0.53]$  for the SUSY 800p50 signal.

Variational Autoencoder							
Small				Large			
450p300	800p50	450p300	800p50	Pre reconstruction error cut significance			
0.017	0.017	0.001	0.001	0.017	0.017	0.001	0.001
Post reconstruction error cut significance							
0.023	0.023	0.002	0.002	0.030	0.030	0.004	0.004
0.021	0.022	0.003	0.003	0.026	0.026	0.004	0.004
0.016	0.016	0.002	0.002	0.018	0.018	0.003	0.003

## 4.5 Challenges with the search method and tools

In the previous sections, the output and results of using the autoencoder for anomaly detection have been shown. The method and results, as produced and shown in this thesis, have yielded some promising results, given the nature of the search method. However, it should be known what the challenges of the task actually are, to truly understand why the results only are promising, and not great compared to other search methods. The challenges can be divided into three main points, all of which are entangled together. The three challenges are listed below.

- Model independence
- Reconstruction error minimization
- Feature engineering

Model independence is the first challenge. By model it is here referring to signal models, which it self might be self explanatory, but is a incredibly difficult criteria to uphold. As mentioned in the theory section for the Standard Model, the Standard Model, all though very successful in certain predictions, lack the ability to explain a whole number of behavior around us, and so there have been made a large numbers of suggested solutions to the issues. These new models are often called extensions to the Standard Model, and all though mathematically consistent, not neccesarly physically possible. And even if they are physically possible, in other words, they adhere to certain fundamental physical principles, they still might not exsist, as several searches at ATLAS have excluded but never found any new physics. The search method is inherently biased as one assumes that the new physics looks like the signal, and thus do analysis, data preparation etc with that signal in mind. But we do not know, at all, what the new physics looks like, even the assumption that we are looking for particles are implying a bias that might not be true. From the collisions in the detector to the analysis, there are biased descisions, we cannot avoid them, but we can minimize them as much as possible, which is a goal with the search strategy in this thesis.

This leads us to the second challenge, which is reconstruction error minimization. The proposed method in this thesis is to learn the signature of the standard model so well, that even subtle anomalous behavior will be picked up by the analysis tool. The autoencoder learns the signature of the standard model, by mimimizing the reconstruction error, and then hopefully, the anomalous data will be picked up and skewed to the right end of the reconstructino error distribution in a signal region. One problem with this is that one first blinds oneself to signals that might be very, very similar to the standard model in some feature space, but with very low statistics. These events will for a given set of features, never be found.

The The third challenge is the choice of features. This thesis utilized the RMM structure by Chekanov et al, as it maximizes the amount of information in the input data by using almost completely uncorrolated features. However, as we do not know the signals we are looking for, there is no way to know if this choice is the optimal choice for new physics. In fact, even if we found an ideal set of features, based on some physical principles or something else, it is not trivial that the reconstruction error calculation should weight the error of each feature equally. It might very well be that some features are less important than others. Essentially, the goal is to optimize for a signal we do not know, using features we dont know are optimal,

and weighting them as "unbiased" as possible, simply taking the average, to dictate how the autoencoder learns and updates its internal weights and biases.



