CHRISTIAN MICHELSEN
NIELS BOHR INSTITUTE
UNIVERSITY OF COPENHAGEN

A PHYSICIST'S APPROACH TO MACHINE LEARNING – UNDERSTANDING THE BASIC BRICKS

SUPERVISOR:
TROELS PETERSEN
NIELS BOHR INSTITUTE
UNIVERSITY OF COPENHAGEN

Christian Michelsen

https://github.com/christianmichelsen

First printing, December 2019

Contents

List of Figures

List of Tables

1. *Abstract*

Here will be a decent abstract at some point™.

2. Introduction

> *"Begin at the beginning," the King said, gravely, "and go on till you come to an end; then stop."*
>
> — Lewis Carroll, *Alice in Wonderland*

NOT ONLY is the title of this project fairly broad, so are the subjects covered in this thesis. The overall goal of this project is to apply machine learning to different datasets and see how well these comparatively new tools might improve classical statistical methods. The project have dealt with two (seemingly) very different datasets: Danish housing prices and Quark-Gluon discrimination in particle physics, and the aim of this section is to provide an initial overview of the scope and relationship of the two sub-projects; two sub-projects which are covered in each part of this book.

The first part of the thesis deals with the problem of estimating housing prices as precisely and accurately as possible. This was the sub-project that was worked on in the beginning of the overall project and worked as an initial introduction to the application of machine learning to real-life datasets. The housing prices dataset thus became the playground in which the subtleties of these new modern tools were examined, where the difference between real life datasets with all its quirks, outliers and bad formatting, and curated toy datasets that works out of the box (such as the famous Iris dataset [2, 12]) were experienced first hand. Since the project started the dataset changed due to a new collaboration with the Danish housing agency Boligsiden where the agreement was, stated shortly, that we would get their data and they would get our results. Boligsiden is a natural collaborator since they are the biggest on the market[1] and have been very helpful the continuos process of providing data but it also should also be noted that they have had no say on the results presented in this thesis. During this initial stage, the author sparred with Simon Gudiksen[2] who also worked on the same dataset, however, both projects were done independently. Where Gudiksen focussed on the prediction of the time evolution of the housing prices using Recurrent Neural Networks (RNN), my work was mostly on the different levels and methods of hyperparameter optimization with some smaller detours into Natural Language Processing (NLP) as an example.

The second part, the Quark-Gluon discrimination in particle

The background for this masters's thesis, is that it is part of a so-called 4+4 Ph.D. project (also known as an integrated Ph.D.). The Ph.D. dissertation is about the use of machine learning and deep learning in the field of ancient genomics. Here ancient DNA is sampled and analysed with the hope of finding patterns, structure, in the genome which were previously unknown. The overall goal is two-fold. On the big scale it is the better understand human history in the broadest sense of the word history. Where did we come from, where did we go. On a much smaller scale, the goal is to understand local history and migration patterns; how did we end up where we did. It is with this background that this project should be seen: as an introduction to the general use of applied machine learning.

[1] Due to being owned by the "Dansk Ejendomsmæglerforening", The Danish Association of Chartered Estate Agents.

[2] Who afterwards went on to get a job at Boligsiden.

physics, was the main part of the project. Not only was most of the time focussed on this sub-project, it was also the work that generated the highest academic output; an article based on this is in the making. This part dealt with data from the Large Electron Positron collider (LEP) which was an underground particle accelerator at CERN built in 1989 and was discontinued in 2000, where the first phase (LEP1), from 1989-1995, is the sole source of data. As the name suggests it collided electrons and positrons together in what is still the largest electron-positron accelerator ever built [1]. During LEP1 it was primarily the decay channels of the Z-boson that were probed where especially the $Z \to q\bar{q}g$ and $Z \to q\bar{q}gg$ were examined in this thesis. It is especially the distributions of these gluon jets and the difference between Data[3] and MC that are of interests to the theoreticians that develop the MC-models. At first an improved b-tagging algorithm was developed based on only the vertex variables since the primary variables of interest to the theoreticians are the shape variables. Here methods and code developed in the hyperparameter optimization process from the housing prices part were used. After the improvement in the b-tagging model, an event-based g-tag model – in comparison to the jet-based b-tagging model – was implemented which (hopefully) allows one to extract useful events of interest. Having found these useful events, one can start looking at how the distributions in the relevant variables differ between Data and MC. Finally XXX **TODO!**.

[3] Where "Data" with capital D refers to the actual, measured data and "data" refers to any arbitrary selection of data.

The thesis is structured such that **??** introduces the needed theoretical Machine Learning (ML) background needed for understanding the methods used throughout the thesis, **??** describes the housing prices part of the project as mentioned above, **??** introduces the basic physics in the standard model and the Lund string model which is used throughout the rest of the theses, chapter 4 explains analysis of the main project in this thesis, i.e. the quark gluon analysis, and finally the two chapters **??** and **??** discusses the overall work in this thesis and concludes on it.

The work presented in this thesis is split up into two parts as presented above, however, it should be noted that during the analysis part of the project they were treated not as two different projects but rather as two different instances of same underlying problem: teaching computers how to find patterns in high-dimensional data automatically and should thus not be seen as two independent projects. This also highlight another key aspect of this project, that the author does not have any background in particle physics other than rudimentary knowledge stemming from an undergraduate education in general physics.

All of the work presented here is performed by the author unless otherwise noted.

Part I

The first part of this thesis deals with the introductory theory of machine learning and its predictive power in estimating Danish housing prices.

Part II

The second part of this thesis deals with particle physics and the discriminatory power of machine learning for quark-gluon identification and subsequent analysis.

3. Particle Physics and LEP

> *"Not only is the Universe stranger than we think, it is stranger than we can think."*
>
> — Werner Heisenberg

The aim of this chapter is to introduce the reader to the level of particle physics required for understanding the following chapter, in particular introducing the Standard Model in section 3.1, the theory behind quark hadronization in section 3.2, and the ALEPH detector at LEP in section 3.3. The goal is not to make a deep and thorough introduction to the field as this is not needed for the following analysis along with the fact that the author is no particle physicist himself.

3.1 The Standard Model

The *Standard Model* (SM) [13, 16, 18] of particle physics is the currently best known description of the elementary particles and thus describes the fundamental building blocks of our Universe. An overview of the particles explained by the Standard Model is shown in the typical tabular form seen in Figure 3.1. In general, particles comes in two categories: *bosons* and *fermions*.

The fermions, the left part of the figure, are particles with half-integer spin that obey Fermi-Dirac statistics and are further subdivided into *quarks* (upper left in figure) and *leptons* (lower left). The quarks interact with all of the four known forces[1], including the strong force. In contrary the leptons do not interact with the strong force. Quarks are never observed freely but are always combined into *hadrons* due to *color confinement* which is further explained in section 3.2. An example of this are protons which consists of two up-quarks and a down-quark. Leptons exist as either the charged leptons[2] or as neutral leptons, the so-called neutrinos[3]. The fermions come in three generations with increasing mass.

[1] Gravity, electromagnetism, and the strong and weak force.

[2] The electron e, the muon μ, and the tau τ.

[3] The electron neutrine ν_e, the muon neutrino ν_μ, and the tau neutrino ν_τ.

The bosons, the right part of the figure, are the force-carrying particles (with integer spin and which obey Bose-Einstein statistics) where the gluon g mediates the strong nuclear force (color charge), the photon γ mediates the electromagnetic force (charge), and the two $W^\pm$ and the Z bosons the weak nuclear force (weak isospin). The Higgs boson H, experimentally discovered in 2012 [10], does

