GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Fakultät für Physik $[q,p]=i\hbar$

Master's Thesis

Improving Event Reconstruction for Dileptonic Decays of Top Quarks Pairs Using Machine Learning

prepared by

Siemen Henning Aulich

from Emden

at the DESY Hamburg

Course:	M.Phy.1610: Development and Realization of Scientific Projects in Theoretical Physics
Course period:	7th April 2025 until 4th December 2026
First referee:	Prof. Dr. Stan Lai
Second referee:	Dr. Katharina Behr

Abstract

Here the key results of the thesis can be presented in about half a page.

Keywords: Physics, Bachelor thesis

Contents

1. Introduction

2. Theoretical Background

2.1. The Standard Model of Particle Physics

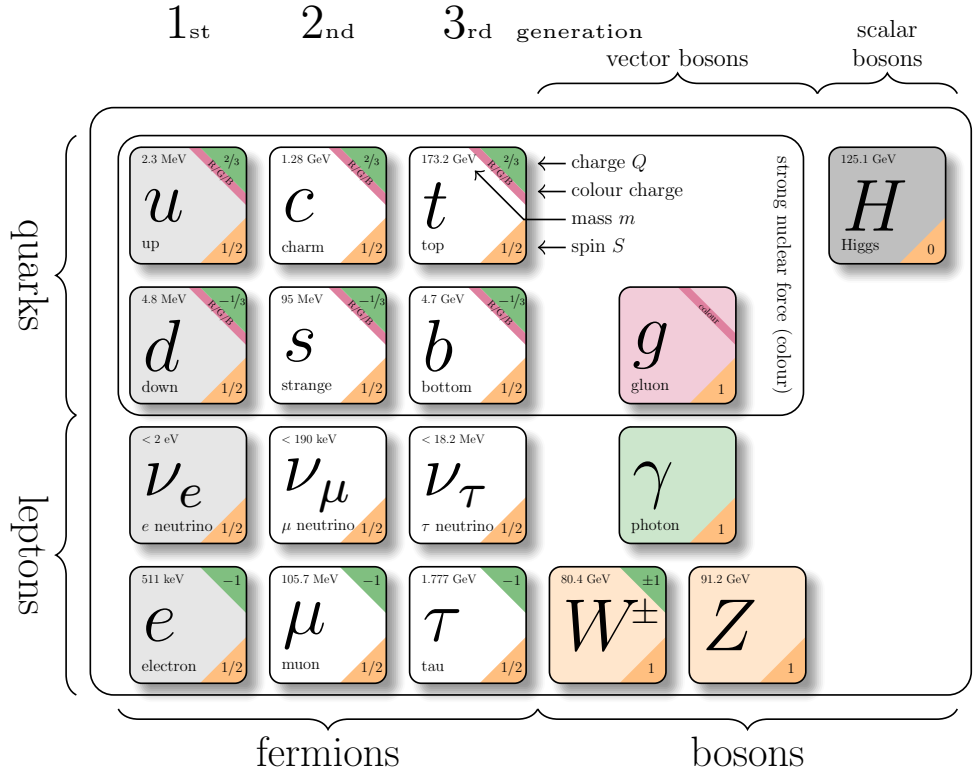

Figure 2.1.: The particles of the Standard Model. Particle properties taken from Ref. [1].

The Standard Model (SM) of particle physics is a quantum field theory describing three of the four fundamental interactions of nature—electromagnetic, weak and strong—and the elementary particles on which they act. It is formulated as a renormalisable, Lorentz-invariant gauge theory with gauge group

$$SU(3)_C \times SU(2)_L \times U(1)_Y,$$

and its structure has been established in a series of seminal works [2–7]. A summary of the particle content is shown in Fig. 2.1.
The elementary particles fall into fermions and bosons. Fermions are spin-$\frac{1}{2}$ fields that constitute matter, organised into three generations of quarks and leptons. Quarks carry colour charge and participate in the strong interaction; leptons do not. Bosons are integer-

spin gauge fields mediating the fundamental forces: the photon for electromagnetism, the $W^{\pm}$ and Z bosons for the weak interaction, and eight gluons for the strong interaction. The Higgs boson completes the SM and is responsible for electroweak symmetry breaking (EWSB), giving mass to the weak gauge bosons and, through Yukawa couplings, to the fermions. The SM Lagrangian can be written schematically as

