Semester Project Update #1 – The Fantastic Four

Portia Allen, Connor Caldwell, Arnaud Filliat, Johnny Murphy

Data Information

The dataset that we intend to work with is the Particle Tracking dataset from the Large Hadron Collider at CERN. Since this data was made publicly available by CERN for the TrackML Particle Tracking Challenge in 2018, there are no restrictions on our use or sharing of the data.

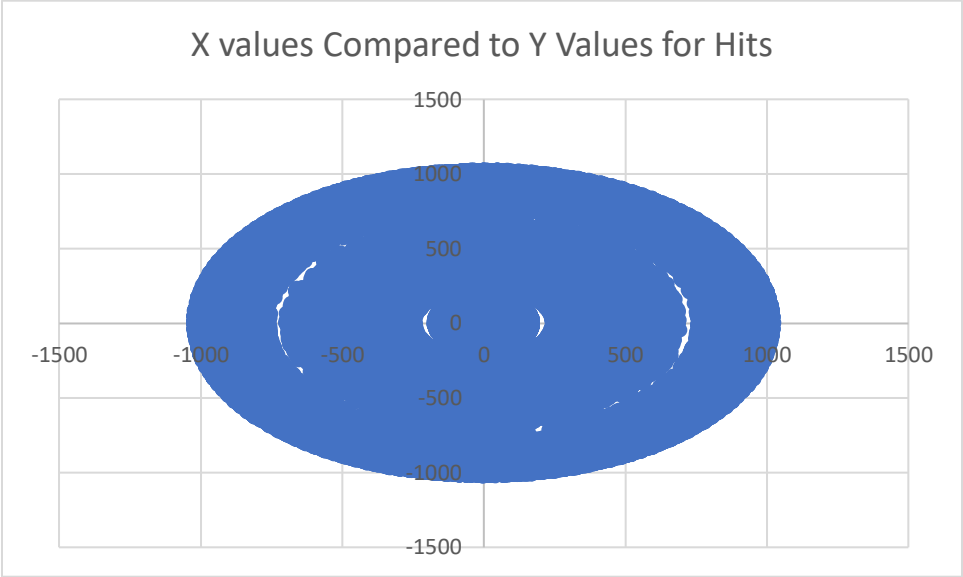
We are given a test dataset with 125 collision events upon which to evaluate our model; a training dataset, 8850 events split into five different files, as well as a sample set (the first 100 events from the training dataset); and some informational files about the geometry of the detectors. These are all provided by CERN in zipped .csv files, that can be downloaded and then read into a Pandas DataFrame for ease of access and manipulability. Additionally, CERN provides a trackml python library in GitHub to simplify some of the data handling, which will most likely be used extensively to aid in data visualization and processing.

Each event has four associated files, containing the hits, the hit cells, the particles, and the ground truth for the event. The hit files contain the identification numbers for the hit itself and the detector group/layer/module location of the hit, as well as the x-y-z coordinates for the hit. The truth files have the hit identification number, the particle identification number, the true hit location and particle momentum, and the weight of the hit (for the scoring metric). The particle files contain the particle ID number, the particle type, initial position and momentum, charge of the particle, and number of hits from this particle. The hit cell files contain the hit ID number, how much charge a particle has deposited on the cell, and the cell coordinates. The cells are the smallest positional identifier on the detectors and can be used to track association more accurately between hits and particles.

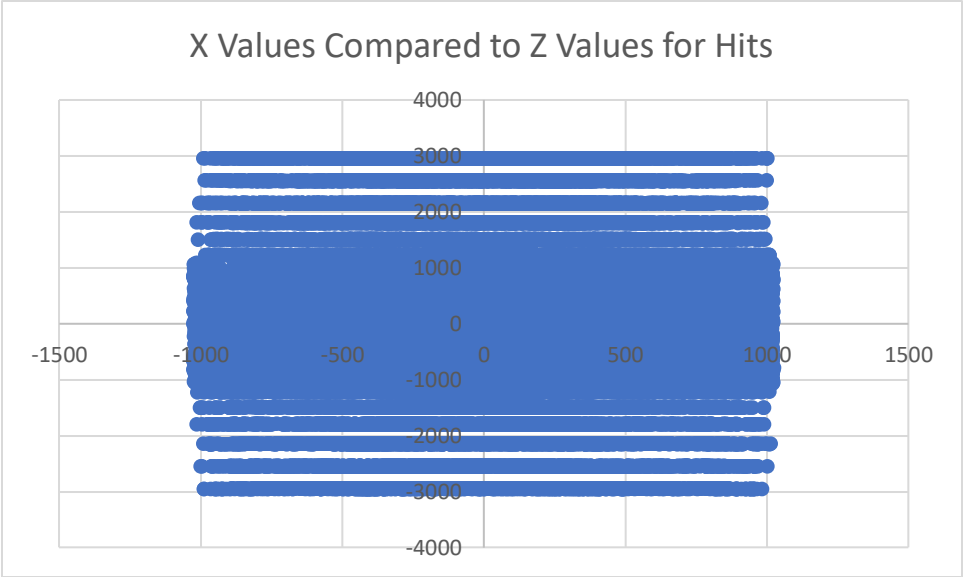
The geometry of the detectors is important information to know because the detector is built from concentric silicon slabs that have been subdivided several times. The largest groups are volumes, subdivided into layers, which are then divided into modules, which are made up of cells. Each of those have ID numbers except the cells, which have a gridded identification system. Each module has a different local position and orientation, so a transformation must be made between the local coordinates of the hit and the global coordinates of the hit to get the actual path of the particle.

