



16 August 2025
kylene.jordan.monaghan@cern.ch

Dijet Angular Distribution Analysis in Search for New Physics

Kylene Monaghan

Supervisors Dr. Conor Henderson and Dr. Nathan Grieser
CERN, CH-1211 Geneva, Switzerland

Summary

This document serves as a record of my CERN Summer Student project and my research into dijet angular distributions with the LHCb. Over the course of the summer, I have investigated the properties of dijets and worked with Monte Carlo simulations to examine potential Beyond Standard Model effects that could be seen by observing dijets. This project is a non-resonant search for new physics which provides sensitivity for Parton Distribution Functions as well as dijet distributions between different types of jets. Moreover, I provide a discussion and training of a Boosted Decision Tree for jet classification between quark and gluon initiated jets.

Contents

1	Introduction	2
2	Simulation and PDF Sensitivity	3
3	Results and Data Analysis	4
3.1	Sensitivity for Parton Distribution Functions	4
3.2	Dijet Angular Distributions	5
3.3	QCD Verification in Split Processes	10
3.4	Boosted Decision Tree and Jet Classification	12
4	Conclusion	14

1 Introduction

Jets are an essential feature in particle physics for reconstructing events and studying the behavior of particles in colliders. A jet is a collimated spray of particles initiated by the scattering of a parton after a collision. This spray forms a cone which has kinematic properties itself, simplifying event analysis and allowing for meaningful reconstructions of particle decays. Dijets, formed in a collision event with two high energy jets, are a powerful tool for probing for new physics. Figure 1 displays a general diagram of the formation of a dijet and the angle off of the beam line. This research seeks to analyze the angular distributions of dijet events and observe Standard Model and Beyond Standard Model simulated effects.

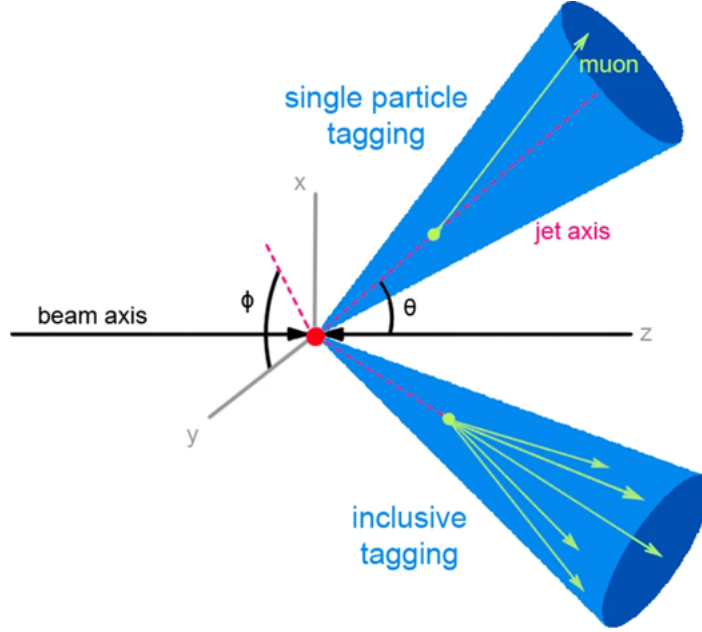


Figure 1: Diagram for basic dijet process

The primary observable in this study is the χ_{dijet} , a function of the angular displacement from the beam line which is given by the following equation:

$$\chi_{dijet} = e^{|\Delta y|} = \frac{1 + \cos \theta^*}{1 - \cos \theta^*} \quad (1)$$

where Δy is the difference in rapidity between the two jets, and θ^* is the polar scattering angle. This variable is Lorentz invariant and is particularly useful as a probe for new physics. It is sensitive to scattering effects off of the beam line, regardless of the type of parton interactions occurring. Importantly, it is insensitive to uncertainties within the Parton Distribution Function (PDF) due to the fact that the angular distributions of dominant QCD processes are similar. Therefore any changes within a parton-parton interaction beyond QCD interactions will modify the angular distributions, typically leading to wider angles relative to standard model QCD predictions [1]. In addition to χ_{dijet} , I examine the $\Delta\phi$, $\Delta\eta$, and M_{jj} variables when studying generated events.

The dijet angular distributions are analyzed to compare the Standard Model with the Arkani-Hamed-Dimopoulos-Dvali (ADD) model. This model offers a potential solution to the hierarchy problem by postulating the existence of compactified large extra dimensions. It predicts signatures of a virtual graviton to appear in nonresonant dijet angular distributions which I am studying [2] [3].

2 Simulation and PDF Sensitivity

For all event simulations, I have used Pythia 8.315 for Monte Carlo event generation. By modifying a C++ script for Pythia, I have then generated millions of events. The general assumptions and jet cuts are shown below. The acceptance incorporates an estimated trigger efficiency jet reconstruction values.

Assumptions	Cuts
Integrated Luminosity = $30fb^{-1}$	Jet PT min = 20 GeV
Efficiency = 30%	$2.0 < \eta < 5.0$
Acceptance = $\frac{1}{1250}$	

The simulations used are at generator level, and I do not move into detector smearing or further simulation techniques. The goal of examining BSM effects using the LHCb sensitivity range for dijet analysis does not require more advanced simulation for the scope of this project. Once I generate my event information and export it into a ROOT file, I use Python and RDataFrame to analyze data and create plots. I first began with the Standard Model and normalized the data for my four observables (see Figure 2).

After this initial simulation, I proceeded to reweight my distributions via Parton Distribution Functions (PDFs) supplied through LHAPDF. The PDF gives the probability that a specific parton will have a certain momentum fraction of the total proton momentum. Therefore given the energy fraction and the flavor ID, I have enough information to convert one PDF into another replica without having to regenerate a new event. This process is called reweighting, and there are several options which can be called via LHAPDF. I have in Figure 3 compared three of these methods: The NNPDF, MSHT, and CT18. The former uses a neural network to divide datasets and through training and validation creates Monte Carlo replicas to do the reweighting, while the latter two use Hessian eigenvalues to determine the error and differences when reweighting. MSHT gives a symmetric error while CT18 can provide asymmetric or symmetric errors. I was interested in comparing the difference of each of these methods to see how they would change the reweighting of my data. I have used the next to leading order option for each reweighting. The default PDF in Pythia 8.315 is “NNPDF23_lo_as_0130_qed” and the PDFs I have converted to using LHAPDF are “MSHT20nlo_as120”, “CT18NLO”, and “NNPDF40_nlo_as_01180”. I will discuss the resulting Figure 3 further in the Results section.