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}_{\mathrm{gauge}} + \mathcal{L}_{\mathrm{fermion}} + \mathcal{L}_{\mathrm{Yukawa}} + \mathcal{L}_{\mathrm{Higgs}},$$

where the first term contains the gauge kinetic terms and self-interactions, the second describes the fermions and gauge covariant derivatives, the third encodes the Yukawa couplings to the Higgs field, and the last term contains the Higgs potential and EWSB mechanism.
Below, the three interaction sectors are briefly summarised, with emphasis on the aspects relevant for top-quark physics and LHC phenomenology.

2.1.1. Electromagnetic Interaction

The electromagnetic interaction is described by quantum electrodynamics (QED), the $U(1)_{\mathrm{EM}}$ gauge theory that remains after EWSB. Its massless gauge boson, the photon, couples to electric charge and does not self-interact. Because QED is an Abelian theory, its coupling runs only mildly with energy and remains perturbative at all experimentally accessible scales.

2.1.2. Weak Interaction

The weak interaction forms, together with electromagnetism, the electroweak theory based on the gauge group $SU(2)_L \times U(1)_Y$. The vacuum expectation value of the Higgs field spontaneously breaks the symmetry to $U(1)_{\mathrm{EM}}$, giving masses to the weak gauge bosons through the Higgs mechanism [8–10]. The resulting massive $W^{\pm}$ and Z bosons mediate a short-range force.
A defining property of the weak interaction is its *chiral* structure: charged-current interactions couple only to left-handed fermions and right-handed antifermions. This feature, predicted in [11] and confirmed in the Wu experiment [12], leads to maximal parity violation and characteristic angular distributions of weak decay products. The vertex factor of the charged-current interaction is given by

$$\frac{-ig}{2\sqrt{2}}\gamma^{\mu}(1-\gamma^{5}), \tag{2.1}$$

where g is the weak coupling constant, γ^{μ} are the Dirac matrices. The projection operator $(1-\gamma^{5})$ leads to a coupling exclusively to the left-handed components. Based on this, the left-handed particles are arranged in $SU(2)_L$ doublets corresponding, while the right-handed particles are singlets.
For quarks, the weak eigenstates d', s', b' are not identical to their mass eigenstates d, s, b. This misalignment is described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix [13, 14], which encodes flavour-changing transitions in charged-current interactions.

2.1.3. Strong Interaction

The strong interaction is described by quantum chromodynamics (QCD), a non-Abelian gauge theory with gauge group $SU(3)_C$. Quarks carry colour charge, while gluons carry a colour and anticolour index, giving rise to self-interactions through three- and four-gluon vertices. These features lead to two fundamental properties:

Asymptotic freedom. The QCD coupling decreases at high momentum transfer [5, 15], allowing perturbative calculations of hard processes such as top-quark pair production. In the high-energy regime of the LHC, quarks and gluons behave as effectively quasi-free particles.

Confinement. At low energies the QCD coupling becomes large, preventing coloured states from existing in isolation. Instead, they form colour-neutral bound states—hadrons. The transition from short-distance partons to long-distance hadrons involves nonperturbative dynamics encapsulated through hadronisation models.
A crucial consequence for collider physics is that quarks and gluons produced in hard interactions undergo a cascade of QCD emissions before hadronising into jets. This process is typically described in terms of:

- **Parton distribution functions (PDFs)**, characterising the proton structure and entering via the QCD factorisation theorem [16].
- **Initial- and final-state radiation (ISR/FSR)**, arising from soft and collinear gluon emission.
- **Parton showers**, which resum logarithmically enhanced emissions and are implemented in Monte Carlo generators such as PYTHIA, first developed in [17, 18] and extended in modern formulations [19].
- **Hadronisation models**, such as the Lund string model [20, 21], which describe the confinement-driven formation of hadrons.

These aspects of QCD are essential for the reconstruction of hadronic final states at the LHC and for modelling top-quark production and decay.

