How to Use CMS Open Data for Beginners

A Beginner Friendly Guide to Exploring Real Particle Physics Data

Internship Project under MacroEdtech

Supervised by: Sagar Sakalley

Project Created by: Atufa Vora, Educational Content Writer (Physics)

May 16, 2026

Abstract

This report gives an elaborate beginner-friendly guide to the CMS Open Data Portal, which has been created by CERN for educational and research purposes. The purpose of this study is to comprehend how students or novice learners will be able to analyze and explore the actual particle collisions' data recorded by CMS experiments conducted at the Large Hadron Collider. In this report, I will describe the structure of CMS Open Data Portal, tools, and datasets that have been made available through it, and the process that was followed to create an actual Higgs boson candidate event display.

In addition to event visualization, I have also explained how histograms as well as simpler CSV datasets can be used to study key physical parameters like invariant mass and particle momentum. Several screenshots have been presented in order to explain the process clearly. Practical learning experiences, challenges, and significance of CMS Open Data for education have also been discussed.

The current research project highlights the potential of open scientific data to make complicated principles related to particle physics comprehensible through practical applications. Through hands-on exposure to actual experimental data, learners will be able to better understand how scientists explore and investigate particles like the Higgs Boson.

Contents

1 Introduction

The European Organization for Nuclear Research (CERN) is one of the biggest centers for conducting experiments in particle physics. The organization possesses the Large Hadron Collider (LHC), in which protons are accelerated and subsequently collided to recreate conditions similar to the moments right after the Big Bang event.

Amongst many other experiments conducted with the help of LHC, CMS, which stands for Compact Muon Solenoid, is worth mentioning. It is a multi-purpose detector that is responsible for recording results of particle collisions.

In 2012, CMS became famous due to being involved in the discovery of the elusive Higgs boson. The CMS Open Data Portal contains various datasets, software tools and materials for studying physics. Thanks to it, one can get access to real-world experimental data.

It goes without saying that open data plays a crucial role in modern science since it increases transparency of the work, facilitates reproduction of experiments and allows more people from all over the globe to participate in discovering new phenomena.

2 Understanding the CMS Open Data Platform

The CMS Open Data Portal is available at:

https://opendata.cern.ch/docs/cms-guide-for-education

The platform provides educational resources for three levels of learners:

1. Beginner: Visualize collisions.
2. Intermediate: Create histograms.
3. Advanced: Perform detailed analyses.

The interface includes a search bar, documentation pages, dataset records, and interactive tools. Each dataset page contains descriptions, metadata, download links, and example analyses.

2.1 Available Tools

- CMS Event Display
- CMS Histogram Visualizer
- Simplified CSV datasets
- Jupyter notebook tutorials
- Spreadsheet exercises

2.2 Available Datasets

Examples include Higgs candidate events, diphoton events, four lepton events, and dimuon events.

3 Step by Step Procedure

The following is an elaborate description of the steps taken to interact with the CMS Open Data Portal and create a visualization of real collisions of particles. The procedure is described in simple terms so that any learner can follow the same steps to achieve the intended purpose.

3.1 Step 1: Open the CMS Guide for Education

The first thing to do was to visit the CMS Guide for Education website. This website contains all materials required to work with CMS Open Data.

The guide was accessed using the following link:

https://opendata.cern.ch/docs/cms-guide-for-education

Once inside the webpage, one should check its content. It consists of three parts according to the complexity of the task:

1. Beginner: Visualize collisions
2. Intermediate: Make histograms with collision data
3. Advanced: Dive deeper into the data

Since this project is intended for beginners, the "Visualize collisions" section was selected.

3.2 Step 2: Launch the CMS Event Display

The beginner part contains a link to the CMS Event Display. CMS Event Display is a software that enables one to see a real collision of protons captured by the CMS detector.

Once one clicks on the link, it takes a few seconds for the software to load. Then one will be shown a 3D model of the CMS detector.

The event display includes:

- A central visualization window.
- A left panel containing display options.

- A toolbar with zoom and navigation controls.
- An "Open file" button used to load event files.

