

UNIVERSITY OF
COPENHAGEN



A PHYSICIST'S APPROACH TO MACHINE LEARNING
Understanding The Basic Bricks

CHRISTIAN MICHELSSEN
Master's Thesis (Cand. Scient.)
January 3rd, 2020

Supervised by
Troels Petersen



UNIVERSITY OF COPENHAGEN

Copyright © 2020
Christian Michelsen

[HTTPS://GITHUB.COM/CHRISTIANMICHelsen](https://github.com/CHRISTIANMICHelsen)

This thesis was inspired by the works of Edward R. Tufte and is based on the Tufte-L^AT_EX package.

First printing, January 2020

Abstract

Here will be a decent abstract at some pointTM.

(This page is intentionally left blank).

Contents

<i>Abstract</i>	iii
<i>Table of Contents</i>	v
<i>Foreword</i>	xi
1 <i>Introduction</i>	1
<i>Part I</i>	3
2 <i>Machine Learning Theory</i>	5
2.1 <i>Statistical Learning Theory</i>	5
2.2 <i>Supervised Learning</i>	6
2.3 <i>Generalization Bound</i>	7
2.3.1 <i>Generalization Bound for Infinite Hypotheses</i>	9
2.4 <i>Avoiding overfitting</i>	10
2.4.1 <i>Model Regularization</i>	10
2.4.2 <i>Cross Validation</i>	12
2.4.3 <i>Early Stopping</i>	13
2.5 <i>Loss functions</i>	14
2.5.1 <i>Evaluation Function</i>	16
2.6 <i>Decision Trees</i>	16
2.6.1 <i>Ensembles of Decision Trees</i>	17
2.7 <i>Hyperparameter Optimization</i>	19
2.7.1 <i>Grid Search</i>	20
2.7.2 <i>Random Search</i>	20
2.7.3 <i>Bayesian Optimization</i>	21
2.8 <i>Feature Importance</i>	22

3	<i>Danish Housing Prices</i>	27
3.1	<i>Data Preparation and Exploratory Data Analysis</i>	28
3.1.1	<i>Correlations</i>	30
3.1.2	<i>Validity of Input Variables</i>	31
3.1.3	<i>Cuts</i>	33
3.2	<i>Feature Augmentation</i>	33
3.2.1	<i>Time-Dependent Price Index</i>	34
3.3	<i>Evaluation Function</i>	35
3.4	<i>Initial Hyperparameter Optimization</i>	36
3.5	<i>Hyperparameter Optimization</i>	38
3.6	<i>Results</i>	40
3.7	<i>Model Inspection</i>	43
3.8	<i>Multiple Models</i>	45
3.9	<i>Discussion</i>	47
3.10	<i>Conclusion</i>	49
<i>Part II</i>		51
4	<i>Particle Physics and LEP</i>	53
4.1	<i>The Standard Model</i>	53
4.2	<i>Quark Hadronization</i>	54
4.3	<i>The ALEPH Detector and LEP</i>	56
4.4	<i>Jet Clustering</i>	58
4.5	<i>The Variables</i>	58
5	<i>Quark Gluon Analysis</i>	63
5.1	<i>Data Preprocessing</i>	63
5.2	<i>Exploratory Data Analysis</i>	64
5.2.1	<i>Dimensionality Reduction</i>	66
5.2.2	<i>Correlations</i>	67
5.3	<i>Loss and Evaluation Function</i>	67
5.4	<i>b-Tagging Analysis</i>	68
5.4.1	<i>b-Tagging Hyperparameter Optimization</i>	69
5.4.2	<i>b-Tagging Results</i>	70
5.4.3	<i>b-Tagging Model Inspection</i>	71
5.5	<i>b-Tagging Efficiency</i>	73
5.6	<i>g-Tagging Analysis</i>	75
5.6.1	<i>Permutation Invariance</i>	75
5.6.2	<i>Truncated Uniform PDF</i>	76

5.6.3	<i>g</i> -Tagging Hyperparameter Optimization	76
5.6.4	<i>PermNet</i>	78
5.6.5	1D Comparison of LGB and <i>PermNet</i>	78
5.6.6	<i>g</i> -Tagging Results	79
5.7	<i>g</i> -Tagging Efficiency	81
5.8	Generalized Angularities in 3-jet events	83
5.9	Gluon splitting	85
5.9.1	Variables	85
5.9.2	Efficiencies	87
5.9.3	Closure Test	88
5.9.4	Results	90
5.10	Discussion	93
5.11	Conclusion	94
A	<i>Housing Prices Appendix</i>	95
A.1	<i>Input Distributions</i>	95
A.2	<i>Correlations</i>	110
A.3	<i>Validity Correlations</i>	111
A.4	<i>Energy Rating Mapping</i>	112
A.5	<i>Facebook Prophet for Houses</i>	112
A.6	<i>Final Variable List</i>	113
A.7	<i>Initial Hyperparameter Results Tables</i>	115
A.8	<i>Initial Hyperparameter Results Parallel Coordinates</i> .	118
A.9	<i>Hyperparameter Results Parallel Coordinates</i>	118
A.10	<i>Hyperparameter Optimization as a Function of Iterations</i>	119
A.11	<i>Histogram of z Values for Apartments with the Tight Cut</i>	121
A.12	<i>Performance Metrics for the Tight Selection</i>	121
A.13	<i>Multiple Models for Houses</i>	121
B	<i>Quarks vs. Gluons Appendix</i>	123
B.1	<i>Fraction of Different Types of Jets</i>	123
B.2	<i>Random Search PDFs for XGB</i>	123
B.3	<i>Global SHAP Feature Importances for the g-Tagging Models in 3-Jet Events</i>	123
B.4	<i>UMAP Parameters Comparison</i>	124
B.5	<i>t-SNE Parameters Comparison</i>	124
B.6	<i>HPO results for 3-jet b-Tagging</i>	125
B.7	<i>HPO results for 4-jet b-Tagging</i>	125
B.8	<i>b-Tag Scores in 3-Jet Events</i>	126

<i>B.9 b-Tag Scores in 4-Jet Events</i>	126
<i>B.10 ROC curve for 3-jet b-tagging</i>	127
<i>B.11 ROC curve for 4-jet b-tagging</i>	127
<i>B.12 Distribution of b-Tags in 3-Jet Events</i>	128
<i>B.13 Distribution of b-Tags in 4-Jet Events</i>	128
<i>B.14 Global Feature Importances for the LGB b-Tagging Algorithm on 3-Jet Events</i>	129
<i>B.15 Global Feature Importances for the LGB b-Tagging Algorithm on 4-Jet Events</i>	129
<i>B.16 b-tag Efficiency for b-Jets in the b-Signal Region for 3-jet events</i>	130
<i>B.17 b-tag Efficiency for g-Jets in the g-Signal Region for 3-jet events</i>	130
<i>B.18 b-tag Efficiency for b-Jets in the g-Signal Region for 3-jet events</i>	131
<i>B.19 b-tag Efficiency for g-Jets in the b-Signal Region for 3-jet events</i>	131
<i>B.20 Parallel Coordinate Plot of HPO Results for 3-Jet g-Tagging for Energy Ordered Jets</i>	132
<i>B.21 Parallel Coordinate Plot of HPO Results for 3-Jet g-Tagging for Shuffled Jets</i>	132
<i>B.22 Parallel Coordinate Plot of HPO Results for 4-Jet g-Tagging for Energy Ordered Jets</i>	133
<i>B.23 Parallel Coordinate Plot of HPO Results for 4-Jet g-Tagging for Shuffled Jets</i>	133
<i>B.24 PermNet Architecture</i>	134
<i>B.25 1D LGB and PermNet Comparison for 4-Jet Events . .</i>	135
<i>B.26 1D LGB and PermNet Comparison for 3-Jet Events . .</i>	136
<i>B.27 g-Tag Scores in 4-Jet Events</i>	137
<i>B.28 g-Tag Scores in 3-Jet Events</i>	137
<i>B.29 ROC Curve for g-Tag in 4-Jet Events</i>	138
<i>B.30 ROC Curve for g-Tag in 3-Jet Events</i>	138
<i>B.31 Distribution of g-Tag Scores in 4-Jet Events for Signal and Background</i>	139
<i>B.32 Distribution of g-Tag Scores in 3-Jet Events for Signal and Background</i>	139
<i>B.33 Generalized Angularities for Charged Gluons Tracks .</i>	140
<i>B.34 Generalized Angularities for Neutral Gluons Clusters</i>	142
<i>B.35 g-Tagging Efficiency for 4-Jet Events in MC</i>	144

<i>B.36 Closure Plot for g-Tagging Model in 4-Jet Events</i>	146
<i>B.37 Gluon Splitting Distribution Comparison in MC and Data for Phase Space Region A</i>	149
<i>B.38 Gluon Splitting Distribution Comparison in MC and Data for Phase Space Region B</i>	150
<i>B.39 Gluon Splitting Distribution Comparison in MC and Data for Phase Space Region C</i>	151
<i>B.40 Gluon Splitting Distribution Comparison in MC and Data for Phase Space Region D</i>	152
<i>List of Figures</i>	159
<i>List of Tables</i>	162
<i>Bibliography</i>	163

(This page is intentionally left blank).

Part I

Part I of this thesis covers the introductory theory of machine learning in [chapter 2](#) along with some extra technical aspects of it. In [chapter 3](#) machine learning is applied to estimate Danish housing prices as precisely and accurately as possible.

(This page is intentionally left blank).

Part II

Part II of this thesis deals with particle physics and the discriminatory power of machine learning for quark-gluon identification and subsequent analysis. In [chapter 4](#) the theory of the Standard Model is introduced together with a description of the ALEPH detector. Theory is applied in [chapter 5](#) where the types of jets and events in each collision is analysed using machine learning to improve the understanding of how gluon jets hadronizes and splits: simply said how they look and behave.

(This page is intentionally left blank).

5. Quark Gluon Analysis

“Research is what I am doing I don’t know know what I’m doing.”

— Wernher von Braun

IN THIS CHAPTER, the analysis of quarks and gluons is presented. The overall goal with this analysis is to improve the knowledge of gluon jets by developing a method to increase the precision at which quark and gluon jets can be distinguished. The overall goal is to improve the understanding of gluon hadronization and splitting by measuring the gluon hadronization in 3-jet events and the gluon splitting in 4-jet events. The method includes two models; a *b*-tagging model which tag individual jets within an event, and a *g*-tagging models which classify entire events.

This chapter is organized as follows. In section [section 5.1](#), the applied datasets are presented and the initial selection criteria for the data are described. Visualizations of the most important variables for the quark and gluon jet events are shown in section [section 5.2](#). The choices of loss and evaluation function are discussed in [section 5.3](#). In [section 5.4](#), the developed *b*-tagging model is presented, and measures of the model’s efficiency are shown in section [section 5.5](#). Similarly, the *g*-tagging model is presented in section [section 5.6](#), and measures of the model’s efficiency are shown in [section 5.7](#). The jet hadronization in 3-jet events is analyzed in [section 5.8](#) and the gluon splitting in 4-jet events in [section 5.9](#). Finally the quark gluon project is discussed in [section 5.10](#) and concluded in [section 5.11](#).

5.1 Data Preprocessing

The data files were acquired from Prof. Peter Hansen (NBI) who worked on the ALEPH experiment. The Data¹ consists of 43 data files from between 1991 and 1995 totalling 3.5 GB (Data). In addition to the datasets measured by the ALEPH experiment, Monte Carlo (MC) simulations of events containing quark and gluon jets are applied. These additional simulations total 125 files (8.4 GB) and additional 42 MC-files with only *b*-quark events (MC_b) (2.1 GB). The data files which are in the form of *Ntuples*, ROOT’s data format [[29](#)], are converted to HDF5-files by using the Python package uproot [[7](#)]. While iterating over the Ntuples, some basic selection criteria are applied before exporting the data to HDF5-format.

¹ Remember that “Data” refers to actual measurements, whereas “data” refers to any data (not necessarily real-life data).

The first selection criteria is that the center of mass energy E of the events has to be close the mass of the Z boson, $90.8 \text{ GeV} \leq E \leq 91.6 \text{ GeV}$. The second selection criteria is used to remove $Z \rightarrow \tau^+ \tau^-$ events by requiring that the sum of the momenta in each event is greater or equal to 32 GeV . The third selection criteria is used to select events of high quality by requiring at least two good tracks in the events to ensure a primary vertex, where a good track is defined as having at least 7 TPC hits and 1 silicon hit or more. Finally it is required that the cosine of the thrust axis polar angle, which is the angle between the thrust axis and the beam, is less than or equal to 0.8 to avoid any low angle events since the detector performance worsens significantly in that region. These initial requirements were standard requirements for the ALEPH experiment.

An additional selection criteria was applied to further improve the quality of the applied data by defining a lower threshold value for *jet matching*. The jet matching is the process of matching the jet with one of the final state quarks. The jet is said to be matched if the dot product between the final quark momentum and the jet momentum is higher than the threshold value. A high jet matching threshold results in events with cleaner jets but at the expense of less statistics. A jet matching threshold of 0.90 was found to be a good compromise between purity and quantity where 97.8% of all 2-jet events are matched and 96.7% of all events with more than two jets were matched².

The structure of the data contained in Ntuples is quite different compared to the “normal” structure of tidy data [103]. The data is organized such that one iterates over each event where the variables are variable-length depending on the number of jets in the events; this is also known as *jagged arrays*. The data is un-jagged³ before exporting to HDF5-format and only the needed variables are kept. This reduces the total output file to a 2.9 GB HDF5-file including both Data, MC, and MCb. The number of events for each number of jets can be seen in Table 5.1.

5.2 Exploratory Data Analysis

The variables applied in the machine learning models are the three vertex variables `projet`, `bqvjet`, and `ptljet`, which were introduced in chapter 4, these variables are the primary focus of this section. For MC-simulated data, both the final state particles are simulated as well as the trajectories of the particles within the detector and the detector interactions. This means, that for MC-simulated data the detector measurements can be directly compared to the “true” particles. For the MC simulated data, each event is generated such that the type of collision, or *flavor*, is known and assigned the variable `flevt`. The mapping from flavor to `flevt` can be seen in Table 5.2.

Unfortunately the truth information about the individual jets in the event is not recorded in the MC files, only the information about which type of quarks were created in the collision and not which jets

² Compare this to 98.5 % and 97.8 % for a threshold of 0.85 or 95.9 % and 93.9 % for a threshold of 0.95.

³ Such that e.g. a 3-jet event will figure as three rows in the dataset.

<i>n</i> -jet	Data	MC
2	1 179 869	3 646 797
3	1 206 430	3 593 630
4	213 584	560 477
5	10 555	20 764
6	85	98
Total	2 610 523	7 821 766

Table 5.1: The number of *n*-jet events for Data and MC.

Flavor:	<i>bb</i>	<i>cc</i>	<i>ss</i>	<i>dd</i>	<i>uu</i>
<code>flevt</code> :	5	4	3	2	1

Table 5.2: Mapping between the flavor and the variable `flevt`.

corresponds to which quarks. We define that an event is *q-matched* if one, and only one, of the jets are assigned to one of the (final) quarks, if one, and only one, of the jets are assigned to the other (final) quark, and if no other jets are matched to any of the (final) quarks. We then define what constitutes a *b*-jet: if it has `flevt` = 5, the entire event is *q*-matched, and the jet is matched to one of the (final) quarks. Similarly, *c*-jets and *uds*-jets are defined using the same range of requirements as *b*-jets with the only change that `flevt` = 4 for a *c*-jet and `flevt` ∈ {1, 2, 3} for a *uds*-jet.

A gluon jet is defined to be a jet in a (any-flavor) *q*-matched event where the jet itself is not assigned to any of the (final) quarks. Strictly speaking, this means that the gluon jet is not 100 % certain of being a gluon. We cannot know this, as not all of the input parameters in the MC simulation are known, only the final clustered jets. Due to the *q*-match criterion this also means that some jets are assigned the label “non-*q*-matched” for which these jets are regarded as background. The distribution of different types of jets can be seen in Table 5.3 and shown as relative numbers in Table B.1.

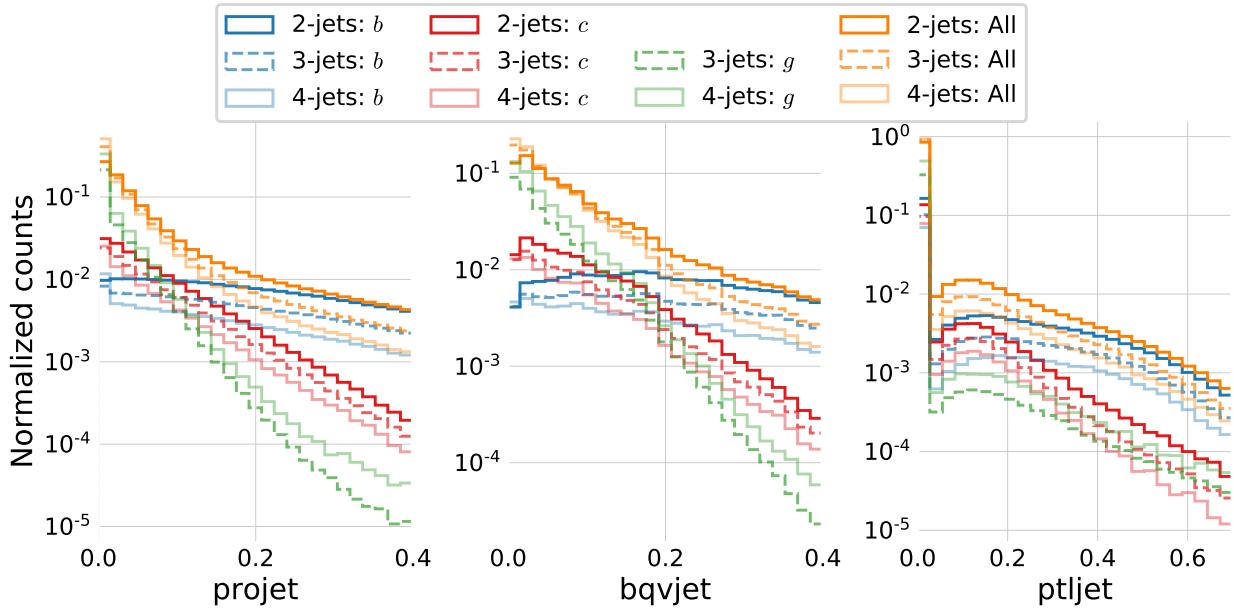
<i>n</i>	<i>b</i>	<i>c</i>	<i>uds/l</i>	<i>g</i>	non- <i>q</i> -matched
2	2 713 454	944 380	2 125 900	0	1 509 860
3	2 433 878	964 212	2 129 218	3 365 969	1 887 613
4	326 264	156 332	336 548	1 012 198	410 566
5	10 332	5960	12 668	54 525	20 335
6	42	26	52	320	148
Total	5 483 970	2 070 910	4 433 012	4 604 386	3 828 522

Table 5.3: Number of different types of jets for MC and MCb for *n*-jet events. See also Table B.1 for relative numbers.

Applying the criteria defined above for what constitutes a specific type of jet, the 1D-distributions for the three vertex variables are plotted in Figure 5.1. The left subplot shows that high values of `projet` tend to indicate *b*-jets. In the middle subplot `bqvjet` is plotted which shares many similarities with the `projet` variables, including that high values indicate *b*-jets. In the right subplot the `ptljet` is plotted. This variable has many zeros in it which correlates mostly with gluons⁴ and large values are mostly due to *b*-jets.

Furthermore, Figure 5.1 shows that there are minor differences between the variables when comparing 2-, 3-, and 4-jet events, except for 2-jet events which do not contain any gluons at all.

⁴ Around 98 % of all *g*-jets are zeros for the variable `ptljet` compared to ∼ 82 % for *c*-jets and ∼ 70 % for *b*-jets.



5.2.1 Dimensionality Reduction

Even though only three vertex variables are applied, it is difficult to evolve a good understanding of how well the variables can separate different kinds of jets. The dataset contains millions of events, so a simple 3D scatter plot quickly becomes overcrowded and hard to interpret. In order to improve the visualization of the applied data, dimensionality reduction is used to reduce the three dimensions down to two dimensions.

Within recent years the field of dimensionality reduction algorithms has grown a lot from just the typical (linear) principal component analysis to also include nonlinear algorithms. The t-SNE algorithm [97] deserves an honorable mention since this algorithm revolutionized the usage of (nonlinear) dimensionality reduction algorithms in e.g. bioinformatics [94, 101] yet its mathematical foundation has strongly been improved with the newer, faster UMAP algorithm [70] which usage is also expanding [21, 22, 42].

The aim of UMAP, short for Uniform Manifold Approximation and Projection, is to correctly identify and preserve the structure, or topology, of the high-dimensional feature space in a lower-dimensional output space. In order to optimally preserve both the global and local structure of the data, the algorithm stitch together local manifolds in the high-dimensional feature space such that the difference between the high- and low-dimensional representations is minimized according to the cross-entropy [70]. Compared to t-SNE the approach in UMAP has a topological background compared to the more heuristic approach taken by t-SNE.

The UMAP algorithm has several hyperparameters, where two of the most important ones are the number of neighbors `n_neighbors` which controls the priority between correctly preserving the global versus the local structure, and the `min_dist` which defines how

Figure 5.1: Normalized histograms of the three vertex variables: `projet`, `bqvjet`, and `ptljet`. In blue colors the variables are shown for **true b-jets**, in red for **true c-jets**, in green for **true g-jets**, and in orange for **all of the jets** (including non q-matched). In fully opaque color are shown the distributions for 2-jet events, in dashed (and lighter color) 3-jet events, and in semi-transparent 4-jet events. Notice the logarithmic y-axis, that there are no g-jets for 2-jet events (as expected), and that all of the distributions are very similar no matter how many jets.

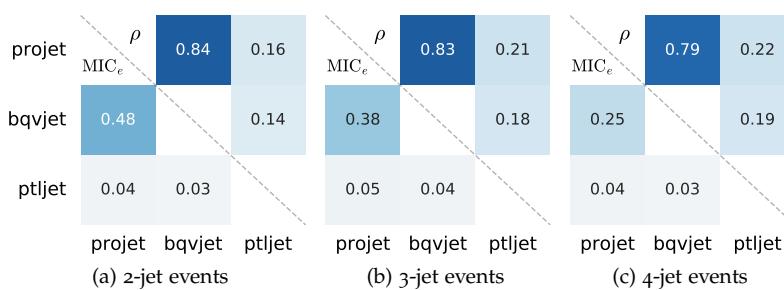
tightly together UMAP is allowed to cluster the points in the low-dimensional representation. In order to determine the best combination of `n_neighbors` and `min_dist` a grid search with `n_neighbors` $\in \{10, 50, 100, 250\}$ and `min_dist` $\in \{0, 0.2, 0.5\}$ is performed. This is shown for 4-jet events in Figure B.1.

The best combination of `n_neighbors` and `min_dist` is subjective at best, but I judged that `n_neighbors` = 250 and `min_dist` = 0.2 gave the best compromise between preserving local and global structure. The results of running UMAP on 4-jet events can be seen in Figure 5.2. Note that the UMAP algorithm is not provided any information about which jets are which types or any other truth information, that information is only provided afterwards to the actual plot. Here the millions of points are plotted using Datashader [8] to avoid overplotting and colored according to the jet type. The figure shows that there are some clear b -jet clusters, however, most of the data seem to be a mix of g -and uds -jets.

Similarly, Figure 5.3 show the two-dimensional visualization for 3-jet events correspondingly. These three figures suggest that it should be possible to discriminate the b -jets from the other jets somewhat, however, no clear separation is expected. The t-SNE algorithm was also tested but showed inferior performance compared to UMAP, see Figure B.2 for an example of this.

5.2.2 Correlations

The correlations between the vertex variables are shown in Figure 5.4, where the upper diagonal shows the linear correlation ρ and the lower diagonal shows the nonlinear correlation MIC_e introduced in subsection 3.1.1. It is found that `projet` and `bqvjet` have a relatively high correlation whereas the other variable combinations correlate much less. Had they all correlated a lot, it would be more difficult to extract any meaningful insights from the system as it would contain less information.



5.3 Loss and Evaluation Function

Contrary to the housing prices project where each house has a price on a continuous scale, the goal of the quark and gluon jet project is to predict types of jets or events where the *signal* observations⁵ are often assigned the label 1 and *background* observations 0. The combination

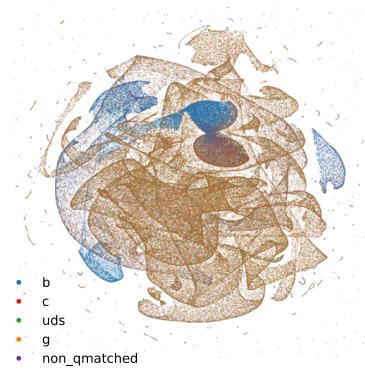


Figure 5.2: Visualization of the vertex variables in 4-jet events for the different categories: **true b -jets** in blue, **true c -jets** in red, **true uds -jets** in green, **true g -jets** in orange, and **non q -matched** events in purple. The clustering is performed with the UMAP algorithm which outputs a 2D-projection. This projection is then visualized using the Datashader which takes care of point size, avoids over and underplotting, and color intensity.

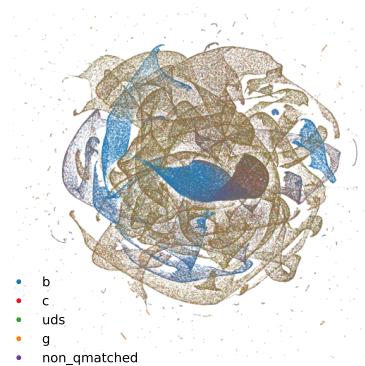


Figure 5.3: UMAP visualization of the vertex variables for 3-jet events.

Figure 5.4: Correlations between the three vertex variables for 2-, 3- and 4-jet events. The upper diagonal shows the linear correlation ρ and the lower diagonal the nonlinear correlations MIC_e .

⁵ Often called signal events, however, since a jet can also be signal, the term signal event is only used when the meaning is clear from the context.

of this being a *classification* problem (compared to a regression problem). Furthermore, all the variables applied in the quark and gluon jet analysis are actual measurements from a particle detector which means that the issue of outliers is negligible. This also means that the problem of finding a robust loss function is less important since for classification models the loss is already bounded in the $[0, 1]$ -interval.

Accuracy, which is simply the fraction of correct predictions, is often used as the loss function in classification, however, accuracy as a metric suffers a lot when handling *imbalanced* data: when the ratio between the number of instances of each class is not approximately fifty-fifty. The problem is that if the sample contains 90 % background and only 10 % signal, then a simple model which simply predicts everything to be background will have a 90 % accuracy.

To circumvent this issue, the area under the ROC⁶ curve (*AUC*) is used, where the ROC curve is the the *signal efficiency* ε_{sig} of the ML model plotted as a function of the *background efficiency* ε_{bkg} . The definition of these two measures are:

$$\varepsilon_{\text{sig}} = \frac{S_{\text{sel}}}{S_{\text{tot}}}, \quad \varepsilon_{\text{bkg}} = \frac{B_{\text{sel}}}{B_{\text{tot}}}, \quad (5.1)$$

where S_{sel} are signal events that were also selected (predicted) as signal by the ML model, $S_{\text{tot}} = S_{\text{sel}} + S_{\text{rej}}$ is the total number of signal events (the selected and rejected), and likewise for background events B . Within the machine learning community the signal efficiency is called the true positive rate (TPR) and the background efficiency the false positive rate (FPR). For the rest of this project, the AUC will be the evaluation function $f_{\text{eval}} = \text{AUC}$, however, since this metric does not work on single observations it cannot be used as the loss function. Instead we will use the *log-loss* as the loss function⁷ which in comparison to the AUC is not only differentiable for single predictions but also takes the certainty of the prediction into account. When using tree-based algorithms or neural networks one can compute not only whether or not a single observation is classified as signal or background but also a prediction score. This is a number in the $[0, 1]$ -interval and the closer to 1 the score is, the more certain the model is of the prediction being signal. Given the prediction score \hat{y} and the true label y , the log-loss ℓ_{\log} is calculated as:

$$\ell_{\log}(\hat{y}, y) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}). \quad (5.2)$$

This is visualized in Figure 5.5. Here it can be seen how the loss changes as a function of the prediction score. Notice that when $y = 0$ the loss for $\hat{y} = 1$ diverges towards ∞ and likewise with $y = 1$ and $\hat{y} = 0$ (since $\log 0$ diverges to $-\infty$).

5.4 *b*-Tagging Analysis

The ability to discriminate between the different types of particles produced in a collision is obviously important in order to correctly interpret the measurements. A lot of research focus on tagging algorithms today, e.g. *b*-tagging in ATLAS and CMS [85]. The reason

⁶ Receiver Operating Characteristic.

Strictly speaking it is not a function of the background efficiency, but rather ε_{sig} and ε_{bkg} plotted parametrically as functions of the threshold cut \hat{y}_{cut} .

⁷ In the context of machine learning this is the same as the *cross entropy*.

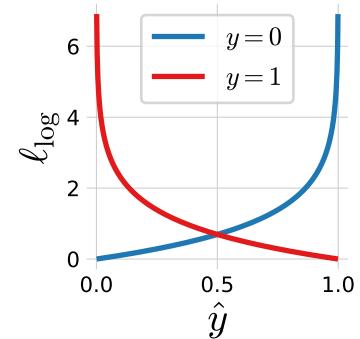


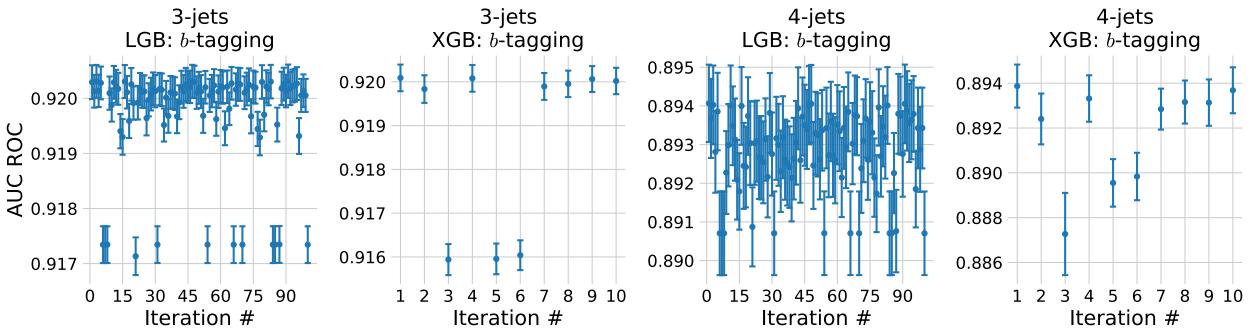
Figure 5.5: Plot of the log-loss ℓ_{\log} for background ($y = 0$) in blue and signal ($y = 1$) in red.

that b -quarks are tagged specifically is both due to b -quarks having more unique characteristics compared to e.g. c -quarks and are thus easier to tag, but also the fact that b -quarks are the second-heaviest of the quarks and are measured to better understand CP-violation⁸ at LHC-b, contributes to the choice of tagging b -quarks. In ALEPH Proriol et al. [77] started the work of comparing different methods for b -tagging already in 1991. They concluded that a neural network had the best performance compared to e.g. a linear (Fisher) discriminant. The neural network used was a 3-layer neural network (NN) trained on nine variables for which the output was the predicted score `nnbjet`. From here on, this pre-trained network will be called NNB. The events are split in a (80-20)% train-test ratio in such a way that the individual jets in an event are not split.

5.4.1 b -Tagging Hyperparameter Optimization

Compared to the housing prices dataset, the number of observations N has significantly increased ($3 \times 10^7 \gg 5 \times 10^5$), but the dimensionality M is much smaller ($3 \ll 143$). Therefore both XGBoost (XGB) and LightGBM (LGB) were included as models initially since their performances in the housing dataset were very similar. LightGBM was expected to be faster on this dataset due to the large N . The models were hyperparameter optimized (HPO) using random search (RS) since the Bayesian optimization (BO) did not show any performance gains compared to RS in the housing project.

They were both hyperparameter optimized with 5-fold cross validation and early stopping with a patience of 100. The PDFs for the random search for the LightGBM model can be seen in Table 5.4, and the ones for XGBoost in Table B.2. The random search has been run with 100 iterations for LightGBM and only 10 iterations for XGBoost since XGBoost turned out to be very slow⁹ at fitting datasets of this size.



The results of the HPO for 3-jet and 4-jet events can be seen in Figure 5.6. For 3-jet it can be seen how most of the iterations share about the same performance (within 1σ), however, some iterations show a significant decrease in performance. The same clear pattern is not found in the 4-jet events.

The relationship between the different hyperparameters in 4-jet events can be seen in the parallel coordinate plot in Figure 5.7. First

⁸ Short for charge-parity.

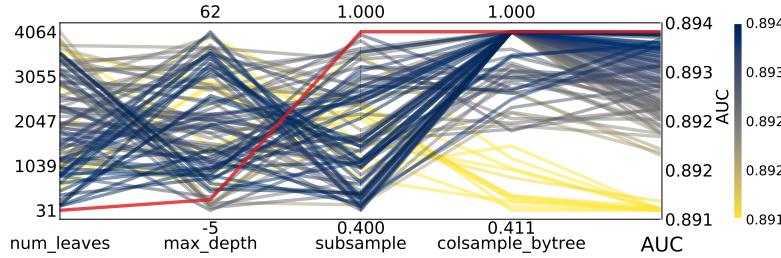
Hyperparameter	Range
subsample	$\mathcal{U}(0.4, 1)$
colsample_bytree	$\mathcal{U}_{\text{trunc}}(0.4, 1, 2)$
max_depth	$\mathcal{U}_{\text{int}}(-5, 63)$
num_leaves	$\mathcal{U}_{\text{int}}(7, 4095)$

Table 5.4: Probability Density Functions for the random search hyperparameter optimization process for the LightGBM model. For an explanation of $\mathcal{U}_{\text{trunc}}$, see subsection 5.6.2. All negative values of `max_depth` are interpreted as no max depth by both LGB and XGB.

⁹ See page 71 for a discussion of the timings.

Figure 5.6: Hyperparameter Optimization results of b -tagging with random search. From left to right, we have A) 100 iterations of RS with LGB on 3-jets, B) 10 iterations of RS with XGB on 3-jets, C) 100 iterations of RS with LGB on 4-jets, D) 10 iterations of RS with XGB on 4-jets.

of all the importance of the column downsampling `colsample_bytree` variable is significant: all of the low-performing hyperparameter sets have a low value of this hyperparameter. Since $M = 3$ for the vertex variables this makes logical sense; using only $\text{int}(\sim 0.5 \cdot 3) = 1$ variable¹⁰ the model cannot properly learn the structure in the data. Compared to the column downsampling, the other hyperparameters are less important. The same overall conclusion can be inferred in the 3-jet case, see Figure B.3.



¹⁰ See subsection 5.6.2 for a deeper discussion about the `colsample_bytree` hyperparameter.

Figure 5.7: Hyperparameter optimization results of b -tagging for 4-jet events. The results are shown as parallel coordinates with each hyperparameter along the x -axis and the value of that parameter on the y -axis. Each line is an event in the 4-dimensional space colored according to the performance of that hyperparameter as measured by AUC. The **single best hyperparameter** is shown in red.

5.4.2 b -Tagging Results

The prediction score \hat{y} is usually called the b -tag for b -tagging models and will be written as β_{tag} . The distribution of the b -tags for the two HPO-optimized models, LGB and XGB, together with the pre-trained neural network, NNB, can be seen in Figure 5.8 for 4-jet events and in B.5 for 3-jet events. Notice the strong match between the NNB and LGB models. The XGB model has almost no high b -tags ($\beta_{\text{tag}} > 0.8$), but a majority of b -tags in the very low end. This indicates that the XGBoost has focussed on the background events compared to the signal events, whereas the NNB and LGB models have focused more on the signal events.

