SMU Data Science Review

Volume 6
Number 2 *Summer 2022*

Article 8

2022

Classification of Pixel Tracks to Improve Track Reconstruction from Proton-Proton Collisions

Kebur Fantahun
Southern Methodist University, kfantahun@smu.edu

Jobin Joseph
Southern Methodist University, jobinj@mail.smu.edu

Halle Purdom
Southern Methodist University, hpurdom@smu.edu

Nibhrat Lohia
Southern Methodist University, nlohia@mail.smu.edu

Classification of Pixel Tracks to Improve Track Reconstruction from Proton-Proton Collisions

Halle Purdom[1], Jobin Joseph[1], Kebur Fantahun[1], Nibhrat Lohia[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] Adjunct Lecturer,
Southern Methodist University, Dallas, TX 75275 USA
{hpurdom, jobinj, kfantahun}@smu.edu
nlohia@smu.edu

Abstract. In this paper, machine learning techniques are used to reconstruct particle collision pathways. CERN (*Conseil européen pour la recherche nucléaire*) uses a massive underground particle collider, called the Large Hadron Collider or LHC, to produce particle collisions at extremely high speeds. There are several layers of detectors in the collider that track the pathways of particles as they collide. The data produced from collisions contains an extraneous amount of background noise, i.e., decays from known particle collisions produce fake signal. Particularly, in the first layer of the detector, the pixel tracker, there is an overwhelming amount of background noise that hinders analysts from seeing true track reconstruction. This paper aims to find and optimize methods that are instrumental in figuring out how the true particle track can be decoupled from the background noise produced at the pixel tracker level of the detector. The results of this study include successful implementation of machine learning techniques to classify signal and background from particle collision data. From these results, it was concluded that neural networks are a successful resource for analyzing and processing particle collision data to reconstruct particle pathways.

1 Introduction

Particle collisions play a part in understanding what the universe is made of and how it works. Colliding bunches of particles at high speeds produces results that include but are not limited to finding new particles, exposing new physics through the decay process of known particles, and gaining a stronger understanding of the Standard Model. To analyze particle collision experiments, the particles' positions must be tracked as they leave a collision on their specific path. The CMS Silicon Pixel Detector at CERN contains many pixel, strip and other types of detectors that

record when particles interact with the detector called a “hit.” The various positions of the hits from a particle collision are tracked with enough precision for the particle track to be reconstructed.

Particle collisions lead scientists to better understand the world. In the early 1900s, physicists believed that the current world's physical understanding was complete. All physical theories related to heat, electricity and general mechanics were thought to be complete at the time. In the 1900s, a quote from Lord Kelvin states “There is nothing new to be discovered in physics now, all that remains is more and more precise measurement.”. It was believed that there was nothing left to discover but more precision to higher decimal points for physical constants such as the charge of the electron. As physicists began to push the fringe of the current theories of the time, they learned of electrons through experiments like J. J. Thomson’s cathode ray tube. Eventually, the picture of the fundamental particle spectrum began to form which paved the way for the study of atomic nuclei. Although there is a disparity in understanding the relationship between gravity and electromagnetism, physicists continue to strive to unify all physical understanding with one theory (Landua, 2006).

Physicists would like to explain the building blocks of matter in a compact, detailed yet elegant way. The state-of-the-art model that describes all the properties of what are thought to be the building blocks of the universe is called the Standard Model. The Standard Model illustrates which particles make up all known matter and how they work together. Protons and neutrons are made up of particles called quarks. Leptons are the family name of the electron. Protons, neutrons, and electrons come together to form all known matter. The mediators of each of the forces between all observable matter are called bosons and are known as the W boson, the Z boson, the photon, and the gluon. They mediate through the forces that physicists are interested in understanding, such as the Strong Nuclear Force (SNF), the Weak Nuclear Force (WNF), and the electromagnetic force (EMF) or “Light” itself which is an electromagnetic wave. The only fundamental force that is not explained by this model is gravity. EMF is mediated by the photon which plays a role in electric and magnetic fields. SNF is mediated by the gluon, these particles connect to “glue” the quarks together which form atomic nuclei. WNF is not as commonly observed as the other forces, this force details reactions such as nuclear decay, e.g., nuclear fusion like what happens inside of a star.

With all that is explained by the Standard Model, there are still fundamental aspects of the universe that are not fully understood. It is currently the best description of all physics, but it is not a complete model. For example, with the recent Higgs Boson discovery, physicists learned that several particles are given mass through the Higgs field interactions they go through. Although most masses are understood, scientists still do not know if neutrinos, which are leptons, are also given mass through the Higgs. The model also does not explain everything observed like dark matter or antimatter. Studying particle collisions can further expand the Standard Model through discovery. Particle collision experiments are where new physics can be observed and discovered, which is why these experiments are relevant and necessary (CERN, 2019).

General neural network model development has the capability to assist future physics studies at the LHC regarding pixel and particle track reconstruction as well as the demystification of background noise. As the previously mentioned models improve, the data produced at the LHC can be better used since more correct predictions of a particle's path will be made. The electricity cost to run the LHC alone is about 23.5 million dollars a year. Utilizing machine learning (ML) models that are trained on the few runs that LHC runs a year, scientists can hypothesize other particle collisions that might follow similar trajectories. This is extremely helpful as a substantial number of resources can be saved. ML models that were formed by training with real experiments could also lead to conclusions of what experiments are not worth exploring in the future.

The research in this study is a necessary step in processing the massive amounts of data produced by the CMS. The aim of this research is to introduce methods that require less computing power to process particle collision data. As machine learning techniques continue to increase in complexity and efficiency, there will be improvements to this process.

The data analyzed in this study was simulated by CERN based on 2018 CMS experiments. Analyzing simulated data in particle physics is extremely useful to build effective models from data sources not limited to collision experiment runs. For example, in the Higgs boson discovery, simulations were used to design detectors and analytical procedures aimed towards the identification of events with Higgs characteristics that were previously predicted in theory (Elvira, 2017). In a similar manner, simulated data in this work will be used to develop the analytical procedures for CMS collision experiments.

In this study, machine learning is employed to assist in path reconstruction for particles that are leaving a particle collision. Because there is a huge combinatorial background in the data collected, the goal is to parse out the noise in the data to isolate the signal. The event classification between signal and background can be modeled using different machine learning techniques like random forests, boosted decision trees, and neural networks. This research aims to use machine learning to assist in improving the signal to background ratio for pixel tracks.

2 Literature Review

2.1 Particle Collision Tracking with the Compact Muon Solenoid

CERN uses the Large Hadron Collider (LHC) to accelerate particles and run particle collision experiments. These collisions are tracked by several different detectors including the Compact Muon Solenoid (CMS) detector. The CMS detector is an all-silicon tracker with pixel and strip detectors (Allport, 2019). As Allport mentions, there are advances that are anticipated soon to help with better particle tracking. There is growth in the CMOS imaging sensor market, which can assist in

producing more precise tracking of future particle collision experiments (Allport, 2019).