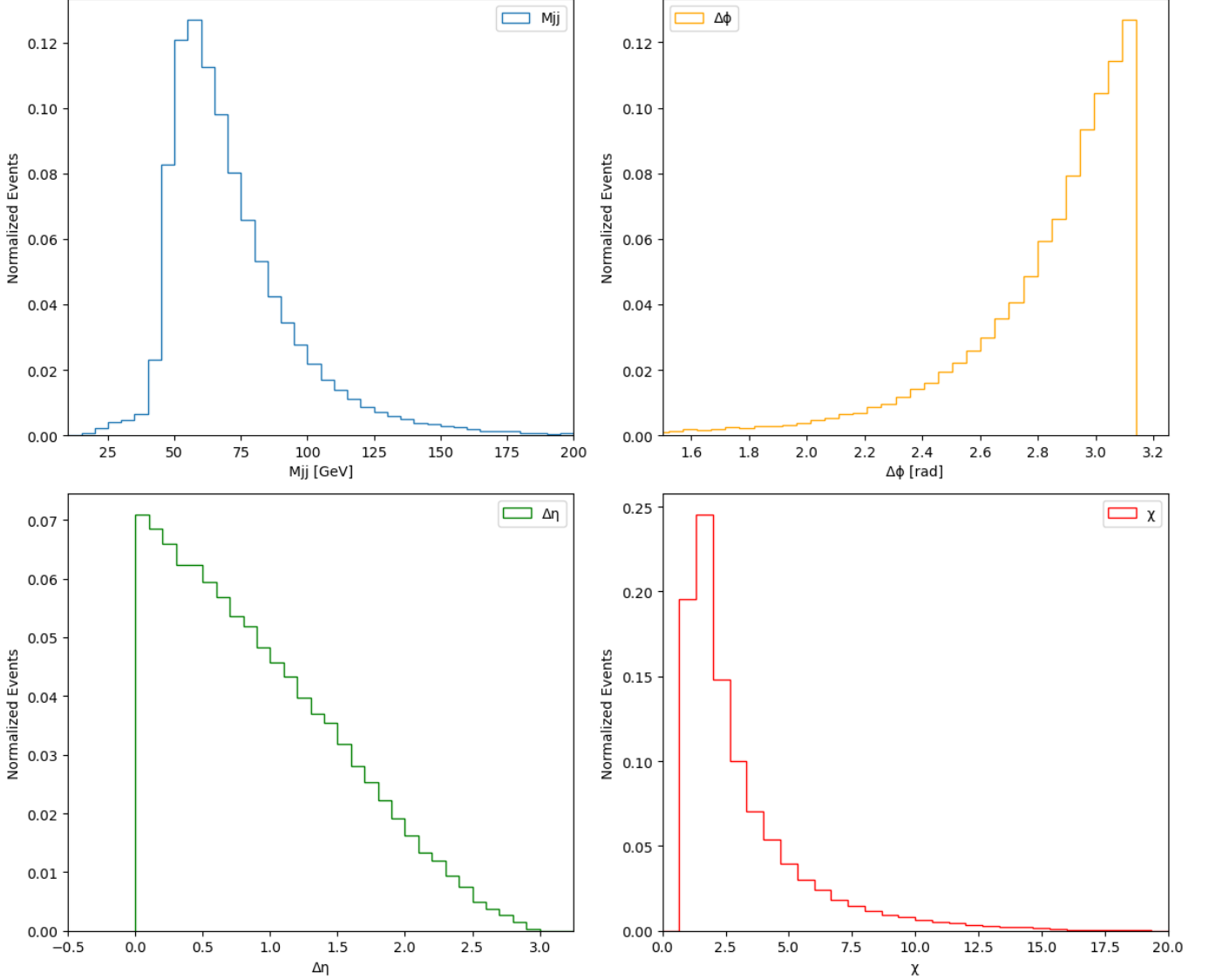


Figure 2: Normalized distributions of four main observables prior to PDF reweighting.

3 Results and Data Analysis

3.1 Sensitivity for Parton Distribution Functions

While there are no dramatic differences between the three methods of reweighting, Figure 3 shows that the LHCb parameters in this study provide enough sensitivity to distinguish between them. This information can be used to further improve PDF models in the future and reduce the errors calculated by the various PDF groups. The ratios are plotted with $\frac{MSHT}{CT18}$ in blue and $\frac{NNPDF}{CT18}$ in green for simplicity.