# Conclusion

Future work, more work



# Appendices



# Appendix A

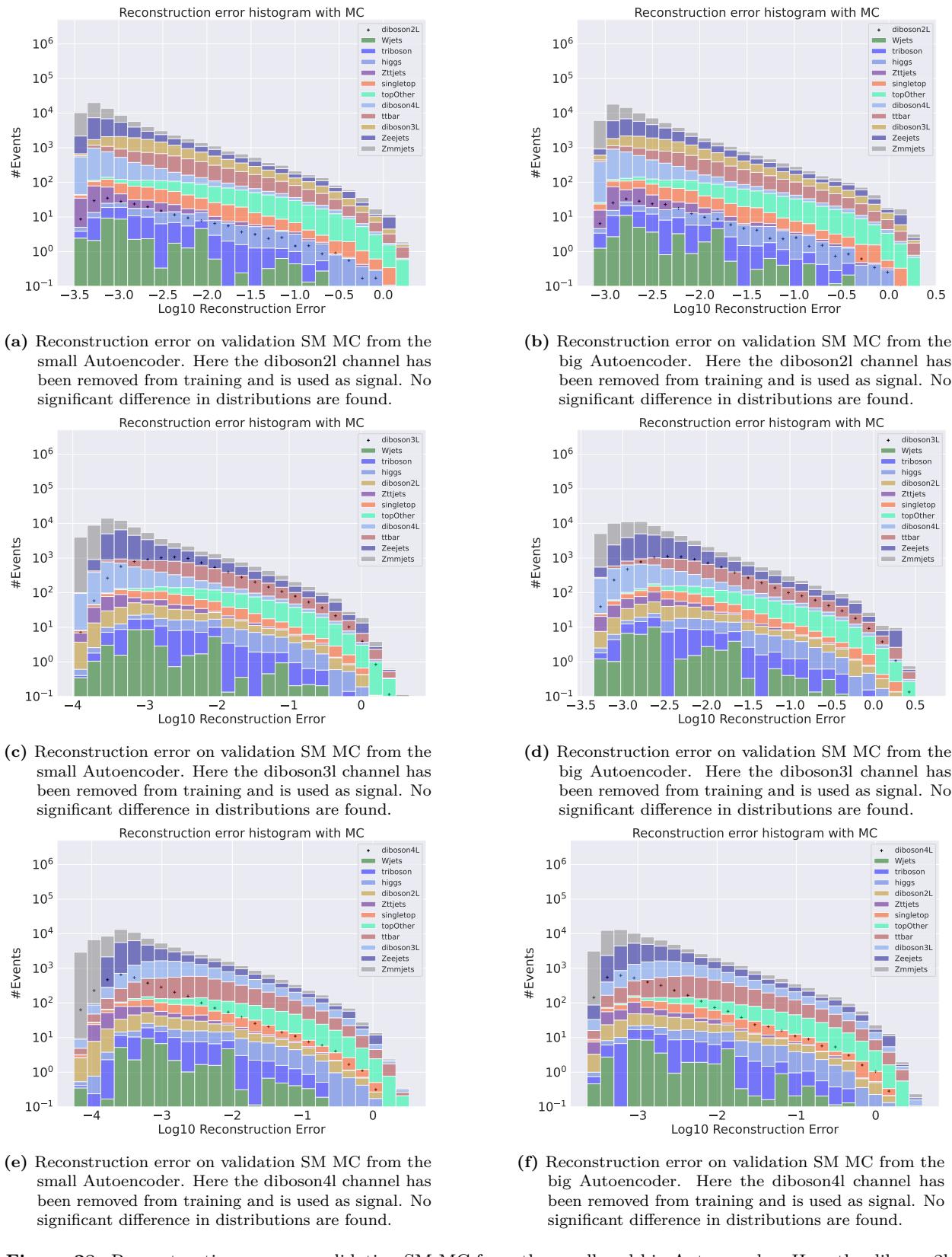


