

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

Search for heavy neutrinos 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

Sakarias Garcia de Presno Frette

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

Search for heavy neutrinos at the LHC with the ATLAS
detector

Supervisors:
Professor Farid Ould-Saada
Dr. Eirik Gramstad
Dr. James Catmore

Abstract

Acknowledgments

- Veilledere
- William og Mikkel
- Foreldre
- NN

Contents

Introduction	1
1 Machine learning	3
1.1 Anomaly detection	3
1.2 Neural Networks	3
1.3 Autoencoders	8
2 Standard model phenomenology	11
2.1 Structure and composition of the Standard Model	12
2.2 Proposal model	13
3 Implementation	15
3.1 ATLAS	15
3.2 ROOT	15
3.3 The dataset features	18
3.4 Code implementation	20
4 Results	25
5 Discussion	27
Conclusion	29
Appendices	31
Appendix A	33
Appendix B	35

Introduction

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. Chapter five is dedicated to the 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

Machine learning

1.1 Anomaly detection

Anomaly detection is a tool with a wide range of uses, from time series data, fraud detection or anomalous sensor data. Its main purpose is to detect data which does not conform to some predetermined standard for normal behavior. The predetermined standard varies from situation to situation, both from the context it self and what is expected as an anomaly. Anomalies are typically classified in three categories [1]:

1. Point anomalies
2. Contextual anomalies
3. Collective anomalies

Point anomalies are singular or few outliers from a larger context or group. These anomalies can occur in many situations, are indeed quite important to detect. One such example is Michael Phelps. Phelps is famous for being one of the best swimmers of all time. Along with extensive training, planning and dedication, he has another tool that has helped him, he does not produce much lactic acid. In fact, his body produces so little that he can swim continuously and much more intensive than most other top swimmers. This ability is not common, infact it is very rare amongst humans, and can be considered a point anomaly. It is important to understand that point anomalies does not have to be singular occurrences. Rather they are extremely rare events that deviate alot from the expected behavior.

Contextual anomalies are another kind of anomalies, and are defined based on the context of the anomaly and data, rather than as a whole. Suppose you have have data on continous stream of gas in a pipe. The extraction of this gas is day dependent, to the point where the delivery on saturdays might oscillate between half and 3/4 of that of monday through friday, due to shorter work day. Should there one saturday suddenly flow the same amount as friday, an analyst think nothing of this, but due to this being a saturday, the context of this behavior dictates that this be categorized as a contextual anomaly.

The last type of anomaly described by Chandola, Banerjee and Kumar [1] are the collective anomalies. The collective anomalies are anomalies that as a group deviate from the expected behavior of the dataset. In particle physics these anomalies are the only type that are of interest. This is because there are so many sources for anomalous behavior in an experiment that only collective one are worth investigating. The major problem for such experiments is noise, and noise can be created from a large number of components. This alone is reason enough to only consider collective anomalies. Another reason is that certain processes in particle physics look much alike, but have different crossection, thus one process is much more likely to occur than another. This was one of the main issues with the discovery of Higgs, as Higgs has a background of

1.2 Neural Networks

There are several categories of statistical algorithms for data analysis within machine learning. Amongst them are neural networks, which have for the last decade exponentially been used within industry and academia for a number of usecases. From image analysis to weather prediction, these models are used extensively.

Neural networks, or feed forward neural networks (FFNN), are based on a few principles. First, the data is feeded forward through the network. The end output is evaluated in some fashion, and corrections are then back propagated through the network, updating the weights and biases. This "training" is done until a sufficient threshold is met. A general layout of a neural network is displayed in figure 1.1.