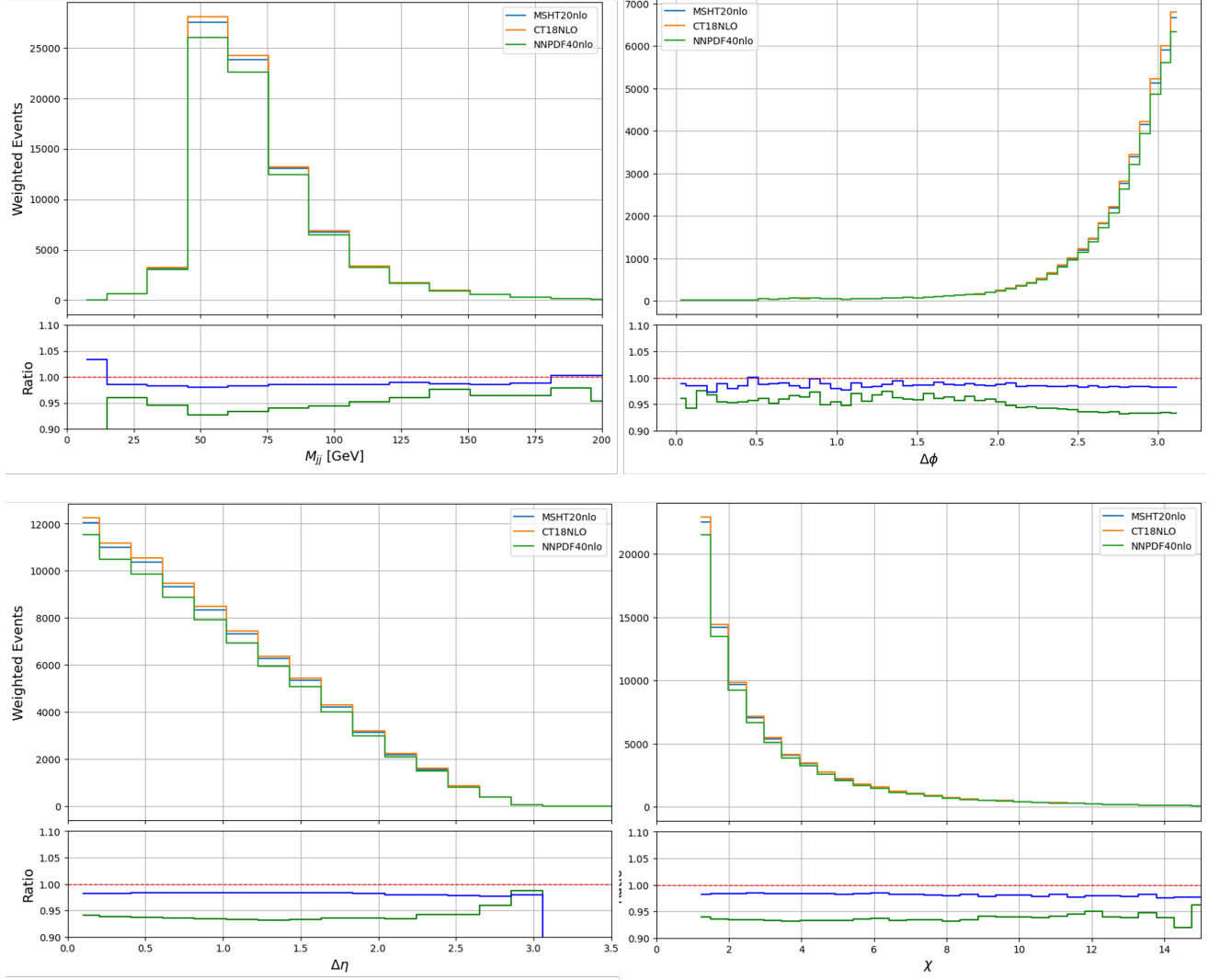


Figure 3: Four observables reweighting using NNPDF, MSHT, and CT18 [4] [5] [6]

3.2 Dijet Angular Distributions

After the Standard Model dijet generation, I then altered my Pythia code to generate dijets in accord with the ADD model via the Large Extra Dimensions (LED) Pythia settings. I generated events for $\Lambda = 15$ TeV, $\Lambda = 10$ TeV, and $\Lambda = 8$ TeV to observe the BSM effects of varying energy scales. I have plotted the χ_{dijet} in bins of invariant dijet mass, M_{jj} , since the behavior differs in lower vs higher mass ranges. The resulting plots can be seen in Figure 4. The 40-100 GeV and 100-200 GeV mass bins show consistent agreement, but increases in variation from the Standard Model prediction appear in the lower and higher mass ranges. From the theory, it is expected that the non-resonant shift will be more apparent from the spin 2 graviton at higher energies, so this trend matches the expectation. A similar study regarding dijet distributions was conducted at CMS [2], but at significantly higher mass. This demonstrates LHCb's ability to have lower mass range sensitivity.

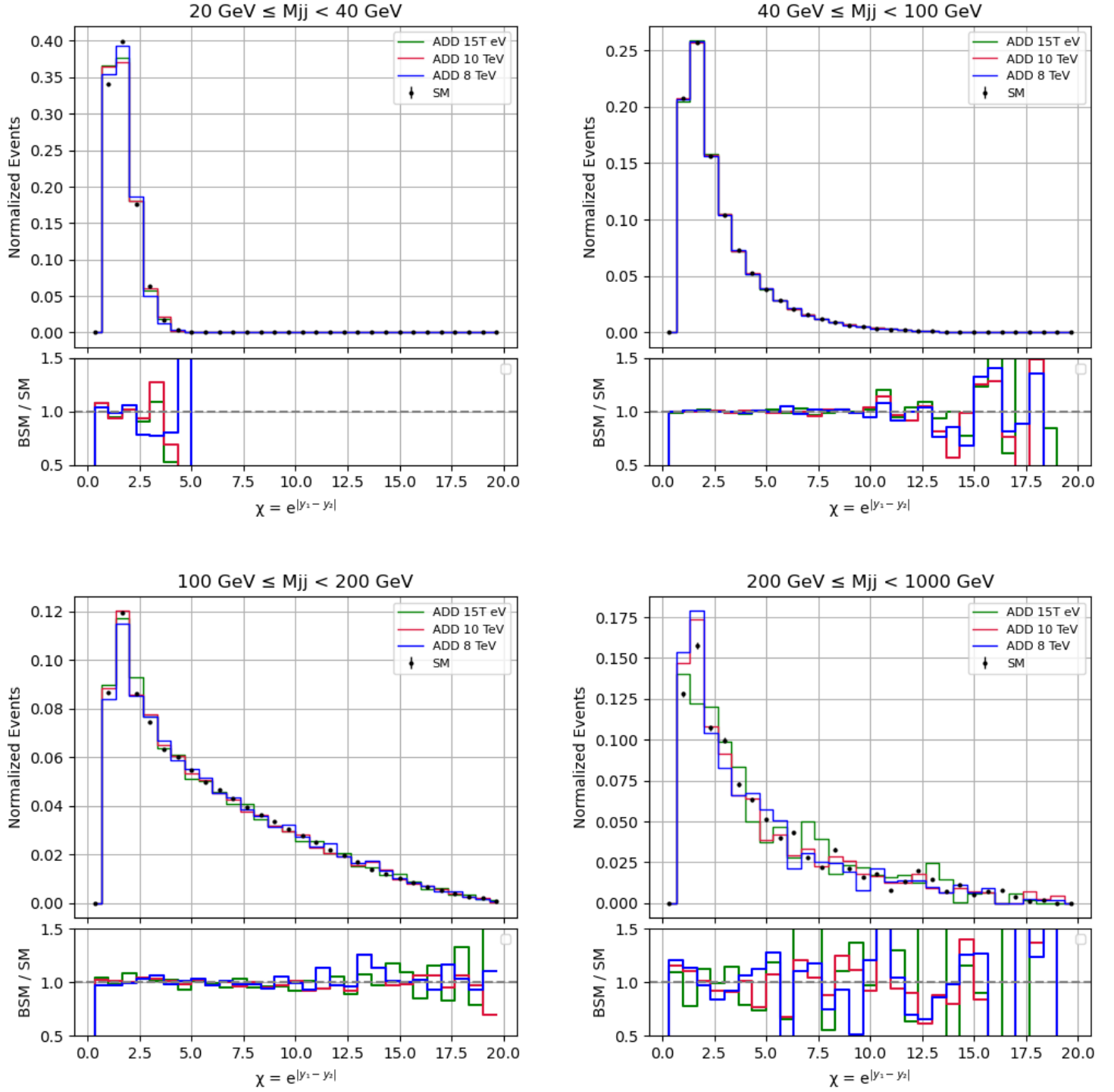


Figure 4: Plots comparing the ADD model at 15, 10, and 8 TeV Lambda energy scales with the Standard Model. The higher mass bin begins to show the most deviation from the SM black data points, which is to be expected from the ADD Model predictions. Of note as well is the apparent sensitivity in the lowest mass bin of 20-40 GeV.

