A Lorentz covariant model of string fragmentation

Jade Abidi

Student ID: 31461964

A thesis submitted for the degree of **Bachelor of Science (Honours)**

November 2025

School of Physics and Astronomy
Monash University

Supervisor: Peter Skands

Abstract

Monte Carlo event generators are extensively used to simulate high-energy particle-collision events. For analytically intractable aspects, they rely on phenomenological models. The so-called Lund model describes the non-perturbative hadronisation process as the fragmentation of a classical string with constant tension. Lorentz covariance then implies the self-similarity of this fragmentation process along the string. The current formulaion of the Lund model, used in the PYTHIA generator, violates this property in terms of both kinematic distributions and hadronic chemistry. We introduce an additional tunable parameter that lessens this violation, and propose a new model for string fragmentation that preserves Lorentz covariance.

Contents

Chapter 1

Introduction

The field of particle physics investigates the most fundamental particles and interactions in nature. It naturally evolved from nuclear and atomic physics in the early 20th century as technological and scientific knowledge allowed scientists to probe matter at higher energies and smaller length scales [1,2]. The physical theory underpinning particle physics developed out of quantum field theory, culminating in the Standard Model of particle physics which was formalised in the 1980s. The Standard Model unifies three of the four fundamental forces of nature (electromagnetism, the weak force, and the strong force) into a single theory, and predicted the existence of the Higgs boson well before its historic discovery at the Large Hadron Collider (LHC) in 2012 [3–5].

The interactions investigated in particle-physics experiments occur at high energies, with low probabilities. Particle colliders, such as the LHC, are designed specifically to enable such experiments by accelerating charged particles, usually electrons or protons (and their antiparticles), to speeds up to 99.99999% of the speed of light. Most modern particle colliders are circular, and use high voltages and strong magnetic fields to accelerate particles and keep them within a thin beamline. The particle beams are then made to collide millions of times every second at interaction points, around which bespoke detector systems collect data on the final state [1, 6, 7].

For comparison against this data, it is very useful to be able to generate ensembles of large numbers of events based on physical models and assumptions. Because the physics involved is so complex and often lacks an analytic solution, Monte Carlo event generators such as PYTHIA provide a way to sample the phase space efficiently [1, 8–10]. Powerful factorisation theorems allow for the various processes that occur at different energy scales to be modelled independently of each other, which greatly simplifies the problem [10, 11]. The focus of this project is the process of hadronisation, in which partons (quarks and gluons) form hadronic bound states. This occurs at momentum transfer scales below the confinement energy of $\Lambda_{\mathrm{QCD}} \sim 0.2\,\mathrm{GeV}$ [10].

Due to asymptotic freedom, the coupling constant of the strong force becomes very large for momentum transfers in this range [12, 13], and therefore the physics of hadronisation cannot be calculated perturbatively. Non-perturbative quantum chromodynamics has not yet been analytically solved, and numerical methods such as lattice QCD have limitations that render them inapplicable to hadronisation [14]. Instead, PYTHIA uses the Lund string model as a description of hadronisation, which is a phenomenological theory (and is not derivable from first principles) [8, 15].

In the Lund string model, the strong field between a quark and antiquark is modelled as a string with constant tension $\kappa \sim 0.9\,\mathrm{GeV}$ per femtometre, as measured in lattice

QCD simulations [10, 16]. At separation distances of $\gtrsim 1\,\text{fm}$, there is enough potential energy stored in the string for it to "break", forming a new quark-antiquark pair. This leaves two quark-antiquark string subsystems, which proceed to break in the same manner as the original system, giving rise to a recursive self-similar process resulting in a set of outgoing bound states of quarks [8, 10, 15, 17].

This self-similarity of the fragmentation process along the string is one of the most fundamental properties of the Lund string model. The string breaks are all spacelike separated and hence cannot causally influence each other, and the tension along the string is constant and invariant under longitudinal boosts. It therefore follows that, away from the endpoints of the string, all fragmentation observables (like the number or species of hadrons produced from a region of the string) must be invariant under Lorentz boosts (or equivalently, rapidity translations) along the string axis. This means that distributions such as hadron density per unit rapidity should be flat except for endpoint effects (denoted the "rapidity plateau"). It also implies that, if considered iteratively, the process by which each string subsystem is fragmented must be scale invariant [8, 15, 17].

((TODO: Rewrite this and next paragraph to be more clear about what finalTwo is, what it does, and its importance)). Despite the importance of the Lorentz covariance of string fragmentation in the Lund model, current simulations in PYTHIA reveal that this property is quite badly broken. Since the release of PYTHIA 8.0, the rapidty plateau has exhibited a significant dip in the central region. Initial investigations revealed this to be a consequence of the hadronisation algorithm used in PYTHIA, where the final two hadrons have kinematics forced by energy-momentum conservation and the mass-shell relation. In the Monash tune of PYTHIA 8.3, the rapidity spacing between these final two hadrons is larger than the typical rapidity spacing. This, combined with the non-uniformity of location of the joining step along the string, gives rise to this central rapidity dip. Furthermore, the `finalTwo` procedure that generates these hadrons fails about 50% of the time, introducing bias that skews the species makeup of these final two hadrons.

PYTHIA offers three tunable parameters that adjust the behaviour of this procedure — `stopMass`, `stopNewFlav`, and `stopSmear`. Prior to this project, it was believed that these parameters could be tuned to give a flat rapidity plateau, and would not need to be retuned when other fragmentation parameters change [8, 18]. This is not the case for the Monash tune, where it is impossible to achieve a flat plateau with these parameters without significantly worsening the failure rate of `finalTwo`.

In this thesis, chapter 2 provides a review of the theoretical background of quantum chromodynamics, PYTHIA, and the Lund string model. In chapter 3, the current behaviour of the PYTHIA hadronisation algorithm is analysed in more detail, and the various problems are described and explained. In chapter 4, the tuning of the `finalTwo` parameters is discussed. An additional tunable parameter `revertFinalBreak` is presented which provides an additional dimension of parameter space and allows for a flatter rapidity plateau to be obtained without as much of an impact on the failure rate. Finally, in chapter 5, an alternative algorithm for hadronisation in PYTHIA is developed, called the accordion join, which achieves an approximately flat rapidity plateau for any tune, and fails more than a hundred times less frequently.

Chapter 2

QCD, the Lund Model, and PYTHIA

2.1 Quantum Chromodynamics and Collider Physics

Before introducing PYTHIA and the Lund string model, we first provide an overview of the theoretical models and experimental techniques used in particle physics, with a focus on quantum chromodynamics.

The field of particle physics developed out of the study of atomic and nuclear physics in the early 20th century. Experiments like the discovery of the nucleus in 1911 [19] or the neutron in 1932 [20] paved the way for the development of more advanced technologies such as particle accelerators and colliders. The high centre-of-mass energies reached in these experiments allowed for the discovery of a slew of particles in the 1950s and 1960s, dubbed the "particle zoo" [2]. Alongside these experiments, the theoretical success of quantum mechanics inspired the quantisation of the electromagnetic field in the original formulation of quantum field theory (QFT) and quantum electrodynamics (QED) by Dirac, who notably predicted the existence of antimatter a number of years before its discovery [21,22]. QED was later unified with a description of the weak force (responsible for phenomena like beta decay) into electroweak theory, a Yang-Mills gauge theory consisting of a local $\mathrm{U}(1) \times \mathrm{SU}(2)$ gauge symmetry [1, 23].