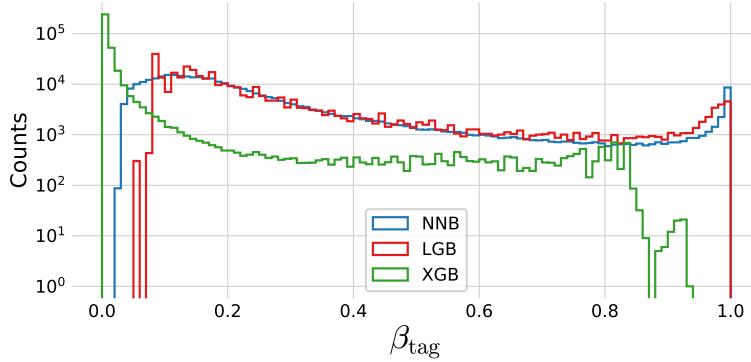
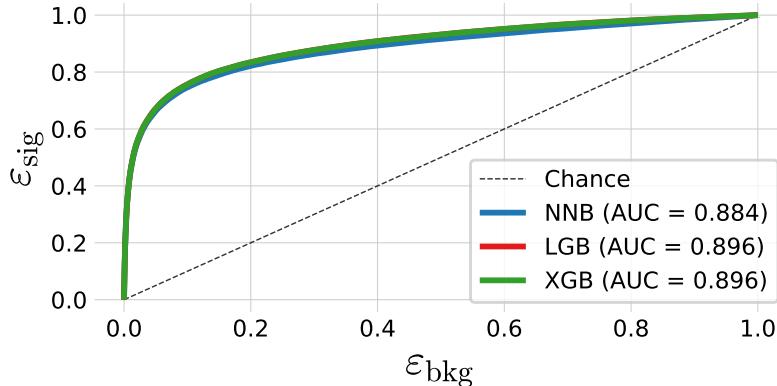


Figure 5.8: Histogram of b -tag scores β_{tag} in 4-jet events for `NNB` (the neural network pre-trained by ALEPH, also called `nnbjet`) in blue, `LGB` in red, and `XGB` in green.

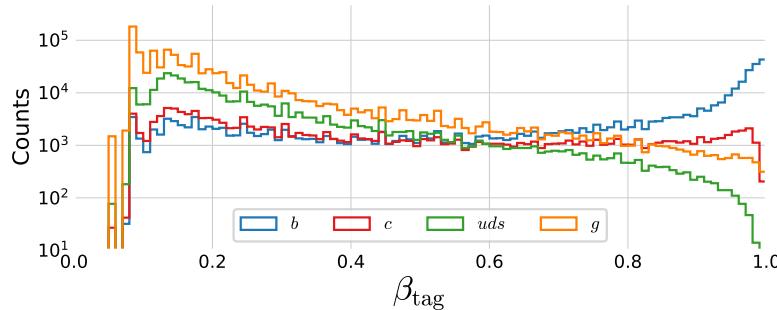
Even though the distributions of b -tags are different between the three models, the real performance plot for classification is the ROC curve seen in Figure 5.9 for 4-jet events. Here the signal efficiency ε_{sig} is plotted as a function of the background efficiency ε_{sig} with the AUC shown in the bottom right corner. The LGB and XGB models performs similarly well with an AUC = 0.896 compared to the NNB with AUC = 0.884. The differences between the models are even smaller for 3-jet events seen in Figure B.7. In general the LGB and XGB models are so similar that they cannot be distinguished from

another in any of the plots and their difference in AUC is on the fourth decimal point.



The LGB model is, however, several times faster than the XGB model. In comparison, 10 iterations of HPO using RS on 3-jet events with XGB took more almost 34 hours on HEP¹¹ compared to just 23 hours for 100 iterations for LGB. The same performance difference was seen in 4-jet events where the timings were 4 hours for XGB compared to 2.5 hours for LGB. Since their performance is similar, XGB is dropped in the subsequent analysis.

The distribution of the b -tag scores β_{tag} for signal and background in 4-jet events can be seen in Figure 5.10. The separation between the heavier quarks and light quarks (and gluons) is clear at high values of β_{tag} , however, a lot of c -quarks also get a high b -tag score. The same is seen for 3-jet events in Figure B.9.



5.4.3 b -Tagging Model Inspection

To get a better understanding of the trained LGB model, the global SHAP feature importances are shown in Figure 5.11 for 4-jet events. First of all it is noted that the `projet` has global feature importance of 57.32 %, `bqvjet` 29.16 %, and `ptljet` 13.52 %. For all three variables it is seen how most of the points have many small feature values which has a (small) negative impact on the model output. Especially the `ptljet` has many features with a low value (0 in fact) yet this does not pull the model too much towards background events. Compared this to a jet with a high value of `ptljet` which has a strong, positive impact on the output prediction.

In regression, the model output is a continuous prediction $\hat{y}_{\text{reg}} \in \mathbb{R}$. In classification what is actually happening under the hood is that

Figure 5.9: ROC curve of the three b -tag models in 4-jet events for `NNB` (the pre-trained neural network trained by ALEPH, also called `nnbjet`) in blue, `LGB` in red, and `XGB` in green. In the legend the area under curve (AUC) is also shown. Notice that the LGB and XGB models share performance and it is thus due to overplotting that only the green line for XGB can be seen.

¹¹ The local computing cluster.

Figure 5.10: Distribution of b -tags in 4-jet events for b -jets in blue, c -jets in red, uds in green and g in orange.

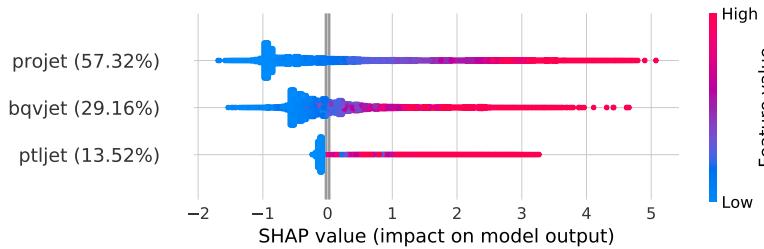


Figure 5.11: Global feature importances for the LGB b -tagging algorithm on 4-jet events. The normalized feature importance is shown in the parenthesis and each dot is an observation showing the dependence between the SHAP value and the feature value.

the model predicts a value $\tilde{y} \in \mathbb{R}$ which is transformed to a number in the $[0, 1]$ -interval via the *expit* function:

$$\text{expit}(\tilde{y}) = \frac{e^{\tilde{y}}}{1 + e^{\tilde{y}}} \equiv p, \quad (5.3)$$

where p is a number in the $[0, 1]$ -interval. The expit function is also sometimes known as the logistic function and is visualized in Figure 5.12. Its inverse is the *logit* function:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \tilde{y}, \quad (5.4)$$

which is visualized in Figure 5.13. The fraction in equation (5.4) is called the *odds* and the logit-transformed value of p , $\text{logit}(p) = \tilde{y}$, is thus sometimes called the *log-odds*. Because LightGBM makes its predictions in this log-odds space, the SHAP values in Figure 5.11 are also in log-odds space¹².

With this in mind, single predictions of the LGB b -tagging model can be understood with SHAP which Figure 5.14 is an example of. This plot, which is my own extension to Figure 3.22, shows the logic behind the model's prediction for a particular jet in a particular 3-jet event. That the bias is negative reflects that there is a majority of background jets compared to signal jets¹³. This particular event has `projet` = 1.003, `bqvjet` = 0.529, and `ptljet` = 0. In the plot it is seen how this high value of `projet` has the greatest impact on the model prediction, while the medium value of `bqvjet` also pushes the model prediction towards a signal-prediction. The four red and green bars in the left part of the plot are all in log-odds space and their sum is shown as the blue bar to right, where the right y -axis shows the value in probability space $p \in [0, 1]$. This jet was in fact a b -jet.

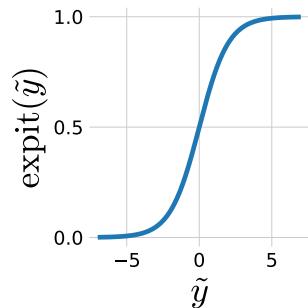


Figure 5.12: The expit function.

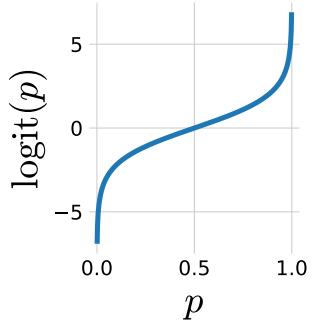


Figure 5.13: The logit function.

¹² The additivity property of SHAP, see section 2.8, is thus also in this log-odds space.

¹³ Only around 22 % of all the jets are b -jets.

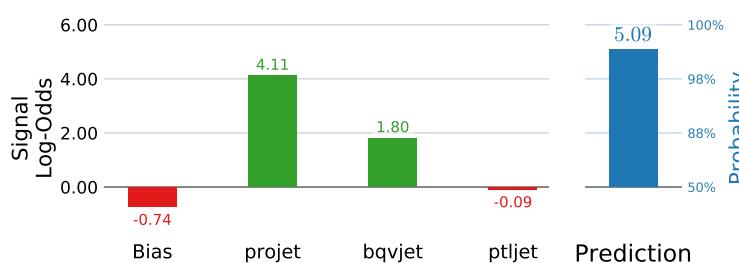


Figure 5.14: Model explanation for the 3-jet b -tagging LGB model for a b -like jet. The first column is the bias which can be seen as the naive prediction baseline, the rest are the input variables. On the right hand side of the plot is the model prediction shown. The left part of the plot is shown in log-odds space, the right part in probability space. The negative log-odd values are shown in red, positive ones in green, and the prediction value in blue.

5.5 *b*-Tagging Efficiency

Before any further analysis can be done, the efficiency of the *b*-tagging model has to be measured. The efficiency ε is defined as the number of particles, events, jets, or any other countable measure, N_{sel} , that are selected by the algorithm divided by the *true* number, N_{truth} :

$$\varepsilon = \frac{N_{\text{sel}}}{N_{\text{truth}}}. \quad (5.5)$$

Of course, the truth is never known in Nature, however, it is for simulated MC events. The efficiency is used to estimate how many particles (e.g.) that were generated even though only a subset of the particles were detected. Imagine a hypothetical experiment where 21 particles were observed and the efficiency of the experiment was $\varepsilon = 50\%$. This means that there were created $21/\varepsilon = 42$ particles in the experiment, yet only 21 of them were observed.

For measuring the *b*-tagging efficiency we apply a Tag-Tag-Probe (TTP) method based on the *b*-tags. In 3-jet events two of the jets will serve as tags and the last one as probe. The tags are jets where, if they are known, the probe is also known (with high probability). One can then apply the cut to the probe and see if it would have passed the cut or not. This method provides a clean and unbiased sample (the probes) and since (with high probability) the truth of the probe jet is known, the efficiency can be measured in this way [37]. Since the TTP method does not depend on real truth, it can be used on both MC and Data.

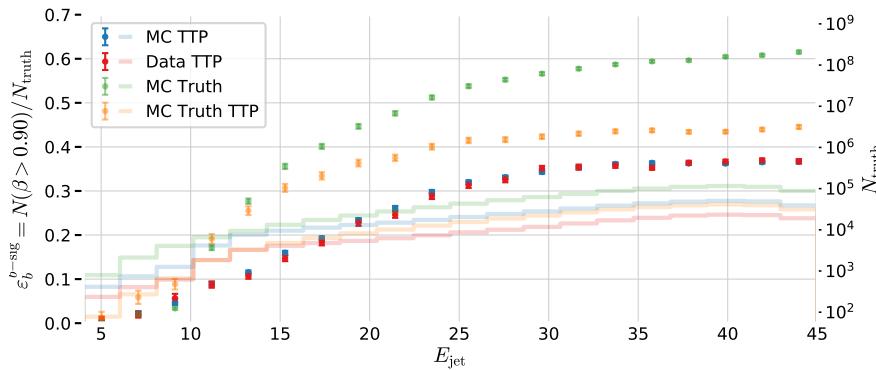
To measure the *b*-tagging efficiency we use that the Z decay has the form $Z \rightarrow q\bar{q}g$. Quantum field calculations based on the Standard Model predicts that the decay $Z \rightarrow ggg$ is extremely unlikely with a branching ratio of $\sim 10^{-6}$ [96]. This decay was actually tested at LEP which did not find any ggg events and concluded that the branching ratio was $< 1.6\%$ with an upper limit at 95% confidence level [41]. This number is further reduced today to $< 1.1\%$ [74].

That the Z decays to $q\bar{q}g$ means that if one of the jets get a high *b*-tag (and is thus likely to be a *b*-jet), and another one of the jets gets a low *b*-tag (and is thus likely to be a *g*-jet), then it is highly probable that the remaining jet is a *b*-jet. To formalize this, sort the jets after their *b*-tags values from high to low such that $\beta_{\text{tag}_1} > \beta_{\text{tag}_2} > \beta_{\text{tag}_3}$ for the jets $\mathbf{J} = [J_1, J_2, J_3]$ where J_i has *b*-tag β_{tag_i} . We then define the two tags T_b and T_g as $\mathbf{J} = [T_b, P, T_g]$ where P is the probe. If the two tags T_b and T_g passes the cuts $\beta_{b\text{-cut}} < \beta_{\text{tag}_1}$ and $\beta_{\text{tag}_3} < \beta_{g\text{-cut}}$, then the probe is selected $P = J_2$. If the probe is selected, then the last cut $\beta_{b\text{-cut}} < \beta_{\text{tag}_2}$ is the one that the efficiency is based on.

Based on Figure 5.10 we define the threshold for the *b*-jet tag to be $\beta_{b\text{-cut}} = 0.9$ and for the *g*-jet to be $\beta_{g\text{-cut}} = 0.4$. The *b*-signal region is thus $0.9 < \beta$ and the *g*-signal region $\beta < 0.4$. With these cuts, the efficiency, denoted $\varepsilon_b^{b\text{-sig}}$, of the *b*-jets being tagged as *b*-signal is computed. The TTP method is applied to both Data (Data TTP) and MC (MC TTP), along with a measurement of the efficiency when

measured using MC truth (MC Truth) and a measurement based on the probes that were actual b -jets according to truth (MC Truth TTP)¹⁴. The efficiency is measured as a function of jet energy E_{jet} to gauge the energy dependence of the efficiency, i.e. computed in a bin-by-bin basis split according to the jet energy (of the probe).

The efficiencies can be seen in Figure 5.15. The efficiency $\varepsilon_b^{b\text{-sig}}$ as a function of jet energy E_{jet} can be seen on the left y -axis, whereas the number of probes in each bin N_{truth} can be seen on the right y -axis. The efficiencies increase as a function of energy and reaches a plateau at $E_{\text{jet}} \sim 30 \text{ GeV}$: high-energy b -jets are easier to classify than low-energy ones. Even though the efficiencies of the MC TTP and Data TTP methods are lower than the MC Truth and MC Truth TTP, the important thing to notice is that they follow each other closely, an indicator of the trained b -tagging model working equally well on both MC and Data (as hoped).



¹⁴ So the probes are selected by the TTP method, however, only true b -jet probes are kept.

The efficiency of g -jets in the g -jet signal region $\varepsilon_g^{g\text{-sig}}$ based on the b -tags can be measured in a similar manner. Again TTP method is used, however, now the two b -jets are the tags and the g -jet is the probe. The cuts are the same as before, however, now it is required that $0.9 < \beta_{\text{tag}_1}$ and $0.9 < \beta_{\text{tag}_2}$ before the probe is selected $P = J_3$. The efficiency is then based on $\beta_{\text{tag}_3} < 0.4 = \beta_{g\text{-cut}}$. The efficiency $\varepsilon_g^{g\text{-sig}}$ is plotted in Figure 5.16. Here the MC TTP and Data TTP also follow each other closely, this time to around $\sim 25 \text{ GeV}$.

Both of the efficiencies so far, $\varepsilon_b^{b\text{-sig}}$ and $\varepsilon_g^{g\text{-sig}}$, can be seen as signal efficiencies. Likewise, there are two background efficiencies, one for b -jets in the g -signal region $\varepsilon_b^{g\text{-sig}}$ seen in Figure B.15 and one for g -jets in the b -signal region $\varepsilon_g^{b\text{-sig}}$ seen in Figure B.16.

This section shows that the b -tagging efficiencies of the LGB b -tagging model shows comparable performance on both MC and Data which indicates that the model is un-biased (w.r.t. energy).

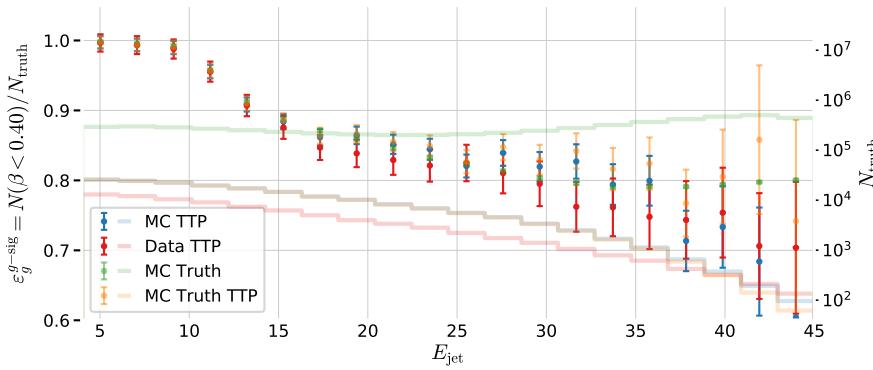


Figure 5.16: b -tag efficiency for g -jets in the g -signal region for 3-jet events, $\epsilon_g^{g\text{-sig}}$, as a function of jet energy E_{jet} . In the plot the efficiencies are shown for **MC TTP** in blue, **Data TTP** in red, **MC Truth TTP** in green, and **MC Truth** in orange. The efficiencies (the errorbars) can be read off on the left y -axis and the counts (histograms) on the right y -axis. Notice how both MC TTP and Data TTP follow each other closely until $\sim 25 \text{ GeV}$.

5.6 g -Tagging Analysis

The b -tagging model is a jet-based model which provides a b -tag score β_{tag} to a jet. This also means that each of the jets in e.g. a 4-jet event can get a b -tag: $\beta_{\text{tag}} = [\beta_{\text{tag}_1}, \beta_{\text{tag}_2}, \beta_{\text{tag}_3}, \beta_{\text{tag}_4}]$. Treating β_{tag} as an individual observation, one can train a new model on the events based on the events compared to the individual jets. This event-based process will be called g -tagging and the trained model will return a g -tag score written as γ_{tag} .

For this model, signal events are defined to be events which are q -matched¹⁵ and where the two jets with the highest b -tags are matched to the two (final) quark jets. Another way to say this, is that the non- q -matched jets are assigned the $n - 2$ lowest b -tag scores for n -jet events. An example of a signal event is $\beta_{\text{tag}} = [0.95, 0.89, 0.15, 0.07]$ for an event with the four jets $[b, \bar{b}, g, g]$. Compared to the b -tagging model, this model will allow one to extract entire events which contains a clear identification of gluons versus non-gluons.

5.6.1 Permutation Invariance

Since the b -tags are only based on the vertex variables, the goal of the g -tag is to also be constructed in an un-biased way with respect to the jet energy E_{jet} . However, even though β_{tag} is independent of E_{jet} and γ_{tag} is a function of β_{tag} , it turned out that γ_{tag} was not independent of E_{jet} . This was because the ordering of the jets within the event was energy-dependent: they are sorted according to their E_{jet} .

This meant that the different variables (b -tags) in β_{tag} had different feature importances when tested, even though they should be equally important. Instead of defining β_{tag} as a vector it should instead be seen as a set¹⁶ $\beta_{\text{tag}} \equiv \{\beta_{\text{tag}_1}, \dots, \beta_{\text{tag}_n}\}$. The g -tagging model trained on the events should thus be *permutation invariant*¹⁷ with regards to the input variables. The category of permutation invariant (and equivariant¹⁸) neural networks has seen an huge development within recent years in the deep learning community. The paper from Zaheer et al. [105] in 2017 was highly influential, however also other examples exists [79, 52]. Yet, the same development

¹⁵ Remember that q -matched events are events with one, and only one, jet that is q -matched to one of the quark-jets, and one, and only one of the jets is q -matched to the other quark-jet.

¹⁶ Since sets have no inherent order.

¹⁷ $f(\mathbf{x}) = f(\tau(\mathbf{x}))$ for any permutation τ on an input vector \mathbf{x} .

¹⁸ $\tau(f(\mathbf{x})) = f(\tau(\mathbf{x}))$ for any permutation τ on an input vector \mathbf{x} .

cannot be said to have happened within the more classic machine learning field.

Although not being a novel software-technical solution, I circumvent the problem with two simple different approaches: 1) by simply shuffling the inputs variables independently for each observation (row) in the dataset, and 2) training on all possible permutations of the variables in the dataset. The second approach can be seen as a feature augmentation technique where the data is artificially increased with factor of n factorial: $N \rightarrow n! \cdot N$ where N is the number of events and $n \in \{3, 4\}$ is the number of jets. These two methods were tested along with keeping the original order of the dataset.

5.6.2 Truncated Uniform PDF

Initially when plotting the HPO performance as a function of iteration, it was seen how there were some very clear plateaus, where the highest plateau (i.e. the highest AUC value and thus the best score) was only seen in the very first iteration. It was quickly realized that this was due to the very first iteration was being run with the default values of the LGB in my HPO setup. However, what was not understood was why this value was performing so much better than the different sets of hyperparameters in the random search. Of course LightGBM have chosen their default parameters wisely, however, one would not expect them to outperform other sets of hyperparameters that clearly. During the debugging process, the column downsampling `colsample_bytree` was diagnosed to be the culprit.

The default value is `colsample_bytree` = 1, however, the probability density function (PDF) used in random search for this parameter was $\mathcal{U}(0.4, 1)$ which was expected to give the same performance as the default value, at least for values of `colsample_bytree` close to 1. By inspecting the source code of LightGBM, I realized that if the column downsampling is less than 1 the model takes the integer of the column downsampling multiplied with the total number of columns (variables/features) [6]. This means that no matter how close to 1 the column downsampling get, the integer value of the total number of columns get floored to maximally 2 in 3-jet events, compared to when the column downsampling is exactly 1 (which it only is for the default values).

To deal with this problem I developed a new PDF¹⁹ on top of the existing ones in Scipy: the truncated uniform PDF $\mathcal{U}_{\text{trunc}}(a, b, c)$. This PDF first generates a random number x from a uniform distribution between a and c . Then, if x is larger than b it is floored to b . In this way, it is possible to both get values of x in the interval $[a, b]$ but also values exactly equal to b . The value of c controls how often these “overflow” values of x are generated.

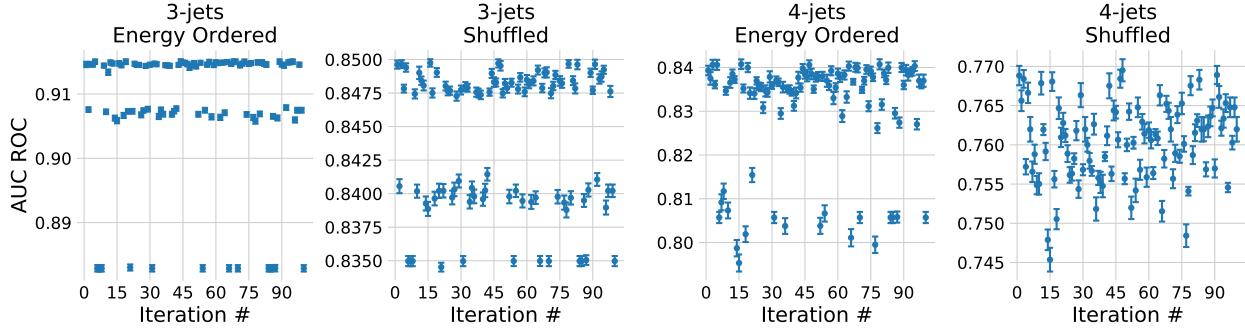
¹⁹ Not strictly a PDF since it is not normalized, but otherwise behaves as one.

5.6.3 g -Tagging Hyperparameter Optimization

Four LightGBM models, two for 3-jet events and two for 4-jet events, were trained and hyperparameter optimized for both the the energy

ordered and shuffled data sets with 100 iterations of random search. The method with all permutations was trained using the same hyperparameters as the best ones found in the HPO for the shuffled model to reduce time²⁰ spent on HPO. The same PDFs as for the b -tagging were used, see Table 5.4, and 5-fold cross validation and early stopping was applied with a patience of 100.

²⁰ The time performance is extra important for the method with all permutations as the dataset is 24 times larger than the method with the shuffling for 4-jet events.



The results of the HPO can be seen in Figure 5.17. Here the two 3-jets models are seen in the two plots to the left, and the two 4-jets to the right. The very left plot shows the performance as a function of iteration number for the 3-jet energy-ordered method (no permutation or shuffling). This was where the issues mentioned in subsection 5.6.2 were first discovered. There are three very noticeable plateaus in this plot which corresponds to running column subsampling with zero, one, or two variables dropped. The three plateaus are also seen in the 3-jet events that were shuffled, however, with more variation in each plateau (along with a drop in performance). For the 4-jet events the plateaus are not as apparent but it can still be seen how some of the iteration show a significantly lower score than others. The parallel coordinate plots for the four plots can be seen in Figure B.17–B.20.

The global SHAP feature importances $\phi_{\beta_i}^{\text{tot}}$ are computed for each one of the three methods, energy-ordered, shuffled, or with all permutations, to verify their permutation invariance properties. The results are seen in Table 5.5 for 4-jet events and in Table B.3 for 3-jet events.

β_{tag_i}	Energy Ordered	Shuffled	All Permutations
$i = 1$	0.986 ± 0.008	0.474 ± 0.005	0.465 ± 0.005
$i = 2$	0.609 ± 0.006	0.467 ± 0.005	0.464 ± 0.005
$i = 3$	0.424 ± 0.004	0.461 ± 0.005	0.452 ± 0.005
$i = 4$	0.244 ± 0.002	0.481 ± 0.005	0.466 ± 0.005

Here it can be seen that the model trained on the energy ordered data learned to attribute the highest weight to the first b -tag, second highest weight to the second b -tag, and so on. In contrary, the weights are uniformly distributed between the different b -tags in both the shuffled and all-permuted datasets (within a few sigma). The same overall pattern is seen for the 3-jet events. Based on the ta-

Figure 5.17: Hyperparameter Optimization results of g -tagging with 100 iterations of random search with LGB. From left to right, we have A) 3-jet events energy-ordered (no permutations), B) 3-jet events row-shuffled, C) 4-jet events energy-ordered, D) 4-jet events row-shuffled. Notice the different ranges on the y-axes.

Table 5.5: Global SHAP feature importances $\phi_{\beta_i}^{\text{tot}}$ for the three g -tagging models in 4-jet events. Each $\phi_{\beta_i}^{\text{tot}}$ is shown for the three methods in the columns and the four b -tags as variables in the rows.

bles, it can be seen that both the shuffling method and all-permuting method are methods for training ML models with permutation invariant properties due to their approximately equal attribution of weight to the different variables (b -tags).

5.6.4 PermNet

In addition to the LGB models, a permutation invariant neural network called PermNet based on the Deep Sets paper [105] implemented in Tensorflow [10] by Faye [47] was also tested. Zaheer et al. [105] showed that $f(X)$ is permutation invariant if and only if it can be decomposed in the following way:

$$f(X) = \rho \left(\sum_{x \in X} \phi(x) \right). \quad (5.6)$$

for suitable transformations ρ and ϕ which the neural network learns. This is possible since neural networks are universal function approximators [56]. The PermNet was trained using three layers²¹ with leaky ReLU [69] as the activation function and ADAM [62] as the optimizer optimizing the log-loss. The network was trained with early stopping with a patience of 50 epochs and a batch size of 128. A visual overview of the PermNet architecture can be seen in Figure B.21. It took around 6 hours to fit the 3-jet events and 4.5 hours for the 4-jets (due to fewer events) for each of the models.

5.6.5 1D Comparison of LGB and PermNet

I made a small study to better understand the LGB and (especially) the PermNet models. This comparison was constructed by summing the b -tag scores in the n -jet event together $\sum_i^n \beta_{\text{tag}_i}$. The β_{tag_i} are summed together since this turns the problem into a 1D problem that is easy to visualize, the sum of numbers is a permutation invariant function. The sum also corresponds to the simplest functions of ρ and ϕ in equation (5.6): the identity function. Both 1D models are fit to the training events. After the fits, a linear scan from $\sum_i^n \beta_{\text{tag}_i} = 0.4$ to 3.1 is made to see how the predicted g -tags distribute. This is shown in Figure 5.18 for 4-jet events. Here the value of γ_{tag} is shown for the two models together with the fraction of signal to background in each bin. If the g -tag score should resemble a true probability it would be expected to follow the signal ratio, e.g. a model should predict $\gamma_{\text{tag}} = 0.9$ if there is 90 % signal in that bin. In the figure it is seen how the PermNet does a great job at fitting the signal fraction and the LGB model also does a decent job. Remember that none of these models were shown the signal fraction explicitly, only the b -tag sum and a truth label. The distribution of signal and background²² together with the distribution of cuts made by the LGB model can be seen in Figure B.23. The similar plots for 3-jet events are plotted in Figure B.24 and B.25.

It can be concluded, at least in 1D, that both LGB and PermNet are able to capture the inherent structure in the 1D

²¹ Where the two hidden layers have 128 and 64 neurons in each.

²² Which the signal fraction is based on.

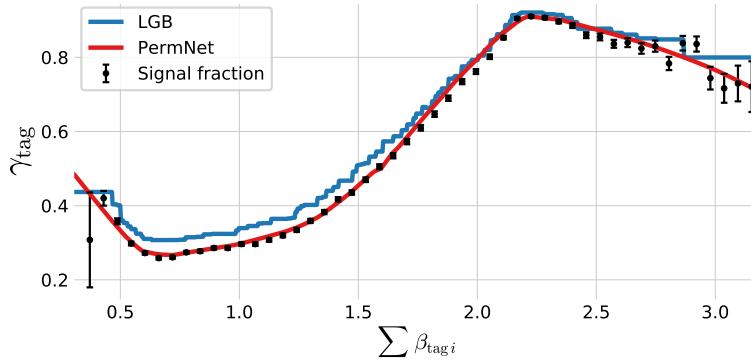
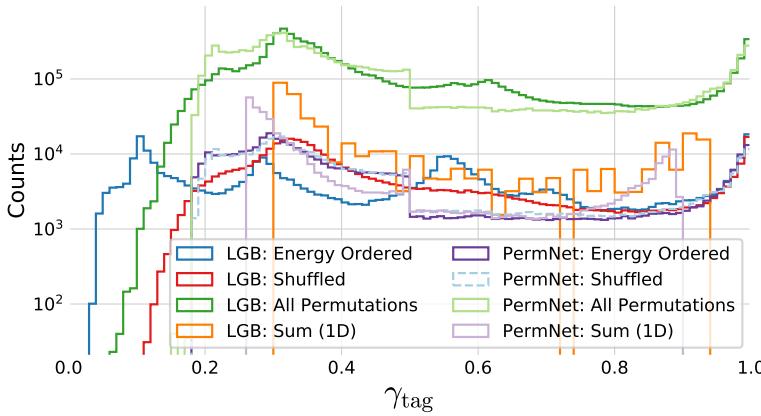


Figure 5.18: Plot of the (1D) g -tag scores for 4-jet events as a function of $\sum \beta_{\text{tag}_i}$ for the LGB model in blue and the PermNet model in red. The signal fraction (based on the signal and background histograms in Figure B.23) is plotted as black error bars where the size of the error bars is based on the propagated uncertainties of the signal and background histogram assuming Poissonian statistics.

5.6.6 g -Tagging Results

The distribution of g -tag scores in 4-jet (training) events can be seen in Figure 5.19 for the eight combinations of the two models (LGB and PermNet) and the four data sorting methods (energy ordered, (row) shuffled, all permutations, and the 1D sum.). At first the increased number of events (a factor of 24 for 4-jet events) with the all-permutation scheme is seen separating the two light green curves from the rest. The energy ordered LGB model is the combination which utilizes most of the γ_{tag} -range, while the two 1D sum models have the most limited range, indicating that the models are more uncertain about their predictions. The energy ordered and shuffled PermNet models can more or less only be distinguished because the latter is plotted with dashed lines. This makes sense, since they are also expected to make the same predictions were they really permutation invariant²³. When plotted with normalized counts it is seen how the shuffled and all-permuted LGB models also follow each other very closely, which can still be partly seen in this plot by comparing the two distributions. The distribution of g -tags in 3-jet training events can be seen in Figure B.27.

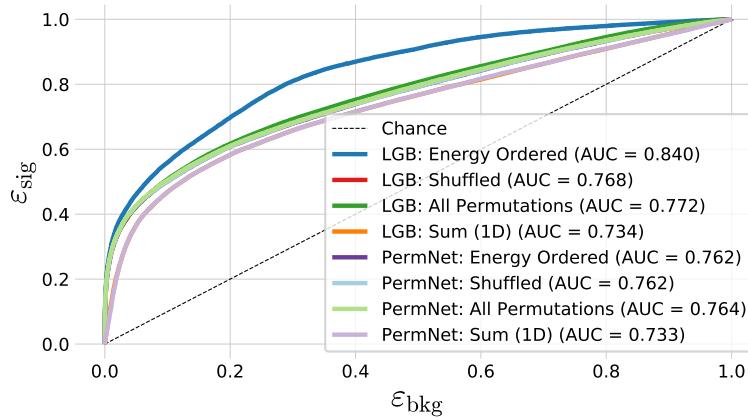


²³ It is only because of the stochasticity in the optimization process of the two networks that they did not converge to exactly the same predictions.

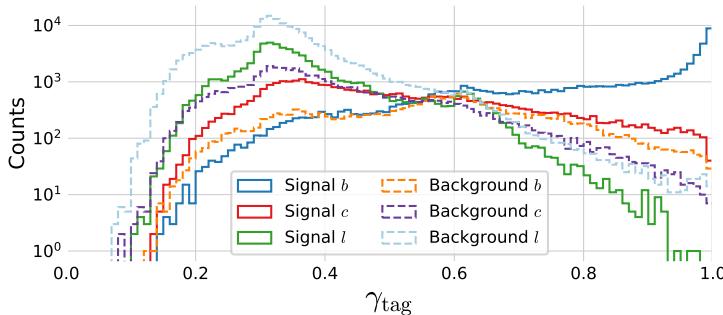
Figure 5.19: Distribution of g -tag scores in 4-jet events shown with a logarithmic y -scale for LGB: Energy Ordered in blue, LGB: Shuffled in red, LGB: All Permutations in green, LGB: Sum 1D in orange, PermNet: Energy Ordered in purple, PermNet: Shuffled in light-blue, PermNet: All Permutations in light-green, PermNet: Sum 1D in light-purple. Here LGB and PermNet are the two different type of models and “Energy Ordered”, “Shuffled”, “All Permutations”, and “Sum 1D” are the different methods used for making the input data permutation invariant (except energy ordered).