As particles travel on a trajectory, the pixels and strips in the CMS record each place the particle passes through as a "hit," and the particle path can then be reconstructed from this data. The resolution of the detectors must be extremely high to be able to differentiate between different particles. Thousands of hits are then detected in the span of nanosecond intervals, outputting a large amount of data to be analyzed for the particle path reconstruction (Erdmann, 2010). The data output from the CMS is of sizes that could be compared to some of the largest industrial data sets today. To interpret this data, machine learning techniques like neural networks and boosted decision trees started the foundation of research into using machine learning in particle physics (Radovic et al., 2018).

The CMS itself has its own tracking algorithm that iterates through different steps to look at hits that have not been assigned to a reconstructed particle pathway. These steps include searching seeds in the inner layers, applying a Kalman filter based on pattern recognition, then fitting the trajectories with the Kalman filter. A compatibility test rejects any outliers and then the particle tracks are selected. The CMS tracking algorithm must continuously evolve to account for the higher luminosity environment of the detector (Sguazzoni, 2016).

2.2 Implications of the High Luminosity Phase of the Large Hadron Collider

Luminosity in terms of the LHC refers to the potential number of collisions per area over time in the collider. The goal behind increasing the luminosity of a collider is to see more collisions and collect more data. Often the particle collision events that scientists look for in the collider are exceedingly rare, so it can be challenging to observe something that rarely happens. By increasing the luminosity, the probability of detecting one of these rare events increases, because there are more collisions overall (CERN Accelerating Science, 2019).

CERN is currently working towards upgrading their LHC to increase its luminosity by a factor of 10, making it a High Luminosity Large Hadron Collider. The upgrade will introduce a drastic increase in the number of collisions happening at a given time, so the detectors will pick up many more "hits" of particles interacting with the detectors. More experiments can then be run; however, this will also introduce an increase in the background data that needs to be parsed through to find the signals. By upgrading the LHC, a large challenge posed for the data analysis will be the efficient reconstruction of particle pathways (Tüysüz, 2021).

The act of particle tracking is manageable now but will soon present a massive burden as the LHC upgrades to the High-Luminosity LHC. Bernius (2019) illustrates through this upgrade that the LHC is projected to produce data that is 10-fold the size of the previous 3 LHC-Runs combined, on the order of exabytes. To maximize particle collision data, the LHC spends a large amount of time and heavy resources for the reconstruction process, specifically with tracking. Reconstruction is the method by which researchers take raw detector data and transform it to build a

physically examinable picture of the original particle collision and its debris. Machine learning algorithms offer an advantage in using the architecture of parallelization to lighten the computational load from tracking and reconstruction.

Another issue that will arise with the LHC's luminosity increase is the radiation damage that will be inflicted on several layers of the tracker as a result of the continuous running of the detector. With extra luminosity, track reconstruction performance at the current level will suffer. To keep up with the growth of pileup, the pixel system must be upgraded to be 4 times as strong to retain a good granularity (Schmidt, 2016). Pileup occurs when there are several collisions happening at the same time, and the background collisions must be separated from the target signal. While this is more of a hardware challenge to be faced when luminosity of the LHC increases, it will also affect the data output from the CMS. Because of that, it is something to take into consideration in this study of track reconstruction, as it could change the algorithm or require an algorithm that functions very accurately.

2.3 Different Machine Learning Methods for Particle Track Reconstruction

Machine-learning plays a practical role in particle physics. Proton-proton collisions at the LHC give way to a complicated data system that requires sophisticated algorithms to organize. After the LHC produces event collision data, the next steps for the data include reconstructing all detected particle's pathways in an event and then looking into the physics that created the particles. This study will focus on the former task, which is reconstructing pathways of all particles in an event. This task is key to getting to the final goal of discovering new physics by analyzing the involved particles in a particular event. The CMS uses multivariate regression to train boosted decision trees (BDTs) to determine particle properties. Graph convolutional networks and RNNs have been employed with wide application in research. Specifically, CNNs and computer vision can be used for neutrino experiments, because of the challenge of finding the small neutrino in the large detector. RNNs are also especially useful for beauty quark identification, and they must be able to identify the jets that are radiated from the beauty quark. To widen the approach of machine learning, other models should be considered. Specifically, Generative Adversarial Networks (GANs) and Variational Auto-encoders (VACs) (Radovic et al. 2018).

The pursuit of refining pixel track reconstruction led many researchers to develop algorithms based on different forms of neural networks. In Andrews et al. (2019), Convolutional Neural Networks (CNNs) are used in an end-to-end format, allowing for event classification by separating the signal versus background in addition for attributes like angular distribution and shape of photon showers and energy scale of nearby hits. End-to-end classifiers are also referred to as image-based classifiers, and all these types of classifiers use Residual Net-type (ResNet-15) CNNs. The benefits of these types of CNNs are their scalability and their simplicity. Andrews et al. (2019) studies five specific models in the central and central-forward models.

CNNs are also used in Florio et al. (2018), for doublet seed filtering with success in dealing with the large combinatorial background. CNNs are also employed as a classifier in Florio's experiment using a novel technique. There are two layers, one consisting of a CNN block which is a normal stack of convolutional layers and another stack which is a dense block. The dense block operates differently as it is a stack of two connected layers which are fed a one-dimensional reduced image. Baranov et al. (2019) compares two different methods for particle track reconstruction, including a CNN-based approach and recurrent neural network (RNN) approach. The end-to-end CNN algorithm overcomes all the disadvantages of the deep RNN model, including the CNN model's ability to run completely end-to-end and its lack of sequential nature (Baranov et al. 2019).

Similarly, Tsaris et al. (2018) approaches the problem of application of deep learning to this problem by using CNNs and RNNs. Tsaris et al. (2018) explains that as the LHC moves into its high luminosity phase, the particle tracking algorithms currently used like the Kalman Filter scale poorly. As the LHC increases in luminosity, it therefore increases its data output because many more collisions are happening at once. These previous algorithms will not be sufficient for the new phase of the LHC. To solve this issue faced, Tsaris et al. (2018) turns to machine learning to provide algorithms that can reconstruct particle pathways. Machine learning techniques will scale better than these previous methods and can also deal with high dimensional data and nonlinear data. By exploring Long Short-Term Memory (LSTM) recurrent neural networks, it is shown that these can toy datasets simulating particles passing in two and three dimensions how the particle passes through the detector layers. The toy datasets are produced through Monte Carlo (MC) simulations meant to simulate LHC data. Another form of RNNs used was a shallow neural network with Grated Recurrent Unit (GRU) layers instead of LSTM. On simulated two-dimensional data, the accuracy performed the best with less tracks, and was also able to be tuned by changing the threshold. To explore how these neural networks scale, an A Common Tracking Software (ACTS) dataset was used rather than the toy MC data simulations originally used. By scaling, it was found that the LSTM performed with 70% accuracy and that accuracy improves with less pileup.

