JOURNAL OF
EMERGING INVESTIGATORS

Article

Jet optimization using a hybrid multivariate regression model and statistical methods in dimuon collisions

Krishna Chunduri[1], Vishnu Srinivas[2], Larry McMahan[3]
[1] Mission San Jose High School, Fremont, California
[2] Foothill High School, Pleasanton, California
[3] Aspiring Scholars Directed Research Program, Computer Science and Engineering Dept., Fremont, California

SUMMARY

Collisions of heavy ions, such as muons result in jets and noise. In high-energy particle physics, researchers use jets as crucial event-shaped observable objects to determine the properties of a collision. However, many ionic collisions result in large amounts of energy lost as noise, thus reducing the efficiency of collisions with heavy ions. The purpose of our study is to analyze the relationships between properties of muons in a dimuon collision to optimize conditions of dimuon collisions and minimize the noise lost. We used principles of Newtonian mechanics at the particle level, allowing us to further analyze different models. We used simple Python algorithms as well as linear regression models with tools such as sci-kit Learn, NumPy, and Pandas to help analyze our results. We hypothesized that since the invariant mass, the energy, and the resultant momentum vector are correlated with noise, if we constrain these inputs optimally, there will be scenarios in which the noise of the heavy-ion collision is minimized.

INTRODUCTION

The Standard Model classifies particles into two categories: fermions and bosons. Fermions make up matter and include quarks, which combine to form protons, neutrons, and other hadrons; and leptons, such as electrons and neutrinos. Bosons are responsible for mediating the fundamental forces and include the photon (electromagnetic force), gluon (strong nuclear force), W and Z bosons (weak nuclear force), and the Higgs boson (responsible for giving particle mass) (1).

Muons share similar properties with electrons but have approximately 207 times more mass (2). Although collisions between other subatomic particles, such as gluons, may apply to our research because gluon collisions may also create noise, our research is based specifically on properties of muons which may not be the same as properties of different subatomic particles. In dimuon collisions, two muons are sent towards each other at high speeds, upon which they collide and release energy. We studied specific events in CERN particle colliders and visualized them using CERN's visualization software **(Figure 1)**.

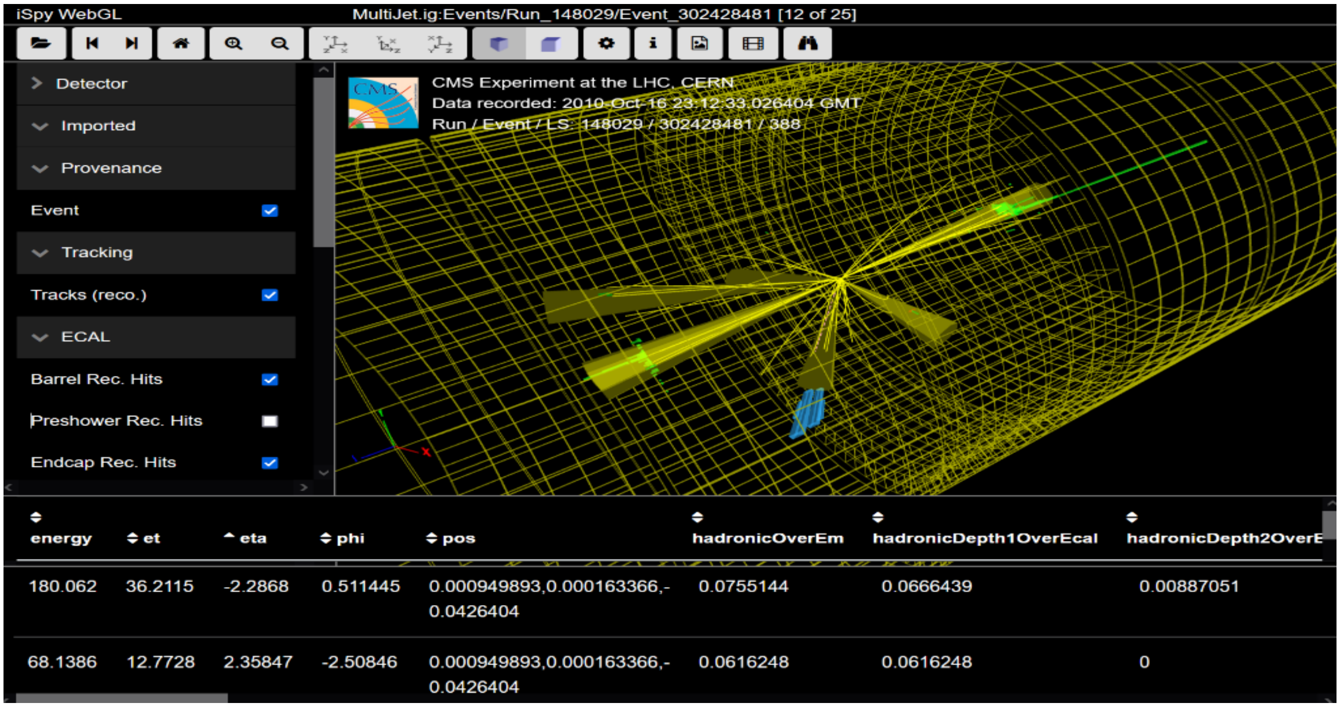
The energy that is produced in these collisions is in the form of jets and noise. Jets in particle physics or heavy ion experiments are slim cones of hadrons that emerge from the hadronization process, wherein quarks or gluons fragment into bound states known as hadrons. During these collisions, energetic quarks and gluons are generated. As both the quarks and gluons travel away from the collision point, they emit additional gluons, which can undergo further splitting, generating additional quarks, antiquarks, and gluons in a cascading manner. These particles and the associated energy are measurable through various detectors in particle physics experiments (3).

