UNIVERSITY OF COPENHAGEN

A Physicist's Approach To Machine Learning

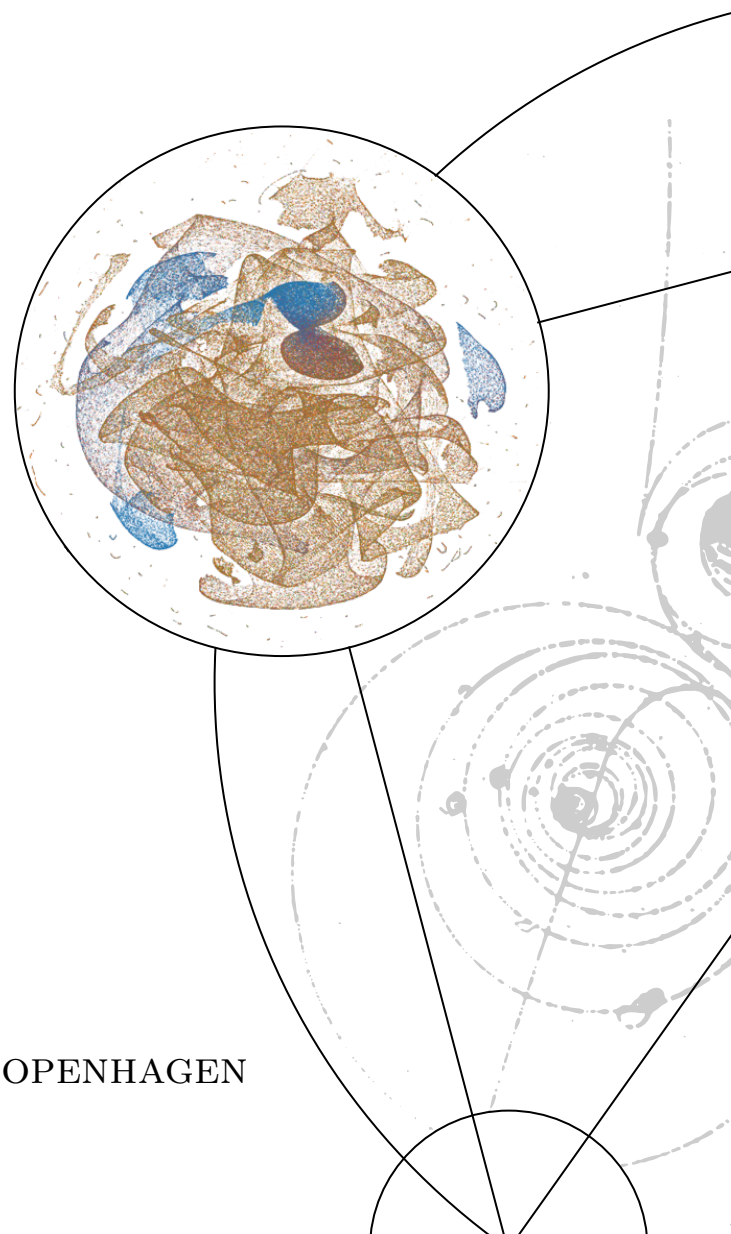
Understanding The Basic Bricks

Christian Michelsen

Master's Thesis (Cand. Scient.)

January 3rd, 2020

Supervised by
Troels Petersen



University of Copenhagen

Christian Michelsen

https://github.com/christianmichelsen

This thesis was inspired by the works of Edward R. Tufte using the Tufte-LaTeX package.

First printing, December 2019

Abstract

Here will be a decent abstract at some point^{TM}.

Contents

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.

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. This 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 look and behave. The results from this chapter are discussed in chapter 6 along with further work, and finally both Part I and Part II are concluded in chapter 7.

5. Quark Gluon Analysis

"Research is what I am doing I don't know know what I'm doing."

— Wernher von Braun

The analysis of the quarks and gluons that were introduced in the previous chapter is described here. The overall goal is to be able to discriminate between quark and gluon jets to better be able to describe the gluon jets. The gluon jet distributions are measured in 3-jet events and how they split in 4-jet events. This chapter is organized as follows. In section 5.1 the data are presented and the initial cuts are described and the variables are visualized in section 5.2. The choices of loss and evaluation function are discussed in section 5.3. In section 5.4 the first of the two overall models developed in this chapter is presented, the b-tagging model. The efficiency of this model is measured in section 5.5. The second model, the g-tagging model is used for classification of entire events, compared to the b-tagging model which classifies individual jets, and is introduced in section 5.6 and its efficiency is measured in section 5.7. The jet distributions in 3-jet events are analyzed in section 5.8 and their the gluon splitting in 4-jet events in section 5.9.

5.1 Data Preprocessing

The data files were acquired from Prof. Peter Hansen (NBI) who worked on the ALEPH expiriment. The Data consists of 43 data files from between 1991 and 1995 totalling 3.5 GB (Data). Along with this comes 125 files based on Monte Carlo (MC) simulations (8.4 GB) and additional 42 MC-files with only b-quark events (MCb) (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 cuts are applied before exporting the data to HDF5-format. The first one being that the (center of mass) energy E in the event has to be within $90.8\,\text{GeV} \leq E \leq 91.6\,\text{GeV}$ to only use the Z peak data. The second one being that the sum of the momenta p_{sum} in each event is $32\,\text{GeV} \leq p_{\text{sum}}$ to remove any $Z \to \tau^+\tau^-$ events. To ensure a primary vertex, at least two good tracks are required 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 trust 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 cuts were standard requirements for the ALEPH experiment.

One last cut which was experimented with was the 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 more than then threshold value. Higher thresholds means 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 other jets were matched[1].

[1] 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.

The data structure is quite differently structured in the Ntuples compared to normal structured data in the form of tidy data [99]. 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 unjagged[2] 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.

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

The number of events for each number of jets can be seen in Table 5.1 for the Data and in Figure 5.2 for the MC and MCb.

n	# jets	# events
2	2 359 738	1 179 869
3	3 619 290	1 206 430
4	854 336	213 584
5	52 775	10 555
6	510	85
Total	6 886 649	2 610 523

Table 5.1: The dimensions of the dataset for the actual Data for n-jet events. The numbers in the jet columns are the number of events multiplied with the number of jets; e.g. $85 \cdot 6 = 510$.

5.2 *Exploratory Data Analysis*

Since the machine learning models are only trained on the three vertex variables `projet`, `bqvjet`, and `ptljet` – see chapter 4 for a deeper introduction to these variables – these variables will be the primary focus of this section. Given the fact that MC-simulated data exists, the truth of each simulated event is also known. This allows us visualize the difference between the different types of quarks. In the MC simulation each event are generated such that the type of quark, or *flavor*, is known and assigned the variable `flevt`. The mapping from flavor to `flevt` is:

Flavor:	*bb*	*cc*	*ss*	*dd*	*uu*
`flevt`:	5	4	3	2	1

n	# jets	# events
2	7 293 594	3 646 797
3	10 780 890	3 593 630
4	2 241 908	560 477
5	103 820	20 764
6	588	98
Total	20 420 800	7 821 766

Table 5.2: The dimensions for the MC and MCb datasets for n-jet events.

In addition to knowing the correct flavor, 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 quarks. Similarly we define *c*-jets only with the change that `flevt` $= 4$, and *uds*-jets[3] with `flevt` $\in \{1, 2, 3\}$. A gluon jet is defined to be a jet in a (any-flavor) q-matched event where

[3] Will also be called a *l*-jet.

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" which is 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.

n	b	c	uds/l	g	non-q-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.

With 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. For all three subplots the histograms are shown with logarithmic y-axes, all b-jets in blue, c-jets in red, g-jets in green, and all of the jets (no matter their type) are shown in orange. The distributions for 2-jet events are shown in fully opaque color, 3-jet events in dashed lines, and 4-jet events in lighter colors. In the left subplot the `projet` variable is plotted where it can be seen 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 with mostly with gluons[4] and large values are mostly due to b-jets. In general it is clear to see how the differences in distribution between the 2-, 3-, and 4-jet events are minor, with the one exception of 2-jet events which does not contain any gluons at all.

[4] Around 98 % of all g-jets are zeros for the variable `ptljet` compared to $\sim$ 82 % for c-jets and $\sim$ 70 % for b-jets.

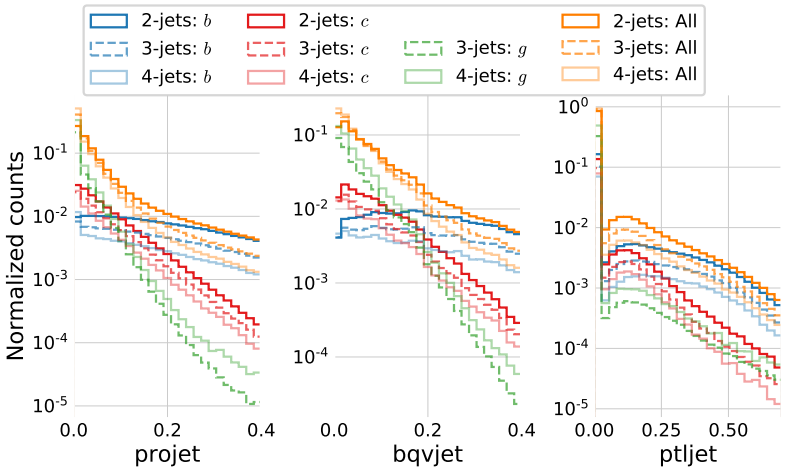


Figure 5.1: Normalized histograms of the three vertex variables: `projet`, `bqvjet`, and `ptljet`. In blue colors the variables are shown for true b-jets, in red for true c-jets, in green for true g-jets, and in orange for all of the jets (including non q-matched). In fully opaque color are shown the 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 not matter how many jets.

5.2.1 *Dimensionality Reduction*

Even though there are only three vertex variables, it is difficult to properly get an intuition about how easily separated the different types of jets are. Since there are millions of points a single 3D scatter plot quickly becomes overcrowded if one plots all jets. We apply dimensionality reduction from the three dimensions down to two dimensions by using the UMAP algorithm [69]. 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 [93] deserves an honorable mention since this algorithm revolutionized the usage of (nonlinear) dimensionality reduction algorithms in e.g. bioinformatics [91, 97] yet its mathematical foundation has strongly been improved with the newer, faster UMAP algorithm [69] which usage is also expanding [21, 22, 41].

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. It does so by trying to 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 such that both global structure and local structure is preserved [69]. Compared to t-SNE the approach in UMAP has a topological background compared to the more heuristic approach taken by t-SNE. Note that the UMAP algorithm is not provided any information about which jets are which types or any other truth information.

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 tightly together UMAP is allowed to cluster the points in the low-dimensional representation. To properly choose 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. Here the millions of points are plotted using Datashader [8] to avoid overplotting and colored according to the jet type. From the figure it is seen how there are some clear *b*-jet clusters, however, most of the data seem to be a mix of *g*-and *uds*-jets. The plots with the same UMAP parameters for 3-jet and 2-jet events are seen in Figure 5.3 and 5.4.

These figures suggest that it should be possible to discriminate the *b*-jets from the other jets somewhat, however, no clear separa-

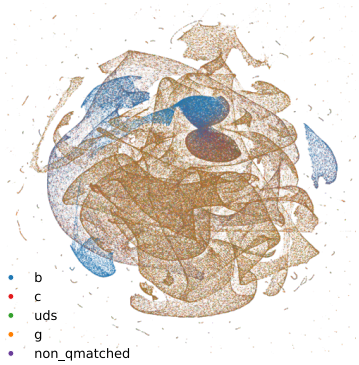


Figure 5.2: Visualization of the vertex variables in 4-jet events for the different categories: true b-jets in blue, true c-jets in red, true uds-jets in green, true g-jets in orange, and non q-matched events in purple. The clustering is performed with the UMAP algorithm which outputs a 2D-projection.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

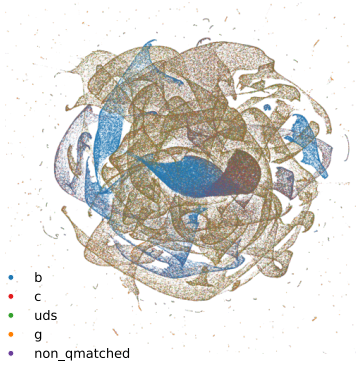


Figure 5.3: UMAP visualization of vertex variables for 3-jet events.

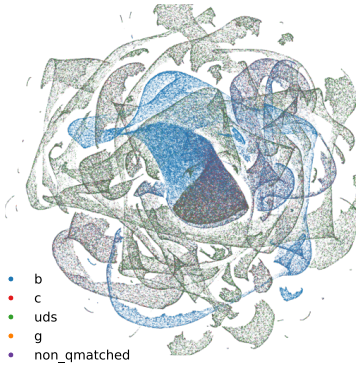
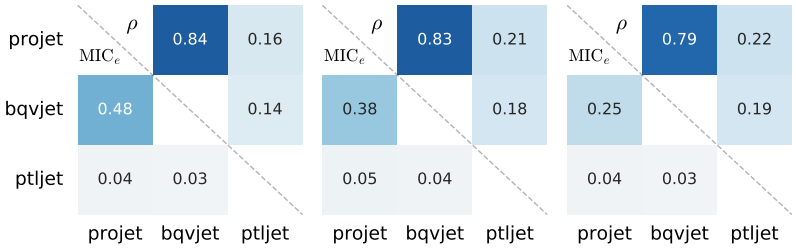


Figure 5.4: UMAP visualization of vertex variables for 2-jet events.

tion 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 correlation between the vertex variables can be seen in Figure 5.5, 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. Here it can be seen that `projet` and `bqvjet` are the two variables that correlate the most whereas the other variables correlate a lot 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.



(a) 2-jet events (b) 3-jet events (c) 4-jet events

Figure 5.5: Correlation of the three vertex variables for 2-, 3- and 4-jet events. Here the upper diagonal shows the linear correlation ρ and the lower diagonal the nonlinear correlations MIC_e.

5.3 *Loss and Evaluation Function*

In contrary to the housing prices project, the goal in this project is to predict types of jets or events where the *signal* observations[5] are often assigned the label 1 and *background* observations 0. The combination of this being a *classification* problem (compared to a regression problem) along with the fact all the variables are actual measurements from a particle physics accelerator means that the issue of outliers is negligible. This also means that the problem of finding a robust loss function is less important since the in classification 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 *ROC*[6] 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:

$$\varepsilon_{\text{sig}} = \frac{S_{\text{sel}}}{S_{\text{tot}}}, \qquad \varepsilon_{\text{bkg}} = \frac{B_{\text{sel}}}{B_{\text{tot}}}, \tag{5.1}$$

[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.

[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 $\hat{y}_{\text{cut}}$.

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 function[7] 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 $\ell_{\log}$ is calculated as:

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

This is visualized in Figure 5.6. 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 -∞).

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

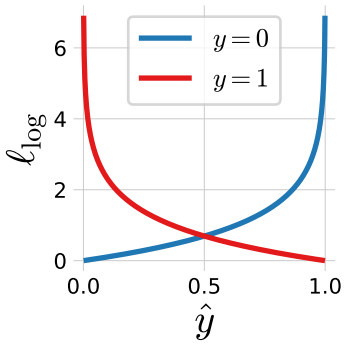


Figure 5.6: Plot of the log-loss $\ell_{\log}$ for background ($y = 0$) in blue and signal ($y = 1$) in red.

5.4 *b-Tagging Analysis*

The ability to discriminate between the different types of particles produced in a collision is obviously import to understand the results. Today much work go into tagging algorithms, e.g. b-tagging in ATLAS and CMS [82]. 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-violation[8] at LHC-b, contributes to the choice of tagging b-quarks. In ALEPH Proriol et al. [75] 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 and the outputted the prediction score `nnbjet`. From here on, this pre-trained network will be called NNB.

[8] Short for charge-parity.

I split the data into training and test sets in such a way that the individual jets in an event are not split. The events are split in a (80-20)% train-test ratio.

5.4.1 *b-Tagging Hyperparameter Optimization*

Compared to the housing prices dataset, the number of observations N is a lot larger ($3 \times 10^7 \gg 5 \times 10^5$), although the dimen-

sionality 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 hypermeter 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 slow[9] at fitting datasets of this size. The results of the HPO for 3-jet and 4-jet events can be seen in Figure 5.7. For 3-jets 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 seen in the 4-jet events.

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 70 for a discussion of the timings.

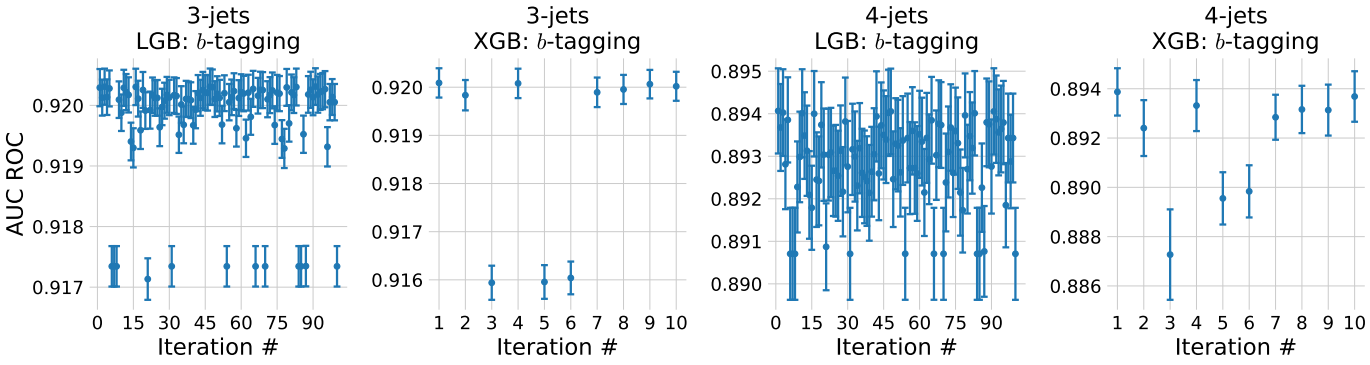


Figure 5.7: Hyperparameter Optimization results of b-tagging with random search. From left to right, we have A) 100 iterations of RS with LGB on 3-jets, B) 10 iterations of RS with XGB on 3-jets, C) 100 iterations of RS with LGB on 4-jets, D) 10 iterations of RS with XGB on 4-jets. Notice the different ranges on the y-axes.

The relationship between the different hyperparameters in 4-jet events can be seen in the parallel coordinate plot in Figure 5.8. First 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 $\text{int}(\sim 0.5 \cdot 3) = 1$ variable[10] 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.

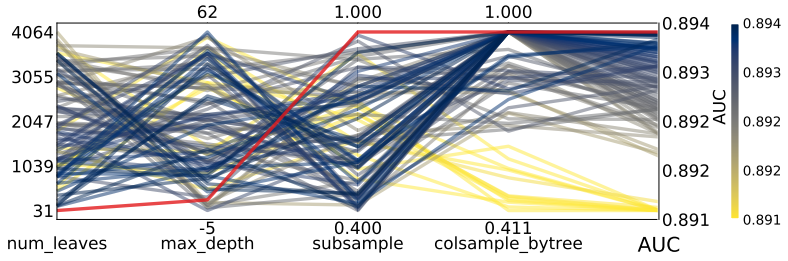


Figure 5.8: Hyperparameter optimization results of b-tagging for 4-jet events. The results are shown as parallel coordinates with each hyperparameter along the x-axis and the value of that parameter on the y-axis. Each line is an event in the 4-dimensional space colored according to the performance of 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 5.9 for 4-jet events and in B.4 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.