The other approach studies have taken to particle track reconstruction is Graph Neural Networks (GNNs). Where CNNs function in Euclidean space, GNNs exist outside of that so exploring how these models compare in event classification is worth investigating. In Duarte and Vlimant (2020), GNNs are used for track reconstruction because of their advantage to CNNs and RNNs of making full use of the relational structure of the data. Tüysüz et al. (2021) uses GNNs as well but converts a novel GNN into a hybrid quantum-classical GNN and incorporates a circuit-based model for track reconstruction. Since the status of quantum hardware is not able to handle the high pile- up conditions of the TrackML dataset, there could only be simulations that are being employed to train the models Tüysüz et al. (2021). The fact that particle tracking datasets are tough to work with could materialize other novel ways of loading data into GNNs. In Tüysüz et al. (2021), experiments were limited by exceptionally large RAM requirements and the significant increase of training times to more than a week for models.

Data produced at the LHC provides researchers at CERN the opportunity to exploit novel machine learning methods to better understand the aftermath of particle

collisions. The problem with attempting to use such new models is that they were initially made to deal with sequences or images. Models such as convolutional neural networks (CNNs) are well suited to dealing with image processing while recurrent neural networks (RNNs) work best with sequential data. Duarte and Vlimant (2020) suggest that it is common to transform particle physics data into images, sequences or even graphs so that models like CNNs, RNNs and graph neural networks can be used. Physical data and processes that are measured in terms of time and space have a structure that does not map easily to images or sequences so researchers must look to models outside CNNs and RNNs to ideally reconstruct particle tracks. Duarte and Vlimant (2020) offer a solution to this problem in that GNNs, which deal with graphical data well, inherently synchronize with the relational structure of particle collision measurement data.

In contrast, simple experiments about modeling tracking of motions can be done using just mathematical algorithms that show direct relationships. This can be clearly seen when looking at how Vilela (2020) conducted their experiment showing the droplet - particle collisions based on a numerical study. This allowed for simple simulations to later be produced easily on computers using numerical methods. The benefit for the reader of the articles might be easier to show what factors influence the outcome for particle collision modeling using the numerical method instead of using a neural network. Statistical learning was used at the end to see how the methods that were used would function for other problems that would be similar.

Finally, another method that could be used to look at the particle collision tracks would be a class of algorithms called Monte Carlo simulations. In Madlener et al. (2018), where the performance of the CMS detector's proton-proton collisions is studied, the researchers were able to model the particle tracks using a Monte Carlo simulation, this was proven to be comparable with the results of the earlier published study in 2010 (CMS Collaboration, 2013). This class of algorithms provides an avenue to explore with this data set that has a great amount of noise (Madlener et al., 2018).

Based on the structure of the data, Dense Neural Networks are predicted to be the most effective at track reconstruction.

3 Methods

3.1 Data

The dataset in this study is from CERN Open Data called "Sample with tracker hit information for tracking algorithm ML studies TTbar_13TeV_PU50_PixelSeeds" (Di Florio et al., 2019). The data consists of pixel doublet seeds to be used in the study of particle tracking algorithms. The attributes of the data include location details of detected particles moving throughout the detector and details about the specific event or collision for each row.

This dataset was derived from the parent dataset "TTToHadronic_TuneCP5_13TeV-powheg-pythia8 in FEVTDEBUGHLT format for 2018 collision data" (CMS Collaboration, 2019). This data was simulated based on 2018 collision data from the CMS using Monte Carlo methods. Pile-up events (collisions) were added to the data in this step of processing to effectively simulate what data would look like from a CMS experiment.

3.1.1 Data Production

Monte Carlo methods use randomness to sample from a known probability distribution. To simulate data, the CMS uses a Monte Carlo based production technique. There are two steps in producing the simulated data, including the generation and simulation stages. These stages are visualized in Figure 1, showing the processing steps from simulation of the event, detection, to reconstruction and analysis.

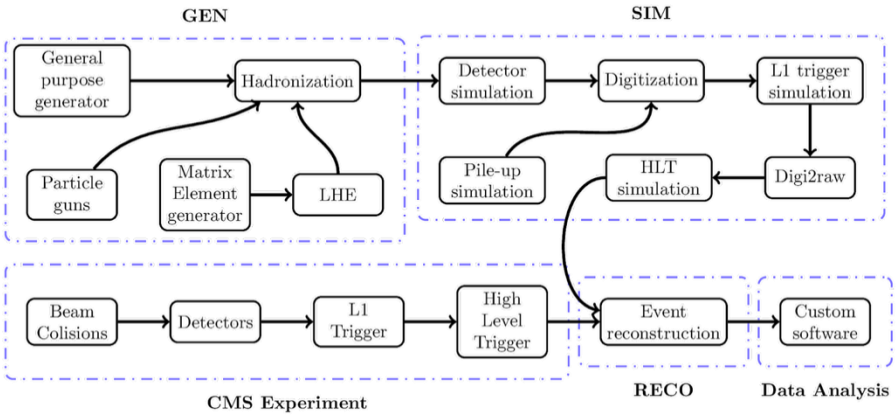

Figure 1. CMS Monte Carlo production overview. An overview of the steps taken to make CMS data ready for analysis. Generation and simulation steps create data based on Monte Carlo production methods. These methods produced the data provided in this study, based on CMS 2018 collected data. (CERN Open Data Portal).

In the generation step, an event generator is used to simulate beam collisions. In this case, an event refers to the collision of two protons in the CMS detector. In stage two, the detectors are simulated, including simulation of the L1 triggers, HLT, and pile-up. Triggers and the HLT determine which particle collisions are kept in the data and which are discarded. The L1 triggers are the first low-level step to this, then the HLT (higher level trigger) is the next step. Pile-up simulation produces the background noise that is seen in the CMS from different event occurrences at the same time or beam overlay. Particularly in high luminosity environments like the CMS, there is a lot of background picked up existing among the target signal.

All these simulations create an output like the data output from the actual CMS experiment. Regardless of whether data came from simulation or a CMS experiment, the data then enters the RECO step where event collisions are reconstructed for analysis. In this study, the data is at this point in processing after it has gone through generation, simulation, and RECO stages (CERN Open Data Portal).

3.2 Data Exploration

To start looking at the data and visualizing the particle pathways and collisions, Figures 2 through 5 use the dataset to explore how the data tracks the location of particle hits. Next steps into analyzing the data further than exploratory methods and visualizations will include beginning to develop machine learning models to classify true versus fake signals.