Many different properties of dimuon collisions are measured by the CERN data sets. The initial energy is a quantity that refers to the total energy of each respective muon before the collision of both muons. The group at CERN measured such using giga-electron volts (GeV) which is represented by E_1 and E_2 in our dataset. Momentum in x-, y-, and z-directions are values that represent the 3D vector components of the muon's momentum, defined by P = mv, where m is the mass and v is the velocity of the object. Transverse momentum is another quantity given in both datasets, described by CERN (European Council for Nuclear Research) as the "amount of a particle's momentum perpendicular to the beam [jet] direction" (4). Invariant mass, also known as rest mass, is a term in particle physics that is directly proportional to a particle's inertia when it is stationary. The inferred mass value remains unchanged regardless of the reference frame used to measure the energies and momenta, making it "invariant." By analyzing the energies and momenta of the decay products, one can calculate the mass of a particle prior to its decay (5). This concept is frequently extended to muons emerging from a collision, allowing invariant mass to be determined in the dataset.

Initially, when examining our dataset, we perceived noise values to be insignificant in dimuon collisions due to their small quantitative nature (ranges from one part in 10 million to one part in 100 million depending on the collision). The fluctuation-dissipation theorem contradicts this observation, stating that even small amounts of noise can lead to large fluctuations (6). Work by Kapusta and Young highlights that heavy ion collisions are subject to second-order dissipative fluid mechanics, which is the study of the densely packed matter and energy resulting from particle collisions. It reveals the importance of these seemingly small amounts of noise (7). Consequently, it becomes crucial to minimize noise levels, even at a small scale, as the matter produced can lead to severe issues if noise is treated non-perturbatively.

Background modeling and event reconstruction are commonly used to interpret collision data. However, due to even small amounts of noise introducing inaccuracies or distortions in the detector signals, it can be challenging to reconstruct these events. These issues could impair the accuracy of these reconstructions and modeling.

Issues related to noise profoundly affect background modeling and event reconstruction. The primary goal of background modeling is to eliminate interference from



CMS Experiment at the LHC, CERN
Data recorded: 2010-Oct-16 23:12:33.026404 GMT
Run / Event / LS: 148029 / 302428481 / 388

energy	et	eta	phi	pos	hadronicOverEm	hadronicDepth1OverEcal	hadronicDepth2OverE
180.062	36.2115	-2.2868	0.511445	0.000949893,0.000163366,-0.0426404	0.0755144	0.0666439	0.00887051
68.1386	12.7728	2.35847	-2.50846	0.000949893,0.000163366,-0.0426404	0.0616248	0.0616248	0

Figure 1: CERN's simulation/visualization of a dimuon collision. This representation shows two muons colliding within the constraints of simulation and displays the properties of such a collision. We created this image using CERN's software.

the background and identify desired signals. In heavy ion collisions, data from detectors are used to reconstruct particle trajectories and properties. Noise can introduce inaccuracies or distortions in the detector signals, making it significantly more difficult to reconstruct events accurately. As a result, these challenges can significantly influence the measurement of quantities such as transverse momentum or invariant mass of the involved particles (8).