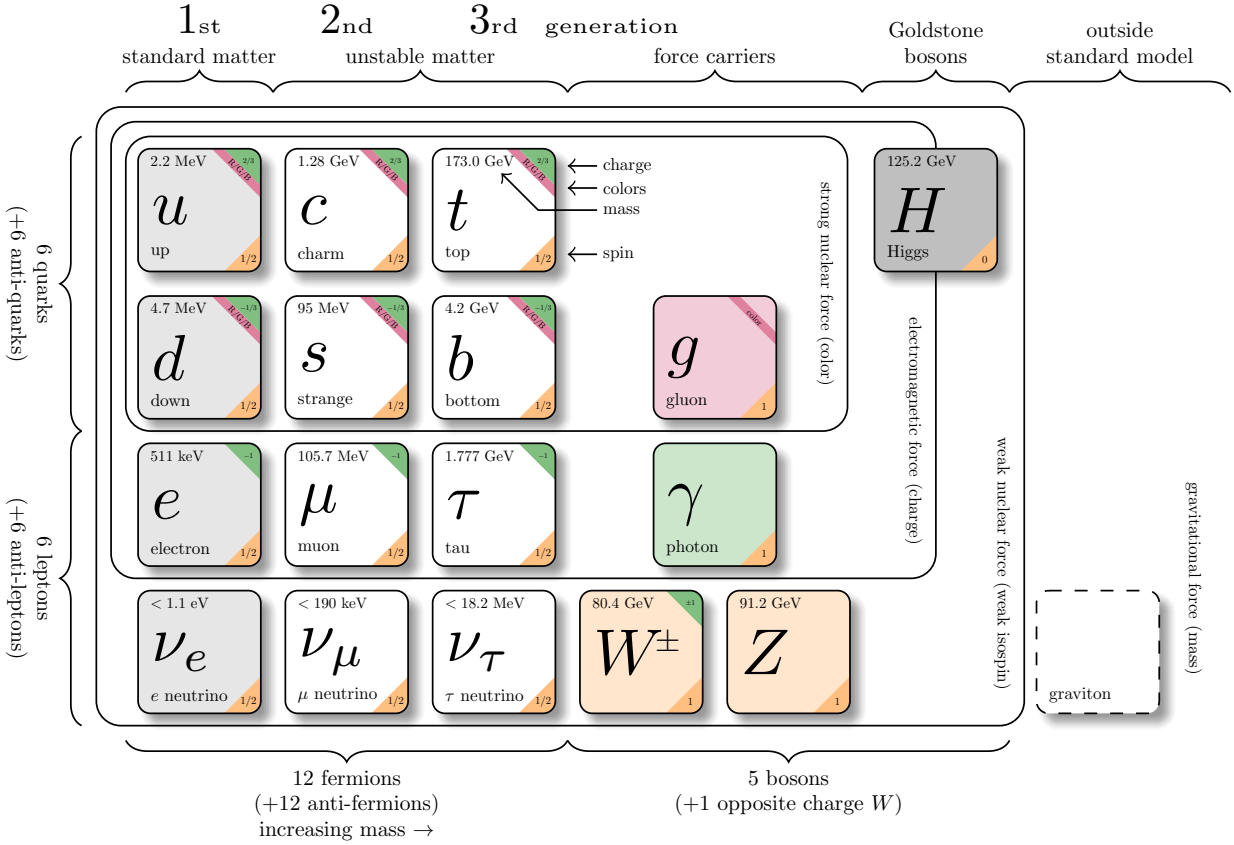

Figure 3.1: The Standard Model. Inspired by Purcell [15] using the template by Burgard [8] with manually updated masses according to Particle Data Group et al. [14].

not mediate any forces but interacts with all massive particles and explains why particles have mass.

All particles have antiparticles which are particles with opposite charge but the same mass. Some particles are their own antiparticles[4], such as the Z. At the Large Electron Positron collider (LEP), see section 3.3, electrons e^- and their antiparticles positrons e^+ were collided at an energy of around 91 GeV. This particular energy was chosen since this is at the resonance peak of the Z which mass distribution follows a Cauchy distribution (also known as Breit-Wigner) with mean[5] $m_Z = (91.1876 \pm 0.0021)\,\text{GeV}$ and a full width of $\Gamma_Z = (2.4952 \pm 0.0023)\,\text{GeV}$: LEP was as such a Z-factory. The Z, however, is only very short-lived with a half-life of $1/\Gamma_Z \sim 2.6 \times 10^{-25}\,\text{s}$. The decay mode for this unstable Z particle is primarily to hadrons ($(69.91 \pm 0.06)\,\%$) where the ratio (R) for b-quarks is $R_b = (Z \to b\bar{b}) = (15.12 \pm 0.05)\,\%$ and $R_g = (Z \to ggg) = (15.12 \pm 0.05)\,\%$ for gluons [14]. The fact that the Z is its own anti-particle forces its decay to be a particle–anti-particle decay (due to charge-conservation) where antiparticles are written with a a bar on top, e.g. the $\bar{b}$-quark is the antiparticle of the b-quark.

[4] The photon, the Z, and the Higgs.

[5] Calculated in natural units where $c = \hbar = 1$ which will also be used throughout this thesis.

3.2 *Quark Hadronization*

The electron-positron e^+e^- annihilations at LEP are complicated events that require advanced high-energy particle physics theory to be properly understood. Most of the aspects of the process is well-described by now, however, especially the hadronization process is still an area of active research. To better get an overview of the different stages of the e^+e^- annihilations, see the Feynman diagram in Figure 3.2.

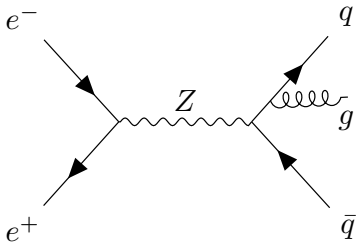

Figure 3.2: Feynman diagram showing the $e^+e^- \to Z^0$ production at LEP. The Z has several decay modes where the $Z \to q\bar{q}g$ is shown here.

Reading from left to right, the electron and the positron annihilates to a Z. This interaction is well-described by quantum electrodynamics (QED), a theory that has been around for more than 60 years by now. As mentioned in the previous section, the Z has several decays modes, yet most of these are background processes of no interest in this project and the focus for now will be the decay mode $Z \to q\bar{q}$ (Z to quark–anti-quark) as seen in the Feynman diagram. The particles produced by the Z-decay are called primary *partons*. Since this process involves quarks, and thus color charge, QED is no longer an adequate theory: quantum chromodynamics (QCD) is needed [4]. The $q\bar{q}$ pairs in this example acts as (color) dipoles from which a gluon can radiate. It can be shown with QCD that the gluon can only be radiated inside the cone that the $q\bar{q}$ pairs spans [6]. As mentioned in the introduction, quarks cannot exist freely (due to *confinement*) and we therefore cannot observe the $q\bar{q}g$ event produced in the Feynman diagram. Confinement is basically the QCD principle saying that quarks are always confined or bound inside hadrons. The initial partons (carrying color charge) are converted to (color-neutral) hadrons by non-perturbative QCD processes in what is called *hadronization*, and these hadrons can be measured.

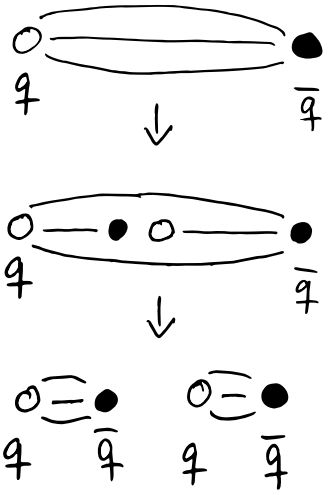

Figure 3.3: Illustration of the quarks splitting as explained by the Lund string model. For large charge separation the (color) field lines seem to be compressed to a tube-like region, where the strong interactions are mediated by the massless gluons (that couple to the color charge of quarks). When the two quarks are separated enough, the potential energy is released by the production of a new $q\bar{q}$ pair.

The hadronization process is not yet fully developed and currently two competing models for predicting the hadronization pattern exists: the Lund string model and the cluster model. In this project only the former of the models will be used. The Lund string model [3] is the theoretical framework underlying the widely used Monte Carlo event generator PYTHIA [17]. The string model is based on the observation that (color) field lines between quarks seem to compress into a tube-like region mediated by gluons, see the top part of Figure 3.3. The field can be described by a linearly rising potential $V(r) = \kappa r$ at large distances[6], where r is the distance and κ is the strength of potential [7]. This field is similar to the (constant) force of a string: $V(r) = \kappa r \Rightarrow F(r) = -\kappa$ where κ is the to be regarded as the spring tension. As quarks move apart, the potential energy stored in the "string" increases until it is large enough to "snap" and convert its potential energy into mass energy. This mass energy is released with the production of a new $q\bar{q}$ pair, see the rest of Figure 3.3.

An example of the hadronization process, or the transition from initial partons to final hadrons is sketched in Figure 3.4. Here the

[6] At small distances a Coulumb term has to be included, however, this term is assumed to be negligible by the Lund string model.

production of two kaons K^- and K^+, and two pions π^- and π^0 are shown. Since particles are created by "splits" in the "string", and the fact that there is energy-momentum conservation, they all have to share the total energy stored in the string. This is described by the fragmentation function:

$$f(z) \propto \frac{(1-z)^a}{z} \exp\left(-\frac{bm^2}{z}\right), \tag{3.1}$$

where $0 \leq z \leq 1$ is the remaining momentum that the new hadron takes, a and b are constants, and m is the mass[7] [6]. When the system runs out of available momentum, it will stop producing new hadrons and the fragmentation function thus explains the distribution of final state particles. The Lund string model can be extended from only $q\bar{q}$ events to $q\bar{q}g$ events where it predicts cones spanning the angular regions qg and $\bar{q}g$ should receive enhanced particle production compared to the $q\bar{q}$ region. This prediction by the Lund string model is also measured in e^+e^- collisions [7].

