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 Copyright © 2019 Christian Michelsen HTTPS://GITHUB.COM/CHRISTIANMICHELSEN Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at http://www.apache.org/licenses/

LICENSE-2.0. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations

under the License.

First printing, December 2019

Contents

1	Abstract 1	
2	Particle Physics and LEP 7	
	2.1 The Standard Model 7	
	2.2 Quark Hadronization 9	
	2.3 The ALEPH Detector and LEP	10
	2.4 Jet clustering 12	
	2.5 The variables 12	
Α	Quarks vs. Gluons Appendix	17
	Index 21	

List of Figures

2.1 The Standard Model 8
2.2 Feynman diagram for the jet production at LEP 9
2.3 Quark Splitting 9
2.4 Hadronization process 10
2.5 The ALEPH detector 11
2.6 Polar Angle 11
2.7 Azimuthal Angle 11

List of Tables

1. Abstract

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

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.

2. 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 2.1, the theory behind quark hadronization in section 2.2, and the ALEPH detector at LEP in section 2.3. 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.

2.1 The Standard Model

The *Standard Model* (SM) [10, 13, 15] 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 2.1. 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 forces¹, 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 2.2. An example of this are protons which consists of two up-quarks and a down-quark. Leptons exist as either the charged leptons² or as neutral leptons, the so-called neutrinos³. The fermions come in three generations with increasing mass.

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 [8], does

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

² The electron e, the muon μ , and the

³ The electron neutrine ν_e , the muon neutrino ν_{μ} , and the tau neutrino ν_{τ} .

Figure 2.1: The Standard Model. Inspired by Purcell [12] using the template by Burgard [6] with manually updated masses according to Particle Data Group et al. [11].

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 antiparticles⁴, such as the Z. At the Large Electron Positron collider (LEP), see section 2.3, 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 mean⁵ $m_Z = (91.1876 \pm 0.0021)$ GeV and a full width of $\Gamma_Z = (2.4952 \pm 0.0023)$ GeV: LEP was as such a Zfactory. The Z, however, is only very short-lived with a half-life of $1/\Gamma_Z \sim 2.6 \times 10^{-25}$ 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 $R_b = (Z \rightarrow b\bar{b}) = (15.12 \pm 0.05)$ % and $R_g = (Z \to ggg) = (15.12 \pm 0.05)$ % for gluons [11]. The fact that the Z is its own anti-particle forces its decay to be a particle-antiparticle 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.

2.2 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 2.2.

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 $Z \rightarrow q\bar{q}$ (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 [2]. 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 [4]. 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.

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 [1] is the theoretical framework underlying the widely used Monte Carlo event generator PYTHIA [14]. 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 2.3. The field can be described by a linearly rising potential $V(r) = \kappa r$ at large distances⁶, where r is the distance and κ is the strength of potential [5]. 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 2.3.

An example of the hadronization process, or the transition from initial partons to final hadrons is sketched in Figure 2.4. Here the

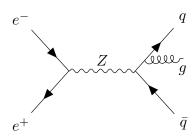


Figure 2.2: Feynman diagram showing the $e^+e^- \to Z^0$ production at LEP. The Z has several decay modes where the $Z \to q\bar{q}g$ is shown here.

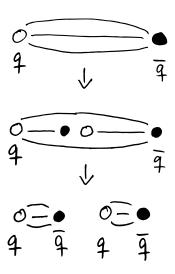


Figure 2.3: Illustration of the quarks splitting as explained by the Lund string model. For large charge separation the (color) field lines seem to be compressed to a tube-like region, where the strong interactions are mediated by the massless gluons (that couple to the color charge of quarks). When the two quarks are separated enough, the potential energy is released by the production of a new $q\bar{q}$ pair.

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:

$$f(z) \propto \frac{(1-z)^a}{z} \exp\left(-\frac{bm^2}{z}\right),$$
 (2.1)

where $0 \le z \le 1$ is the remaining momentum that the new hadron takes, a and b are constants, and m is the mass⁷ [4]. 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 [5].

⁷ Where $m \rightarrow m_{\perp}$ for particles with transverse momentum

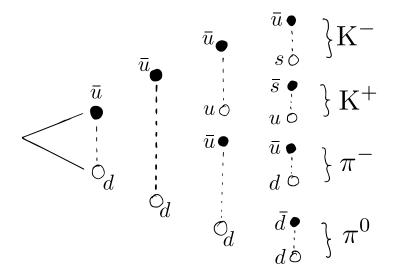


Figure 2.4: Illustration of the hadronization process by which \bar{u} -and d-quarks decay into four different mesons. The theoretical strings are shown as dashed lines and particles as circles, where filled circles are antiparticles.