Following the inclusive M_{jj} binning, I decided to examine in more detail the types of jets produced. Jets can be classified as being quark or gluon initiated based upon their constituents and features. To determine if one dijet type is more sensitive to BSM effects than another, I split inclusive binning into gg, qq, and qg dijet sources. The resulting plots are shown in Figures 5, 6, and 7.

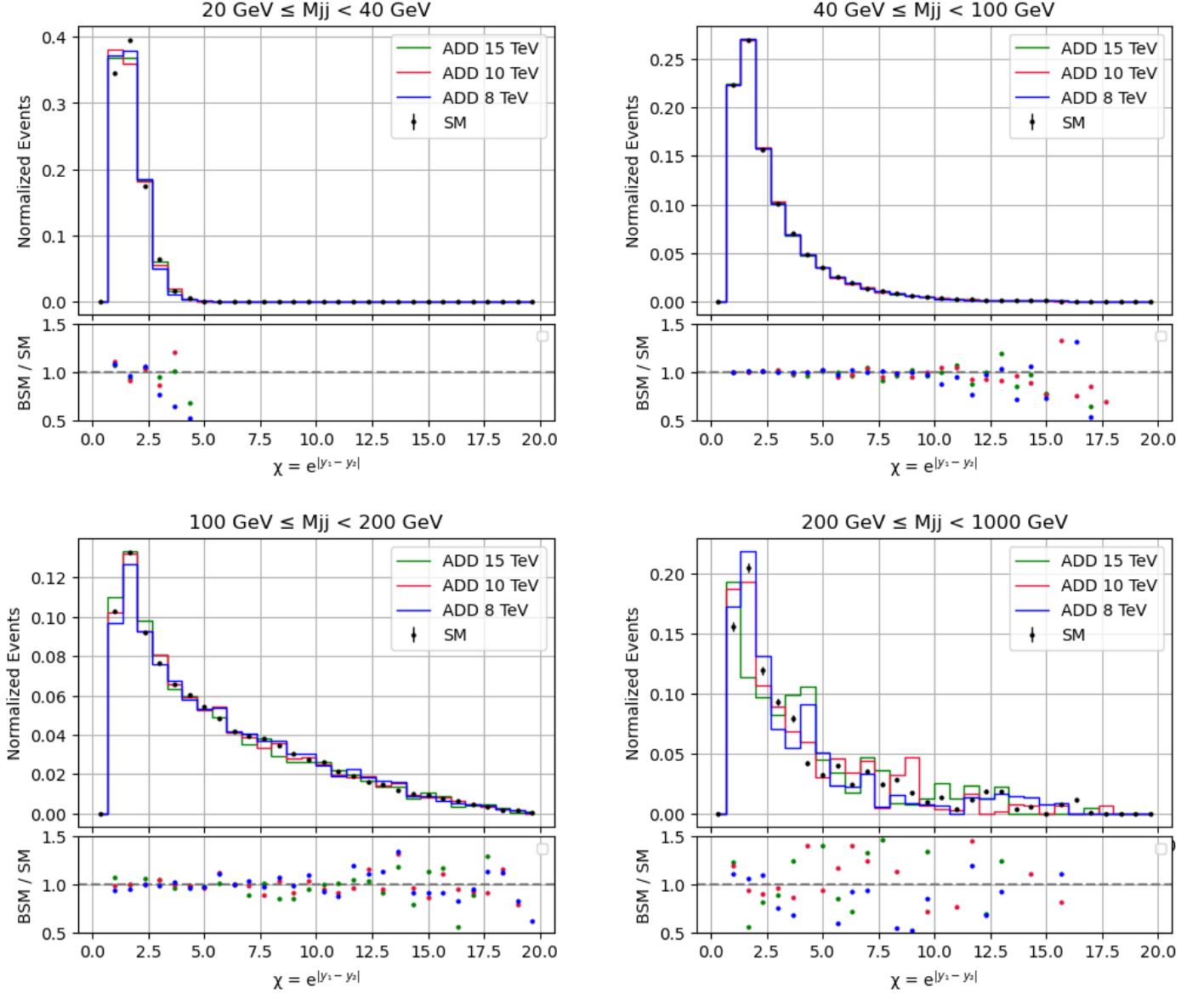


Figure 5: Angular distributions for the gluon-gluon (gg) initiated dijets. These distributions most closely resemble the inclusive plots as gg-initiated dijets comprise around 60% of the produced dijets.