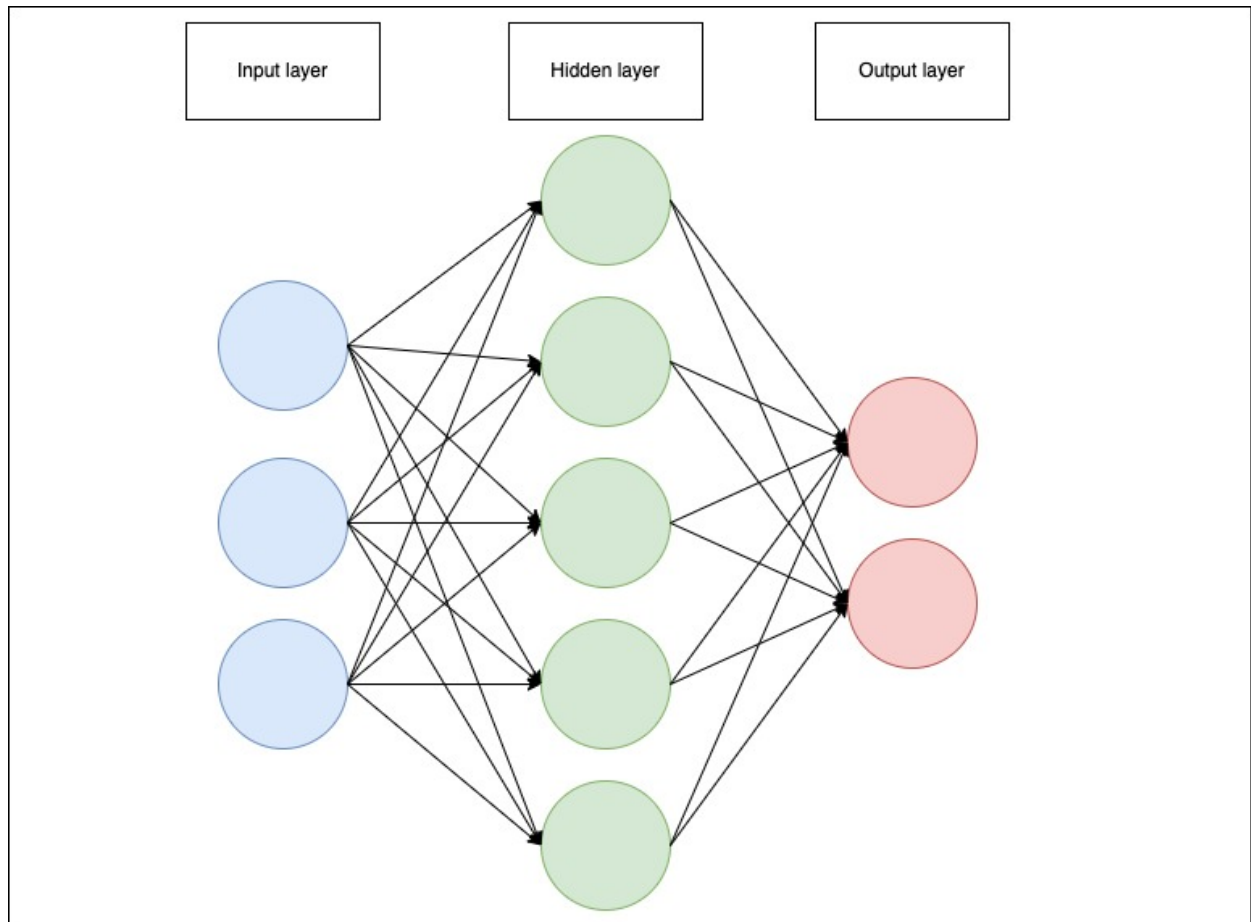


Figure 1.1: Simple neural network diagram drawn using Draw.io. Here the blue dots are the input layer, the green dots are a hidden layer, and the red dots are the output layer. The arrows show the connections between each node.

The input layer has the same shape of the dataset one uses to train or predict on, with one node for each feature in the dataset. The next layer is the hidden layers. For a given network, the amount of hidden layers can be tuned, as well as the number of nodes per layer. Finally, the last hidden layer is connected to the output layer, which is determined by the aim of the problem. In the case of figure 1.1, this neural network would represent a binary classification problem, in other words, two categories. The nodes in the network interacts through so called weights w and biases b . These are known as tunable parameters, which needs to be trained on the dataset before any prediction can be made.

In order to avoid confusion, we will adhere to table 1.1 for the notation used in the following sections.

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 k 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 .	
a_j^l	Activated node output on node j on layer l .	\mathbb{R}
Functions		
C	Cost function	
σ^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 [2]

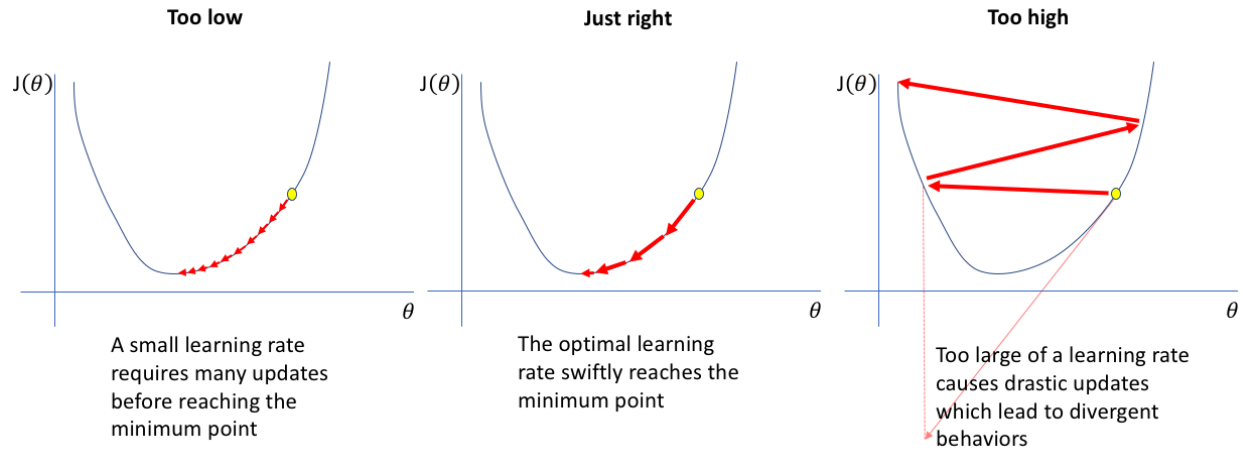


Figure 1.2: Figures showing different choice of learning rate for a given costfunction, with respect to the tunable parameters. Source: [Jeremy Jordan](#), accessed 03.10.22.

Gradient descent

Let us now consider a general n-dimansional problem, with parameters $\theta = \{\theta_1, \theta_2, \dots, \theta_n\}$. We want the set of θ such that we minimize a costfunction with respect to the data and target. One way to solve this problem is using ordinary least squares. For this approach, the optimal paramters θ_{opt} are derived from minimizing the cost function, as shown here:

$$\theta_{opt} = (X^T X)^{-1} X^T t,$$

where X is the design matrix containing the data, and 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 X . Thus, an alternative approch is to iteratively approximate the the ideal parameters.

Suppose we we have a cost function $C(\theta)$ for a given problem. We can approximate the minimum of the cost function by calculating the gradient $\nabla_{\theta} C$ with respect to θ . The negative of this gradient indicates the direction for the minimum of C when evaluating it in a specific point θ_i in the parameter space [2]. This is formulized as follows

$$\theta_{i+1} = \theta_i - \eta \nabla_{\theta} C(\theta_i), \quad (1.1)$$

where η is a step size, also called the learning rate. The choice of η is not a trivial case. It is one of several hyperparameters¹ 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 1.2.

Figure 1.2 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 prefered 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 local minima. To circumvent this issue, mini batches are introduced.

Feed forwarding

Inference (prediction) and training both use the same feed-forward algorithm. 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 [2]:

- The data is recieved by the input nodes in teh 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 .