2.2. Top Quark Physics

The top quark is the heaviest known elementary particle. Its predicted existence arose from the already observed CP-Violiation in the kaon system, which required a third generation of quarks to be explained within the SM framework [14]. Furthermore, the discovery of the bottom quark in 1977 [22] implied the existence of its weak isospin partner, the top quark. The top quark was eventually discovered in 1995 by the CDF and DØ collaborations at the Tevatron collider [23, 24]. Currently, the world average of the top quark mass is measured to be [1]

$$m_t = 172.56 \pm 0.31\,\text{GeV}. \tag{2.2}$$

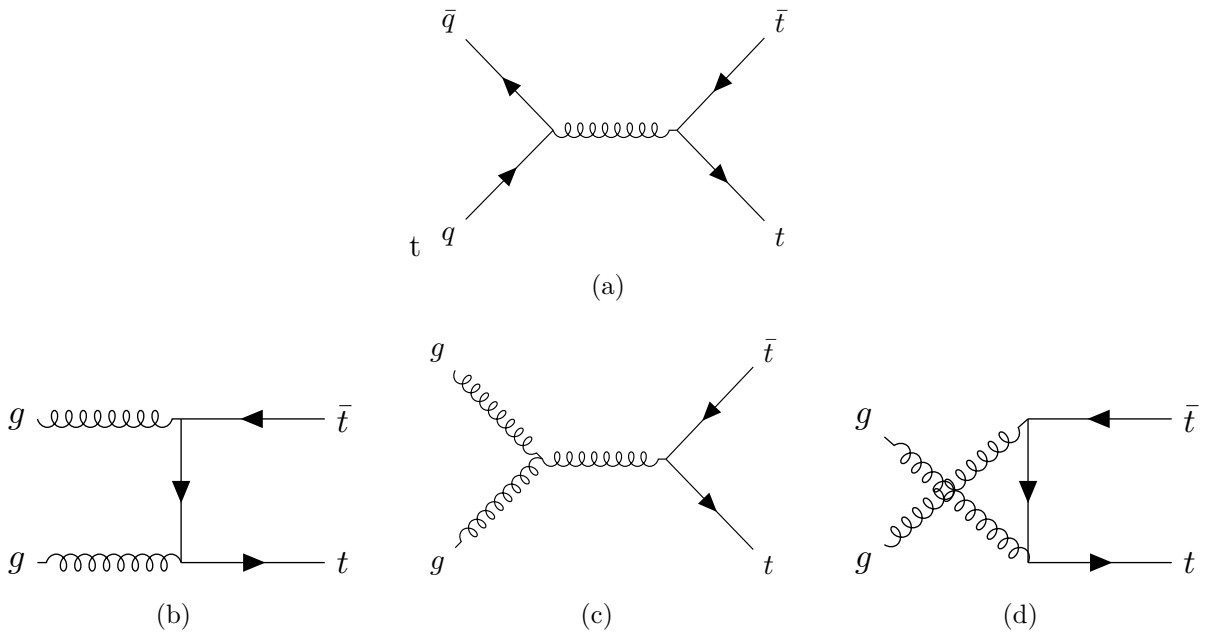

Figure 2.2.: Leading order Feynman diagrams for top quark pair production in (a) quark-antiquark annihilation and (b), (c) & (d) gluon fusion (ggF).

Along with the highest mass, the top quark also has the strongest Yukawa coupling to the Higgs boson, making it interesting for studies of the Higgs mechanism and electroweak symmetry breaking.
Due to its large mass, the top quark also has the shortest lifetime of all quarks, about 5×10^{-25} s [1], which is shorter than the typical hadronisation timescale of 3×10^{-24} s [25]. Therefore, the top is the only quark that decays before it can hadronise, allowing for direct studies of its properties through its decay products.
Notably, the top quark also decays before its spin can decorrelate, preserving spin information in the angular distributions of its decay products [26]. This unique feature enables precise measurements of top quark spin correlations and polarisation. At hadron colliders like the LHC, top quarks are predominantly produced in pairs via the strong interaction. The leading order (LO) Feynman diagrams for top quark pair production are shown in Fig. 2.2. At the LHC, the dominant production mechanism is gluon fusion (ggF). This is due to the high gluon density in the proton at the relevant Bjorken-x values for top quark production.
The CKM matrix element V_{tb} is close to unity [1], leading to a nearly exclusive decay of the top quark into a W boson and a b quark. The W can subsequently decay either leptonically into a charged lepton and a neutrino, or hadronically into a pair of quarks, with branching ratios of about $\mathrm{BR}(W \to \ell\nu) = 1/3$ and $\mathrm{BR}(W \to q\bar{q}) = 2/3$, respectively [1].
Based on these W decay modes, top quark pair events are categorised into three channels. In the **dileptonic channel**, both W bosons decay leptonically ($W^+ \to \ell^+\nu_\ell$, $W^- \to \ell^-\bar{\nu}_\ell$), resulting in two charged leptons, two neutrinos and two b quarks in the final state. This channel has a branching ratio of approximately 1/9. The **semileptonic channel**