3.3 Step 3: Open Event Files from the Web

For loading data into the event display, the following procedures were done:

1. Clicked the **Open file** button.
2. Selected **Open file(s) from web**.
3. Waited for the list of available event collections to appear.
4. Selected the folder named `HiggsCandidates/`.

This directory has collision events that played a significant role in the detection of the Higgs boson.

3.4 Step 4: Choose a Higgs Candidate File

Inside the `HiggsCandidates/` folder, two important files are available:

- `4lepton.ig`
- `diphoton.ig`

The `4lepton.ig` file contains events where the Higgs boson decays into four leptons, typically electrons and muons.

The `diphoton.ig` file contains events where the Higgs boson decays into two photons.

Both of these files were loaded for the comparison of different Higgs boson decay modes.

3.5 Step 5: Load a Specific Event

After selecting a file, a list of available events appeared on the right side of the screen.

The following procedure was followed for visualization of the event:

1. Selected one event from the list.
2. Clicked the **Load** button.
3. Waited for the detector display to update.

Thus, the event was displayed by the colorful trajectories and energy depositions within the CMS detector.

3.6 Step 6: Display Relevant Particle Types

The left sidebar gives a menu for displaying various types of reconstructed objects.

For `4lepton.ig`:

1. Expanded the **Physics** section.
2. Checked **Electrons**.
3. Checked **Muons**.
4. Expanded the **Tracking** section.
5. Unchecked **Tracks (reco)** to reduce visual clutter.

For `diphoton.ig`:

1. Expanded the **Physics** section.
2. Checked **Photons**.

With these settings, the interesting decay products became easily identifiable.

3.7 Step 7: Explore the Event in Three Dimensions

The event display allows full three dimensional interaction.

The following actions were performed:

- Rotated the detector by dragging the mouse.
- Zoomed in and out using the mouse wheel.
- Changed the viewing angle.
- Inspected the paths of particles from the collision point outward.

It is possible to interact with the event display in three dimensions.

3.8 Step 8: Observe the Higgs Boson Signatures

The loaded events were analyzed visually.

In the four lepton events:

- Four high energy leptons were observed.
- Their tracks originated from a common point.

- The event topology was consistent with a Higgs boson decay.

In the diphoton events:

- Two energetic photons were visible.
- The photons appeared as electromagnetic energy deposits.
- The event matched the expected signature for Higgs to two photon decay.

These observations helped connect theory with real detector data.

3.9 Step 9: Access the Histogram Visualizer

The CMS Open Data guide also provides access to the histogram visualizer.

The following steps were performed:

1. Opened the histogram visualizer link.
2. Selected a sample dataset.
3. Chose a variable such as invariant mass.
4. Generated the histogram.

The resulting plot displayed the frequency distribution of the selected variable.

3.10 Step 10: Interpret the Histogram

The histogram was examined for identifying key features.

A peak at around 125 GeV suggests the existence of Higgs boson candidate events.

Peaks at other regions could be associated with known particles such as:

- J/ψ
- Υ
- Z boson

This was an illustration of how physicists identify new particles by analyzing statistics.

3.11 Step 11: Download Simplified CSV Datasets

The portal offers easy-to-understand CSV files that facilitate the learning process.

Below is the procedure that was followed:

1. Opened a dataset record.
2. Located the download section.
3. Downloaded the CSV file.
4. Saved it to the local computer.

The above files could have been analyzed using:

- Python
- Microsoft Excel
- Google Sheets

3.12 Step 12: Explore Jupyter Notebook Tutorials

The educational package has Jupyter notebooks that teach data analysis through actual CMS data.

The notebooks include details about how to use:

- Loading datasets.
- Cleaning data.
- Plotting histograms.
- Performing statistical analysis.

The notebooks can either be run online or offline.

3.13 Step 13: Capture Screenshots and Plots

Screenshots were taken during the activity from:

- CMS Open Data Educational Homepage.