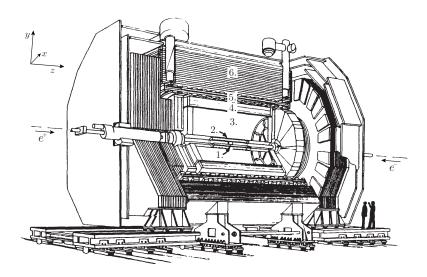
The initial partons produced as *Z* decay therefore decay themselves to final state hadrons⁸ which create a whole "shower" in the direction of the initial parton: this is called a *parton shower* and it is this parton shower, a *jet*, that is measured in the detector. The reverse computation from tracks 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.

2.3 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 [2], 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 ALEPH9.



9 Together with DELPHI, L3, and OPAL.

Figure 2.5: 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.

The apparatus for LEP physics (ALEPH) was a particle detector at LEP with a wide coverage, almost 4π , consisting of cylindrical subdetectors, see Figure 2.5, with the coordinate system shown in the upper left corner¹⁰. The polar angle θ is illustrated in Figure 2.6 together with the transverse (longitudinal) momentum $p_{\perp}(p_L)$ and the azimuthal angle ϕ in Figure 2.7. 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) [7]. As can be seen in Figure 2.5, 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)).

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 fitted¹¹ 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 ¹⁰ The z-axis pointing along the beam direction, the y-axis pointing upwards, and the x-axis pointing towards the center of LEP.

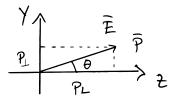


Figure 2.6: The polar angle θ defined in the zy coordinate system

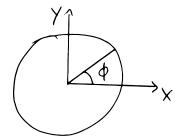


Figure 2.7: The azimuthal angle ϕ defined in the xy coordinate system. 11 the process of fitting tracks is called track reconstruction in high energy particle physics.

parametrization [7]:

$$\sigma(E) = ((0.59 \pm 0.03) \cdot \sqrt{E/\text{GeV}} + (0.6 \pm 0.3)) \text{ GeV}.$$
 (2.2)

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.

2.4 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 [3]. 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:

$$d_{ij}^2 = \frac{2E_i E_j (1 - \cos \theta_{ij})}{E_{\text{vis}}^2},$$
 (2.3)

where $E_{\rm vis}$ is the visible energy¹² 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 $\min(d_{ij}^2) > d_{\rm cut}^2$ for some predefined value of $d_{\rm cut}^2$. In the dataset at hand, only the final jets were available and not the jet constituents, unfortunately.

2.5 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

¹² The total sum of energies in the event.

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 S – which is δ/σ_{δ} – is computed and is thus a measure of the certainty of a measured track. High values of 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 udsjets might as well have negative values of S. From S the track probability \mathcal{P}_{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}_{jet} which projet is a function of [9]. Whether or not \mathcal{P}_{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 good¹³ 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 [2]. 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 correlated¹⁴.

ptlrel: If any leptons (in the case of e^{\pm} or μ^{\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

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

 $^{^{14}}$ With a linear correlation of ho=0.82 in the dataset.

magnitudes written as the CKM matrix V [11]:

$$\mathbf{V} = \begin{pmatrix} d & s & b \\ u & 0.97446 & 0.22452 & 0.00365 \\ 0.22438 & 0.97359 & 0.04214 \\ t & 0.00896 & 0.04133 & 0.99911 \end{pmatrix}. \tag{2.4}$$

The matrix element $|V_{ii}|^2$ 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 2.6: $\cos\theta$.

phijet : The angle ϕ of defined in Figure 2.7: ϕ .

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

$$S^{(\alpha\beta)} = \frac{\sum_{i=1}^{N} p_i^{(\alpha)} p_i^{(\beta)}}{\sum_{i=1}^{N} |p_i|^2} \quad \alpha, \beta \in \{x, y, z\}, \tag{2.5}$$

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 [2].

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.

A. Quarks vs. Gluons Appendix

Bibliography

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] C. Burgard. Standard model of physics | TikZ example. URL http://www.texample.net/tikz/examples/model-physics/.
- [7] 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-o. URL http://www.sciencedirect.com/science/ article/pii/0370269394911770.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] A. Purcell. Go on a particle quest at the first CERN webfest. URL https://cds.cern.ch/record/1473657.
- [13] 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.
- [14] 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.
- [15] 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.

Index

license, ii