The ROC curve in Figure B.28 shows the performance of the different models on 4-jet events with the AUC shown in the legend. First of all it is easy to see that the energy ordered LGB model is significantly higher-performing than the rest of the models, however, this model is also energy-biased (not permutation invariant in the

b -tags) and is only included to see how large a performance drop the permutation invariance criterion causes. The worst performing models are the two 1D sum models since they only have a single dimension to learn from, compared to the four dimensions that the other models have. Overall it can be seen that the rest of the models are performing almost identically, with the LGB model trained on all permutations to be the highest-performing of them all by a small margin. For 3-jet events a similar picture is seen, see Figure ??, however, here the LGB model trained on the shuffled events performs the best, yet this performance improvement is so small compared to the all-permutations LGB model that it is expected to be due to statistical fluctuations and not a real performance difference.



Based on the AUC scores seen in the ROC curves in Figure B.28 and ??, the LGB-model trained on all permutations will be the g -tagging model choice. To see how this model's predictions of the g -tags distribute for signal and background events, see Figure 5.21. Here the distribution of γ_{tag} is shown for 4-jet signal events and background events. Remember that in g -tagging, the signal events are defined as events where the two jets with the highest b -tags are also the two q -matched jets (and the entire event is q -matched). In the figure it can be seen that at high values of γ_{tag} primarily $b\bar{b}gg$ events (signal b) are tagged where the jets are sorted according to their b -tags. Next after signal b events is $c\bar{c}gg$ (signal c) and $b\bar{g}g$ -events²⁴ (background b). At low values of γ_{tag} light quarks (uds) dominate.



The similar plot for 3-jet events is seen in Figure B.31. This plot has some surprising bumps for mainly l -quark events. When comparing

Figure 5.20: ROC curve of the eight g -tag models in 4-jet events. First one in dashed black is the ROC curve that you get by random chance. The colors are the same as in Figure 5.19 and in the legend also the Area Under the ROC curve (AUC) is shown.

²⁴ Or any other permutation of b, \bar{b}, g, g which is not $b\bar{b}gg$.

Figure 5.21: Histogram of g -tag scores from the LGB-model in 4-jet events for b signal in blue, c signal in red, l (uds) signal in green, b background in orange, c background in purple, l (uds) background in light-blue.

l -quark events in the high- γ_{tag} bump with the ones getting a low γ_{tag} -value, see Figure 5.22, one can see that l -quark events with high γ_{tag} has only two jets with high b -tags, compared to low- γ_{tag} l -quark events which more often has three jets with high b -tags.

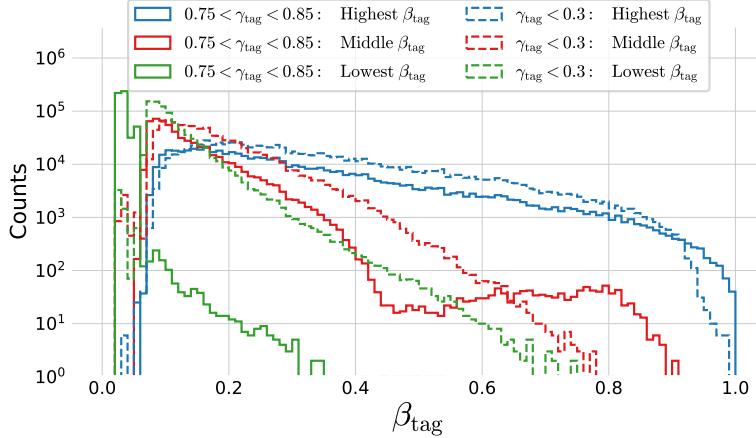
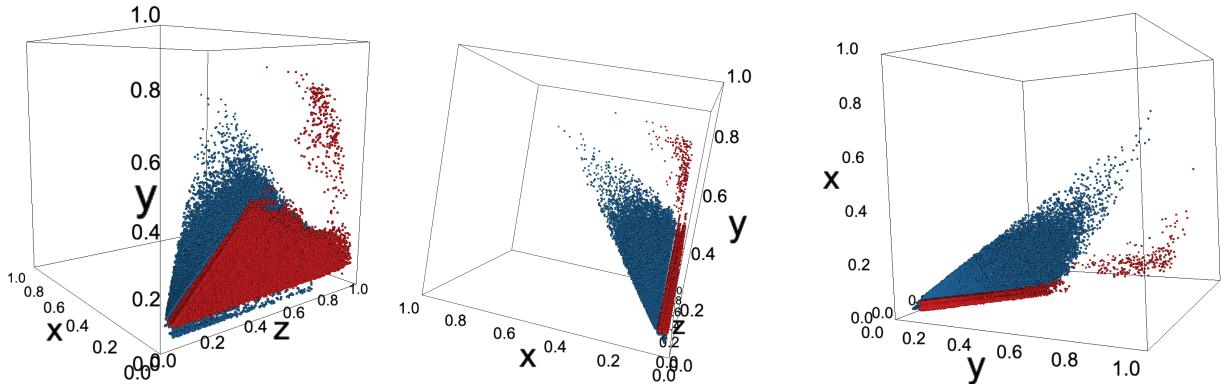


Figure 5.22: Distribution of b -Tag Scores in 3-Jet l -Quark Events for low and high g -tags values. Here l -quark events with $0.75 < \gamma_{\text{tag}} < 0.85$, so the high peak in Figure B.31, are plotted in fully connected lines, and events with $\gamma_{\text{tag}} < 0.3$ are plotted in dashed lines. For each of these two selection of events the value of the jet with the **highest β_{tag}** is shown in blue, the jet with the **middle β_{tag}** in red, and the jet with the **lowest β_{tag}** in green.

This is even more visible once seen in a 3D scatter plot with the lowest β_{tag} on the x -axis, the middle on the y -axis, and highest on the z -axis. Three small views from the 3D visualization can be seen in Figure 5.23. Here it is easily seen how the separating variable is the lowest b -tag: if an event where all three jets have high b -tags are used as input to the g -tagging model it gives it a low g -tag compared to if only two of the three jets have high b -tags.



5.7 g -Tagging Efficiency

These efficiencies are only possible to measure for MC-generated data as the truth labels are required. The Tag-Tag-Probe (TTP) method in section 5.5 is not possible for whole events as every event is completely independent of the other and thus one event cannot work as a tag for another event. We can, however, construct a pseudo g -tagging efficiency based on the b -tagging efficiencies. This efficiency will be computed by looking at 3-jet events with two jets with a high b -tag and one jet with a low b -tag, i.e. events where two jets has $0.9 < \beta_{\text{tag}}$

Figure 5.23: 3D scatter plot of β_{tag} -values for high and low γ_{tag} l -quark events. Here the x -axis is the lowest b -tag, the y -axis the middle, and the z -axis the highest. Here the **high- γ_{tag} l -quark events** are plotted in red and the **low ones** in blue.

and one jet has $\beta_{\text{tag}} < 0.4$. This indicates a $b\bar{b}g$ event where all of the jets have been correctly identified by the b -tagging algorithm. The pseudo efficiency $\varepsilon_{b\bar{b}g}$ is then defined as:

$$\varepsilon_{b\bar{b}g} = \varepsilon_b^{\text{b-sig}}(b) \cdot \varepsilon_{\bar{b}}^{\text{b-sig}}(\bar{b}) \cdot \varepsilon_g^{\text{g-sig}}(g). \quad (5.7)$$

This is only a pseudo efficiency since this number is based on the jets in the event and not the event itself, however, by plotting it as a function of an event variable and comparing MC to Data, we can gauge the validity of the g -tagging algorithm. The first of the event variables used is the g -tag of the event γ_{tag} , see Figure 5.24. Here the pseudo efficiency is plotted as a function of γ_{tag} for Data and MC together with the counts in each bin and the ratio between $\varepsilon_{b\bar{b}g}$ for Data and MC is plotted below. At low values of γ_{tag} the uncertainties dominate due to low statistics, however, at higher γ_{tag} $\varepsilon_{b\bar{b}g}$ plateaus until very high values of γ_{tag} where it increases again. The important thing to note in this figure is the high agreement between Data and MC which converges to (almost) 1 at high γ_{tag} -values. This is an important indication of the 3-jet g -tagging model working.

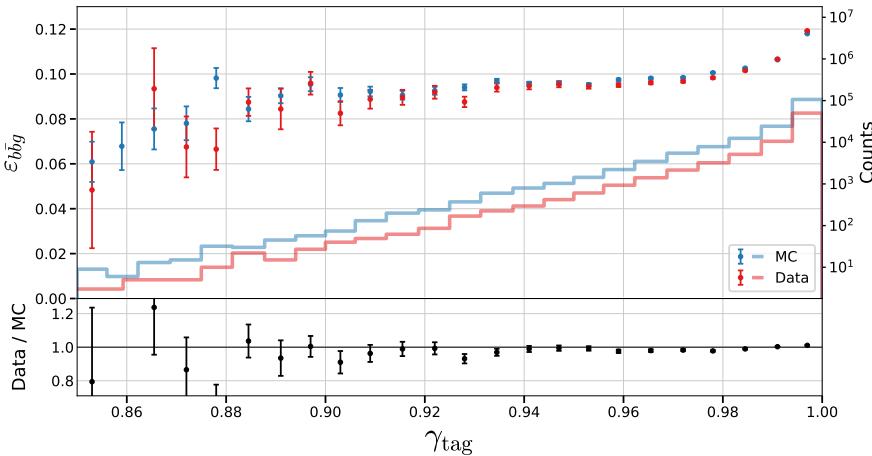


Figure 5.24: Pseudo efficiency of the g -tags for $b\bar{b}g$ 3-jet events as a function of the event's g -tag γ_{tag} . In the top plot the pseudo efficiency $\varepsilon_{b\bar{b}g}$ is shown for MC in blue and Data in red where the counts in each bin can be read on right y-axis. In the bottom plot the ratio between Data and MC is shown. The pseudo efficiency is measured by finding $b\bar{b}g$ -events where $\beta_b > 0.9$, $\beta_{\bar{b}} > 0.9$, and $\beta_g < 0.4$, and then calculating $\varepsilon_{b\bar{b}g} = \varepsilon_b^{\text{b-sig}} \cdot \varepsilon_{\bar{b}}^{\text{b-sig}} \cdot \varepsilon_g^{\text{g-sig}}$.

Another event variable to look at is the mean of the two invariant masses m_{bg} and $m_{\bar{b}g}$. The invariant mass between two quantities²⁵ is:

$$m_{12} = \sqrt{(E_1 + E_2)^2 - \|\mathbf{p}_1 + \mathbf{p}_2\|^2}, \quad (5.8)$$

where E_i is the energy of the i^{th} quantity and \mathbf{p}_i its momentum. The invariant masses between all three jets $m_{b\bar{b}g}$, which otherwise might have been the first intuition to use as the event variable, is non-informative (in this context) since this is just the total event energy which is kept (approximately) constant at ~ 91 GeV.

The pseudo efficiency is plotted as a function of the mean of m_{bg} and $m_{\bar{b}g}$ in Figure 5.25. Here the overall correspondence between Data and MC is lower than in Figure 5.24, especially for high values of the mean invariant mass. That the ratio is close to 1 for most of the data (from 20 GeV to 50 GeV) is a good sign, whereas the tails indicate the model is biased here which could be corrected for with correction factors, however, this would increase the systematic uncertainties.

²⁵ When measured in natural units.

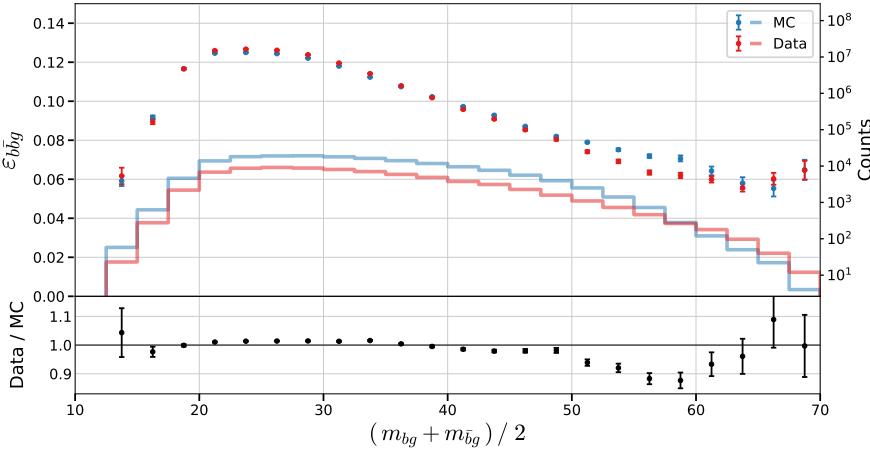


Figure 5.25: Pseudo efficiency of the g -tags for $b\bar{b}g$ 3-jet events as a function of the mean of the two invariant masses $m_{b\bar{b}g}$ and $m_{\bar{b}\bar{b}g}$ in the event. In the top plot the pseudo efficiency $\varepsilon_{b\bar{b}g}$ is shown for MC in blue and Data in red where the counts in each bin can be read on right y-axis. In the bottom plot the ratio between Data and MC is shown.

These two figures strengthens the claim that the trained b -tagging and g -tagging models provide un-biased models that not only work in Data but also in MC.

5.8 Generalized Angularities in 3-jet events

To measure how gluon jet hadronizes, i.e. their jet distributions, number of tracks, etc., the *generalized angularities* provide an overall framework for doing so. The generalized angularities is a two-parameter family of variables depending on the angular weighting $\beta \leq 0$ and an energy weighting factor $\kappa \leq 0$:

$$\lambda_{\beta}^{\kappa} = \sum_{i \in \text{jet}} z_i^{\kappa} \theta_i^{\beta}, \quad (5.9)$$

where $z_i \equiv E_i / E_{\text{jet}}$ is the momentum fraction, i.e. $0 \leq z_i \leq 1$, $\theta_i \equiv \Omega_i / R$ is the normalized angle with respect to the jet axis where R is the jet radius such that $0 \leq \theta_i \leq 1$, and i runs over all the jet constituents [51, 64]. Different values of (β, κ) probe different parts of the (gluon) jet fragmentation phase space. I will limit the analysis to the five sets of (β, κ) -values shown in Figure 5.26, where each of the sets of variables are related to the following aspects:

(β, κ)

$(0, 0)$: Hadron Multiplicity.

$(0, 2)$: Transverse Momentum Distribution p_T^D :

$$\lambda_0^2 = \sum z_i^2 \equiv (p_T^D)^2 [38].$$

$(\frac{1}{2}, 1)$: Les Houches Angularity (LHA) [91].

$(1, 1)$: Width or broadening [35].

$(2, 1)$: Mass.

We will look at the generalized angularity distributions for gluons in 3-jet events. We do so by using the g -tag from the g -tagging model to select events with a high g -tag and then select the jet in the

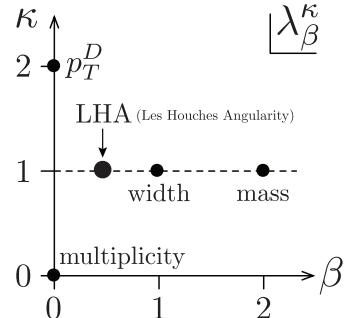


Figure 5.26: Generalized angularities. Adapted from Larkoski et al. [64].

event with the lowest of the b -tags since this jet is expected to be the gluon jet. From the plot in Figure B.31, the γ_{tag} cut off threshold is set to $\gamma_{\text{cutoff}} = 0.9$ for 3-jet events. This cut corresponds to selecting 340 476 events in MC as gluon events with a signal efficiency of $\epsilon_g^{\text{3-jet}} = 19.68\%$ and a signal purity of $\rho_g^{\text{3-jet}} = 98.77\%$. Here a gluon event is defined as an event with $0.9 < \gamma_{\text{tag}}$.

The distribution of λ_0^2 , i.e. $(\beta, \kappa) = (0, 2)$, related to the transverse momentum distribution, in gluon events is seen in Figure 5.27. Here the distribution of λ_0^2 is shown for MC Truth (actual gluons jets using truth-label), MC selected (gluons jets selected using the g -tag) and Data (gluons jets selected using the g -tag). The generalized angularities are computed for both charged jets and neutral clusters, where this figure shows the distribution of λ_0^2 for charged jets. The MC Selected has been scaled to Data according to the fraction of the number of events in each: $w_{\text{MC}} = N_{\text{Data}} / N_{\text{MC}}$. The MC Truth has been scaled with the same weight multiplied with the gluon efficiency $w_{\text{MC-Truth}} = w_{\text{MC}} \cdot \epsilon_g^{\text{3-jet}}$. The $\lambda_\beta^\kappa = 0$ values has been removed before plotting for all sets of values of (β, κ) .

In Figure 5.27 it is seen how MC Truth and Data does not match perfectly well which indicates that this is not the perfect method, however, that MC Selected and Data distributions matches each other quite well indicates that the model is equally biased for both MC and Data. The errorbars in the ratio plot are fitted with both a constant $f(x) = b$ and a straight line $f(x) = ax + b$ with the fit results shown as text in the plot. In the text box `DOF` is the number of degrees of freedom and `P` is the χ^2 -probability²⁶. I compute the systematic error σ_{sys} that would have to be added in quadrature to the standard deviation, $\sigma \rightarrow \sqrt{\sigma^2 + \sigma_{\text{sys}}^2}$, such that the χ^2 -probability would be 50%. I do this using Brent's method [27] and the systematic error is written as `sigma_sys` in the figure.

²⁶ $P(\chi^2; N_{\text{DOF}}) = \int_{\chi^2}^{\infty} f_{\chi^2}(x; N_{\text{DOF}}) dx$, where $f_{\chi^2}(x; N_{\text{DOF}})$ is the χ^2 distribution with N_{DOF} degrees of freedom.

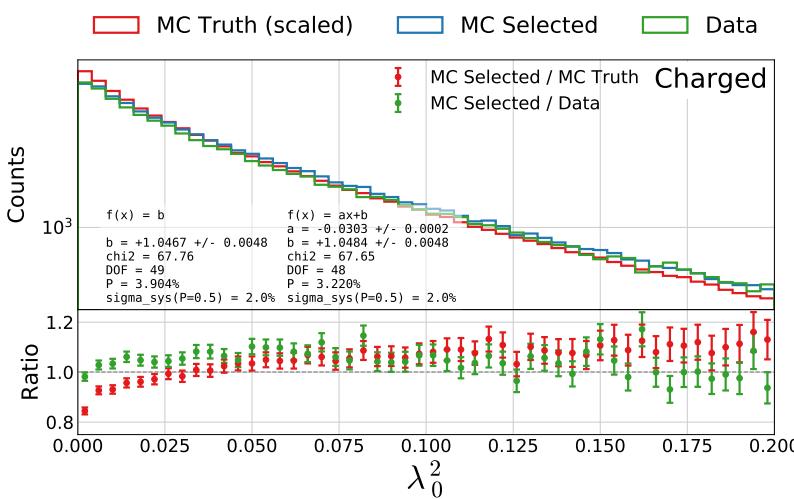


Figure 5.27: Distribution of the generalized angularity λ_0^2 for charged gluons jets in 3-jet events: λ_0^2 . The distributions for **MC Truth** is shown in red, **MC Selected** in blue, and **Data** in green in the top plot and in the bottom plot the ratio between **MC Selected** and **MC Truth** is shown in red and between **MC Selected** and **Data** in green.

The rest of the plots can be seen in Figure B.32–B.41. For the charged tracks it can be seen that the first few bins generally has ratios just below 1 with a upwards trend until around the middle of the λ range and then it decreases again (and the uncertainties

increases with lower and lower statistics). The discrepancy between MC selected and Data is larger for the neutral clusters, which are also less well-defined than the charged jets (at the energies used at LEP), and with the same trend as for the charged jets.

These trends are quantified in by the fits to the ratio plots. For λ_0^2 , Figure 5.27 above, we see that a constant fit does a reasonable job with a $P_\chi^2 = 3.9\%$. The small decrease in χ^2 value, $\chi^2 = 67.76$ for the constant fit but only 67.65 for the linear fit, is attributed to the linear fit being a more advanced fit (more fit parameters). In this case the difference would be expected to be $\Delta\chi^2 \sim 1$ due to difference in the number of fit parameters being 1, which is also what is seen. As such, the systematic error will be based on the constant fit here $\sigma_{\text{sys}}(\lambda_0^2) = 2.0\%$.

In general there is a $\sim 4\%$ constant offset for the charged tracks, which is easy to correct for. For the neutral clusters it is seen to be a lot higher at $\sim 15\%$. For charged tracks the systematic error is around $\sim 2\%$ for most of the generalized angularities, while it is around $\sim 4\%$ for the neutral clusters. A summary of all the systematic errors for the charged tracks can be seen in Table 5.6 and in Table 5.7 for the neutral clusters. Overall, the figures show a reasonable good correspondence between MC selected and Data for the generalized angularities.

5.9 Gluon splitting

In addition to measuring how the gluon jets hadronizes, we are also interested in measuring how they split: $g \rightarrow gg$. We do so by looking at 4-jet events with high g -tag values and then identify the two gluon jets (the ones with the lowest b -tag values). From Figure 5.21, the γ_{tag} cut off threshold is set to $\gamma_{\text{cutoff}} = 0.8$ for 4-jet events. This corresponds to selecting 41117 events in MC as gluon events with a signal efficiency of $\varepsilon_g^{4\text{-jet}} = 23.30\%$ and a signal purity of $\rho_g^{4\text{-jet}} = 92.67\%$.

5.9.1 Variables

The variables we use to measure the gluon splitting are each meant to probe different, smaller parts of the phase space. It is difficult to parametrize the basis of this phase space, but in discussion with one of the authors of the MC simulator Pythia [87], Peter Skands, the first dimension of the phase space will be the energy asymmetry. The energy asymmetry describes how the two gluons share their available energy between them; do they share it equally or does one of them get the majority of it? The second dimension is the resolution scale which probes the hardness of the gluons by looking at their invariant mass and the angle between the their jets. The third axis is the azimuthal angle which measures the rotational distribution of the gluons compared to the quarks. In addition to these, some extra variables were proposed by Peter Skands [88]. These “Peter

	σ_{sys}	$f(x)$
λ_0^2	2.0 %	b
$\lambda_{1/2}^1$	2.1 %	b
λ_1^1	1.1 %	$ax + b$
λ_2^1	0.4 %	$ax + b$
λ_0^0	7.5 %	b

Table 5.6: Systematic errors for the generalized angularities for charged tracks. The last column, $f(x)$, denotes which fit the systematic error is based on.

	σ_{sys}	$f(x)$
λ_0^2	4.2 %	$ax + b$
$\lambda_{1/2}^1$	4.2 %	$ax + b$
λ_1^1	3.9 %	$ax + b$
λ_2^1	1.6 %	$ax + b$
λ_0^0	2.0 %	b

Table 5.7: Systematic errors for the generalized angularities for neutral clusters. The last column, $f(x)$, denotes which fit the systematic error is based on.

Note that the number of selected events in the 4-jet case is much lower than in the 3-jet case and that the signal purity is lower as well, $\rho_g^{3\text{-jet}} = 98.77\%$ vs. $\rho_g^{4\text{-jet}} = 92.67\%$, at the selected cut value.

Skands"-variables are aimed at other probing other interesting areas of the phase space.

The gluon splitting variables are variables where current MC generators, such as Pythia, and Data show some discrepancies. Better measurements of these variables might lead to new theoretical insights that can reduce these discrepancies. The gluon splitting variables are:

Energy Asymmetry:

E_{diff} : The relative difference in energy between the gluon jet with the highest and lowest energy: $E_{\text{diff}} = \frac{E_{\text{max}} - E_{\text{min}}}{E_{\text{max}} + E_{\text{min}}}$.

$E_{\text{rel}_{\text{min}}}$: The relative energy of the gluon jet with the lowest energy and the sum: $E_{\text{rel}_{\text{min}}} = \frac{E_{\text{min}}}{E_{\text{max}} + E_{\text{min}}}$.

E_{rel} : The relative energy of the gluon jet with the lowest energy and the highest energy: $E_{\text{rel}} = \frac{E_{\text{min}}}{E_{\text{max}}}$.

Resolution Scale:

$\Delta\theta$: The angle between the two gluon jets.

m_{gg} : The invariant mass of the two gluon jets.

Azimuthal Angle

ϕ_{\parallel} : The angle between the plane spanned by the two b -jets and the plane spanned by the two gluon jets. This is the same angle as the angle between the b -jet cross product $\mathbf{p}_{b_1} \times \mathbf{p}_{b_2}$ and the gluon-jet cross product $\mathbf{p}_{g_1} \times \mathbf{p}_{g_2}$ where \mathbf{p} is the jet momentum.

Peter Skands:

$\ln(k_t^2/m_{\text{vis}}^2)$: Logarithm of the ratio between the k_t value, see equation (5.12), of the two gluon jets and the visible mass of the event.

$p_{\perp,\text{A}}^2$: The p_{perp} antenna defined as:

$$p_{\perp,\text{A}}^2 = \tilde{m}_{12}^2 \cdot \min \left(\frac{\min(\tilde{m}_{b1}^2, \tilde{m}_{b2}^2) - m_b^2}{\tilde{m}_{b12}^2 - m_b^2}, \frac{\min(\tilde{m}_{b1}^2, \tilde{m}_{b2}^2) - m_b^2}{\tilde{m}_{b12}^2 - m_b^2} \right), \quad (5.10)$$

where $\tilde{m}^2 = m^2 \cdot m_Z^2 / m_{\text{vis}}^2$ and m_Z is the mass of the Z boson.

The gluon splitting variables will be analyzed in distinct regions defined by the k_t [34, 45] and Cambridge/Aachen (CA) [43, 104] jet clustering algorithms for e^+e^- collisions which will first be described. The two algorithms uses the following²⁷ jet distance measure:

$$d_{ij}^2(p) = \min \left(E_i^{2p}, E_j^{2p} \right) \left(\frac{1 - \cos \theta_{ij}}{2} \right), \quad (5.11)$$

²⁷ When using $R = 1$ in eq. (9a) in Ref. [33].

where E is the (pseudo)jet energy and θ_{ij} is the angle between (pseudo)jet i and j [33]. For $p = 1$ equation (5.11) is called the k_t algorithm and the Cambridge/Aachen for $p = 0$. Both the k_t and CA algorithms are newer jet clustering algorithms than JADE, see section 4.4, and their distance measures are also pretty similar to JADE's. The value $d_{ij}(p = 1)$ is also known as the k_t :

$$k_t = \sqrt{\min(E_i^2, E_j^2) \left(\frac{1 - \cos \theta_{ij}}{2} \right)}. \quad (5.12)$$

Based on the two algorithms, we define the ratio R_{gg} between the two gluon jets d_{gg}^2 and the lowest value of d_{ij}^2 not including the two gluon jets:

$$R_{gg}(p) \equiv \frac{d_{gg}^2(p)}{\min_{(i,j) \neq (g,g)} d_{ij}^2(p)}. \quad (5.13)$$

We further define $R_{gg}^{k_t} \equiv R_{gg}(p = 1)$ and $R_{gg}^{CA} \equiv R_{gg}(p = 0)$.

Since the CA algorithm is energy-independent and the k_t algorithm is not, they describe different parts of the phase space. One example of this is the case of soft wide gluon jets illustrated in Figure 5.28. Here the two gluon jets are low-energy (soft) jets but with a high angle between them (wide). In this case the k_t algorithm would probably cluster the two gluon jets together but the CA algorithm would not. This means that $R_{gg}^{k_t}$ would be less than 1 since the distance measure between the two gluon jets would be the smallest of them all, whereas to R_{gg}^{CA} would be larger than 1 since $d_{\bar{b}g} < d_{gg}$. In addition to the soft, wide angle gluon jets, we have the case of soft, collinear gluon jets which the two algorithms agree to cluster together, see Figure 5.29 or the opposite case where the two algorithms both agree on not to cluster the gluon jets together, see Figure 5.30. These figures are just illustrations of how to better understand the R_{gg} phase space.

The variables are computed on an event-by-event basis. From now on primarily events with with “good” g -tags, $\gamma_{tag} > 0.8$, are used. The efficiency of these signal events are measured in the following subsection.

5.9.2 Efficiencies

It is not possible to take advantage of the Tag-Tag-Probe method for whole events as it was for individual jets in the 3-jet case. As such, we are also unable to estimate the g -tagging efficiency in Data of the gluon splitting variables defined in the previous subsection. However, it is still possible to do so for MC using the truth labels. The efficiency $\epsilon_{gg}^{4\text{-jet}}$ for the gluon splitting variable E_{diff} is shown in Figure 5.31 for signal and background based on MC Truth. Note that even though the efficiency here is quite low, it is close to constant (as a function of E_{diff}).

The plot for the rest of the g -tagging efficiencies can be seen in Figure B.42–B.49.

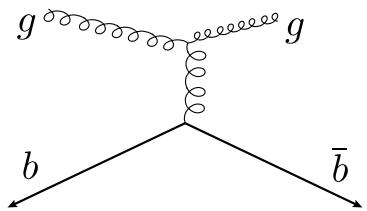


Figure 5.28: Soft, wide angle gluons in 4-jet events.

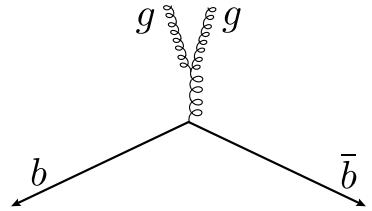


Figure 5.29: Soft, collinear gluons in 4-jet events.

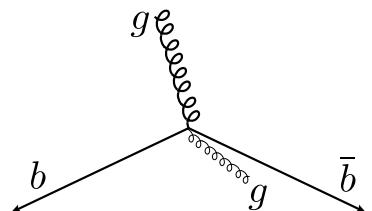


Figure 5.30: Hard, non $g \rightarrow gg$ gluons in 4-jet events.

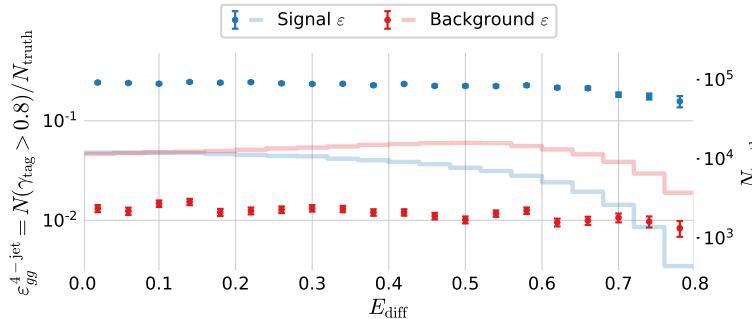


Figure 5.31: Efficiency of the g -tagging algorithm for 4-jet events as a function of normalized gluon-gluon jet energy difference (asymmetry) E_{diff} in MC. The efficiency is measured as the number of events with a g -tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for [signal events](#) according to MC Truth in blue and [background events](#) according to MC Truth in red.

5.9.3 Closure Test

I perform a closure test to validate the g -tagging model for the gluon variables in 4-jet events. Closure tests compare the developed method after corrections²⁸ to MC Truth. Any discrepancies can then be investigated and finally the closure test can gauge the systematic uncertainties of the analysis.

The closure test is based on the distribution of the gluon splitting variables, see [subsection 5.9.1](#), for events in the signal region or the so-called *sideband region*. The sideband region²⁹ is a region close to the signal region where the background is expected to behave approximately similar to the background in the signal region (and likewise for the signal). From Figure 5.21, the signal region was defined to be $0.8 < \gamma_{\text{tag}}$ for 4-jet events. This region is expected to contain primarily signal events. For 4-jet events, we define the sideband region to be $0.6 < \gamma_{\text{tag}} < 0.8$.

The aim is to fully reconstruct the true distribution $\mathcal{P}_{gg}(x)$ of any gluon splitting variable x . To first order, this is given by $\mathcal{P}_{gg} \approx \mathcal{P}_{\text{sig}}/\varepsilon_{gg}^{\text{4-jet}}$, where $\mathcal{P}_{\text{sig}}(x)$ is the distribution of x for all events in the signal region. However, this expression completely ignores the background events that are also found in the signal region. To correct for the assumption of no background, we introduce $\mathcal{P}_{\text{bkg}}(x)$ which is the distribution of x for background events:

$$\mathcal{P}_{gg} = \frac{\mathcal{P}_{\text{sig}} - \alpha \cdot \mathcal{P}_{\text{bkg}}}{\varepsilon_{gg}^{\text{4-jet}}}. \quad (5.14)$$

Here α is the fraction of background events in the signal region $N_{\text{bkg}}^{\text{sig}}$ relative to the background events in the sideband $N_{\text{bkg}}^{\text{side}}$. Letting f_{gg}^{sig} denote the fraction of signal in the signal region and $f_{\text{bkg}}^{\text{side}}$ the fraction of background in the sideband region, α is defined as:

$$\alpha = \frac{N_{\text{bkg}}^{\text{sig}}}{N_{\text{bkg}}^{\text{side}}} = \frac{(1 - f_{gg}^{\text{sig}}) \cdot N_{\text{sig}}}{f_{\text{bkg}}^{\text{side}} \cdot N_{\text{side}}}, \quad (5.15)$$

where N_i is the number of events in region i (either signal or sideband). The background distribution $\mathcal{P}_{\text{bkg}}(x)$ itself can be approximated to be $\mathcal{P}_{\text{bkg}} \approx \mathcal{P}_{\text{side}}$ if assuming no signal events in the sideband region, yet this assumption is also not satisfied and is thus

²⁸ E.g. Applying the efficiencies found in the previous subsection.

²⁹ Also sometimes known as the control region.

corrected for:

$$\mathcal{P}_{\text{bkg}} = \mathcal{P}_{\text{side}} - \beta \cdot \mathcal{P}_{gg} \cdot \varepsilon_{gg}^{4\text{-jet}}, \quad (5.16)$$

where β is the fraction of signal events in the sideband region relative to the signal events in the signal region and is defined as:

$$\beta = \frac{(1 - f_{\text{bkg}}^{\text{side}}) \cdot N_{\text{side}}}{f_{gg}^{\text{sig}} \cdot N_{\text{sig}}} \quad (5.17)$$

Plugging equation (5.16) into (5.14) and solving for \mathcal{P}_{gg} yields:

$$\mathcal{P}_{gg} = \frac{\mathcal{P}_{\text{sig}} - \alpha \cdot \mathcal{P}_{\text{side}}}{\varepsilon_{gg}^{4\text{-jet}} \cdot (1 + \alpha\beta)}. \quad (5.18)$$

The advantage of this equation is that only $\varepsilon_{gg}^{4\text{-jet}}$ and the two constants³⁰ f_{gg}^{sig} and $f_{\text{bkg}}^{\text{side}}$ depend on MC truth, the rest can be applied to data without any truth label. It should be made clear that the final expression, equation (5.18), is based on the assumption that the signal distribution of x is (approximately) similar in the signal and sideband regions $\mathcal{P}_{gg}^{\text{sig}} \approx \mathcal{P}_{gg}^{\text{side}}$ and likewise for the background distributions $\mathcal{P}_{\text{bkg}}^{\text{sig}} \approx \mathcal{P}_{\text{bkg}}^{\text{side}}$.

The distribution of \mathcal{P}_{gg} in MC is shown in Figure 5.32 together with \mathcal{P}_{sig} , $\mathcal{P}_{\text{side}}$, and the distribution for MC Truth $\mathcal{P}_{gg}^{\text{Truth}}$. In this figure the distributions are shown for the gluon splitting variable E_{diff} with histograms shown in the top plot and a ratio plot between \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ in the bottom part. By just looking at the distributions, it can be seen that the distributions in the signal and sideband regions follow each quite closely even though they start to differ at large values of E_{diff} . Similarly, also \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ follow each other quite closely, however, by looking at the ratio plot it can be seen that \mathcal{P}_{gg} generally has fewer counts in each bin than $\mathcal{P}_{gg}^{\text{Truth}}$.

³⁰ Which are found to be:
 $f_{gg}^{\text{sig}} = \rho_g^{4\text{-jet}} = 92.67\%$
 $f_{\text{bkg}}^{\text{side}} = 62.5\%$.