[7] Where $m \to m_\perp$ for particles with transverse momentum.

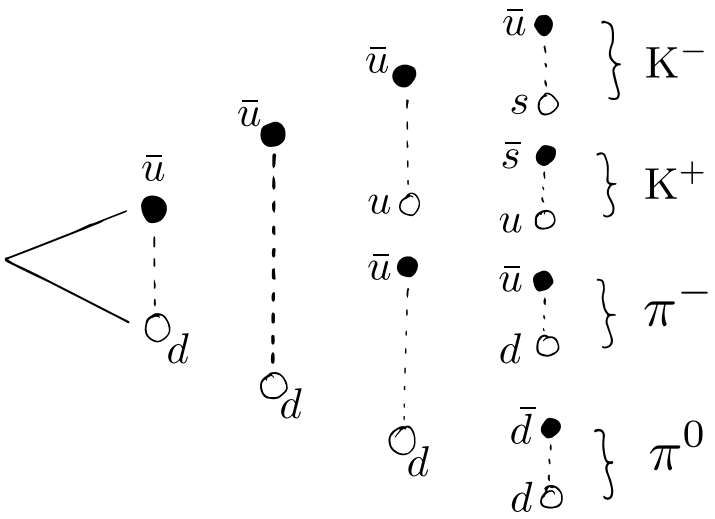

Figure 3.4: Illustration of the hadronization process by which $\bar{u}$- and d-quarks decay into four different mesons. The theoretical strings are shown as dashed lines and particles as circles, where filled circles are antiparticles.

The initial partons produced as Z decay therefore decay themselves to final state hadrons[8] which create a whole "shower" in the direction of the initial parton: this is called a *parton shower* and it is this parton shower, a *jet*, that is measured in the detector. The reverse computation from tracks measured in the detector is done with the use of *jet clustering* algorithms. The detector and the clustering algorithms are described in the following section.

[8] To either mesons which consist of two quarks (color–anti-color) or baryons (r-g-b) which consist of three quarks.

3.3 The ALEPH Detector and LEP

The Large Electron Positron collider (LEP) was a particle collider at CERN in Switzerland operating from 1989 to 2000. It collided counter-rotating bunches of electrons and positrons in a giant ring with a circumference of more than 26 km. The first phase, LEP1, ran from 1989 to 1995 at the Z resonance 91 GeV and the second phase, LEP2, continued afterwards closer to 200 GeV for W^+W^-

pair production [4], however, it is only the data collected at the energy $\sqrt{s} = 91.3\,\text{GeV}$ called the *Z peak data* that is used throughout the rest of this project. There were four independent detectors at the LEP experiment, one of them ALEPH[9].

[9] Together with DELPHI, L3, and OPAL.

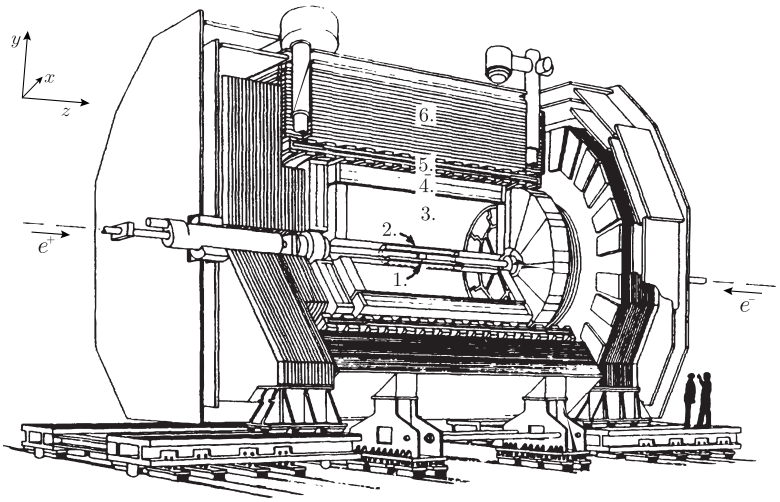

Figure 3.5: The ALEPH detector at LEP. 1) Vertex detector (VDET). 2) Drift chamber (ITC). 3) Time projection chamber (TPC). 4) Electromagnetic calorimeter (ECAL). 5) Superconducting magnet coil. 6) Hadron calorimeter (HCAL). Adapted from Buskulic et al. [9].

The *apparatus for LEP physics* (ALEPH) was a particle detector at LEP with a wide coverage, almost 4π, consisting of cylindrical subdetectors, see Figure 3.5, with the coordinate system shown in the upper left corner[10]. The polar angle θ is illustrated in Figure 3.6 together with the transverse (longitudinal) momentum $p_\perp$ (p_L) and the azimuthal angle ϕ in Figure 3.7. The goal of ALEPH was to measure the energy deposited in calorimeters by charged and neutral particles, measure the momenta of charged particles, measure the distance of travel of short-lived particles, and to identify the three lepton flavors (electron, muon, tau) [9]. As can be seen in Figure 3.5, ALEPH consisted of three detectors (the vertex detector (VDET), the drift chamber (ITC), and the time projection chamber (TPC)) and two calorimeters (the electromagnetic (ECAL) and the hadronic calorimeters (HCAL)).

[10] The z-axis pointing along the beam direction, the y-axis pointing upwards, and the x-axis pointing towards the center of LEP.

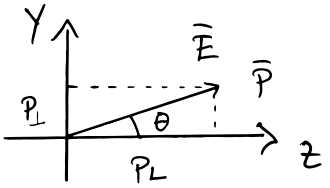

Figure 3.6: The polar angle θ defined in the zy coordinate system

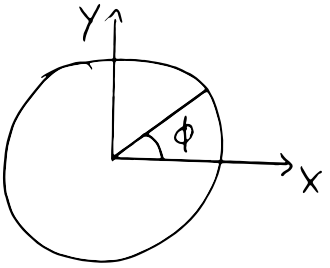

Figure 3.7: The azimuthal angle ϕ defined in the xy coordinate system.

The detectors allow for precise tracking of the charged particles produced in the parton shower and the calorimeters of precise energy measurements for both charged and neutral particles going through the detector.

A hadronic event from a parton shower may leave a score of charged tracks resulting in hundreds of hits in the detectors (VDET, ITC, and TPC) which are fitted[11] with Kalman filters to obtain global track fits, of which bad charged tracks are discarded for further analysis. The tracks are helical due to the presence of a 1.5 T magnetic field which curves the charged particles according to their $p_\perp$.

The energy resolution σ of the calorimeters, or the *calorimeter performance*, is expected to increase with $\sqrt{E}$. In fact, it was found at ALEPH that the energy dependence of the resolution follows the

[11] the process of fitting tracks is called *track reconstruction* in high energy particle physics.

parametrization [9]:

$$\sigma(E) = \Big((0.59 \pm 0.03) \cdot \sqrt{E/\mathrm{GeV}} + (0.6 \pm 0.3)\Big)\,\mathrm{GeV}. \tag{3.2}$$

Even though $\sigma(E)$ increases with E, the relative resolutions improves with higher energies. Since one never measures Nature directly, the results one obtains in a measurement are thus products of both model and experimental uncertainties folded together. To unfold the measurements to obtain experiment-independent results, the uncertainties are important to understand. Of course there are dozens of other uncertainties in an advanced experiment like LEP, however, the energy dependence is the primary focus in this project.

3.4 *Jet clustering*

Since the initial partons created as decay products from the Z are unstable themselves, what is measured in the detector is a whole shower of hadrons seen as charged tracks in the detectors and energy deposits in the calorimeters. However, say that the Z decayed to a $b\bar{b}$ event. In this case the two b's would be back-to-back and the final hadrons would be observed approximately in the same direction as the b's were created. The interest of the experiment is not to measure the final hadrons, but rather to infer information about the initial quarks and gluons. This is done via the reverse-engineering process called *jet clustering*. Over the years many clustering algorithms have been developed, however, most of these are younger than LEP. In the ALEPH experiment the JADE algorithm was used [5]. JADE is a sequential recombination algorithm where final state particles are initially described as individual so-called pseudo-jets which are then recursively merged to larger jets according to their inter-jet distance d_{ij}^2. The distance measure for JADE is:

$$d_{ij}^2 = \frac{2E_iE_j(1-\cos\theta_{ij})}{E_{\mathrm{vis}}^2}, \tag{3.3}$$

where E_{vis} is the visible energy[12] and θ_{ij} is the angle between jet i and j. The JADE algorithm computes d_{ij}^2 for all combinations of jets and merges the two jets with the lowest d_{ij}^2, continuing like that recursively until $\min(d_{ij}^2) > d_{\mathrm{cut}}^2$ for some predefined value of d_{cut}^2. In the dataset at hand, only the final jets were available and not the jet constituents, unfortunately.