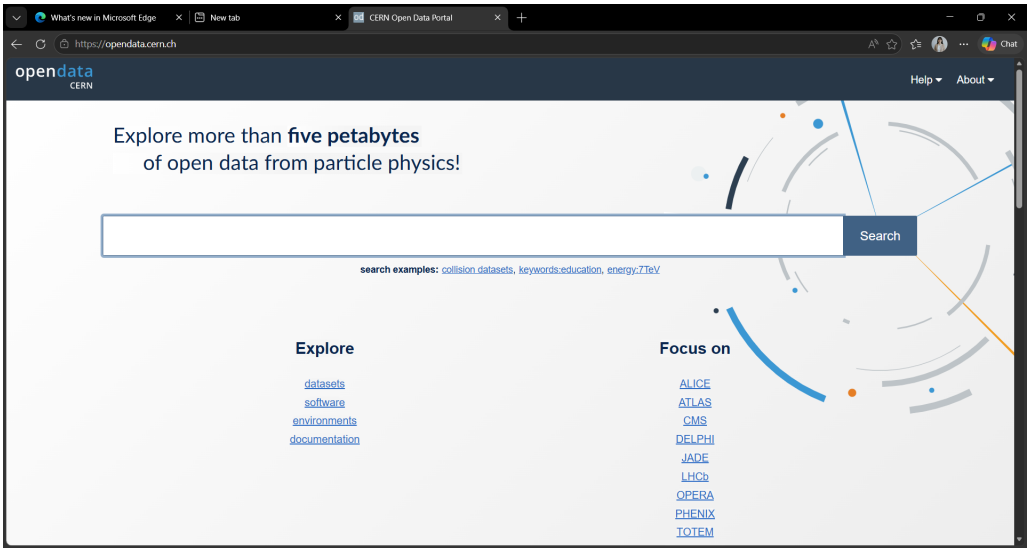

Figure 1: CMS Guide for Education page on CERN Open Data Portal. The page offers user-friendly tutorials and educational materials to help visualize events in collisions, create histograms, and analyze real data from the CMS experiment.

- The event display interface and Higgs candidate events.

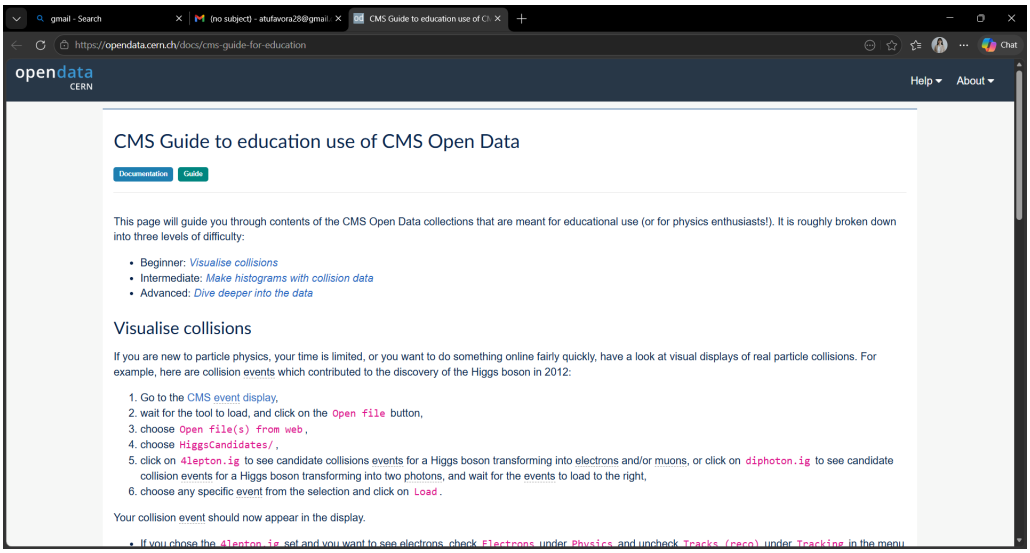

Figure 2: The Screenshot of the CMS Guide for Education page (https://opendata.cern.ch/docs/cms-guide-for-education) from the CERN Open Data Portal. This guide served as the main source of information while using CMS Event Display and choosing `HiggsCandidates/` dataset as well as Higgs boson candidates' events, including `4lepton.ig` and `diphoton.ig`.