¹Give reference to hyperparameters

- This value z is then sent through an activation function σ , 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[3]. This is a bit misleading, as the backprop algorithm actually only refers to how to compute the gradient[4]. The algorithm allows us to alter the weights and biases such that we get an ideal output. Assuming a costfunction 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:

Notethatthiscalculationisnotageneralizedalgorithmforbackpropagation, butratherforaspecialcaseusingMS

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

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

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

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:

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

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 δ^l be the vector of all the errors in the l 'th layer, and δ^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:

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

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:

$$\delta_j^l \equiv \frac{\partial C}{\partial z_j^l} = \sum_k \frac{\partial C}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial z_j^l} = \sum_k \frac{\partial z_k^{l+1}}{\partial z_j^l} \delta_k^{l+1}. \quad (1.3)$$

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

$$\frac{\partial z_k^{l+1}}{\partial z_j^l} = w_{j,k}^{l+1} \sigma'(z_j^l). \quad (1.4)$$

This allows us to substitute equation into equation to get the following expression:

$$\delta_j^l = \sum_k w_{j,k}^{l+1} \sigma'(z_j^l) \delta_k^{l+1}. \quad (1.5)$$

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

- Compute the local gradient for the last layer, δ^L .
- Recursively compute the local gradient for the remaining layers, δ^l for $l = L - 1, L - 2, \dots, 1$.
- Update the weights and biases for all layers, $l = 1, 2, \dots, L$. 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 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 denoising, illustrated in figure 1.3. Suppose you have a noised image, and want to denoise it. By feeding the encoder a noised image, and comparing the decoder output to the actual image, the autoencoder can tune itself to denoise images.

Mathematically this is represented as follows. Using the annotations of each component in figure 1.3 we have that the decoded information is defined as follows

$$\mathbf{c} = \mathbf{E}_\phi(\mathbf{x}),$$

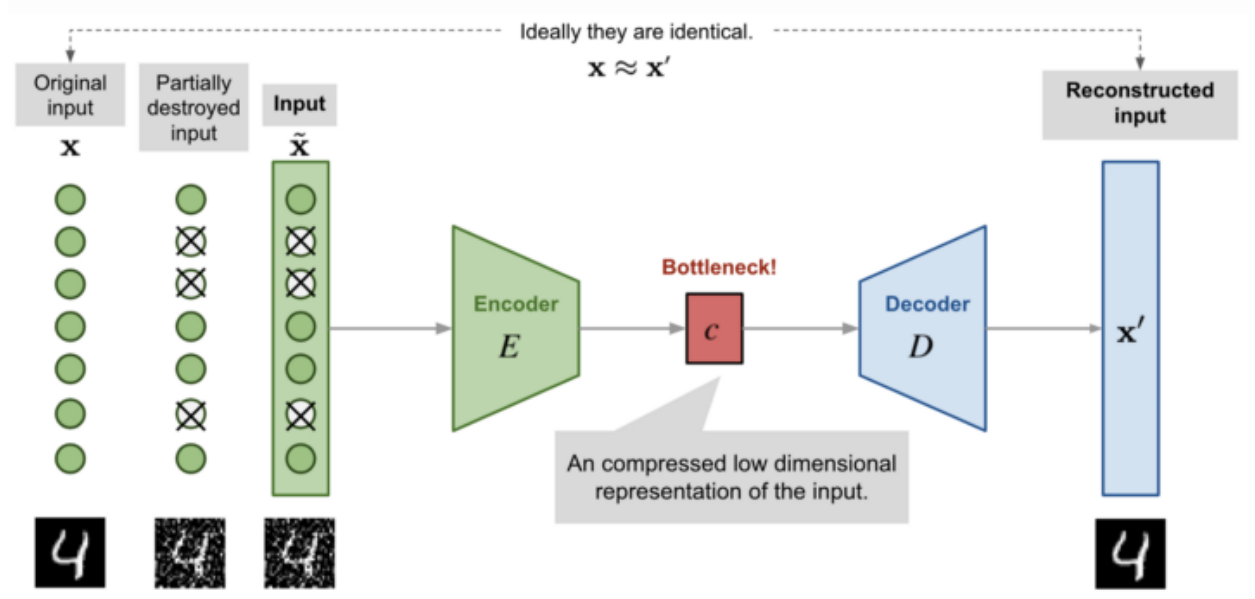


Figure 1.3: Figure depicting a model for an image denoising autoencoder. Here the input \mathbf{x} is the original image, $\tilde{\mathbf{x}}$ is a noisy version of \mathbf{x} , E is the encoder, D is the decoder, and c is the latent space. Found 27.09.22 [here](#).

and the reconstruction given as

$$\mathbf{x}' = \mathbf{D}_\theta(\mathbf{E}_\phi(\mathbf{x})).$$

The parameters (ϕ, θ) 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

$$L_{AE}(\phi, \theta) = \frac{1}{N} \sum_{i=0}^{N-1} (\mathbf{x}^i - \mathbf{D}_\theta(\mathbf{E}_\phi(\mathbf{x}^i)))^2. \quad (1.6)$$