occurs when one W boson decays leptonically and the other hadronically ($W \to q\bar{q}'$), yielding one charged lepton, one neutrino, four quarks (including two b quarks) in the final state. This is the most common channel with a branching ratio of about 4/9. Finally, in the **all-hadronic channel**, both W bosons decay hadronically, producing six quarks (including two b quarks) in the final state. This channel has the highest branching ratio of approximately 4/9 but suffers from large multijet background contamination, especially in the busy environment of hadron colliders.

2.2.1. Dileptonic Decay Channel

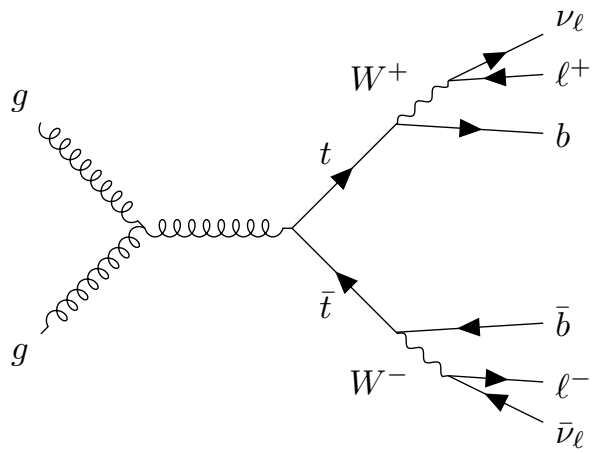

Figure 2.3.: Leading order Feynman diagram for top quark pair production with dileptonic decay.

The dileptonic decay channel, where both W bosons decay leptonically, has a branching ratio of about 1/9 and results in a final state with two charged leptons, two neutrinos and two b quarks, as illustrated in Fig. 2.2.1. Due to the lower branching ratio, the dileptonic channel has a smaller event yield compared to the other channels. However, it offers a cleaner signature with reduced background contamination, making it particularly suitable for precision measurements of top quark properties. The presence of two neutrinos in the final state leads to missing transverse energy ($\not{E}_{\mathrm{T}}$) in the event, complicating the full reconstruction of the top quark kinematics. Advanced techniques, such as kinematic fitting and multivariate analyses, are often employed to address these challenges and extract maximum information from the dileptonic events.

Kinematic Constraints

Assuming both W bosons and top quarks are on-shell, the following kinematic constraints can be applied to the dileptonic decay channel

$$(p_{\ell_1} + p_{\nu_1})^2 = m_W^2, \tag{2.3}$$
$$(p_{\ell_2} + p_{\nu_2})^2 = m_W^2, \tag{2.4}$$
$$(p_{b_1} + p_{\ell_1} + p_{\nu_1})^2 = m_t^2, \tag{2.5}$$
$$(p_{b_2} + p_{\ell_2} + p_{\nu_2})^2 = m_t^2, \tag{2.6}$$
$$(\vec{p}_{\nu_1} + \vec{p}_{\nu_2})_x = \not{E}_{\mathrm{T}x}, \tag{2.7}$$
$$(\vec{p}_{\nu_1} + \vec{p}_{\nu_2})_y = \not{E}_{\mathrm{T}y}. \tag{2.8}$$

If one assumes, the neutrinos to be massless, these six equations provide constraints on the six unknown components of the two neutrino momenta. Therefore, additional assumptions or techniques are required to fully reconstruct the event kinematics in the dileptonic channel. Due to the algebraic nature of the constraints, multiple solutions can exist for a given event, leading to ambiguities in the reconstruction [27, 28]. Various methods have been developed to address these challenges, including likelihood-based approaches [29], matrix element methods, and machine learning techniques [30], which aim to optimally utilise the available information and improve the accuracy of the top quark kinematic reconstruction in dileptonic events.