At the same time, quantum chromodynamics (QCD) emerged, describing the variety of particles discovered in the "particle zoo" as composed of more fundamental subatomic particles — quarks and gluons. Initial models like the Eightfold Way classified these various strongly interacting particles as hadrons emerging from a SU(3) flavour symmetry between three quark flavours [24, 25], as shown in Figure 2.1. These flavours were later denoted the up (u), down (d), and strange (s) quarks, and three more (charm, beauty, and top) were eventually found. Hadrons were classified into two groups — mesons, which are bosonic bound states consisting of a quark and an antiquark, and baryons, which are fermions consisting of three quarks or antiquarks. The experimental reality of these quarks outside of phenomenology was confirmed by deep inelastic scattering experiments, which showed pointlike constitutents within the proton, called partons [6, 23].

It was later found that these partons come in two types — the spin $\frac{1}{2}$ quark, and the spin 1 gluon [1]. The existence of the gluon was confirmed in the JADE experiment [26], which was notably one of the first uses of Monte Carlo event generators in the analysis of experimental data. In addition to their electromagnetic charge, quarks and gluons also carry colour charge, which has three components labelled red, green, and blue (with cyan,

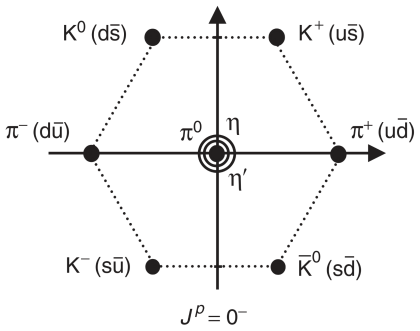


Figure 2.1: A visual depiction of the pseudoscalar mesons in the Eightfold Way, which form an octet and a singlet in SU(3) flavour space [24, 25]. The horizontal axis represents I_3, an isospin component, and the vertical axis represents Y, the hypercharge, both of which are quantum numbers arising from the SU(3) flavour symmetry. Figure taken from [23].

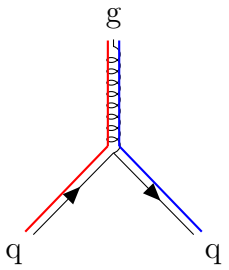


Figure 2.2: An illustration of a qqg vertex in QCD, showing the colour flow. The incoming and outgoing quarks have one colour, while the gluon has two (a colour and anticolour).

magenta, and yellow as their anticolour opposites) [1, 23].

In the 1970s, a formal theory of QCD based on a SU(3) gauge symmetry was developed, and by 1980, the Standard Model was formalised, which unified the electromagnetic, weak, and strong forces into a single unified gauge theory with three generations of fermions (6 quarks and 6 leptons). Collider experiments throughout the following years continued to confirm predictions of the Standard Model, and it has since become our most successful and fundamental model of physical reality [3, 23].

A full theoretical description of QFT and QCD is outside the scope of this review, and we will only provide a brief summary in order to establish the necessary theoretical background. The reader is directed to the textbook by Peskin and Schroeder [27] for a general overview of quantum field theory, or the textbook by Ellis, Stirling, and Webber [6] for closer detail on quantum chromodynamics or collider physics.

Quantum field theories model particles as excitations of underlying quantised fields, which are operator-valued at every point in space and time. Particles interact via the exchange of bosons as virtual particles, and in this way the concept of a "force" is reduced to an allowed interaction between particles. As in quantum mechanics, these particles can be more or less localised in coordinate or momentum space, with these two uncertainties linked by Heisenberg's relation

$$\Delta x \Delta p_x \geq \frac{\hbar}{2}. \tag{2.1}$$

Typically, in particle physics, particles are described as entirely delocalised plane waves with precisely defined energies and momenta. This is justified because the length scales involved in particle collisions are much smaller than typical De Broglie wavelengths, and in this limit the physical extent of the particle's wave packet is not relevant.

The physical laws governing these quantised fields are captured in the Lagrangian, and can be derived via the Euler-Lagrange equations. In a gauge theory, these Lagrangians are derived to obey underlying local gauge symmetries. QCD is based on a local SU(3) gauge symmetry, which implies that the wavefunction must carry three additional degrees of freedom representing the colour charge. The three fundamental colour states

corresponding to red, green, and blue are

$$r = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad g = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{2.2}$$

A local SU(3) gauge transformation is a rotation of states in colour space, where the axis and amount of rotation can vary throughout time and space. The Noether invariant corresponding to this symmetry is colour charge, which is a conseved quantity in QCD.

In QCD, the strong interaction is mediated by the eight gluons, corresponding to the eight generators of the SU(3) group. Unlike quarks, gluons carry a combination of colour charge and anticharge. Since quarks are the only fermions that carry colour charge in the Standard Model, only quarks can couple to gluons. The Lagrangian density describing QCD is given by

$$\mathcal{L} = \bar{\psi}_q^i (i\gamma^\mu)(D_\mu)_{ij} \psi_q^j - m_q \bar{\psi}_q^i \psi_{qi} - \frac{1}{4} F_{\mu\nu}^\alpha F^{\alpha\mu\nu}. \tag{2.3}$$

Here, the first term describes the kinetic energy of quarks and their interaction with gluons, the second term endows quarks with their mass, and the third term allows for gluon-gluon interactions. Figure 2.2 shows how colour flows through a quark-gluon Feynman interaction vertex. Because the generators of SU(3) do not commute, QCD is a non-Abelian gauge theory and therefore gluon-gluon interations are also possible. This is in contrast to quantum electrodynamics, where there is no photon-photon coupling term in the Lagrangian [1, 6, 23].

An important property of QCD is confinement. The quarks and gluons that make up hadrons have never been observed on their own as free particles. In nature, they seem to always be confined to hadronic bound states, and cannot be separated beyond the typical hadron size of $\sim 1\,\text{fm}$. Confinement has not been mathematically proven—indeed, its proof is one of the Millenium Prize Problems (a consequence of the Yang-Mills existence and mass gap). However, it can still be understood as a consequence of the running coupling of QCD. Figure 2.3 shows how the coupling constant (describing the strength of the strong interaction) decreases with the energy scale of the interaction [1]. As such, at short distances (and high energies) the strong interaction is weak, but at longer distance the strong interaction becomes strong enough to forbid the separation of quarks and gluons. This phenomenon is known as asymptotic freedom [12, 13].

The strong potential between quarks and antiquarks is given by the Cornell potential,

$$V(r) = -\frac{4}{3}\frac{\alpha_s}{r} + \kappa r, \tag{2.4}$$

which combines a short-distance Coloumb potential with a stronger, longer-distance linear potential. Here, α_s is the strong coupling constant, and κ is a constant measured to be approximately $1\,\text{GeV}\,\text{fm}^{-1}$. Figure 2.4 shows lattice QCD calculations of the strong potential between a $q\bar{q}$ pair, exhibiting Coulombic behaviour at short distances and becoming linear for longer distances. Because QCD allows for gluon-gluon interactions, the linear potential can be understood as a result of the self-attraction of strong field lines, which compresses the field into a flux tube with constant tension. It is this linear potential that gives rise to the confinement distance of approximately 1 fm, at which it is energetically favourable for the strong field to break into a new quark-antiquark pair, forming a new hadron. This "string-breaking" model of hadron formation is the basis for the Lund string model, described further in section 2.3.