Optimizers

Optimizing [5],

Chapter 2

Standard model phenomenology

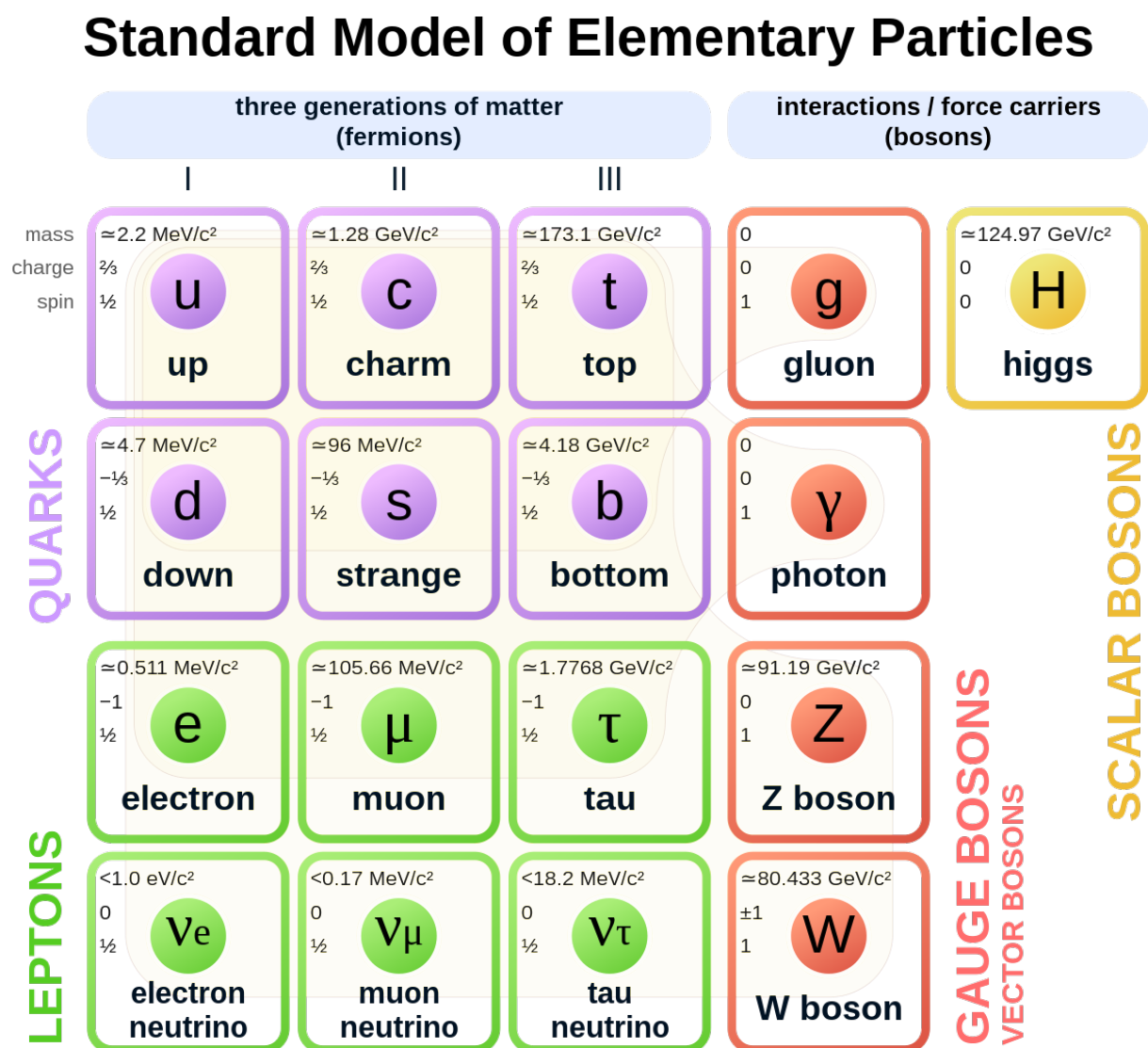


Figure 2.1: The standard model of elementary particles. Source [here](#). Accessed 07.10.22

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, nor the results or discussion of them. For a more technical explanation, read (Pich, 2008)[6] for a well written paper containing the 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)[7] is highly recommended. Finally, for more understanding of particle physics, the nature of interactions and more, see Thomson's "Modern Particle physics" (Thomson, 2013)[8].

The standard model is to physicists what the periodic table is to chemists. It is comprised of two parent class particles, fermions and bosons, where fermions are comprised of quarks and leptons. The model contains 17 particles, 6 quarks, 6 leptons and 5 bosons, and are shown in figure 2.1.

Fermions

The fermions are the building blocks of matter, and contain two types of particles, leptons and quarks. Together they form protons and neutrons, atoms all around us. Fermions, unlike bosons, are spin half particles, and are also the only particles to have anti 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 ν 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 fermions.

Quarks are fractional charge particles, with defined charge of either $2/3$ or $-1/3$. 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.

Bosons

Bosons are integer number spin particles, with spin $0, 1, 2, \dots$. Within bosons there are so called elementary bosons, some of which are force carriers or mediators such as the W, Z and the photon. The Higgs boson is also an elementary boson, but is not a force carrier. It provides masses for the fermions via spontaneous symmetry breaking[6]. Other bosons are so called composite bosons, which are particles constructed by an even amount of fermions yielding the integer spin.

Left and right handedness

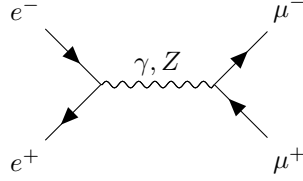
Particles in the standard model are subject to a quantum mechanical property called chirality. Chirality is a property that describes a particle's ability to be superimposed on its mirror image. If a particle has chirality, it cannot be superimposed on its mirror image by any combination of translation, 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[6][9], and each component can be linked to a part in the Lagrangian for the system.

In figure 2.3 we have a Feynman diagram describing two scenarios: electron scattering into muon pairs and electron muon scattering into electron muon. This is because we have not yet set the direction for time evolution. In this thesis, all diagrams will be interpreted from left to right, i.e figure 2.3 then only represents electron scattering into muon pair. The diagram contains all the components in the Lagrangian, and arrows,

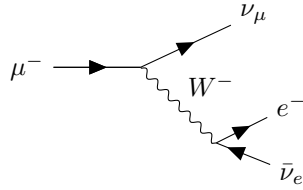
Figure 2.2: Feynman diagram of muon pair production from electron scattering. Here, both Z and γ can work as the propagator.



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) or particle, showing the momentum direction essentially. The is often also a propagator between the input and the output of such processes, 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[6], 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 below.

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.

Some limitations

All though the standard model have 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 explain gravity in a quantized way. There are models that try 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[8]:

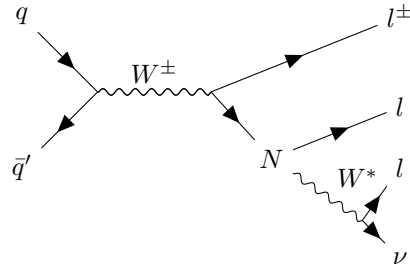
$$x \rightarrow -x. \quad (2.1)$$

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

This asymmetry is strange, and hints towards new physics that perhaps can restore the parity breaking. Another note to make is that it has been experimentally verified that the neutrinos are massive[10], with an upper limit on the mass for the anti electron neutrino of $m_\nu < 0.8 \text{ eV} c^2$ at 90% confidence level. This is somewhat problematic, as the masses of the neutrinos are not predicted by the standard model.

2.2 Proposal model

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 CERN, thus evidence of such a boson would come from the detection



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.