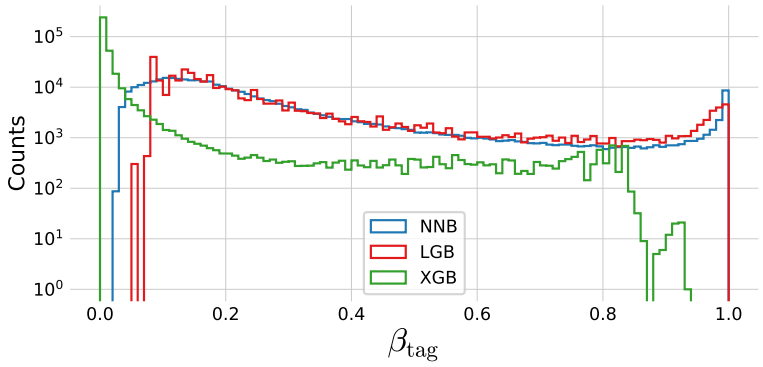


Figure 5.9: Histogram of b-tag scores β_{tag} in 4-jet events for NNB (the neural network pre-trained by ALEPH, also called `nnbjet`) in blue, LGB in red, and XGB in green.

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 5.10 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.5. 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.

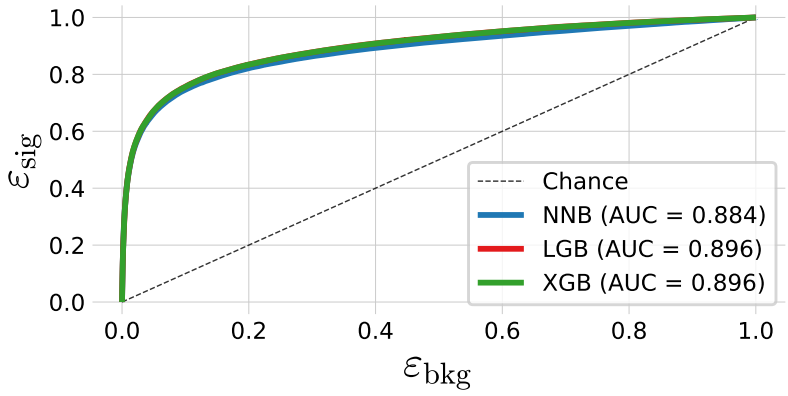


Figure 5.10: ROC curve of the three b-tag models in 4-jet events for NNB (the pre-trained neural network trained by ALEPH, also called `nnbjet`) in blue, LGB in red, and XGB in green. In the legend the area under curve (AUC) is also shown. Notice that the LGB and XGB models share performance and it is thus due to overplotting that only the green line for XGB can be seen.

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 HEP[11] compared to just 23 hours for 100 iterations for LGB. The same performance difference

[11] The local computing cluster.

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 5.11. 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.6.

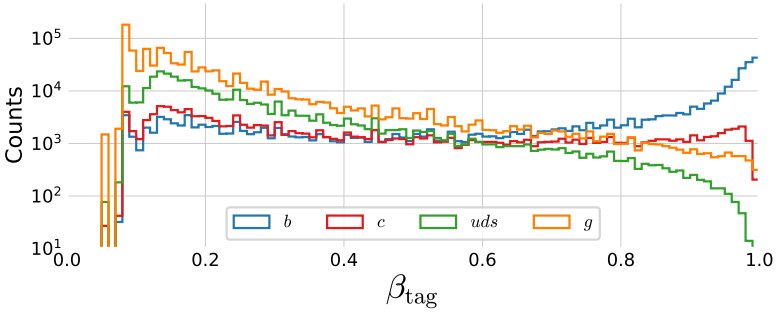


Figure 5.11: Distribution of b-tags in 4-jet events for b-jets in blue, c-jets in red, uds in green and g in orange.

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 5.12 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.

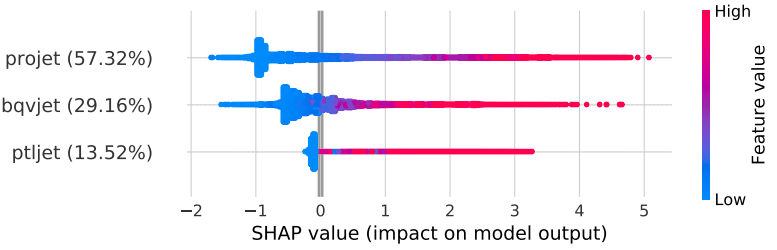


Figure 5.12: Global feature importances for the LGB b-tagging algorithm on 4-jet events. The normalized feature importance is shown in the parenthesis and the each dot is an observation showing the dependance between the SHAP value and the feature value.

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 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:

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

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

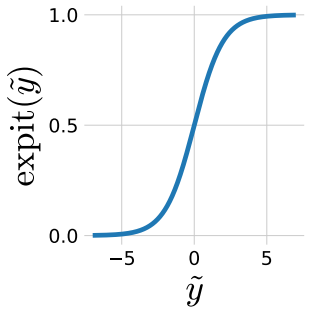


Figure 5.13: The expit function.

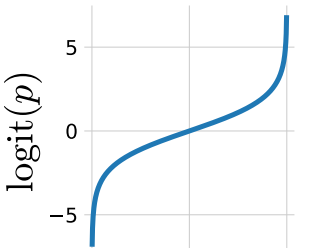


Figure 5.13. Its inverse is the *logit* function:

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

which is visualized in Figure 5.14. 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 5.12 are also in log-odds space[12].

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

With this in mind, single predictions of the LGB b-tagging model can be understood with SHAP which Figure 5.15 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 jets[13]. 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.

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

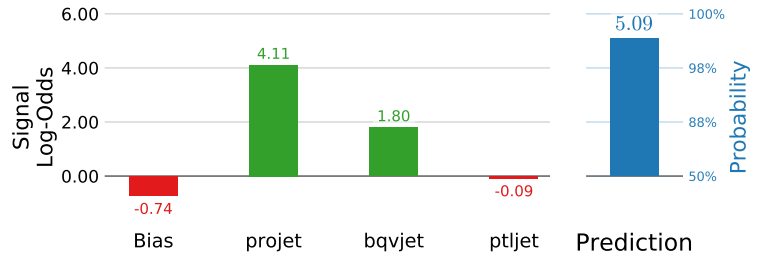


Figure 5.15: Model explanation for the 3-jet b-tagging LGB model for a b-like jet. The first column is the bias which can be seen as the naive prediction baseline, the rest are the input variables. On the right hand side of the plot is the model prediction shown. The left part of the plot is shown in 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}:

$$\varepsilon = \frac{N_{\text{sel}}}{N_{\text{truth}}}. \tag{5.5}$$

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 make use of the characteristic signature of the Z decay that the clear majority[14] of 3-jet decays are $Z \to q\bar{q}g$. This 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.

[14] The fraction of $Z \to ggg$ events are $< 1.1\,\%$ [72].

Based on Figure 5.11 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)[15]. 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).

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

The efficiencies can be seen in Figure 5.16. 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).

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 $P = J_3$.

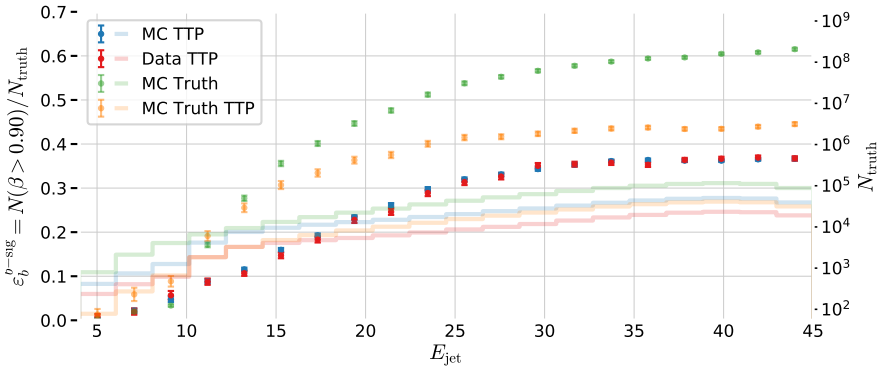


Figure 5.16: b-tag efficiency for b-jets in the b-signal region for 3-jet events, $\varepsilon_b^{b\text{-sig}}$, as a function of jet energy E_{jet}. In the plot the efficiencies are shown for MC TTP in blue, Data TTP in red, MC Truth TTP in green, and MC Truth TTP in orange. The efficiencies (the errorbars)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

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 5.17. 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.20 and one for g-jets in the b-signal region $\varepsilon_g^{b\text{-sig}}$ seen in Figure B.21.

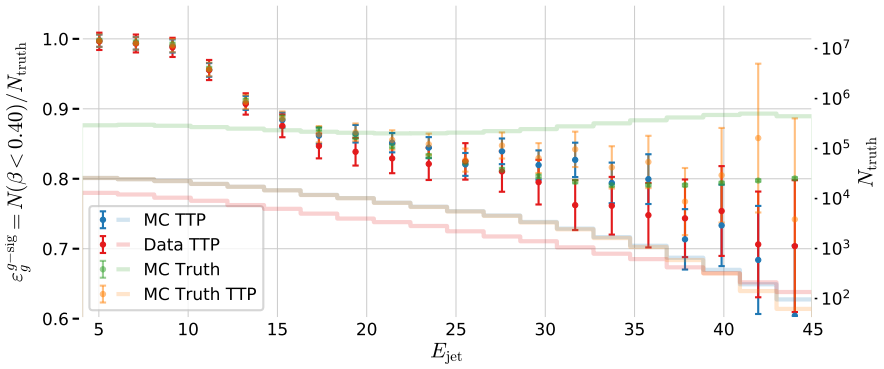


Figure 5.17: b-tag efficiency for g-jets in the g-signal region for 3-jet events, $\varepsilon_g^{g\text{-sig}}$, as a function of jet energy E_{jet}. In the plot the efficiencies are shown for MC TTP in blue, Data TTP in red, MC Truth TTP in green, and MC Truth TTP in orange. The efficiencies (the errorbars) can be read off on the left y-axis and the counts (histograms) on the right y-axis. Notice how both MC TTP and Data TTP follow each other closely until $\sim 25\,\text{GeV}$.

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).

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: $\boldsymbol{\beta}_{\text{tag}} = [\beta_{\text{tag}_1}, \beta_{\text{tag}_2}, \beta_{\text{tag}_3}, \beta_{\text{tag}_4}]$. Treating $\boldsymbol{\beta}_{\text{tag}}$ 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-matched[16] 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 $n-2$ low-

[16] 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.

est b-tag scores for n-jet events. An example of a signal event is $\boldsymbol{\beta}_{\text{tag}} = [0.95, 0.89, 0.15, 0.07]$ 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 $\boldsymbol{\beta}_{\text{tag}}$ had different feature importances when tested, even though they should be equally important. Instead of defining $\boldsymbol{\beta}_{\text{tag}}$ as a vector it should instead be seen as a set[17] $\boldsymbol{\beta}_{\text{tag}} \equiv \{\beta_{\text{tag}_1}, \ldots, \beta_{\text{tag}_n}\}$. The g-tagging model trained on the events should thus be *permutation invariant*[18] with regards to the input variables. The category of permutation invariant (and equivariant[19]) neural networks has seen an huge development within recent years in the deep learning community. The paper from Zaheer et al. [101] in 2017 was highly influential, however also other examples exists [77, 51]. Yet, the same development 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: $N \to n! \cdot N$ 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.

[17] Since sets have no inherent order.

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

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

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 PDF[20] 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.

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

5.6.3 *g-Tagging Hyperparameter Optimization*

Four LightGBM models[21], 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 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. The results of the HPO can be seen in Figure 5.18. 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.8–B.11.

[21] The method with all permutations was trained using the same hyperparameters as the best ones founds in the HPO for the shuffled model to reduce time spent on HPO. The time performance is extra important for the method with all permutations as the dataset is 24 times larger the method with the shuffling for 4-jet events.

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. Here it can be seen that the model trained on the

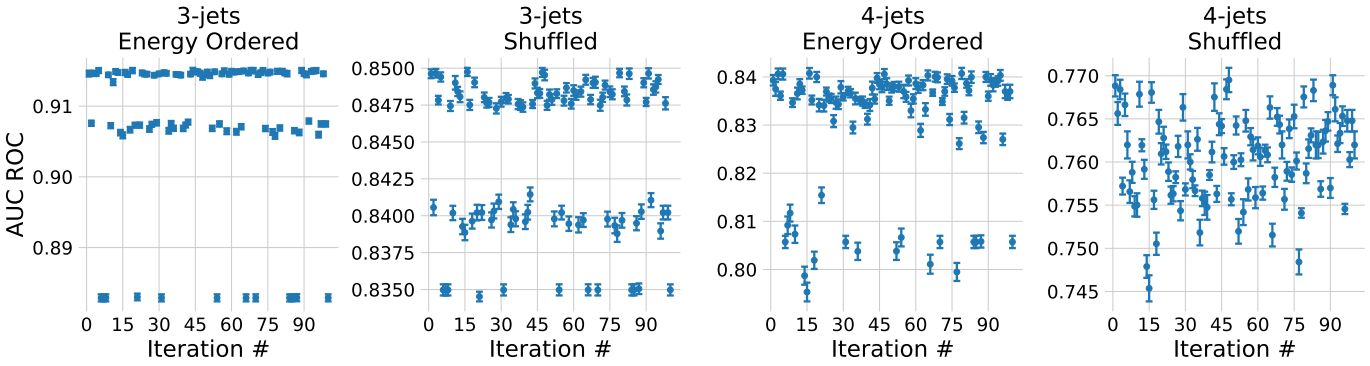


Figure 5.18: Hyperparameter Optimization results of g-tagging with 100 iterations of random search with LGB. From left to right, we have A) 3-jet events energy-ordered (no permutations), B) 3-jet events row-shuffled, C) 4-jet events energy-ordered, D) 4-jet events row-shuffled. Notice the different ranges on the y-axes.

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 tables, 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).

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

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

5.6.4 *PermNet*

In addition to the LGB models, a permutation invariant neural network called PermNet based on the Deep Sets paper [101] implemented in Tensorflow [10] by Faye [46] was also tested. Zaheer et al. [101] showed that $f(X)$ is permutation invariant if and only if it can be decomposed in the following way:

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

for suitable transformations ρ and ϕ (which the neural network learns[22]). The PermNet was trained using three layers[23] with leaky ReLU [68] as the activation function and ADAM [61] 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.12. 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.

[22] This is possible since neural networks are universal function approximators [55].

[23] Where the two hidden layers have 128 and 64 neurons in each.

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 5.19 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 background[24] together with the distribution of cuts made by the LGB model can be seen in Figure B.13. The similar plots for 3-jet events are plotted in Figure B.14 and B.15.

[24] Which the signal fraction is based on.

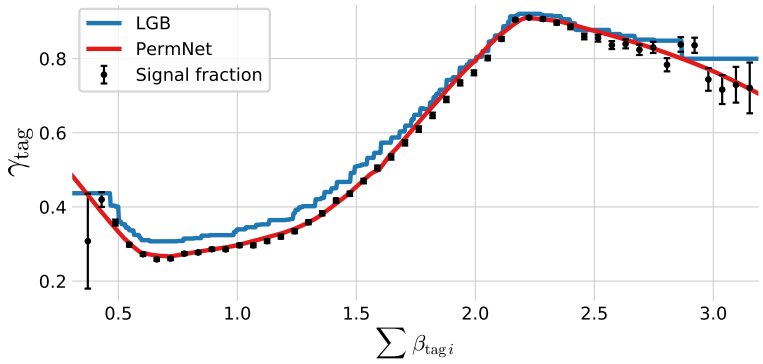


Figure 5.19: Plot of the (1D) g-tag scores for 4-jet events as a function of $\sum \beta_i$ for the LGB model in blue and the PermNet model in red. The signal fraction (based on the signal and background histograms in Figure B.13) is plotted as black error bars where the size of the error bars is based on the propagated uncertainties of the signal and background histogram assuming Poissonian statistics.

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

5.6.6 g-Tagging Results

The distribution of g-tag scores in 4-jet (training) events the can be seen in Figure 5.20 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 invariant[25]. 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.16.

[25] 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.

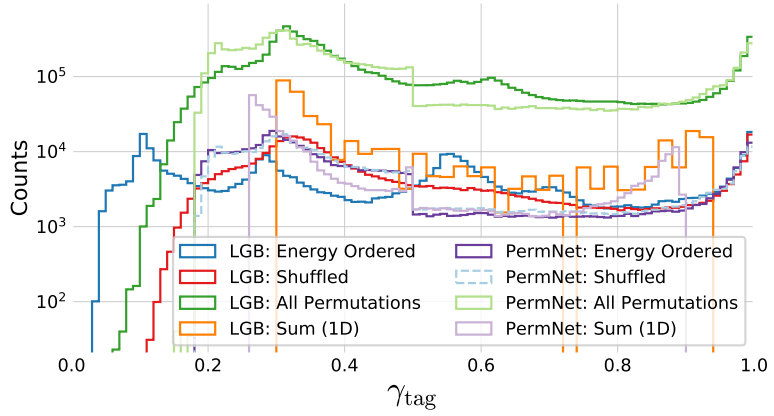


Figure 5.20: Distribution of g-tag scores in 4-jet events shown with a logarithmic y-scale for LGB: Energy Ordered in blue, LGB: Shuffled in red, LGB: All Permutations in green, LGB: Sum 1D in orange, PermNet: Energy Ordered in purple, PermNet: Shuffled in light-blue, PermNet: All Permutations in light-green, 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 5.21 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 B.18, 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 5.21 and B.18, 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 5.22. 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 accord-

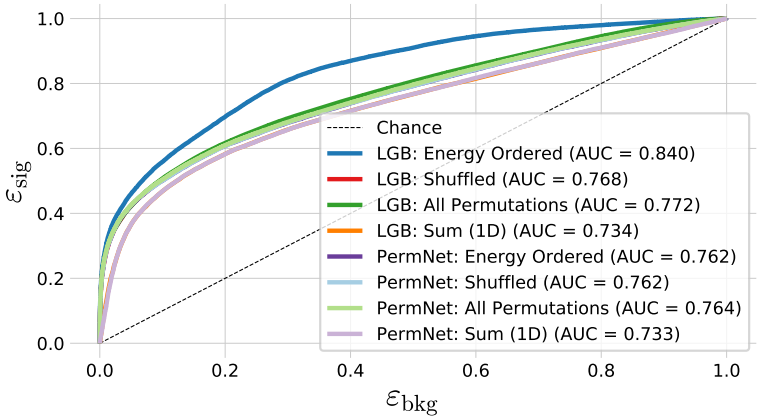


Figure 5.21: ROC curve of the eight g-tag models in 4-jet. First one in dashed black is the ROC curve that you get by random chance. The colors are the same as in Figure 5.20 and in the legend also the Area Under the ROC curve (AUC) is shown.

ing to their b-tags. Next after signal b events is $c\bar{c}gg$ (signal c) and $bg\bar{b}g$-events[26] (background b). At low values of γ_{tag} light quarks (uds) dominate.

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

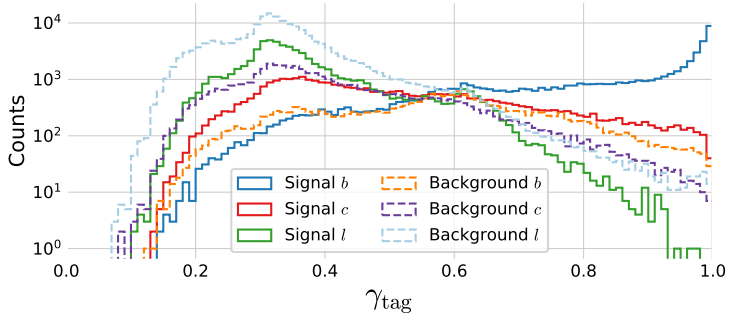


Figure 5.22: Histogram of g-tag scores from the LGB-model in 4-jet events for b signal in blue, c signal in red, l (uds) signal in green, b background in orange, c background in purple, l (uds) background in light-blue.