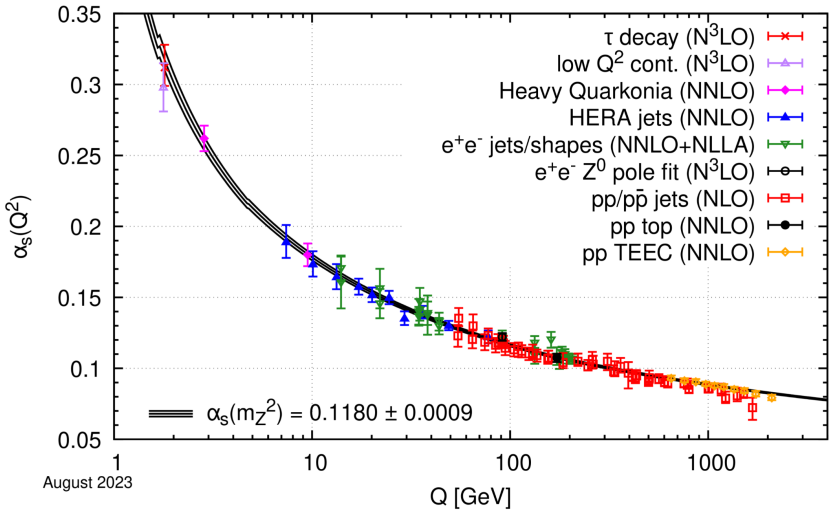


Figure 2.3: The running of the QCD coupling constant $\alpha_S(Q^2)$ with respect to the energy scale Q. The PDG average is shown as a solid line alongside data points from numerous experiments. The asymptotic freedom of QCD is evident from how the coupling constant is sufficiently small for perturbation theory at large values of Q, but becomes prohibitively large for smaller values of Q. Figure taken from [1].

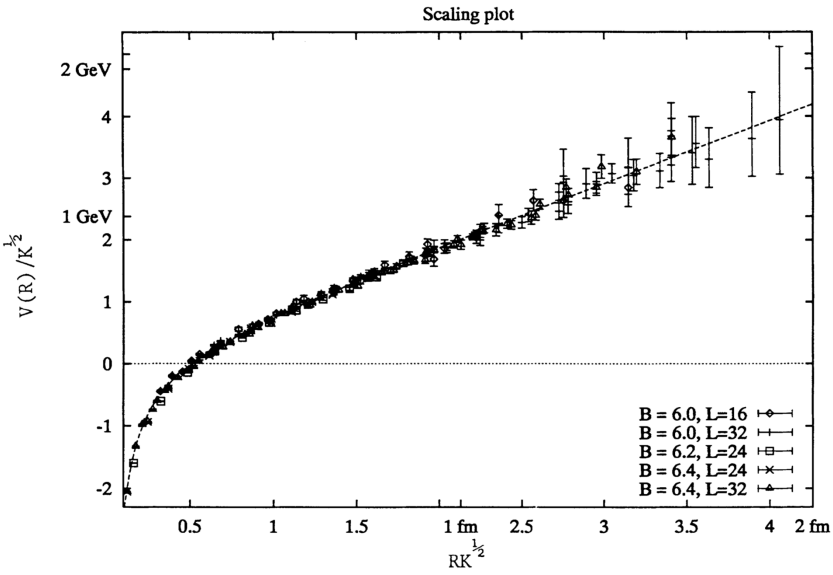


Figure 2.4: The potential between a quark and antiquark as a function of distance. Data points are obtained from lattice QCD simulations [16], and the dotted line shows the Cornell potential as defined in equation (2.4). Except for at short distances, the potential is approximately linear, motivating the treatment of the colour field as a classical string with constant tension. Figure taken from [16].

As mentioned in the introduction, experiments in particle physics require very high centre-of-mass energies for the interactions under observation to be kinematically allowed. If two particle beams have energies E_1 and E_2, then the resulting relativistic centre-of-mass (CM) energy is given by

$$E_{\text{CM}} \approx 2\sqrt{E_1 E_2}, \tag{2.5}$$

which reduces to $E_{\text{CM}} = 2E_b$ if the two beam energies are the same, which is common in modern colliders. The Large Hadron Collider has achieved CM energies up to 13.6 TeV.

Another important property of particle colliders is luminosity. While the beam and centre-of-msas energies describe the energies achieved by individual particles in the collider, the luminosity quantifies the flux of particles in the beam. Luminosities in modern colliders are quite large; the most recent LHC experiments have luminosities around $5 \times 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$.

The "probability" of a given interaction occuring is quantified by the cross section σ. The cross section has units of area, but typical cross sections are so small that the standard unit is barns, where 1 barn is equal to $1 \times 10^{-28}\,\text{m}^{-2}$. Despite not actually representing a physical area, the cross section generalises the notion of the area of a target representing its probability of being struck by an incoming particle. The variation of the interaction probability over the values of a variable, such as the solid angle Ω, is often expressed using the differential cross section $d\sigma/d\Omega$, which is related to the overall cross section by

$$\sigma = \int \frac{d\sigma}{d\Omega}\, d\Omega. \tag{2.6}$$

Here, Ω can be replaced with any number of variables, such as the Mandelstam variable s which quantifies the CM energy of the collision. If $\mathcal{L}(t)$ is the instantaneous luminosity of a particle beam, then the expected number of events with cross section σ is

$$N = \sigma \int \mathcal{L}(t)\, dt. \tag{2.7}$$

In this way, the cross section solely represents the underlying physics of an interaction, with the specific experimental collider setup factorised out.

The most important particle colliders for experimental QCD are typically proton-proton (pp) colliders, also called hadron colliders. The first hadron collider was the Intersecting Storage Rings (ISR) at CERN, which were operational from 1971, with a centre-of-mass energy of 62 GeV [1, 28, 29]. It was succeeded by the Tevatron which, as suggested by its name, reached centre-of-mass energies of 1 TeV [30]. Notable discoveries such as that of the top quark were made by the Tevatron [31, 32]. Currently, the largest hadron collider is the creatively named Large Hadron Collider (LHC) described above, which was operational from 2008 and is responsible for the monumental discovery of the Higgs boson in 2012 [4, 5, 7]. A photo of the Compact Muon Solenoid (CMS) detector, one of the particle detectors in the LHC, is shown in Figure 2.5.

Detectors are only able to measure the final state of an event, and ultimately output a list of the types, masses, and four-momenta of the outgoing particles. For high-luminosity colliders like the LHC, many billions of these event records are recorded. For many theoretical aspects of QCD, it is only possible to compare experimental data to theory by using Monte Carlo techniques to generate ensembles of events based on theoretical assumptions. As mentioned earlier, the discovery of the gluon in the JADE experiment utilised the JETSET generator. The next section outlines the functionality of these event generators, with a focus on PYTHIA, the modern successor to JETSET.

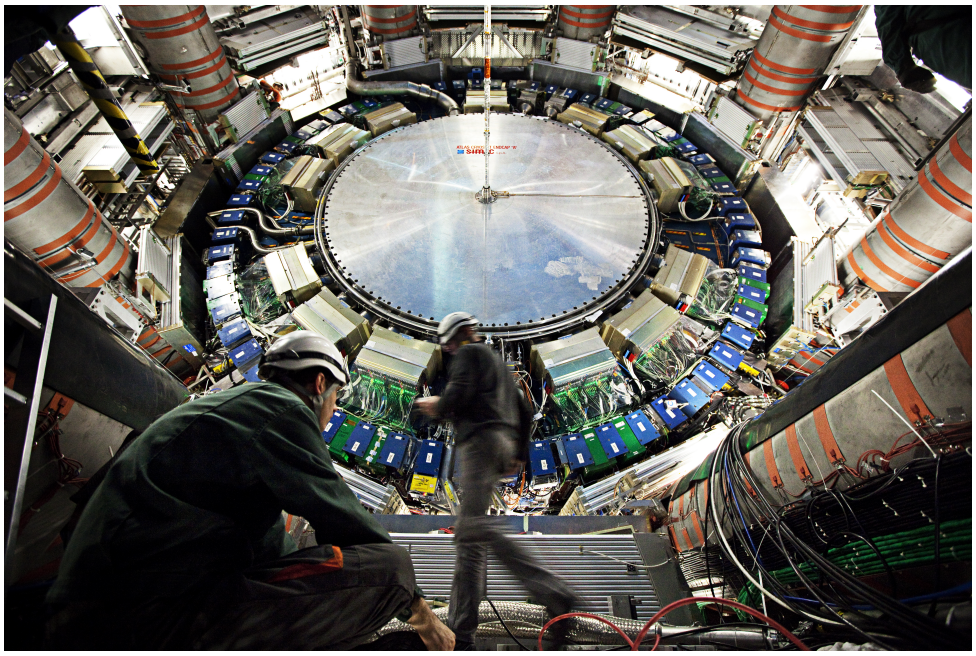


Figure 2.5: The ATLAS detector, the largest of the particle detectors at the LHC. It was involved in the discovery of the Higgs boson in 2012 [4]. The particle beams are contained within the thin tube visible at the top centre of the image, and the rest of the structure contains detectors that measure the energy and momentum of final state particles. Figure from [33].