Chapter 3

Implementation

3.1 ATLAS

, [11], [12]

Data collection

The ATLAS detector three 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[13]. [14]

Triggers

Data preparation

3.2 ROOT

It is a lot. [15]

ROOT, memory management etc.

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

[16]

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 lazily² apply constraints 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

²In this context lazily means that the functions and or cuts are done first after all have been registered, see [ROOT guidelines](#) for more.

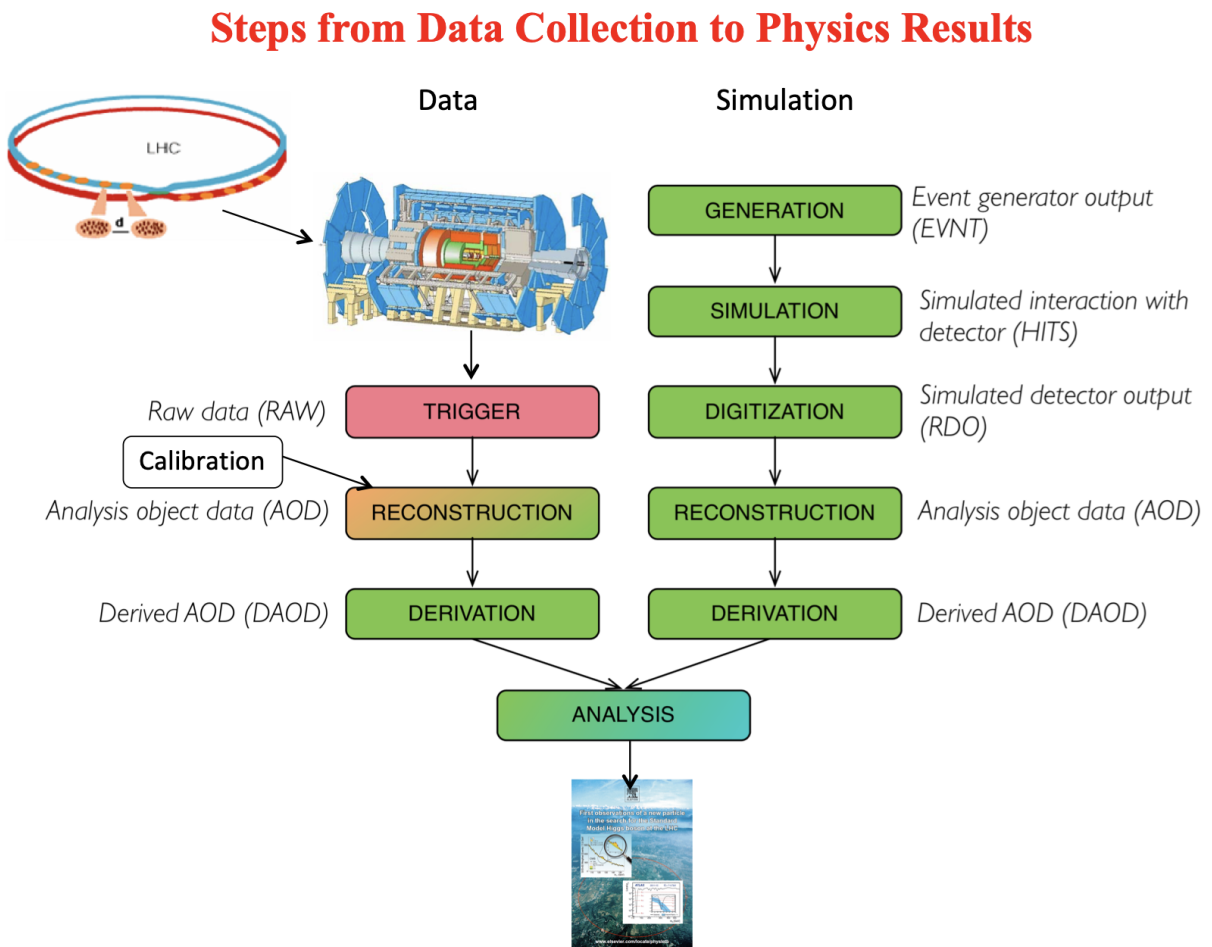


Figure 3.1: Figure describing the steps to take for data collection at ATLAS, fetched from [Hybrid ATLAS Induction Day + Software Tutorial workshop](#), part [Computing and Data preparation](#), held by S.M Wang [13] .

given channel³ 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.

```

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"h_flcomp_{k}",
39             len(fldic.keys()),
40             0,
41             len(fldic.keys()),
42         ),
43         "flcomp",
44         "wgt_SG",
45     )
46

```

```

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();

```

³A 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.

```

21
22     return inv_mass;
23 }