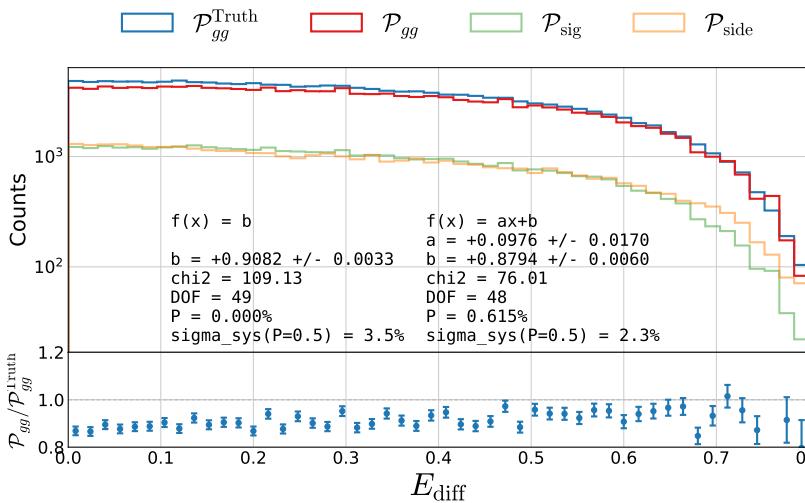


Figure 5.32: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for the normalized gluon gluon jet energy asymmetry E_{diff} . In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the \mathcal{P}_{gg} based on MC but without Truth in red, the distribution in the signal region \mathcal{P}_{sig} in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

The errorbars in the ratio plot are fitted with both a constant $f(x) = b$ and a straight line $f(x) = ax + b$ with the fit results shown as text in the plot. The closure plot for the rest of the gluon splitting variables in Figure B.50–B.57. In general it can be seen that the

\mathcal{P}_{gg} matches $\mathcal{P}_{gg}^{\text{Truth}}$ pretty well for all the gluon splitting variables, however, with a 10 % offset. This offset is easy to correct for with a constant correction factor.

However, we see that for the energy asymmetry variables and Δ_θ there are clear, linear trends in the ratio plots. This is seen e.g. for E_{diff} in Figure 5.32 where the χ^2 value for the constant fit is 109.13 but only 76.01 for the linear fit. This difference in χ^2 value cannot be attributed to just the linear fit being a more advanced fit in which case the difference would be expected to be $\Delta\chi^2 \sim 1$ due to difference in the number of fit parameters being 1. For these four variables the systematic uncertainty will thus be based on the linear fit, whereas it will be based on the constant for the rest of the variables.

The g -tagging algorithm perform well in general where most of the systematic attributed are 2 % to 3 %. The highest systematic uncertainty is for the invariant mass of the two gluons: $\sigma_{\text{sys}}(m_{gg}) = 3.6\%$. A summary of all the systematic errors from the closure tests can be seen in Table 5.8 for all of the gluon splitting variables.

The closure test shows that the g -tagging algorithm can also be trusted in the 4-jet case, however, with high systematic uncertainties for the m_{gg} variable.

5.9.4 Results

The comparison between the gluon splitting distributions will be done in four distinct regions of the $R_{gg}^{k_t}$ - R_{gg}^{CA} phase space. These four regions, A, B, C, and D, are defined in Table 5.9. Three of these regions have already been described, that is region A which is the soft, collinear region with an example shown in Figure 5.29, region C which is the soft, wide angle region in Figure 5.28, and the hard non- $(g \rightarrow gg)$ region D in Figure 5.30. Region B is a region where neither the k_t algorithm nor the CA algorithm is totally sure whether or not the two gluons should be clustered together.

The four regions are visualized in Figure 5.33 for both MC and Data. Here each dot shown in the scatter plot is an event with $0.8 < \gamma_{\text{tag}}$, totalling 41 117 events in the MC sample and 22 473 events in the data sample. The number of events that fall into each of the four regions are shown in the figure. The events are split up into these four regions because they each concern different physical interactions. The number of events in each region were taken into account when creating the ranges in Table 5.9 to minimize the statistical uncertainties.

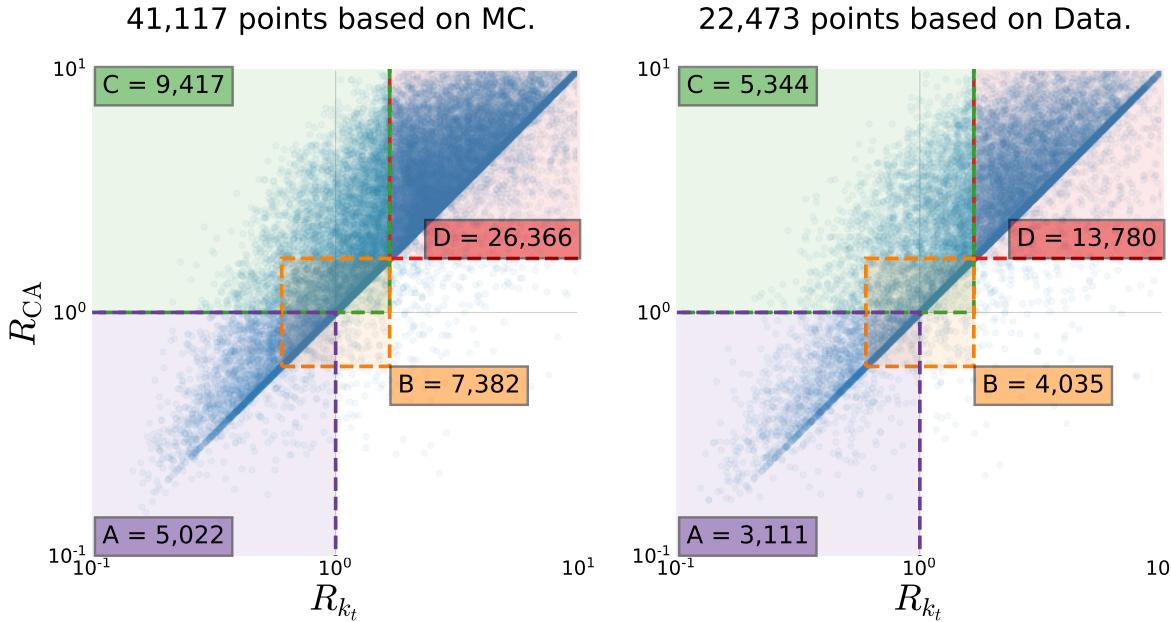
The distributions for the different gluons splitting variables for region A, the soft, collinear region, is shown in Figure 5.34 for both MC (scaled to Data) and Data. The in addition to scaling MC to Data, the samples have been efficiency corrected in the histograms, that is given a weight $w(x) = 1/\varepsilon_{gg}^{\text{4-jet}}(x)$. Region A contains 5022 events in the MC sample and 3111 in the Data samples. Generally the distributions match pretty well between the MC and Data, however, there seems to be small discrepancies in the energy asymmetry variables,

	σ_{sys}	$f(x)$
E_{diff}	2.3 %	$ax + b$
$E_{\text{rel,min}}$	2.3 %	$ax + b$
E_{rel}	2.5 %	$ax + b$
Δ_θ	2.0 %	$ax + b$
m_{gg}	3.6 %	b
ϕ_{\parallel}	1.7 %	b
$\ln(k_t^2/m_{\text{vis}}^2)$	2.1 %	b
$p_{\perp,A}^2$	3.1 %	b

Table 5.8: Systematic errors for the gluon splitting variables based on the closure test, see subsection 5.9.3. The last column, $f(x)$, denotes which fit the systematic error is based on.

Region	$R_{gg}^{k_t}$	R_{gg}^{CA}
A	$[0, 1]$	$[0, 1]$
B	$[\frac{3}{5}, \frac{5}{3}]$	$[\frac{3}{5}, \frac{5}{3}]$
C	$[0, \frac{5}{3}]$	$[1, \infty]$
D	$[\frac{5}{3}, \infty]$	$[\frac{5}{3}, \infty]$

Table 5.9: Definitions of the four regions in the $R_{gg}^{k_t}$ - R_{gg}^{CA} phase space.



however, this might just be due to statistical fluctuations (notice the low bin count in each bin).

When looking at the other regions, see Figure B.58–B.61, the energy discrepancies seem to disappear. To compare the similarity of the MC and Data distributions, I perform a Kolmogorov–Smirnov (KS) two-sample test. This nonparametric hypothesis test compares two samples based on the null hypothesis that they are both samples coming from the same (but unknown) parent distribution. The KS test outputs the probability P_{KS} of the null hypothesis. I add this probability to the plot if it is less than 100 %. This is only the case in region D, Figure B.61, which I suspect is due to low statistics in the other regions as the KS test is known for its low power [72].

To properly test the assumption that the variations in E_{diff} are due to statistical noise, I make a ratio plot between MC and Data and compare them between region A and D. The plot for region A is shown in Figure 5.35. When looking at the ratio plot in the bottom part of the figure, it can be seen that the almost all of the bins are with 1σ of 1. Indeed, when calculating the χ^2 value between the errorbars and a constant ratio of 1, that is to say $f(x) = 1$, one finds a χ^2 probability of $P_{\chi^2} = 51.3\%$. This value strongly suggests that the fluctuations in the two histograms are due to statistical noise (at least we cannot disprove that it is not). Remember, that this was not even a fit, but simply a comparison to a (fixed) constant.

In Figure 5.36 one can compare the ratio in region A with the one in region D. Even though region D contains a lot more data, the same pattern is seen as for region A. In region D the χ^2 probability is even a bit too high which indicates too little variation between the two histograms, however, not by too much to just happen by chance. In comparison to the χ^2 probability, the KS probability between the two samples is $P_{\text{KS}}(E_{\text{diff}}) = 0.25\%$. This difference in probability is an

Figure 5.33: Overview of the four regions, A, B, C, and D, in the $R_{gg}^{k_t}$ - R_{gg}^{CA} phase space. The regions are shown as colored rectangles together with a scatter plot showing the 2D-distribution for signal gluon events ($0.8 < \gamma_{\text{tag}}$). The left plot shows is for MC and the right one for Data.

Cut region A with 5022 MC events and 3111 data events.

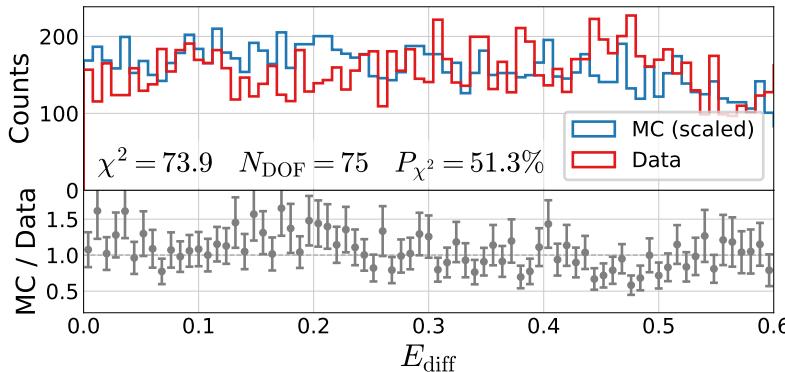
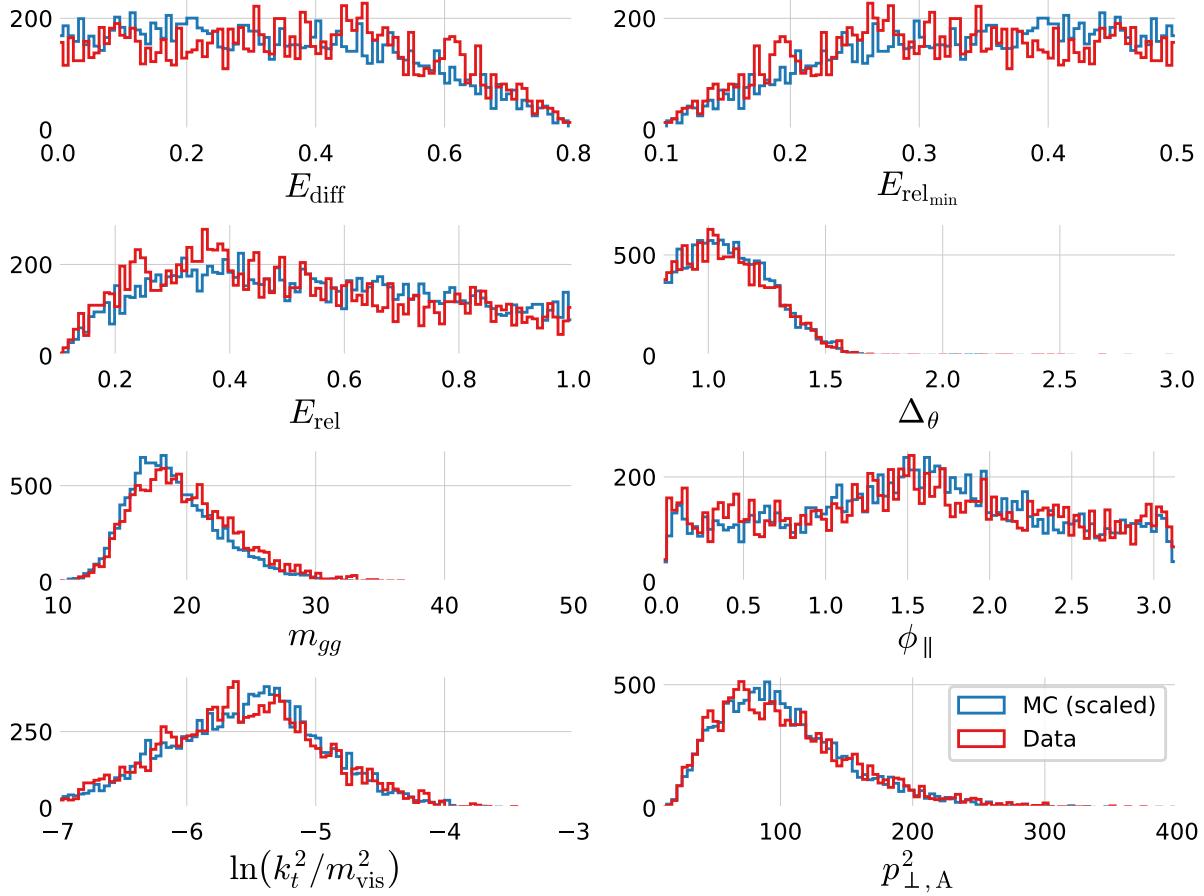


Figure 5.34: Comparison of the gluon splitting distributions in MC and Data for $E_{\text{diff}}^{k_t, \text{RCA}}$ phase space region A, see Figure 5.35. The distribution for MC splitting variable E_{diff} in MC and Data (scaled to Data) is shown in blue and for Data in red. Both MC and Data has the distribution of E_{diff} for MC (scaled to Data) is shown in blue and for Data distributions are for the $E_{\text{diff}}^{k_t, \text{RCA}}$ region. Both MC and Data has region A, which has 5022 events in the MC sample and 3111 in the Data sample. In text is written the χ^2 between the errorbars and a constant ratio equal to 1. That is to say that $f(x) = 1$, which is thus not a fit. The χ^2 probability is written as P_{χ^2} and it indicates an almost perfect correspondence between Data and MC.

implication of the fact that the KS test and χ^2 test are sensitive to different characteristics, e.g. the χ^2 test is only sensitive to absolute differences between distributions, whereas the KS test is not bound to only compare absolute differences.

In general there seems to be a high degree of correspondence between the distributions for MC and Data for all the gluon splitting variables in all four of the regions of phase space tested.

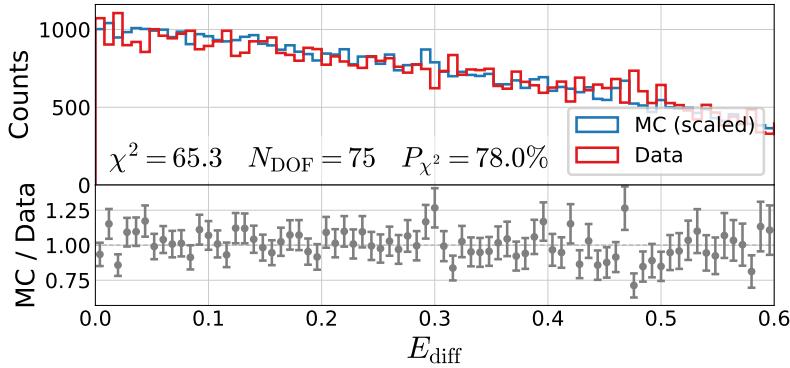


Figure 5.36: Ratio plot of the gluon splitting variable E_{diff} in MC and Data for $R_{gg}^{k_l} - R_{gg}^{\text{CA}}$ in region D. In the top plot, the distribution of E_{diff} for MC (scaled to Data) is shown in blue and for Data in red. Both MC and Data has been efficiency corrected. Below are the ratio plot between MC and Data. In text is written the χ^2 between the errorbars and a constant ratio equal to 1. That is to say that $f(x) = 1$, which is thus not a fit. The χ^2 probability is written as P_{χ^2} and it indicates an high correspondence between Data and MC, almost too high.

5.10 Discussion

The distributions of both the generalized angularities and the gluon splitting variables are based on the g -tagging model which itself is based on the b -tagging model. At first this way of training two different ML models, where one is trained on the other³¹, might seem a bit counterintuitive: why not just build a single, combined model that is able to extract gluons? The reason why the method used here is set up the way it is, is to be able to better understand the intermediate steps while also being able to compare the results to others. With the b -tagging model we are able to compare our model to the neural network trained by ALEPH (NNB) and measure the b -tagging efficiency for both gluons and b -quarks individually. Neither of these would not have been possible with a combined model which just outputs single gluons.

Regarding the NNB model, it is interesting to compare the b -tagging ROC curves in Figure 5.9 for 4-jets and B.7 for 3-jets. Even though the performance for the old NNB model is comparable to our model³², especially for 3-jet events, one has to take into account that the NNB model is trained on nine variables compared to only three variables by our model. Yet, remember that this is a model that was trained more than 25 years ago by now. I doubt many other areas of science implemented machine models that long ago with a performance that is almost comparable to today's.

That we only train on the vertex variables is to reduce any potential bias that might be introduced when training on the shape variables (in addition to the vertex variables) and then afterwards used to compare the same shape variables between MC and Data afterwards. Another approach to this would be to train on all variables and then afterwards quantify the potential bias and correct for it. This would also allow one to compare the loss in performance by only using the vertex variables. This was not done in this project due to time limitations.

As mentioned above, with the modularized approach we apply in this project we are able to measure the b -tagging efficiencies not only in MC but also in Data. The efficiencies were measured using a tag-tag-probe method in 3-jet events. In 4-jet events this problem

³¹ This is very different to the ensemble model used in section 3.8 which can basically be treated as a single, advanced ML model.

³² Technically models, both the XGBoost model and the LightGBM model.

is a lot harder, however, it might have been possible to apply a tag-tag-probe method, although the results of this are very likely to be limited by large uncertainties due to much lower statistics in 4-jet events compared to 3-jet events.

Some sources of error in this project not yet mentioned are the fact that we do not have access to the full MC simulation and that sometimes particles do not get detected in the detector. That we do not have access to the full MC simulation means that we have to q -matching to the final quarks; we do not know the real truth for the jets, even in MC. On the other hand, particles going undetected is something that also happens in Data. When this happens in e.g. a 4-jet event it will be classified as a 3-jet event in the data which it was clearly not originally.

Even after taking the sources of error into account along with the statistical and systematic errors, the result would still not reflect Nature. This is due to the fact that the observations $g(y)$ are a combination of the truth in Nature $f(x)$ combined with the detector $A(x, y)$. This combination is mathematically a convolution [84]:

$$\int f(x)A(x, y) dx = g(y). \quad (5.19)$$

The goal is thus to estimate truth $f(x)$ given knowledge about the detectors response $A(x, y)$. Another way to look at this is that $f(x)$ is smeared, or distorted, by detector and binning effects (e.g. finite resolutions). The process of inverting equation (5.19) is known as *unfolding* and is basically a deconvolution process.

5.11 Conclusion

blabla

B. Quarks vs. Gluons Appendix

B.1 Fraction of Different Types of Jets

	b	c	uds	g	non- q -matched
2	37.2 %	12.9 %	29.1 %	0.0 %	20.7 %
3	22.6 %	8.9 %	19.7 %	31.2 %	17.5 %
4	14.6 %	7.0 %	15.0 %	45.1 %	18.3 %
5	10.0 %	5.7 %	12.2 %	52.5 %	19.6 %
6	7.1 %	4.4 %	8.8 %	54.4 %	25.2 %

Table B.1: Number of different types of jets for MC and MCb written in relative numbers such that each row sum to 100 %. See also Table 5.3.

B.2 Random Search PDFs for XGB

Hyperparameter	Range
<code>subsample</code>	$\mathcal{U}(0.4, 1)$
<code>colsample_bytree</code>	$\mathcal{U}_{\text{trunc}}(0.4, 1, 2)$
<code>max_depth</code>	$\mathcal{U}_{\text{int}}(1, 20)$
<code>min_child_weight</code>	$\mathcal{U}_{\text{int}}(0, 10)$

Table B.2: Probability Density Functions for the random search hyperparameter optimization process for the XGBoost model.

B.3 Global SHAP Feature Importances for the g -Tagging Models in 3-Jet Events

β_{tag_i}	Energy Ordered	Shuffled	All Permutations
1	0.827 ± 0.006	0.924 ± 0.006	0.923 ± 0.006
2	0.749 ± 0.006	0.909 ± 0.006	0.918 ± 0.005
3	1.198 ± 0.006	0.878 ± 0.005	0.906 ± 0.005

Table B.3: Global SHAP feature importances $\phi_{\beta_i}^{\text{tot}}$ for the three g -Tagging Models in 3-Jet Events. Each $\phi_{\beta_i}^{\text{tot}}$ is shown for the three methods in the columns and the three b -tags as variables in the rows.

B.4 UMAP Parameters Comparison

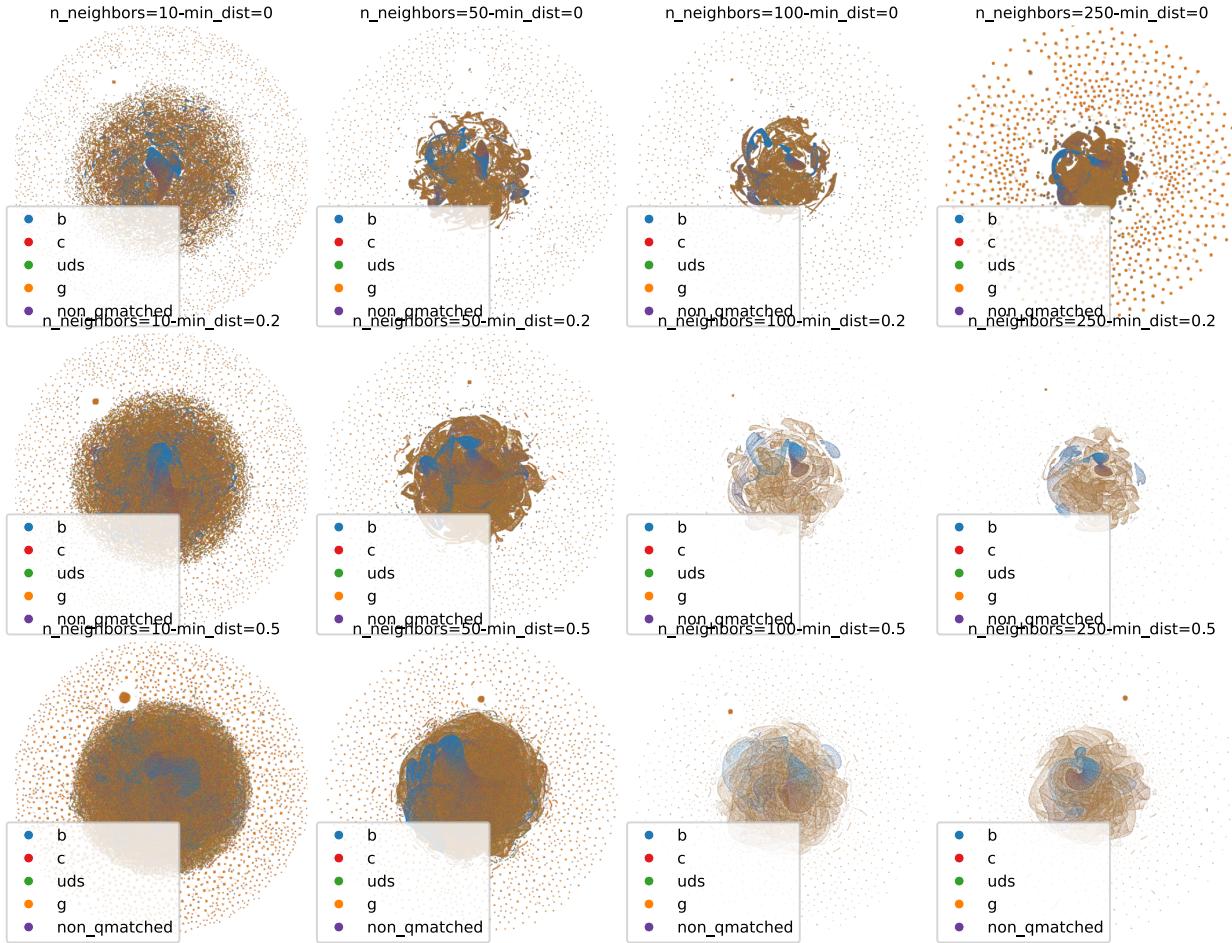


Figure B.1: Grid search of the two parameters `n_neighbors` and `min_dist` for the UMAP algorithm run on 4-jet events. For an explanation of these, see section 5.2.

B.5 t-SNE Parameters Comparison

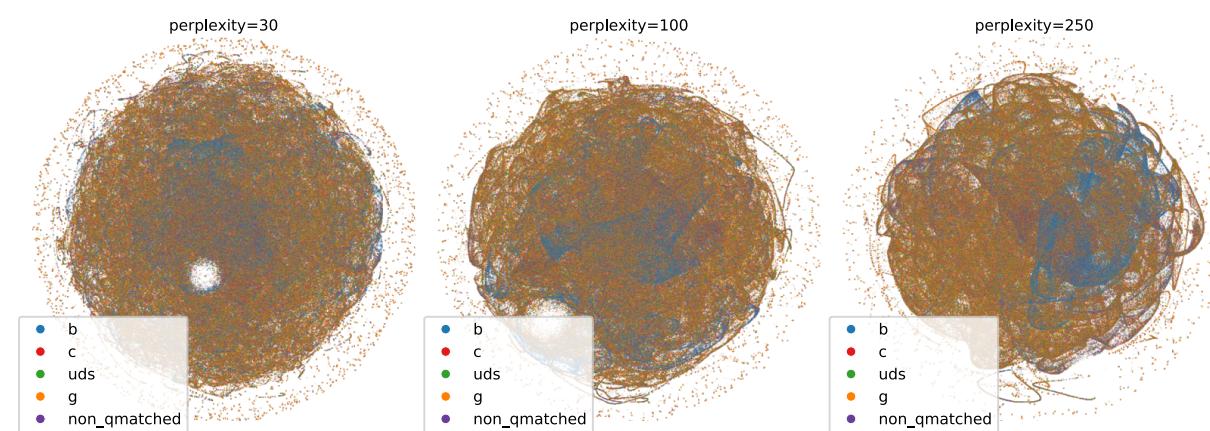


Figure B.2: Visualization of the t-SNE algorithm as a function of the `perplexity` parameters for 4-jet events.

B.6 HPO results for 3-jet b-Tagging

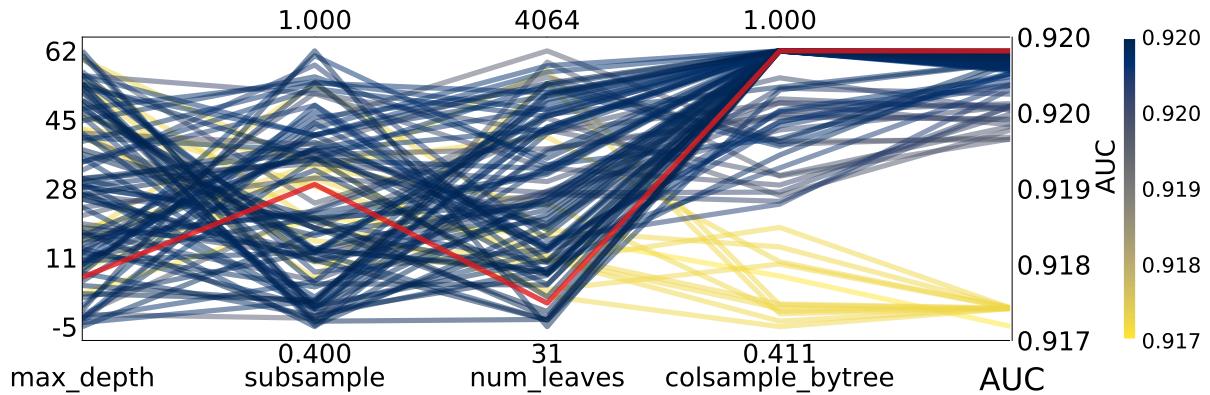


Figure B.3: Hyperparameter optimization results of b -tagging for 3-jet events. The results are shown as parallel coordinates with each hyperparameter along the x -axis and the value of that parameter on the y -axis. Each line is an event in the 4-dimensional space colored according to the performance of that hyperparameter as measured by AUC from highest AUC in dark blue to lowest AUC in yellow. The **single best hyperparameter** is shown in red.

B.7 HPO results for 4-jet b-Tagging

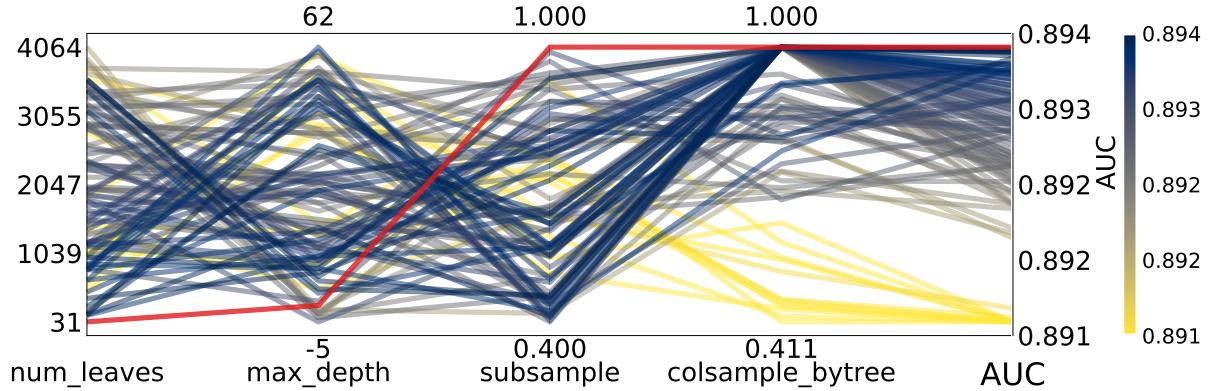


Figure B.4: Hyperparameter optimization results of b -tagging for 4-jet events. The results are shown as parallel coordinates with each hyperparameter along the x -axis and the value of that parameter on the y -axis. Each line is an event in the 4-dimensional space colored according to the performance of that hyperparameter as measured by AUC from highest AUC in dark blue to lowest AUC in yellow. The **single best hyperparameter** is shown in red.

B.8 b -Tag Scores in 3-Jet Events

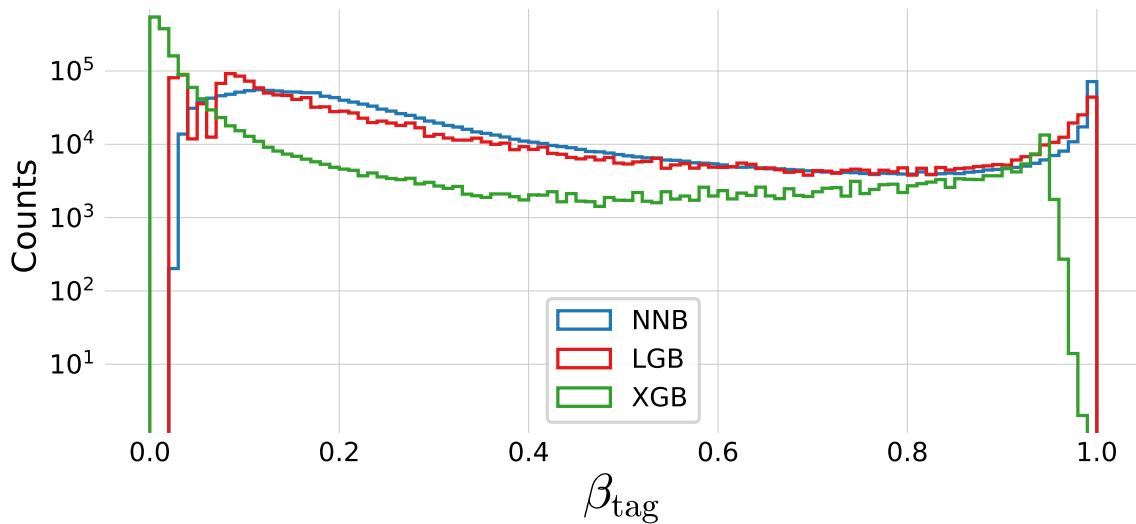


Figure B.5: Histogram of b -tag scores β_{tag} in 3-jet events for **NNB** (the neural network pre-trained by ALEPH, also called `nnbjet`) in blue, **LGB** in red, and **XGB** in green.

B.9 b -Tag Scores in 4-Jet Events

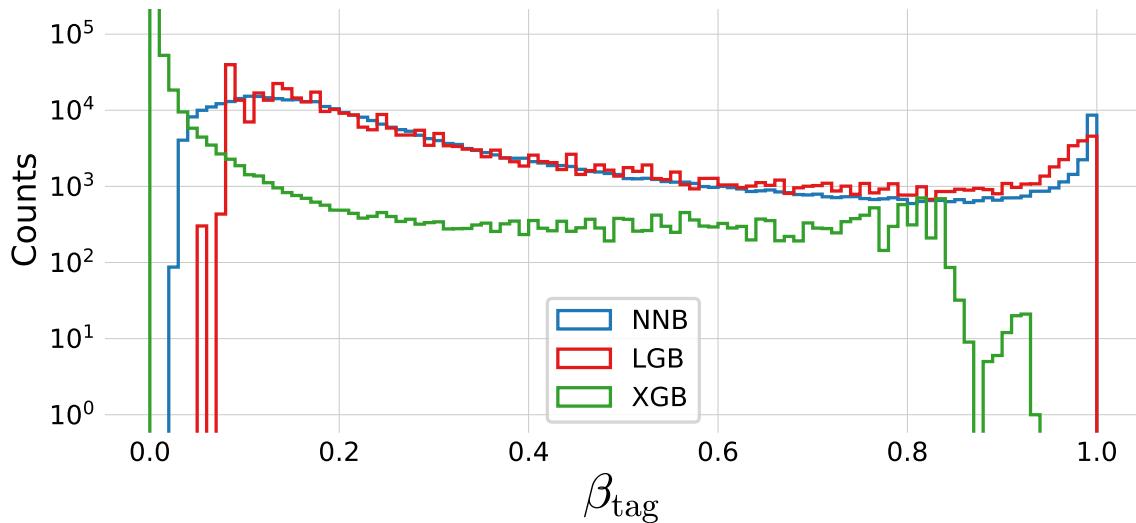


Figure B.6: Histogram of b -tag scores β_{tag} in 4-jet events for **NNB** (the neural network pre-trained by ALEPH, also called `nnbjet`) in blue, **LGB** in red, and **XGB** in green.