2.2 Monte Carlo Event Generators and PYTHIA

Essential to any scientific undertaking is the ability to compare the predictions of a theoretical model with the results of an actual experiment. In many fields of physics, making a prediction of experimental results is a simple matter of a mathematical derivation or computation. Even within particle physics, discoveries such as that of the Higgs boson often have clear experimental signatures, such as resonances in distributions of particles over observables like invariant mass [4, 5], and these signatures can be identified with model fitting or machine learning to extract the signal from the background. On the other hand, some experimental predictions are very far removed from the underlying theory. For example, in the JADE experiment, the existence of the gluon was inferred not from any resonant peak, but rather from a comparison of the jet mass distribution with the results of detailed simulations of particle collisions with and without gluons. Similarly, the discovery of the top quark required use of simulations to determine the expected signal and background of various distributions [31, 32].

Simulating a particle collision (often referred to as an "event") is not a simple task, for a variety of reasons. QCD has the property of asymptotic freedom, meaning the coupling constant at lower energies is too large for perturbation theory to apply [12, 13]. Non-perturbative QCD has not been analytically solved [1], and while techniques like lattice QCD exist to obtain numerical solutions, they are not suitable for event generation for a number of reasons. The energy scales involved in particle collisions span many orders of magnitudes beyond the energy ranges possible in lattice QCD simulations. Furthermore, the computational cost of lattice QCD is very high, and not appropriate for a situation where the number of required events is in the millions, or even billions. Lattice QCD also uses Euclidean spacetime, which is a very poor approximation when relativistic effects are prominent [14].

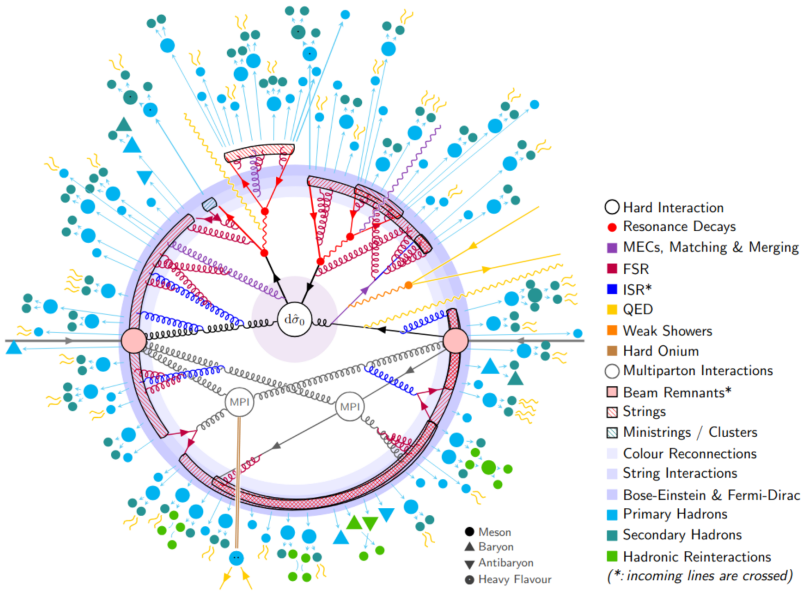


Figure 2.6: A visual representation of a $pp \to t\bar{t}$ event as generated in the PYTHIA MC event generator. The various processes involved are illustrated here with the hardest processes in the centre and the softest processes on the outer edge. Note in particular how the resultant partons from the parton shower are combined into strings, which go on to fragment into sets of outgoing hadrons (which in turn decay and emit further bremsstrahlung). Figure taken from [8].

Because of the implausability of using an analytic or numerical solution to generate events, event generators usually use Monte Carlo (MC) techniques to sample the phase space of possible events. Such event generators exploit the fact that many aspects of particle collisions are described by random distributions. By randomly sampling from these distributions, MC event generators can generate ensembles of simulated events that closely approximate real-world events [1, 9, 10]. One of the first MC event generators to be developed was JETSET, which was based on the Field-Feynman model of hadronisation [34, 35] and played an essential role in the aformentioned discovery of the gluon [26]. In 1996, JETSET was merged into PYTHIA, which uses the Lund string model for hadronisation [8, 34]. Other commonly used event generators include HERWIG [36] and SHERPA [37], which both use the cluster model of hadronisation.

In this thesis, we focus on the PYTHIA event generator. Figure 2.6 shows a visual representation of the various processes that take place in a single $pp \to t\bar{t}$ event in PYTHIA. Here, the radial coordinate represents the hardness (momentum transfer) scale — so the hardest processes (like the initial scattering) take place in the centre, while soft processes like particle decays take place on the outer edge [8].

The hardest process in a particle collision is the hard scattering between the incoming particles (or constitutent partons, in the case of pp collisions), which can be computed using perturbative techniques including Feynman diagrams and matrix elements. After this initial hard scattering, the outgoing particles will continue to emit gluon and photon ratiation, which can in turn evolve into more partons (quarks and gluons) in a self-similar process known as the "parton shower". This process continues until the energy scale is below the confinement scale of $\Lambda_{\text{QCD}} \sim 0.2\,\text{GeV}$, at which point colour confinement becomes relevant and the partons become bound within colour netural states [8, 9].

The primary focus of this thesis is the energy scales below the confinement scale. At this scale, the parton shower is finished, leaving a large amount (often hundreds)

of outgoing quarks and gluons. These quarks and gluons will go on to combine into bound states — mesons and baryons — in a process called hadronisation. Because the physics involved is non-perturbative, hadronisation is poorly understood, and the lack of an analytic solution necessitates the use of phenomenological models to describe the underlying physics [8, 10]. A number of such models have been developed, the most commonly used of which are the Lund [15, 17] and cluster models [1, 9].

Many of the hadrons produced in hadronisation are unstable and have short lifetimes, and will therefore go on to decay into stable states. This is the final, softest step of event generation, and results in a list of outgoing particles and their four-momenta as they are observed in detectors [8,9]. Integral to the functionality of event generators are the various factorisation theorems, which together imply that processes at different energy scales can be considered independently of one another [10, 11]. More specifically, the differential cross section for pp collisions with respect to an observable $\mathcal{O}$ can be written as

$$\frac{d\sigma}{d\mathcal{O}} = \sum_{i,j} \int_0^1 dx_i\, dx_j \sum_f \int d\Phi_f f_{i/h_1}(x_i, \mu_F^2) f_{j,h_2}(x_j, \mu_F^2) \frac{d\hat{\sigma}_{ij\to f}}{d\hat{\mathcal{O}}} D_f(\hat{\mathcal{O}} \to \mathcal{O}, \mu_F^2), \quad (2.8)$$

where $\hat{\mathcal{O}}$ denotes the observable as evaluated on the final partonic state (after the parton shower, and before hadronisation). Here, f_{i/h_1} and f_{j/h_2} denote the parton densities (which describe the probabilities of interactions involving the various partons inside the colliding protons), $d\hat{\sigma}_{ij\to f}$ is the partonic cross section (describing the parton shower), and D_f describes the fragmentation of partons into final-state hadrons [10]. Each of these different processes are factorised out and different models and approxmations can be used to describe each of them.