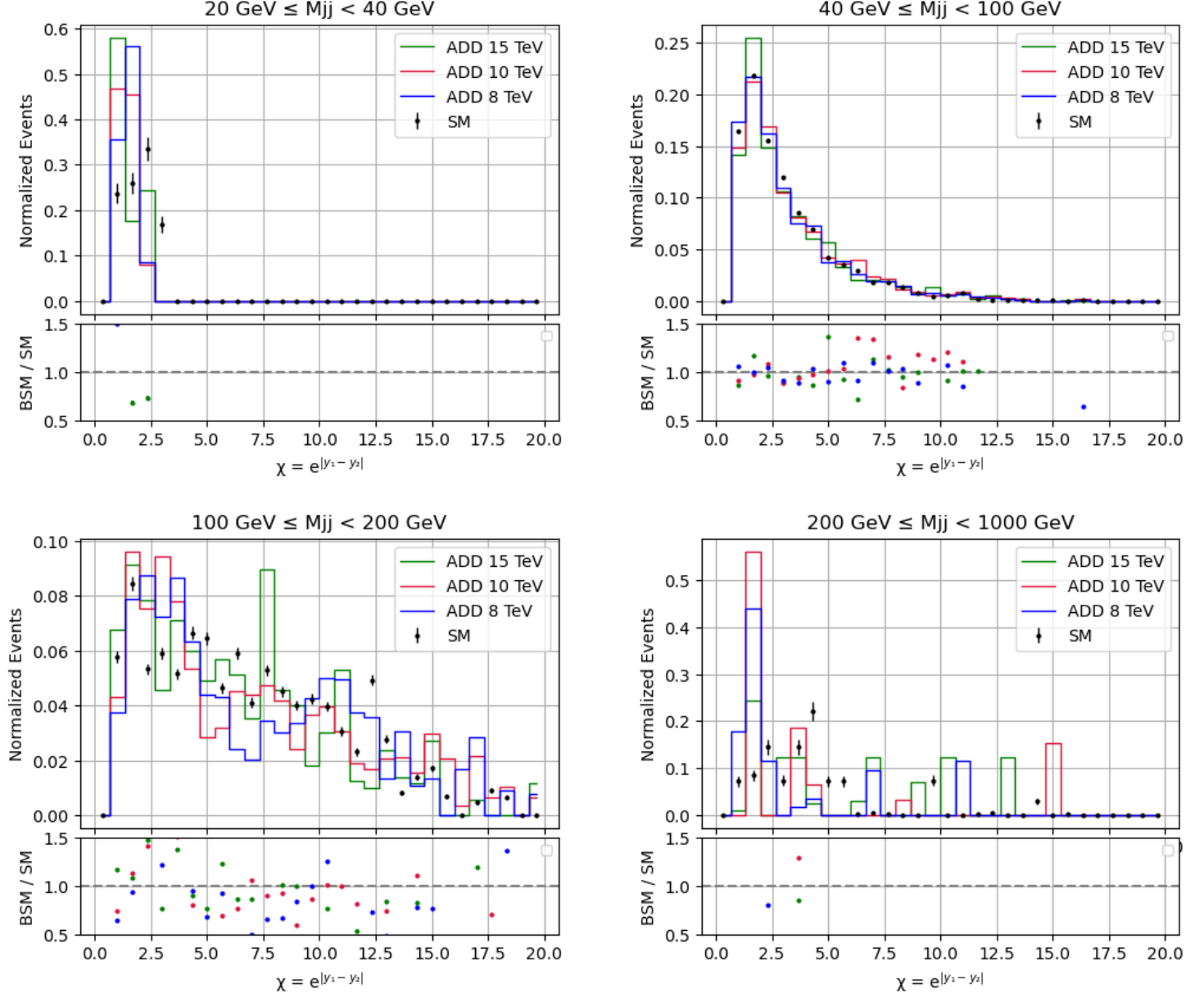


Figure 6: Angular distributions for the quark-quark (qq) initiated dijets. These dijets comprise only around 3% of produced dijets, so the higher mass range has less statistics. Nonetheless, we can observe some sensitivity between the ADD and SM models.

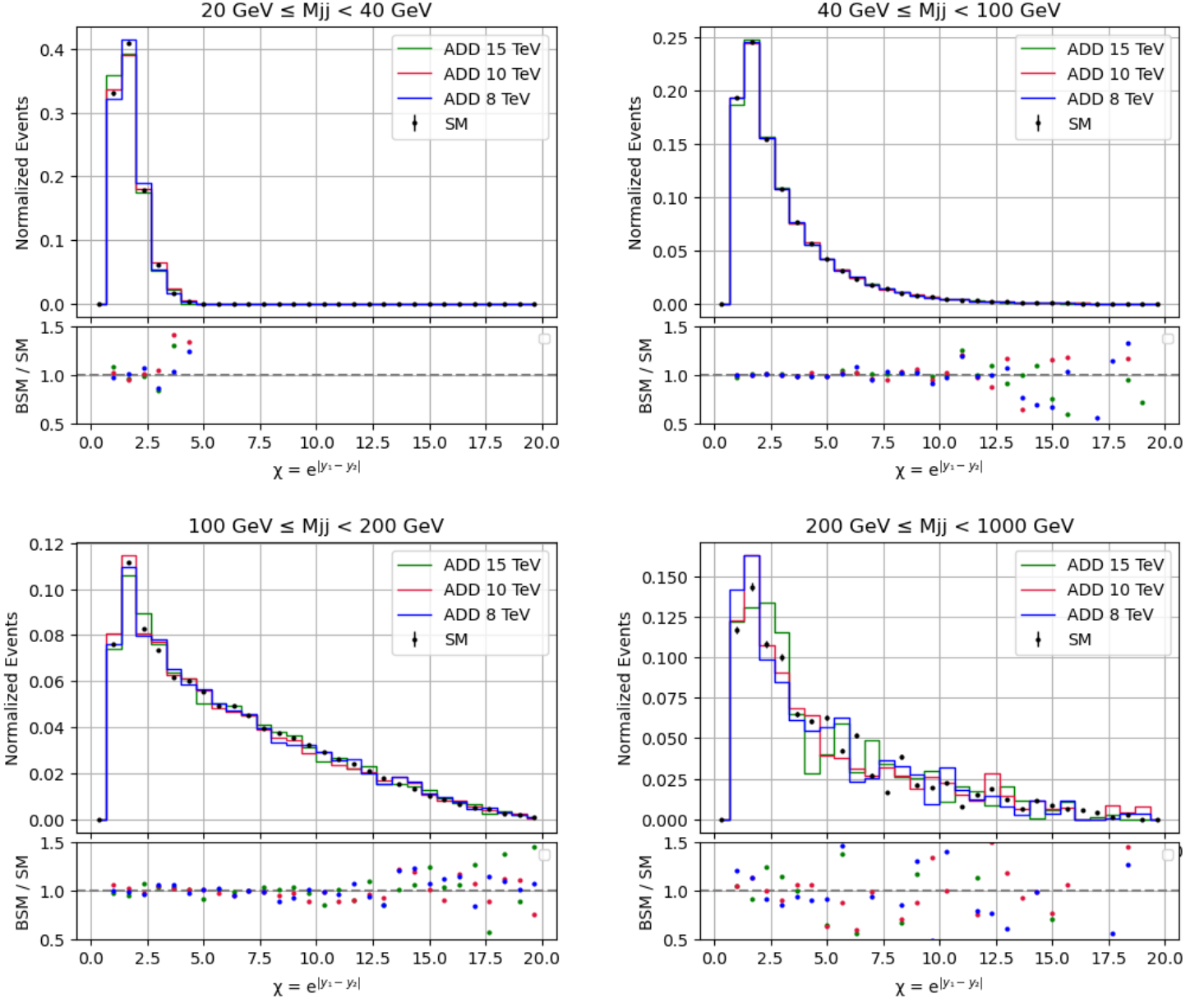


Figure 7: Angular distributions for the quark-gluon (qg) initiated dijets. These dijets comprise approximately 25% of produced dijets.