# Appendix B

Channel removal testing

Regular autoencoder

Variational autoencoder



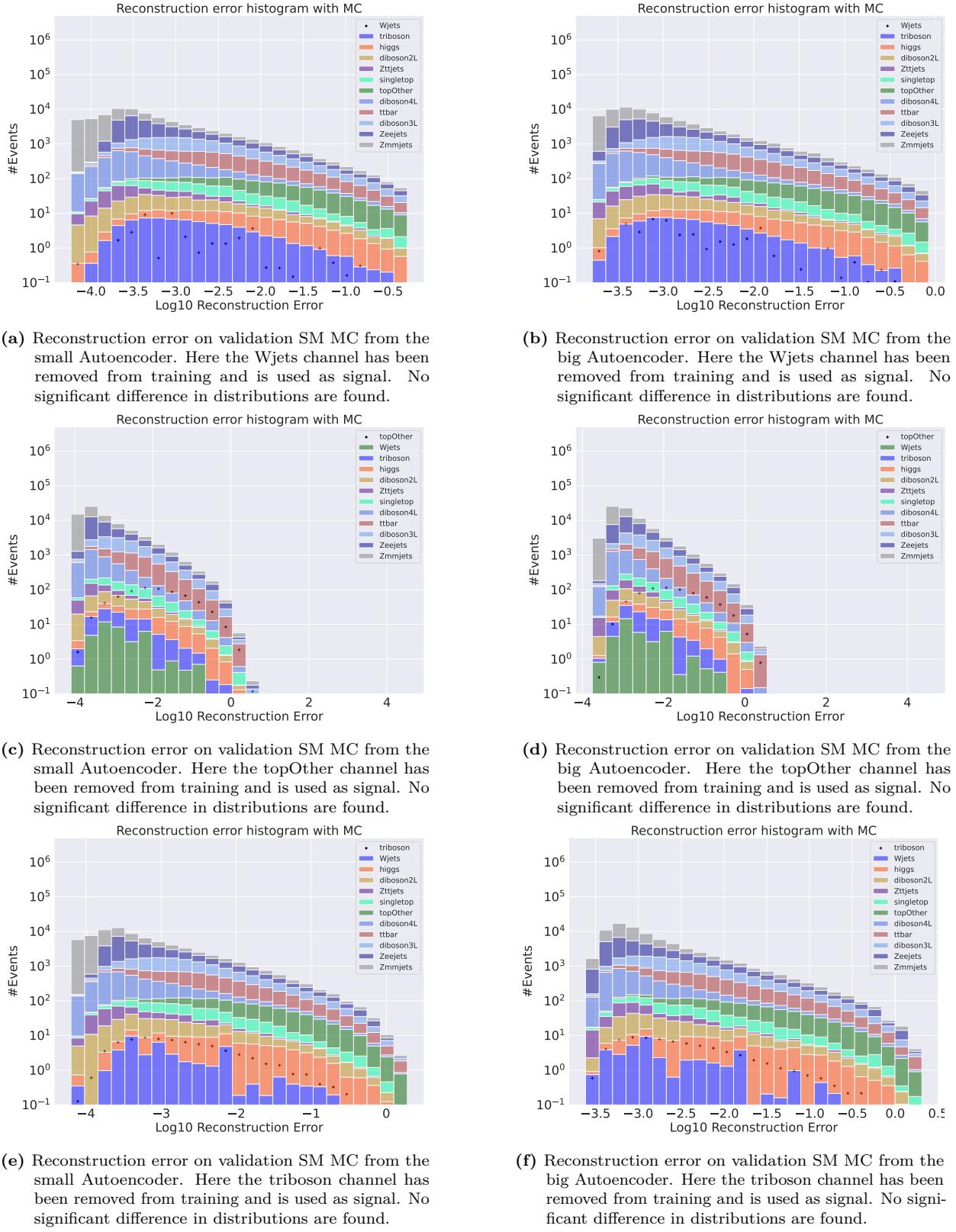


Figure 29

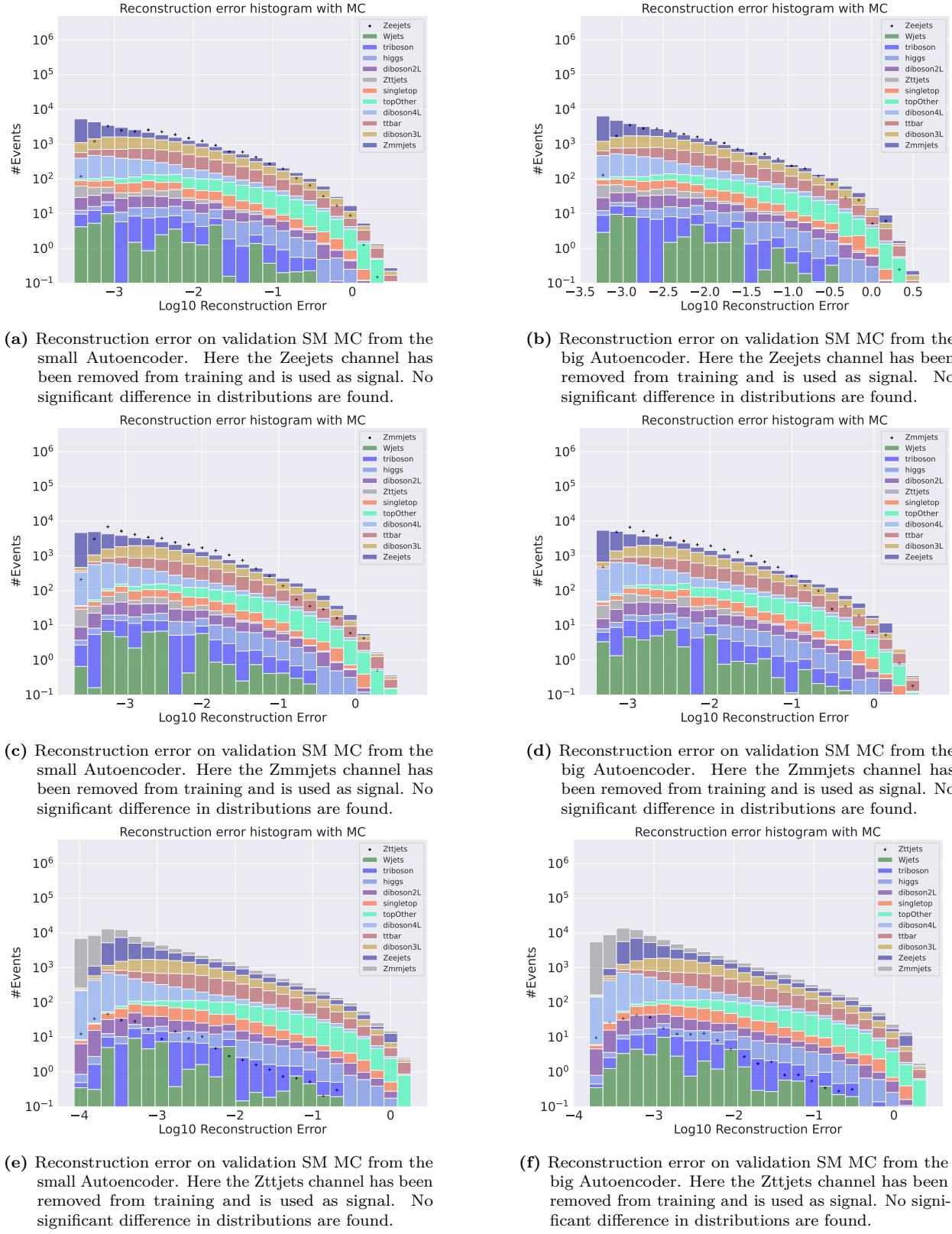