[12] The total sum of energies in the event.

3.5 *The variables*

The overall goal of the project is to be able to discriminate quarks and gluons using only vertex variables. The reason for the last condition is that the goal is to better understand the shape distributions of gluons in which there is still significant differences between Monte Carlo (MC) simulations and Data. Therefore only vertex

variables will be used to avoid any biases introduced by using shape-related variables to detect differences in shape-distributions. The vertex variables are a subset of all variables which include the three variables `projet`, `bqvjet`, and `ptlrel`. These three particular variables have each shown discriminatory power in separating b-quarks from light quarks and gluons.

`projet`: For each track in the jet an impact parameter δ is computed. This parameter is the minimum distance between the estimated Z decay point and the track itself and its sign depends on whether or not the point of closest approach is in front of or behind the Z decay point (relative to the momentum). From δ the significance $\mathcal{S}$ – which is δ/σ_δ – is computed and is thus a measure of the certainty of a measured track. High values of $\mathcal{S}$ is typically an indicator of b jets, since long-lived particles typically decay in front of the Z relative to the jet direction, while uds-jets might as well have negative values of $\mathcal{S}$. From $\mathcal{S}$ the track probability $\mathcal{P}_{\text{track}}$ of a track originating at the decay point of the Z can be computed, which can further be aggregated across all tracks within a jet to form the jet probability $\mathcal{P}_{\text{jet}}$ which `projet` is a function of [11]. Whether or not $\mathcal{P}_{\text{jet}}$ is strictly a probability can be discussed but it is related to the probability of all tracks within a jet to originate from long-lived particles, which itself is a good indicator of being a b- (or c-) jet. This variable further has the advantage of being independent of any vertex algorithm.

`bqvjet`: For any jet with good[13] charged tracks, a fit with a (hypothetical) secondary vertex is performed. The difference in χ^2 between the null hypothesis that all good tracks originate from the same primary vertex and the alternative hypothesis that a secondary vertex exists in addition to the primary one is calculated. For the long-lived massive b and c quarks this typically results in large differences in χ^2 compared to uds- and gluon jets which have much lower $\Delta\chi^2$-values [4]. The `bqvjet` is related to the $\Delta\chi^2$-value from the secondary vertex algorithm. This value is dependent of the vertex algorithm, but still explores other areas of phase space than `projet`, however, they are still very correlated[14].

[13] Meaning that there are at least four TPC hits and the fit has a reduced χ^2 of less than four [4]

[14] With a linear correlation of $\rho = 0.82$ in the dataset.

`ptlrel`: If any leptons (in the case of $e^\pm$ or $\mu^\pm$) are measured by the detector, this is a good sign of the jet originating from a b-quark. The high mass of the b quark leads to high $p_\perp$ for the leptons relative to the jet axis which is exactly measured by `ptlrel`.

The fact that the heavy b-quarks have much longer lifetimes than the lighter uds-quarks stems from their much lower inter-coupling

magnitudes written as the CKM matrix **V** [14]:

$$\mathbf{V} = \begin{array}{c} \\ u \\ c \\ t \end{array} \begin{array}{c} \begin{array}{ccc} d & s & b \end{array} \\ \begin{pmatrix} 0.97446 & 0.22452 & 0.00365 \\ 0.22438 & 0.97359 & 0.04214 \\ 0.00896 & 0.04133 & 0.99911 \end{pmatrix} \end{array}. \tag{3.4}$$

The matrix element $|V_{ij}|^2$ is proportional to the transition-probability of quark i transitioning to quark j. From the CKM matrix it can be seen that u and d quarks couples strongly together, likewise with c-s and b-t quark pairs. When a Z decays into a b-quark, this quark couples strongly with the top quark, however, due to the high mass of the top quark compared to the center-of-mass energies at LEP1, the b-quark cannot decay into a t-quark but must (almost always) decay to a c, however, still with low probability, $V_{bc} \ll 1$. This, together with the fact that $V_{bu} \ll V_{bc}$ explains the long life-time of b quarks. This is also why the three variables above are very common variables for b-tagging algorithms. That c-quarks also have relative long life-times are not due to the CKM elements, as for b-quarks, but rather due to the c-decay being governed by the weak force through virtual W^* bosons, a force that is much slower than the strong force. This also happens for b-quarks which further explains why c-quarks share many similarities with b-quarks but also resembles resembles light-quarks (which are very short-lived.).

The rest of the non-vertex variables are:

`ejet` : The energy of the jet E_{jet}

`costheta` : The cosine of the θ angle defined in Figure 3.6: $\cos\theta$.

`phijet` : The angle ϕ of defined in Figure 3.7: ϕ.

`sphjet` : The sphericity tensor **S** is defined as:

$$S^{(\alpha\beta)} = \frac{\sum_{i=1}^{N} p_i^{(\alpha)} p_i^{(\beta)}}{\sum_{i=1}^{N} |p_i|^2} \quad \alpha, \beta \in \{x, y, z\}, \tag{3.5}$$

and the sphericity is determined as $S = \frac{3}{2}(\lambda_2 + \lambda_3)$ where $\lambda_1 \geq \lambda_2 \geq \lambda_3$, $\lambda_1 + \lambda_2 + \lambda_3 = 1$ are the three eigenvalues of the sphericity tensor. The sphericity $0 \leq S \leq 1$ is a measure of the angular distribution of the tracks in a jet. When $S = 0$ the jets form a perfect sphere, compared to $S = 1$ for a perfect jet. The `sphjet` variable is the sphericity of the jet when calculated in its boosted rest frame, also known as *boosted sphericity*.

`pt2jet` : The sum of the square of transverse momentum w.r.t. the jet axis: $\sum_i p_{\perp,i}^2$.

`muljet` : The multiplicity of the jet.

For further details about the variables, see Armstrong [4].

The variables explained above are all used in the following analysis where the machine learning model is trained on onluy the vertex variables to probe differences in the shape-distributions of the shape-variables. The goal of this is to better understand the gluon hadronization process to minimize differences in MC simulations and ultimately get a better understanding of the rules governed by Nature.

4. Quark Gluon Analysis

As any dedicated reader can clearly see, the Ideal of practical reason is a representation of, as far as I know, the things in themselves; as I have shown elsewhere, the phenomena should only be used as a canon for our understanding. The paralogisms of practical reason are what first give rise to the architectonic of practical reason. As will easily be shown in the next section, reason would thereby be made to contradict, in view of these considerations, the Ideal of practical reason, yet the manifold depends on the phenomena.

4.1 Data Preprocessing

The data consists of 43 data files taken 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) simulated (2.1 GB). The data files which are in the form of *Ntuples*, ROOT's data format [**?**], are converted to HDF5-files by using uproot [**?**]. While iterating over the Ntuples, some basic cuts are applied before exporting the data to HDF5. 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 $\mathbf{p}$ in each event is $32\,\text{GeV} \leq p$ to remove any $Z \rightarrow \tau^+\tau^-$ events since XXX. To ensure a primary vertex, at least two good tracks are required where a good track is defined as having 7 TPC hits and $\geq$ 1 silicon hit. 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 XXX. These cuts were standard requirements for the ALEPH experiment (P. Hansen, personal communication, December, 2019, XXX).

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 of 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].

The data structure is quite differently structured in the Ntuples

[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.

compared to normal structured data in the form of tidy data [19]. 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 for both Data, MC, and MCb.

The number of events for each number of jets can be seen in Table 4.1 for the Data and in Figure 4.2 for the MC and MCb.

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

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

	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 4.2: The dimensions for the MC and MCb datasets.

4.2 *Explanatory Data Analysis*

Since the machine learning models are only trained on the three vertex variables `projet`, `bqvjet`, and `ptljet` – see chapter 3 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

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 quarks, one, and only one, of the jets are assigned to the other quark, and no other jets are matched to any of the quarks. We can then define what constitutes a *b*-jet: if it has `flevt` $= 5$, the entire event is q-matched, and the jet is matched to on of the quarks. Similarly we define *c*-jets only with the change that `flevt` $= 5$, and *uds*-jets with `flevt` $\in \{1, 2, 3\}$. A gluon jet is defined as an any-flavor event which is q-matched but the jet is not assigned to any of the quarks. Strictly speaking, this means that *g*-jet is not 100 % certain of being a gluon, however, since the MC simulation does not contain this information this is the only option. Due to the q-match criterion this also means that some jets are assigned the label "non-q-matched" which is regarded as background.