B.10 ROC curve for 3-jet b -tagging

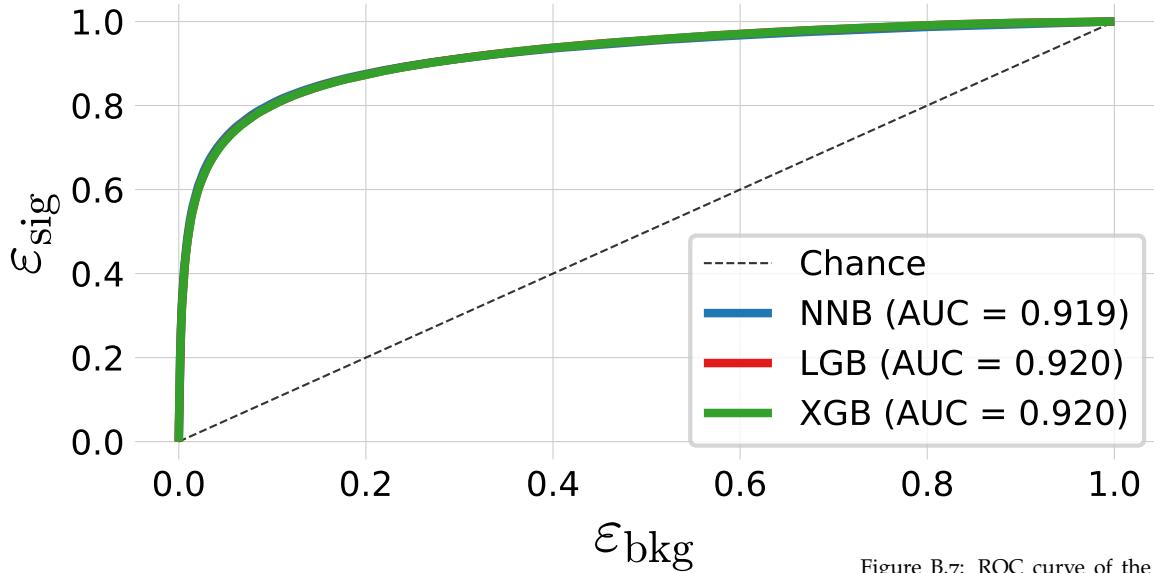


Figure B.7: ROC curve of the three b -tag models in 3-jet events for **NNB** (the pre-trained neural network trained by ALEPH, also called `nnbjet`) in blue, **LGB** in red, and **XGB** in green. In the legend the area under curve (AUC) is also shown. Notice that the LGB and XGB models share performance and it is thus due to overplotting that only the green line for XGB can be seen.

B.11 ROC curve for 4-jet b -tagging

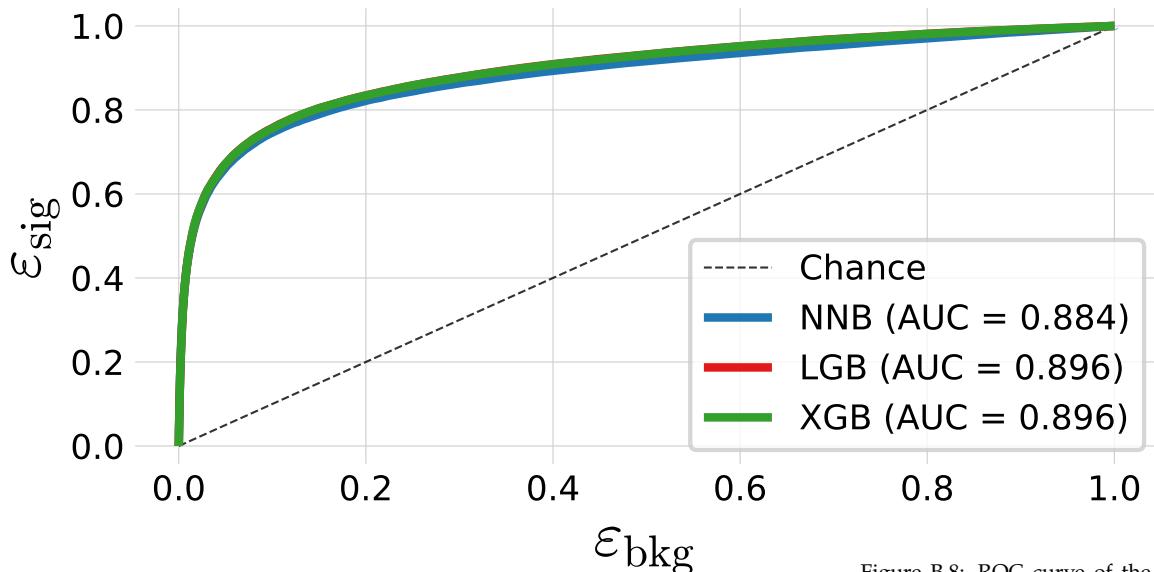


Figure B.8: ROC curve of the three b -tag models in 4-jet events for **NNB** (the pre-trained neural network trained by ALEPH, also called `nnbjet`) in blue, **LGB** in red, and **XGB** in green. In the legend the area under curve (AUC) is also shown. Notice that the LGB and XGB models share performance and it is thus due to overplotting that only the green line for XGB can be seen.

B.12 Distribution of b -Tags in 3-Jet Events

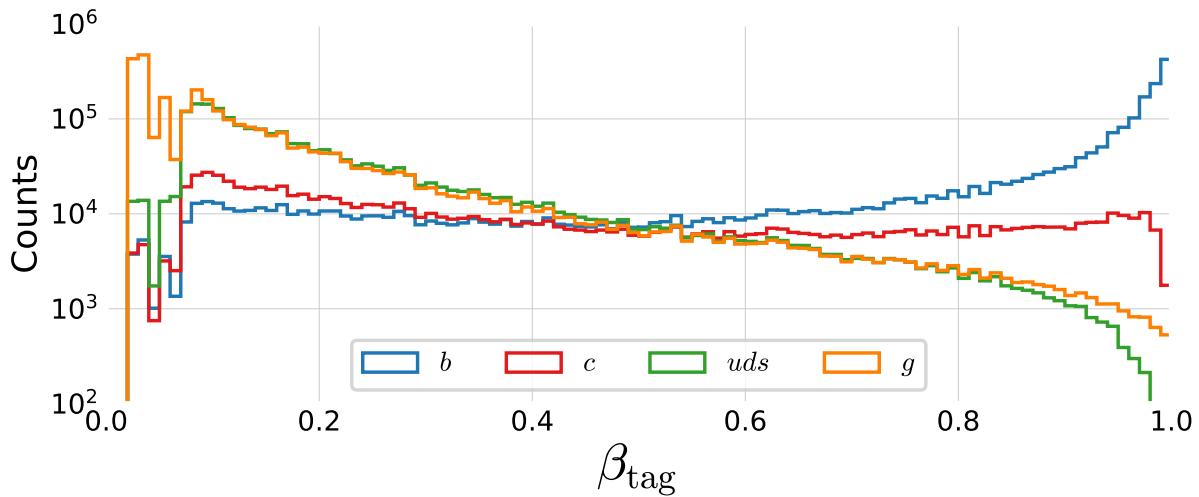


Figure B.9: Distribution of b -tags in 3-jet events for b -jets in blue, c -jets in red, uds in green and g in orange.

B.13 Distribution of b -Tags in 4-Jet Events

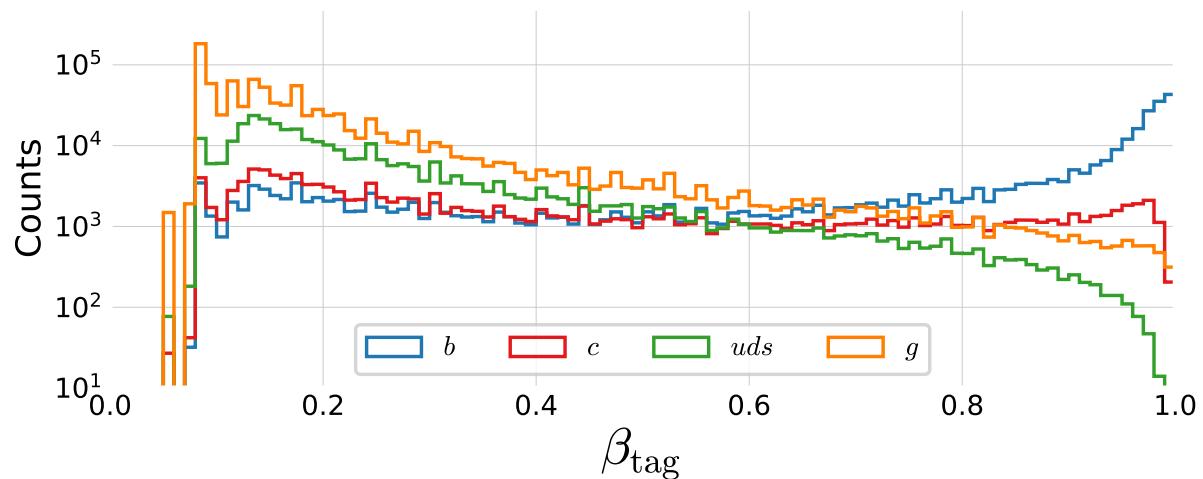


Figure B.10: Distribution of b -tags in 4-jet events for b -jets in blue, c -jets in red, uds in green and g in orange.

B.14 Global Feature Importances for the LGB b-Tagging Algorithm on 3-Jet Events

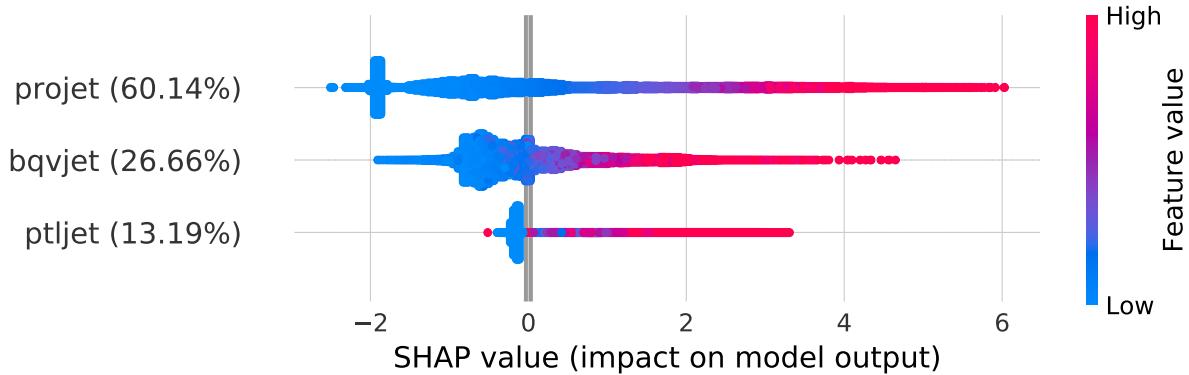


Figure B.11: Global feature importances for the LGB b -tagging algorithm on 3-jet events. The normalized feature importance is shown in the parenthesis and the each dot is an observation showing the dependance between the SHAP value and the feature's value.

B.15 Global Feature Importances for the LGB b-Tagging Algorithm on 4-Jet Events

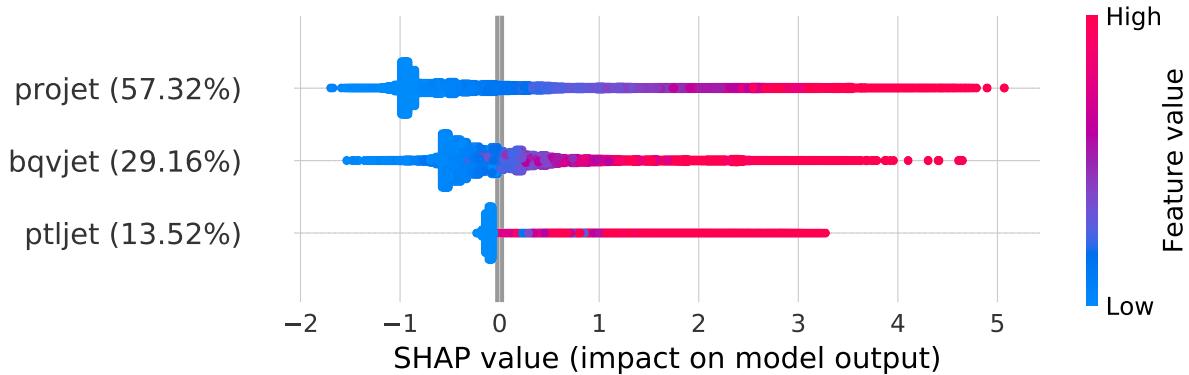


Figure B.12: Global feature importances for the LGB b -tagging algorithm on 4-jet events. The normalized feature importance is shown in the parenthesis and the each dot is an observation showing the dependance between the SHAP value and the feature's value.

B.16 b -Tagging Efficiency $\varepsilon_b^{b\text{-sig}}$ as a Function of Jet Energy

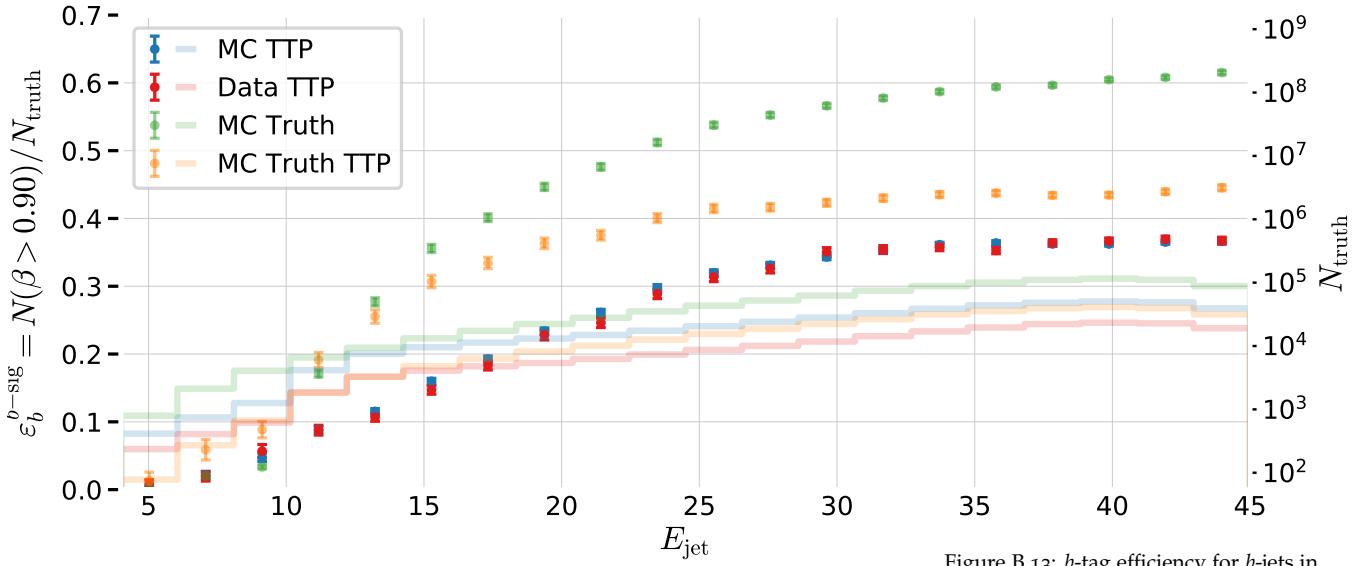


Figure B.13: b -tag efficiency for b -jets in the b -signal region for 3-jet events, $\varepsilon_b^{b\text{-sig}}$, as a function of jet energy E_{jet} . In the plot the efficiencies are shown for MC TTP in blue, Data TTP in red, MC Truth TTP in green, and MC Truth in orange. The efficiencies (the errorbars) can be read off on the left y -axis and the counts (histograms) on the right y -axis.

B.17 b -Tagging Efficiency $\varepsilon_g^{g\text{-sig}}$ as a Function of Jet Energy

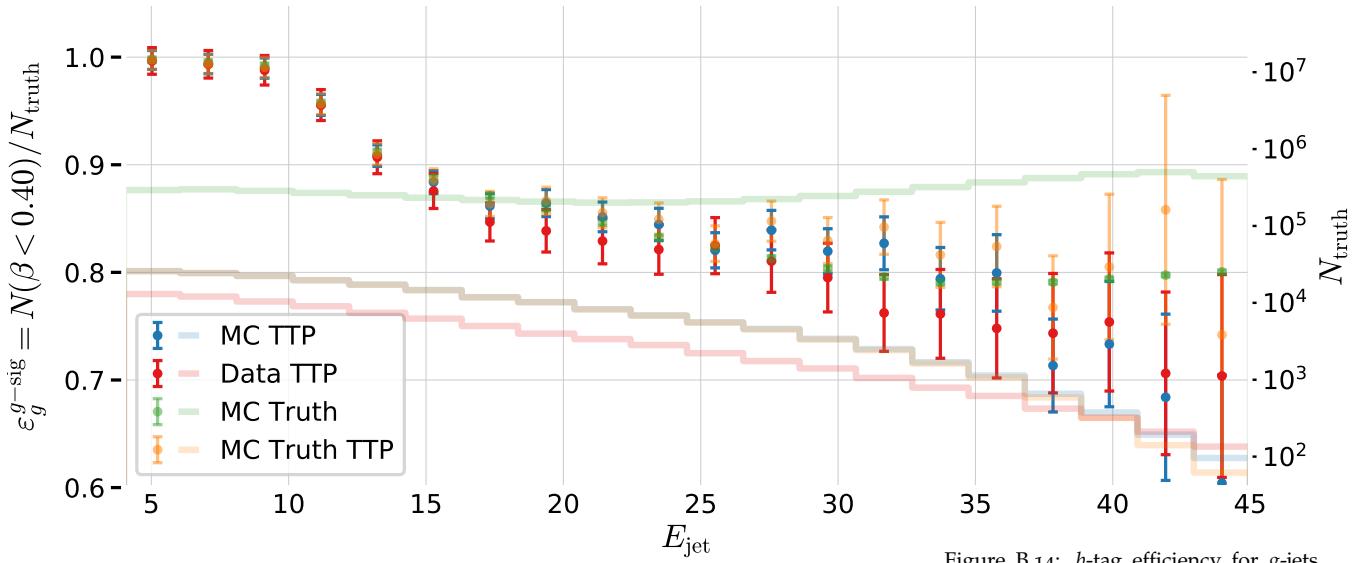


Figure B.14: b -tag efficiency for g -jets in the g -signal region for 3-jet events, $\varepsilon_g^{g\text{-sig}}$, as a function of jet energy E_{jet} . In the plot the efficiencies are shown for MC TTP in blue, Data TTP in red, MC Truth TTP in green, and MC Truth in orange. The efficiencies (the errorbars) can be read off on the left y -axis and the counts (histograms) on the right y -axis.

B.18 b -Tagging Efficiency $\varepsilon_b^{g\text{-sig}}$ as a Function of Jet Energy

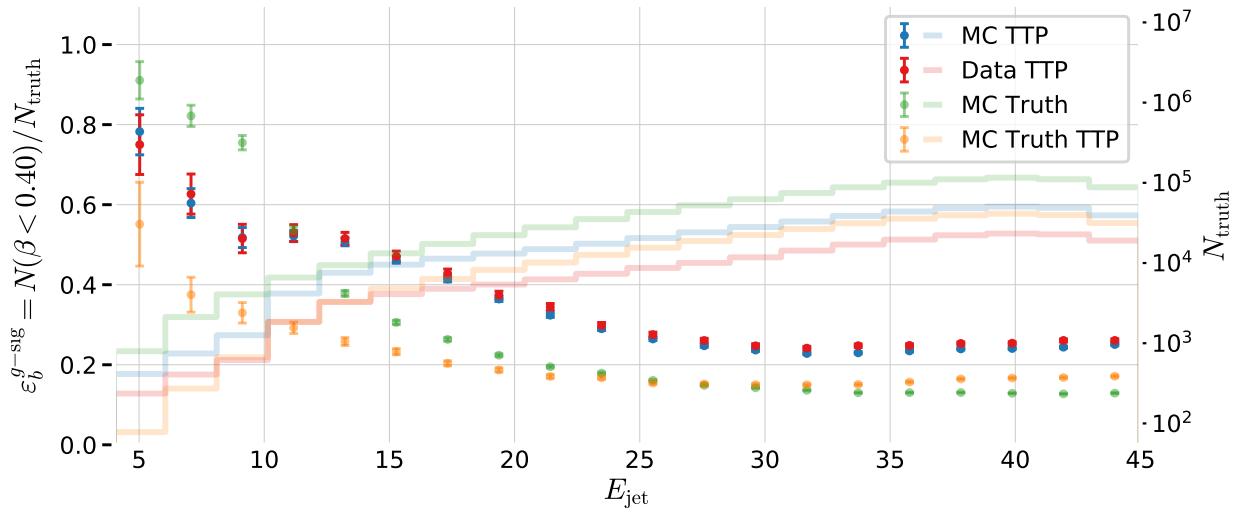


Figure B.15: b -tag efficiency for b -jets in the g -signal region for 3-jet events, $\varepsilon_b^{g\text{-sig}}$, as a function of jet energy E_{jet} . In the plot the efficiencies are shown for MC TTP in blue, Data TTP in red, MC Truth in green, and MC Truth TTP in orange. The efficiencies (the errorbars) can be read off on the left y -axis and the counts (histograms) on the right y -axis.

B.19 b -Tagging Efficiency $\varepsilon_g^{b\text{-sig}}$ as a Function of Jet Energy

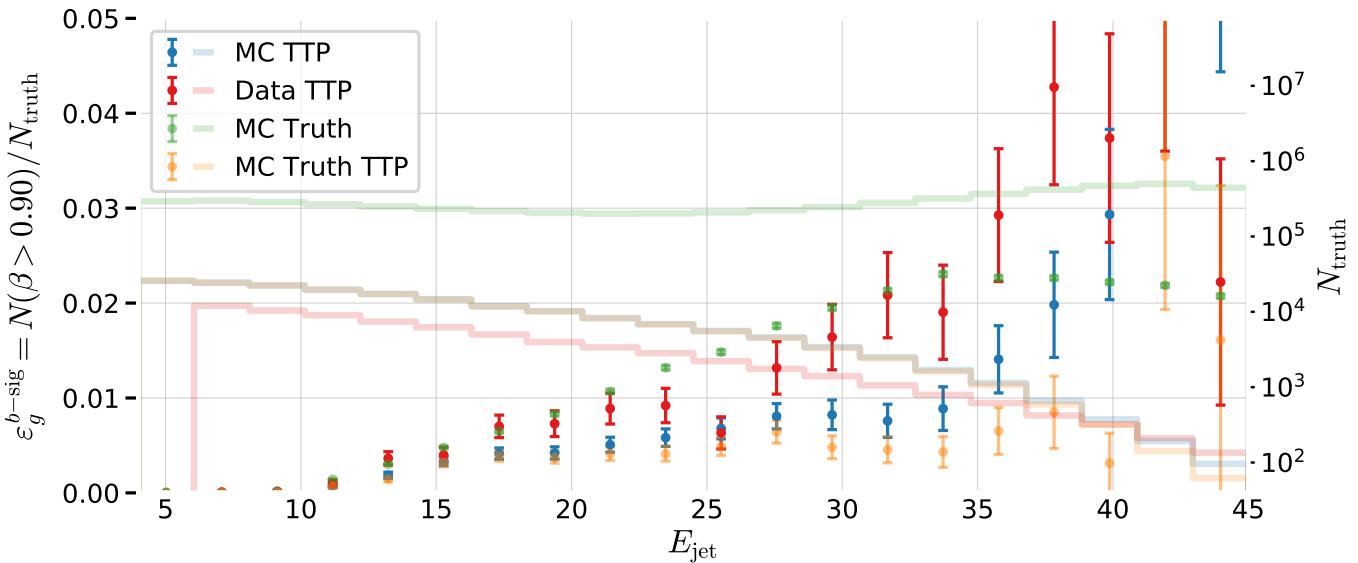


Figure B.16: b -tag efficiency for g -jets in the b -signal region for 3-jet events, $\varepsilon_g^{b\text{-sig}}$, as a function of jet energy E_{jet} . In the plot the efficiencies are shown for MC TTP in blue, Data TTP in red, MC Truth in green, and MC Truth TTP in orange. The efficiencies (the errorbars) can be read off on the left y -axis and the counts (histograms) on the right y -axis.

B.20 Parallel Coordinate Plot of HPO Results for 3-Jet g-Tagging for Energy Ordered Jets

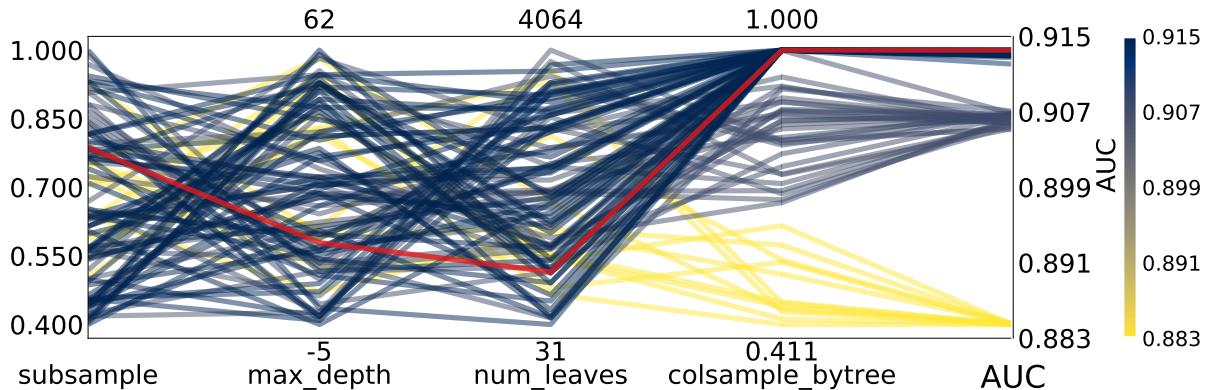


Figure B.17: Hyperparameter optimization results of g -tagging for 3-jet events for energy ordered jets shown as parallel coordinates.

B.21 Parallel Coordinate Plot of HPO Results for 3-Jet g-Tagging for Shuffled Jets

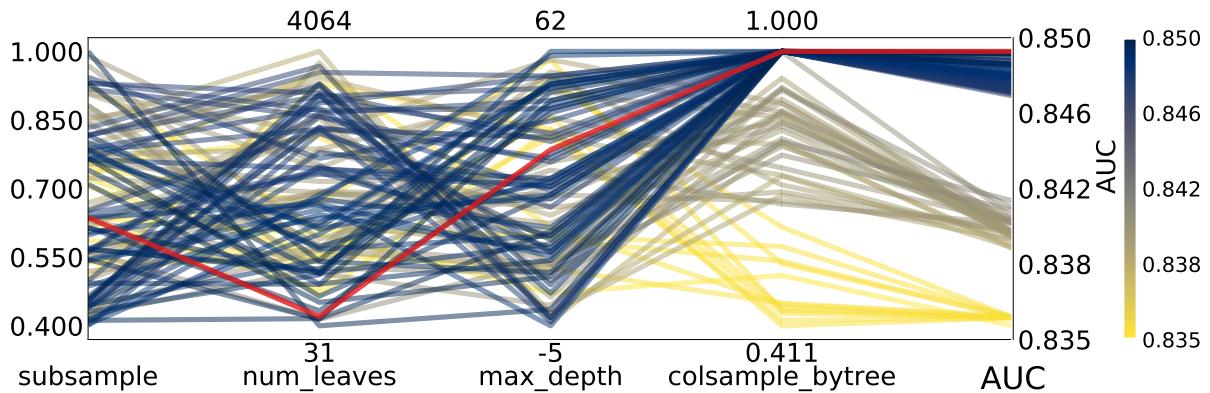


Figure B.18: Hyperparameter optimization results of g -tagging for 3-jet events for (row) shuffled jets shown as parallel coordinates.

B.22 Parallel Coordinate Plot of HPO Results for 4-Jet g -Tagging for Energy Ordered Jets

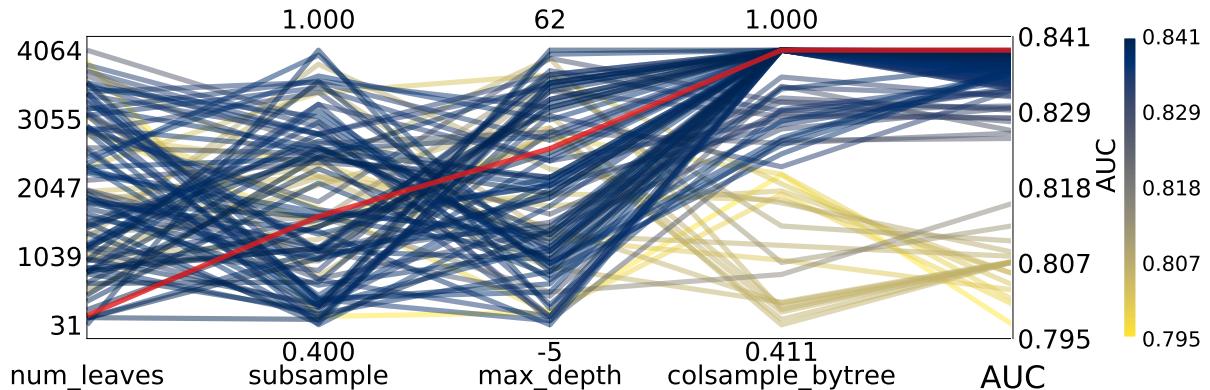


Figure B.19: Hyperparameter optimization results of g -tagging for 4-jet events for energy ordered jets shown as parallel coordinates.

B.23 Parallel Coordinate Plot of HPO Results for 4-Jet g -Tagging for Shuffled Jets

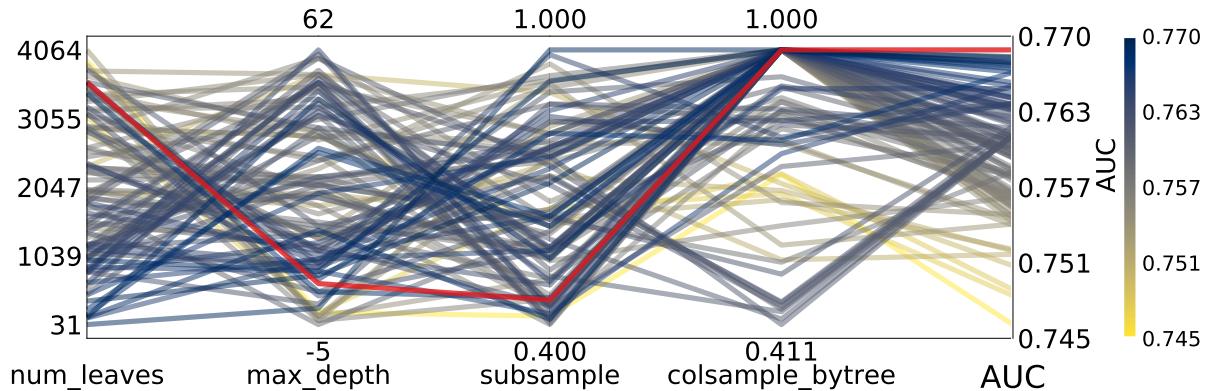


Figure B.20: Hyperparameter optimization results of g -tagging for 4-jet events for (row) shuffled jets shown as parallel coordinates.