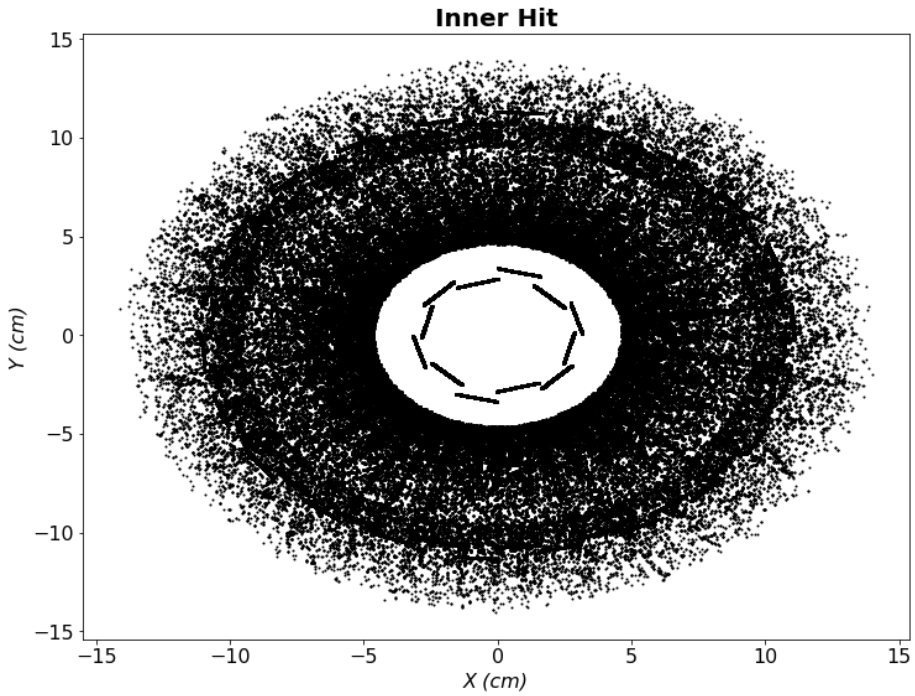

Figure 2. Visualization of a particle hit on the inner layer of the tracker.

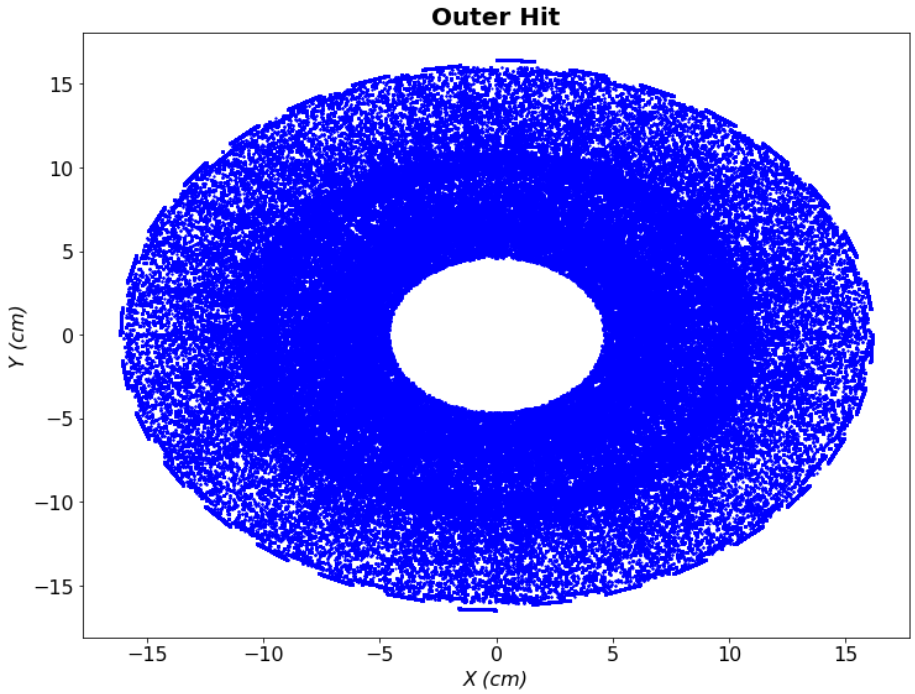

Figure 3. Visualization of a particle hit on the outer layer of the tracker.

Figures 2 and 3 above provide a visual for particle hits that occur in the CMS part of the LHC. Using the x and y coordinates it was possible to map out particle hits that were able to be picked up from the silicon sensors that line the collider. It is important to note that even in a simulated data set there are many hits that are recorded which do not give much information about the trajectory of the collision.

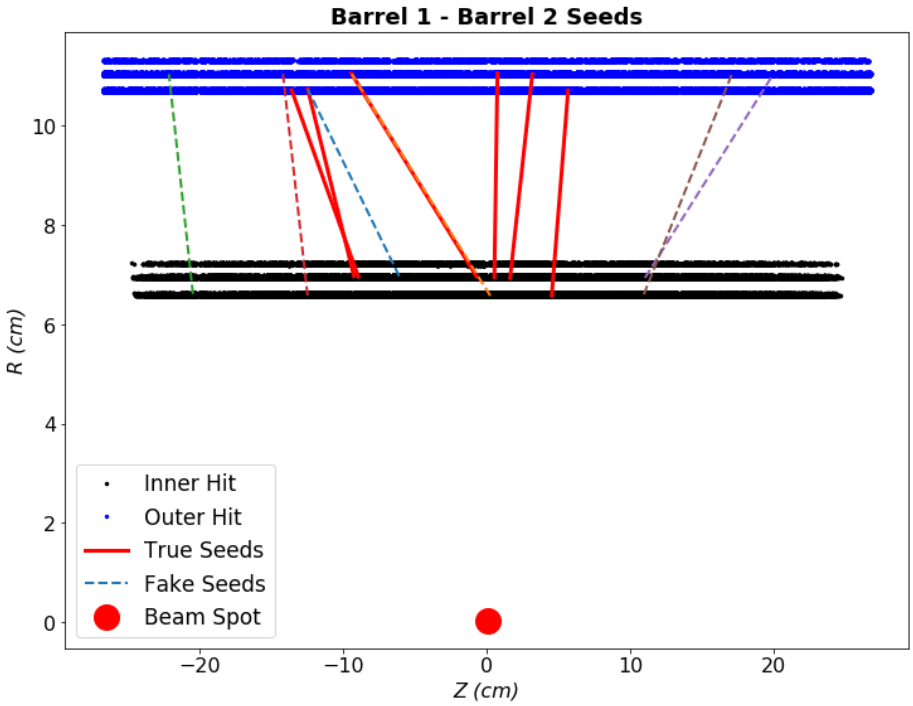

Figure 4. Visualization of the true and fake particle pathways between the inner and outer tracking layers of the detector.

Figure 4 above shows examples of true and fake particle pathways detected in the tracking layers. The black dots represent the inner tracking layer, and the blue dots represent the outer tracking layer. The beam that the event collisions are generated from is shown with the large red dot at Z = 0 cm. To show the difference between mapping background collisions versus the target collision, the red solid lines represent true particle pathways, and the dotted lines represent fake pathways. This helps visualize the goal of this study, which is sorting the target collisions from background noise. That background noise is from other collisions from pile-up sources, and scales with luminosity in the detector.