Because of the various approximations involved, MC generators involve dozens of free parameters which must be adjusted to fit experimental data. A set of such parameters is known as a "tune" [8, 9]. The JETSET generator, as well as earlier versions of PYTHIA, used the JETSET tune — the most up-to-date version of PYTHIA, PYTHIA 8.316, uses the Monash tune, which was developed at Monash University in 2014 [8, 38].

PYTHIA has performed very well across a wide range of use cases, and played an integral role in experiments like the discovery of the top quark [31, 32] or gluon [26]. However, there are a number of recent findings that indicate discrepancies between how PYTHIA models hadronisation and recent LHC findings. For example, ALICE results from 2017 show an increase in the proportion of strange hadrons produced as the charged particle multiplicity goes up in pp collisions [39]. More recently, there have also been discrepancies found in various jet parameters between PYTHIA simulations and data from the ATLAS experiment [40]. These results indicate that further improvements to the phenomenological models, approximations, and tuning of PYTHIA can be done, and research in this area has resulted in the development of numerous new models for pp collisions [41–43].

Having established the fundamentals of event generators and quantum chromodynamics, we can now proceed to a treatment of the Lund string model, the phenomenological model developed at Lund University in the 1980s that forms the basis of the PYTHIA generator.

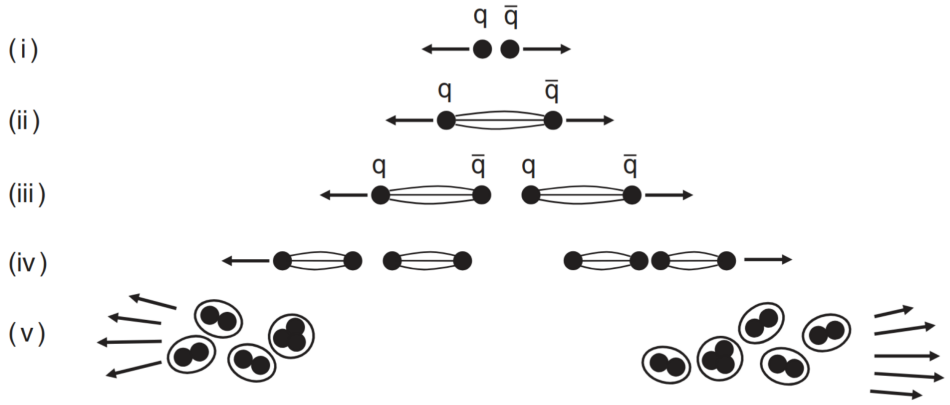


Figure 2.7: A schematic representation of hadronisation in the Lund string model. (i) A $q\bar{q}$ pair is created. (ii) As the q and $\bar{q}$ move away from each other, the strong field between them forms a string. (iii) The string breaks, forming a new $q\bar{q}$ pair and leaving two strings. (iv) These smaller strings proceed to break in the same way, creating further $q\bar{q}$ pairs. (v) The final state consists of jets of hadrons (bound quark-antiquark states) in either direction. Figure taken from [23].

2.3 The Lund String Model

The Lund string model is a phenomenological model of hadronisation which models the strong colour field between colour charged particles as a classical string with constant tension, and describes hadronisation in terms of the fragmentation of these strings. In this section, we provide an overview of the Lund model, with a focus on the underlying symmetry of Lorentz covariance. This review broadly follows the textbook by Andersson [17], as well as the primary review [15], and the reader is encouraged to consult these sources for a more detailed description.

The fundamental assumption of the Lund model is that when a colour charge and anticharge are separated by a distance r, the potential for large separations takes the asymptotic form of the Cornell potential (2.4),

$$V(r) = \kappa r, \tag{2.9}$$

where $\kappa \sim 1\,\mathrm{GeV\,fm^{-1}}$ is the string tension [16]. Because of the gluon-gluon interaction in QCD, this strong field is compressed into a thin flux tube, justifying the approximation of a classical string with constant tension.

At a high level, the process of hadronisation in the Lund model is illustrated in Figure 2.7. First, a $q\bar{q}$ pair is created in a particle collision, and a string forms between them. At a sufficient separation distance, the string has enough energy to break, forming a new $q\bar{q}$ pair. There are now two $q\bar{q}$ pairs, each separated by a string, each of which will continue to fragment in a self-similar manner until there is no longer enough energy for further string breaks to occur. At this point, the final state consists of a number of outgoing string pieces in the two jet directions, which are the outgoing hadrons.

In the Lund model, mesons are described as "yo-yo modes", which consist of a quark and an antiquark connected by a string and oscillating back and forth, as depicted in Figure **??**. Baryons, which consist of three quarks or three antiquarks, are also modelled as yo-yo modes where one of the string endpoints is a diquark (a grouping of two quarks or antiquarks).

Before describing the longitudinal momentum selection process, we must first introduce the longitudinal lightcone coordinates and rapidity variable. In position space, the

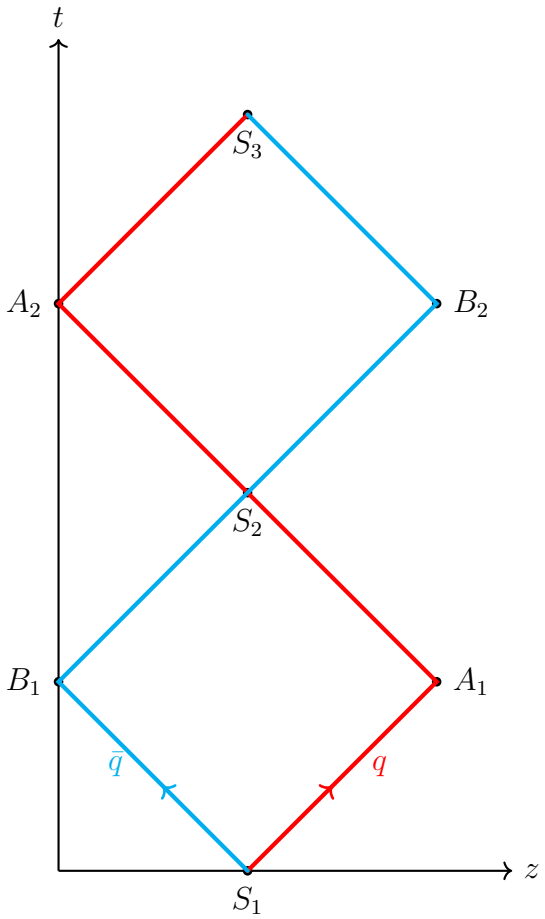


Figure 2.8: A spacetime diagram of the yo-yo mode of the Lund model.

lightcone coordinates are defined as

$$z^{\pm} = z \pm t, \tag{2.10}$$

where z^+ describes the position of a particle along the positive worldline and z^- describes the position of a particle along the negative worldline. In momentum space, lightcone coordinates are defined as

$$p_z^{\pm} = E \pm p_z, \tag{2.11}$$

where p_z is the momentum in the longitudinal direction. Finally, rapidity is defined (in the longitudinal direction) as

$$\begin{aligned} y &= \frac{1}{2}\ln\Big(\frac{1+v}{1-v}\Big) \\ &= \frac{1}{2}\ln\Big(\frac{E+p_z}{E-p_z}\Big) \\ &= \frac{1}{2}\ln\Big(\frac{p^+}{p^-}\Big), \end{aligned} \tag{2.12}$$

where v is velocity and $p^{\pm}$ are the lightcone momenta as defined in Equation (2.11). Rapidity can be thought of as a dimensionless transformation of velocity, scaled to range from $-\infty$ to $+\infty$ rather than $-c$ to $+c$. Energy and momentum can be expressed in terms of rapidity as

$$E = m_\perp \cosh(y) \tag{2.13}$$

and

$$p_z = m_\perp \sinh(y), \tag{2.14}$$

where $m_\perp$ is the transverse mass, given by

$$m_\perp^2 = m^2 + p_x^2 + p_y^2. \tag{2.15}$$

Rapidity can therefore be understood as the hyperbolic angle of a particle along the mass-shell relation $m_\perp^2 = E^2 - p_z^2$. The main benefit of using rapidity instead of velocity is that rapidities transform additively under Lorentz boosts. A rapidity boost of y_{boost} gives energy $E' = m_\perp \cosh(y + y_{\text{boost}})$ and longitudinal momentum $p_z' = m_\perp \sinh(y + y_{\text{boost}})$. Since Lorentz boosts amount to translations in rapidity space, rapidity differences are invariant under Lorentz boosts.