The similar plot for 3-jet events is seen in Figure B.19. This plot has some surprising bumps for mainly l-quark events. When comparing l-quark events in the high-γ_{tag} bump with the ones getting a low γ_{tag}-value, see Figure 5.23, 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.

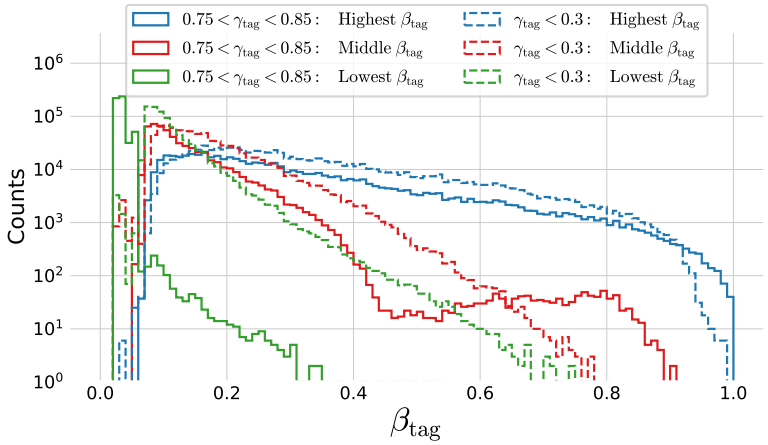


Figure 5.23: Distribution of b-Tag Scores in 3-Jet l-Quark Events for low and high g-tags values. Here l-quark events with $0.75 < \gamma_{\text{tag}} < 0.85$, so the high peak in Figure B.19, are plotted in fully connected lines, and events with $\gamma_{\text{tag}} < 0.3$ are plotted in dashed lines. 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 5.24. 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.

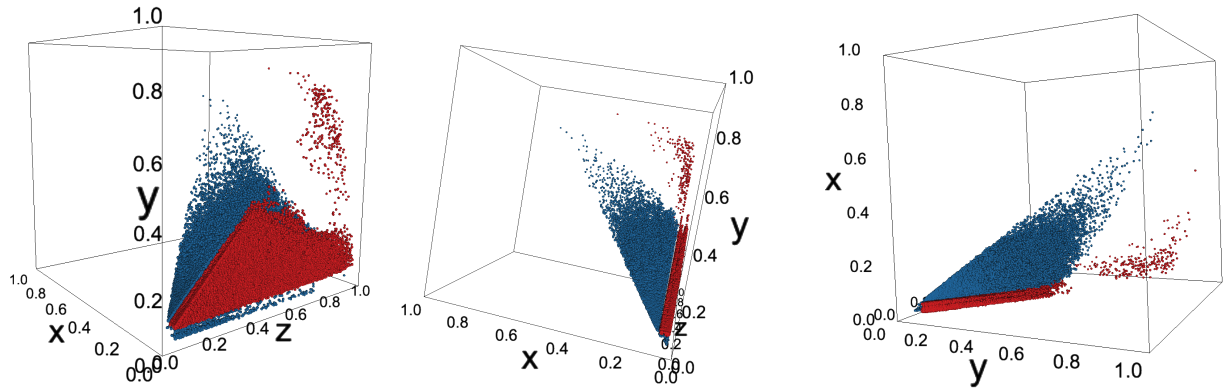


Figure 5.24: 3D scatter plot of β_{tag}-values for high and low γ_{tag} l-quark events. Here the x-axis is the lowest b-tag, the y-axis the middle, and the z-axis the highest. Here the high-γ_{tag} l-quark events are plotted in red and the low ones in blue.

5.7 *g-Tagging Efficiency*

blablabla intro here XXX.

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 $0.9 < \beta_{\text{tag}}$ 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:

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

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 5.25. 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.

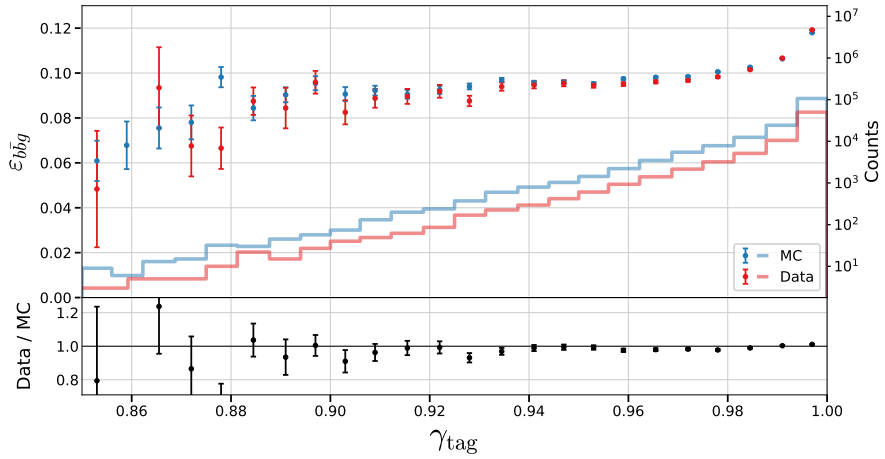


Figure 5.25: Proxy efficiency of the g-tags for $b\bar{b}g$ 3-jet events as a function of the event's g-tag γ_{tag}. In the top plot the proxy efficiency $\varepsilon_{b\bar{b}g}$ is shown for MC in blue and Data in red where the counts in each bin can be read on right y-axis. In the bottom plot the ratio between Data and MC is shown. The proxy 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 $\varepsilon_{b\bar{b}g} = \varepsilon_b^{b\text{-sig}} \cdot \varepsilon_{\bar{b}}^{b\text{-sig}} \cdot \varepsilon_g^{g\text{-sig}}$.

Another event variable to look at is the mean of the two invariant masses[27] m_{bg} and $m_{\bar{b}g}$. The invariant mass between two quantities[28] is:

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

where E_i is the energy of the i^{th} quantity and $\mathbf{p}_i$ its momentum. The pseudo efficiency is plotted as a function of the mean of m_{bg} and $m_{\bar{b}g}$ in Figure 5.26. Here the overall correspondence between Data and MC is lower than in Figure 5.25, especially for high values of the mean invariant mass. XXX is this a problem?.

[27] 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.

[28] When measured in natural units.

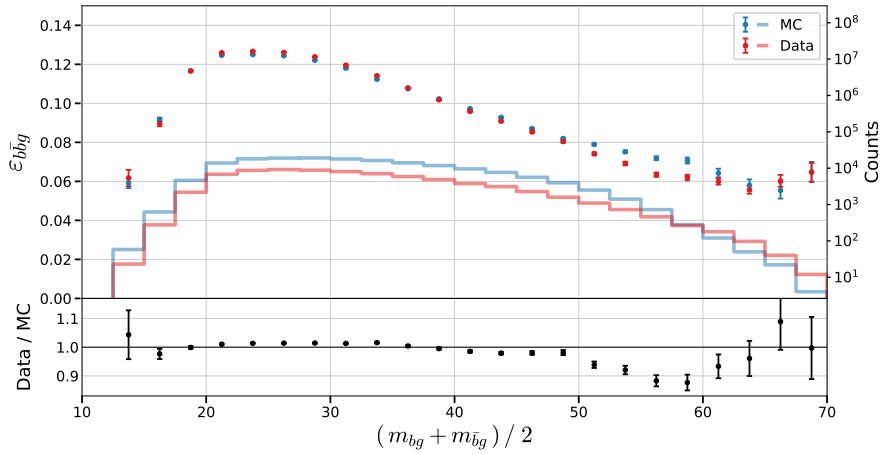


Figure 5.26: Proxy efficiency of the g-tags for $b\bar{b}g$ 3-jet events as a function of the mean of the two invariant masses m_{bg} and $m_{\bar{b}g}$ in the event. In the top plot the proxy efficiency $\varepsilon_{b\bar{b}g}$ is shown for MC in blue and Data in red where the counts in each bin can 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 "looks", 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 $\kappa \leq 0$:

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

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 [50, 63]. 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 5.27, where each of the sets of variables are related to the following aspects:

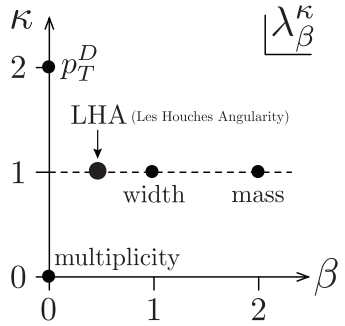


Figure 5.27: Generalized angularities. Adapted from Larkoski et al. [63].

(β, κ)

$(0, 0)$: Hadron Multiplicity.

$(0, 2)$: Transverse Momentum Distribution p_T^D: $\lambda_0^2 = \sum z_i^2 \equiv \left(p_T^D\right)^2$ [38].

$(\frac{1}{2}, 1)$: Les Houches Angularity (LHA) [88].

$(1, 1)$: 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 event with the lowest of the b-tags since this jet is expected to be the gluon jet. From the plot in Figure B.19, 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 $\varepsilon_g^{\text{3-jet}} = 19.68\,\%$ and a signal purity of $\rho_g^{\text{3-jet}} = 98.77\,\%$. Here a gluon event is defined as an event with a $0.9 < \gamma_{\text{tag}}$.

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 5.28. 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 and neutral jets, 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 $w_{\text{MC-Truth}} = w_{\text{MC}} \cdot \varepsilon_g^{\text{3-jet}}$. The $\lambda_\beta^\kappa = 0$ values has been removed before plotting for all sets of values of (β, κ).

In Figure 5.28 it is seen how MC Selected and Data distributions matches each other quite well. That the MC Truth and Data not matches equally well indicates that XXX **TODO!**. The rest of the plots can be seen in Figure B.22–B.31.

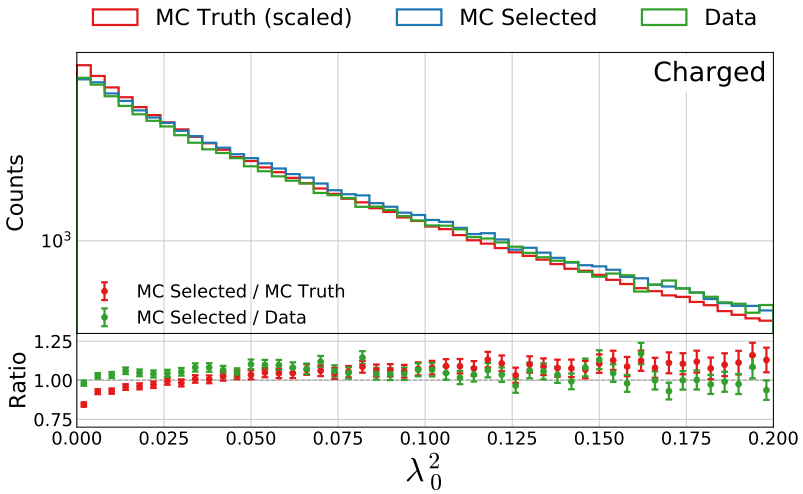


Figure 5.28: Distribution of the generalized angularity λ_0^2 for charged gluons jets in 3-jet events: λ_0^2. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

5.9 *Gluon splitting*

In addition to measuring how the gluon jets "look", 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 5.22, the γ_{tag} cut off threshold is set to $\gamma_{\text{cutoff}} = 0.8$ for 4-jet events. This corresponds to selecting 41 117 events in MC as gluon events with a signal efficiency of $\varepsilon_g^{\text{4-jet}} = 23.30\,\%$ and a signal purity of $\rho_g^{\text{4-jet}} = 92.67\,\%$. The variables related to measuring the gluon splitting will be introduced in subsection 5.9.1 and their efficiencies in MC will be computed in subsection 5.9.2. The method will be tested with a closure test in subsection 5.9.3 and the results shown in subsection 5.9.4.

5.9.1 *Variables*

The variables we use to measure gluon splitting are related to the energy asymmetry, the resolution scale, and the angle between the gluon-jets and the b-jets. In addition to these, some extra variables were proposed by Peter Skands in private communication, e.g. the $p_\perp$-antenna variable. These "Peter Skands"-variables are aimed at XXX **TODO!**. The gluon splitting variables are variables where current MC generators, such as Pythia [84], and Data show significant differences. Better measurements of these variables might lead to new theoretical insights that can reduce this discrepancy [85]. 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:

Δ_θ: The angle between the two gluon jets.

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

Angle

$\phi_\parallel$: 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:

$\ln\left(k_t^2/m_{\text{vis}}^2\right)$: Logarithm of the ratio between the k_t value of the two gluon jets and their visible mass.

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

$$p_{\perp,\text{A}}^2 = \widetilde{m}_{12}^2 \cdot \min\left(\frac{\min\left(\widetilde{m}_{b1}^2, \widetilde{m}_{b2}^2\right) - m_b^2}{\widetilde{m}_{b12}^2 - m_b^2}, \frac{\min\left(\widetilde{m}_{\bar{b}1}^2, \widetilde{m}_{\bar{b}2}^2\right) - m_b^2}{\widetilde{m}_{\bar{b}12}^2 - m_b^2}\right), \tag{5.10}$$

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

The gluon splitting variables will be analyzed in distinct areas of the phase space. This phase space is defined by the k_t [34, 44] and Cambridge/Aachen (CA) [42, 100] jet clustering algorithms for e^+e^- collisions which will first be described. The two algorithms uses the following[29] *jet distance measure*:

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

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

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. 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:

$$R_{gg}(p) \equiv \frac{d_{gg}^2(p)}{\min_{(i,j)\neq(g,g)} d_{ij}^2(p)}. \tag{5.12}$$

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 5.29. 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 5.30 or the opposite case where the two algorithms both agree on not to cluster the gluon jets together, see Figure 5.31. These figures are just illustrations of how to better understand the R_{gg} phase space.

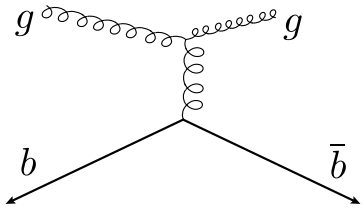


Figure 5.29: Soft, wide angle gluons in 4-jet events.

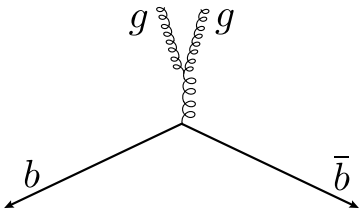


Figure 5.30: Soft, collinear gluons in 4-jet events.

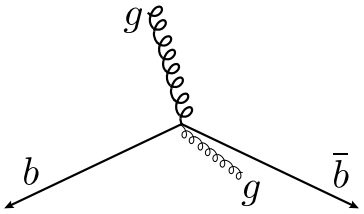


Figure 5.31: Hard, non $g \to gg$ gluons in 4-jet events.

The variables are computed for each event, and their relationship with γ_{tag} is shown in Figure B.32–B.39. From now on primarily events with with "good" g-tags, $\gamma_{\text{tag}} > 0.8$, 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 $\varepsilon_{gg}^{\text{4-jet}}$ for the gluon splitting variable E_{diff} is shown in Figure 5.32 for signal and background based on MC Truth.

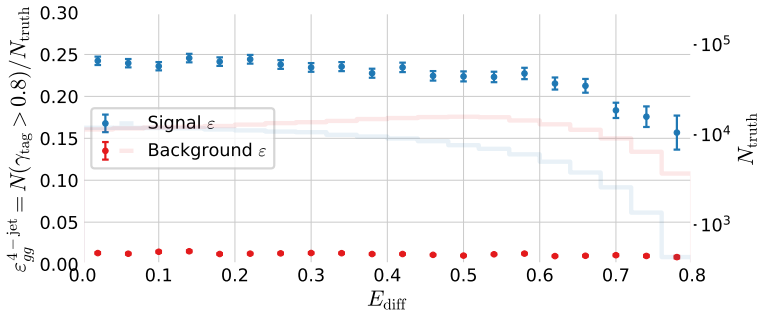


Figure 5.32: Efficiency of the g-tagging algorithm for 4-jet events as a function of normalized gluon-gluon jet energy difference (asymmetry) E_{diff} in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

The plot for the rest of the g-tagging efficiencies can be seen in Figure B.40–B.49. It can be concluded that XXX.

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 corrections[30] 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

[30] E.g. Applying the efficiencies found in the previous subsection.

so-called *sideband region*. From Figure 5.22, the signal region was defined to be $0.8 < \gamma_{\text{tag}}$ for 4-jet events. This region is expected to containing primarily signal events, according to the figure. Remember, signal events are here defined to be events where the two quark jets are the jets with the highest b-tag values. The sideband region[31] 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). For 4-jet events, we the sideband region is defined to be $0.6 < \gamma_{\text{tag}} < 0.8$.

[31] Also sometimes known as a control region.

The aim is to fully reconstruct the true distribution of the gluon splitting variable x, called $\mathcal{P}_{gg}(x)$. To first order, this is given by $\mathcal{P}_{gg} \approx \mathcal{P}_{\text{sign}}/\varepsilon_{gg}^{\text{4-jet}}$, where $\mathcal{P}_{\text{sign}}(x)$ 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 in background events:

$$\mathcal{P}_{gg} = \frac{\mathcal{P}_{\text{sign}} - \alpha \cdot \mathcal{P}_{\text{bkg}}}{\varepsilon_{gg}^{\text{4-jet}}}. \tag{5.13}$$

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 $N_{\text{bkg}}^{\text{side}}$. Defining f_{gg}^{sig} to be the fraction of signal in the signal region, $f_{\text{bkg}}^{\text{side}}$ the fraction of background in the sideband region, α is defined as:

$$\alpha = \frac{N_{\text{bkg}}^{\text{sig}}}{N_{\text{bkg}}^{\text{side}}} = \frac{\left(1 - f_{gg}^{\text{sig}}\right) \cdot N_{\text{sig}}}{f_{\text{bkg}}^{\text{side}} \cdot N_{\text{side}}}, \tag{5.14}$$

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 it is thus corrected for:

$$\mathcal{P}_{\text{bkg}} = \mathcal{P}_{\text{side}} - \beta \cdot \mathcal{P}_{gg}\varepsilon_{gg}^{\text{4-jet}}, \tag{5.15}$$

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:

$$\beta = \frac{\left(1 - f_{\text{bkg}}^{\text{side}}\right) \cdot N_{\text{side}}}{f_{gg}^{\text{sig}} \cdot N_{\text{sig}}}. \tag{5.16}$$

Plugging equation (5.15) into (5.13) and solving for $\mathcal{P}_{gg}$ yields:

$$\mathcal{P}_{gg} = \frac{\mathcal{P}_{\text{sign}} - \alpha \cdot \mathcal{P}_{\text{side}}}{\varepsilon_{gg}^{\text{4-jet}} \cdot (1 + \alpha\beta)}. \tag{5.17}$$

The advantage of this equation is that it is that only $\varepsilon_{gg}^{\text{4-jet}}$ and the two constants[32] f_{gg}^{sig} and $f_{\text{bkg}}^{\text{side}}$ depend on MC truth, the rest can be applied to data without any truth label. The assumptions equation (5.17) are based on are that the signal distribution of x is similar in

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

the signal and sideband regions $\mathcal{P}_{gg}^{\text{sig}} = \mathcal{P}_{gg}^{\text{side}}$ and likewise for the background distributions $\mathcal{P}_{\text{bkg}}^{\text{sig}} = \mathcal{P}_{\text{bkg}}^{\text{side}}$.