Figure 30

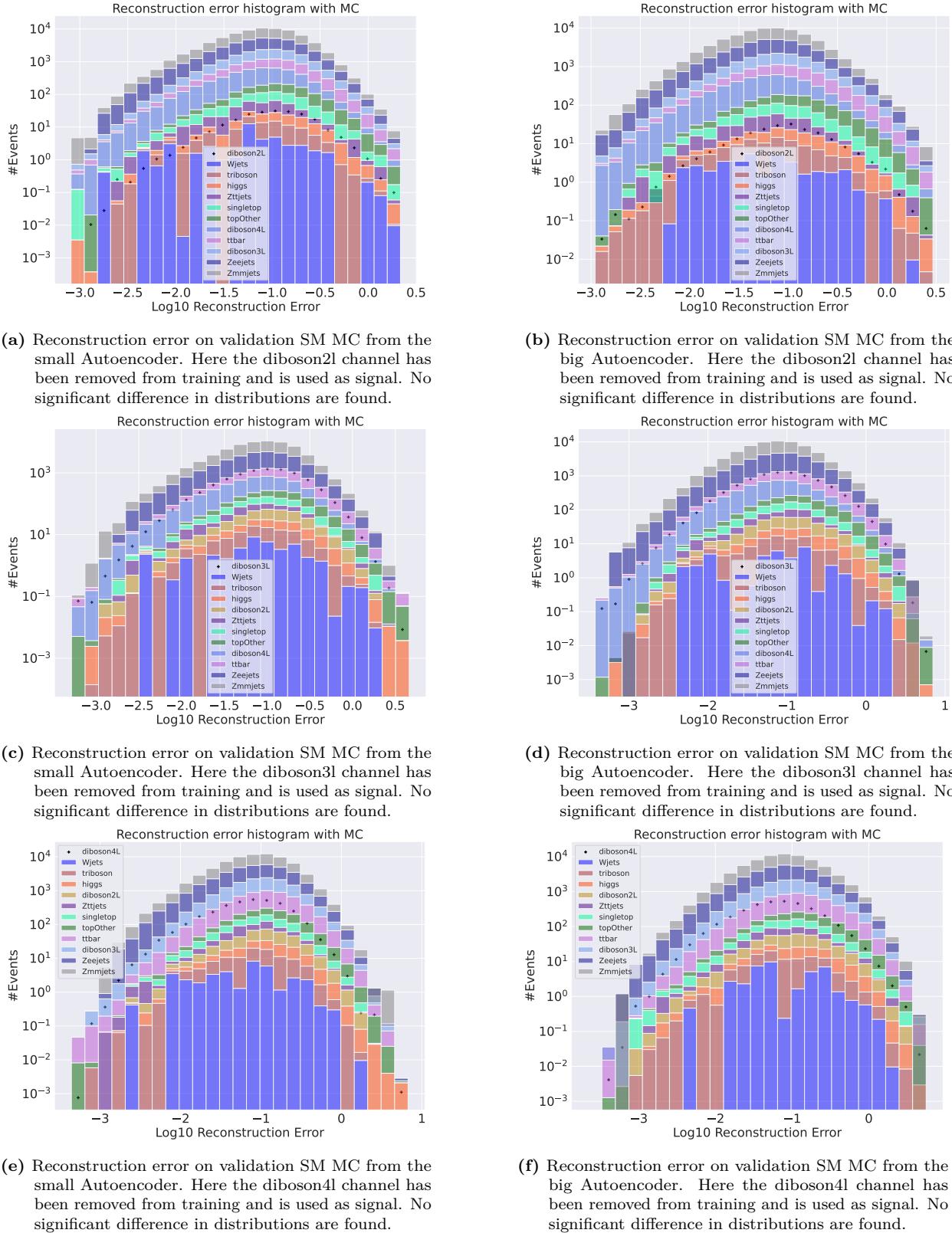


Figure 31