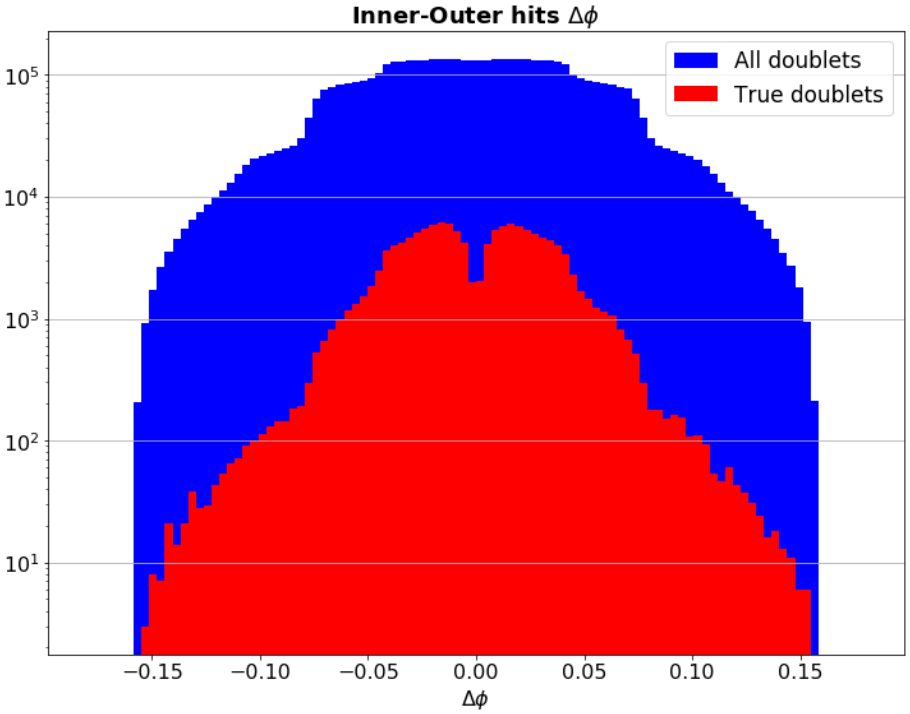

Figure 5. Change in angle between only the true signal and all the signals including background.

Figure 5 above shows the change of doublet inner hit azimuthal angle. The doublet is categorized as when there is an unpaired electron. Phi describes the angle of the spectral lines. This can provide some insight about how to factor the change of phi into particle collision tracking for future datasets. This graph also helps visualize the proportion of true doublets to all the doublets.

4 Results

This paper intended to find the best algorithm for the path reconstruction of particles from the data. This is an optimization problem for a classification task. To find the best algorithm for the data, different machine learning methods were compared to find the most accurate algorithm for the data used in this research and then optimize parameters of the algorithms.

The models implemented a 58-feature subset of the data. The features of the entire dataset are divided into 566 features comprised of two sparse matrices of in and out pixels as well as 58 features that describe numerical attributes of the event. The sparse

matrices produced the images seen in Figure 6 below. Examples of these 58 features used in the models include different coordinates of the row for the in and out pixel, attributes of the event, and details about the pixel clusters. Using this 58-feature subset of the data allowed for reduced computing time and eliminated any issues stemming from sparse matrices.

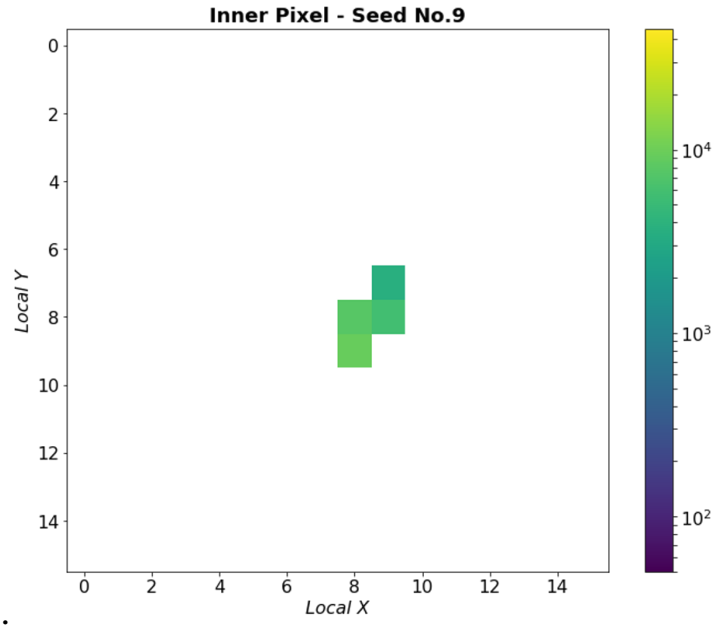

Figure 6a. Visualization of a sample event from the dataset. Showing the location and intensity of the sample event on the inner pixels.

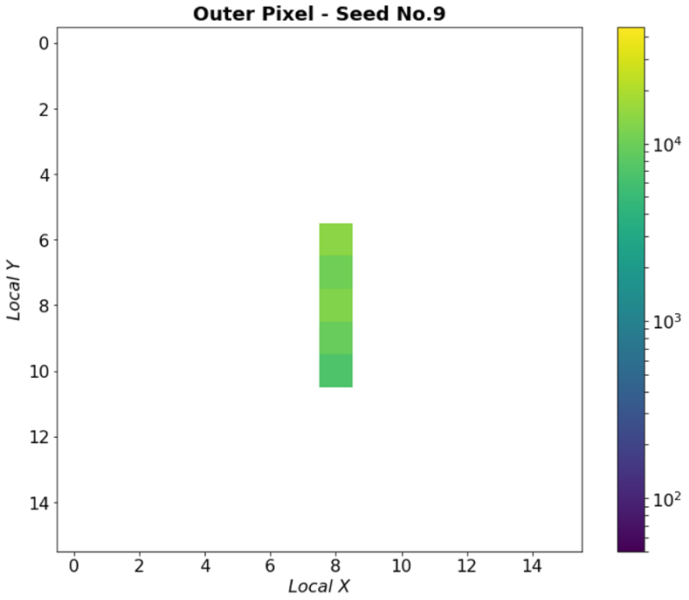

Figure 6b. Visualization of a sample event from the dataset. Showing the location and intensity of the sample event on the outer pixels.

The first model tested was a boosted decision tree as a baseline for the neural network models. In this study, the AdaBoost classifier from the python library scikit-learn, which uses the AdaBoost-SAMME algorithm, was used to classify several features that make up particle tracks into either a true or fake class. AdaBoost is a boosting method commonly used with a random forest model. AdaBoost uses a tree with just two nodes (stump), making them weak learners. In AdaBoost, the forest has trees that influence each other and learn from the previous mistakes. Finally, based on each of their accuracies, each weak learner is assigned a greater weight in the classification than others. Higher weights are assigned to the most accurate weak learners until AdaBoost combines all the accurate weak learners into a single strong classifier. This is important as this collation of learners can help locate real particles compared to particles that are not in scope.