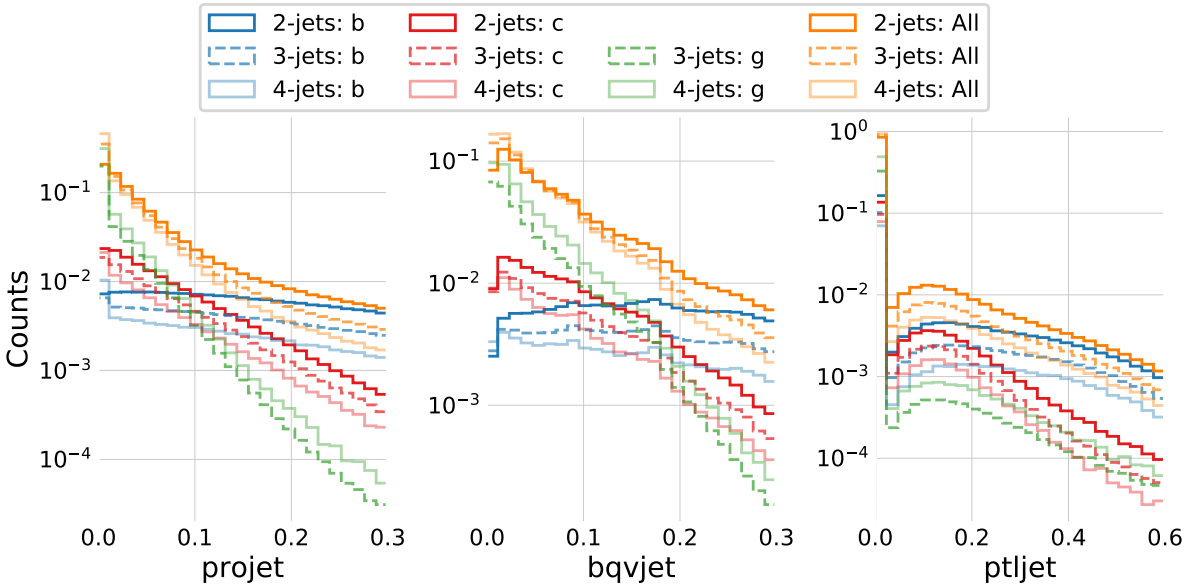

Figure 4.1: Histograms of the three vertex variables, `projet`, `bqvjet`, and `ptljet`, used as input variables in the b-tagging models. In blue colors the variables are shown for true b-jets, in red for true c-jets, in green for true g-jets, and in orange for all of the jets (including non q-matched). 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.

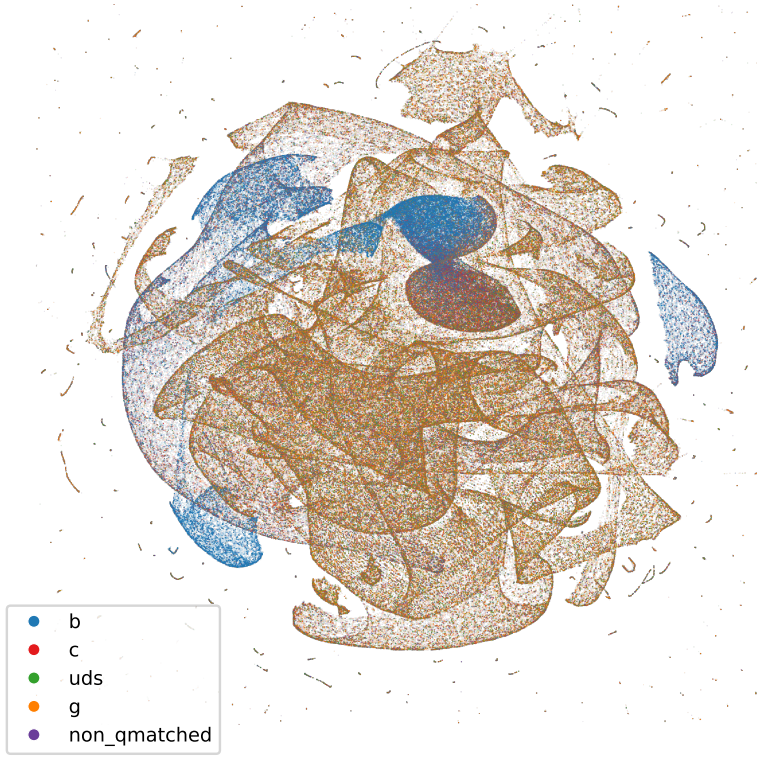

Figure 4.2: Vizualisation of the vertex variables for the different categories: true b-jets in blue, true c-jets in red, true uds-jets in green, true g-jets in orange, and non q-matched. The clustering is performed with the UMAP algorithm which outputs a 2D-projection. This projection is then visualized using the Datashader which takes takes care of point size, avoids over- and under-plotting, and color intensity.

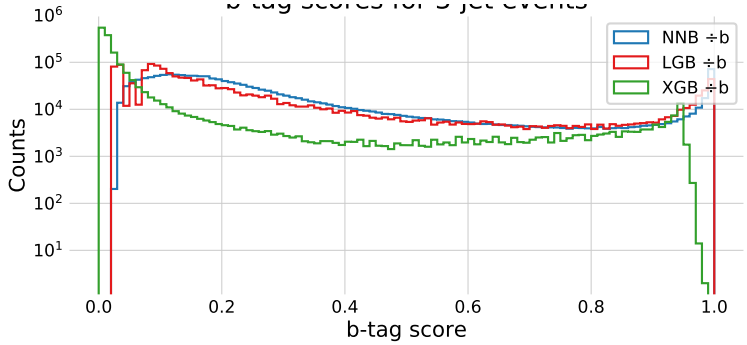

Figure 4.3: Histogram of b-tag scores (model prediction) in 3-jet events for NNB (the neural network trained by ATLAS, also called `nnbjet`) in blue, XGB in red, and XGB in green. We see that the XGB predictions closely match those of NNB which is a good confirmation of a successful fit.

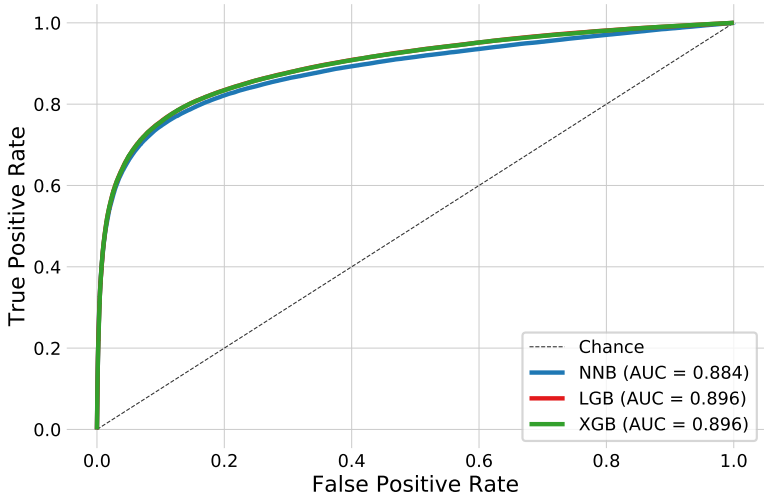

Figure 4.4: ROC curve of the three b-tag models in 3-jet events for NNB (the neural network trained by ATLAS, also called `nnbjet`) in blue, XGB in red, and XGB in green. In the legend the Area Under Curve (AUC) is also shown. Notice that the XGB and XGB models share performance and it is thus due to overplotting that only the green line for XGB can be seen. In the particle physics community False Positive Rate (FPR) is sometimes better known as background efficiency and True Positive Rate (TPR) as signal efficiency.

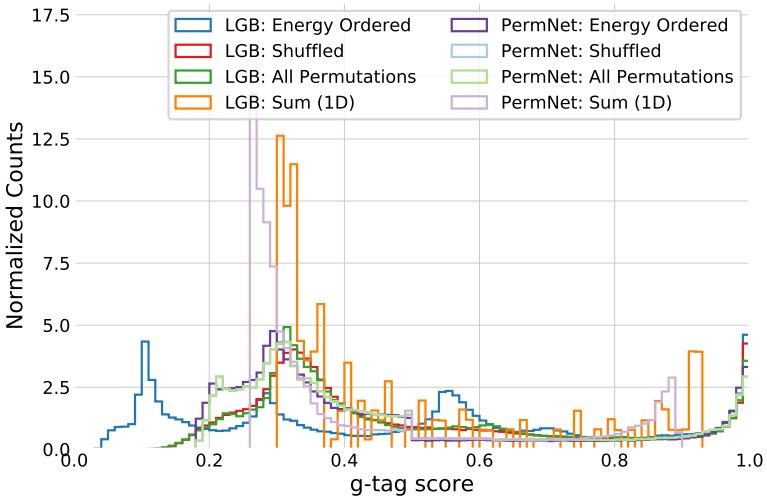

Figure 4.5: Histogram of g-tag scores (model prediction) in 4-jet events for XGB: Energy Ordered in blue, XGB: Shuffled in red, XGB: All Permutations in green, XGB: Sum 1D in orange, PermNet: Energy Ordered in purple, PermNet: Shuffled in light-blue, PermNet: All Permutations in light-green, PermNet: Sum 1D in light-purple. Here XGB 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.