2.2.2. Spin Correlations in Top Quark Pairs

As before elaborated, the top quark decays before it can hadronise or its spin becomes decorrelate. Therefore, the spin information of the top quark is directly transferred to its decay products. In top quark pair production, the spins of the top and antitop quarks are correlated due to the production mechanism. These spin correlations manifest in the angular distributions of the decay products, providing a unique opportunity to study the spin dynamics of the top quark.

Recent studies at the LHC began to explore quantum entanglement in top quark pairs [31]. Entanglement generally describes a quantum mechanical phenomenon where the quantum states of two or more particles become correlated in such a way that the state of one particle cannot be described independently of the state of the other(s), even when the particles are separated by large distances [32]. This property is expressed as the system having a non-seperable density matrix. While a general description of this phenomenon is beyond the scope of this thesis, entanglement can be tested using specific criterions. One such criterion is the *Peres-Horodecki criterion* [33, 34], which states that if the partial transpose of the density matrix of a bipartite system has at least one negative eigenvalue, then the system is entangled. Considering top quark pairs as a bipartite system of two spin-$\frac{1}{2}$ particles (qubits), this criterion can be applied to test for entanglement in their spin states.

In Ref. [35] it was proposed to measure entanglement in top quark pairs produced at the LHC by analysing the angular distributions of their decay products, specifically the charged leptons in the dileptonic decay channel. The study showed that by measuring

the opening angle between the two leptons, one can construct an observable sensitive to the entanglement of the top quark spins. Due to the kinematics of the top quark pair production, the entanglement is expected to be more pronounced in certain regions of phase space, particularly at high invariant masses of the top quark pair system. One particular region of interest is the so-called *threshold region*, where the top quark pair is produced with low relative velocity. In this region, the top quark is produced as a spin singlet state in about $\sim 80\%$ of the cases, leading to a strong entanglement signature [36]. The ATLAS collaboration has recently reported the first experimental evidence of quantum correlations consistent with entanglement in top quark pairs [31].

2.2.3. Top Quark Pair Bound State Effects

Even though the top quark decays before it can hadronise, near the production threshold of top quark pairs, the strong interaction between the top and antitop quarks can lead to the formation of transient bound states, coined as toponium [36–38]. These bound state effects can influence the production cross section and kinematic distributions of top quark pairs near threshold. The CMS collaboration has recently observed an excess in the invariant mass distribution of top quark pairs near threshold, consistent with the presence of toponium bound state effects [39]. The Modelling of these effects is based on non-relativistic QCD (NRQCD) calculations, which account for the strong interaction dynamics between the top and antitop quarks in the near-threshold region [40, 41]. Understanding and accurately modelling these bound state effects is crucial for precision measurements of top quark properties and for searches for new physics in top quark pair production.
This study is also investigating the threshold region of top quark pair production and makes use of angular correlations between the decay products as a probe of the underlying dynamics, including potential bound state effects.

3. Machine Learning in High Energy Physics

Since this work focuses on improving the event reconstruction of dileptonic $t\bar{t}$ decays using machine learning techniques, this chapter provides an overview of the fundamental concepts and methodologies employed in machine learning.

3.1. Supervised Learning

Machine learning can be broadly categorized into supervised and unsupervised learning. In supervised learning, the model is trained on a labeled dataset, where each input data point is associated with a corresponding target output. The goal of the model is to learn a mapping from inputs to outputs, enabling it to make accurate predictions on unseen data.