The accuracy of the AdaBoost model was 82.2%, and the output of the model is visualized below in Figure 7. The boosted decision tree classifies an event as either the “true” signal, or the “fake” background noise.

An example of the AdaBoost decision tree is seen below in Figure 8. Each attribute of the particle track, e.g. “inX,” “inY” and “inPixelZero,” has a corresponding y-value that classifies the track as “True” (Y=+1) or “Fake” (Y=-1). The classifier employed is trained on the y-values +1 and –1 to build a model that will be able to classify new particle tracks.

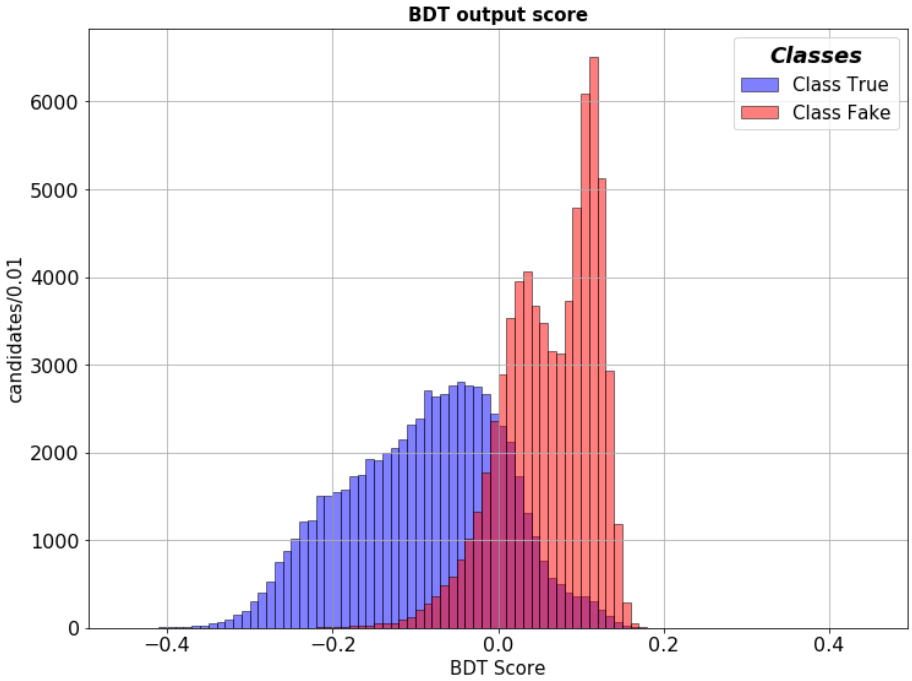

Figure 7. Output of the boosted decision tree model as a histogram visualizing the two classes. The true signal is shown in blue and the fake background is shown in red.

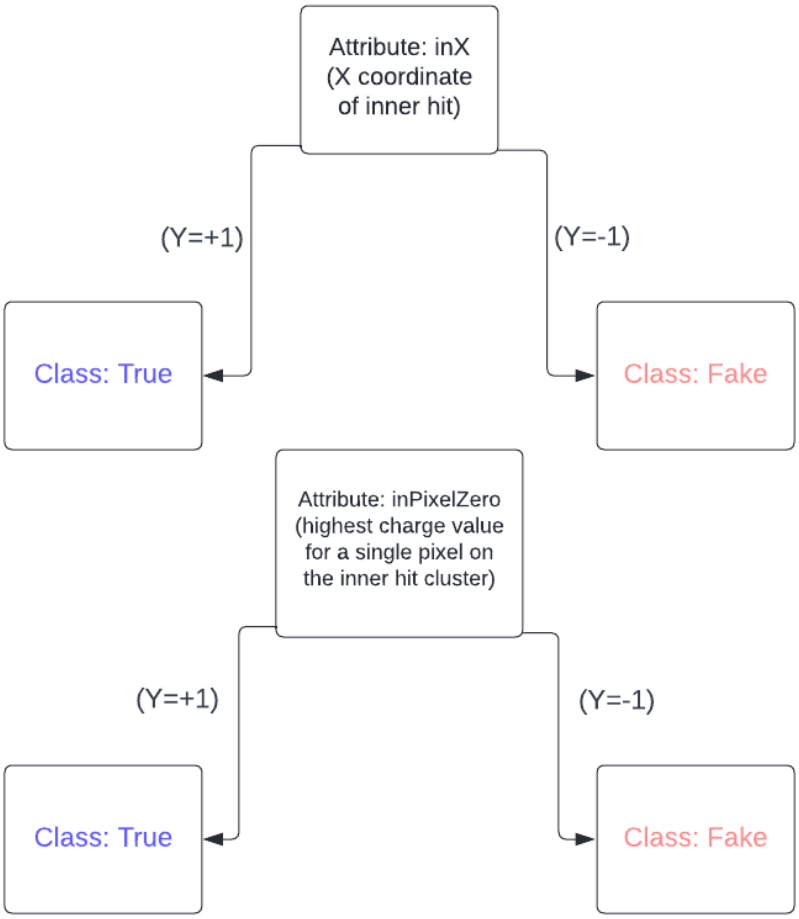

Figure 8. Two decision tree stumps where the AdaBoost classifier learns by looking at each attributes y-values leads to a true or fake classification. As in Figure 7, blue corresponds to a true track classification and red corresponds to a fake track classification.