- Higgs vs Non-Higgs Count Plot.

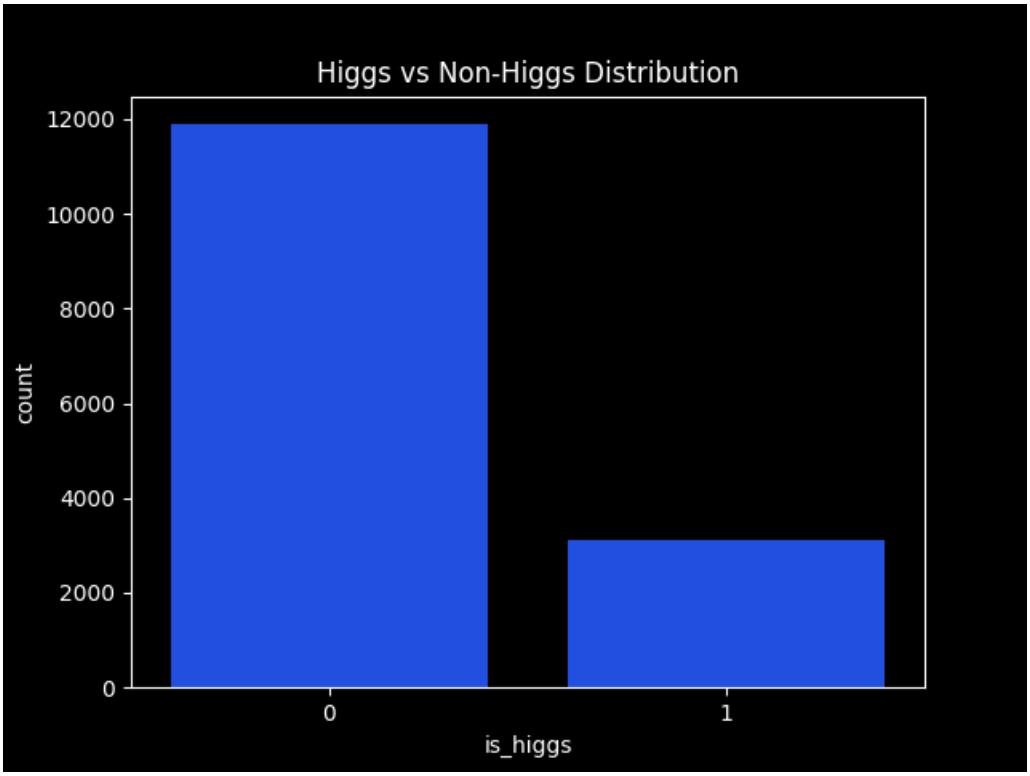

Figure 3: The graph shows the distribution of Higgs-like and non-Higgs events in the synthetic data sample. As it is shown, there are more non-Higgs events than Higgs-like ones; hence, the problem can be considered moderately imbalanced.

- Mass Distribution Histogram.