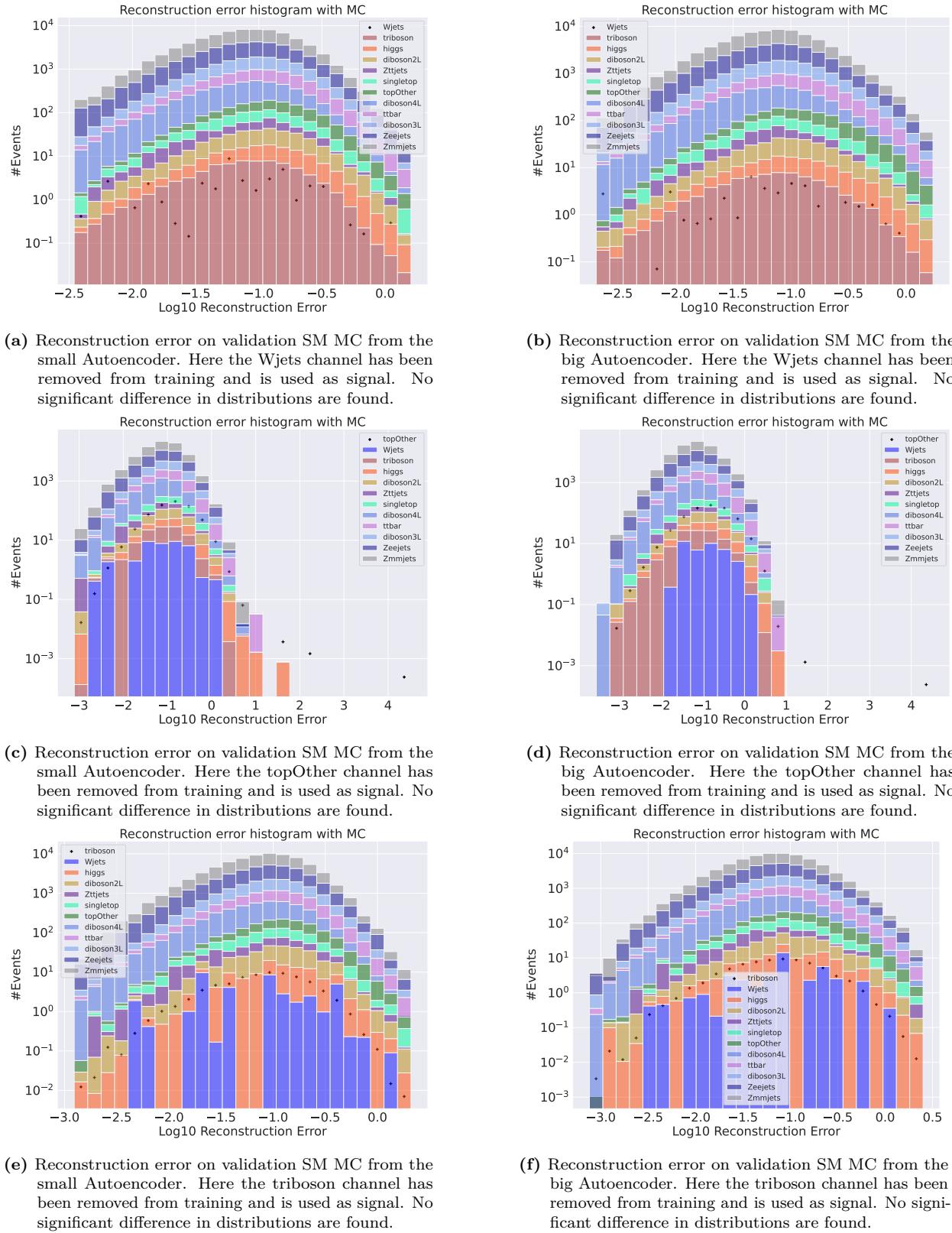


Figure 32

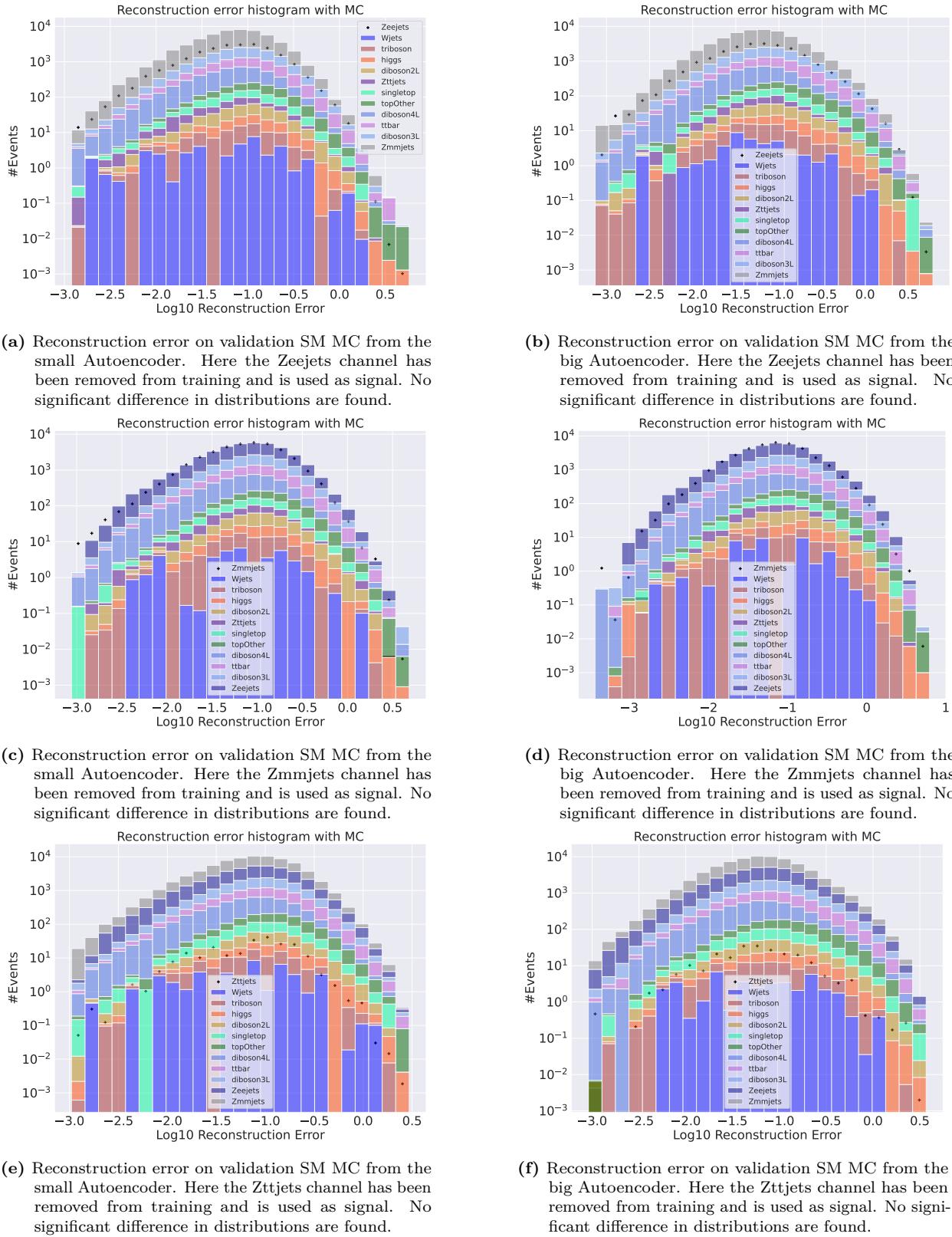


Figure 33