Figure (((REF))) shows a spacetime diagram of Lund string fragmentation in the longitudinal direction. There are a number of string breaks where new $q\bar{q}$ pairs are formed, and these quarks screen the colour field so that the string no longer runs between them. These string breaks result in a set of outgoing yo-yo modes, representing the outgoing hadrons. If the quark-antiquark pairs are enumerated $q_i\bar{q}_i$ along the string, then the hadrons from right to left are formed from $q_0\bar{q}_1$, $q_1\bar{q}_2$, and in general, $q_i\bar{q}_{i-1}$. Each hadron has positive lightcone momentum p_z^+ and negative lightcone momentum p_z^- that are linked by

$$p_z^+ p_z^- = m_\perp^2, \tag{2.16}$$

where $m_\perp$ is the transverse mass, defined by $m_\perp^2 = m^2 + p_x^2 + p_y^2$. It is therefore possible to describe the longitudinal momentum selections of all hadrons along the string just by their positive or negative lightcone momenta, since the other is fixed by the mass-shell requirement.

Figure (((REF))) shows this same string fragmentation process in a frame boosted to the right, where the string breaks are now ordered right-to-left. A critical feature of the Lund model is that all the string breakup vertices are spacelike separated, and therefore causally independent. The production of the hadrons can be ordered from right-to-left or left-to-right depending on the frame of reference. It is also worth noting that, because of the constant string tension, the string breakup vertices obey a constant probability area law, where the probability of a breakup occuring in a worldsheet region of area A is proportional to $\exp(-bA)$, where b is a tunable parameter. This implies the string fragmentation process must be covariant under Lorentz boosts along the string axis. Because of the constant tension and causal independence of breakup vertices, nothing that happens at one point on the string can causally affect anything at any other point, and in a sense the string is the "same" along its length. We therefore expect that observables — the density, rapidity spacing, and chemistry (species spectra) — should be the same for hadrons irrespective of where along the string they are formed. Since Lorentz boosts are equivalent to translations in rapidity space, these observables must be Lorentz invariant.

If string fragmentation is considered right-to-left (or left-to-right), then the self-similiarty of the fragmentation implied by Lorentz covariance is guaranteed if the probability of each newly formed hadron taking a fraction z of the remaining positive (or negative) lightcone momentum is the same for all hadrons. This is encapsulated in the fragmentation function $f(z)$, which gives a probability distribution of selecting values of z ranging from 0 to 1, normalised such that

$$\int_0^1 f(z)\,dz = 1. \tag{2.17}$$

If positive (or negative, if fragmenting left-to-right) lightcone momentum fractions are selected according to the same fragmentation function, then the string fragmentation will be self-similar and therefore Lorentz covariant. The lightcone momentum fractions z are Lorentz invariant, since the effect of a rapidity boost y_{boost} on the lightcone momenta amounts to a dilation.

The rapidity difference $\Delta y_i = y_i - y_{i-1}$ between subsequent hadrons produced in left-to-right fragmentation (such that the sign is correct) is given by

$$\Delta y_i = -\ln\Big(\frac{z_i}{z_{i-1}}\frac{m_{\perp,i-1}}{m_{\perp,i}}(1-z_{i-1})\Big), \tag{2.18}$$

where $m_{\perp,i}$ is the mass of hadron i and z_i is the lightcone momentum fraction selected for hadron i. If the same fragmentation function is used for all hadrons, then the distribution of rapidity spacings Δy is the same along the string, implying that the distribution dN/dy of hadron density per unit rapidity should be flat. This is called a "rapidity plateau", and its flatness also follows from the Lorentz invariance (and therefore invariance under rapidity translations) of the hadron density along the string.

To summarise, the constant tension and causal indepndence of breakups along the string implies that final state observables are Lorentz invariant and that the string fragmentation is therefore self-similar or Lorentz covariant. This implies a flat rapidity plateau, and a consistent distribution of particle species and rapidity differences along the string away from the endpoints. When string fragmentation is considered right-to-left or left-to-right, the Lorentz invariant lightcone momentum fractions z must be selected from a consistent fragmentation function $f(z)$ to preserve Lorentz covariance. The requirement that the same results are obtained from right-to-left or left-to-right fragmentation (left-right symmetry) constrains the allowed form of the fragmentation function to the

Lund symmetric fragmentation function

$$f(z) = \frac{N}{z}(1-z)^a \exp\Big(-\frac{bm_\perp^2}{z}\Big), \tag{2.19}$$

where a is a tunable parameter that can in principle vary with quark flavour, b is a tunable parameter equivalent to that of the area law, and N is a normalisation factor to enforce Equation (2.17).

String breaking in the Lund model is modelled analogously to the Schwinger mechanism of quantum electrodynamics [44], where the tunnelling probability of forming a quark-antiquark pair of mass m_q and transverse momentum $p_{\perp,q}$ is given by

$$\begin{aligned}\Pr(m_q, p_{\perp,q}) &\propto \exp\Big(\frac{-\pi m_q^2}{\kappa}\Big)\exp\Big(\frac{-\pi p_{\perp,q}^2}{\kappa}\Big) \\ &= \exp\Big(\frac{-\pi m_{\perp,q}^2}{\kappa}\Big).\end{aligned} \tag{2.20}$$

Here, $m_{\perp,q}$ is defined as the transverse momentum of the quark, where $m_{\perp,q}^2 = m_q^2 + p_{\perp,q}^2$. There is therefore a Gaussian suppression of string breaks in both transverse momentum and quark mass. Practically, PYTHIA implements the mass suppression factor using quark-specific probabilities that are tunable parameters, rather than using the Gaussian selection here, but the probabilities are approximately similar.

Equation (2.20) governs the flavour and transverse momentum selection of the quarks created in string breaks in the Lund model, which in turn determines the species and transverse momentum distributions of hadrons produced along the string in hadronisation. Equation (2.19) governs the longitudinal momentum selection of these hadrons, and in combination, they describe how hadrons and their four-momenta are selected in $q\bar{q}$ hadronisation.

There are a number of other string topologies that can emerge in the evolution of a partonic state to a set of primary hadrons, some of which are illustrated in Figure (((REF))). For example, strings can stretch between gluons in between a quark and antiquark, and these gluons are treated as "kinks" on the string. Strings can also form between multiple gluons in so-called "gluon loops", and SU(3) junction structures can form between three quarks or antiquarks.

Chapter 3

String Fragmentation in PYTHIA

3.1 The PYTHIA Fragmentation Algorithm

Having established how the Lund string model describes hadronisation, we can now proceed to a description of how this model is implemented in the PYTHIA event generator. Actual events involve many hundreds of outgoing partons from the parton shower, and these will combine into various string topologies. For the purposes of this project, we will focus solely on the fragmentation of a $q\bar{q}$ string into a number of outgoing primary hadrons, prior to decays, and without any gluon kinks, loops, or junctions. As we will see, Lorentz invariance is entirely violated in PYTHIA even in this minimal situation.

In the Lund model, the Schwinger mechanism describes the relative probabilities of quark flavours in a string breakup. In turn, the probabilities of forming the various possible hadrons from a combination of quarks and antiquarks are theoretically dependent only on the various mixings of quark states that comprise the hadrons. In PYTHIA, these probabilities are implemented as tunable parameters, but the overall process is the same. Also described by the Schwinger mechanism is the transverse momenta of the quarks formed in string breakups, and therefore the transverse momenta of all final-state hadrons.