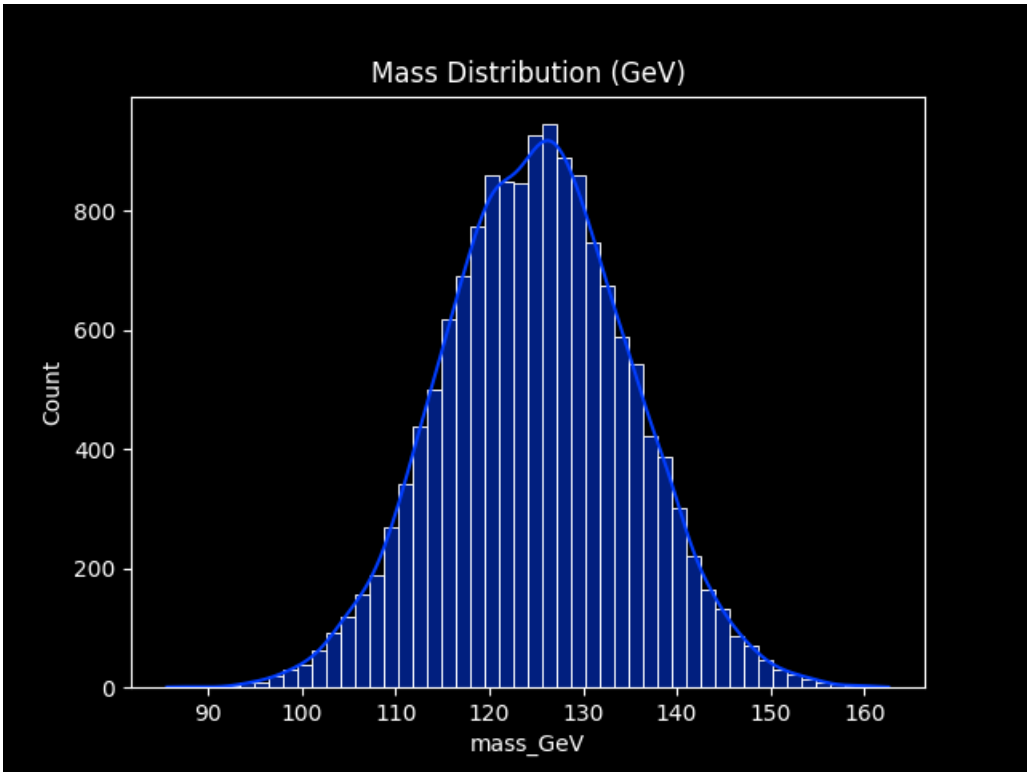

Figure 4: Graphs depicting the histogram of invariant mass values for each event. As one may see, the histogram shows the presence of a peak near the value of 125 GeV, which is a distinctive feature of the Higgs boson.

- Correlation Heatmap.

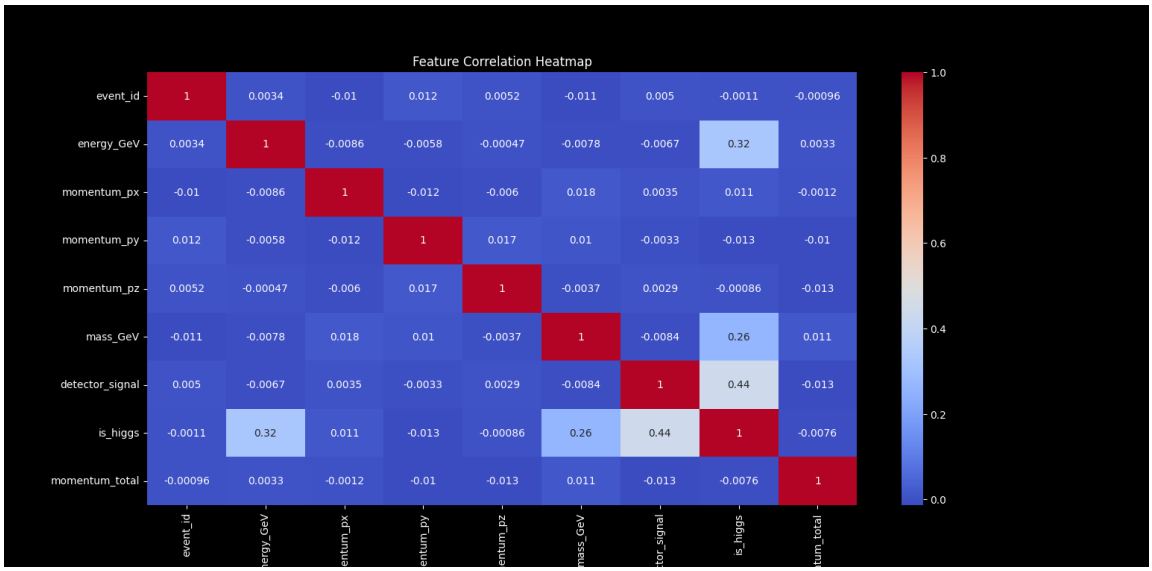

Figure 5: Heatmap that displays the correlation coefficients between each numerical feature and target variable `is_higgs`. There exist strong correlations between detector signal, energy, and invariant mass of particles.