3.3 QCD Verification in Split Processes

Having simulated millions of events, I now look to take advantage of my python scripts which separate dijets by their processes. This can be extremely useful in a jet classification application. Quantum Chromodynamics (QCD) predicts certain behaviors and interactions between quarks and gluons, and I can examine these within the context of the Standard Model. I have therefore plotted the Standard Model χ_{dijet} distribution as split into the four most common QCD process types (Figure 8). While the shape is still the general decay profile, the gluon-gluon to gluon-gluon ($gg \rightarrow gg$) process has a sharper descent, particularly compared with the distribution for quark-quark to quark-quark ($qq \rightarrow qq$) process.

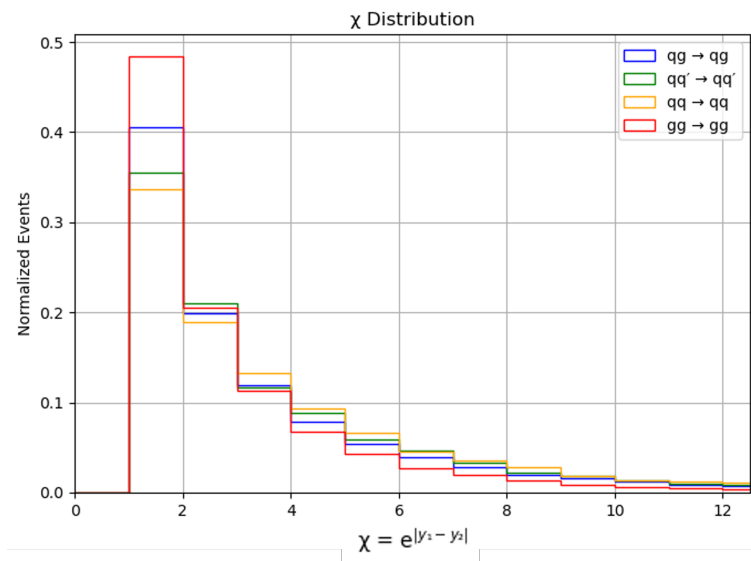


Figure 8: Distribution of χ_{dijet} split into the primary QCD processes.

In this new context, however, there are several other variables in addition to the χ_{dijet} which can be interesting to study as well. These are shown in Figure 9. A variable of note here is the Jet Girth, which is written as

$$g = \frac{1}{p_T^{jet}} \sum_i p_T^i \Delta R_{i,jet} \quad \text{where} \quad \Delta R_{i,jet} = \sqrt{(\Delta y_{i,jet})^2 + (\Delta \phi_{i,jet})^2}.$$

This variable is a function of the transverse momentum, p_T , the difference in rapidity, Δy , the difference in azimuthal angle, $\Delta \phi$, and the radius of the jet. [7] This combination makes it differ more strongly for gluon-gluon jets, which the plot in the upper right of Figure 9 reflects. I therefore use it in the subsequent Boosted Decision Tree as a strong feature for classification. The $\Delta \phi$ plot also shows significant difference between the $gg \rightarrow gg$ and $qq \rightarrow qq$ distributions. Being able to clearly distinguish these processes is greatly advantage for QCD and jet classification analyses.

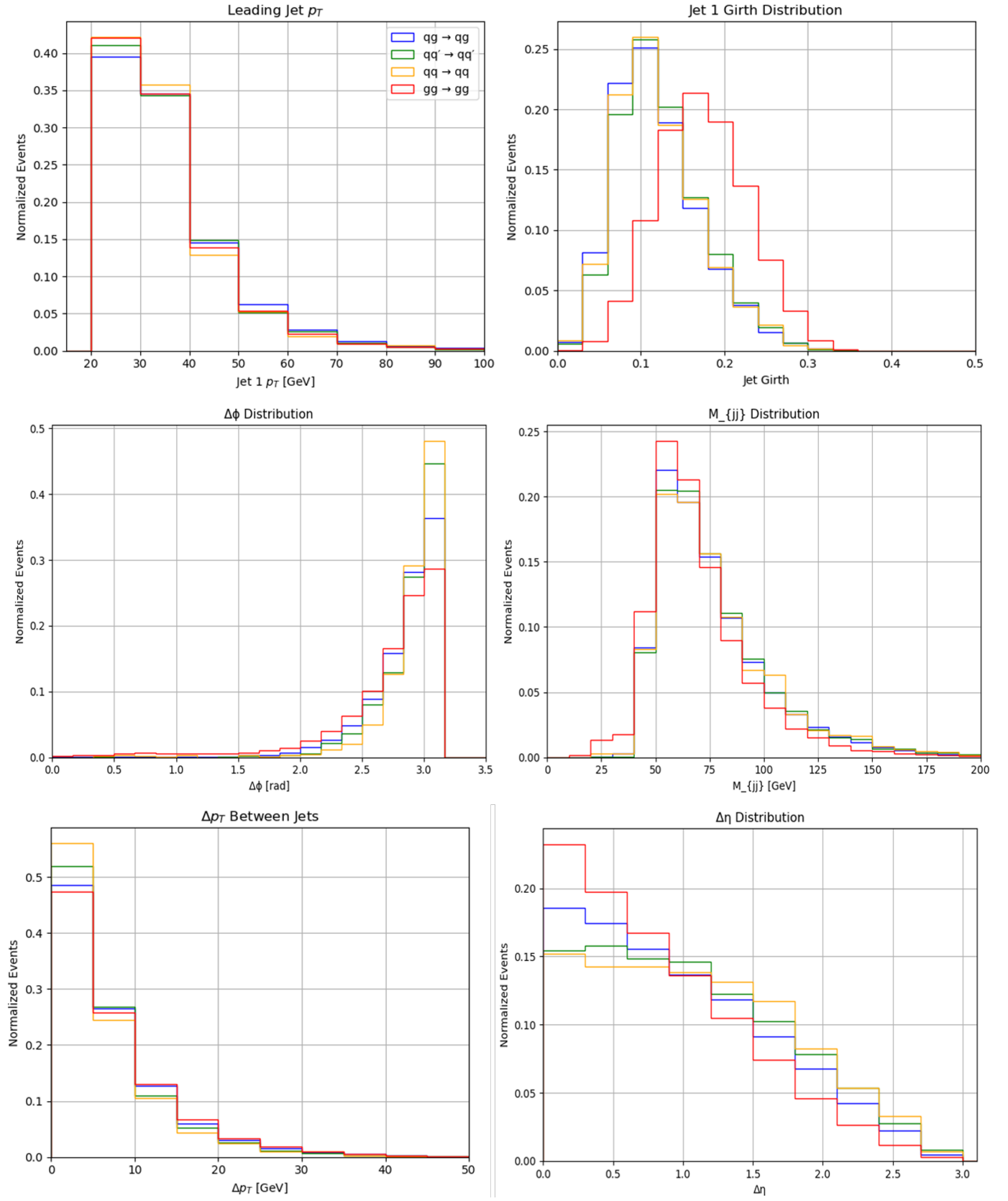


Figure 9: Distributions of dijet measurables split into the primary QCD processes.