The distribution of $\mathcal{P}_{gg}$ in MC is shown in Figure 5.33 together with $\mathcal{P}_{\text{sig}}$, $\mathcal{P}_{\text{side}}$, 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 $\mathcal{P}_{gg}^{\text{Truth}}$ follows 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 $\mathcal{P}_{gg}^{\text{Truth}}$. The errorbars in the ratio plot is 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, `P` is the χ^2-probability[33]. I compute the systematic error σ_{sys} that would have to be added in quadrature to the standard deviation, $\sigma \to \sqrt{\sigma^2 + \sigma_{\text{sys}}^2}$, 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.

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

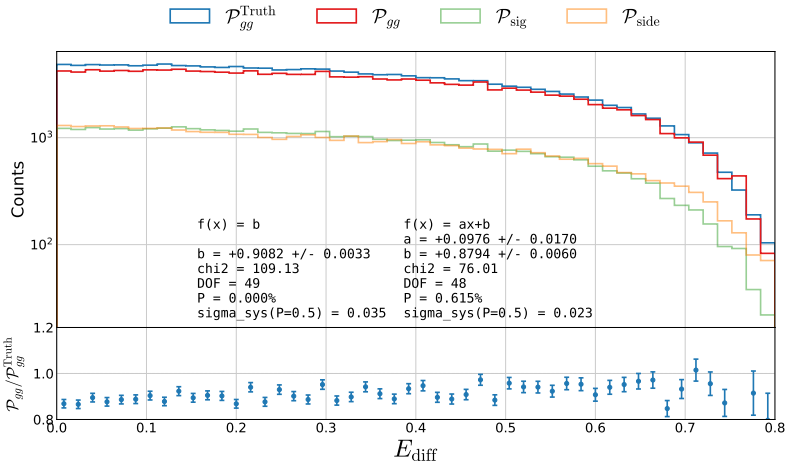


Figure 5.33: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for the normalized gluon gluon jet energy asymmetry E_{diff}. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the signal region $\mathcal{P}_{\text{sig}}$ 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 closure plot for E_{diff} is shown in Figure 5.33 and for the rest of the gluon splitting variables in Figure B.50–B.57. In general it can be seen that the g-tagging algorithm perform well on the energy asymmetry variables with $\sigma_{\text{sys}}(E) \approx 3\,\%$, the angle-dependant variables, Δ_θ and $\phi_\parallel$ with $\sigma_{\text{sys}}(\Delta_\theta) \approx 5\,\%$ and $\sigma_{\text{sys}}(\phi_\parallel) \approx 2\,\%$. When the mass enters in the variables, there is a much higher bias, however, still small for the $p_\perp$-antenna variable $\sigma_{\text{sys}}(p_{\perp,\text{A}}^2) \approx 5\,\%$. Worst is the gluon-gluon invariant mass where $\sigma_{\text{sys}}(m_{gg}) \approx 23\,\%$, see Table 5.6 for all of the values.

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.

	σ_{sys}
E_{diff}	3.5 %
$E_{\text{rel}_{\text{min}}}$	3.5 %
E_{rel}	3.2 %
Δ_θ	5.6 %
m_{gg}	22.6 %
$\phi_\parallel$	1.7 %
$\ln\left(k_T^2/m_{\text{vis}}^2\right)$	12.9 %
$p_{\perp,\text{A}}^2$	5.0 %

Table 5.6: Systematic errors for the gluon splitting variables based on the closure test, see subsection 5.9.3.

5.9.4 4-jet results

The comparison between the gluon splitting distributions will be done in four distinct areas of the $R_{gg}^{k_t}$-R_{gg}^{CA} phase space. These four areas, A, B, C, and D, are defined in Table 5.7. Three of these regions have already been described, that is area A which is the soft, collinear region with an example shown in Figure 5.30, area C which is the soft, wide angle region in Figure 5.29, and the hard non-$g \to gg$ area D in Figure 5.31. Area 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 or not. The four areas are visualized in Figure 5.34 for both MC and Data. Here each dot shown in the scatter plot is an event with $0.8 < \gamma_{\text{tag}}$, with a total of 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 areas are shown in the figure. The events are split up into these four areas because they each concern different physical interactions, yet one also has to take the number of events in each area into account such that the statistical uncertainties does not prevent any conclusions to be drawn.

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 5.7: Definitions of the four areas in the $R_{gg}^{k_t}$-R_{gg}^{CA} phase space.

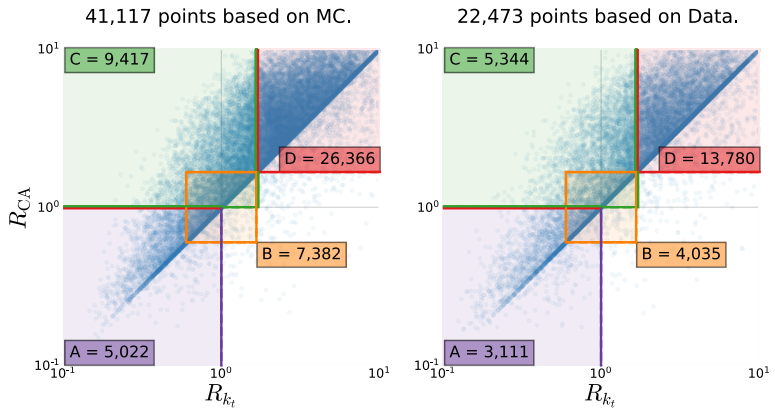


Figure 5.34: Overview of the four areas, A, B, C, and D, in the $R_{gg}^{k_t}$-R_{gg}^{CA} phase space. The areas are shown as colored rectangles together with a scatter plot showing the 2D-distribution for signal gluon events ($0.8 < \gamma_{\text{tag}}$). The left plot shows is for MC and the right one for Data.

The distributions for the different gluons splitting variables for area A, the soft, collinear region, is shown in Figure 5.35 for both MC (scaled to Data) and Data. Area A contains 5022 events in the MC sample and 3111 in the Data sample, both for signal ($0.8 < \gamma_{\text{tag}}$) events. Generally the distributions match pretty well between the MC and Data, however, there seems to small discrepancies in the energy assymetry variables, however, this might just be due to statistical fluctuations (notice the low bin count in each bin). Furthermore, there is a mismatch for the $\ln\left(k_T^2/m_{\text{vis}}^2\right)$ variable at around 16. When looking at the the other areas, see Figure B.58–B.61, the energy discrepancies seem to disappear, whereas the mismatch between MC and Data in $\ln\left(k_T^2/m_{\text{vis}}^2\right)$ continues to exist.

5.10 Un-folding

And then lastly they should be unfolded? Or it should at least be mentioned?

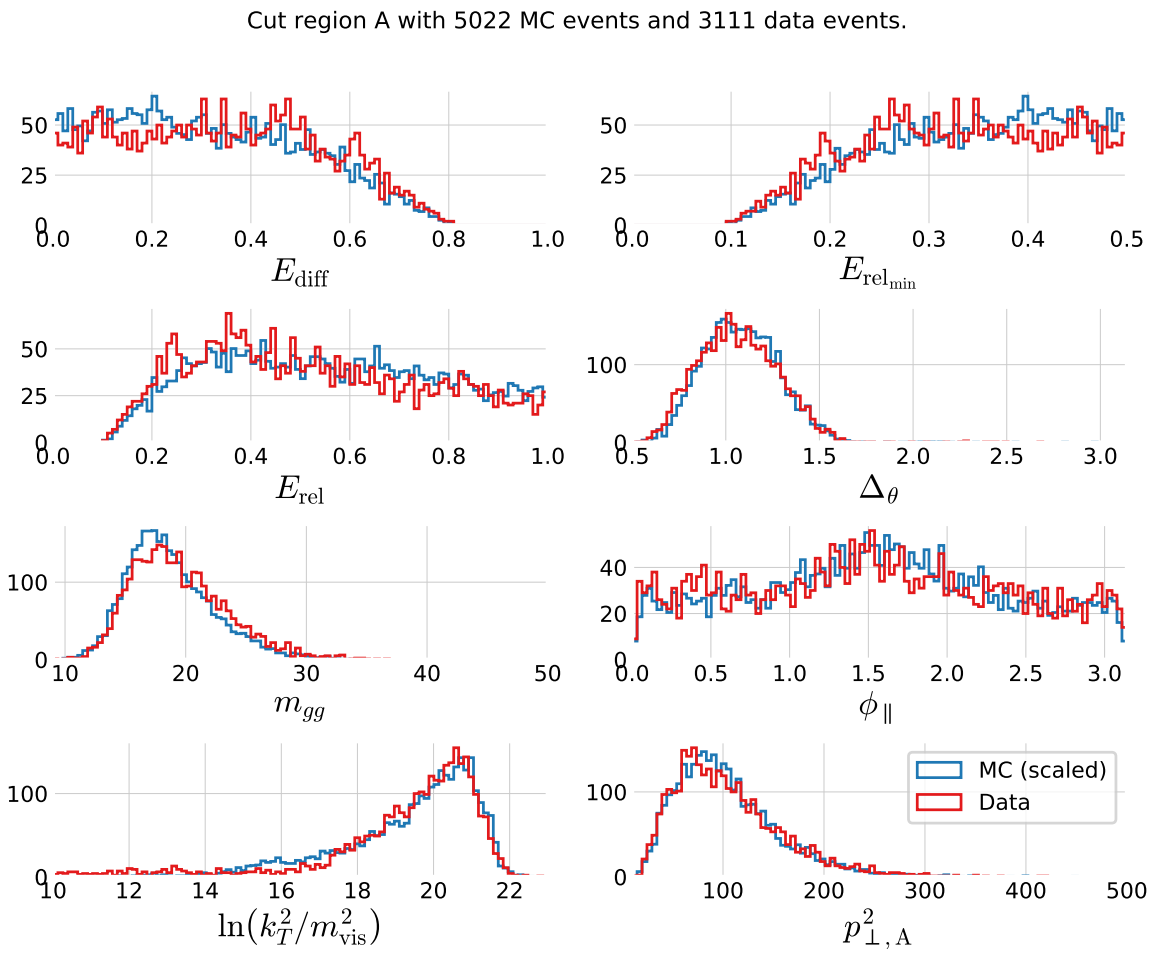


Figure 5.35: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area A, see Table 5.7. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area A which has 5022 events in the MC sample and 3111 in the Data sample.

B. Quarks vs. Gluons Appendix

	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.

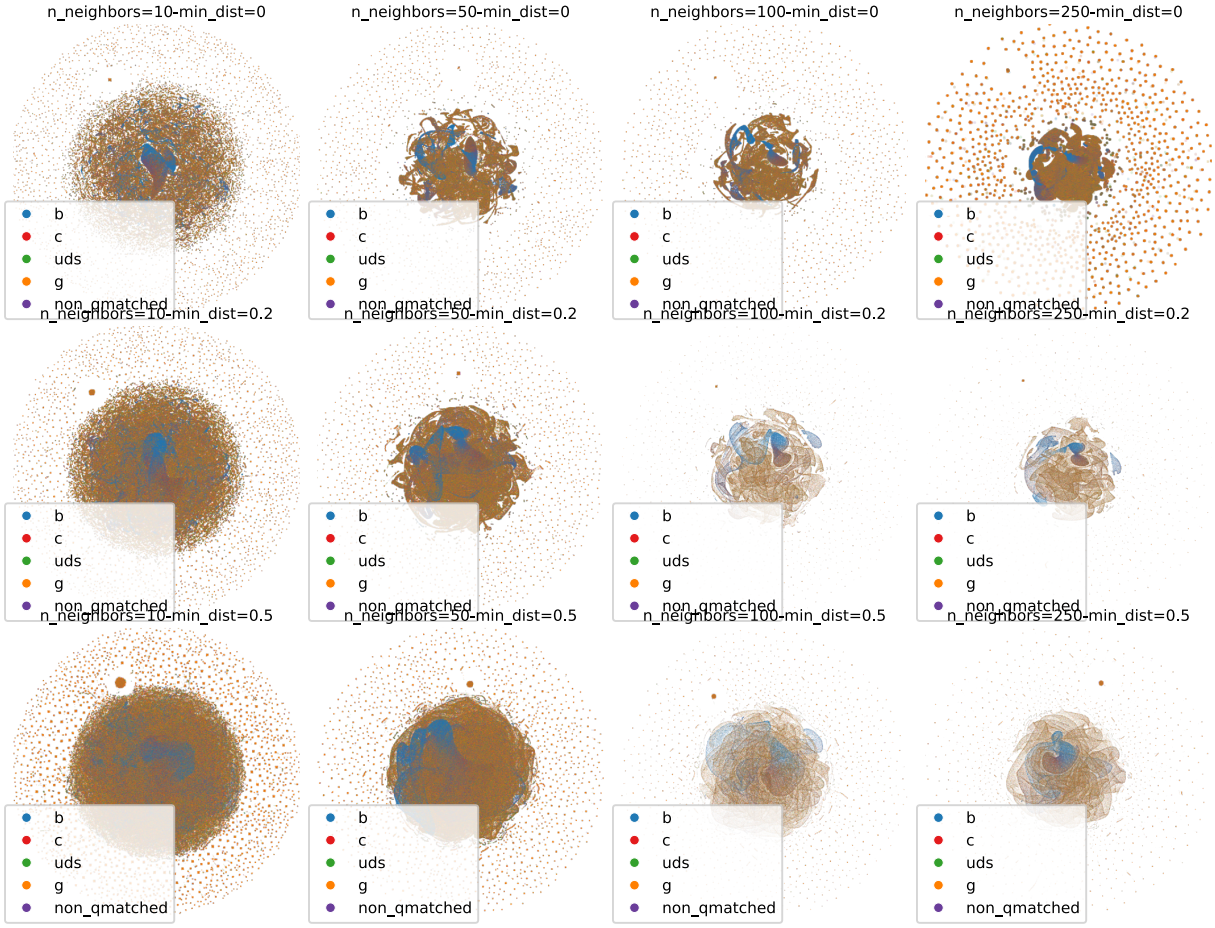


Figure B.1: Grid search of the two parameters `n_neighbors` and `min_dist` for the UMAP algorithm run on 4-jet events. For an explanation of these, see section 5.2.

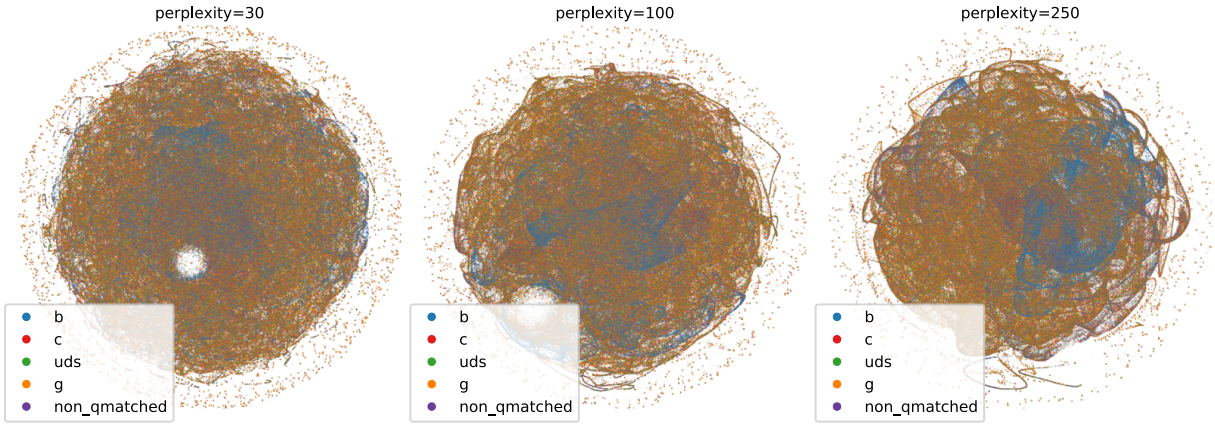


Figure B.2: Visualization of the t-SNE algorithm as a function of the `perplexity` parameters for 4-jet events.

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

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

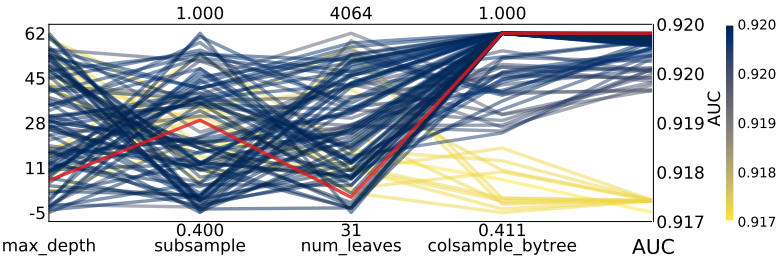


Figure B.3: Hyperparameter optimization results of b-tagging for 3-jet events. The results are shown as parallel coordinates with each hyperparameter along the x-axis and the value of that parameter on the y-axis. Each line is an event in the 4-dimensional space colored according to the performance of 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.

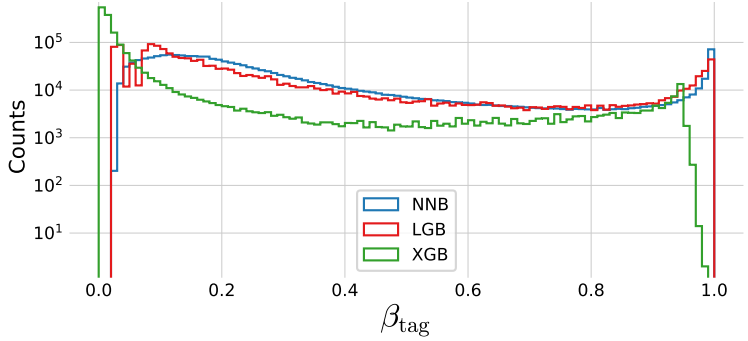


Figure B.4: Histogram of b-tag scores β_{tag} in 3-jet events for NNB (the neural network pre-trained by ALEPH, also called `nnbjet`) in blue, LGB in red, and XGB in green.

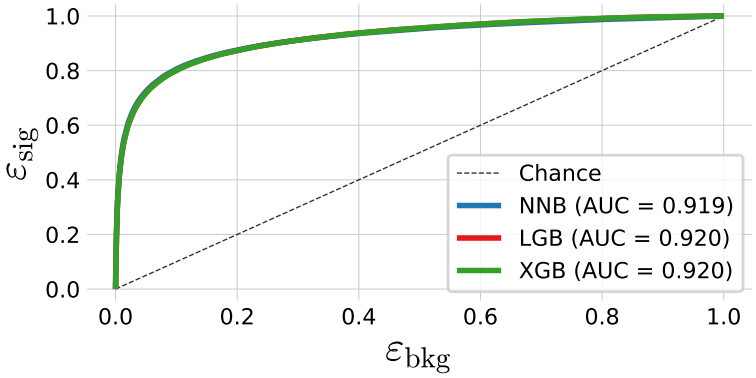


Figure B.5: ROC curve of the three b-tag models in 3-jet events for NNB (the pre-trained neural network trained by ALEPH, also called `nnbjet`) in blue, LGB in red, and XGB in green. In the legend the area under curve (AUC) is also shown. Notice that the LGB and XGB models share performance and it is thus due to overplotting that only the green line for XGB can be seen. In the machine learning community the background efficiency ε_{bkg} is sometimes know as the false positive rate (FPR) and the signal efficiency ε_{sig} as the true positive rate (TPR).