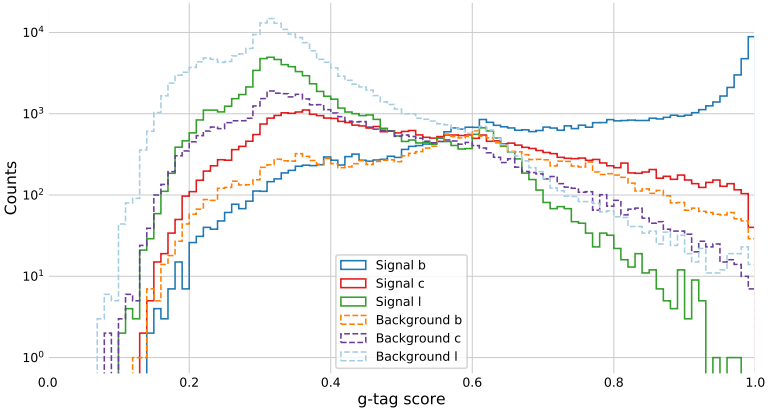

Figure 4.6: Histogram of g-tag scores (model prediction) from the XGB-model in 4-jet events for b signal in blue, c signal in red, l signal in green, b background in orange, c background in purple, l background in light-blue.

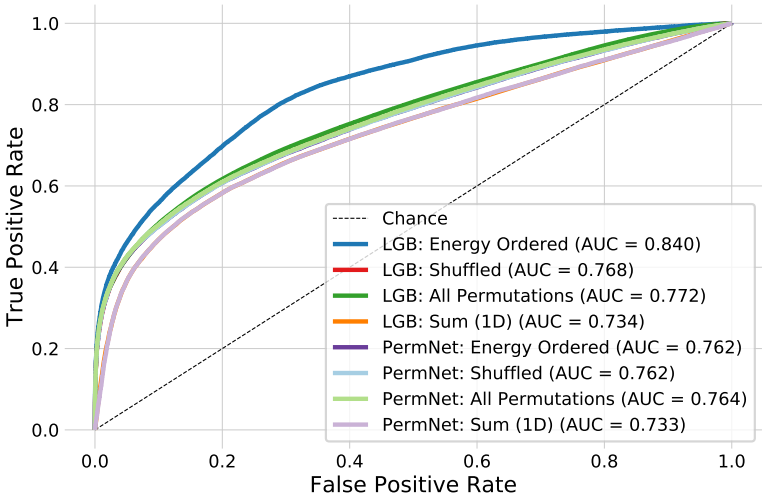

Figure 4.7: ROC curve of the eight g-tag models in 4-jet events. First one in dashed black is the ROC curve that you get by random chance. The colors are the same as in Figure 4.5 and in the legend also the Area Under the ROC curve (AUC) is shown. Notice that the XGB model which uses the energy ordered 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.

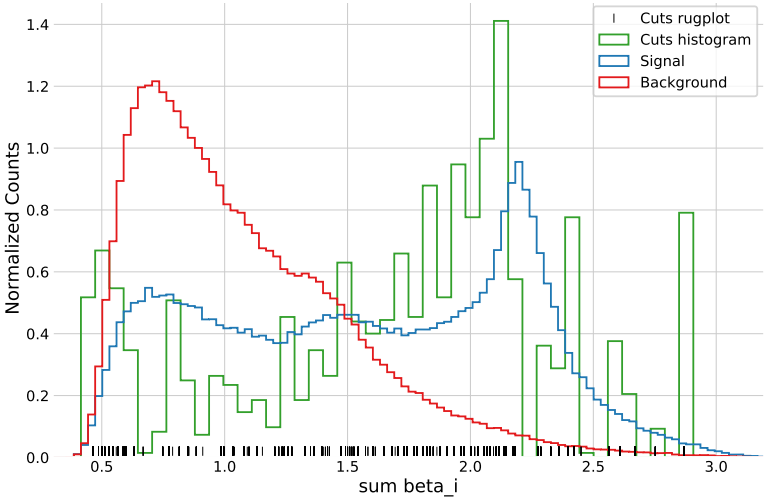

Figure 4.8: Histogram of the distribution of signal in blue and background in red for 1-dimensional sum of b-tags training data. A histogram of the cut values from the XGB model trained on this data is shown in green together with a rug plot of the cut values in black. Notice how most of the cuts match up with the signal peak at around a $\sum \beta_i \sim 2.1$, however, there are also quite a lot of cuts around $\sum \beta_i \sim 0.5$.

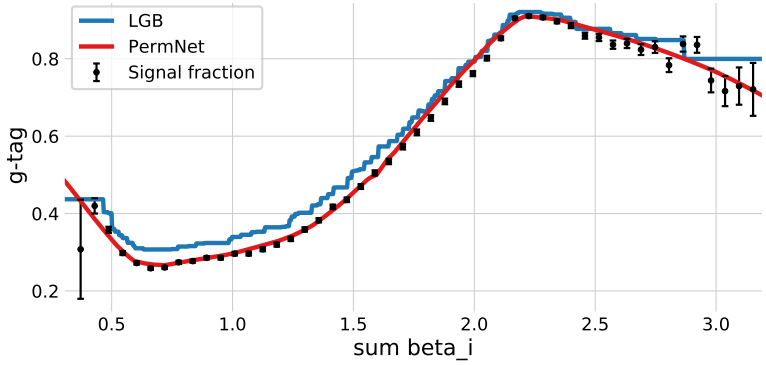

Figure 4.9: Plot of the (1D) g-tag scores as a function of $\sum \beta_i$ for the XGB model in blue and the PermNet model in red. Here the g-tag scores are just the models' output values when input a uniformly spaced grid of $\sum \beta_i$ values between 0 and 4. The signal fraction (based on the signal and background histograms in Figure 4.8) 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. Notice how both models capture the overall trend of the signal fraction with the PermNet being particularly close.

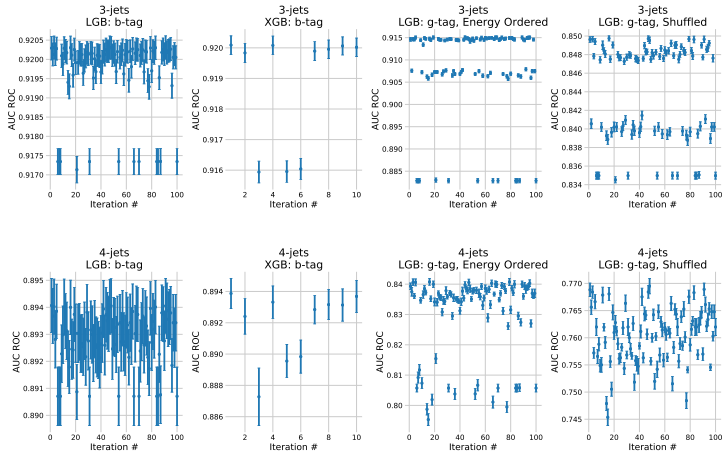

Figure 4.10: Hyperparameter Optimization (HPO) results after running 100 iterations of Random Search (only 10 for XGB). In the top row are the results of the 3-jet models and in the bottom row the results of the 4-jet models. From left to right, we have first) the b-tagging results of XGB, second) the b-tagging results of XGB using only 10 iterations of RS, third) the g-tagging results of XGB fit on the Energy Ordered b-tags, and forth) the g-tagging results of XGB fit on the shuffled b-tags. Notice the different ranges on the y-axes.

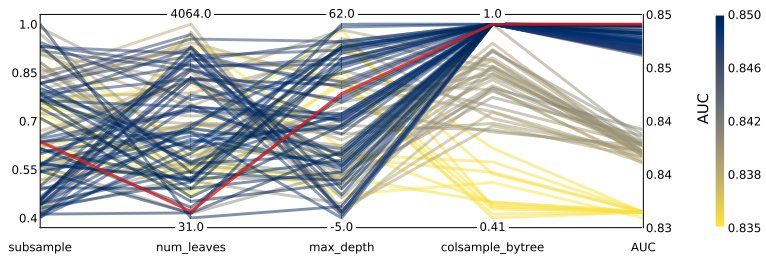

Figure 4.11: Hyperparameter optimization results of g-tagging for 3-jet shuffled events. The results are shown as parallel coordinates with each hyperparameter along the x-axis and the value of that parameter on the y-axis. Each line is an event in the 4-dimensional space colored according to the performance 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.