3.4 Boosted Decision Tree and Jet Classification

Having seen how distributions change based off of jet type, it would be useful to be able to classify unknown jets as quark or gluon initiated. To do this, I have shifted into using machine learning with a Boosted Decision Tree. First developed by Breiman [8], a decision tree is a cut-based classification approach where a series of binary splits are performed in order to separate signal from background. Each binary decision point is called a node, after which the tree branches into two daughter branches. The process continues until a pre-determined stopping condition is reached. A Boosted Decision Tree (BDT) is a machine learning technique which combines many smaller decision trees, known as weak classifiers, in order to create a strong classifier. It feeds misclassified information and correctly classified information back into the algorithm in order to greatly improve accuracy [9].

To train a BDT, you separate a dataset into training and testing data. The test data is kept separate and hidden from the algorithm, while the training data (with known signal and background) are supplied to the BDT. Each node is a feature which you also supply to the BDT from which it will make classifying decisions. Once the training is complete, the testing dataset is inputted into the algorithm, which then returns a weighted classification score to evaluate how accurately the BDT performed. To improve the BDT's score, one can change the input features, increase the training dataset, alter the number of classifiers, or try changing other parameters. Because of this process, BDTs fall under the category of supervised machine learning since it uses pre-labeled datasets in the training process.

For my purposes, I have trained a BDT to classify quark or gluon jets. These two categories represent a binary split, so a BDT was a viable option for this task. I used the following features for training: Jet p_T , η , mass, number of constituents, number of charged constituents, jet girth, and constituent ID. This last feature is unique to LHCb's detector setup in being able to identify jet constituent information. In addition, I have taken the top 10 p_T jets and extracted their three vector momentum, energy, and ID, and fed these all as features into my BDT. Using a learning rate of 0.05, 2000 estimators, and a maximum tree depth of 8, I have successfully trained my BDT to distinguish between quark and gluon jets using generator level data. The score distribution, ROC curve, and feature importances are shown in Figures 10-12. I have used XGBoost and SciKitLearn for the training and evaluation of this model.[10] [11]

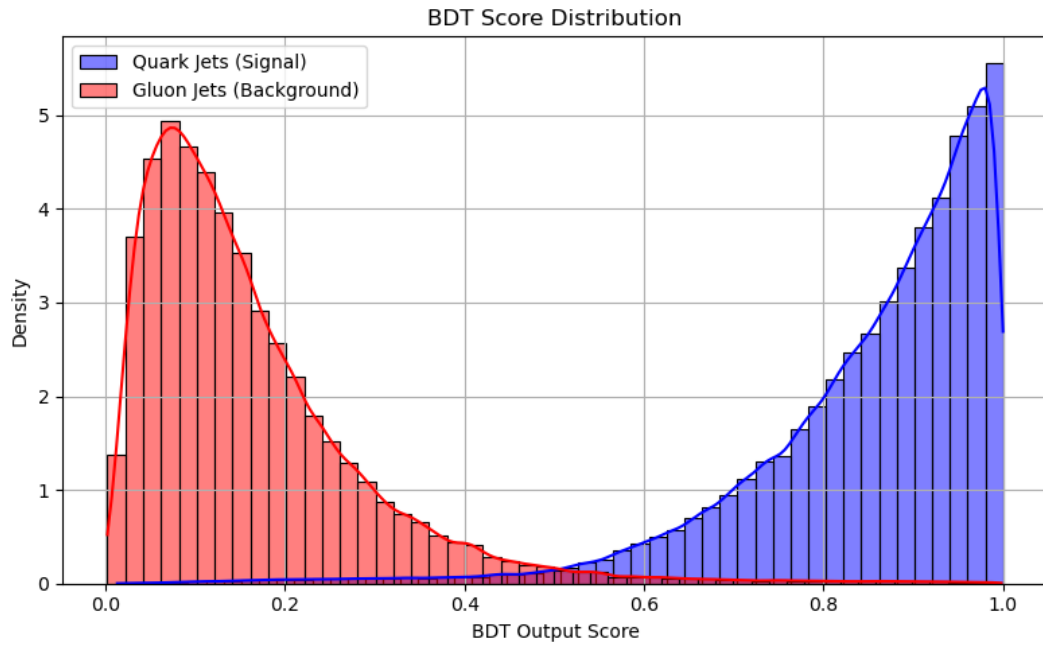


Figure 10: Score distribution for the quark and gluon jet classification.

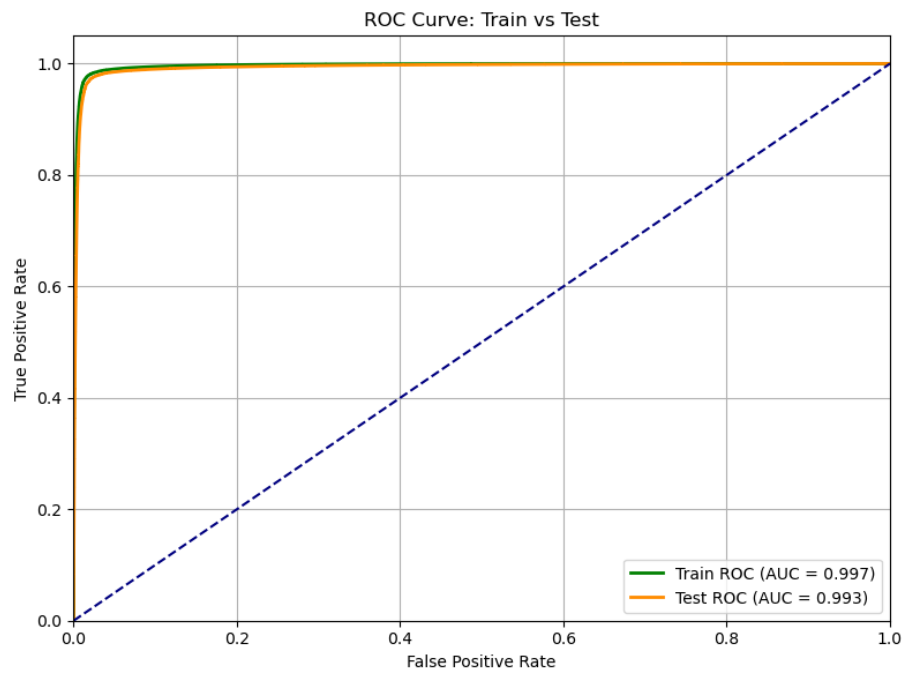


Figure 11: ROC Curve for Training and Testing performance. The testing curve is just under the training curve, confirming the BDT is not over training.