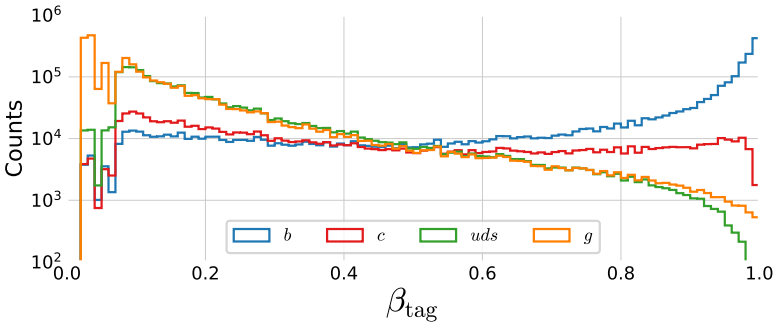


Figure B.6: Distribution of b-tags in 3-jet events for b-jets in blue, c-jets in red, uds in green and g in orange.

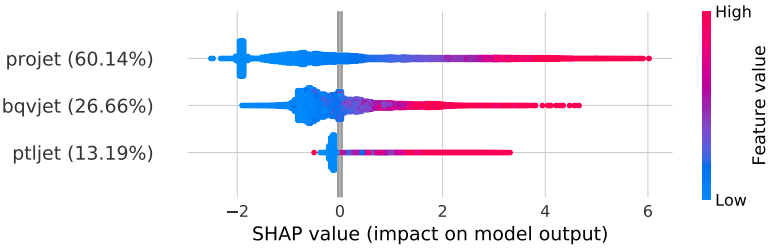


Figure B.7: Global feature importances for the LGB b-tagging algorithm on 3-jet events. The normalized feature importance is shown in the parenthesis and the each dot is an observation showing the dependance between the SHAP value and the feature's value.

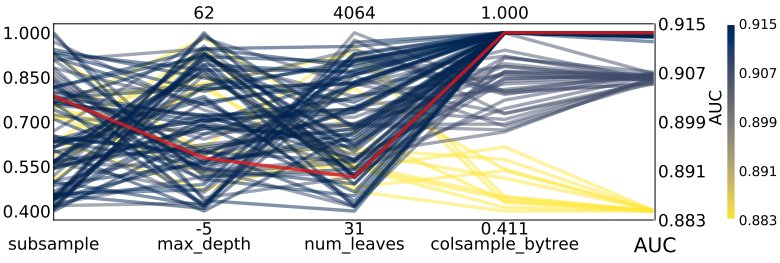


Figure B.8: Hyperparameter optimization results of g-tagging for 3-jet events for energy ordered jets.

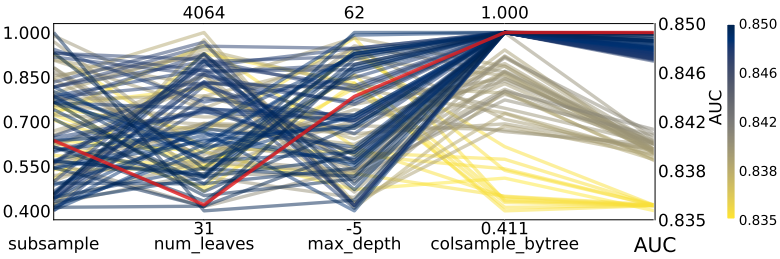


Figure B.9: Hyperparameter optimization results of g-tagging for 3-jet events for (row) shuffled jets.

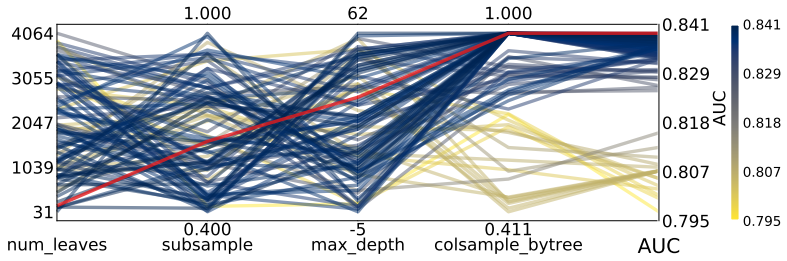


Figure B.10: Hyperparameter optimization results of g-tagging for 4-jet events for energy ordered jets.

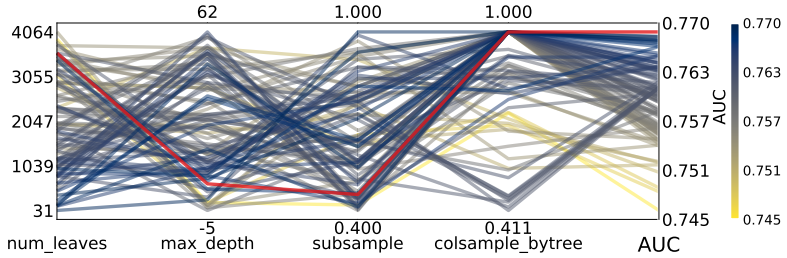


Figure B.11: Hyperparameter optimization results of g-tagging for 4-jet events for (row) shuffled jets.

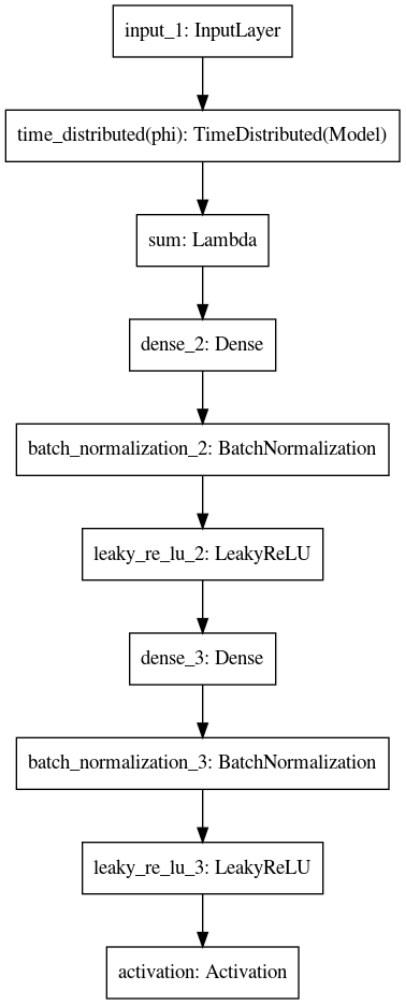


Figure B.12: Architecture of the PermNet neural network.

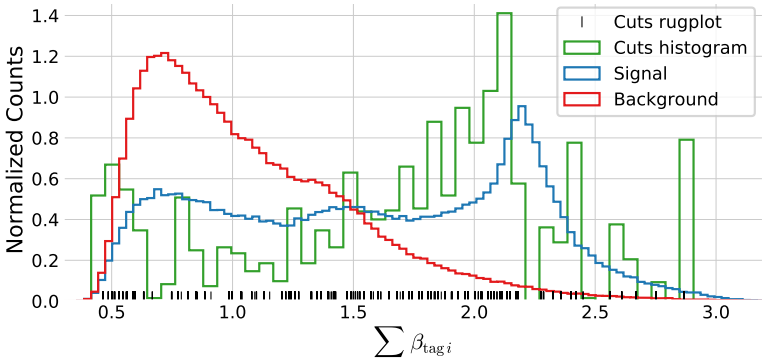


Figure B.13: Histogram of the distribution of signal in blue and background in red for the 1-dimensional sum of b-tags for 4-jet events. A histogram of the cut values from the LGB model trained on this data is shown in green together with a rug plot of the cut values in black. Notice how most of the cuts match up with the signal peak at around a $\sum \beta_i \sim 2.1$.

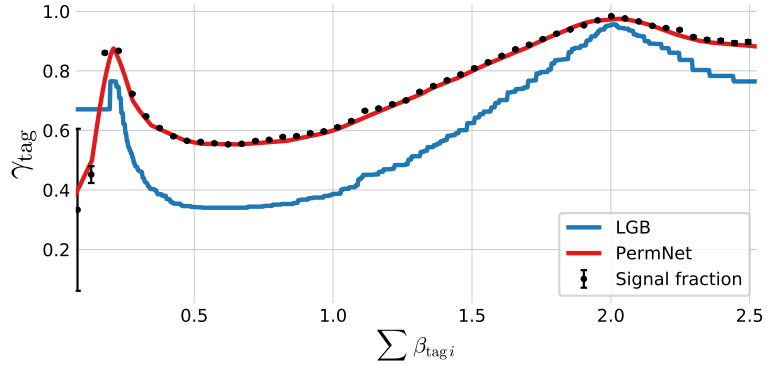


Figure B.14: Plot of the (1D) g-tag scores for 3-jet events as a function of $\sum \beta_i$ for the LGB model in blue and the PermNet model in red. The signal fraction (based on the signal and background histograms in Figure B.15) is plotted as black error bars where the size of the error bars is based on the propagated uncertainties of the signal and background histogram assuming Poissonian statistics.

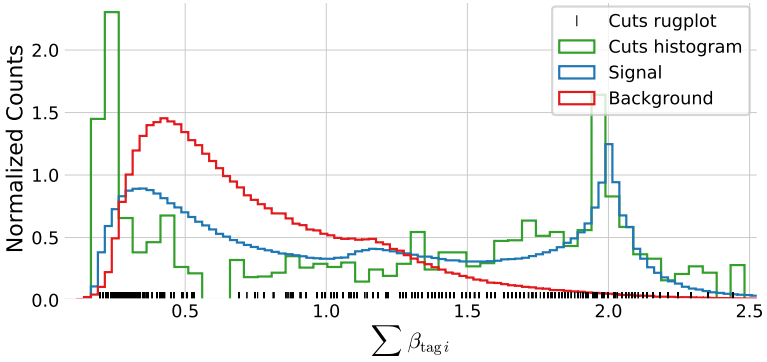


Figure B.15: Histogram of the distribution of signal in blue and background in red for the 1-dimensional sum of b-tags for 3-jet events. A histogram of the cut values from the LGB model trained on this data is shown in green together with a rug plot of the cut values in black. Notice how most of the cuts match up with the signal peak at around a $\sum \beta_i \sim 2.1$.

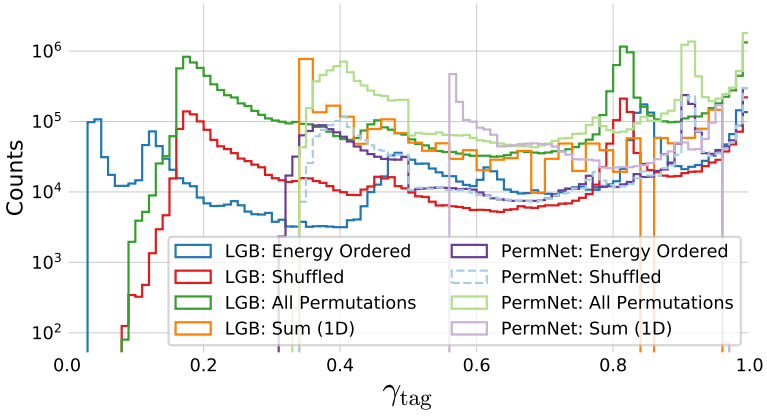


Figure B.16: Distribution of g-tag scores in 3-jet events shown with a logarithmic y-scale for LGB: Energy Ordered in blue, LGB: Shuffled in red, LGB: All Permutations in green, LGB: Sum 1D in orange, PermNet: Energy Ordered in purple, PermNet: Shuffled in light-blue, PermNet: All Permutations in light-green, 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).

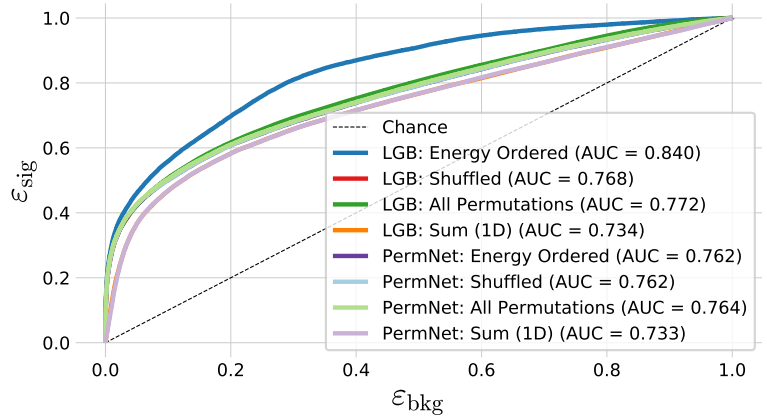


Figure B.17: ROC curve of the eight g-tag models in 4-jet events. First one in dashed black is the ROC curve that you get by random chance. The colors are the same as in Figure 5.20 and in the legend also the Area Under the ROC curve (AUC) is shown. Notice that the XGB model which uses the energy ordered data produced the best model, however, this model is not permutation invariant. Of the permutation invariant models (the rest), the XGB model trained on all permutations of the b-tags performs highest. The lowest performing models are the two models trained only on the 1-dimensional sum of b-tags, as expected, however, still with a better performance than expected by the author.

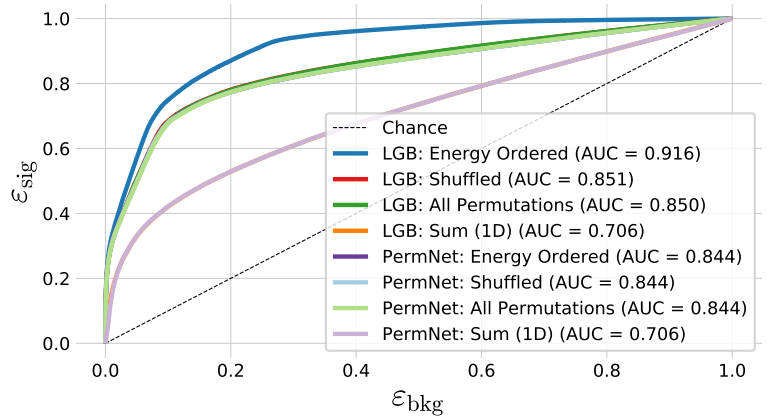


Figure B.18: ROC curve of the eight g-tag models in 3-jet. First one in dashed black is the ROC curve that you get by random chance. The colors are the same as in Figure 5.20 and in the legend also the Area Under the ROC curve (AUC) is shown.

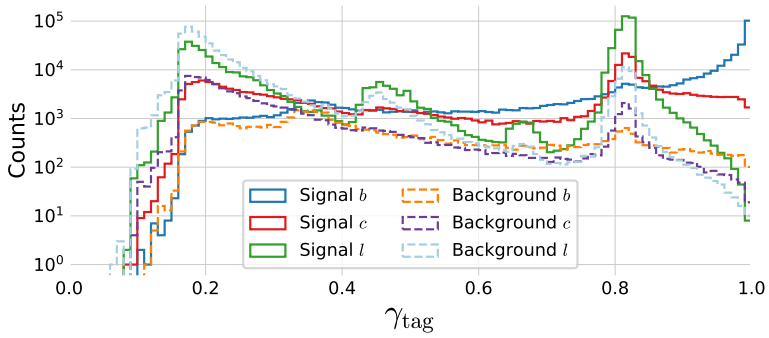


Figure B.19: Histogram of g-tag scores from the LGB-model in 3-jet events for b signal in blue, c signal in red, l (uds) signal in green, b background in orange, c background in purple, l (uds) background in light-blue.

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 $\phi^{\text{tot}}_{\beta_i}$ for the three g-Tagging Models in 3-Jet Events. Each $\phi^{\text{tot}}_{\beta_i}$ is shown for the three methods in the columns and the three b-tags as variables in the rows.

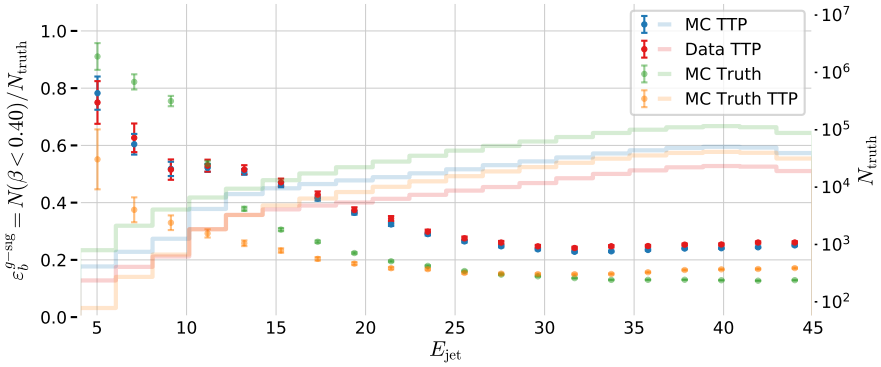


Figure B.20: b-tag efficiency for b-jets in the g-signal region for 3-jet events, $\varepsilon_b^{g\text{-sig}}$, as a function of jet energy E_{jet}. In the plot the efficiencies are shown for MC TTP in blue, Data TTP in red, MC Truth TTP in green, and MC Truth TTP in orange. The efficiencies (the errorbars) can be read off on the left y-axis and the counts (histograms) on the right y-axis.

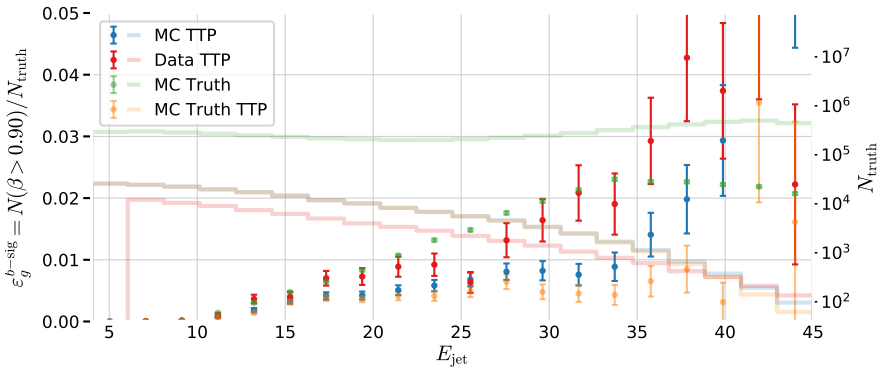


Figure B.21: b-tag efficiency for g-jets in the b-signal region for 3-jet events, $\varepsilon_g^{b\text{-sig}}$, as a function of jet energy E_{jet}. In the plot the efficiencies are shown for MC TTP in blue, Data TTP in red, MC Truth TTP in green, and MC Truth TTP in orange. The efficiencies (the errorbars) can be read off on the left y-axis and the counts (histograms) on the right y-axis.