```

Using ROOT we can create Lorentzvectors to calculate a number of properties, such as the invariant mass of two particles.

```

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 a dataframe has been created, the columns chosen and histograms are drawn, the dataframe can be converted to a pandas dataframe, which is a very popular choice for data structure when doing data analysis in python. Here the new pandas dataframe is stored as hdf5[17] files, for later use.

3.3 The dataset features

The RMM matrix

Most of the features in the analysis are elements in the so called Rapidity-Mass (RMM) matrix inspired by the work of Chekanov [18].

!! Motivation for using such a matrix in machine learning \rightarrow hint to highly uncorrelated feats!!

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. An example is shown in matrix 3.1.

$$\begin{pmatrix}
 e_T^{miss} & m_T(j_1) & m_T(j_2) & m_T(e_1) & m_T(e_2) \\
 h_L(j_1) & e_T(j_1) & m(j_1, j_2) & m(j_1, e_1) & m(j_1, e_2) \\
 h_L(j_2) & h(j_2, j_1) & \delta e_T(j_2) & m(j_2, e_1) & m(j_2, e_2) \\
 h_L(e_1) & h(e_1, j_1) & h(e_1, j_2) & e_T(e_1) & m(e_1, e_2) \\
 h_L(e_2) & h(e_2, j_1) & h(e_2, j_2) & h(e_2, e_1) & \delta e_T(j_2)
 \end{pmatrix} \quad (3.1)$$

In matrix 3.1 we have the RMM matrix for a T2N2 system, in other words we have two types of particles, jets 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 δe_T . e_T^{miss} is placed in the (0,0) in the matrix. It accounts for the missing 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

$$e_T = \sqrt{m^2 + p_T^2}$$

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

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

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

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

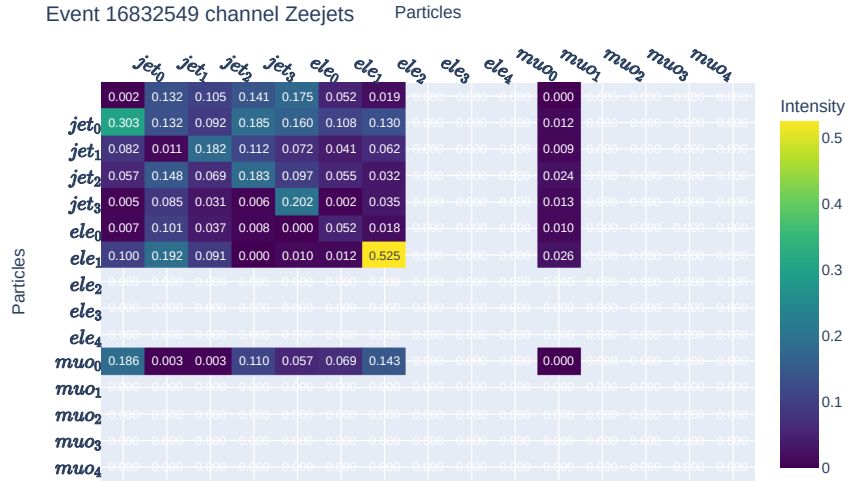
where C is a constant to ensure that the average $h_L(i_n)$ values do not deviate too much from the ranges of the invariant masses of the transverse masses, found to be 0.15[18]. y is the rapidity of the particle, and i_n is the particle number. On the lower triangle we have the longitudinal properties 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 .

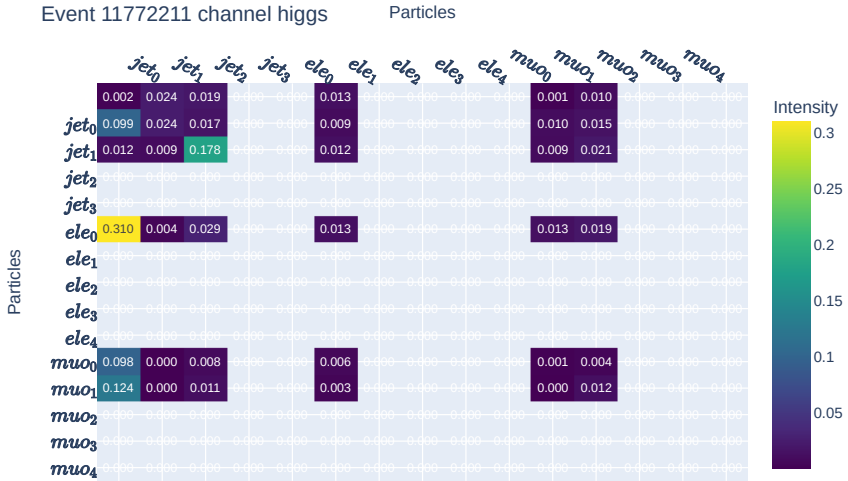
Implementation of the RMM matrix

An example of the RMM matrices used in this thesis is shown in figure 3.2 below:



Loading [MathJax]/extensions/MathMenu.js

- (a) RMM matrix for event number 16832549 from the MonteCarlo Zeejets sample. Each feature is scaled based on a fit for that feature for all events in the training set ($\approx 80\%$ of total MC)



Loading [MathJax]/extensions/MathMenu.js

- (b) RMM matrix for event number 11772211 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)

Figure 3.2

In figure 3.2 we see two RMM matrices created from two different channels in the MonteCarlo samples. This RMM is of type T3N5⁴. These RMM matrices have scaled the features that a minimum value less than -3 and or a maximum value of above 3 . These scalers were fitted to the specific value using the `".fit_transform()"` and `".transform()"` functions from the Scikit-learn library[19]. For easier interpretability, the gray area corresponds to a missing value, leading to so called "islands" in the RMM matrix.

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 training samples we use are actually useful. As described by R. Stuart Geiger et al. [20], 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 learns 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 the are numerical approximations, and can sometimes be off. Thus, every feature that well 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. Assuming that there are no issues with the data collection from ATLAS, we then assume that the data from ATLAS is the "ground truth", as it is the only dataset we can analyze, and the only dataset we can compare to. The MonteCarlo is then compared to the data, and if the distributions are similar, we can assume that the training samples are good enough.

In figure 3.3 two features have been selected to vizualize the comparison between Monte Carlo and ATLAS data, e_T^{miss} and $flcomp$. We see that both e_T^{miss} and $flcomp$ 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 appendix B.

Tabular data

3.4 Code implementation

The machine learning analysis was written with Keras[21] using the Tensorflow api[22]. The machine learning structure was written using a functional structure⁵. In practise, this model could just as well have been written as a Sequential model⁶, 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.