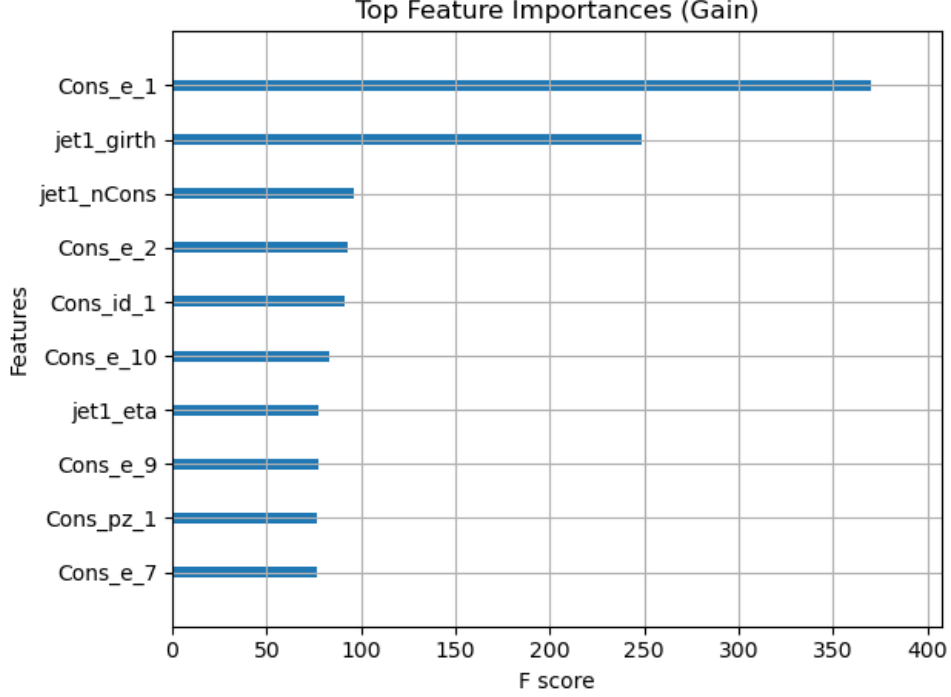


Figure 12: Top 10 features in the training algorithm. The constituent energy of the highest p_T jet is the most contributing feature, followed by the previously discussed jet girth.

4 Conclusion

In this analysis, I have demonstrated that the LHCb has sensitivity to variations within Parton Distribution Functions. Furthermore, I have shown LHCb is potentially sensitive to Beyond Standard Model effects, as tested with the ADD model. Future analysis for this topic could extend this study to a larger variety of new physics models and continue exploring additional energy scales. This research also illustrates that QCD behavior can be modeled and studied by parton type at the LHCb through jet-flavor quark/gluon classification. I have successfully trained and tested a Boosted Decision Tree to separate gluon-initiated from quark-initiated jets. The next step for the classification will involve retraining the classifier on a full detector simulation to take into account detector and reconstruction effects. This jet-flavor classification can allow for more precision measurements of QCD within the standard model and potentially reduce background in searches for new physics. Overall, the LHCb and its unique forward region of detection allow for never before tested jet classifications and continue to offer sensitivity for Parton Distribution Functions. This research could be adapted into further BSM searches, and additionally can be used for the Run 3 data of the LHC which is currently being collected.

References

- [1] UA1 Collaboration. Angular distributions for high mass jet pairs and a limit on the energy scale of compositeness for quarks from the cern $\bar{p}p$ collider. *Phys. Lett. B*, 177:244–250, 1986.
- [2] CMS Collaboration. Search for new physics in dijet angular distributions using proton-proton collisions at 13 tev and constraints on dark matter and other models. *JHEP*, 07:013, 2017.
- [3] Nima Arkani-Hamed, Savas Dimopoulos, and Gia Dvali. The hierarchy problem and new dimensions at a millimeter. *Phys. Lett. B*, 429:263–272, 1998.
- [4] Richard D. Ball, Stefano Carrazza, Juan Cruz-Martinez, Luigi Del Debbio, Stefano Forte, Tommaso Giani, Shayan Iranipour, Zahari Kassabov, Jose I. Latorre, Emanuele R. Nocera, Rosalyn L. Pearson, Juan Rojo, Roy Stegeman, Christopher Schwan, Maria Ubiali, Cameron Voisey, and Michael Wilson. The path to proton structure at 1 *The European Physical Journal C*, 82(5), May 2022.
- [5] S. Bailey, T. Cridge, L. A. Harland-Lang, A. D. Martin, and R. S. Thorne. Parton distributions from lhc, hera, tevatron and fixed target data: Msht20 pdfs. *The European Physical Journal C*, 81(4), April 2021.
- [6] Tie-Jiun Hou, Jun Gao, T.J. Hobbs, Keping Xie, Sayipjamal Dulat, Marco Guzzi, Joey Huston, Pavel Nadolsky, Jon Pumplin, Carl Schmidt, Ibrahim Sitiwaldi, Daniel Stump, and C.-P. Yuan. New cteq global analysis of quantum chromodynamics with high-precision data from the lhc. *Physical Review D*, 103(1), January 2021.
- [7] CMS Collaboration. Girth and groomed radius of jets recoiling against isolated photons in lead-lead and proton-proton collisions at $\sqrt{s} = 5.02$ tev. *Physics Letters B*, 861:139088, February 2025.
- [8] Leo Breiman, Jerome Friedman, R.A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.
- [9] Paolo Calafiura, David Rousseau, and Kazuhiro Terao. *Artificial Intelligence for High Energy Physics*. World Scientific Publishing Company, 2022.
- [10] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, USA, 2016. ACM.
- [11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.