The Lund model describes the selection of hadron longitudinal momenta in terms of an iterative left-right (or right-left) ordering of string breakups, and the fragmentation function $f(z)$. This model is derived from the causal independence of breakup vertices and the resulting Lorentz covariance and self-similarity of the string fragmentation. It is important to note, however, that the Lund model provides no description of how energy and momentum are globally conserved in string fragmentation. If string fragmentation is considered iteratively, there will eventually come a point where there is insufficient energy left in the string to create a hadron with the required mass. PYTHIA enforces energy conservation by halting the string fragmentation when the remaining string energy is below a certain value (with some smearing), and then creating two final hadrons. Energy-momentum conservation and the on-shell condition fix the kinematics of these final two hadrons entirely. In the event where there is insufficient energy to create these two hadrons, the event is thrown out and string fragmentation is restarted.

This implementation of energy conservation manifestly breaks Lorentz covariance. The manner in which the kinematics of these final two hadrons are determined is different from the rest of the hadrons along the string. Furthermore, refragmenting the entire string when it is not possible to create the final two hadrons introduces bias in the species makeup of these final two. This is in contradiction with the expectation that the string fragmentation

process should be self-similar along the string. The PYTHIA documentation (and online manual, prior to updates that were part of this project) describe how this violation of Lorentz covariance is compensated for in a few ways. Firstly, instead of fragmenting left-to-right or right-to-left, hadrons are fragmented from either end with equal probability, in order to smear the position of the joining step uniformly along the string. Secondly, the parameters that control when string fragmentation is stopped are carefully chosen in order to give a flat rapidity plateau [8, 18]. It was thought that the dependence of these final two stopping parameters on the rest of the fragmentation parameters (such as the Lund a and b parameters, or the quark and hadron probabilities) was minimal, and that as such these parameters would not need to be adjusted between different PYTHIA tunes [45]. As we will see, none of these statements are true.

Algorithm 1 shows the algorithm used in PYTHIA 8.316 [1] for $q\bar{q}$ string fragmentation. It takes in the two string endpoint flavours as well as the centre-of-mass energy, and populates the event record event with a set of hadrons, each with a specified species, mass, and four-momentum. The transverse momenta of the string break quarks (or antiquarks) $q^{\pm}_{\text{break}}$ are selected according to (2.20), and the selection of their flavours is encapsulated by the procedure FlavSel. The details of this procedure depend on the various PYTHIA parameters that govern flavour selection, which have the prefix `StringFlav`. The dependence on the current string end quark flavour is only to determine whether the new parton should be a quark or antiquark (or a diquark or antidiquark), such that the resulting hadron is either a meson or a baryon.

The procedure combine takes two quark flavours/antiflavours and combines them into a resultant hadron. Again, this procedure depends on a number of PYTHIA parameters with the prefix `StringFlav`. Neither the details of FlavSel or combine are within the scope of this project, beyond the fact that they encapsulate the species makeup of the hadrons produced along the string. Similarly, the procedure zFrag combines the Lund fragmentation function (2.19) with other fragmentation functions used depending on the tune and quark flavours, and generally selects a z fraction for the new hadron depending on the quark flavours and hadron transverse mass $m_{\perp}$. The mass of each hadron is selected based on its species, usually according to a Breit-Wigner distribution, in behaviour contained within the procedure mSel.

This z fraction specifies the fraction of the positive (or negative, depending on which end fragmentation is currently from) lightcone momentum remaining in the string. The product p^+p^- of the lightcone momenta of the string gives the remaining string mass

$$W_{\text{rem}} = \sqrt{p^+p^-}. \tag{3.1}$$

If $W_{\text{rem}} \leq W_{\text{min}}$, where W_{min} is selected according to the procedure Wmin, then string fragmentation is stopped and the final two hadrons are created. Note that if creating the most recent hadron used more energy than was available in the string, then string fragmentation will also be stopped at this step, and FinalTwo will fail automatically. If FinalTwo does fail, which can occur for this reason or if there is insufficient energy in the string to create the final two hadrons, then the entire string fragmentation process is restarted.

[1] Version 8.316 of PYTHIA was released on October 3, 2025, during the writing of this thesis. It contains a number of changes, including a fix to a bug where tunes were incorrectly loaded when reading subrun settings from a file, which was uncovered as part of this project. The default value of `StringFragmentation:stopMass` was also changed, for reasons explained later in this thesis. There are no actual changes to the fundamental algorithm used in fragmentation or the final two between PYTHIA 8.316 and PYTHIA 8.315.

Algorithm 1 The default PYTHIA 8.316 algorithm for $q\bar{q}$ hadronisation.

procedure FRAGMENT($E_{\mathrm{CM}}, \mathrm{flav}_{+\mathrm{end}}, \mathrm{flav}_{-\mathrm{end}}$)
  repeat
    initialise event record `event`
    $\vec{p}_{\perp,+\mathrm{end}} \leftarrow \vec{0}$
    $\vec{p}_{\perp,-\mathrm{end}} \leftarrow \vec{0}$
    $p^+ \leftarrow E_{\mathrm{CM}}$
    $p^- \leftarrow E_{\mathrm{CM}}$
    loop
      fromPos $\leftarrow$ true or false with equal probability ▷ *$\pm$ and $\mp$ reflect this selection*
      $\vec{p}_{\perp,\pm\mathrm{break}} \leftarrow$ transverse momentum selected according to (2.20) ▷ *Select transverse momentum*
      $\vec{p}_{\perp,\mp\mathrm{break}} \leftarrow -\vec{p}_{\perp,\pm\mathrm{break}}$
      $\vec{p}_{\perp} \leftarrow \vec{p}_{\perp,\pm\mathrm{end}} + \vec{p}_{\perp,\pm\mathrm{break}}$
      $\vec{p}_{\perp,\pm\mathrm{end}} \leftarrow \vec{p}_{\perp,\mp\mathrm{break}}$
      repeat ▷ *Select string break flavour and new hadron species, mass, and z fraction*
        $\mathrm{flav}_{\pm\mathrm{break}} \leftarrow$ FLAVSEL($\mathrm{flav}_{\pm\mathrm{end}}$)
        $\mathrm{flav}_{\mp\mathrm{break}} \leftarrow \mathrm{anti}(\mathrm{flav}_{\pm\mathrm{end}})$
        species $\leftarrow$ COMBINE($\mathrm{flav}_{\pm\mathrm{end}}, \mathrm{flav}_{\pm\mathrm{break}}$)
        $m \leftarrow$ MSEL(species)
        $m_\perp \leftarrow \sqrt{m^2 + |\vec{p}_\perp|^2}$
        $z \leftarrow$ ZFRAG($\mathrm{flav}_{\pm\mathrm{end}}, \mathrm{flav}_{\pm\mathrm{break}}, m_\perp$)
        $\mathrm{flav}(q_{\pm\mathrm{end}}) \leftarrow \mathrm{flav}(q_{\mp\mathrm{break}})$
      until COMBINE succeeds
      if $\sqrt{p^+p^-} \leq$ WMIN($\mathrm{flav}_{+\mathrm{end}}, \mathrm{flav}_{-\mathrm{end}}, \mathrm{flav}_{+\mathrm{break}}, \mathrm{flav}_{-\mathrm{break}}$) **then**
        break
      $p^{\pm}_{\mathrm{hadron}} \leftarrow zp^{\pm}$ ▷ *Calculate longitudinal momentum*
      $p^{\pm} \leftarrow (1-z)p^{\pm}$
      $p^{\mp}_{\mathrm{hadron}} \leftarrow p^{\pm}/m^2_\perp$
      $p^{\mp} \leftarrow p^{\mp} - p^{\mp}_{\mathrm{hadron}}$
      $E \leftarrow p^{+}_{\mathrm{hadron}} + p^{-}_{\mathrm{hadron}}$
      $p_z \leftarrow p^{+}_{\mathrm{hadron}} - p^{-}_{\mathrm{hadron}}$
      append new hadron with $(\mathrm{species}, m, \vec{p}_\perp, p_z, E)$ to `event`
    FINALTWO(`event`, $p^+, p^-, q_{+\mathrm{end}}, q_{-\mathrm{end}}$, species, $m, \vec{p}_\perp$)
  until FINALTWO succeeds
  return `event`

	PYTHIA 6	PYTHIA 8
`stopMass`	0.8	1.0
`stopNewFlav`	2.0	2.0
`stopSmear`	0.2	0.2

Table 3.1: Default values for final two procedure parameters in PYTHIA 6 and PYTHIA 8 [18,45]. Note the change in `stopMass`.