```

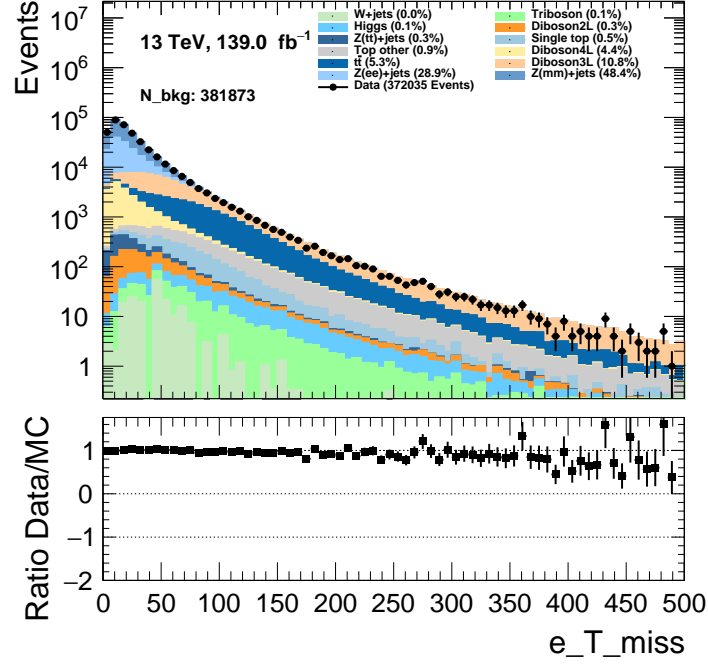
1 import tensorflow as tf
2
3
4 inputs = tf.keras.layers.Input(shape=data_shape, name="input")
5
6 # First hidden layer
7 First_layer = tf.keras.layers.Dense(
8     units=30,
9     activation="relu"
10 )(inputs)
11
12 # Second hidden layer
13 Second_layer = tf.keras.layers.Dense(
14     units=45,
15     activation="relu"
16 )(First_layer)
17

```

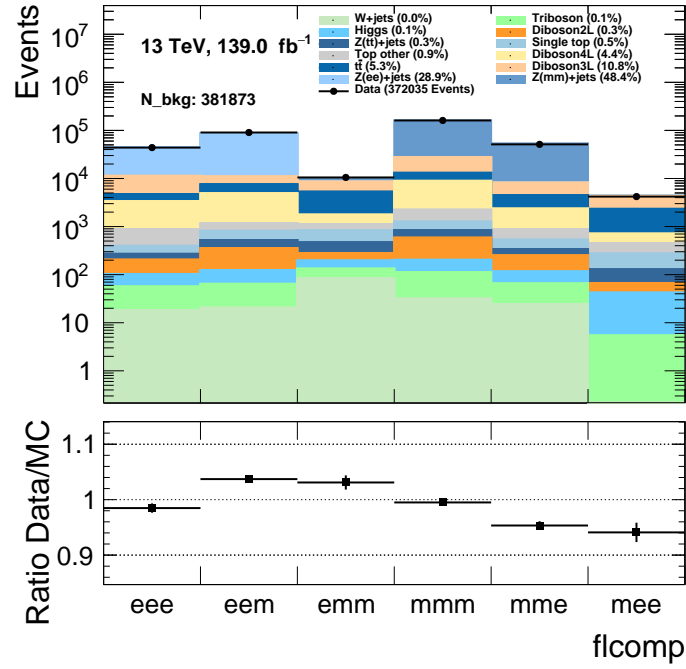
⁴T3 \rightarrow 3 particle types: jets, electrons and muons. N5 \rightarrow 5 particles per particle type. Note here that we have 5 particles only for the leptons, and 4 particles for the jets.

⁵Functional 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 $a \rightarrow b$. This allows for more flexible structures. More on the functional api can be found [here](#).

⁶Sequential structure adds layers in sequence, i.e for layers a, b, c we have that $a \rightarrow b \rightarrow c$, with a strict structure. This allows for more organized code. More on sequential models can be found [here](#).



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



(b) Flavor combination for the three lepton final state. This histogram only contains the flavor combinations for the good leptons, denoted lep_{SG} . The histogram contains the entire Run 2 dataset.

Figure 3.3: Comparison of the MonteCarlo and data for the three lepton final state with the features e_T^{miss} and flavor composition.

```

18 # Second hidden 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 this example, using a 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"])

```

In the case of this project, we want to create an autoencoder, as shown in figure 1.3. A general network architecture is proposed below, where the nodes, activation functions, and other hyperparameters can be tuned.

```

1 import tensorflow as tf
2
3 class HyperParameterTuning(RunAE):
4     def __init__(self, data_structure: object, path: str)->None:
5         super().__init__(data_structure, path)
6
7     def AE_model_builder(self, hp: kt.engine.hyperparameters.HyperParameters):
8
9         """_summary_
10
11         Args:
12             hp (kt.engine.hyperparameters.HyperParameters): _description_
13
14         Returns:
15             _type_: _description_
16         """
17         ker_choice = hp.Choice("Kernel_reg", values=[0.5, 0.1, 0.05, 0.01])
18         act_choice = hp.Choice("Atc_reg", values=[0.5, 0.1, 0.05, 0.01])
19
20         alpha_choice = hp.Choice("alpha", values=[1.0, 0.5, 0.1, 0.05, 0.01])
21
22         # Activation functions
23         activations = {
24             "relu": tf.nn.relu,
25             "tanh": tf.nn.tanh,
26             "leakyrelu": "leaky_relu",
27             "linear": tf.keras.activations.linear,
28         } # lambda x: tf.nn.leaky_relu(x, alpha=alpha_choice),
29
30         # Input layer
31         inputs = tf.keras.layers.Input(shape=self.data_shape, name="encoder_input")
32
33         # First hidden layer
34         x = tf.keras.layers.Dense(
35             units=hp.Int(
36                 "num_of_neurons1", min_value=60, max_value=self.data_shape - 1, step=1
37             ),
38             activation=activations.get(
39                 hp.Choice("1_act", ["relu", "tanh", "leakyrelu", "linear"])
40             ),
41             kernel_regularizer=tf.keras.regularizers.L1(ker_choice),
42             activity_regularizer=tf.keras.regularizers.L2(act_choice),
43         )(inputs)
44

```