Another issue caused by noise from collisions is impaired particle identification and measurement accuracy. Detectors in particle colliders detect the distinct signatures that certain particles leave behind. Noise can even mask or even change these signatures, causing the identity of a particle to be mistaken for another (9).

To address these problems, we aimed to determine how optimizing collision parameters could reduce noise. Specifically, we focused on identifying commonalities between the top quartile of the events in our data with the least noise. Comparing the ranges of the input parameters for the full dataset and the top quartile showed us that the spreads of the distribution in the top quartile were generally between 10% and 50% of the spreads in the full dataset (Appendix A). While this seems reasonable, these ranges didn't account for the other parameters such as the momenta in different directions and the invariant mass. By addressing which factors play a role in noise reduction, our research aims to enhance our understanding of noise reduction in particle collisions and provide valuable insights for particle accelerators to optimize collision processes effectively. In the end, these noise reductions can help us understand how to conserve energy in particle accelerators which would lead to better collision data from the same.

RESULTS

We used the mathematical and Newtonian relationships between the different properties of the muon (i.e. invariant mass, momenta, and energies of the muons). We used the matplotlib library in python to graph the relationships between different inputs and check for correlations (10). In the scatter plot illustrating the relationship between noise (energy lost) and each invariant mass and resultant momentum sum, we were able to see that the invariant mass did not appear to correlate with noise, while the resultant momentum sum seemed to have a positive linear correlation with noise **(Figure 2)**.

The second scatter plot shows the relationship between the initial energy of the first muon (E_1) and the invariant mass (m), revealing a relatively linear correlation **(Figure 3)**. However, it is important to keep in mind that the error is large for this initial test as it hasn't been linearized, which in this case is very difficult due to the complexity of the equation below.

Similarly, the relationship between the initial energy of the second muon (E_2) and the invariant mass (m) shown in the third scatterplot mirrors the relationship in the scatterplot before it **(Figure 4)**.

The next step was the derivation of an equation using principles of Newtonian mechanics and principles of the conservation of energy to understand which inputs influence the noise and which inputs are not necessary to analyze in our model. This relationship is a critical portion of our research.

$$Noise = E_1 + E_2 - \sqrt{((px_1 + px_2)^2 + (py_1 + py_2)^2 + (pz_1 + pz_2)^2 + m^2)}$$

In this equation, the components of p in the x direction (px_1 and px_2) represent the momentum of the first and second muon along the x-axis in the 3d plane. The components of p in the y direction (py_1 and py_2) represent the momentum of the