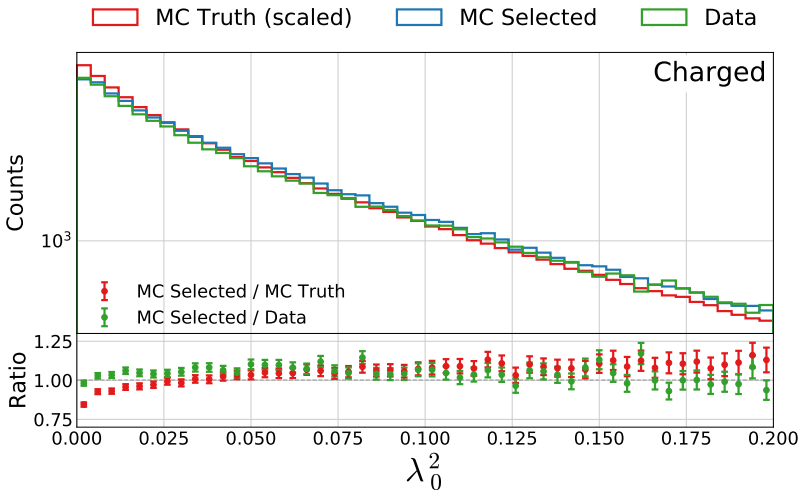


Figure B.22: Distribution of the generalized angularity λ_0^2 for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

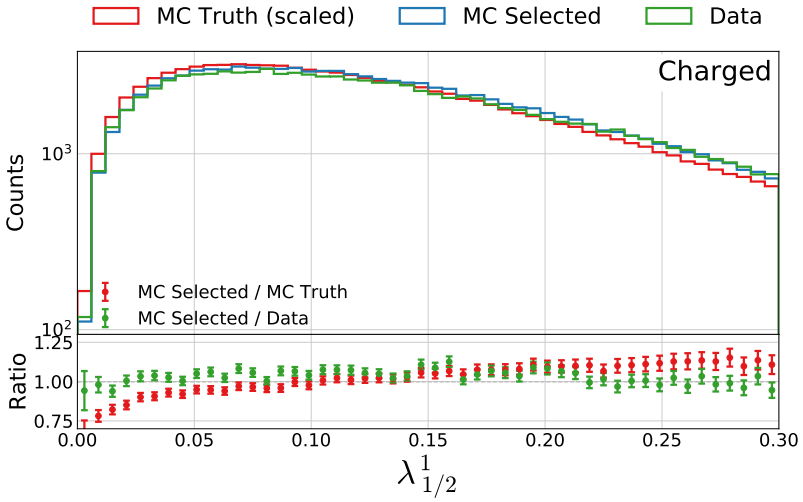


Figure B.23: Distribution of the generalized angularity $\lambda_{\frac{1}{2}}^1$ for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

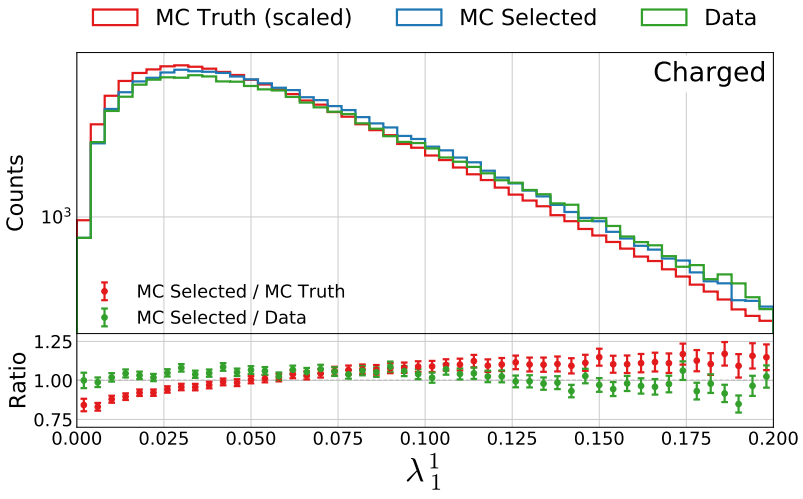


Figure B.24: Distribution of the generalized angularity λ_1^1 for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

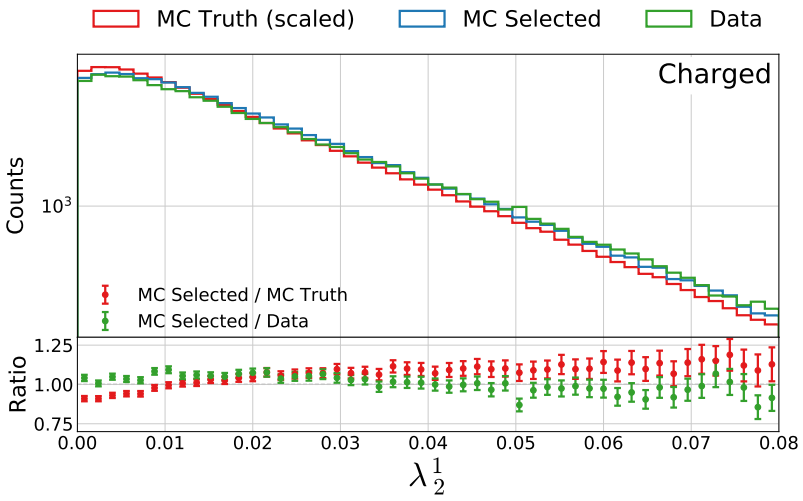


Figure B.25: Distribution of the generalized angularity λ_1^2 for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

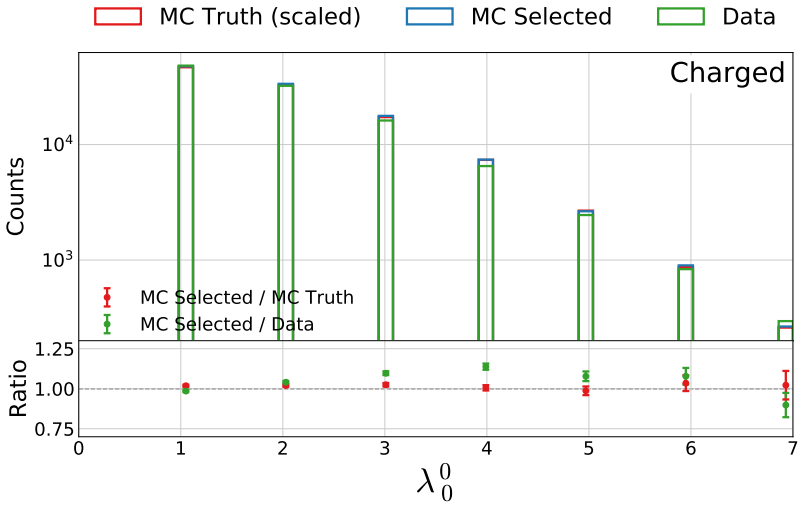


Figure B.26: Distribution of the generalized angularity λ_0^0 for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

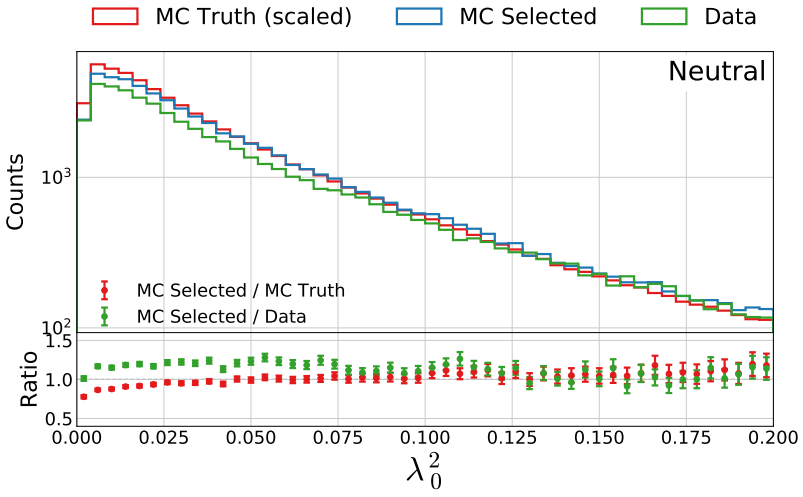


Figure B.27: Distribution of the generalized angularity λ_0^2 for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

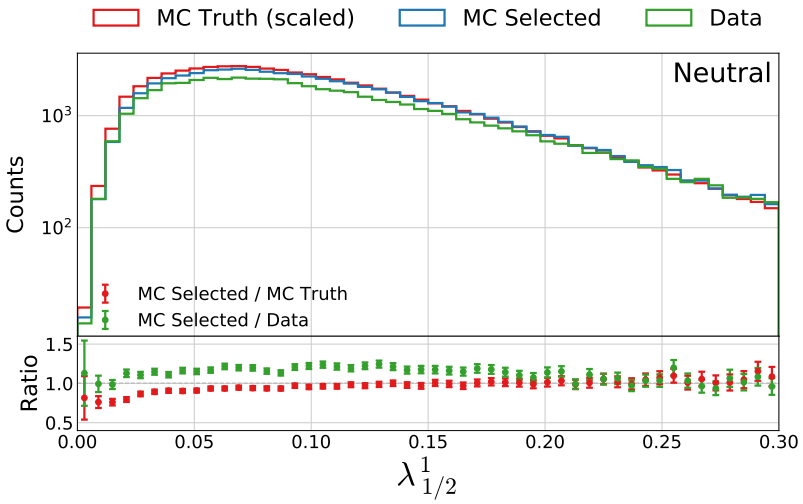


Figure B.28: Distribution of the generalized angularity $\lambda^1_{\frac{1}{2}}$ for neutral gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

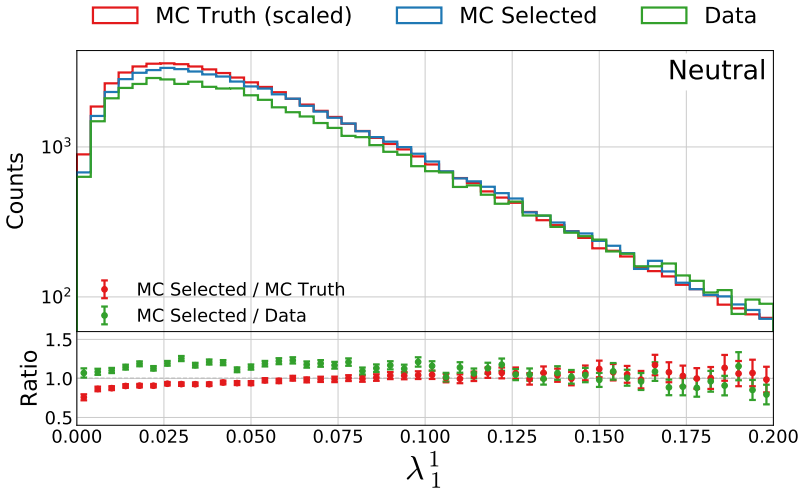


Figure B.29: Distribution of the generalized angularity λ^1_1 for neutral gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

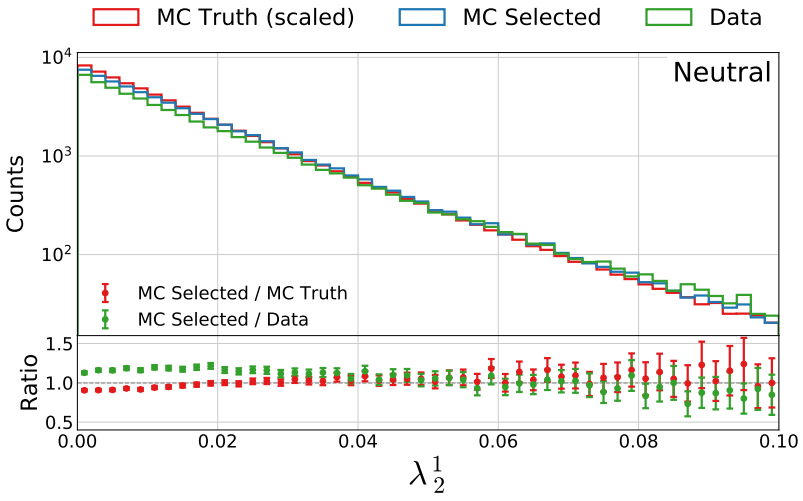


Figure B.30: Distribution of the generalized angularity λ^2_1 for neutral gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

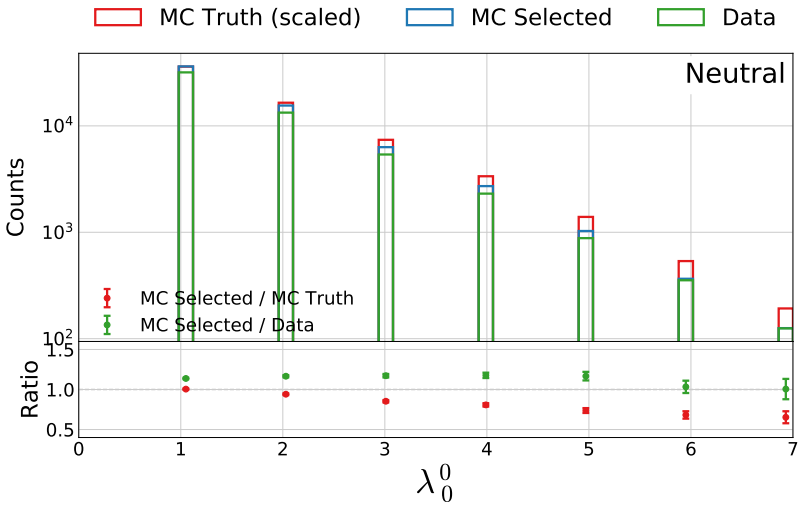


Figure B.31: Distribution of the generalized angularity λ_0^0 for neutral gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

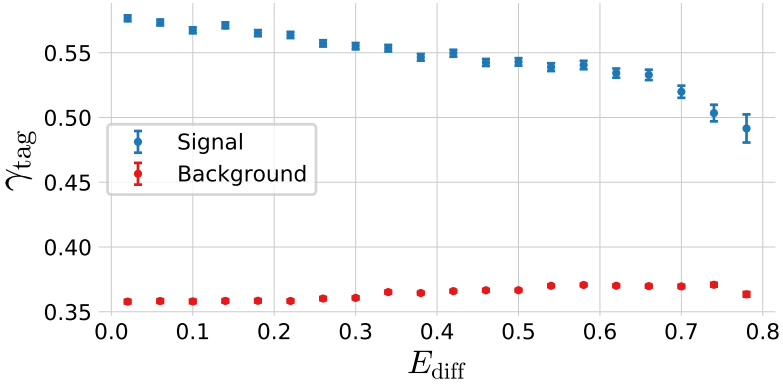


Figure B.32: Relationship between the g-tag value γ_{tag} and the gluon splitting variable E_{diff}. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

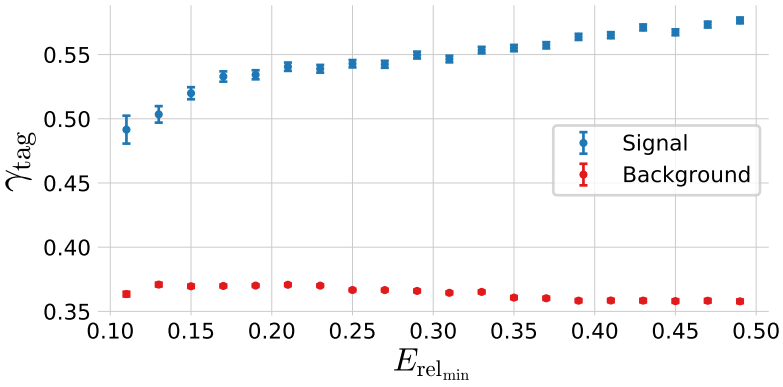


Figure B.33: Relationship between the g-tag value γ_{tag} and the gluon splitting variable $E_{\text{rel}_{\text{min}}}$. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

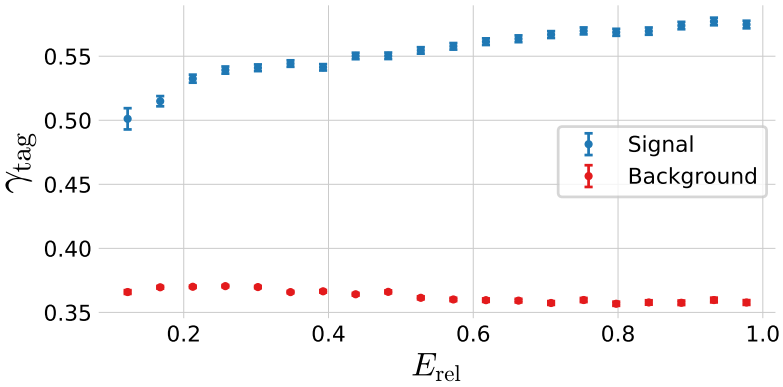


Figure B.34: Relationship between the g-tag value γ_{tag} and the gluon splitting variable E_{rel}. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

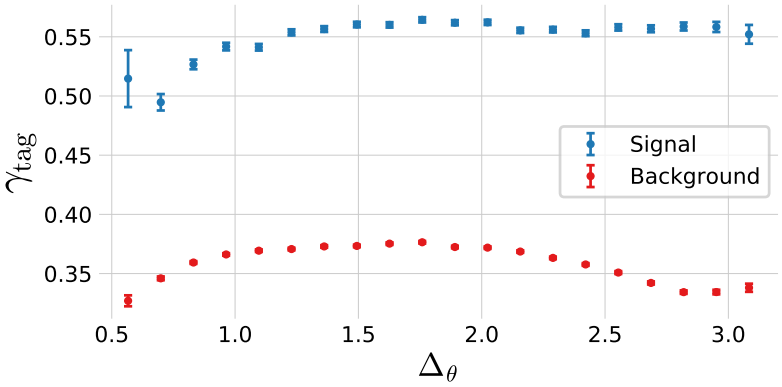


Figure B.35: Relationship between the g-tag value γ_{tag} and the gluon splitting variable Δ_θ. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

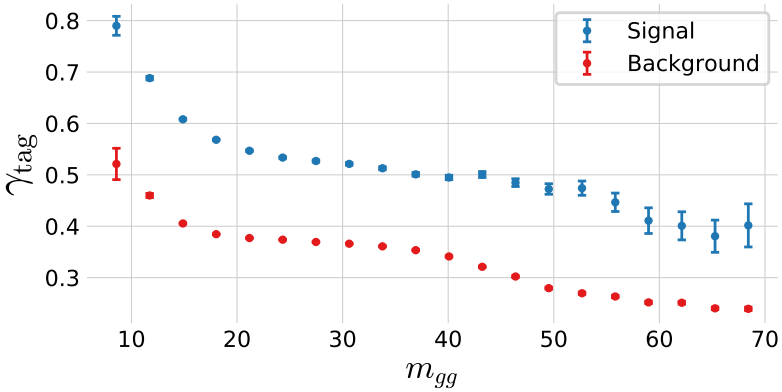


Figure B.36: Relationship between the g-tag value γ_{tag} and the gluon splitting variable m_{gg}. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

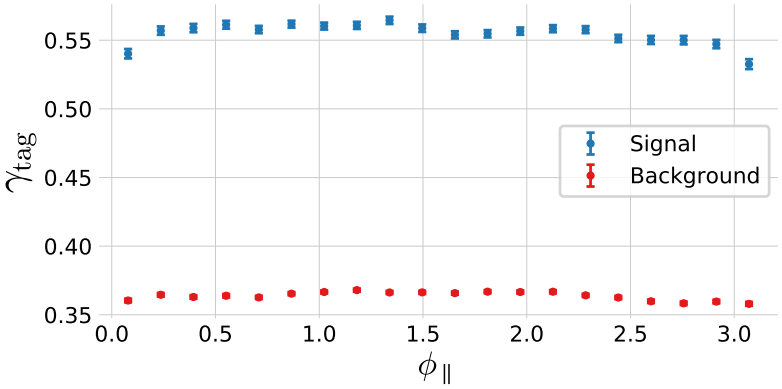


Figure B.37: Relationship between the g-tag value γ_{tag} and the gluon splitting variable $\phi_\parallel$. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

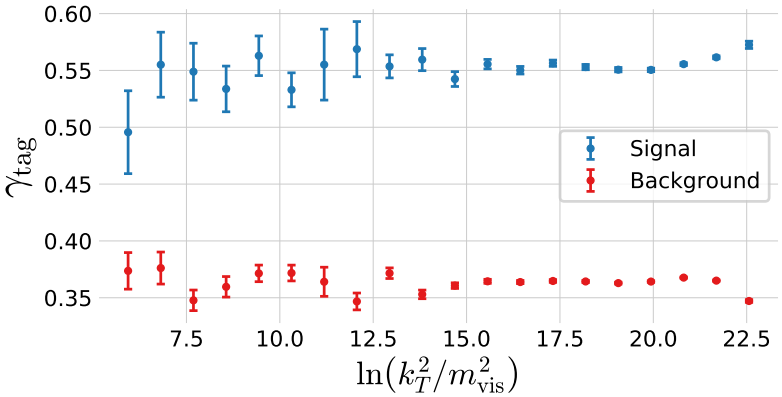


Figure B.38: Relationship between the g-tag value γ_{tag} and the gluon splitting variable $\ln\left(k_t^2/m_{\text{vis}}^2\right)$. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