```

45     # Second hidden layer
46     x_ = tf.keras.layers.Dense(
47         units=hp.Int("num_of_neurons2", min_value=30, max_value=59, step=1),
48         activation=activations.get(
49             hp.Choice("2_act", ["relu", "tanh", "leakyrelu", "linear"])
50         ),
51     )(x)
52
53     # Third hidden layer
54     x1 = tf.keras.layers.Dense(
55         units=hp.Int("num_of_neurons3", min_value=10, max_value=29, step=1),
56         activation=activations.get(
57             hp.Choice("3_act", ["relu", "tanh", "leakyrelu", "linear"])
58         ),
59         kernel_regularizer=tf.keras.regularizers.L1(ker_choice),
60         activity_regularizer=tf.keras.regularizers.L2(act_choice),
61     )(x_)
62
63     val = hp.Int("lat_num", min_value=1, max_value=9, step=1)
64
65     # Forth hidden layer
66     x2 = tf.keras.layers.Dense(
67         units=val,
68         activation=activations.get(
69             hp.Choice("4_act", ["relu", "tanh", "leakyrelu", "linear"])
70         ),
71     )(x1)
72
73     # Encoder definition
74     encoder = tf.keras.Model(inputs, x2, name="encoder")
75
76     # Latent space
77     latent_input = tf.keras.layers.Input(shape=val, name="decoder_input")
78
79     # Fifth hidden layer
80     x = tf.keras.layers.Dense(
81         units=hp.Int("num_of_neurons5", min_value=10, max_value=29, step=1),
82         activation=activations.get(
83             hp.Choice("5_act", ["relu", "tanh", "leakyrelu", "linear"])
84         ),
85         kernel_regularizer=tf.keras.regularizers.L1(ker_choice),
86         activity_regularizer=tf.keras.regularizers.L2(act_choice),
87     )(latent_input)
88
89     # Sixth hidden layer
90     x_ = tf.keras.layers.Dense(
91         units=hp.Int("num_of_neurons6", min_value=30, max_value=59, step=1),
92         activation=activations.get(
93             hp.Choice("6_act", ["relu", "tanh", "leakyrelu", "linear"])
94         ),
95     )(x)
96
97     # Seventh hidden layer
98     x1 = tf.keras.layers.Dense(
99         units=hp.Int(
100             "num_of_neurons7", min_value=60, max_value=self.data_shape - 1, step=1
101         ),
102         activation=activations.get(
103             hp.Choice("7_act", ["relu", "tanh", "leakyrelu", "linear"])
104         ),
105         kernel_regularizer=tf.keras.regularizers.L1(ker_choice),
106         activity_regularizer=tf.keras.regularizers.L2(act_choice),
107     )(x_)
108
109     # Output layer
110     output = tf.keras.layers.Dense(
111         self.data_shape,
112         activation=activations.get(
113             hp.Choice("8_act", ["relu", "tanh", "leakyrelu", "linear"])
114         ),
115     )(x1)
116
117     # Encoder definition
118     decoder = tf.keras.Model(latent_input, output, name="decoder")
119

```

```
120     # Output definition
121     outputs = decoder(encoder(inputs))
122
123     # Model definition
124     AE_model = tf.keras.Model(inputs, outputs, name="AE_model")
125
126     hp_learning_rate = hp.Choice(
127         "learning_rate", values=[9e-2, 9.5e-2, 1e-3, 1.5e-3]
128     )
129     optimizer = tf.keras.optimizers.Adam(hp_learning_rate)
130
131     AE_model.compile(loss="mse", optimizer=optimizer, metrics=["mse"])
132
133     return AE_model
134
```

This function creates a tensorflow.keras model that has tuneable structure, thus allowing for optimized tuning with the Keras-Tuner library[23].

Chapter 4

Results

Chapter 5

Discussion

Conclusion

Future work, more work

Appendices

Appendix A

Appendix B

Feature checks between Monte Carlo and ATLAS data

Bibliography

- [1] V. Chandola, A. Banerjee and V. Kumar, *Anomaly detection: A survey*, *ACM Comput. Surv.* **41** (07, 2009) .
- [2] S. Frette, W. Hirst and M. M. Jensen, *A computational analysis of a dense feed forward neural network for regression and classification type problems in comparison to regression methods*, .
- [3] D. Rumelhart, G. Hinton and R. Williams, *Learning representations by back-propagating errors*, *Nature* (1986) .
- [4] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016.
- [5] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. 10.48550/ARXIV.1412.6980.
- [6] A. Pich, *The Standard Model of Electroweak Interactions; rev. version*, .
- [7] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory*. Westview Press, 1995.
- [8] M. Thomson, *Modern particle physics*. Cambridge University Press, New York, 2013.
- [9] K. Kumericki, *Feynman diagrams for beginners*, 2016. 10.48550/ARXIV.1602.04182.
- [10] M. Aker, A. Beglarian, J. Behrens, A. Berlev, U. Besserer, B. Bieringer et al., *Direct neutrino-mass measurement with sub-electronvolt sensitivity*, *Nature Physics* **18** (Feb, 2022) 160–166.
- [11] The ATLAS collaboration, *Operation of the ATLAS trigger system in run 2*, *Journal of Instrumentation* **15** (oct, 2020) P10004–P10004.
- [12] ATLAS collaboration, R. E. Owen, *The ATLAS Trigger System*, .
- [13] S.-M. Wang, *ATLAS Computing and Data Preparation. ATLAS Induction Day + Software Tutorial*, .
- [14] C. Bernius, *ATLAS Trigger and Data Acquisition. ATLAS Induction Day + Software Tutorial*, .
- [15] R. Brun, F. Rademakers, P. Canal, A. Naumann, O. Couet, L. Moneta et al., *root-project/root: v6.18/02*, Aug., 2019. 10.5281/zenodo.3895860.
- [16] E. Manca and E. Guiraud, *Using RDataFrame, ROOT’s declarative analysis tool, in a CMS physics study. Using RDataFrame, ROOT’s declarative analysis tool, in a CMS physics study*, .
- [17] The HDF Group, *Hierarchical Data Format, version 5*, 1997-2022.
- [18] S. Chekanov, *Imaging particle collision data for event classification using machine learning*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **931** (jul, 2019) 92–99.
- [19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825–2830.
- [20] R. S. Geiger, D. Cope, J. Ip, M. Lotosh, A. Shah, J. Weng et al., *”garbage in, garbage out” revisited: What do machine learning application papers report about human-labeled training data?*, *CoRR* **abs/2107.02278** (2021) , [2107.02278].
- [21] F. Chollet et al., *Keras*, 2015.
- [22] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015.
- [23] T. O’Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi et al., *”Kerastuner.”* <https://github.com/keras-team/keras-tuner>, 2019.