https://doi.org/10.59720/23-180

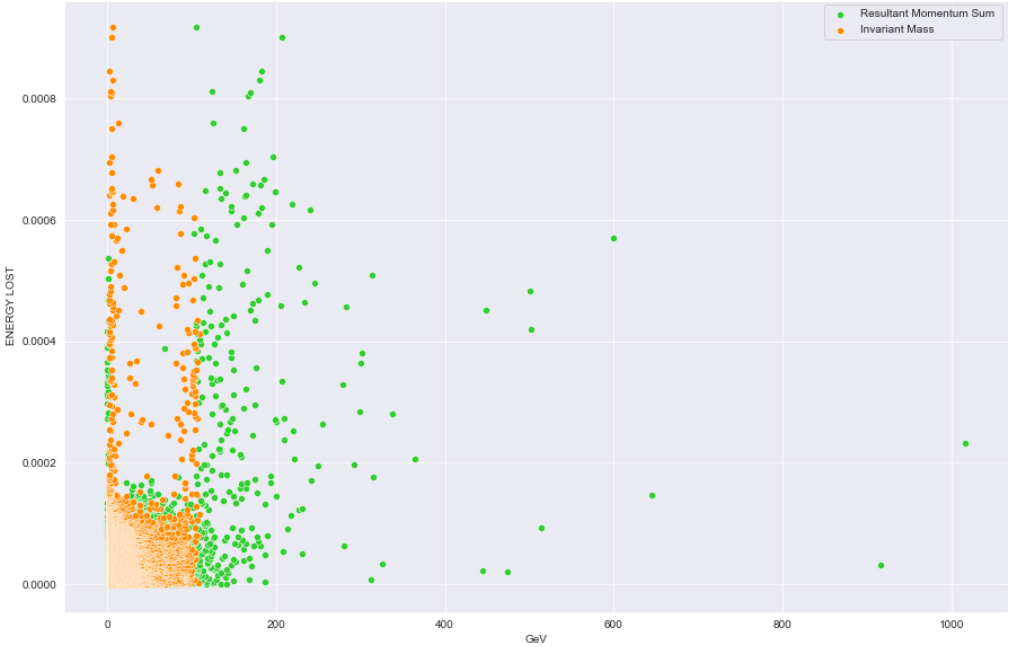


Figure 2: Scatter plot of resultant momentum and invariant mass with noise. This figure is a visualization of the relationship between Resultant Momentum Sum (green) and Invariant Mass (orange) with Noise (energy lost). The X-axis and Y-axis are measured in GeV. The figure shows steep linear relations for the resultant momentum sum but just a cluster for the invariant mass.

first and second muon along the y-axis in the 3d plane. The components of p in the z direction (pz$_1$ and pz$_2$) represent the momentum of the first and second muon along the z-axis in the 3d plane. Additionally, m represents the invariant mass and E$_1$ and E$_2$ represent the initial energies of the first and second muon respectively.

To confirm that the experimentally determined noise can be described by this equation, we used a multivariate linear regression model to analyze the trends in the data and found that the equation above fits our data with high accuracy. The accuracy for the model with feature engineering was roughly around 99% which shows our data follows our theoretical model with little error. Additionally, we ran the model on the 2011 dimuon dataset as well (raw dataset with 100,000 observations) and got scores of 91% accuracy, confirming that the formula is consistent for data sets for dimuon collisions outside of the 2010 dimuon collision dataset (contains 50,000 observations).

To find the experimental ranges of our dataset, we used mathematical Python functions (Appendix B) as well as the

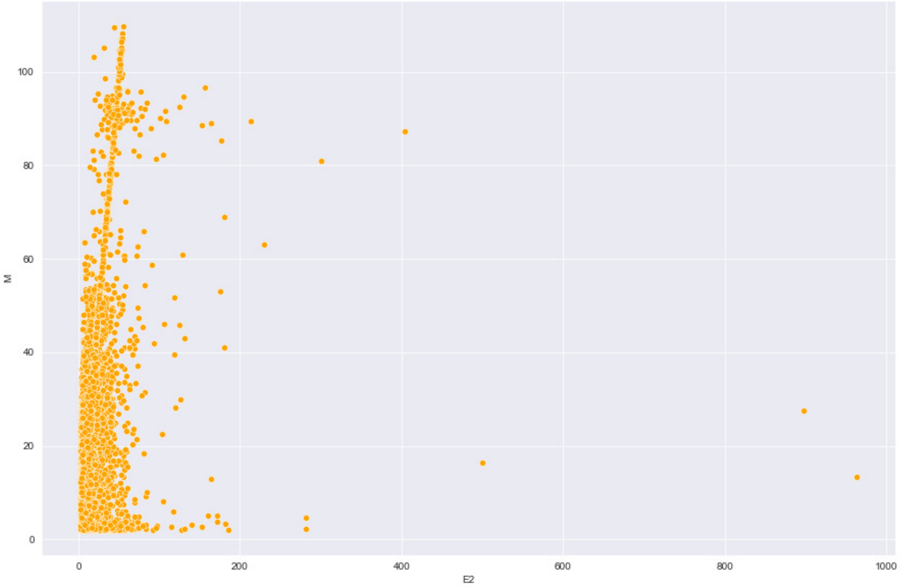


Figure 3: Scatter plot between E$_1$ and M. This figure is a visualization of the correlation between E$_1$ and M. It was useful for visually confirming trends with the energy of the first muon and Invariant Mass (M) with respect to GeV. Additionally, it showed a strong linear correlation between E$_1$ and M with a large slope magnitude.