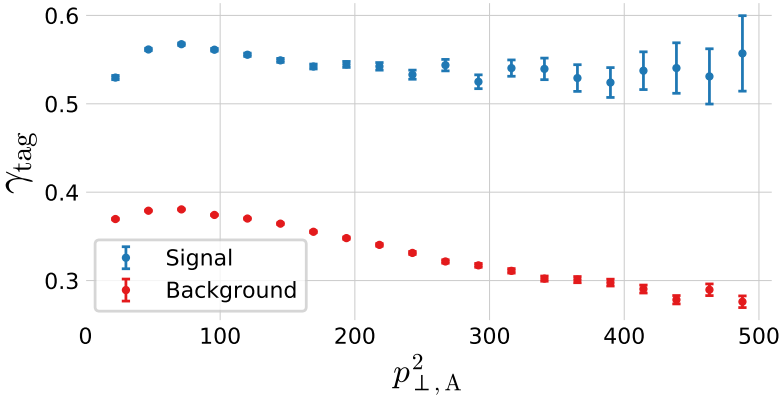


Figure B.39: Relationship between the g-tag value γ_{tag} and the gluon splitting variable $p^2_{\perp,\text{A}}$. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

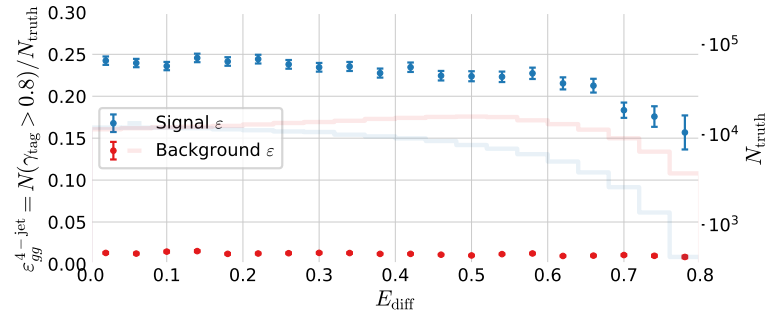


Figure B.40: Efficiency of the g-tagging algorithm for 4-jet events as a function of normalized gluon-gluon jet energy difference (asymmetry) E_{diff} in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

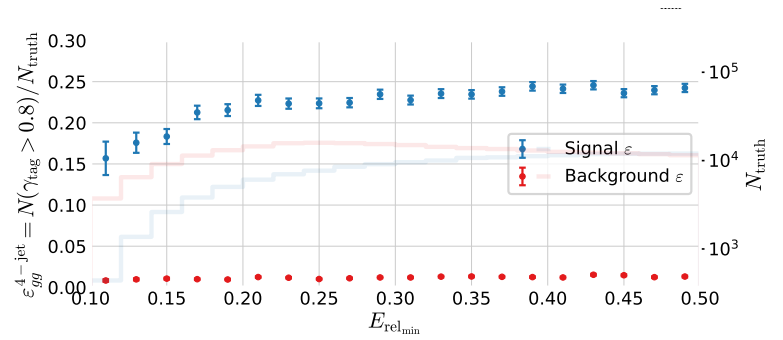


Figure B.41: Efficiency of the g-tagging algorithm for 4-jet events as a function of $E_{\text{rel}_{\text{min}}}$ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

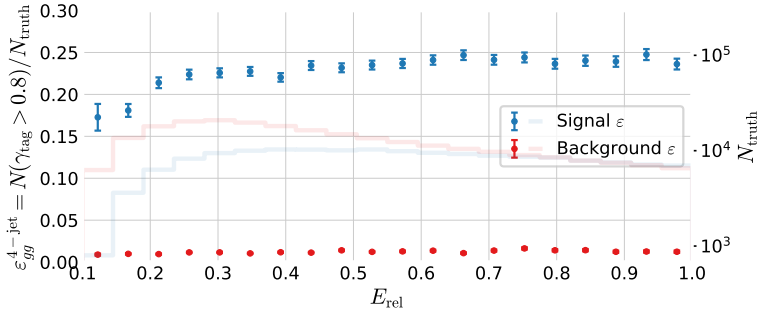


Figure B.42: Efficiency of the g-tagging algorithm for 4-jet events as a function of E_{rel} in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background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

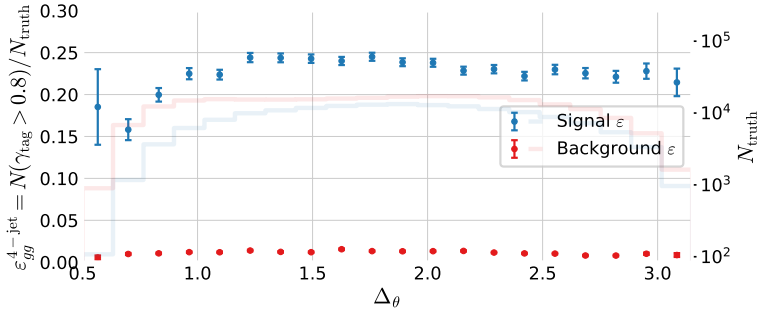


Figure B.43: Efficiency of the g-tagging algorithm for 4-jet events as a function of Δ_θ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background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

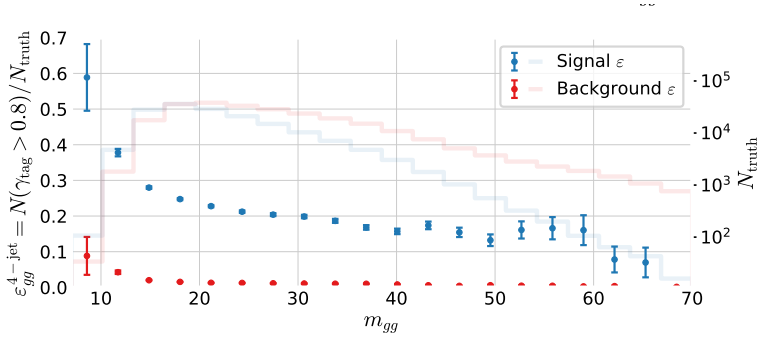


Figure B.44: Efficiency of the g-tagging algorithm for 4-jet events as a function of m_{gg} in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background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

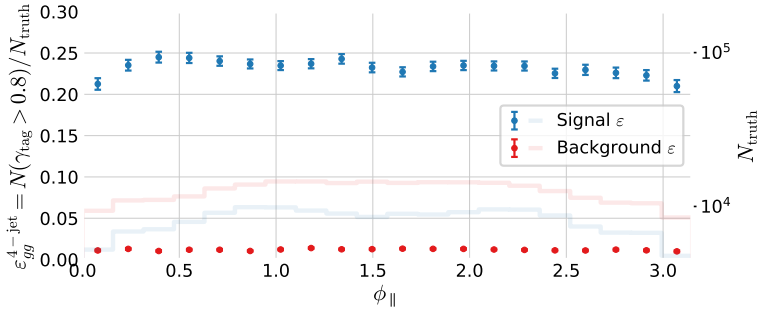


Figure B.45: Efficiency of the g-tagging algorithm for 4-jet events as a function of $\phi_\parallel$ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue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

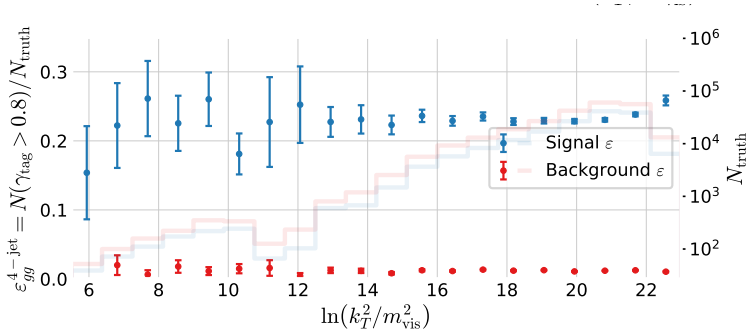


Figure B.46: Efficiency of the g-tagging algorithm for 4-jet events as a function of $\ln\left(k_t^2/m_{\text{vis}}^2\right)$ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

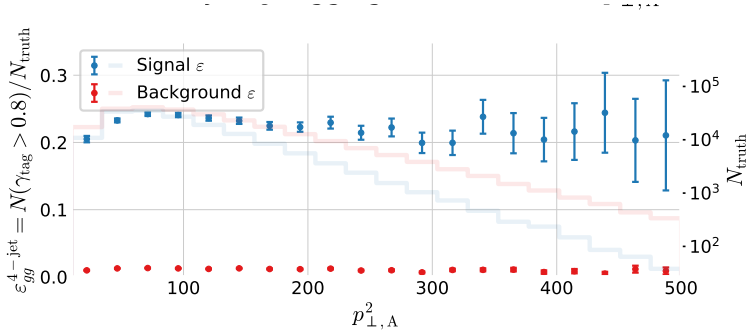


Figure B.47: Efficiency of the g-tagging algorithm for 4-jet events as a function of $p_{\perp,\text{A}}^2$ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

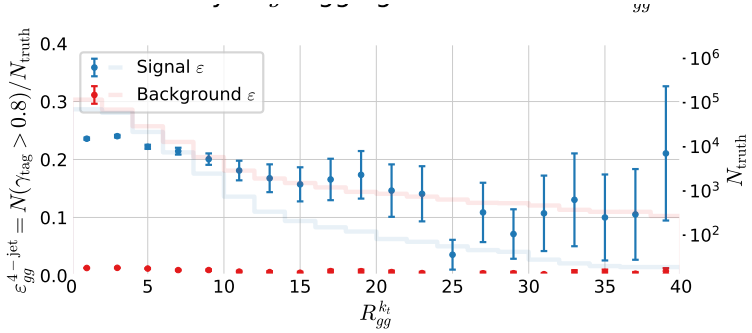


Figure B.48: Efficiency of the g-tagging algorithm for 4-jet events as a function of $R_{gg}^{k_t}$ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue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

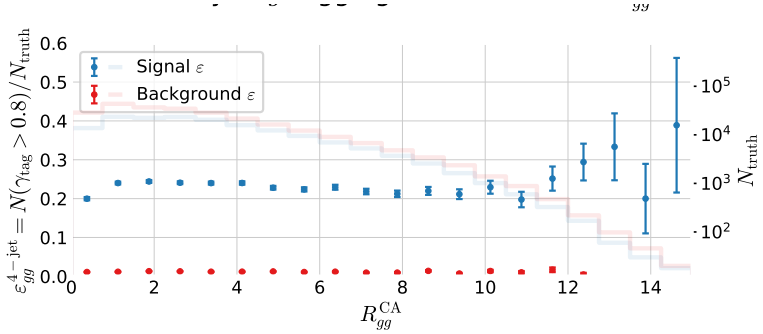


Figure B.49: Efficiency of the g-tagging algorithm for 4-jet events as a function of E in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background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

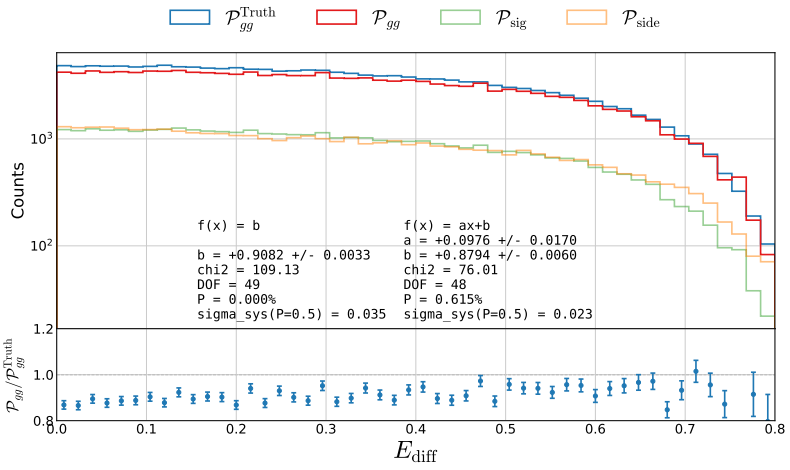


Figure B.50: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for E_{diff}. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the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

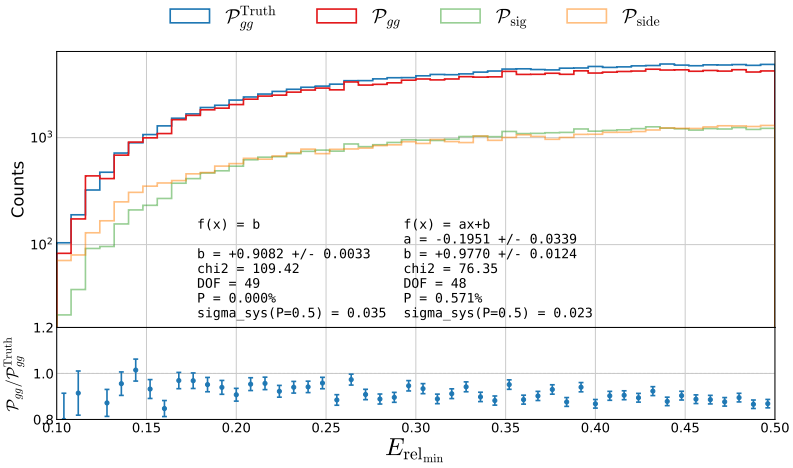


Figure B.51: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for $E_{\text{rel}_{\text{min}}}$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the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

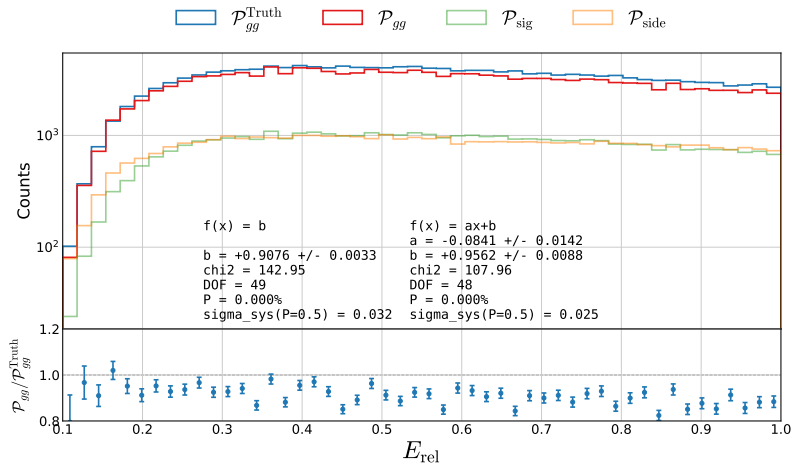


Figure B.52: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for E_{rel}. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the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

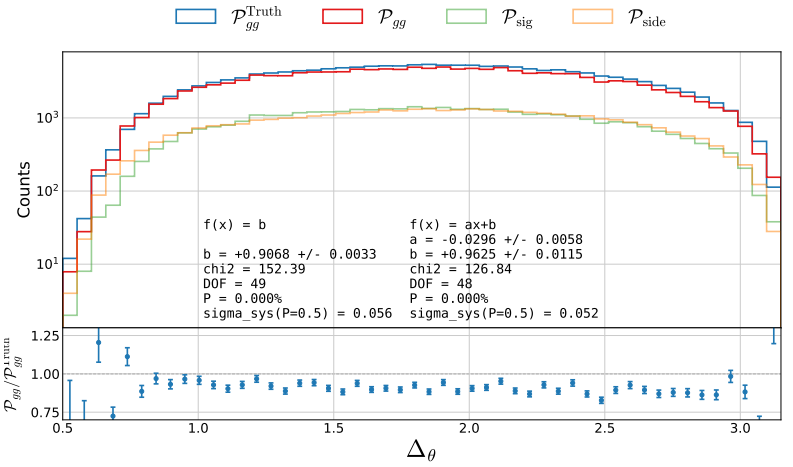


Figure B.53: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for Δ_θ. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the signal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

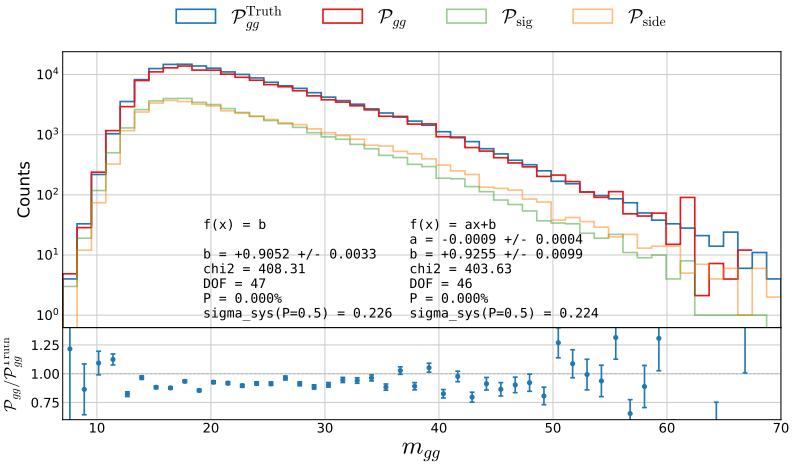


Figure B.54: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for m_{gg}. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the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

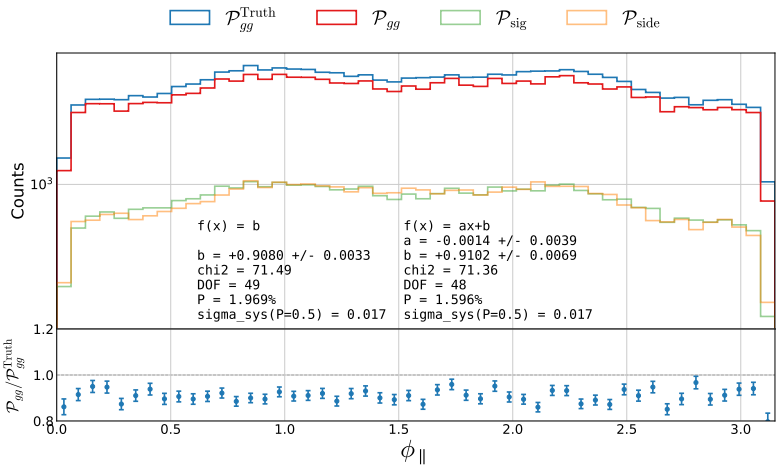


Figure B.55: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for $\phi_{\parallel}$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution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

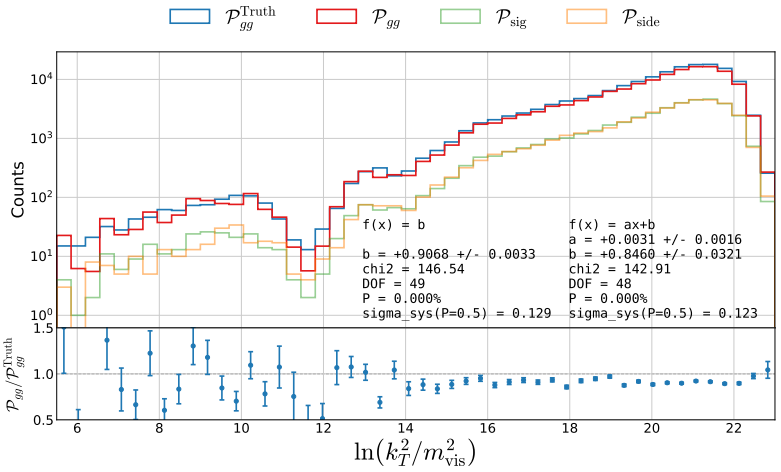


Figure B.56: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for $\ln\left(k_t^2/m_{\text{vis}}^2\right)$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth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