B.24 PermNet Architecture

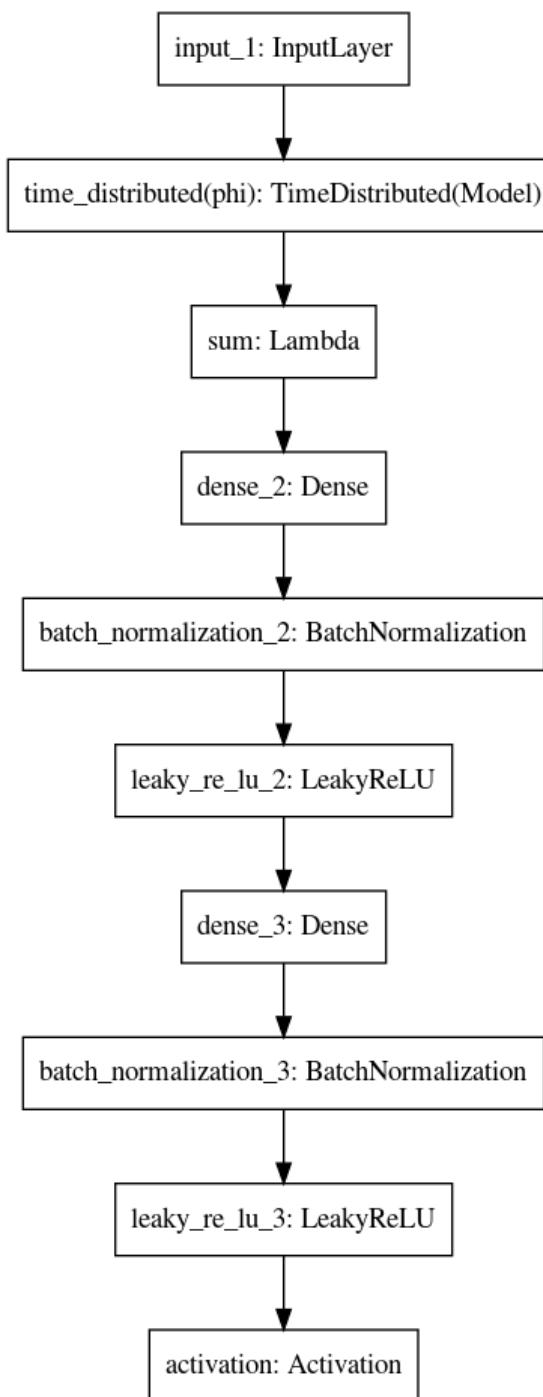
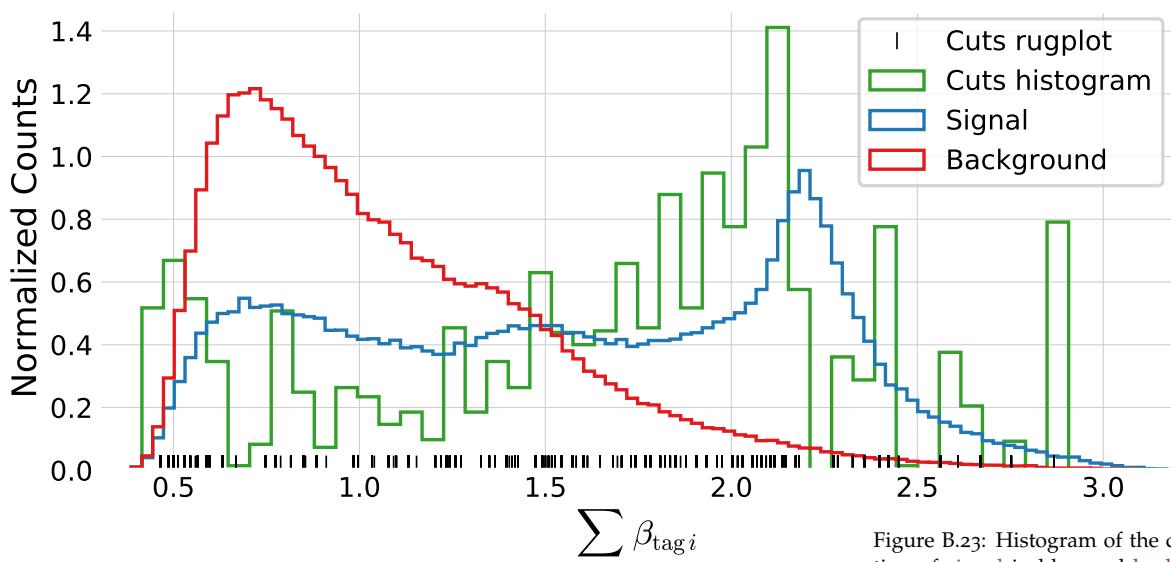
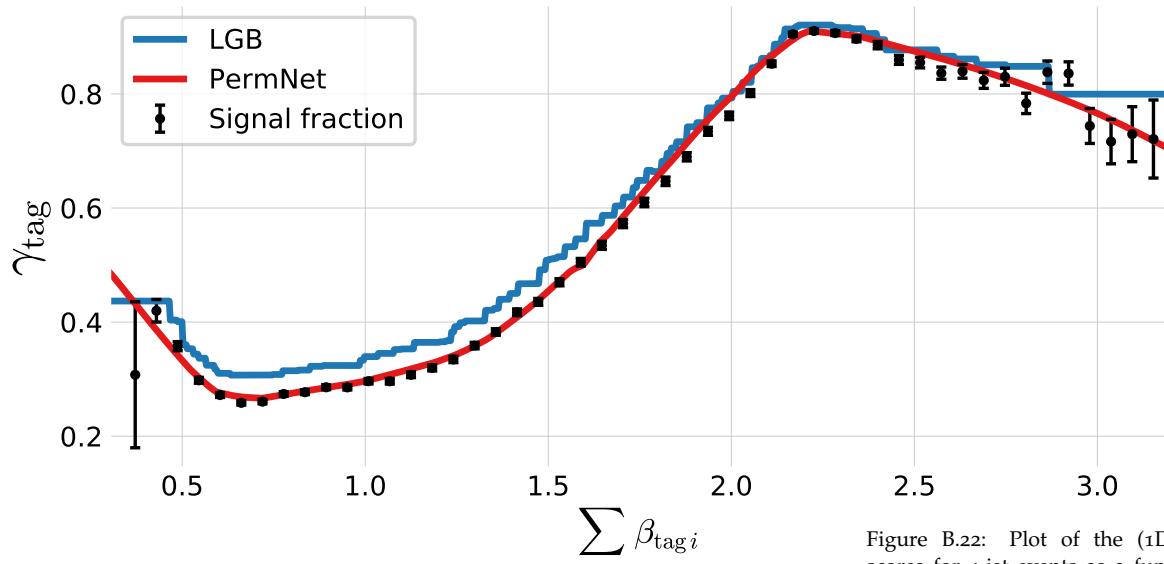
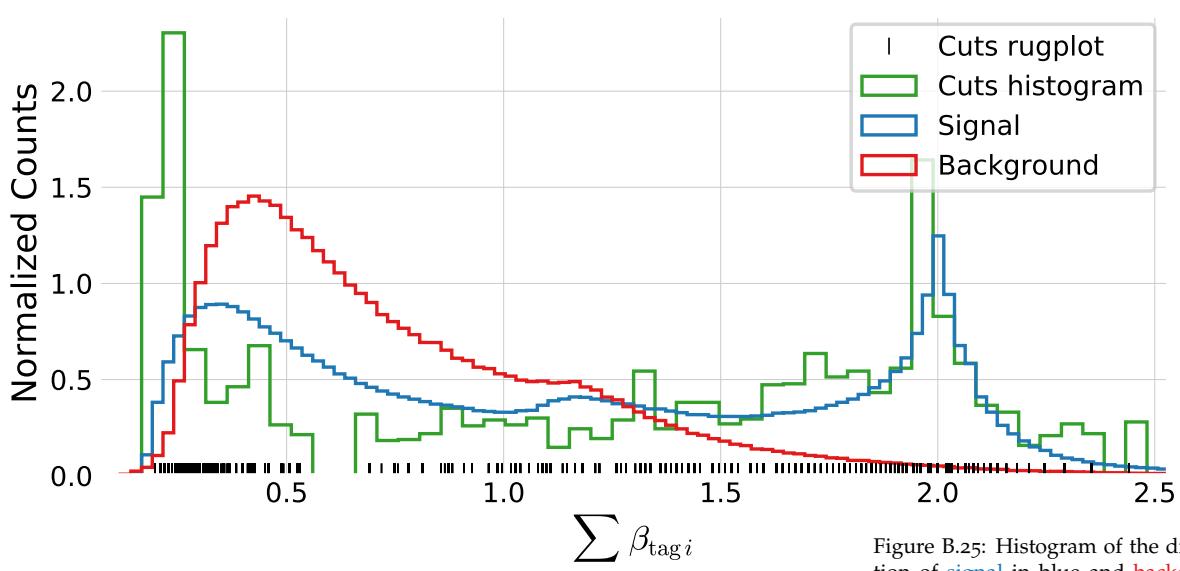
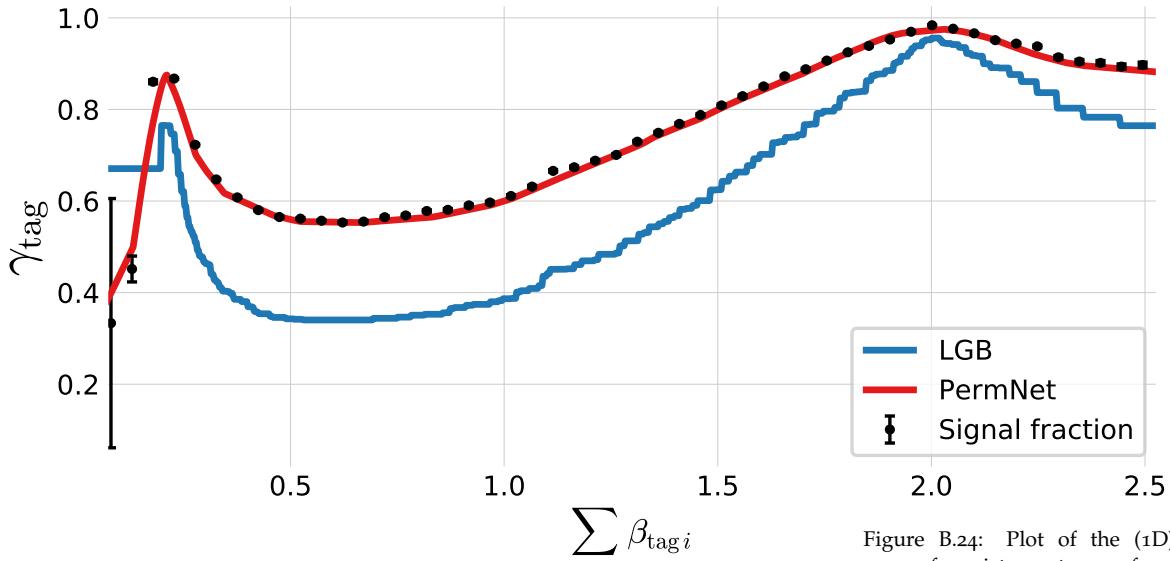


Figure B.21: Architecture of the PermNet neural network.

B.25 1D LGB and PermNet Comparison for 4-Jet Events



B.26 1D LGB and PermNet Comparison for 3-Jet Events



B.27 g -Tag Scores in 4-Jet Events

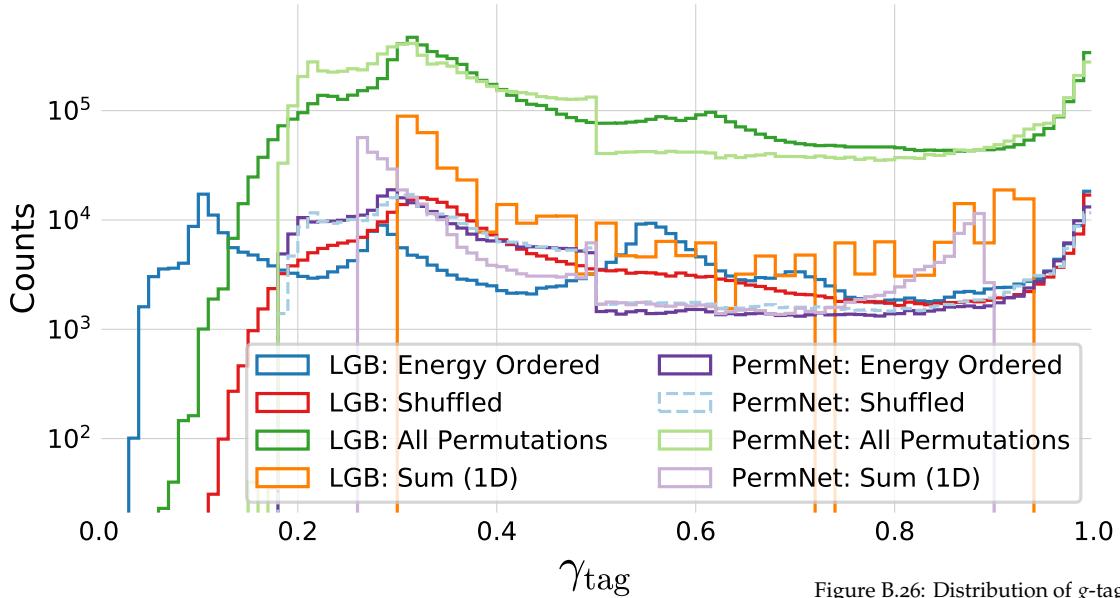


Figure B.26: Distribution of g -tag scores in 4-jet events shown with a logarithmic y -scale for LGB: Energy Ordered in blue, LGB: Shuffled in red, LGB: All Permutations in green, LGB: Sum 1D in orange, PermNet: Energy Ordered in purple, PermNet: Shuffled in light-blue, PermNet: All Permutations in light-green, PermNet: Sum 1D in light-purple. Here LGB and PermNet are the two different type of models and “Energy Ordered”, “Shuffled”, “All Permutations”, and “Sum 1D” are the different methods used for making the input data permutation invariant (except energy ordered).

B.28 g -Tag Scores in 3-Jet Events

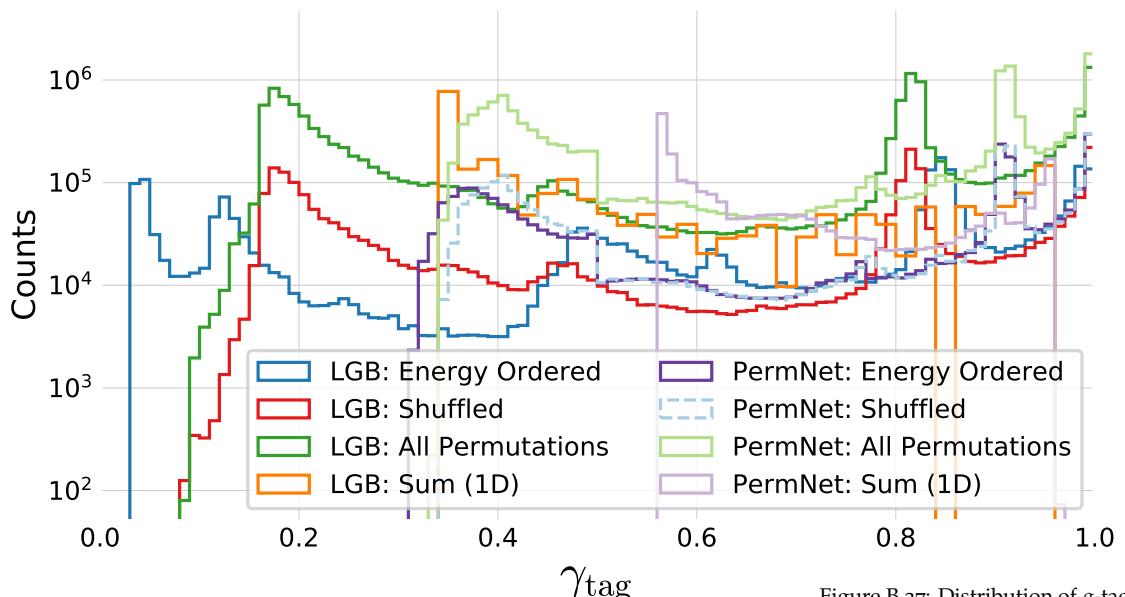


Figure B.27: Distribution of g -tag scores in 3-jet events shown with a logarithmic y -scale.

B.29 ROC Curve for g -Tag in 4-Jet Events

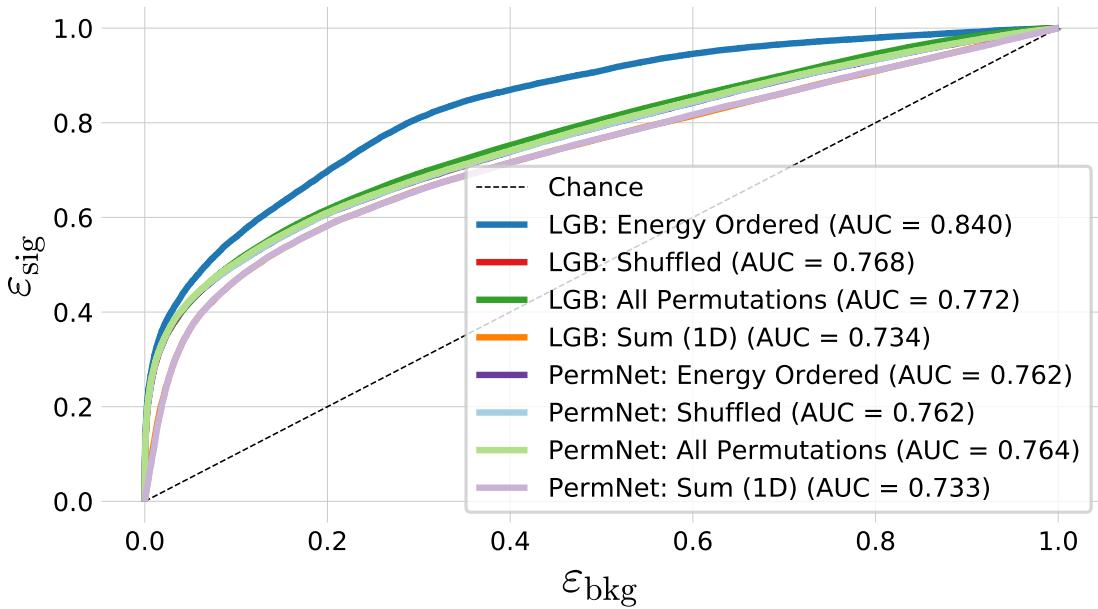


Figure B.28: ROC curve of the eight g -tag models in 4-jet events. First one in dashed black is the ROC curve that you get by random chance. The colors are the same as in Figure 5.19 and in the legend also the Region Under the ROC curve (AUC) is shown.

B.30 ROC Curve for g -Tag in 3-Jet Events

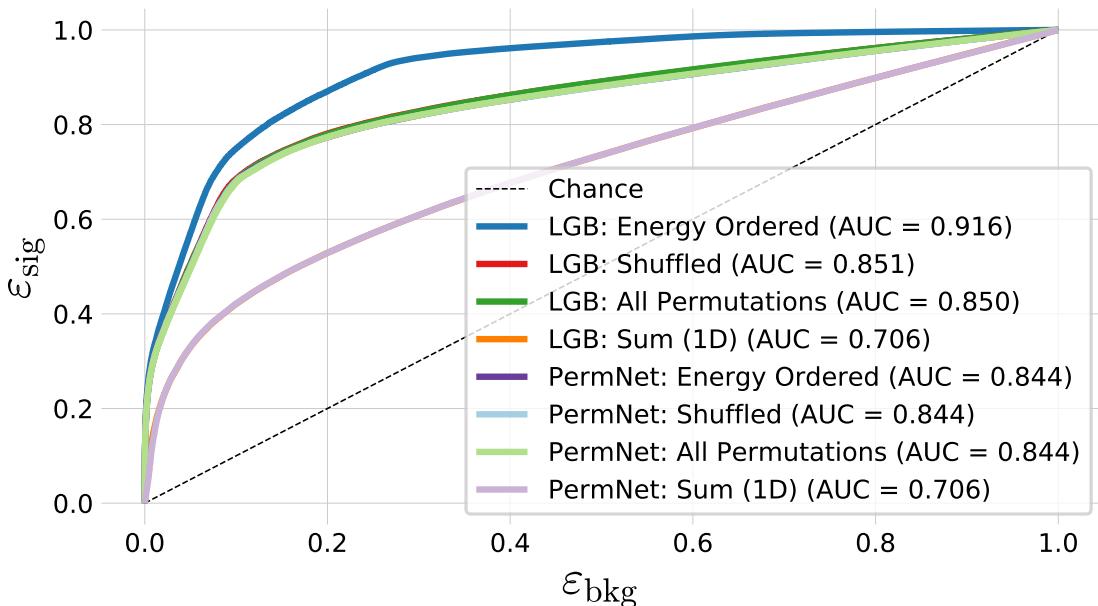


Figure B.29: ROC curve of the eight g -tag models in 3-jet events. First one in dashed black is the ROC curve that you get by random chance. The colors are the same as in Figure 5.19 and in the legend also the Region Under the ROC curve (AUC) is shown.

B.31 Distribution of g -Tag Scores in 4-Jet Events for Signal and Background

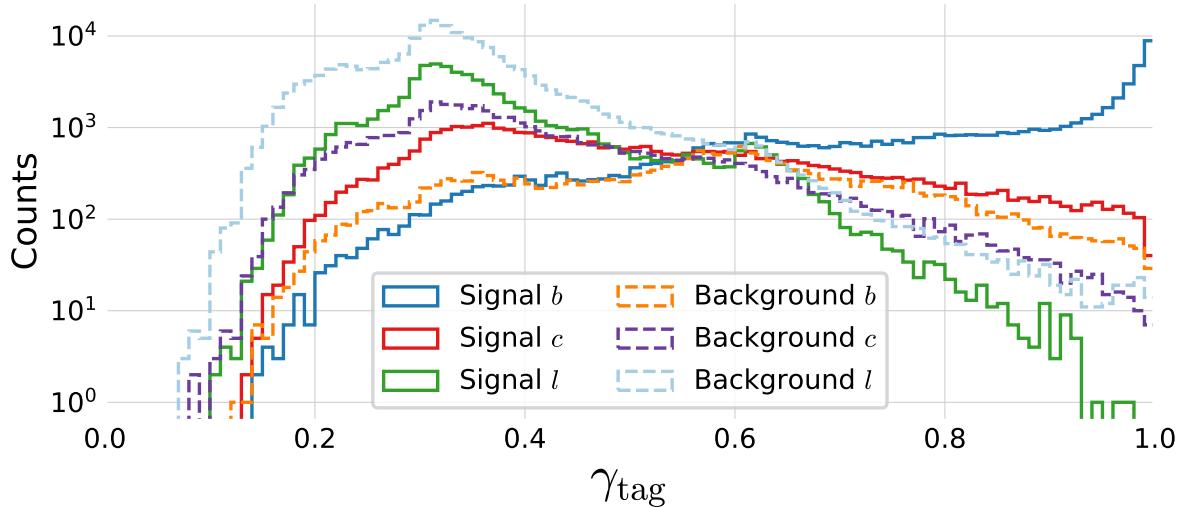


Figure B.30: Histogram of g -tag scores from the LGB-model in 4-jet events for b signal in blue, c signal in red, l (uds) signal in green, b background in orange, c background in purple, l (uds) background in light-blue.

B.32 Distribution of g -Tag Scores in 3-Jet Events for Signal and Background

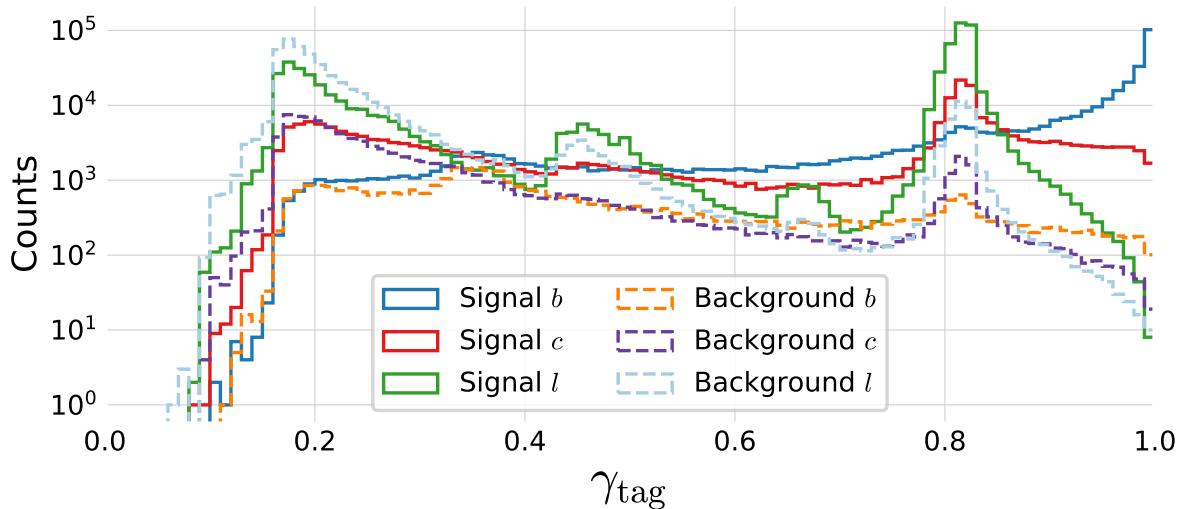
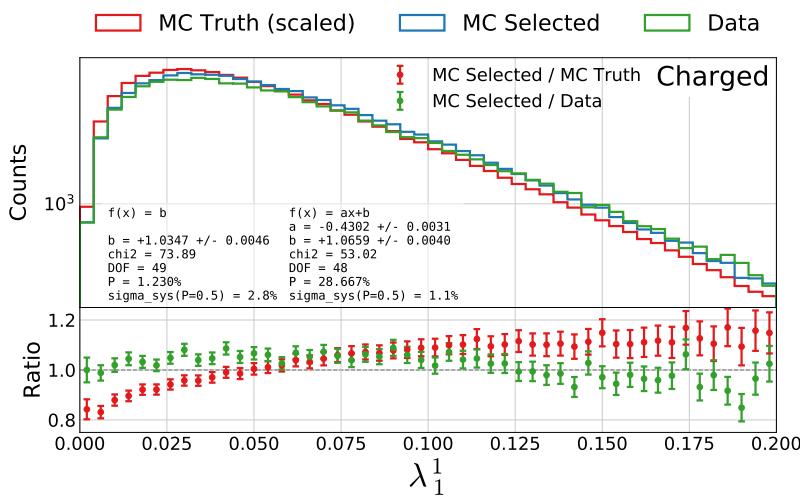
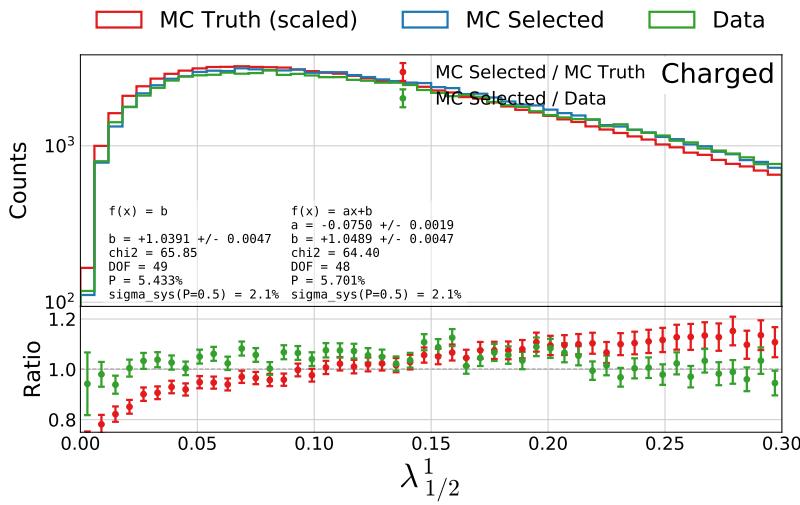
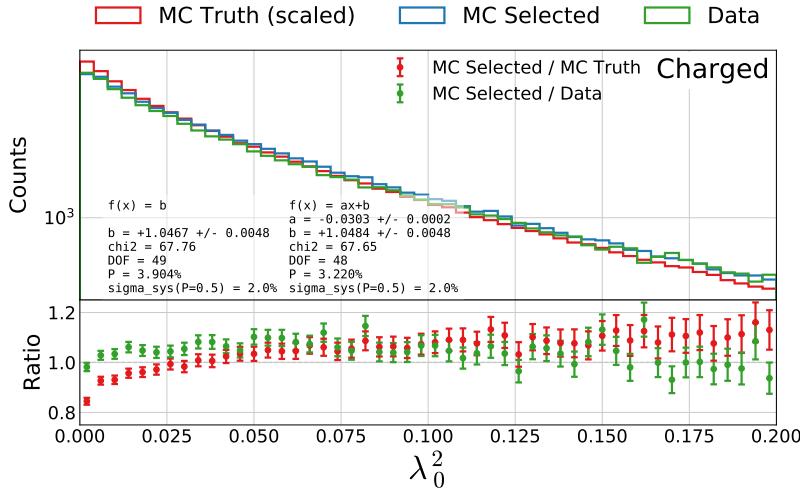


Figure B.31: Histogram of g -tag scores from the LGB-model in 3-jet events for b signal in blue, c signal in red, l (uds) signal in green, b background in orange, c background in purple, l (uds) background in light-blue.

B.33 Generalized Angularities for Charged Gluons Tracks



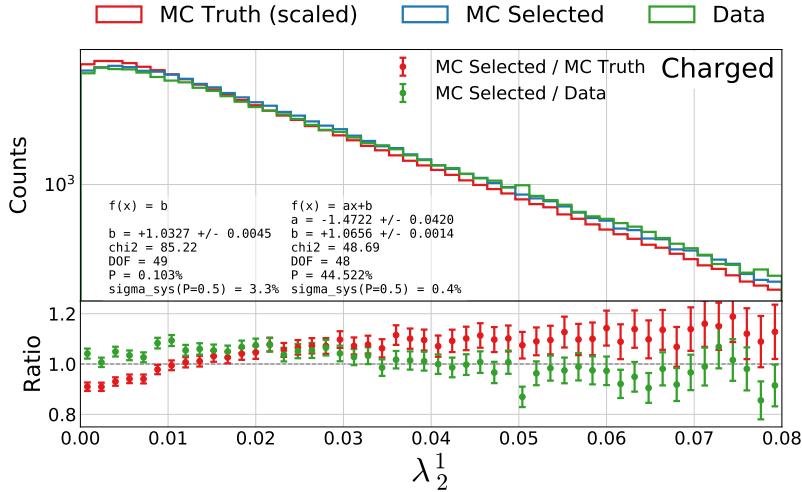


Figure B.35: Distribution of the generalized angularity λ_2^1 for charged gluons jets in 3-jet events. The distributions for MC Selected (blue) and MC Truth (red) are shown in the top plot, while Data (green) is shown in the bottom plot. The ratio between MC Selected and MC Truth is shown in red, and the ratio between MC Selected and Data is shown in green.

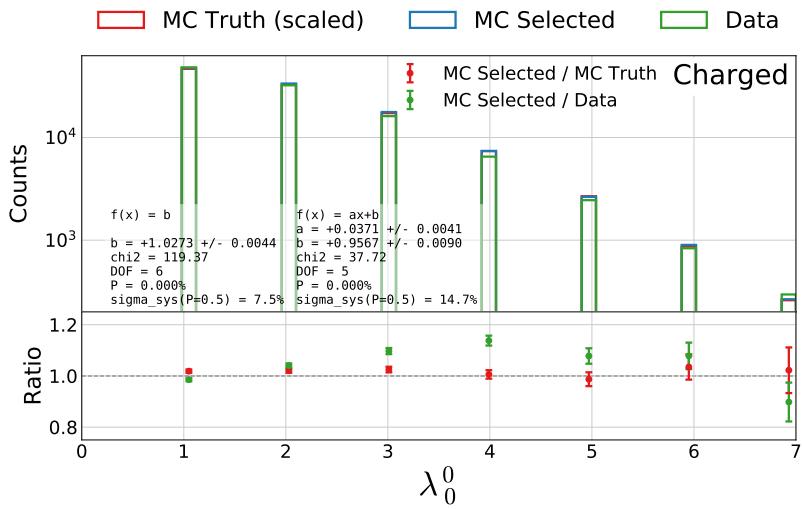
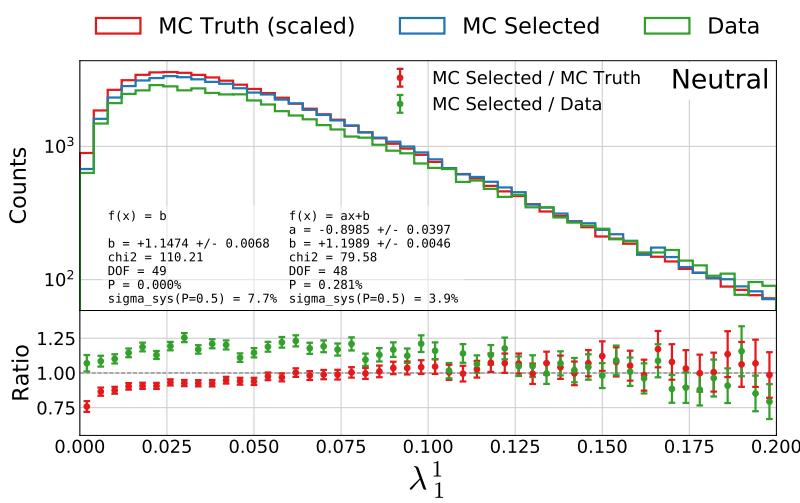
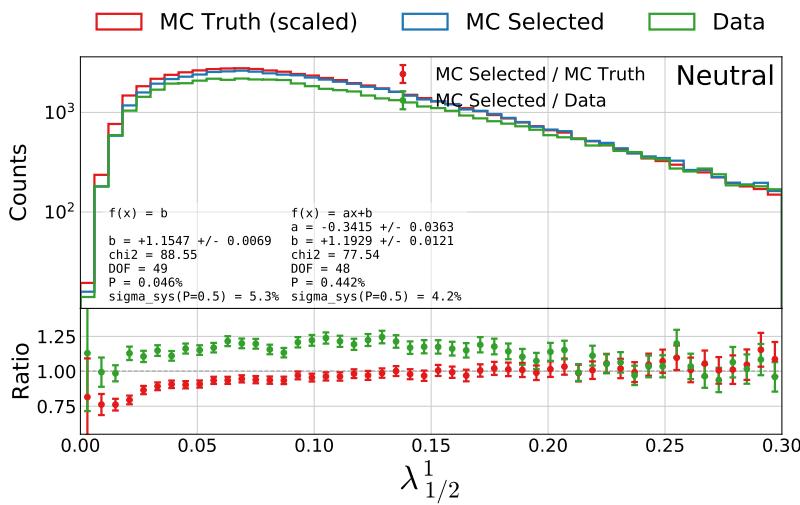
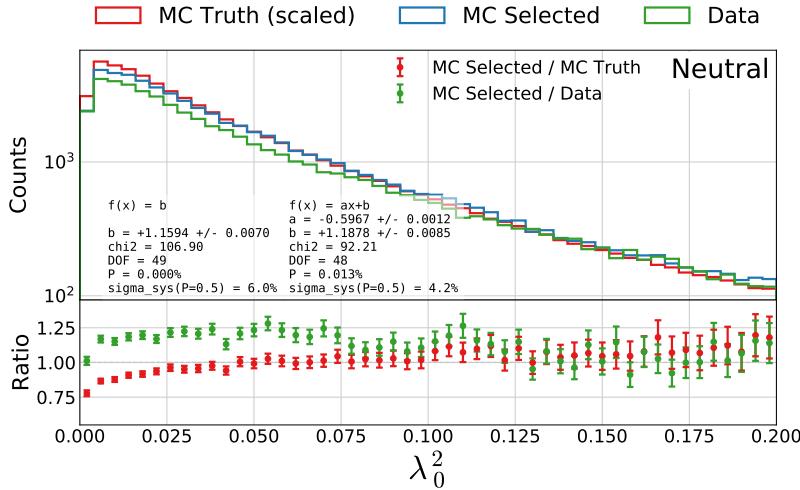
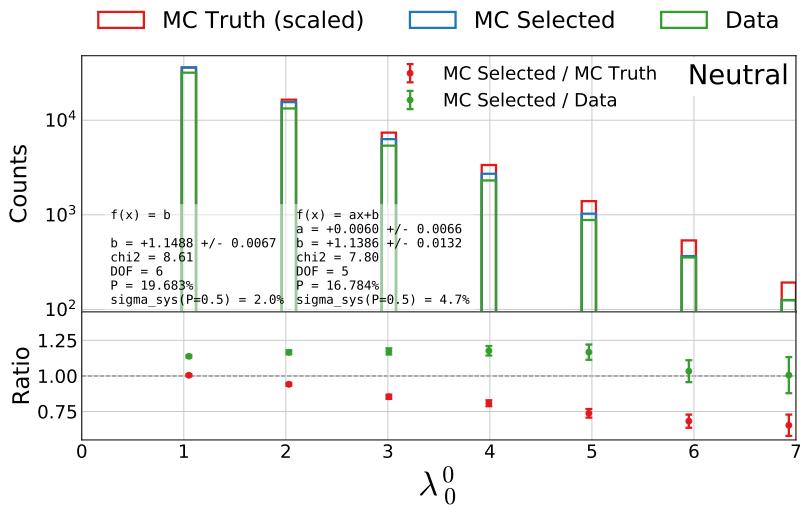
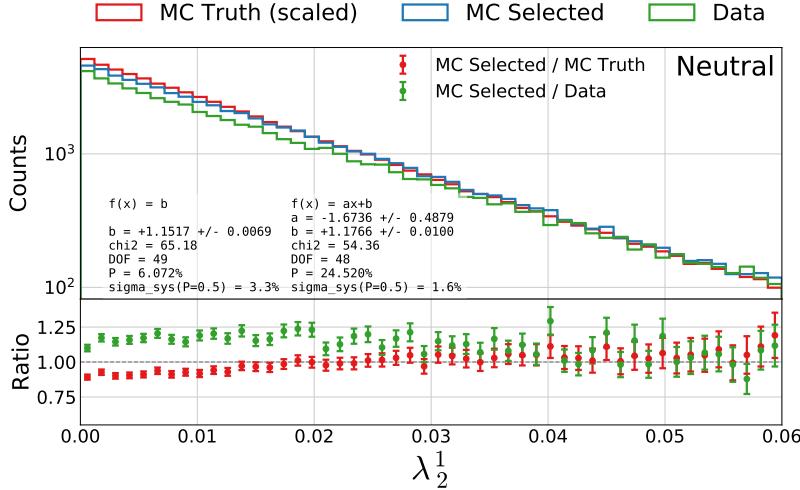


Figure B.36: Distribution of the generalized angularity λ_0^0 for charged gluons jets in 3-jet events. The distributions for MC Selected (blue) and MC Truth (red) are shown in the top plot, while Data (green) is shown in the bottom plot. The ratio between MC Selected and MC Truth is shown in red, and the ratio between MC Selected and Data is shown in green.