Initial Visualizations

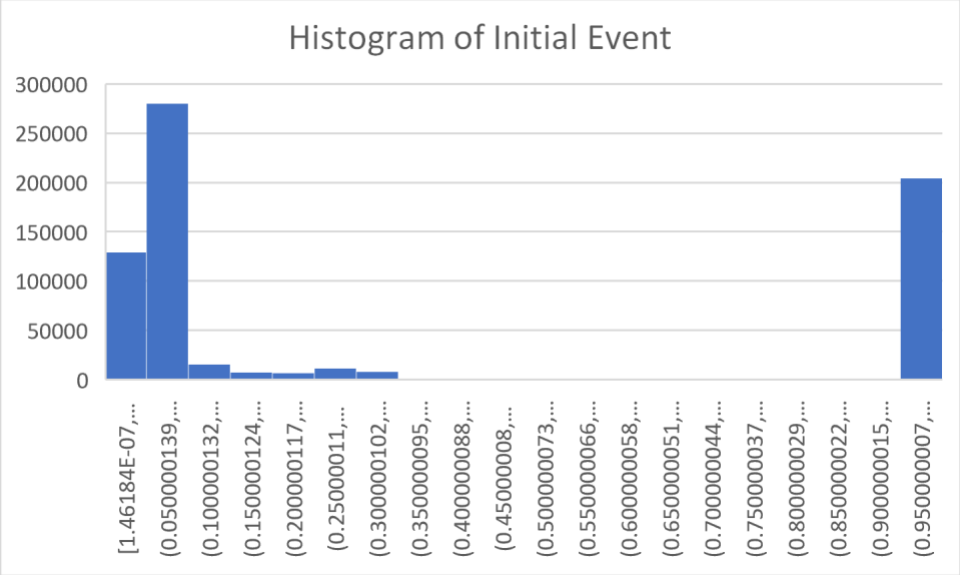
Below is a graph of the X values compared to the Y values in one event. As you can see this looks like a sphere which is expected since the CERN is a cylindrical model, and this allows for us to view the points in which a data point can be constrained in, and then shows that there are a couple of rings where it looks like some events where not recorded, which could help us in our training in order to solve some issues that could be a cause for concern in this model.



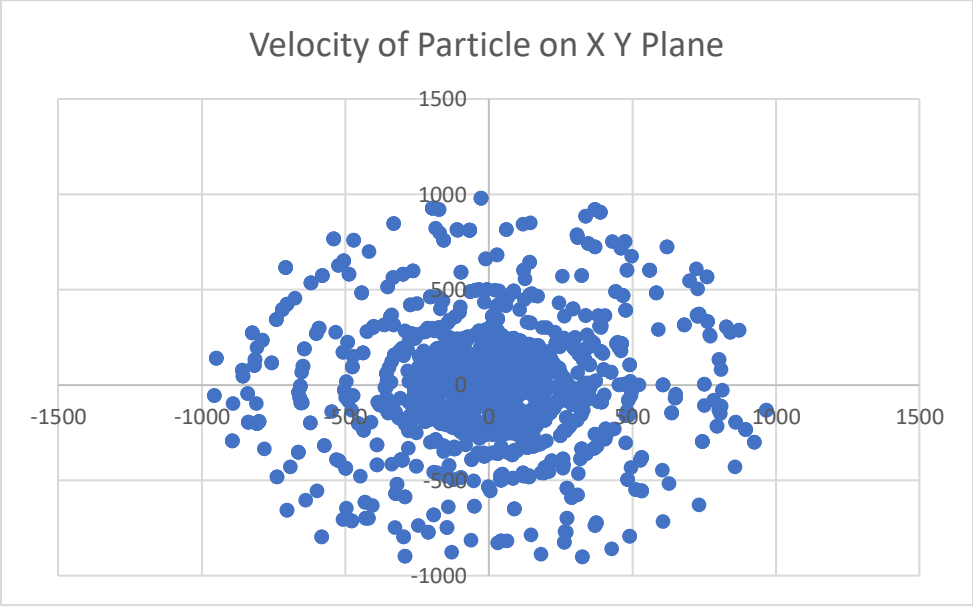
In this visualization below, it is shown that the X Values compared with the y values are in a cylindrical shape, which is expected. This shows that there are certain heights that were omitted in the measurements, which could be a cause for concern when trying to classify particles within the models.



This histogram below shows the values of this hit as a histogram with about 20 bins. This shows that the data is very skewed to the right, which means a certain particle could be much heavier, classifying it as something else, such as the Higgs Boson. The data on the left can be abstracted to particles such as beta particles, which are lighter and much more abundant, same with Z particles as well.



The graph below displays the particle's velocity on an X and Y plane, this displays that the particles were going very rapidly and randomly in a circle. With this, this allows for us to analyze the total mass of the particle and being able to classify since the momentum will be less with higher mass particles.



The graph below shows the number of hits on certain particles, this shows that each particle either had a likely chance of getting hit once or many times, so this allows us to parse and sort our data in a way that allows for us to take out massive outliers within the data set, and will be able to make a model that then complies with classification of certain particle points within our data.