3.1.1. Monte Carlo Training Data

In high energy physics, supervised learning is commonly performed by training mdoels on simulated datasets. The way the Monte Carlo Event modelling is performed in particle physics, makes it naturally suited for supervised learning tasks.
Events are usually simulated using a cascade of different programs, each simulating a different state of the event. First the hard scattering process is simulated using matrix element generators like MadGraph [42] or Powheg [43]. These programs calculate the probabilities of different particle interactions based on the underlying physics theories, such as the Standard Model. Using these probabilities, they generate events that represent the initial state of the particles after the collision. The possible kinematic phase space is sampled according to these probabilities, resulting in a set of particles with specific momenta and energies. This type of event variables is often referred to as *parton-level*.
Next, the parton showering and hadronization processes are simulated using programs like Pythia [19]. Parton showering models the emission of additional particles from the initial partons, while hadronization simulates the formation of hadrons from quarks and gluons. These processes are crucial for accurately modelling the final state particles observed in detectors.
Finally, detector simulation programs like Geant4 [44] are used to simulate the interaction of particles with the detector material, producing realistic detector responses. This includes simulating the energy deposits in calorimeters, hits in tracking detectors, and other relevant signals.
While the actual detector response is simulated using complex detector simulation soft-

ware, for many machine learning applications, a simplified representation of the detector response is sufficient. This can involve smearing the particle momenta and energies according to the detector resolution, applying efficiency corrections, and simulating the effects of pile-up.
This is because, for the reconstruction of the physics objects, such as jets, leptons, and missing transverse energy, highly sophisticated algorithms are used that already take into account the detector effects (Note, that these algortihms may also be machine learning based,). Therefore, the input features for machine learning models can often be derived directly from these reconstructed objects, rather than relying on the raw detector signals. The reconstructed object event variables are called *reco-level*.

3.1.2. Training

During the training phase, the model is presented with a set of input features derived from the reco-level event variables, along with their corresponding target outputs, which are typically derived from the parton-level event variables. The model learns to map the input features to the target outputs by minimizing a loss function that quantifies the difference between the predicted and true values. Common loss functions include mean squared error for regression tasks and cross-entropy [45] loss for classification tasks.

A. Additional Plots

Bibliography

[1] S. Navas et al. (Particle Data Group Collaboration), *Review of Particle Physics*, Phys. Rev. D **110**, 030001 (2024)

[2] S. L. Glashow, *Partial Symmetries of Weak Interactions*, Nucl. Phys. **22**, 579 (1961)

[3] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19**, 1264 (1967)

[4] A. Salam, *Weak and Electromagnetic Interactions*, Almqvist & Wiksell, Stockholm, nobel symposium 8 edition (1968)

[5] D. J. Gross et al., *Asymptotically Free Gauge Theories*, Phys. Rev. D **8**, 3633 (1973)

[6] H. D. Politzer, *Reliable Perturbative Results for Strong Interactions*, Phys. Rev. Lett. **30**, 1346 (1973)

[7] G. 't Hooft, *Renormalizable Lagrangians For Massive Yang-Mills Fields*, Nucl. Phys. B **35**, 167 (1971)

[8] P. W. Higgs, *Broken Symmetries, Massless Particles and Gauge Fields*, Phys. Lett. **12**, 132 (1964)

[9] F. Englert et al., *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13**, 321 (1964)

[10] G. S. Guralnik et al., *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13**, 585 (1964)

[11] T. D. Lee et al., *Question of Parity Conservation in Weak Interactions*, Phys. Rev. **104**, 254 (1956)

[12] C. S. Wu et al., *Experimental Test of Parity Conservation in Beta Decay*, Phys. Rev. **105**, 1413 (1957)

[13] N. Cabbibo, *Unitary Symmetry and Leptonic Decays*, Phys. Rev. Lett. **10**, 531 (1963)

[14] M. Kobayashi et al., *CP Violation in the Renormalizable Theory of Weak Interaction*, Prog. Theor. Phys. **49**, 652 (1973)

[15] H. D. Politzer, *Asymptotic Freedom: An Approach to Strong Interactions*, Phys. Rept. **14**, 129 (1974)

[16] J. C. Collins et al., *Heavy Particle Production in High-Energy Hadron Collisions*, Nucl. Phys. B **263**, 37 (1986)

[17] T. Sjöstrand, *A model for initial state parton showers*, Physics Letters B **157(4)**, 321 (1985)

[18] G. Marchesini et al., *Simulation of QCD jets including soft gluon interference*, Nuclear Physics B **238(1)**, 1 (1984)

[19] T. Sjöstrand et al., *PYTHIA*, Comput. Phys. Commun. **135**, 238 (2001)