https://doi.org/10.59720/23-180

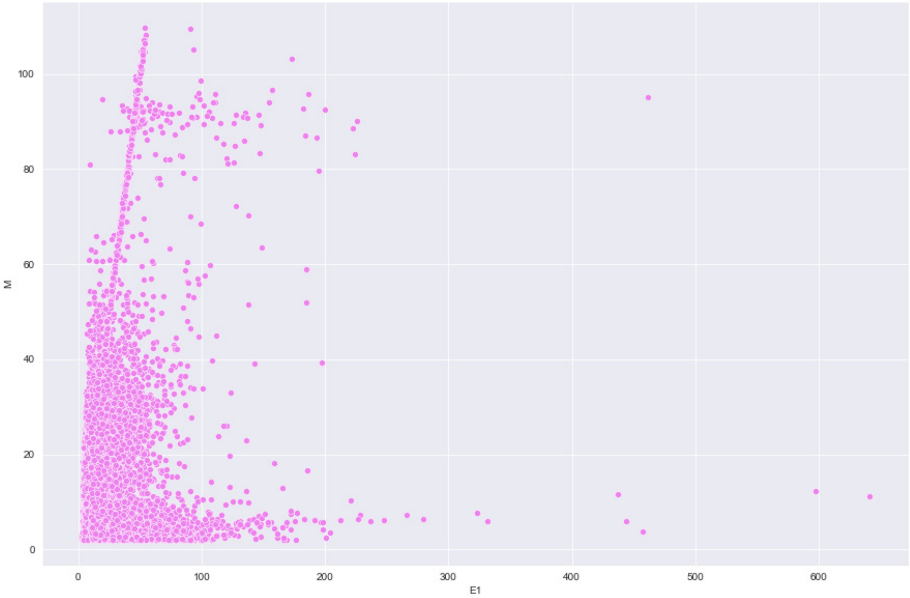


Figure 4: Scatter Plot Between E_2 and M. This figure is a visualization of the correlation between E_2 and M. It was useful for visually confirming trends with the energy of the second muon and Invariant Mass (M), each in GeV. Additionally, it showed a strong linear correlation between E_2 and M with a large slope magnitude.

noise equation we derived using methods mentioned in materials and methods section to find the ideal ranges that would minimize the noise as in the top quartile of our data. For our initial parameters, we used the median values of each variable within the top quartile and selected the noise to be the median of the top quartile of noise values. We then solved for each target variable, concerning the other assumed values. We then worked backwards to find the ideal ranges for the rest of the input parameters.

DISCUSSION

Elevated noise levels have a pronounced impact on jet and particle identification, rendering the differentiation between various particle types and the accurate measurement of their momenta and energies challenging. Initially, our dataset suggested insignificant noise in dimuon collisions, but the fluctuation-dissipation theorem revealed that even minor noise could lead to significant fluctuations. Consequently, it becomes imperative to minimize noise levels, no matter how minor they may seem, as the produced matter could lead to inefficient collisions if noise is not treated perturbatively.

Given the risks associated with high noise levels, our research aimed to identify the optimal property ranges of muons to avoid the aforementioned problems. To conduct dimuon collisions in a particle collider, these ranges would help with minimal noise loss and jet optimization. By examining the top quartile for noise (noise range), we found the ideal ranges for properties of dimuon collisions (Appendix C).

Our model was tested and fitted on the 2010 and 2011 dimuon collision CERN datasets (11, 12). One limitation to this is that the data may not be as accurate as newer datasets released by CERN, which were taken after significant improvements in the accuracy and precision of their particle detectors and data storage. Another limitation of our research is that our model was only tested on dimuon collisions. While these dimuon collisions do share many similarities to other heavy ion collisions, we might not be able to explicitly apply our research to other heavy ion collisions. It may be possible in future studies to generalize our approach to other types of particle collisions.