- Feature Importance Plot.

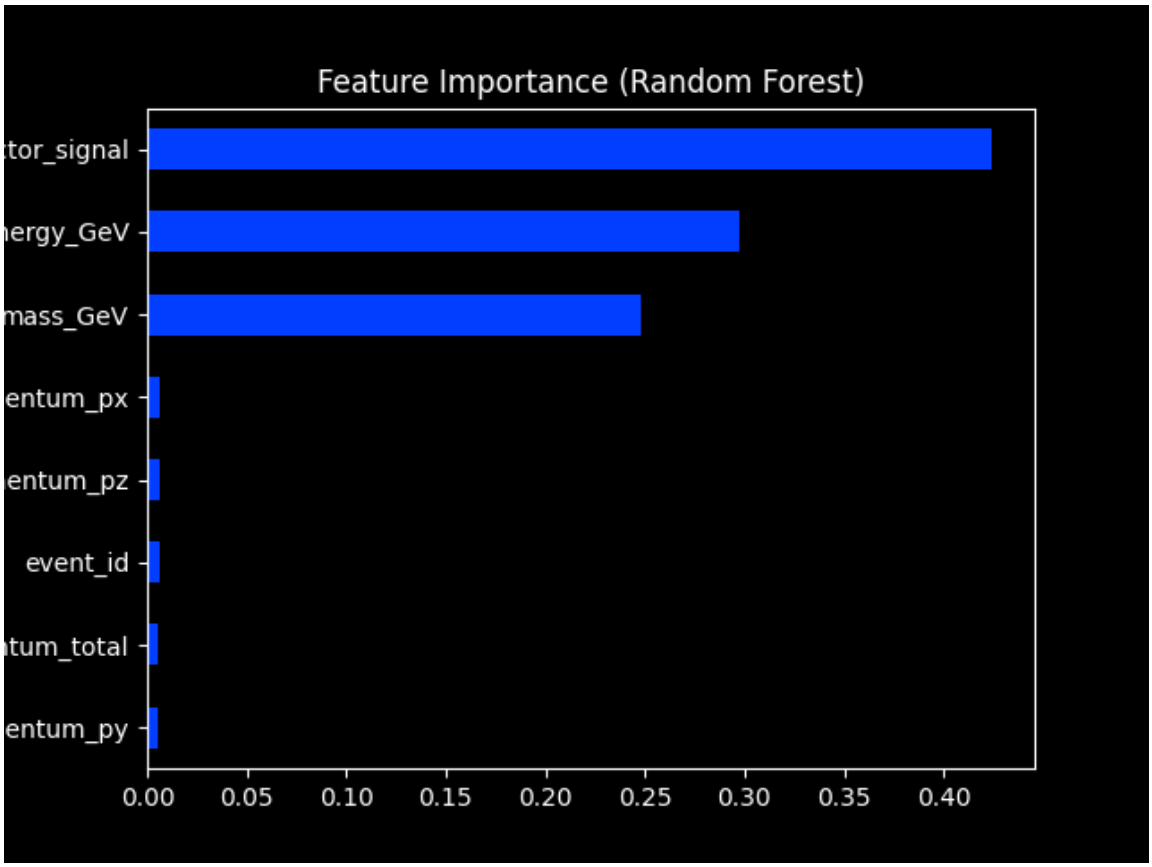

Figure 6: Feature importance measures are derived from the Random Forest classifier. Signal generated by the detector holds the highest significance among features, followed by the values for energy and invariant mass, whereas those of momentum are comparatively less important.

These screenshots were later inserted into the report.

3.14 Step 14: Summarize Observations

At the end of the activity, essential observations were made.

It should be noted that:

- Real particle collisions can be visualized interactively.
- Different particles have unique signatures.
- Histograms reveal hidden patterns in large datasets.
- Open data makes advanced science accessible to students.