[20] B. Andersson et al., *Parton fragmentation and string dynamics*, Physics Reports **97(2)**, 31 (1983)

[21] T. Sjöstrand, *Jet fragmentation of multiparton configurations in a string framework*, Nuclear Physics B **248(2)**, 469 (1984)

[22] J. J. Aubert et al., *Experimental Observation of a Heavy Particle J*, Phys. Rev. Lett. **33**, 1404 (1974)

[23] F. Abe et al. (CDF), *Observation of Top Quark Production in $p\bar{p}$ Collisions with the Collider Detector at Fermilab*, Phys. Rev. Lett. **74**, 2626 (1995)

[24] S. Abachi et al. (DØ), *Observation of the Top Quark*, Phys. Rev. Lett. **74**, 2632 (1995)

[25] I. Bigi et al., *Production and decay properties of ultra-heavy quarks*, Physics Letters B **181(1)**, 157 (1986)

[26] G. Mahlon et al., *Spin correlation effects in top quark pair production at the LHC*, Phys. Rev. D **81**, 074024 (2010)

[27] L. Sonnenschein, *Analytical solution of $t\bar{t}$ dilepton equations*, Phys. Rev. D **73**, 054015 (2006)

[28] V. Abazov et al., *Measurement of the top quark mass in the dilepton channel*, Physics Letters B **655(1)**, 7 (2007)

[29] V. M. Abazov et al. (The D0 Collaboration), *Measurement of the top quark mass in final states with two leptons*, Phys. Rev. D **80**, 092006 (2009)

[30] J. A. Raine et al., *Fast and improved neutrino reconstruction in multineutrino final states with conditional normalizing flows*, Phys. Rev. D **109**, 012005 (2024)

[31] *Observation of quantum entanglement with top quarks at the ATLAS detector*, Nature **633**, 542 (2024)

[32] E. Schrödinger, *Discussion of Probability Relations between Separated Systems*, Mathematical Proceedings of the Cambridge Philosophical Society **31(4)**, 555 (1935)

[33] A. Peres, *Separability Criterion for Density Matrices*, Phys. Rev. Lett. **77**, 1413 (1996)

[34] W. K. Wootters, *Entanglement of Formation of an Arbitrary State of Two Qubits*, Phys. Rev. Lett. **80**, 2245 (1998)

[35] Y. Afik et al., *Entanglement and quantum tomography with top quarks at the LHC*, The European Physical Journal Plus **136(9)**, 907 (2021)

[36] Y. Kiyo et al., *Top-quark pair production near threshold at LHC*, Eur. Phys. J. C **60**, 375 (2009)

[37] Y. Sumino et al., *Bound-state effects on kinematical distributions of top quarks at hadron colliders*, Journal of High Energy Physics **2010(9)**, 34 (2010)

[38] K. Hagiwara et al., *Bound-state effects on top quark production at hadron colliders*, Physics Letters B **666(1)**, 71 (2008)

[39] *Observation of a pseudoscalar excess at the top quark pair production threshold*, Rep. Prog. Phys. **88**, 087801 (2025)

[40] B. Fuks et al., *Signatures of toponium formation in LHC run 2 data*, Phys. Rev. D **104**, 034023 (2021)

[41] B. Fuks et al., *Simulating toponium formation signals at the LHC*, The European Physical Journal C **85(2)**, 157 (2025)

[42] T. Stelzer et al., *Automatic Generation of Tree Level Helicity Amplitudes*, Comput. Phys. Commun. **81**, 337 (1994)

[43] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, J. High Energy Phys. **11**, 040 (2004)

[44] S. Agostinelli et al. (Geant4), *GEANT4: A Simulation Toolkit*, Nucl. Instrum. Methods A **506**, 250 (2003)

[45] C. E. Shannon, *A mathematical theory of communication*, The Bell System Technical Journal **27(3)**, 379 (1948)

Danksagung

Dank. . .

Erklärung nach §17(9) der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen:
Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe.
Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, im Rahmen einer nichtbestandenen Prüfung an dieser oder einer anderen Hochschule eingereicht wurde.

Göttingen, den 8. Dezember 2025

(Siemen Henning Aulich)