By finding ideal ranges for different properties of particles in collisions that minimize the amount of noise produced, the methods highlighted in this paper can be utilized in future experiments so that the outcome of these collisions is not masked. This in turn will make it easier to study what happens to particles and jets in dimuon collisions. Furthermore, background modeling and event reconstruction will be less difficult with reduced noise. This could potentially lead to more efficient setups in future experiments to minimize the amount of noise produced from these collisions.

While the ranges found here will apply solely to other dimuon collisions of similar nature, noise ranges for many other types of particle collisions can be found using similar methods. However, for these methods to be viable, there must be enough data such that two different sets can be used: one to train the model and one to test the model. For example, we used the 2010 and 2011 CERN datasets. It is also important that the model is not overfitted to the data when making a new model for another data set. As a result, it is vital that the model should have a number of parameters justified by the variables (factors found in the equation).

Our proposed ranges for the ideal feature values suggests that the ranges for E_2 are lower than that of E_1 which implies that the muon with higher energy should be assigned to the first muon and the muon with lower energy should be assigned to the second.

Our research suggests many other possibilities for future experiments. One approach could use a neural network instead of a linear regression model. A neural network using a similar data set could lead to more accurate findings and possibly recognition of other patterns that we could not find in the original experiment. Overall, the implications of our research may help optimize future experiments in the field of particle physics by providing precise ranges of input value that decreases the noise of dimuon collisions.

MATERIALS AND METHODS

Data from the CERN heavy-ion 2010 and 2011 data sets were imported and then cleaned, deleting duplicate values, incomplete data rows, and unphysical values. Outliers were then isolated and removed by utilizing Python libraries such as NumPy, Pandas, Matplotlib, and Seaborn by computing the z-score of each data point and comparing it against each other (13, 14, 15). These libraries were also useful for visualization to see physical patterns and relationships. The z-scores were computed through z-score normalization.

Implementing Machine Learning

A variety of machine learning algorithms were implemented using the sci-kit-learn library in Python (16). Mainly, we were able to utilize multivariate linear regression, in order for our model to function. We built the models using the sci-kit-learn library and made appropriate modifications to fit our data. Initial accuracies were between 30-40% but through methods like normalization and feature engineering, we were able to raise that score to above 98% for the 2010 dataset and above 90% on the 2011 data. We confirmed the 2011 data set with both models to make sure this wasn't a case of overfitting.

Physical Methods/Newtonian Derivation

When considering a system of two separate particles, the law of conservation of energy states that the sum of the final energy and the energy lost is equal to the sum of the initial energies of both particles. In other words:

$$E_1 + E_2 - Final\ Energy = Energy\ Lost = Noise$$

We make the following substitution for the Final Energy (17,18):

$$E_1 + E_2 - \sqrt{((resultant\ momentum)^2 + m^2)} = Noise$$

Finally, we can consider the resultant momentum of each particle in each direction (x, y, z), and write the momentum vector in terms of these quantities. This yields a final result of

$$E_1 + E_2 - \sqrt{((px_1 + px_2)^2 + (py_1 + py_2)^2 + (pz_1 + pz_2)^2 + m^2)} = Noise$$

for the total energy lost (19).

Statistical Methods

Initially, our data, as displayed by the graphs, was noisy and didn't show clear relationships. Accuracy of the linear regression model was also considerably lower. To fix this, we found and removed outliers in our data. To identify outliers in our data set, we computed the z-score of the noise of each collision, using the formula for standard deviation below:

$$\sigma = \sqrt{\left(\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}\right)}$$

In this formula, μ is the mean of the population, N is the size of the population, and x_i is each particular value. Then, we removed any value outside 3 standard deviations above or below the mean (a z-score of 3 or -3). This was the z-score method we used to make the data less skewed and more normal.

Our original 2010 dataset had exactly 50,000 events, but after the z-score method to remove outliers, our dataset lost 981 values, retaining 49,021 observations (roughly 98% of the data). Generally, these outliers were above the threshold (z-score greater than 3), and there were very few in comparison with the total number of observations. However, we judged by empirical rule that the number of outliers we excluded was not excessive.

ACKNOWLEDGEMENTS

We would like to thank the Aspiring Scholars Directed Research Program, the European Council for Nuclear Research, and the Olive Children's Foundation for their contributions in facilitating this research and providing data to use in this study.