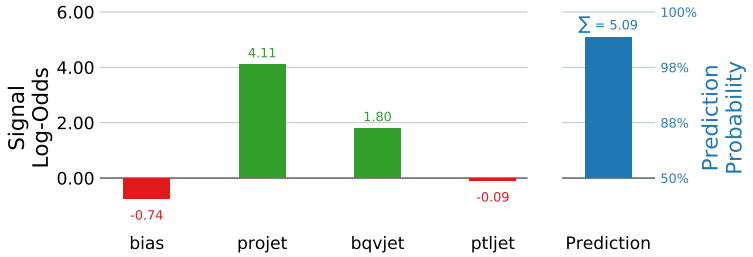

Figure 4.12: Model explanation for the 3-jet b-tagging model for a b-like jet. The first column is the bias of the training set which acts as the naive prediction baseline, the rest are the input data variables. On the right hand side of the plot is the model prediction shown. The left part of the plot is shown in log-odds space, the right part in probability space. The model prediction is the sum of the log-odds (5.09 in this example) transformed into probability space. The negative log-odd values are shown in red, positive ones in green, and the prediction value in blue.

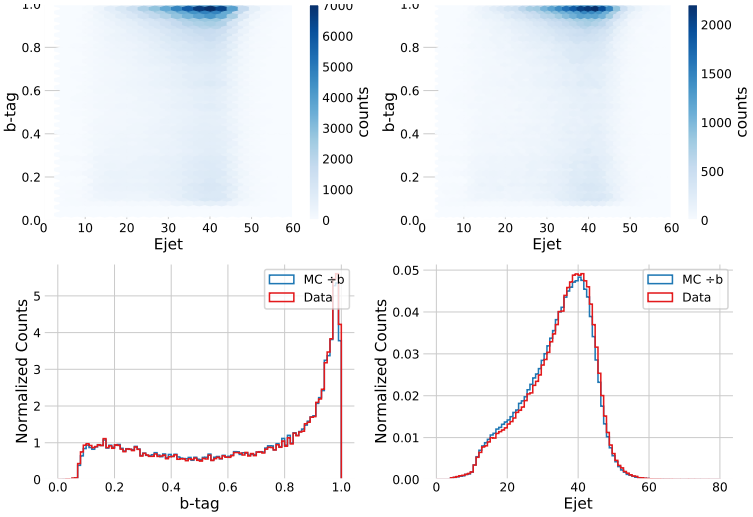

Figure 4.13: Comparison of the b-tag and jet energy (`Ejet`) distributions for Monte Carlo (MC) versus data. In the top row the 2D-distributions are shown for MC on the left (without the extra MCb samples) and data on the right. In the bottom row the 1D marginal distrubtions are shown for the b-tag and the jet energy with data in red and Monte Carlo ones in blue. Notice the the almost identical distributions in b-tag.

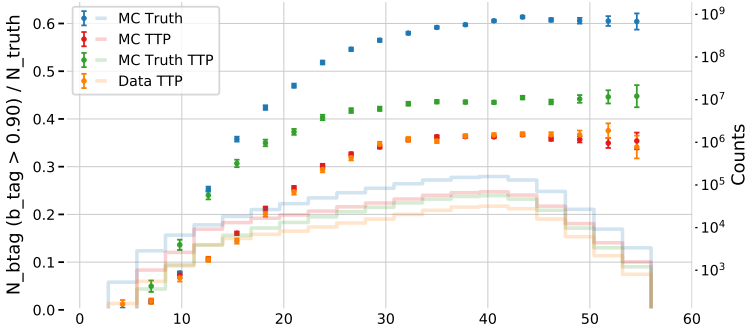

Figure 4.14: Efficiency of the b-tags for b-jets in the b-signal region for 3-jet events, $\varepsilon_b^{b-\text{sig}}$, as a function of jet energy `Ejet` . The b-signal region is defined as $\beta > 0.9$. In the plot the efficiencies are shown for MC Truth in blue, MC TTP in red, MC Truth TTP in green, and Data 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. The abbreviation TTP is short for "Tag, Tag, Probe" where two jets in a event are used as tags and the probe is then used for further analysis. Notice how both MC TTP and Data TTP follow each other closely.

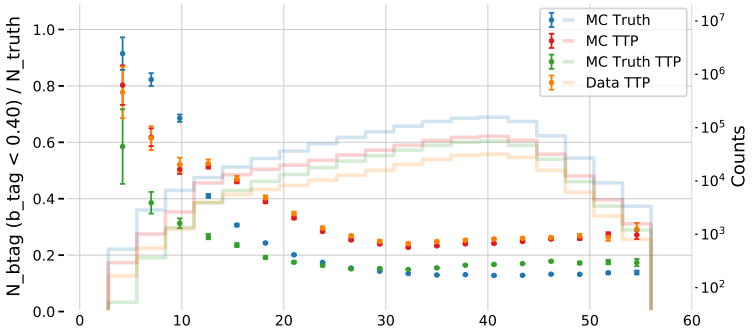

Figure 4.15: Efficiency of the b-tags for b-jets in the g-signal region for 3-jet events, $\varepsilon_b^{g-\text{sig}}$, as a function of jet energy `Ejet` . The g-signal region is defined as $\beta < 0.4$. In the plot the efficiencies are shown for MC Truth in blue, MC TTP in red, MC Truth TTP in green, and Data 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. The abbreviation TTP is short for "Tag, Tag, Probe" where two jets in a event are used as tags and the probe is then used for further analysis. Notice how both MC TTP and Data TTP follow each other closely.

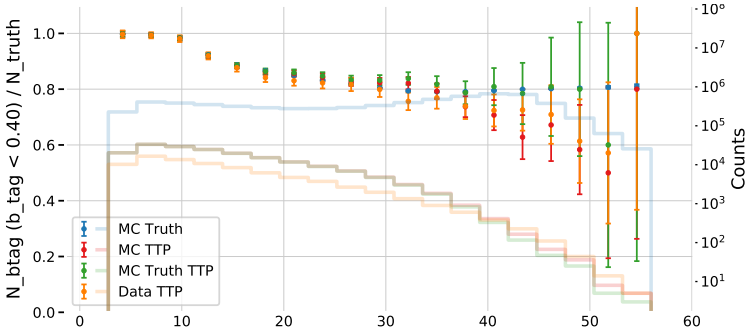

Figure 4.16: Efficiency of the b-tags for g-jets in the g-signal region for 3-jet events, $\varepsilon_g^{g-\text{sig}}$, as a function of jet energy `Ejet`. The g-signal region is defined as $\beta < 0.4$. In the plot the efficiencies are shown for MC Truth in blue, MC TTP in red, MC Truth TTP in green, and Data 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. The abbreviation TTP is short for "Tag, Tag, Probe" where two jets in a event are used as tags and the probe is then used for further analysis. Notice how both MC TTP and Data TTP follow each other closely.

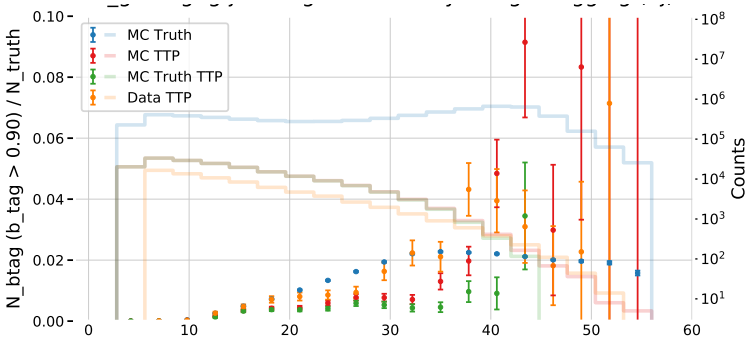

Figure 4.17: Efficiency of the b-tags for g-jets in the b-signal region for 3-jet events, $\varepsilon_g^{b-\text{sig}}$, as a function of jet energy `Ejet`. The b-signal region is defined as $\beta > 0.9$. In the plot the efficiencies are shown for MC Truth in blue, MC TTP in red, MC Truth TTP in green, and Data 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. The abbreviation TTP is short for "Tag, Tag, Probe" where two jets in a event are used as tags and the probe is then used for further analysis.