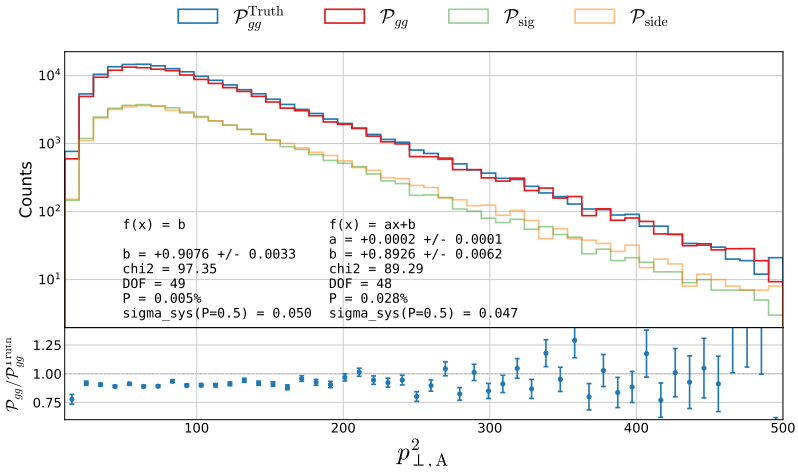


Figure B.57: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for $p_{\perp,\text{A}}^2$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution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

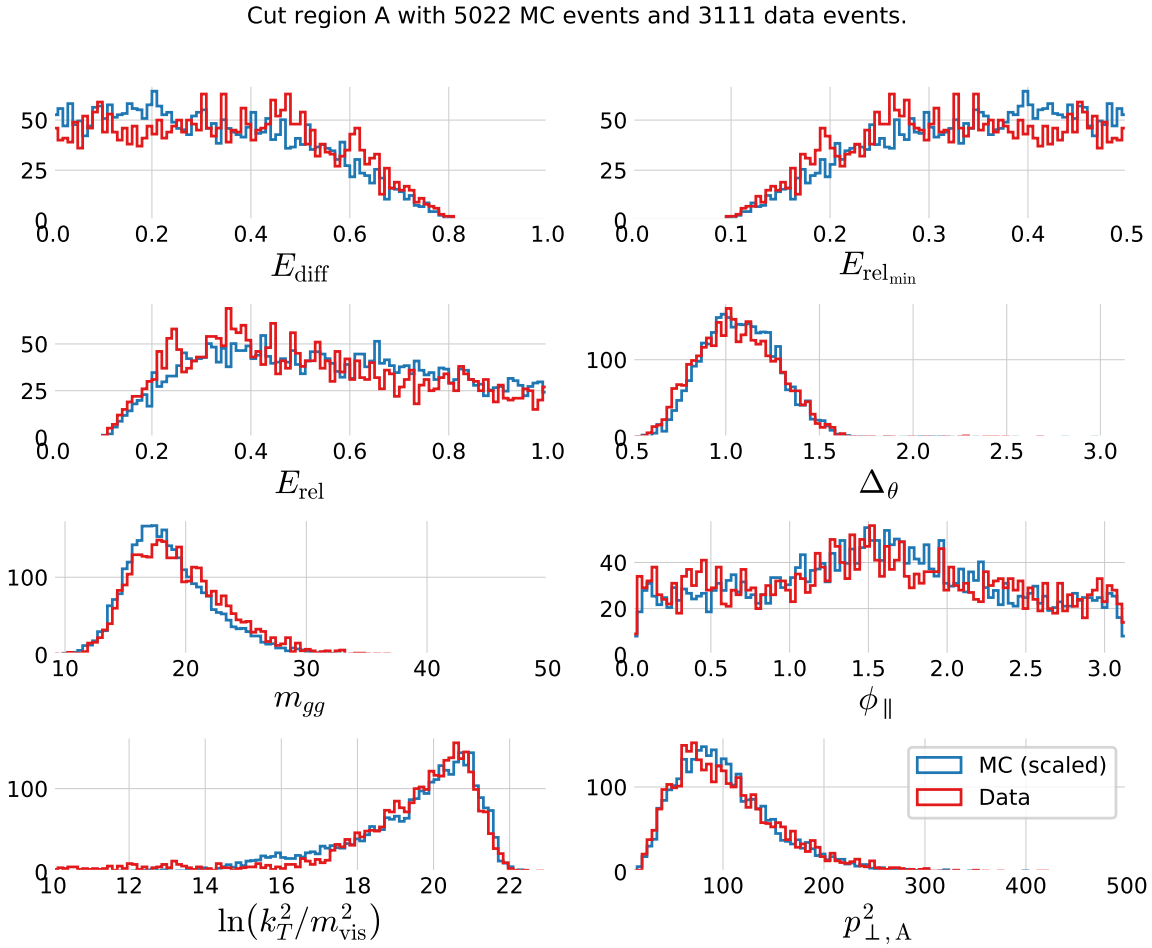


Figure B.58: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area A, see Table 5.7. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area A which has 5022 events in the MC sample and 3111 in the Data sample.

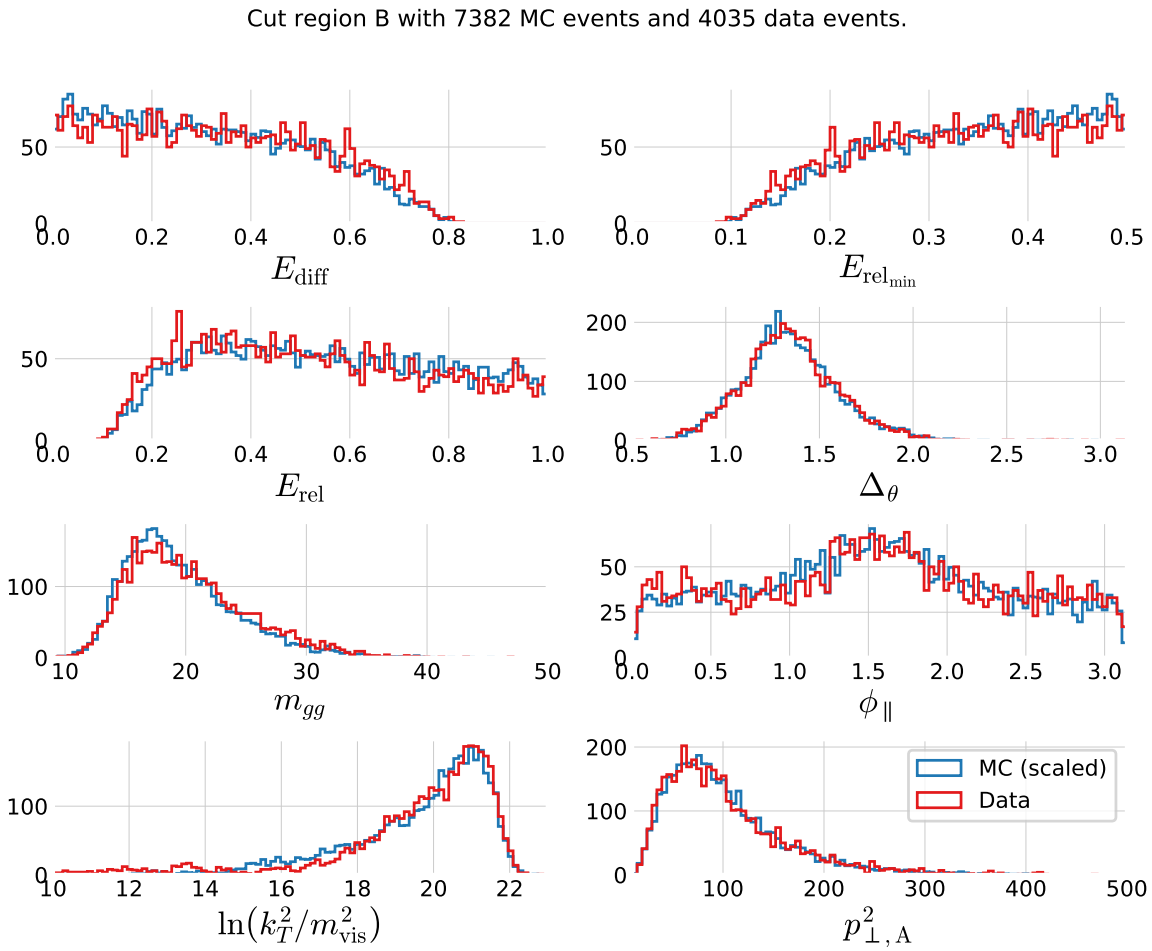


Figure B.59: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area B, see Table 5.7. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area B which has 7382 events in the MC sample and 4035 in the Data sample.

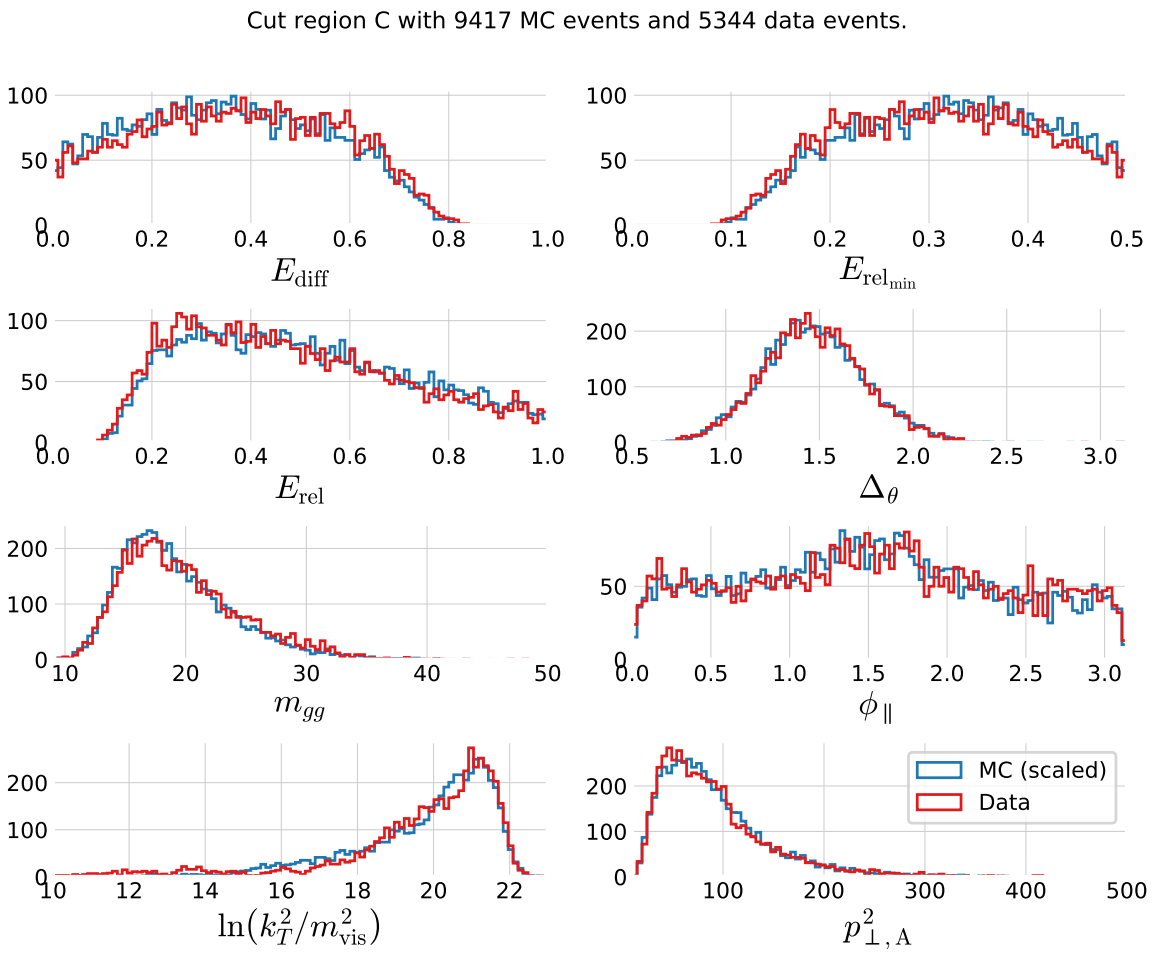


Figure B.60: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area C, see Table 5.7. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area C which has 9417 events in the MC sample and 5344 in the Data sample.

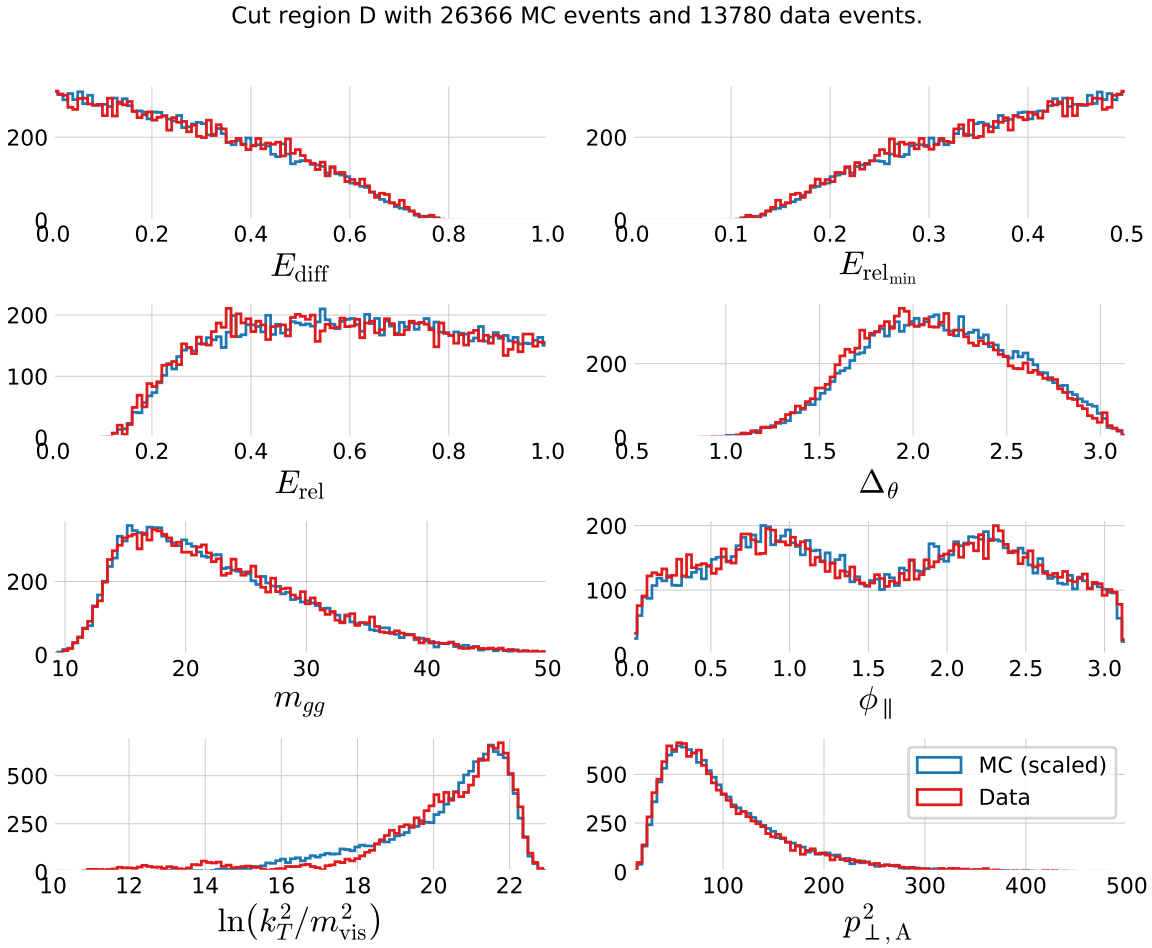


Figure B.61: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area D, see Table 5.7. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area D which has 26 366 events in the MC sample and 13 780 in the Data sample.

List of Figures

List of Tables

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/b3975555d7023fd05f8e56326905fe7b185109de/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. . Www.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. Bittner, 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%3Dxm2%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 Bdo and Bso 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] 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/.

[42] 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.

[43] S. D. DST. Price Index (EJ14) - Statistics Denmark. URL https://www.dst.dk/en/Statistik/emner/priser-og-forbrug/ejendomme.

[44] 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.

[45] 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.

[46] F. Faye. Frederik Faye / deepcalo. URL https://gitlab.com/ffaye/deepcalo.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] A. E. Harvey and S. Peters. Estimation Procedures for Structural Time Series Models. doi: 10.1002/for.3980090203.

[53] T. Hastie and R. Tibshirani. Generalized Additive Models: Some Applications. 82(398):371–386. ISSN 01621459. doi: 10.2307/2289439. URL www.jstor.org/stable/2289439.

[54] 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.

[55] K. Hornik. Approximation capabilities of multilayer feedforward networks. 4(2):251–257. ISSN 0893-6080. doi: 10.1016/0893-6080(91)90009-T. URL http://www.sciencedirect.com/science/article/pii/089360809190009T.

[56] 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=j1OhquR_j88C.

[57] 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.

[58] 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.

[59] R. E. Kalman. A new approach to linear filtering and prediction problems. 82:35–45.

[60] 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.

[61] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. URL http://arxiv.org/abs/1412.6980.

[62] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-Normalizing Neural Networks. URL http://arxiv.org/abs/1706.02515.

[63] 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. URL http://arxiv.org/abs/1408.3122.

[64] 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.

[65] 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.

[66] 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.

[67] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. . URL http://arxiv.org/abs/1802.03888.

[68] A. L. Maas. Rectifier nonlinearities improve neural network acoustic models.

[69] L. McInnes and J. Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. URL http://arxiv.org/abs/1802.03426.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] E. Polley and M. van der Laan. Super Learner In Prediction. URL https://biostats.bepress.com/ucbbiostat/paper266.

[75] 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.

[76] A. Purcell. Go on a particle quest at the first CERN webfest. URL https://cds.cern.ch/record/1473657.

[77] S. Ravanbakhsh, J. Schneider, and B. Poczos. Deep Learning with Sets and Point Clouds. URL http://arxiv.org/abs/1611.04500.

[78] 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.

[79] J. W. Rohlf. *Modern Physics from A to Z*. John Wiley and Sons. ISBN 978-0-471-57270-1.

[80] 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.

[81] 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.

[82] L. Scodellaro. B tagging in ATLAS and CMS. URL http://arxiv.org/abs/1709.01290.

[83] 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.

[84] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, 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.

[85] P. Skands. Peter Skands.

[86] S. J. Taylor and B. Letham. Forecasting at scale. doi: 10.7287/peerj.preprints.3190v2. URL https://peerj.com/preprints/3190.

[87] i. team. Iminuit – A python interface to minuit. URL https://github.com/scikit-hep/iminuit.

[88] J. Thaler. Report of the Les Houches Quark/Gluon Subgroup. (1):28.

[89] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. 58(1):267–288. ISSN 0035-9246. URL www.jstor.org/stable/2346178.

[90] A. Tikhonov. *On the Stability of Inverse Problems*, volume vol. 39 of *Doklady Akademii Nauk SSSR*.

[91] 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.

[92] 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.

[93] 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.

[94] F. van Veen. The Neural Network Zoo. URL http://www.asimovinstitute.org/neural-network-zoo/.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] H. Wickham. Tidy data. 59(10):1–23. ISSN 1548-7660. doi: 10.18637/jss.v059.i10. URL https://www.jstatsoft.org/v059/i10.

[100] M. Wobisch and T. Wengler. Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering. URL http://arxiv.org/abs/hep-ph/9907280.

[101] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola. Deep Sets. URL http://arxiv.org/abs/1703.06114.