B.34 Generalized Angularities for Neutral Gluons Clusters





B.35 *g*-Tagging Efficiency for 4-Jet Events in MC

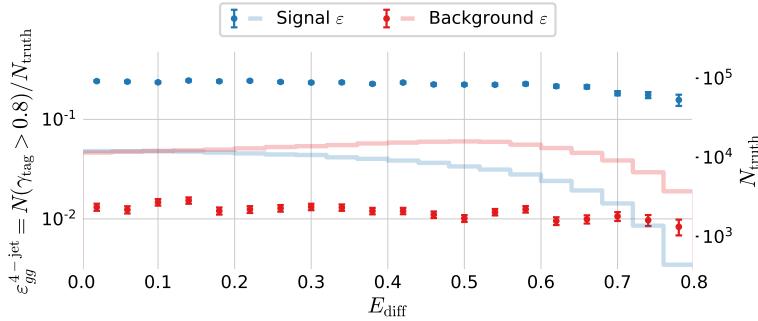


Figure B.42: Efficiency of the *g*-tagging algorithm for 4-jet events as a function of normalized gluon-gluon jet energy difference (asymmetry) E_{diff} in MC. The efficiency is measured as the number of events with a *g*-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for **signal events** according to MC Truth in blue and **background events** according to MC Truth in red.

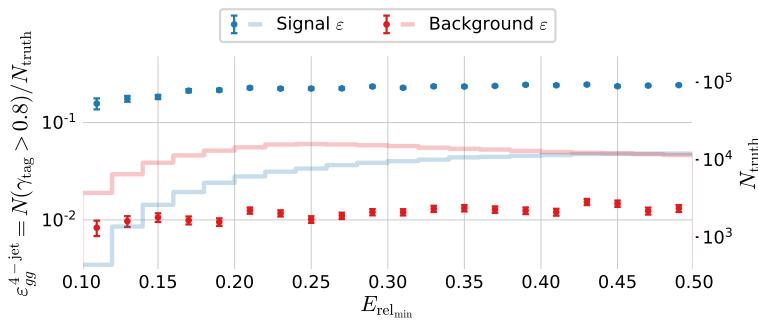


Figure B.43: Efficiency of the *g*-tagging algorithm for 4-jet events as a function of $E_{\text{rel_min}}$ in MC. The efficiency is measured as the number of events with a *g*-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for **signal events** according to MC Truth in blue and **background events** according to MC Truth in red.

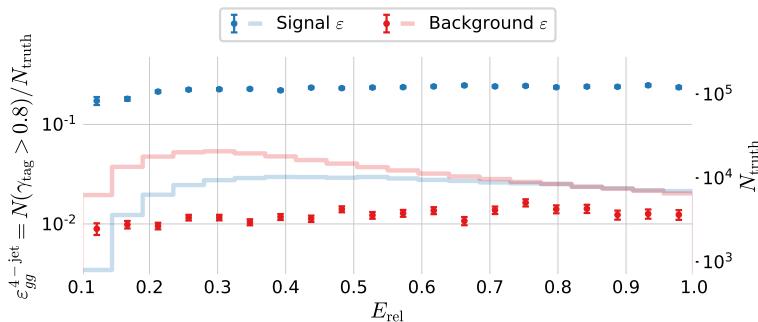


Figure B.44: Efficiency of the *g*-tagging algorithm for 4-jet events as a function of E_{rel} in MC. The efficiency is measured as the number of events with a *g*-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for **signal events** according to MC Truth in blue and **background events** according to MC Truth in red.

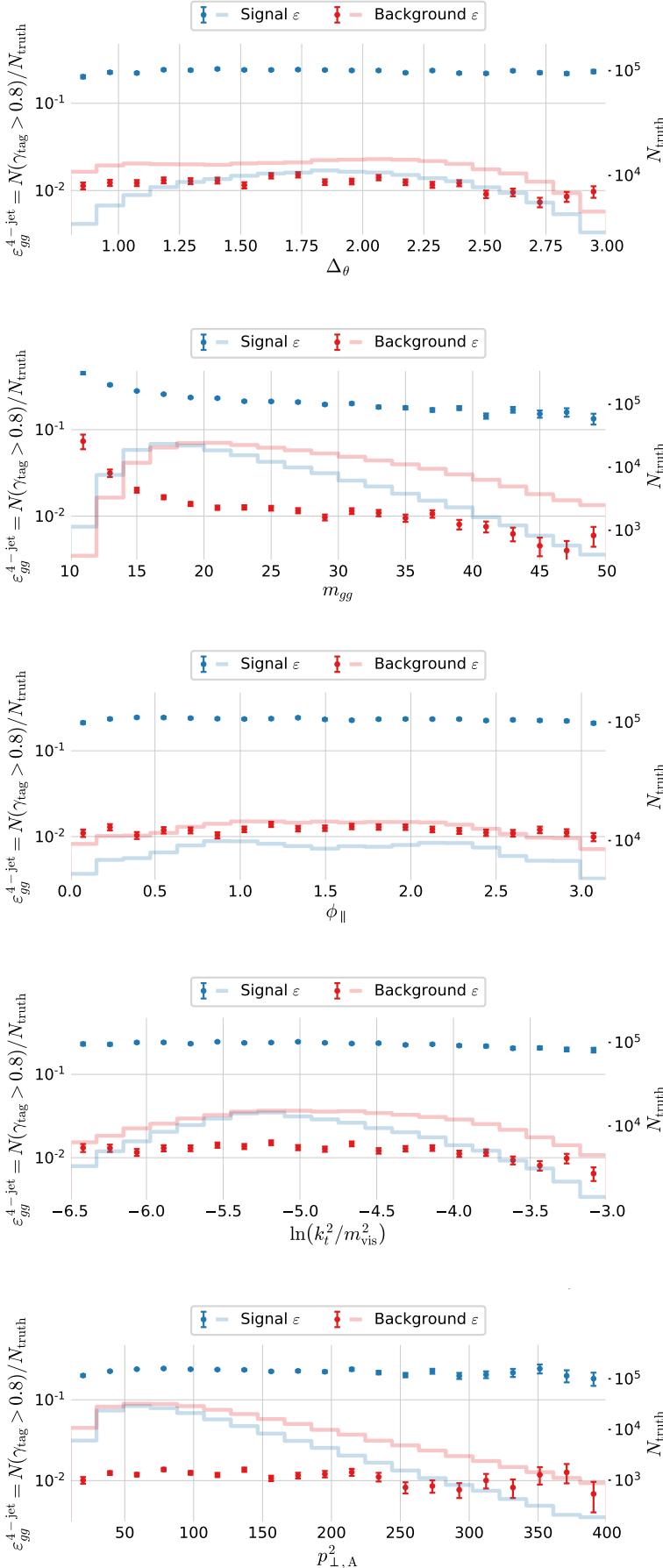


Figure B.45: Efficiency of the g -tagging algorithm for 4-jet events as a function of Δ_θ in MC. The efficiency is measured as the number of events with a g -tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for **signal events** according to MC Truth in blue and **background events** according to MC Truth in red.

Figure B.46: Efficiency of the g -tagging algorithm for 4-jet events as a function of m_{gg} in MC. The efficiency is measured as the number of events with a g -tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for **signal events** according to MC Truth in blue and **background events** according to MC Truth in red.

Figure B.47: Efficiency of the g -tagging algorithm for 4-jet events as a function of ϕ_{\parallel} in MC. The efficiency is measured as the number of events with a g -tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for **signal events** according to MC Truth in blue and **background events** according to MC Truth in red.

Figure B.48: Efficiency of the g -tagging algorithm for 4-jet events as a function of $\ln(k_t^2/m_{\text{vis}}^2)$ in MC. The efficiency is measured as the number of events with a g -tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for **signal events** according to MC Truth in blue and **background events** according to MC Truth in red.

Figure B.49: Efficiency of the g -tagging algorithm for 4-jet events as a function of $p_{\perp,A}^2$ in MC. The efficiency is measured as the number of events with a g -tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for **signal events** according to MC Truth in blue and **background events** according to MC Truth in red.

B.36 Closure Plot Comparing MC Truth and the Efficiency Corrected g -Tagging Model in 4-Jet Events

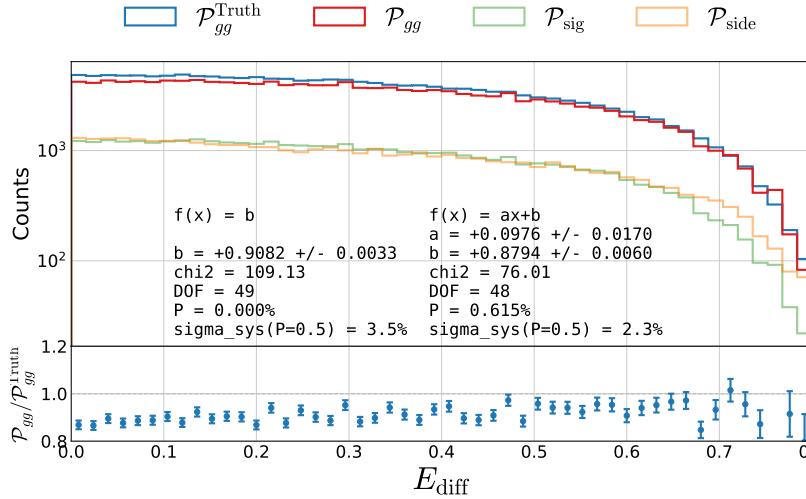


Figure B.50: Closure plot comparing MC Truth and the efficiency corrected g -tagging model in 4-jet events for E_{diff} . In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the \mathcal{P}_{gg} based on MC but without Truth in red, the distribution in the signal region \mathcal{P}_{sig} in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

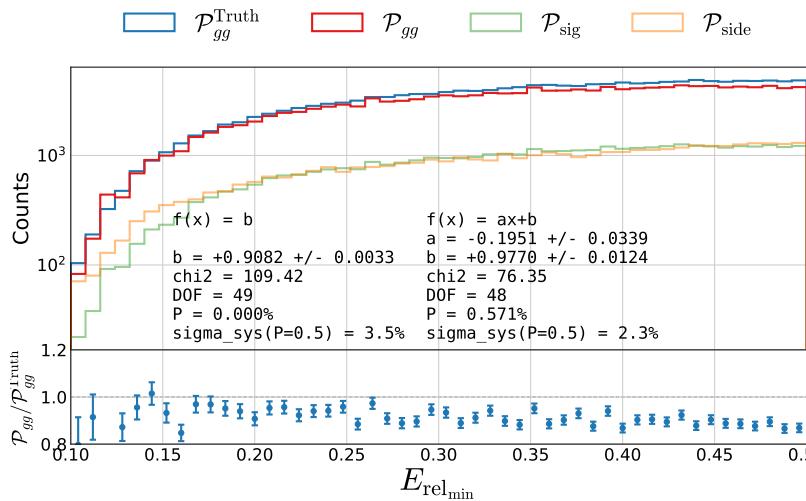


Figure B.51: Closure plot comparing MC Truth and the efficiency corrected g -tagging model in 4-jet events for $E_{\text{rel_min}}$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the \mathcal{P}_{gg} based on MC but without Truth in red, the distribution in the signal region \mathcal{P}_{sig} in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

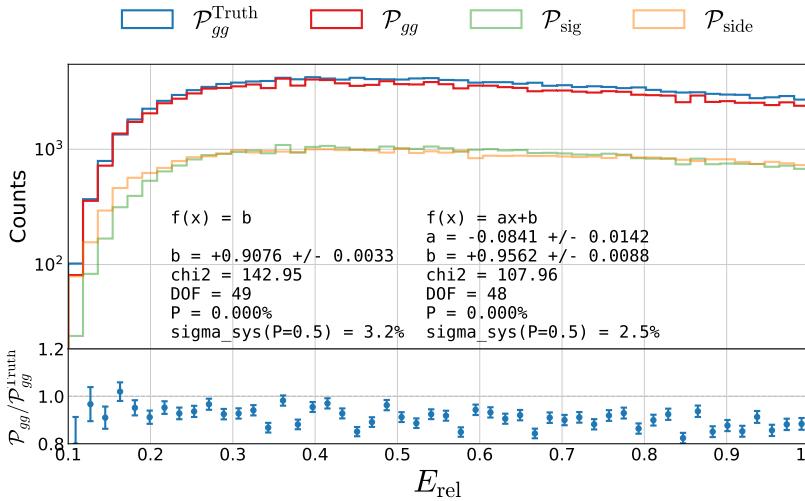


Figure B.52: Closure plot comparing MC Truth and the efficiency corrected g -tagging model in 4-jet events for E_{rel} . In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the \mathcal{P}_{gg} based on MC but without Truth in red, the distribution in the signal region \mathcal{P}_{sig} in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

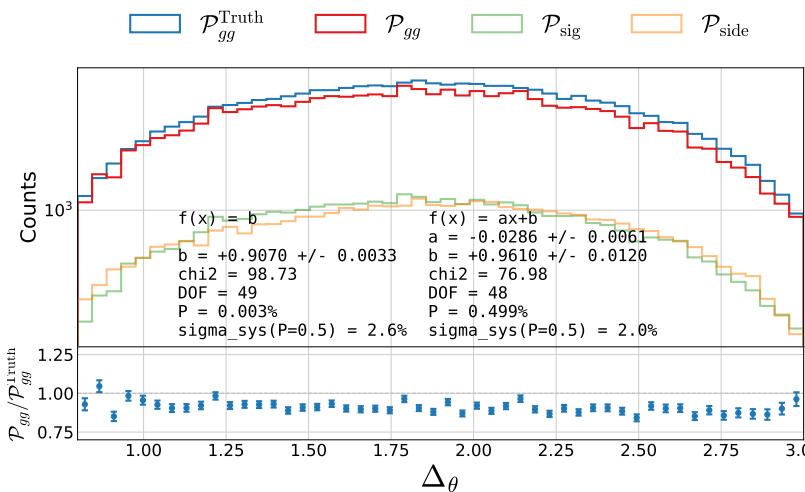


Figure B.53: Closure plot comparing MC Truth and the efficiency corrected g -tagging model in 4-jet events for Δ_{θ} . In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the \mathcal{P}_{gg} based on MC but without Truth in red, the distribution in the signal region \mathcal{P}_{sig} in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

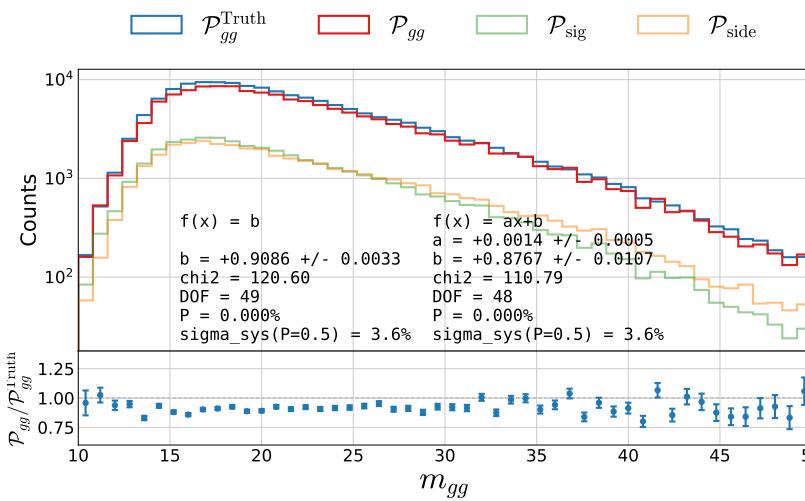


Figure B.54: Closure plot comparing MC Truth and the efficiency corrected g -tagging model in 4-jet events for m_{gg} . In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the \mathcal{P}_{gg} based on MC but without Truth in red, the distribution in the signal region \mathcal{P}_{sig} in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

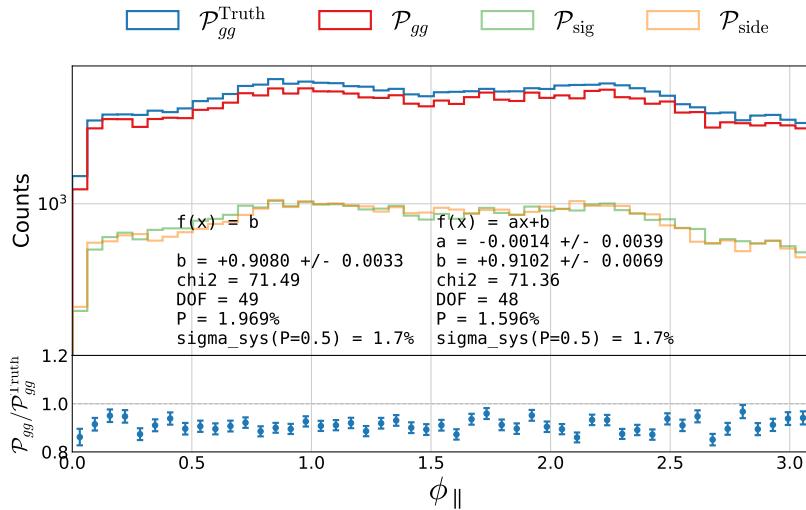


Figure B.55: Closure plot comparing MC Truth and the efficiency corrected g -tagging model in 4-jet events for ϕ_{\parallel} . In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the \mathcal{P}_{gg} based on MC but without Truth in red, the distribution in the signal region \mathcal{P}_{sig} in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

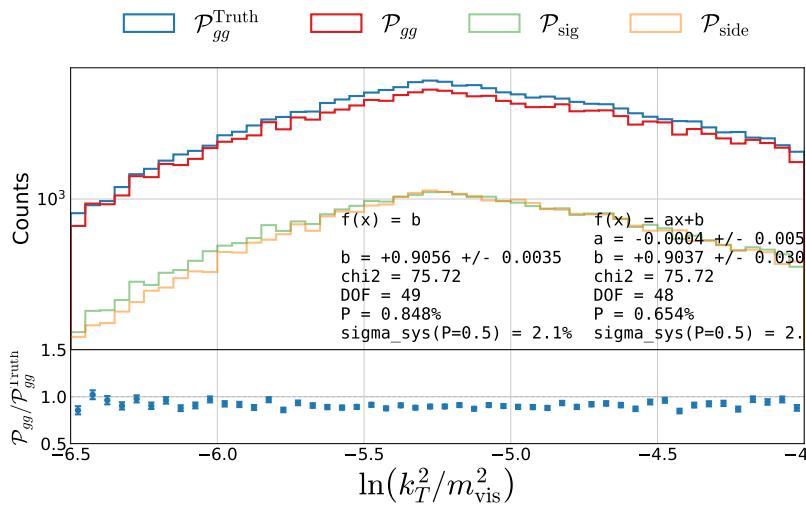


Figure B.56: Closure plot comparing MC Truth and the efficiency corrected g -tagging model in 4-jet events for $\ln(k_T^2/m_{\text{vis}}^2)$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the \mathcal{P}_{gg} based on MC but without Truth in red, the distribution in the signal region \mathcal{P}_{sig} in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

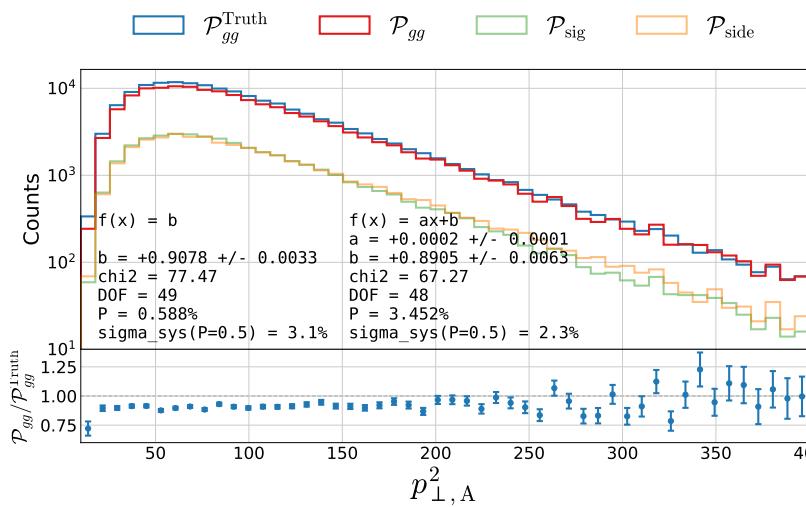


Figure B.57: Closure plot comparing MC Truth and the efficiency corrected g -tagging model in 4-jet events for $p_{\perp,A}^2$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the \mathcal{P}_{gg} based on MC but without Truth in red, the distribution in the signal region \mathcal{P}_{sig} in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between \mathcal{P}_{gg} and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

B.37 Gluon Splitting Distribution Comparison in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region A

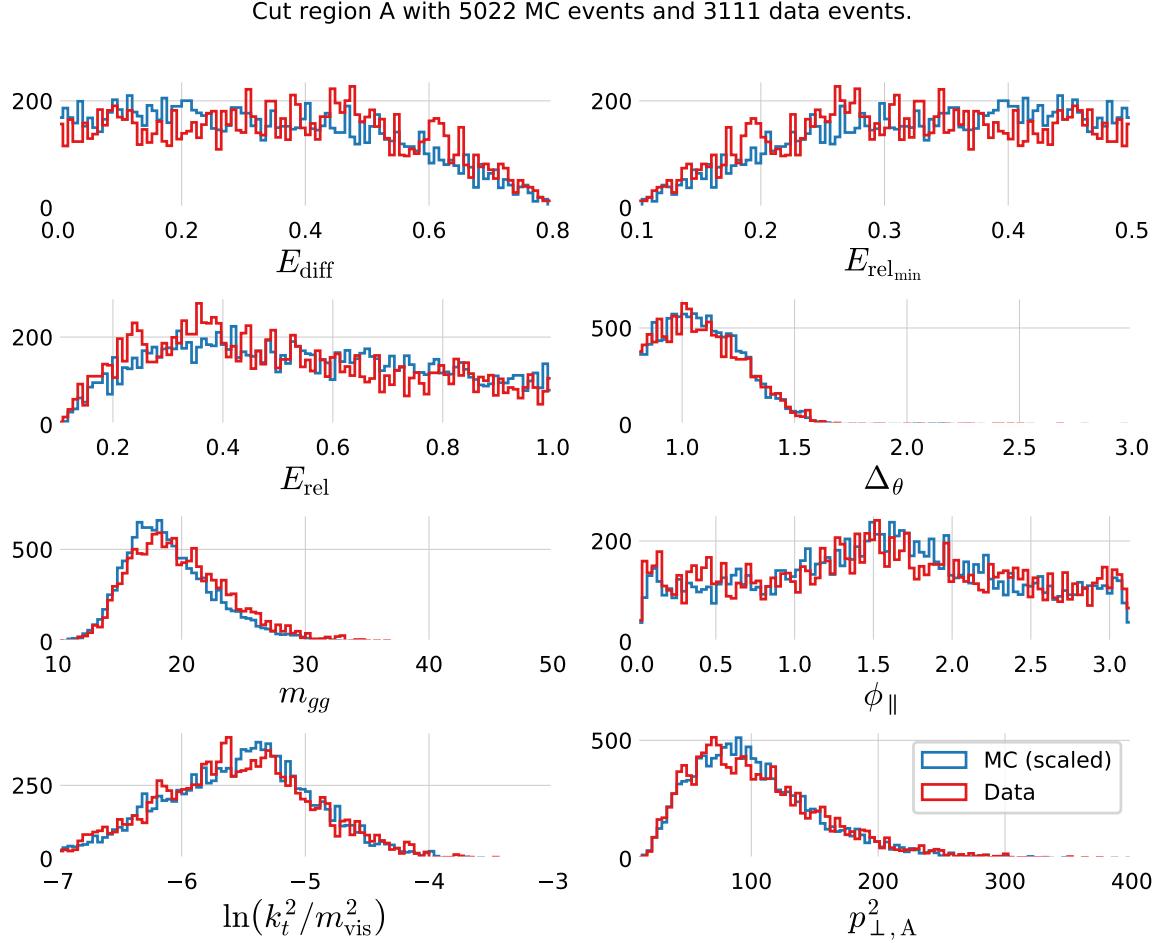


Figure B.58: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region A, see Table ???. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region A which has 5022 events in the MC sample and 3111 in the Data sample.

B.38 Gluon Splitting Distribution Comparison in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region B

Cut region B with 7382 MC events and 4035 data events.

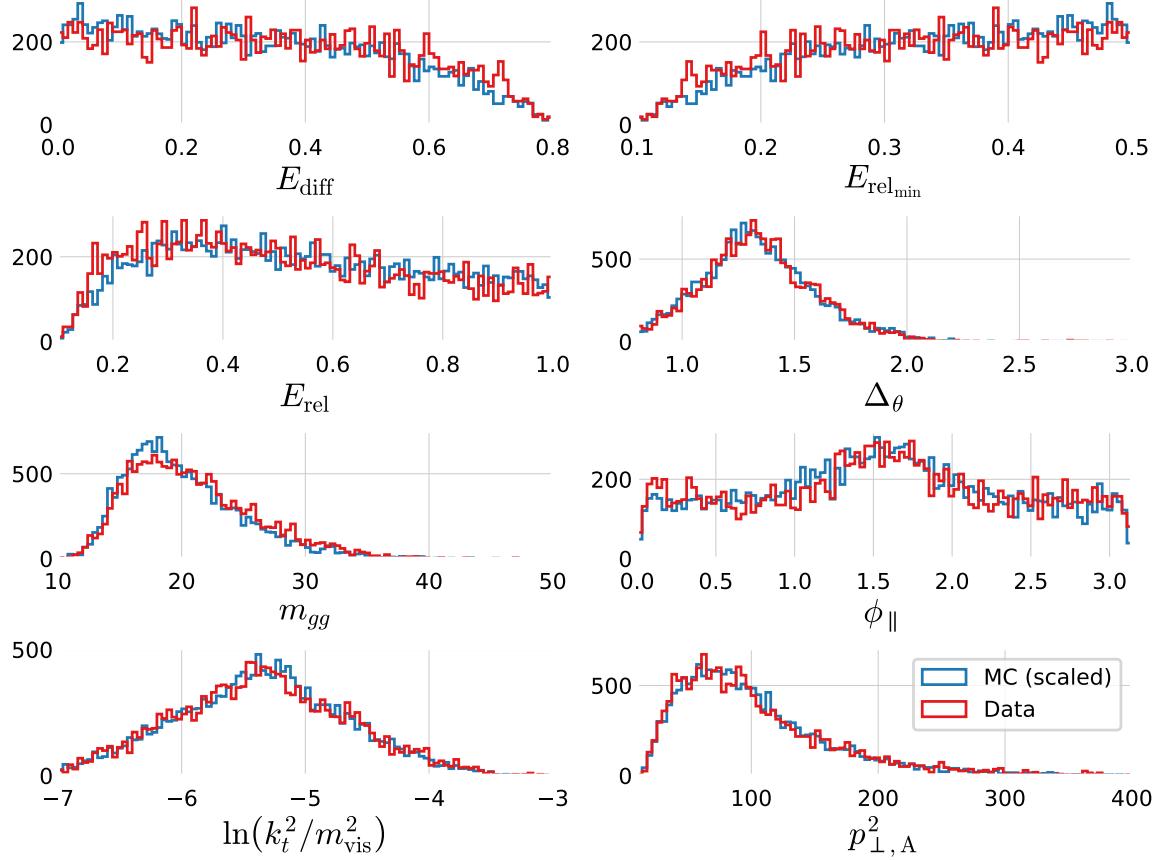


Figure B.59: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region B, see Table ???. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region B which has 7382 events in the MC sample and 4035 in the Data sample.

B.39 Gluon Splitting Distribution Comparison in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region C

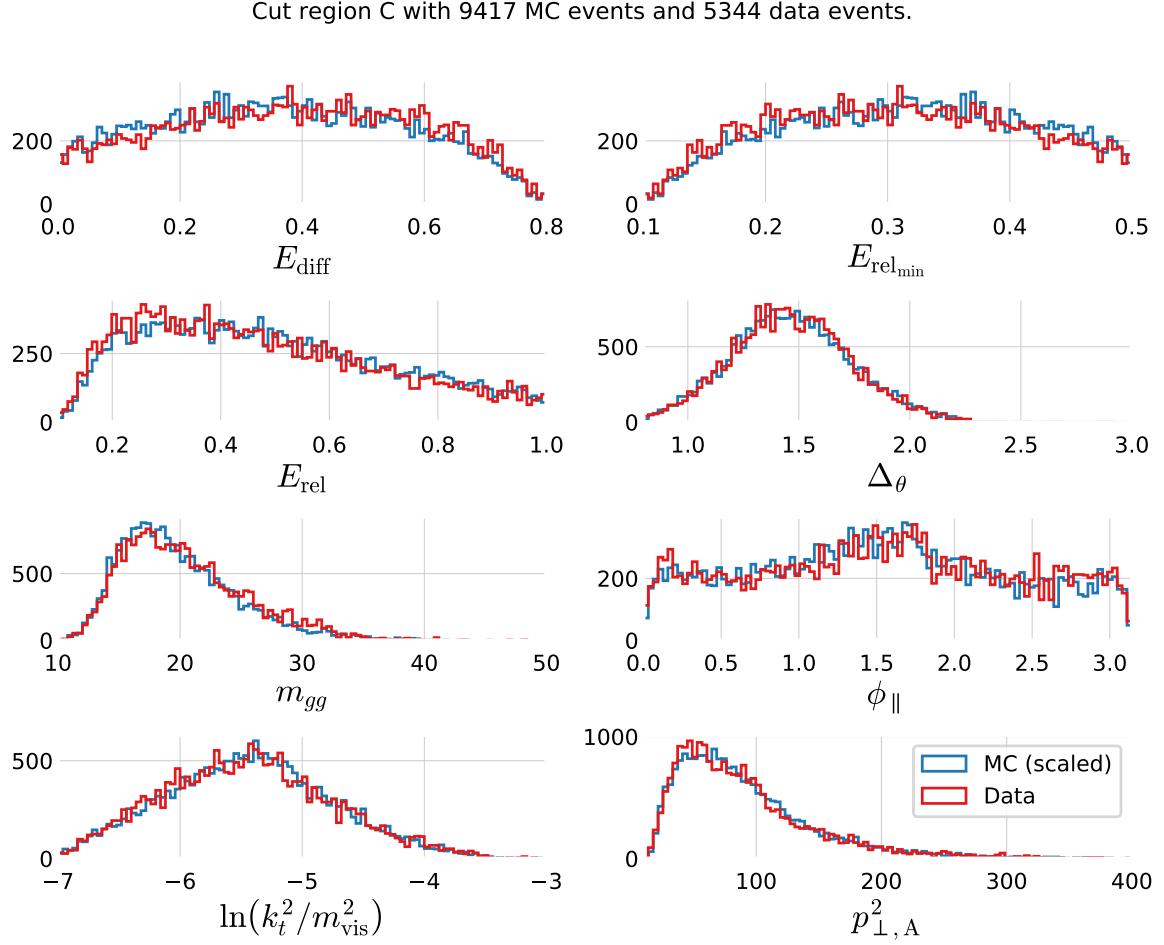


Figure B.60: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region C, see Table ???. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region C which has 9417 events in the MC sample and 5344 in the Data sample.

B.40 Gluon Splitting Distribution Comparison in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region D

Cut region D with 26366 MC events and 13780 data events.

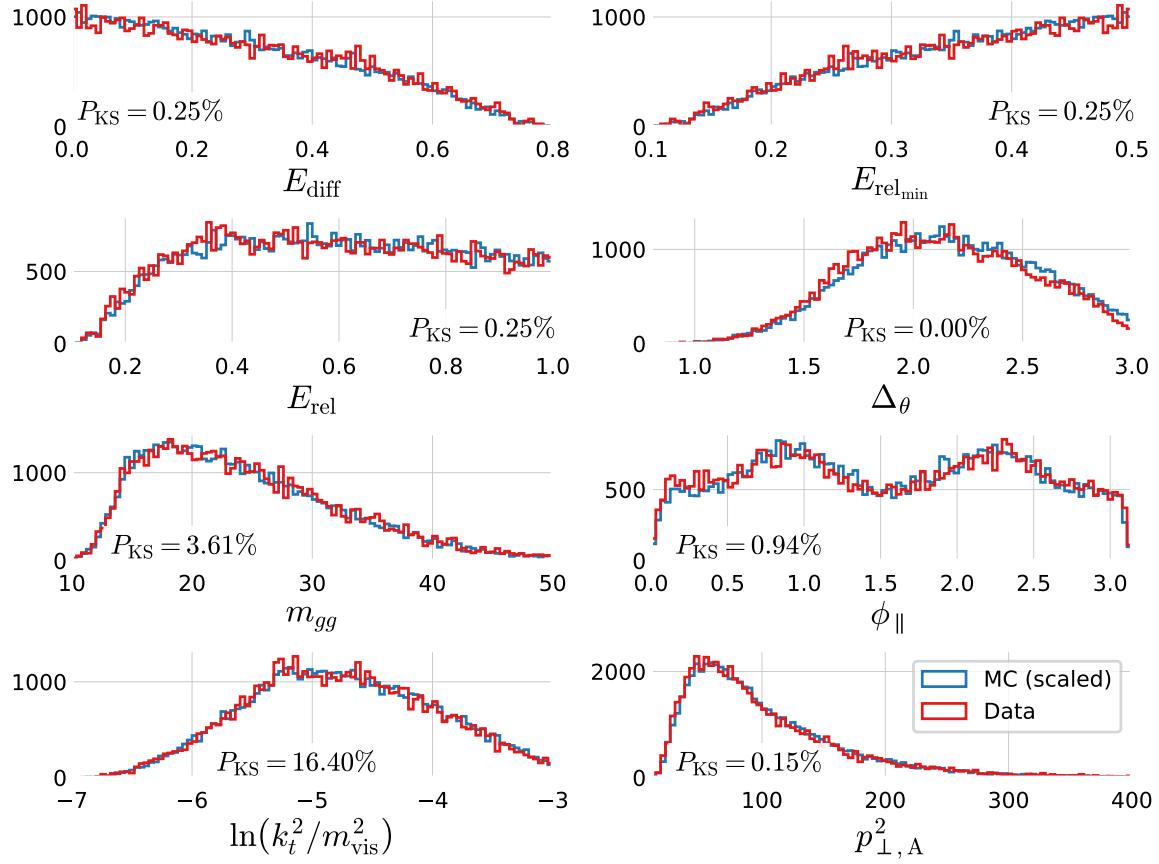


Figure B.61: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region D, see Table ???. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region D which has 26 366 events in the MC sample and 13 780 in the Data sample.

(This page is intentionally left blank).