3.15 Step 15: Final Reflection

The CMS Open Data portal fills the void between the classroom and real-world research. The above procedures provide an opportunity for novices to familiarize themselves with real-world data from one of the most significant experiments in modern physics.

4 Practical Learning Outcomes

This experiment allowed me to gain practical insight into the field of particle physics experimentation and data analysis. Before working on this assignment, my knowledge about high energy physics remained purely theoretical. With the help of the CMS Open Data Portal, I was able to work with real collision events collected by the CMS detector at CERN. Theoretical concepts became tangible visuals.

One of the main lessons that I learned was that proton-proton collisions create a great number of secondary particles. These particles traverse various detector layers, where their properties are measured. Using the CMS Event Display, I saw the distinctive features of electrons, muons, and photons. Electrons create energy showers in the electromagnetic calorimeter, muons go deep inside the detector, and photons register as electromagnetic showers with no charge track.

Moreover, I got the knowledge of using the invariant mass of a system of particles in detecting short lived particles like the Higgs boson. In the four-lepton and diphoton events, the chosen particles were created by a common process from which the characteristics could be extracted as if they were decayed from the Higgs boson. Thus, we can conclude that reconstruction of such unstable particles is done through their decay products.

Another very important lesson in this activity was how scientists used histograms as tools for finding out certain information from large numbers of measurements. Histograms represent large data sets visually, while the peaks of these distributions correspond to certain resonances, which, in turn, correspond to certain particles. For example, the resonance at 125 GeV on the invariant mass graph indicates the existence of the Higgs boson. Thus, the discovery of a certain particle is always based on statistics.

Furthermore, this task was beneficial because it helped me improve my skills in dealing with scientific databases and visualization software. Specifically, in the course of working on the project, I gained experience in browsing data on scientific portals, loading special data sets, and analyzing graphical results.

Most importantly, it made me realize that open science is an invaluable process. The decision by CERN to provide students with actual data from experiments makes it possible for them to engage in genuine scientific investigation. This project helped me understand that scientific investigation is not only reserved for professional scientists. As

long as one has the right resources, he/she can engage in the investigation of actual data.

5 How Visualization Helps in Particle Physics

Experiments in particle physics yield huge quantities of data expressed as numbers that typically run into millions. If we did not have visualization techniques available to us, interpretation of the information obtained would become impossible. Visualization turns abstract numbers into graphical information that helps discover physical correlations and patterns.

One of the best educational visualization techniques is the CMS Event Display. It creates an image of the CMS detector in three dimensions and displays the tracks and energy deposits created by the particles in the process of their interaction within the detector. With this tool, you can rotate and zoom the image of the detector and get insight into how particles behave within the detector. Thus, it becomes easier for you to see the geometry of the detector and signatures for particular particles.

In addition, visualization allows you to differentiate between charged and neutral particles. For instance, the trajectory of charged particles like electrons and muons will be curved due to the influence of the magnetic field; on the other hand, neutral particles like photons will not leave any traces in the form of tracks but rather energy deposits.

Another vital type of visualization is the histogram. Rather than analyzing each event individually, histograms allow us to look at the characteristics of hundreds, thousands, or millions of events at once. The peaks found in histograms represent the most common values for any measurable quantity and usually relate to the masses of known particles. For instance, the spike found at about 125 GeV in the graph of invariant masses is due to the presence of Higgs boson candidates.

In addition to making complex ideas understandable, visualization also plays an important role in detecting any abnormalities or discrepancies in results. Researchers use charts and event displays to help themselves understand their data and then present these results to other scientists. In educational settings, visualization helps students easily comprehend complicated topics.

In summary, visualization acts as a connection between numbers and physics. It enables us to see how events unfold in reality and understand how scientists make breakthroughs.

6 Benefits of CMS Open Data for Students

The CMS Open Data Portal can provide numerous benefits for students from schools and universities in terms of education. The portal gives the possibility of working with

actual experimental data which is much more effective than doing theoretical tasks in classrooms.