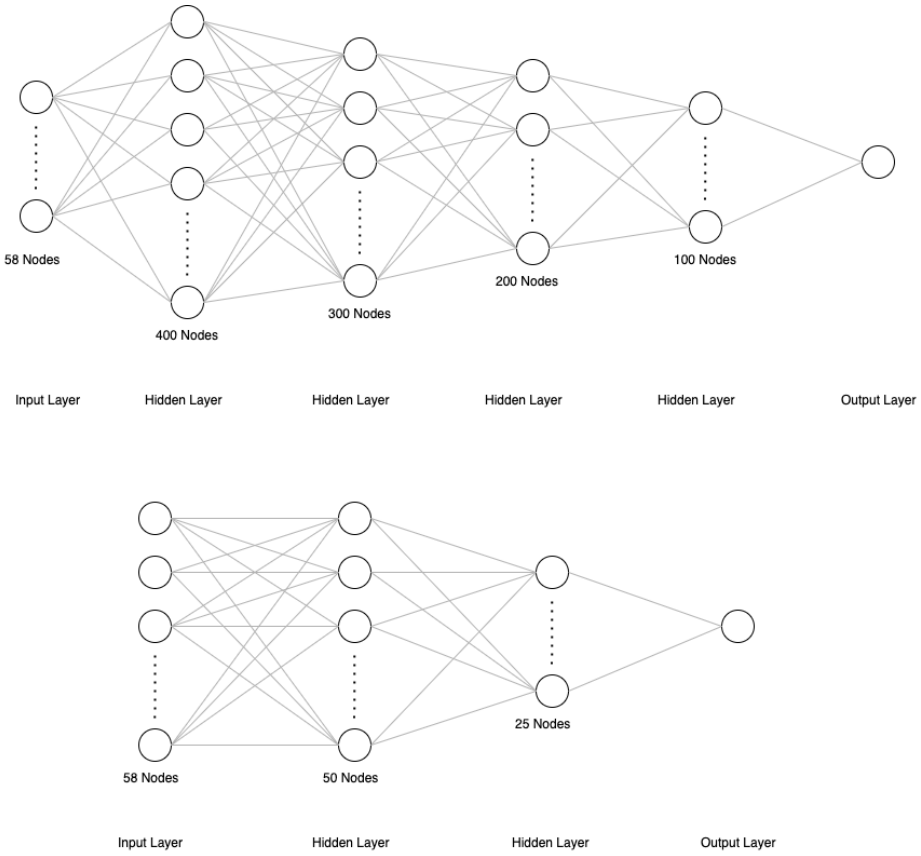

Figure 9. Two versions of the employed Dense Neural Networks. Above is the 4 layered DNN and below is the 2 layered DNN. Each circle in the diagrams represents a node. Because there were sometimes hundreds of nodes in each layer, the ones not pictured above are represented by the dotted lines. In the 4 layered DNN, the first hidden layer has 400 nodes, the second hidden layer 300 nodes, the third hidden layer 200 nodes and the fourth hidden layer 100 nodes. In the 2 layered DNN, the first hidden layer has 50 nodes, and the second hidden layer has 25 nodes.

The second model tested was an Extra Tree Classifier. This model boasts accuracy improvements and control of over-fitting by using randomized decision trees and averaging over sub-samples of the data.

The final models tested were dense neural networks (DNN). The networks consisted of an input layer, dense layers, and the final prediction output layer. 2 DNN's were considered, a 2 layered DNN and a 4 layered DNN. Batch normalization was also implemented after the input layer as well as preceding the final prediction output layer. The activation function for each dense layer was the ReLu (Rectified Linear Unit) function, and the final prediction layer was the sigmoid function. The

Adam optimizer was used and a binary cross entropy loss function. All these parameters were tuned to get the best accuracy for the data. Early stopping for the model was set to monitor the validation set loss, and the final model accuracies can be seen in the table below.

Model	Accuracy	Precision	Recall
Extra Trees Classifier	84.30%	83.95%	84.67%
Boosted Decision Tree	82.20%	80.98%	84.00%
Dense Neural Network (2 layers)	80.09%	79.73%	80.66%
Dense Neural Network (4 layers)	89.05%	91.21%	86.4%

As seen in the table above, the 4 layered dense neural network performed best in accuracy. The extra tree classifier was slightly below the 4-layer DNN, then the boosted decision tree and then the 2-layer DNN.

5 Discussion

5.1 Interpretations and Implications

The results in applying machine learning techniques to the classification task of particle track reconstruction were successful. With the 4-layer neural network, an accurate model was created to take on this classification task. Because particle collision experiments produce such a massive amount of data output, the analyzing of the huge volume of data is vital to drawing out meaningful results from these experiments.

As different machine learning methods are developed, these methods should be tested and utilized in particle collision data analysis. As referenced previously, the LHC at CERN will continue to increase in luminosity over time, meaning a huge increase in data output from particle collision experiments. To continue learning from these experiments, machine learning methods that can process data more efficiently in terms of time and resources are necessary.

The 4-layer dense neural network yielded the highest accuracy of the four models. When looking at the training times of machine learning models, it should be noted that decision trees are often going to be trained much faster than a neural network and be much more interpretable, which is worth noting as the output from the LHC continues to grow in data size. While the decision tree methods tested in this study did not yield the highest accuracy, they still prove useful because of their computing efficiency.

5.2 Ethics

The classification of the pixel tracks to help improve the track reconstruction from proton-proton collisions is part of leveraging data science modeling with particle physics. It is important to consider how ethics can play a role in research and how the research itself is conducted.

There are some basic principles that are important for ethical research. One of these basic principles exemplified in this research is the minimization of the risk of harm. The risk of harm was reduced by utilizing an open-source dataset from CERN. Another ethical research guideline considered was obtaining informed consent. This data set was publicly available through CERN's website, with the intent for scientists to use this data set for research. Another guideline considered was anonymity and confidentiality. As the state did not include people in the research, this was not applicable. Also, other ethical guidelines such as giving the right to withdraw and avoiding deceptive practices are not applicable to the study as it was particles that were part of the data set not humans.

The research strategy implemented used the practice data sets produced from Monte Carlo simulations that were openly sourced from the CERN Open Data Portal. Although this data was created from CMS 2018 simulated data, it was similar enough that it allowed us to do proper research without having complex data mistakes that might have occurred during the actual runs of the LHC. Analysis on simulated particle data allows future researchers to save time because there will be experience on the use of models on similar data. The CERN Open Data Portal is invaluable in that it allows scientists and researchers that are not directly involved to be helpful and push the science of physics.

5.3 Future Research

There is more research that needs to be done for particle physics tracking to be better understood. CERN has recently started running particle collision experiments again after a long stop. Newly upgraded equipment will push the LHC to new energies and allow for the possibility of discovery. All of this is important as the data that comes from the experiments can be analyzed once put in open sources, in research like this paper. Secondly, different modeling techniques can be used to help classify the particle tracks. As the field of data science continues to expand, there is more computing power that is available and ever-improving models can bring new perspectives to particle physics analysis. Another avenue that can be explored to help with the modeling could be the addition of more pre-processing of the data with subject matter experts that have knowledge about particle collisions extensively. This is an exciting area of physics and should be continuously researched as advances in computing and upgrades on the LHC are changing the results and the understanding of physics in real-time.

6 Conclusion

This research investigates how machine learning can be used to reconstruct particle pathways from particle collision experiments at CERN. In classifying the true signal versus background noise, the particle pathways from these experiments can be reconstructed so that scientists can then draw insights and discover new phenomena about the way the world functions. In this research, the 4-layer dense neural network was found to be the most successful classifier in terms of accuracy to process the data after experimentation. Machine learning methods will be increasingly useful as the data output from the LHC continues to grow and more efficient data processing methods are needed to reconstruct the particle pathways of collision experiments.

Acknowledgments. We would like to thank Nibhrat Lohia, Dr. Jacquelyn Cheun, and all our SMU (Southern Methodist University) Data Science professors.