List of Figures

2.1	The Learning Problem	6
2.2	Approximation-Estimation Tradeoff	10
2.3	Regularization Strength	11
2.4	Regularization Effect of L_2	12
2.5	Regularization Effect of L_1	12
2.6	k -Fold Cross Validation	13
2.7	k -Fold Cross Validation for Time Series Data	13
2.8	Objective Functions Zoom In	15
2.9	Objective Functions	15
2.10	Decision Tree Cuts In Feature Space	16
2.11	Decision Tree	16
2.12	Grid Search	20
2.13	Random Search	20
2.14	Bayesian Optimization	22
3.1	Danish Housing Price Index	27
3.2	Distributions of the Variables in the Housing Prices	28
3.3	Geographic Distribution of the Sold Residences	29
3.4	Histogram of Prices of Houses and Apartments Sold in Denmark	30
3.5	Linear Correlation Between Variables and Price	31
3.6	Comparison of the Linear Correlation ρ and the Nonlinear MIC	31
3.7	Nonlinear Correlation Between Variables and Price	32
3.8	Validity of Input Features	32
3.9	Validity Dendrogram	33
3.10	Prophet Forecast for Apartments	34
3.11	Prophet Trends	35
3.12	Sample Weight as a Function of Time for Different Half-Lives.	36
3.13	Parallel Coordinate Plot of the Initial Hyperparameter Optimization for Apartments	37
3.14	Initial HPO Results for the Weight Half-life $T_{\frac{1}{2}}$	38
3.15	Initial HPO Results for the Loss Function	38
3.16	Parallel Coordinate Plot of the Random Search Hyperparameter Optimization Results of XGBoost for Apartments	39
3.17	Hyperparameter Optimization: Random Search Results	40
3.18	Early Stopping Results	40
3.19	Performance of XGB-model for Apartments	41
3.20	Standard Deviation and MAD of the Static and Dynamic XGB Forecasts	41
3.21	Market Index based on the Static and Dynamic XGB Forecasts	42
3.22	SHAP Prediction Explanation for Apartments	44
3.23	Feature importance of Apartments Prices	44

3.24	Feature Importance Interaction Plot for Apartments	45
3.25	Performance Comparison of Multiple Models for Apartments	46
3.26	Feature Importance of Villas With Descriptions	49
4.1	The Standard Model	54
4.2	Feynman Diagram for the Jet Production at LEP	55
4.3	Quark Splitting	55
4.4	Hadronization Process	56
4.5	The ALEPH Detector	57
4.6	Polar Angle	57
4.7	Azimuthal Angle	57
4.8	Track Significance	59
5.1	Histograms of the Vertex Variables	66
5.2	UMAP Visualization of the Vertex Variables for 4-Jet Events	67
5.3	UMAP Visualization of the Vertex Variables for 3-Jet Events	67
5.4	Correlations Between of the Vertex Variables	67
5.5	Plot of the Log-Loss ℓ_{\log}	68
5.6	Hyperparameter Optimization of b -Tagging	69
5.7	Parallel Plot of HPO Results for 4-Jet b -Tagging	70
5.8	b -Tag Scores in 4-Jet Events	70
5.9	ROC Curve for 4-Jet b -Tagging	71
5.10	Distribution of b -Tags in 4-Jet Events	71
5.11	Global Feature Importances for the LGB b -Tagging Algorithm on 4-Jet Events	72
5.12	The Expit Function	72
5.13	The Logit Function	72
5.14	SHAP 3-Jet Model Explanation for b -Like Jet	72
5.15	b -Tagging Efficiency $\varepsilon_b^{b\text{-sig}}$ as a Function of Jet Energy	74
5.16	b -Tagging Efficiency $\varepsilon_g^{g\text{-sig}}$ as a Function of Jet Energy	75
5.17	Hyperparameter Optimization of g -Tagging	77
5.18	1D Sum Models Predictions and Signal Fraction for 4-jets events	79
5.19	g -Tag Scores in 4-Jet Events	79
5.20	ROC Curve for g -Tag in 4-Jet Events	80
5.21	Distribution of g -Tag Scores in 4-Jet Events for Signal and Background	80
5.22	Distribution of b -Tag Scores in 3-Jet l -Quark Events for Low and High g -Tag Values	81
5.23	3D Scatter Plot of β_{tag} -Values for High and Low γ_{tag} l -Quark Events	81
5.24	g -Tagging Pseudo Efficiency for $b\bar{b}g$ -Events as a Function of g -Tag	82
5.25	g -Tagging Pseudo Efficiency for $b\bar{b}g$ -Events as a Function of The Mean Invariant Mass	83
5.26	Generalized Angularities	83
5.27	Generalized Angularities for Charged Gluons Jets in 3-Jet Events: λ_0^2	84
5.28	Soft Wide Angle Gluons in 4-Jet Events	87
5.29	Soft Collinear Gluons in 4-Jet Events	87
5.30	Hard Non $g \rightarrow gg$ Gluons in 4-Jet Events	87
5.31	g -Tagging Efficiency for 4-Jet Events in MC as a Function of the Normalized Gluon-Gluon Jet Energy Difference Asymmetry E_{diff}	88
5.32	Closure Plot Comparing MC Truth and the Efficiency Corrected g -Tagging Model in 4-Jet Events for the Normalized Gluon Gluon Jet Energy Asymmetry	89
5.33	Overview of the Four Regions in the $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space	91
5.34	Gluon Splitting Distribution Comparison in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space Region A	92

5.35 Ratio Plot of E_{diff} in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} for Region A	92
5.36 Ratio Plot of E_{diff} in MC and Data for $R_{gg}^{k_t}$ - R_{gg}^{CA} for Region D	93
A.1 Distributions of the Input Variables – Part 1	96
A.2 Distributions of the Input Variables – Part 2	97
A.3 Distributions of the Input Variables – Part 3	98
A.4 Distributions of the Input Variables – Part 4	99
A.5 Distributions of the Input Variables – Part 5	100
A.6 Distributions of the Input Variables – Part 6	101
A.7 Distributions of the Input Variables – Part 7	102
A.8 Distributions of the Input Variables – Part 8	103
A.9 Distributions of the Input Variables – Part 9	104
A.10 Distributions of the Input Variables – Part 10	105
A.11 Distributions of the Input Variables – Part 11	106
A.12 Distributions of the Input Variables – Part 12	107
A.13 Distributions of the Input Variables – Part 13	108
A.14 Distributions of the Input Variables – Part 14	109
A.15 Linear Correlations	110
A.16 Comparison of the Linear Correlation ρ and the Nonlinear MIC for Different Levels of Noise	111
A.17 Validity Heatmap	111
A.18 Prophet Forecast for apartments	112
A.19 Prophet Trends	113
A.20 Parallel Coordinate Plot of the Initial Hyperparameter Optimization for Apartments	118
A.21 Parallel Coordinate Plot of the Initial Hyperparameter Optimization for Houses	118
A.22 Parallel Coordinate Plot of the Hyperparameter Optimization for Apartments Using Random Search	118
A.23 Parallel Coordinate Plot of the Hyperparameter Optimization for Apartments Using Bayesian Optimization	119
A.24 Parallel Coordinate Plot of the Hyperparameter Optimization for Houses Using Random Search	119
A.25 Parallel Coordinate Plot of the Hyperparameter Optimization for Houses Using Bayesian Optimization	119
A.26 Random Search Results as a Function of Iteration for Apartments	120
A.27 Bayesian Optimization Results as a Function of Iteration for Apartments	120
A.28 Random Search Results as a Function of Iteration for Houses	120
A.29 Bayesian Optimization Results as a Function of Iteration for Houses	120
A.30 Performance of XGB-model on apartment prices	121
A.31 Performance Comparison of Multiple Models for Houses	122
B.1 UMAP Parameter Grid Search	124
B.2 Visualization of the t-SNE algorithm	124
B.3 Parallel Coordinate Plot of HPO results for 3-Jet b -Tagging	125
B.4 Parallel Coordinate Plot of HPO results for 4-Jet b -Tagging	125
B.5 b -Tag Scores in 3-Jet Events	126
B.6 b -Tag Scores in 4-Jet Events	126
B.7 ROC curve for 3-jet b -tagging	127
B.8 ROC curve for 4-jet b -tagging	127
B.9 Distribution of b -Tags in 3-Jet Events	128
B.10 Distribution of b -Tags in 4-Jet Events	128

B.11	Global Feature Importances for the LGB b -Tagging Algorithm on 3-Jet Events	129
B.12	Global Feature Importances for the LGB b -Tagging Algorithm on 4-Jet Events	129
B.13	b -Tagging Efficiency $\varepsilon_b^{b\text{-sig}}$ as a Function of Jet Energy	130
B.14	b -Tagging Efficiency $\varepsilon_g^{b\text{-sig}}$ as a Function of Jet Energy	130
B.15	b -Tagging Efficiency $\varepsilon_b^{g\text{-sig}}$ as a Function of Jet Energy	131
B.16	b -Tagging Efficiency $\varepsilon_g^{g\text{-sig}}$ as a Function of Jet Energy	131
B.17	Parallel Coordinate Plot of HPO Results for 3-Jet g -Tagging for Energy Ordered Jets	132
B.18	Parallel Coordinate Plot of HPO Results for 3-Jet g -Tagging for Shuffled Jets	132
B.19	Parallel Coordinate Plot of HPO Results for 4-Jet g -Tagging for Energy Ordered Jets	133
B.20	Parallel Coordinate Plot of HPO Results for 4-Jet g -Tagging for Shuffled Jets	133
B.21	PermNet Architecture	134
B.22	1D Sum Models Predictions and Signal Fraction for 4-jets events	135
B.23	1D LGB Model Cuts for 4-jets events	135
B.24	1D Sum Models Predictions and Signal Fraction for 3-jets events	136
B.25	1D LGB Model Cuts for 3-jets events	136
B.26	g -Tag Scores in 4-Jet Events	137
B.27	g -Tag Scores in 3-Jet Events	137
B.28	ROC Curve for g -Tag in 4-Jet Events	138
B.29	ROC Curve for g -Tag in 3-Jet Events	138
B.30	Distribution of g -Tag Scores in 4-Jet Events for Signal and Background	139
B.31	Distribution of g -Tag Scores in 3-Jet Events for Signal and Background	139
B.32	Generalized Angularities for Charged Gluons Jets: λ_0^2	140
B.33	Generalized Angularities for Charged Gluons Jets: $\lambda_{1/2}^1$	140
B.34	Generalized Angularities for Charged Gluons Jets: λ_1^1	140
B.35	Generalized Angularities for Charged Gluons Jets: λ_1^2	141
B.36	Generalized Angularities for Charged Gluons Jets: λ_0^0	141
B.37	Generalized Angularities for Neutral Gluons Jets: λ_0^2	142
B.38	Generalized Angularities for Neutral Gluons Jets: $\lambda_{1/2}^1$	142
B.39	Generalized Angularities for Neutral Gluons Jets: λ_1^1	142
B.40	Generalized Angularities for Neutral Gluons Jets: λ_1^2	143
B.41	Generalized Angularities for Neutral Gluons Jets: λ_0^0	143
B.42	g -Tagging Efficiency for 4-Jet Events in MC as a Function of E_{diff}	144
B.43	g -Tagging Efficiency for 4-Jet Events in MC as a Function of $E_{\text{rel}_{\min}}$	144
B.44	g -Tagging Efficiency for 4-Jet Events in MC as a Function of E_{rel}	144
B.45	g -Tagging Efficiency for 4-Jet Events in MC as a Function of Δ_θ	145
B.46	g -Tagging Efficiency for 4-Jet Events in MC as a Function of m_{gg}	145
B.47	g -Tagging Efficiency for 4-Jet Events in MC as a Function of ϕ_{\parallel}	145
B.48	g -Tagging Efficiency for 4-Jet Events in MC as a Function of $\ln(k_t^2/m_{\text{vis}}^2)$	145
B.49	g -Tagging Efficiency for 4-Jet Events in MC as a Function of $p_{\perp,A}^2$	145
B.50	Closure Plot Comparing MC Truth and the Efficiency Corrected g -Tagging Model in 4-Jet Events for E_{diff}	146
B.51	Closure Plot Comparing MC Truth and the Efficiency Corrected g -Tagging Model in 4-Jet Events for $E_{\text{rel}_{\min}}$	146
B.52	Closure Plot Comparing MC Truth and the Efficiency Corrected g -Tagging Model in 4-Jet Events for E_{rel}	147
B.53	Closure Plot Comparing MC Truth and the Efficiency Corrected g -Tagging Model in 4-Jet Events for Δ_θ	147

B.54 Closure Plot Comparing MC Truth and the Efficiency Corrected g -Tagging Model in 4-Jet Events for m_{gg}	147
B.55 Closure Plot Comparing MC Truth and the Efficiency Corrected g -Tagging Model in 4-Jet Events for ϕ_{\parallel}	148
B.56 Closure Plot Comparing MC Truth and the Efficiency Corrected g -Tagging Model in 4-Jet Events for $\ln(k_t^2/m_{vis}^2)$	148
B.57 Closure Plot Comparing MC Truth and the Efficiency Corrected g -Tagging Model in 4-Jet Events for $p_{\perp,A}^2$	148
B.58 Gluon Splitting Distribution Comparison in MC and Data for Phase Space Region A	149
B.59 Gluon Splitting Distribution Comparison in MC and Data for Phase Space Region B	150
B.60 Gluon Splitting Distribution Comparison in MC and Data for Phase Space Region C	151
B.61 Gluon Splitting Distribution Comparison in MC and Data for Phase Space Region D	152

(This page is intentionally left blank).

List of Tables

3.1	Mapping between the Code in <code>SagTypeNr</code> and the Type of Residence	29
3.2	Basic Cuts	33
3.3	Side Door Mapping.	33
3.4	Street Mapping	33
3.5	Number of Observations in the Housing Dataset	36
3.6	Number of Observations in the Housing Dataset for the Tight Selection	36
3.7	Results from the Initial Hyperparameter Optimization for Apartments	37
3.8	Results from the Initial Hyperparameter Optimization for Houses	37
3.9	PDFs Used in the Random Search	39
3.10	Realtors' MAD	41
3.11	Performance Metrics for Apartments	43
3.12	Performance Metrics for Houses	43
5.1	The number of n -jet events for Data and MC.	64
5.2	Mapping Between the Flavor and the Variable Flevt	64
5.3	Number of Different Types of Jets for MC and MCb for n -Jet Events	65
5.4	Random Search PDFs for LGB	69
5.5	Global SHAP Feature Importances for the g -Tagging Models in 4-Jet Events	77
5.6	Generalized Angularities Systematic Errors, Charged Tracks	85
5.7	Generalized Angularities Systematic Errors, Neutral Clusters	85
5.8	Gluon Splitting Systematic Errors	90
5.9	Rgion Definition in the $R_{gg}^{k_t}$ - R_{gg}^{CA} Phase Space	90
A.1	Energy Rating Mapping	112
A.2	Final Variable List	114
A.3	Initial Hyperparameter Optimization Results for Apartments – SE Loss Function	115
A.4	Initial Hyperparameter Optimization Results for Apartments – Log-Cosh Loss Function	115
A.5	Initial Hyperparameter Optimization Results for Apartments – Cauchy Loss Function	115
A.6	Initial Hyperparameter Optimization Results for Apartments – Welsch Loss Function	116
A.7	Initial Hyperparameter Optimization Results for Apartments – Fair Loss Function	116
A.8	Initial Hyperparameter Optimization Results for Houses – SE Loss Function	116
A.9	Initial Hyperparameter Optimization Results for Houses – Log-Cosh Loss Function	117
A.10	Initial Hyperparameter Optimization Results for Houses – Cauchy Loss Function	117
A.11	Initial Hyperparameter Optimization Results for Houses – Welsch Loss Function	117
A.12	Initial Hyperparameter Optimization Results for Houses – Fair Loss Function	117
A.13	Performance Metrics for Apartments with the Tight Selection	121
A.14	Performance Metrics for Houses with the Tight Selection	121

B.1	Number of different types of jets for MC and MC _b written in relative numbers such that each row sum to 100 %. See also Table 5.3.	123
B.2	Random Search PDFs for XGB	123
B.3	Global SHAP Feature Importances for the <i>g</i> -Tagging Models in 3-Jet Events	123

Bibliography

- [1] Advanced Topics in Machine Learning (ATML). URL <https://kurser.ku.dk/course/ndak15014u>.
- [2] Allstate Claims Severity - Fair Loss. URL <https://kaggle.com/c/allstate-claims-severity>.
- [3] Dmlc/xgboost. URL <https://github.com/dmlc/xgboost>.
- [4] HEP meets ML award | The Higgs Machine Learning Challenge. URL <https://higgsml.lal.in2p3.fr/prizes-and-award/award/>.
- [5] The Large Electron-Positron Collider | CERN. URL <https://home.cern/science/accelerators/large-electron-positron-collider>.
- [6] Microsoft/LightGBM. URL https://github.com/microsoft/LightGBM/blob/b397555d7023fd05f8e56326905fe7b185109de/src/treelearner/serial_tree_learner.cpp#L282.
- [7] Scikit-hep/uproot. URL <https://github.com/scikit-hep/uproot>.
- [8] Datashader: Revealing the Structure of Genuinely Big Data. URL <https://github.com/holoviz/datashader>.
- [9] O . Wwww.OIS.dk - Din genvej til ejendomsdata. URL <https://www.ois.dk/>.
- [10] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems. URL <http://tensorflow.org/>.
- [11] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning From Data*. AMLBook. ISBN 978-1-60049-006-4.
- [12] D. Albanese, S. Riccadonna, C. Donati, and P. Franceschi. A practical tool for maximal information coefficient analysis. 7. ISSN 2047-217X. doi: 10.1093/gigascience/giy032. URL <https://doi.org/10.1093/gigascience/giy032>.
- [13] H. Albrecht, H. Ehrlichmann, T. Hamacher, R. P. Hofmann, T. Kirchhoff, A. Nau, S. Nowak, H. Schröder, H. D. Schulz, M. Walter, R. Wurth, C. Hast, H. Kolanoski, A. Kosche,

- A. Lange, A. Lindner, R. Mankel, M. Schieber, T. Siegmund, B. Spaan, H. Thurn, D. Töpfer, D. Wegener, M. Bittrner, P. Eckstein, M. Paulini, K. Reim, H. Wegener, R. Eckmann, R. Mundt, T. Oest, R. Reiner, W. Schmidt-Parzefall, W. Funk, J. Stiewe, S. Werner, K. Ehret, W. Hofmann, A. Hüpper, S. Khan, K. T. Knöpfle, M. Seeger, J. Spengler, D. I. Britton, C. E. K. Charlesworth, K. W. Edwards, E. R. F. Hyatt, H. Kapitza, P. Krieger, D. B. MacFarlane, P. M. Patel, J. D. Prentice, P. R. B. Saull, K. Tzamariudaki, R. G. Van de Water, T. S. Yoon, D. Reßing, M. Schmidtler, M. Schneider, K. R. Schubert, K. Strahl, R. Waldi, S. Weseler, G. Kernel, P. Križnič, T. Podobnik, T. Živko, V. Balagura, I. Belyaev, S. Chechelnitsky, M. Danilov, A. Droutskoy, Y. Gershtein, A. Golutvin, G. Kostina, D. Litvintsev, V. Lubimov, P. Pakhlov, F. Ratnikov, S. Semenov, A. Snizhko, V. Soloshenko, I. Tichomirov, and Y. Zaitsev. A model-independent determination of the inclusive semileptonic decay fraction of B mesons. 318(2): 397–404. ISSN 0370-2693. doi: 10.1016/0370-2693(93)90146-9. URL <http://www.sciencedirect.com/science/article/pii/0370269393901469>.
- [14] E. Anderson. The Species Problem in Iris. 23(3):457–509. ISSN 00266493. doi: 10.2307/2394164. URL www.jstor.org/stable/2394164.
 - [15] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand. Parton fragmentation and string dynamics. 97(2): 31–145. ISSN 0370-1573. doi: 10.1016/0370-1573(83)90080-7. URL <http://www.sciencedirect.com/science/article/pii/0370157383900807>.
 - [16] S. R. Armstrong. A Search for the standard model Higgs boson in four jet final states at center-of-mass energies near 183-GeV with the ALEPH detector at LEP. URL <http://wwwlib.umi.com/dissertations/fullcit?p9910371>.
 - [17] M. Awad and R. Khanna. Support Vector Regression. In M. Awad and R. Khanna, editors, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, pages 67–80. Apress. ISBN 978-1-4302-5990-9. doi: 10.1007/978-1-4302-5990-9_4. URL https://doi.org/10.1007/978-1-4302-5990-9_4.
 - [18] R. J. Barlow. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series)*. WileyBlackwell, reprint edition. ISBN 0-471-92295-1. URL <http://www.amazon.co.uk/Statistics-Statistical-Physical-Sciences-Manchester/dp/0471922951%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0471922951>.

- [19] J. T. Barron. A General and Adaptive Robust Loss Function. URL <http://arxiv.org/abs/1701.03077>.
- [20] W. Bartel et al. Experimental study of jets in electron-positron annihilation. 101(1):129–134. ISSN 0370-2693. doi: 10.1016/0370-2693(81)90505-0. URL <http://www.sciencedirect.com/science/article/pii/0370269381905050>.
- [21] E. Becht, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Evaluation of UMAP as an alternative to t-SNE for single-cell data. page 298430, . doi: 10.1101/298430. URL <https://www.biorxiv.org/content/10.1101/298430v1>.
- [22] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. 37(1):38–44, . ISSN 1546-1696. doi: 10.1038/nbt.4314. URL <https://www.nature.com/articles/nbt.4314>.
- [23] J. Bergstra and Y. Bengio. Random Search for Hyper-parameter Optimization. 13:281–305. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=2188385.2188395>.
- [24] C. Bierlich. Rope hadronization, geometry and particle production in pp and pA Collisions. URL <https://lup.lub.lu.se/search/ws/files/18474576/thesis.pdf>.
- [25] Bolighed. Bolighed - usikkerhed i data-vurderingen. URL <https://bolighed.dk/om-bolighed/spoergsmaal-og-svar/#boligvaerdi>.
- [26] L. Breiman. Random Forests. 45(1):5–32. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL <https://doi.org/10.1023/A:1010933404324>.
- [27] R. P. Brent. *Algorithms for Minimization without Derivatives*. Prentice-Hall Series in Automatic Computation. Prentice-Hall. URL <https://cds.cern.ch/record/113464>.
- [28] E. Brochu, V. M. Cora, and N. de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. URL <http://arxiv.org/abs/1012.2599>.
- [29] R. Brun and F. Rademakers. ROOT — An object oriented data analysis framework. 389(1):81–86. ISSN 0168-9002. doi: 10.1016/S0168-9002(97)00048-X. URL <http://www.sciencedirect.com/science/article/pii/S016890029700048X>.
- [30] A. Buckley, J. Butterworth, S. Gieseke, D. Grellscheid, S. Hoche, H. Hoeth, F. Krauss, L. Lonnblad, E. Nurse, P. Richardson, S. Schumann, M. H. Seymour, T. Sjostrand,

- P. Skands, and B. Webber. General-purpose event generators for LHC physics. 504(5):145–233. ISSN 03701573. doi: 10.1016/j.physrep.2011.03.005. URL <http://arxiv.org/abs/1101.2599>.
- [31] C. Burgard. Standard model of physics | TikZ example. URL <http://www.texample.net/tikz/examples/model-physics/>.
 - [32] D. Buskulic et al. An investigation of B_{d0} and B_{s0} oscillation. 322(4):441–458. ISSN 0370-2693. doi: 10.1016/0370-2693(94)91177-0. URL <http://www.sciencedirect.com/science/article/pii/0370269394911770>.
 - [33] M. Cacciari, G. P. Salam, and G. Soyez. FastJet user manual. 72(3):1896. ISSN 1434-6044, 1434-6052. doi: 10.1140/epjc/s10052-012-1896-2. URL <http://arxiv.org/abs/1111.6097>.
 - [34] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber. Longitudinally-invariant kt-clustering algorithms for hadron-hadron collisions. 406(1):187–224, . ISSN 0550-3213. doi: 10.1016/0550-3213(93)90166-M. URL <http://www.sciencedirect.com/science/article/pii/055032139390166M>.
 - [35] S. Catani, G. Turnock, and B. R. Webber. Jet broadening measures in e+ e- annihilation. B295:269–276, . doi: 10.1016/0370-2693(92)91565-Q.
 - [36] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. pages 785–794. doi: 10.1145/2939672.2939785. URL <http://arxiv.org/abs/1603.02754>.
 - [37] A. Collaboration. Electron efficiency measurements with the ATLAS detector using 2012 LHC proton-proton collision data. 77(3):195, . ISSN 1434-6044, 1434-6052. doi: 10.1140/epjc/s10052-017-4756-2. URL <http://arxiv.org/abs/1612.01456>.
 - [38] C. Collaboration. Search for a Higgs boson in the decay channel H to ZZ(*) to q qbar l-l+ in pp collisions at sqrt(s) = 7 TeV. 2012(4):36, . ISSN 1029-8479. doi: 10.1007/JHEP04(2012)036. URL <http://arxiv.org/abs/1202.1416>.
 - [39] T. A. Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. 716(1):1–29, . ISSN 03702693. doi: 10.1016/j.physletb.2012.08.020. URL <http://arxiv.org/abs/1207.7214>.
 - [40] T. C. Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. 716(1):30–61, . ISSN 03702693. doi: 10.1016/j.physletb.2012.08.021. URL <http://arxiv.org/abs/1207.7235>.
 - [41] M. Dam. An upper limit for Br(Z->ggg) from symmetric 3-jet zo hadronic decays. 389(2):405–415. ISSN 0370-2693. doi: 10.1016/S0370-2693(96)01450-5.

- [42] A. Diaz-Papkovich, L. Anderson-Trocmé, C. Ben-Eghan, and S. Gravel. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. 15(11). ISSN 1553-7390. doi: 10.1371/journal.pgen.1008432. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6853336/>.
- [43] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber. Better Jet Clustering Algorithms. 1997(08):001–001. ISSN 1029-8479. doi: 10.1088/1126-6708/1997/08/001. URL <http://arxiv.org/abs/hep-ph/9707323>.
- [44] S. D. DST. Price Index (EJ14) - Statistics Denmark. URL <https://www.dst.dk/en/Statistik/emner/priser-og-forbrug/ejendomme>.
- [45] S. D. Ellis and D. E. Soper. Successive combination jet algorithm for hadron collisions. 48(7):3160–3166. doi: 10.1103/PhysRevD.48.3160. URL <https://link.aps.org/doi/10.1103/PhysRevD.48.3160>.
- [46] D. et al. Buskulic. A precise measurement of hadrons. 313(3): 535–548. ISSN 0370-2693. doi: 10.1016/0370-2693(93)90028-G. URL <http://www.sciencedirect.com/science/article/pii/037026939390028G>.
- [47] F. Faye. Frederik Faye / deepcalo. URL <https://gitlab.com/ffaye/deepcalo>.
- [48] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. 7(2):179–188. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x>.
- [49] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In P. Vitányi, editor, *Computational Learning Theory*, pages 23–37. Springer Berlin Heidelberg. ISBN 978-3-540-49195-8. Adaboost.
- [50] S. L. Glashow. Partial-symmetries of weak interactions. 22(4): 579–588. ISSN 0029-5582. doi: 10.1016/0029-5582(61)90469-2. URL <http://www.sciencedirect.com/science/article/pii/0029558261904692>.
- [51] P. Gras, S. Höche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer, A. Siódmok, P. Skands, G. Soyez, and J. Thaler. Systematics of quark/gluon tagging. 2017(7):91. ISSN 1029-8479. doi: 10.1007/JHEP07(2017)091. URL <http://arxiv.org/abs/1704.03878>.
- [52] N. Guttenberg, N. Virgo, O. Witkowski, H. Aoki, and R. Kanai. Permutation-equivariant neural networks applied to dynamics prediction. URL <http://arxiv.org/abs/1612.04530>.

- [53] A. E. Harvey and S. Peters. Estimation Procedures for Structural Time Series Models. doi: [10.1002/for.3980090203](https://doi.org/10.1002/for.3980090203).
- [54] T. Hastie and R. Tibshirani. Generalized Additive Models: Some Applications. 82(398):371–386. ISSN 01621459. doi: [10.2307/2289439](https://doi.org/10.2307/2289439). URL www.jstor.org/stable/2289439.
- [55] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer-Verlag, 2 edition. ISBN 978-0-387-84857-0. URL [//www.springer.com/la/book/9780387848570](http://www.springer.com/la/book/9780387848570).
- [56] K. Hornik. Approximation capabilities of multilayer feedforward networks. 4(2):251–257. ISSN 0893-6080. doi: [10.1016/0893-6080\(91\)90009-T](https://doi.org/10.1016/0893-6080(91)90009-T). URL <http://www.sciencedirect.com/science/article/pii/089360809190009T>.
- [57] P. Huber and E. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley. ISBN 978-1-118-21033-8. URL https://books.google.dk/books?id=j10hquR_j88C.
- [58] S. Hviid, Juul. Working Paper: A regional model of the Danish housing market. URL <http://www.nationalbanken.dk/en/publications/Pages/2017/11/Working-Paper-A-regional-model-of-the-Danish-housing-market.aspx>.
- [59] F. James and M. Roos. Minuit – a system for function minimization and analysis of the parameter errors and correlations. 10:343–367. doi: [10.1016/0010-4655\(75\)90039-9](https://doi.org/10.1016/0010-4655(75)90039-9).
- [60] R. E. Kalman. A new approach to linear filtering and prediction problems. 82:35–45.
- [61] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc. URL <http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf>.
- [62] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. URL <http://arxiv.org/abs/1412.6980>.
- [63] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-Normalizing Neural Networks. URL <http://arxiv.org/abs/1706.02515>.
- [64] A. J. Larkoski, J. Thaler, and W. J. Waalewijn. Gaining (Mutual) Information about Quark/Gluon Discrimination. 2014 (11). ISSN 1029-8479. doi: [10.1007/JHEP11\(2014\)129](https://doi.org/10.1007/JHEP11(2014)129). URL <http://arxiv.org/abs/1408.3122>.

- [65] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. 49(4):764–766. ISSN 0022-1031. doi: 10.1016/j.jesp.2013.03.013. URL <http://www.sciencedirect.com/science/article/pii/S0022103113000668>.
- [66] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, pages 4768–4777. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL <http://dl.acm.org/citation.cfm?id=3295222.3295230>.
- [67] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles - SHAP. . URL <http://arxiv.org/abs/1802.03888>.
- [68] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. . URL <http://arxiv.org/abs/1802.03888>.
- [69] A. L. Maas. Rectifier nonlinearities improve neural network acoustic models.
- [70] L. McInnes and J. Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. URL <http://arxiv.org/abs/1802.03426>.
- [71] T. C. Mills. *Time Series Techniques for Economists / Terence c. Mills*. Cambridge University Press Cambridge [England] ; New York. ISBN 0-521-34339-9 0-521-40574-2. URL <http://www.loc.gov/catdir/toc/cam031/89007187.html>.
- [72] N. Mohd Razali and B. Yap. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. 2.
- [73] I. Mulalic, H. Rasmussen, J. Rouwendal, and H. H. Woltmann. The Financial Crisis and Diverging House Prices: Evidence from the Copenhagen Metropolitan Area. ISSN 1556-5068. doi: 10.2139/ssrn.3041272. URL <https://www.ssrn.com/abstract=3041272>.
- [74] Particle Data Group et al. Review of Particle Physics. 98(3):030001. doi: 10.1103/PhysRevD.98.030001. URL <https://link.aps.org/doi/10.1103/PhysRevD.98.030001>.
- [75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. 12:2825–2830.

- [76] E. Polley and M. van der Laan. Super Learner In Prediction. URL <https://biostats.bepress.com/ucbbiostat/paper266>.
- [77] J. Proriol, J. Jousset, C. Guicheney, A. Falvard, P. Henrard, D. Pallin, P. Perret, and B. Brandl. TAGGING B QUARK EVENTS IN ALEPH WITH NEURAL NETWORKS (comparison of different methods : Neural Networks and Discriminant Analysis). page 27.
- [78] A. Purcell. Go on a particle quest at the first CERN webfest. URL <https://cds.cern.ch/record/1473657>.
- [79] S. Ravanbakhsh, J. Schneider, and B. Poczos. Deep Learning with Sets and Point Clouds. URL <http://arxiv.org/abs/1611.04500>.
- [80] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting Novel Associations in Large Data Sets. 334(6062):1518–1524. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1205438. URL <http://science.sciencemag.org/content/334/6062/1518>.
- [81] J. W. Rohlfs. *Modern Physics from A to Z*. John Wiley and Sons. ISBN 978-0-471-57270-1.
- [82] P. J. Rousseeuw and C. Croux. Alternatives to the Median Absolute Deviation. 88(424):1273–1283. ISSN 0162-1459. doi: 10.1080/01621459.1993.10476408. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476408>.
- [83] A. Salam. Weak and electromagnetic interactions. In *Selected Papers of Abdus Salam*, volume Volume 5 of *World Scientific Series in 20th Century Physics*, pages 244–254. WORLD SCIENTIFIC. ISBN 978-981-02-1662-7. doi: 10.1142/9789812795915_0034. URL https://www.worldscientific.com/doi/abs/10.1142/9789812795915_0034.
- [84] S. Schmitt. Data Unfolding Methods in High Energy Physics. 137:11008. ISSN 2100-014X. doi: 10.1051/epjconf/201713711008. URL <http://arxiv.org/abs/1611.01927>.
- [85] L. Scodellaro. B tagging in ATLAS and CMS. URL <http://arxiv.org/abs/1709.01290>.
- [86] L. Shapley. A value for n-person games. In *The Shapley Value*, volume 28 of *Annals of Math Studies*, pages 307–317. doi: 10.1017/CBO9780511528446.003.
- [87] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. De-sai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands. An Introduction to PYTHIA 8.2. 191:159–177. ISSN 00104655. doi: 10.1016/j.cpc.2015.01.024. URL <http://arxiv.org/abs/1410.3012>.

- [88] P. Skands. Peter Skands.
- [89] S. J. Taylor and B. Letham. Forecasting at scale. doi: 10.7287/peerj.preprints.3190v2. URL <https://peerj.com/preprints/3190>.
- [90] i. team. Iminuit – A python interface to minuit. URL <https://github.com/scikit-hep/iminuit>.
- [91] J. Thaler. Report of the Les Houches Quark/Gluon Subgroup. (1):28.
- [92] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. 58(1):267–288. ISSN 0035-9246. URL www.jstor.org/stable/2346178.
- [93] A. Tikhonov. *On the Stability of Inverse Problems*, volume vol. 39 of *Doklady Akademii Nauk SSSR*.
- [94] S. Toghi Eshghi, A. Au-Yeung, C. Takahashi, C. R. Bolen, M. N. Nyachienga, S. P. Lear, C. Green, W. R. Mathews, and W. E. O’Gorman. Quantitative Comparison of Conventional and t-SNE-guided Gating Analyses. 10. ISSN 1664-3224. doi: 10.3389/fimmu.2019.01194. URL <https://www.frontiersin.org/articles/10.3389/fimmu.2019.01194/full>.
- [95] d. L. M. J. van, E. C. Polley, and A. E. Hubbard. Super Learner. 6(1). ISSN 1544-6115. doi: 10.2202/1544-6115.1309. URL <https://www.degruyter.com/view/j/sagmb.2007.6.issue-1/sagmb.2007.6.1.1309/sagmb.2007.6.1.1309.xml>.
- [96] J. J. van der Bij and E. W. N. Glover. Z boson production and decay via gluons. 313(2):237–257. ISSN 0550-3213. doi: 10.1016/0550-3213(89)90317-9. URL <http://www.sciencedirect.com/science/article/pii/0550321389903179>.
- [97] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. 9:2579–2605. ISSN ISSN 1533-7928. URL <http://www.jmlr.org/papers/v9/vandermaaten08a.html>.
- [98] F. van Veen. The Neural Network Zoo. URL <http://www.asimovinstitute.org/neural-network-zoo/>.
- [99] V. Vapnik. Principles of Risk Minimization for Learning Theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS’91, pages 831–838. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-222-9. URL <http://dl.acm.org/citation.cfm?id=2986916.2987018>.
- [100] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore,

J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0—Fundamental Algorithms for Scientific Computing in Python. page arXiv:1907.10121. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv190710121V/abstract>.

- [101] I. Wallach and R. Lilien. The protein–small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. 25(5):615–620. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp035. URL <https://academic.oup.com/bioinformatics/article/25/5/615/183421>.
- [102] S. Weinberg. A Model of Leptons. 19(21):1264–1266. doi: 10.1103/PhysRevLett.19.1264. URL <https://link.aps.org/doi/10.1103/PhysRevLett.19.1264>.
- [103] H. Wickham. Tidy data. 59(10):1–23. ISSN 1548-7660. doi: 10.18637/jss.v059.i10. URL <https://www.jstatsoft.org/v059/i10>.
- [104] M. Wobisch and T. Wengler. Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering. URL <http://arxiv.org/abs/hep-ph/9907280>.
- [105] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola. Deep Sets. URL <http://arxiv.org/abs/1703.06114>.