First of all, using the same information as professionals do while working in CERN makes students feel confident that modern science is within everyone's reach. In addition, such experience will be useful for future work in laboratories since students would already be familiar with this kind of research.

Secondly, students get the possibility of using inquiry based learning which includes looking through data, drawing histograms, and analyzing results. Such hands-on approach is much better than just learning something in theory.

Thirdly, students learn how to perform computations and analyze data as well as how to use certain software tools. They develop useful skills in the sphere of data visualization and presentation. Moreover, those skills will be needed in the future during studying some technical subjects including physics.

CMS Open Data provides very flexible materials that could be utilized in various forms. New users can make use of browser-based programs, while the more advanced users can do data analysis through Python, R, or even spreadsheets.

Another thing that CMS Open Data has provided is the inspiration among students who could replicate the discoveries like the discovery of the Higgs boson, which pushes them to study more in the fields of science and technology.

In conclusion, CMS Open Data provides:

- Access to authentic particle physics data.
- Hands on scientific exploration.
- Development of computational and analytical skills.
- Exposure to modern research techniques.
- Opportunities for interdisciplinary learning.
- Inspiration for future careers in STEM.

7 Challenges Faced

Even though the main purpose of the CMS Open Data Portal is education, several difficulties have been met while working on this project.

The first difficulty is connected to the terms that I had never come across before because of my lack of prior knowledge about the particle physics. Invariant mass, calorimeter, muon chamber, and event reconstruction are examples of terms that needed further explanations.

Secondly, the navigation of the CMS Event Display itself turned out to be quite hard since the application has lots of settings and requires one to spend additional time figuring out how to load events and switch on particular types of particles.

Furthermore, at first, the interpretation of the detector images was a problem for me since the colored tracks and energy deposits looked confusing. But, after exploring different events and finding appropriate documentation, things became much more clear.

Finally, my computer performance and loading time of browsers were factors that could affect the process.

Still, all educational documentation provided by the portal has a good structure and seems to be easy for beginners.

8 Applications in Education

CMS Open Data portal holds great educational potential and may be incorporated into the curriculum on different levels.

At school, the portal can be applied for illustrating the key topics related to contemporary physics – elementary particles, interactions between them, as well as the structure of matter. Visualization of event displays makes the material more vivid and easier to grasp.

For undergraduates, CMS Open Data provides an excellent ground for practical work in such areas as particle physics, statistics, and programming. Students can study real datasets, create their own histograms, and compare experimental data with theoretical predictions.

CMS Open Data allows working on projects, when students can carry out independent research, choose interesting decay channels, and even write reports similar to scientific papers.

Moreover, the portal may be used by teachers for demonstrating the process of doing science – starting from collecting data and ending with drawing conclusions based on evidence.

Furthermore, the resource might prove handy in the process of self-education or organizing events for clubs, science days, etc. Given the availability of the material in the open web, everybody has free access to it.

9 Conclusion

CMS Open Data Portal is an outstanding educational tool that gives young people the opportunity to get acquainted with real physics research. By making publicly available collision data and visualization tools, CERN managed to create a resource where advanced physics concepts can be comprehended by beginners.

During this experiment, I got familiar with the process of recording proton-proton collisions, methods for identifying various particles inside the CMS detector, and how statistics tools like histograms help recognize the existence of particles like the Higgs boson. Event display converted raw data from the detector to comprehensible graphics, and histogram visualization showed how discoveries can be made based on massive data.

Also, this project emphasizes the significance of open science. Having access to experimental data from one of the most vital scientific investigations makes the scientific process transparent and reproducible.

Overall, this project has helped me to enhance my knowledge in the field of particle physics, and to develop my skills in analyzing data, and get insight into the techniques used by researchers in CERN. CMS Open Data portal is a brilliant entry point for any individual who wants to delve into high energy physics.

10 References

1. CERN Open Data Portal. https://opendata.cern.ch
2. CMS Guide for Education Use of CMS Open Data. https://opendata.cern.ch/docs/cms-guide-for-education
3. CERN Official Website. https://home.cern