References

1. Allport. (2019). Applications of silicon strip and pixel-based particle tracking detectors. *Nature Reviews Physics*, *1*(9), 567–576. https://doi.org/10.1038/s42254-019-0081-z
2. Andrews, Paulini, M., Gleyzer, S., & Poczos, B. (2019). Exploring End-to-end Deep Learning Applications for Event Classification at CMS. *EPJ Web of Conferences*, *214*, 6031–. https://doi.org/10.1051/epjconf/201921406031
3. Baranov, Mitsyn, S., Goncharov, P., & Ososkov, G. (2019). The Particle Track Reconstruction based on deep Neural networks. *EPJ Web of Conferences*, *214*, 6018–. https://doi.org/10.1051/epjconf/201921406018
4. Bernius, C. (2019). HL-LHC prospects from ATLAS and CMS. *Journal of Physics: Conference Series, 1271*, 012004. https://doi.org/10.1088/1742-6596/1271/1/012004
5. *CERN accelerating science*. (2019) CERN. (n.d.). Retrieved 2022, from https://home.cern/resources/faqs/high-luminosity-lhc#:~:text=Luminosity%2C%20which%20is%20the%20measure,to%20100%20million%20million%20collisions.
6. CERN. (2019, April 8). *The Standard Model | CERN*. Home.cern. https://home.cern/science/physics/standard-model
7. Chiochia, Swartz, M., Fehling, D., Giurgiu, G., & Maksimovic, P. (2008). A novel technique for the reconstruction and simulation of hits in the CMS pixel detector. *2008 IEEE Nuclear Science Symposium Conference Record*, 1909–1912. https://doi.org/10.1109/NSSMIC.2008.4774762
8. Chowdhury, Ray, M., Passalacqua, A., Mehrani, P., & Sowinski, A. (2021). Electrostatic charging due to individual particle-particle collisions. *Powder Technology*, *381*, 352–365. https://doi.org/10.1016/j.powtec.2020.12.012
9. CMS collaboration. (2013). The performance of the CMS muon detector in proton-proton collisions at√ s= 7 TeV at the LHC. *Journal of Instrumentation*, *8*(11), P11002.

10. CMS Collaboration. (2019). *TTToHadronic_TuneCP5_13TeV-powheg-pythia8 in FEVTDEBUGHLT format for 2018 collision data*. CERN Open Data Portal. Retrieved from https://opendata.cern.ch/record/12303
11. *CMS Monte Carlo Production Overview*. CERN Open Data Portal. (n.d.). Retrieved from https://opendata.cern.ch/docs/cms-mc-production-overview
12. Daniel Elvira. (2017). Impact of detector simulation in particle physics collider experiments. Physics Reports, 695(C), 1–54. https://doi.org/10.1016/j.physrep.2017.06.002
13. Di Florio, A., Pantaleo, F., & Pierini, M. (2019). *Sample with tracker hit information for Tracking Algorithm ML studies ttbar_13tev_pu50_pixelseeds*. CERN Open Data Portal. Retrieved from https://opendata.cern.ch/record/12320
14. Duarte, & Vlimant, J.-R. (2020). *Graph Neural Networks for Particle Tracking and Reconstruction*.
15. ERDMANN. (2010). THE CMS PIXEL DETECTOR. *International Journal of Modern Physics. A, Particles and Fields, Gravitation, Cosmology*, *25*(7), 1315–1337. https://doi.org/10.1142/S0217751X10049098
16. Florio, Pantaleo, F., & Carta, A. (2018). Convolutional Neural Network for Track Seed Filtering at the CMS High-Level Trigger. *Journal of Physics. Conference Series*, *1085*(4), 42040–. https://doi.org/10.1088/1742-6596/1085/4/042040
17. Kar, Choudhury, S., Zhang, X., & Zhou, D. (2020). *Examining the event-shape dependent modifications to charged-particle transverse momentum spectra and elliptic flow in p-Pb collisions at energies available at the CERN Large Hadron Collider*. https://doi.org/10.1103/PhysRevC.102.044901
18. Landua, R. (2006, July 2). HISTORY OF PARTICLE PHYSICS Rolf Landua CERN Lecture. https://indico.cern.ch/event/3324/sessions/152190/attachments/648059/891313/Rolf_Lectures_1_and_2.pdf. Retrieved April 10, 2022, from https://indico.cern.ch/event/3324/sessions/152190/attachments/648059/891313/Rolf_Lectures_1_and_2.pdf
19. Madlener, Van Remortel, N., Tavernier, S., Marinov, A., Bruno, G., Krintiras, G., Chinellato, J., Ahuja, S., Zhang, F., Cabrera, A., Godinovic, N., Ather, M. W., Dewanjee, R. K., Siikonen, H., Grenier, G., Schulz, J., Borras, K., Kleinwort, C., Zenaiev, O., … Heindl, S. M. (2018). Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at root s=13 TeV. *Journal of Instrumentation*, *13*. https://doi.org/10.1088/1748-0221/13/06/P06015
20. Radovic, Williams, M., Rousseau, D., Kagan, M., Bonacorsi, D., Himmel, A., Aurisano, A., Terao, K., & Wongjirad, T. (2018). Machine learning at the energy and intensity frontiers of particle physics. *Nature (London)*, *560*(7716), 41–48. https://doi.org/10.1038/s41586-018-0361-2
21. Schmidt. (2016). The High-Luminosity upgrade of the LHC: Physics and Technology Challenges for the Accelerator and the Experiments. *Journal of Physics. Conference Series*, *706*(2), 22002–. https://doi.org/10.1088/1742-6596/706/2/022002
22. Sguazzoni. (2016). Track reconstruction in CMS high luminosity environment. *Nuclear and Particle Physics Proceedings*, *273-275*, 2497–2499. https://doi.org/10.1016/j.nuclphysbps.2015.09.437
23. *The standard model*. CERN. (n.d.). Retrieved from https://home.cern/science/physics/standard-model
24. Tsaris, Anderson, D., Bendavid, J., Calafiura, P., Cerati, G., Esseiva, J., Farrell, S., Gray, L., Kapoor, K., Kowalkowski, J., Mudigonda, M., Prabhat, Spentzouris, P., Spiropoulou, M., Vlimant, J.-R., Zheng, S., & Zurawski, D. (2018). The HEP.TrkX Project: Deep Learning for Particle Tracking. *Journal of Physics. Conference Series*, *1085*(4), 42023–. https://doi.org/10.1088/1742-6596/1085/4/042023

25. Tüysüz, Rieger, C., Novotny, K., Demirköz, B., Dobos, D., Potamianos, K., Vallecorsa, S., Vlimant, J.-R., & Forster, R. (2021). Hybrid quantum classical graph neural networks for particle track reconstruction. *Quantum Machine Intelligence*, *3*(2). https://doi.org/10.1007/s42484-021-00055-9
26. Vilela, & de Souza, F. J. (2020). A Numerical Study on Droplet-Particle Collision: Lamella Characterization. *Flow, Turbulence and Combustion*, *105*(4), 965–987. https://doi.org/10.1007/s10494-020-00153-x