Received: July 31, 2023
Accepted: November 26, 2023
Published: June 09, 2024

REFERENCES

1. "The Standard Model." *CERN*. home.cern/science/physics/standard-model. Accessed 23 Mar. 2024.
2. "What Is a Muon?" *Vanderbilt University*. www.hep.vanderbilt.edu/~gabellwe/qnweb/qnpptr/What_is_a_Muon.pdf. Accessed 24 Mar. 2024.
3. Kovner, Alexander and Urs Achim Wiedemann "Gluon Radiation and Parton Energy Loss." *Quark–Gluon Plasma 3,* 2004, pp. 192-248. https://doi.org/10.1142/9789812795533_0004.
4. "Glossary: CMS Experiment" *CERN*. cms.cern/content/glossary#:~:text=Transverse%20momentum%20(and%20energy),from%20the%20observed%20transverse%20momenta. Accessed 27 Jul. 2023.
5. "Mass / Invariant Mass." *ATLAS Experiment at CERN*. atlas.cern/glossary/mass. Accessed 23 Mar. 2024.
6. Deserno, Markus. "Fluctuation-Dissipation Theorem for Brownian Motion." *Carnegie Mellon University*, 14 Sept. 2004. www.cmu.edu/biolphys/deserno/pdf/FD.pdf. Accessed 24 Mar. 2024.
7. Kapusta, J.I., and C. Young. "Causal Baryon Diffusion and Colored Noise for Heavy Ion Collisions." *Nuclear Physics A*, vol. 931, 1 Sept. 2014, pp. 1051-1055. https://doi.org/10.1016/j.nuclphysa.2014.08.095.
8. Singh, Manyak, et al. "Hydrodynamic Fluctuations in Relativistic Heavy-Ion Collisions." *Nuclear Physics A*, vol. 982, 22 Jan. 2019, pp. 319-322. https://doi.org/10.1016/j.nuclphysa.2018.10.061.
9. "CERN Accelerating Science." *CERN*. home.cern/science/experiments/how-detector-works. Accessed 23 Mar. 2024.
10. Hunter, John D. "Matplotlib: A 2D Graphics Environment" *Computing in Science & Engineering*, vol. 9, no. 3, May-June 2007. https://doi.org/10.1109/MCSE.2007.55.
11. McCauley, Thomas; (2014). "Events with two muons from 2010." *CERN Open Data Portal.* doi: https://doi.org/10.7483/OPENDATA.CMS.4M97.3SQ9
12. McCauley, Thomas. "Datasets derived from the Run2011A SingleElectron, SingleMu, DoubleElectron, and DoubleMu primary datasets", opendata.cern.ch/record/545. Accessed 9 Nov. 2023.
13. The Pandas development team. "Pandas-Dev/Pandas: Pandas." *Zenodo*, 23 Feb. 2024. https://doi.org/10.5281/

https://doi.org/10.59720/23-180

zenodo.3509134

14. McKinney, W. (2010) Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, Austin, 28 June-3 July 2010, 56-61. https://doi.org/10.25080/Majora-92bf1922-00a
15. Waskom, M. L., (2021). “seaborn: statistical data visualization.” *Journal of Open Source Software*, 6(60), 3021, 06 Apr. 2021. https://doi.org/10.21105/joss.03021.
16. Pedregosa, Fabian, et al. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research*, Oct. 2011. https://doi.org/10.48550/arXiv.1201.0490.
17. J. Beringer, et al. “Review of Particle Physics.” *Physical Review D*, American Physical Society, 20 Jul. 2012. 10.1103/PhysRevD.86.010001.
18. Jackson, J.D. and D.R. Tovey (Sheffield)“38. Kinematics - Particle Data Group.” Particle Data Group. https://pdg.lbl.gov/2008/reviews/rpp2008-rev-kinematics.pdf. Accessed 27 Jul. 2023.
19. “DiMuon Histogram Excel Instructions.” CERN. indico.cern.ch/event/614923/attachments/1433684/2203373/3_DiMuonHistogramExcelInstructions.pdf. Accessed 27 Jul. 2023.