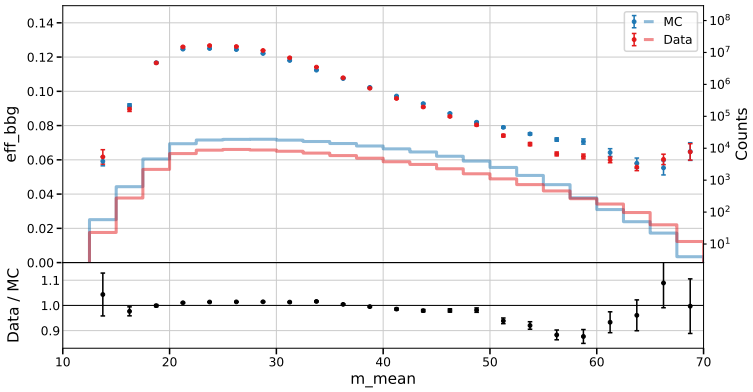

Figure 4.18: Proxy efficiency of the g-tags for $b\bar{b}g$ 3-jet events as a function of the mean of the two invariant masses m_{bg} and $m_{\bar{b}g}$. The proxy efficiency $\varepsilon_{b\bar{b}g}$ is measured by finding $b\bar{b}g$-events where $\beta_b > 0.9$, $\beta_{\bar{b}} > 0.9$, and $\beta_g < 0.4$. and then calculating $\varepsilon_{b\bar{b}g} = \varepsilon_b^{b-\text{sig}} \cdot \varepsilon_{\bar{b}}^{b-\text{sig}} \cdot \varepsilon_g^{g-\text{sig}}$. In the top plot $\varepsilon_{b\bar{b}g}$ 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.

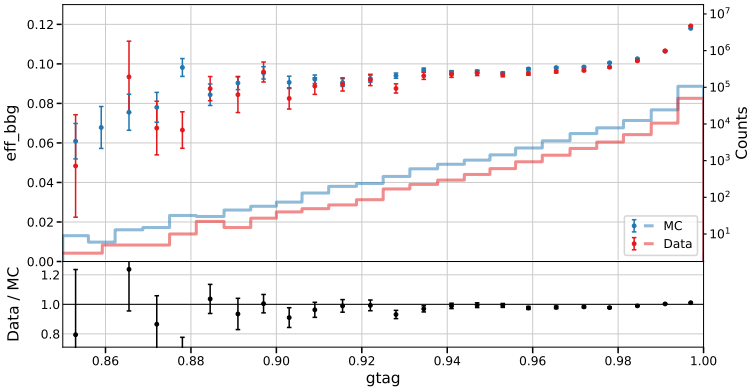

Figure 4.19: Proxy efficiency of the g-tags for $b\bar{b}g$ 3-jet events as a function of the event's g-tag. The proxy efficiency $\varepsilon_{b\bar{b}g}$ is measured by finding $b\bar{b}g$-events where $\beta_b > 0.9$, $\beta_{\bar{b}} > 0.9$, and $\beta_g < 0.4$. and then calculating $\varepsilon_{b\bar{b}g} = \varepsilon_b^{b-\text{sig}} \cdot \varepsilon_{\bar{b}}^{b-\text{sig}} \cdot \varepsilon_g^{g-\text{sig}}$. In the top plot $\varepsilon_{b\bar{b}g}$ 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.

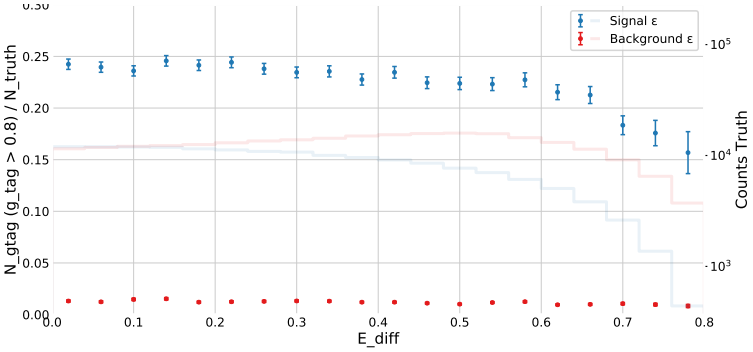

Figure 4.20: Efficiency of the g-tags for 4-jet events as a function of normalized gluon gluon jet energy difference in Monte Carlo. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number and the normalized gluon gluon jet energy difference A is $A = \frac{E_{g_{\max}} - E_{g_{\min}}}{E_{g_{\max}} + E_{g_{\min}}}$ where $E_{g_{\max}}$ ($E_{g_{\min}}$) refers to the energy of the gluon with the highest (lowest) energy. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

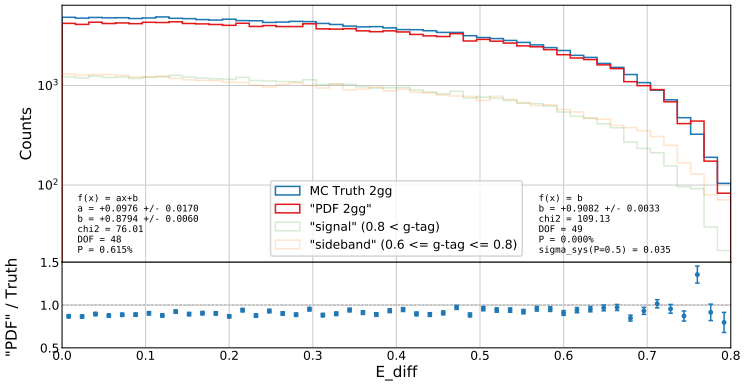

Figure 4.21: Closure plot between MC Truth and the corrected g-tagging model in 4-jet events for the normalized gluon gluon jet energy difference. The corrected g-taggingg model is described in further detail in section XXX **TODO!**. In the top part of the plot the MC Truth is shown in blue, the corrected g-tagging model `"PDF 2gg"` in red, the g-signal distribution in semi-transparent green and the g-sideband distribution in semi-transparent orange. In the bottom part of the plot the ratio between MC Truth and the output of the corrected g-tagging model is shown. The normalized gluon gluon jet energy difference A is $A = \frac{E_{g_{\max}} - E_{g_{\min}}}{E_{g_{\max}} + E_{g_{\min}}}$ where $E_{g_{\max}}$ ($E_{g_{\min}}$) refers to the energy of the gluon with the highest (lowest) energy.

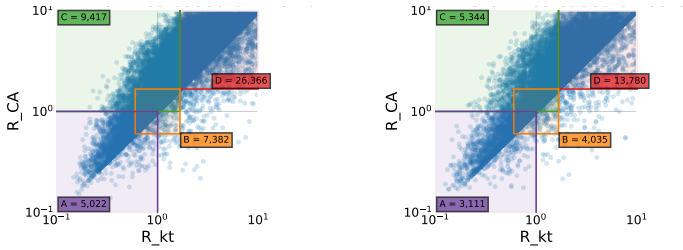

Figure 4.22: R kt CA overview XXX **TODO!**

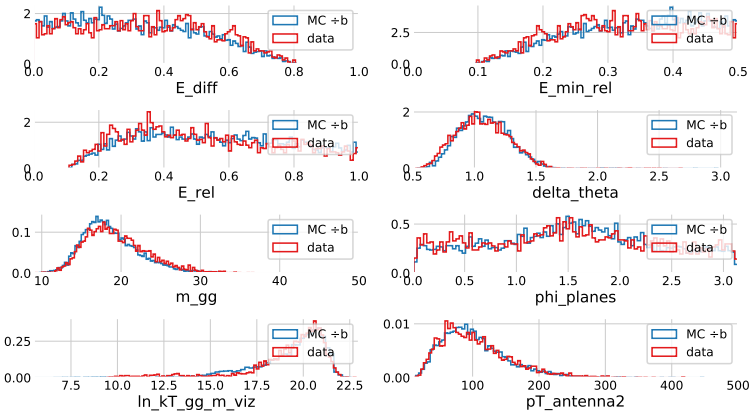

Figure 4.23: R kt CA cut region A XXX **TODO!**

A. *Quarks vs. Gluons Appendix*

Bibliography

[1] The Large Electron-Positron Collider | CERN. URL https://home.cern/science/accelerators/large-electron-positron-collider.

[2] E. Anderson. The Species Problem in Iris. 23(3):457–509. ISSN 00266493. doi: 10.2307/2394164. URL www.jstor.org/stable/2394164.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] C. Burgard. Standard model of physics | TikZ example. URL http://www.texample.net/tikz/examples/model-physics/.

[9] 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.

[10] 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.

[11] 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/0370269393900028G.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

Index