Not included in Algorithm 1 or the pseudocode above is a description of the so-called "popcorn model", which gives a better fit to experimental data by allowing mesons to be formed between baryons when a diquark string break occurs. The effects of this model do not affect the discussion in the rest of this thesis, but figures generated are generated with the default popcorn model settings in order to reflect the typical behaviour of PYTHIA.

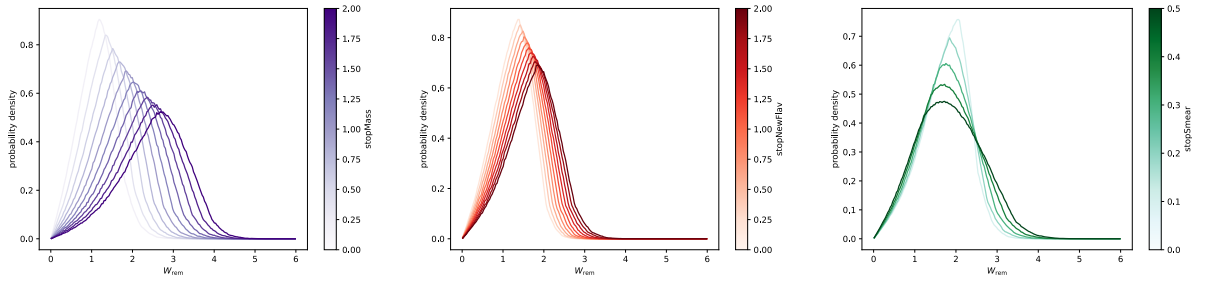
Worth mentioning is a paper by Edén investigating the deviations from the Lund model area law inherent in a procedure where string ends are joined once the remaining energy is below some threshold, even when the final two hadrons are created in a Lorentz invariant manner [46]. Edén presents theoretical solutions to these deviations, but they are a separate but related problem to the issue we will see of the kinematics and hadronic chemistry of the final two being different to the hadrons along the rest of the string. As such, the problems presented by Edén will not be considered further within the scope of this thesis.

3.2 The Joining Step

Having established the overall structure of the PYTHIA fragmentation algorithm, we can now look more closely at how and when the final two hadrons are created. The remaining string mass at which the FinalTwo procedure fails is stipulated by the procedure Wmin, shown in Algorithm 2. The behaviour of Wmin depends on three parameters in PYTHIA. The first is `StringFragmentation:stopMass`, which allows the mean W_{min} to be adjusted. The next is `StringFragmentation:stopNewFlav`, which also adjusts the mean of W_{min} but with a dependence on the new string break quark mass. This allows for fragmentation to be stopped earlier in order to have sufficient energy to create heavier final hadrons (although there is still not always enough energy). Finally, `StringFragmentation:stopSmear` controls the spread of the W_{min} values that are used. The default values of these parameters in PYTHIA are shown in Table 3.1. Note how `stopMass` was changed from 0.8 to 1.0 between PYTHIA 6 and 8.315. This change has since been reverted in PYTHIA 8.316.

Figures 3.1a, 3.1b, and 3.1c show the effects of adjusting these three parameters on the probability distribution of actual W_{rem} values encountered at the final two step, as simulated in PYTHIA 8.315 for $d\bar{d}$ hadronisation at 200 GeV (with default settings, including the Monash tune). These plots show how `stopMass` and `stopNewFlav` affect the mean of W_{min}, while `stopSmear` affects the spread.

When the remaining string energy is below W_{min}, the FinalTwo procedure creates the final two hadrons, as shown in Algorithm 3. In this procedure, the already-selected string break flavour and transverse momenta are used, and hence the species, mass, and transverse momenta $(\text{species}, m, \vec{p}_\perp)$ of one of the final two hadrons is already determined. The leftover quark flavours and transverse momenta are used to determine the properties $(\text{species}_{\text{other}}, m_{\text{other}}, \vec{p}_{\perp,\text{other}})$ of the other final hadron. Finally, the remaining string energy



(a) For stopMass. (b) For stopNewFlav. (c) For stopSmear.

Figure 3.1: Distributions of the remaining available string mass W_{rem} when entering the `finalTwo` procedure, for $d\bar{d}$ hadronisation at 200 GeV in PYTHIA 8.316. Each plot shows how one of the three `finalTwo` parameters affects the distribution of W_{rem}. `stopMass` and `stopNewFlav` have the effect of shifting the centre of the distribution (with some impact on the spread), while `stopSmear` primarily affects the spread. Note that these W_{rem} values include runs of `finalTwo` that would eventually be vetoed, such that the plots represent the actual distribution of values encountered at the start of the `finalTwo` procedure.

Algorithm 2 The procedure to determine the W_{min} string mass threshold to stop fragmentation.

procedure WMIN($\text{flav}_{+\text{end}}, \text{flav}_{-\text{end}}, \text{flav}_{+\text{break}}, \text{flav}_{-\text{break}}$)
- $W_{\text{min}} \leftarrow$ `stopMass` $+ m_{+\text{end}} + m_{-\text{end}}$
- $W_{\text{min}} \leftarrow W_{\text{min}} +$ `stopNewFlav` $* m_{+\text{break}}$
- $W_{\text{min}} \leftarrow (1 +$ `stopSmear` $*$ RANDOM$(-1, 1)) * W_{\text{min}}$
- **return** W_{min}

is used to give these final hadrons their longitudinal momenta, which are fixed by energy conservation and the on-shell requirement.

There are three possible ways for this procedure to fail, as shown in the pseudocode. Firstly, if the remaining string energy is already negative when FinalTwo is called, it will fail. This occurs if the previous hadron created in the fragmentation loop took more energy than was remaining in the string. Secondly, if the two leftover flavours for the other final hadron are both diquarks, then FinalTwo also fails, since there are no tetraquark states in PYTHIA. Finally, if there is insufficient energy in the string to create the final two hadrons with their required transverse momenta and masses, FinalTwo fails. In all of these cases, the event is discarded, and fragmentation begins again from the beginning.

3.3 Performance of the Current finalTwo Procedure

The fragmentation algorithm used in PYTHIA 8.316 manifestly violates Lorentz covariance in how it handles the final two hadrons. The longitudinal kinematics of these hadrons are determined to enforce energy conservation, without the degrees of freedom available to the rest of the hadrons produced along the string. Unless the three parameters controlling W_{min} are tuned perfectly (if such a thing is possible — we will find it is not), the kinematics of the final two hadrons will therefore be differently distributed than the hadrons produced along the string. We will denote these hadrons produced outside of the final two and not adjacent to the endpoints as "regular" hadrons. On top of a deviation in kinematics, the fact that FinalTwo can fail introduces biases in the species composition of the final two, which we will denote the "hadrochemistry". As such, the hadrochemistry

Algorithm 3 The procedure to create the final two hadrons in PYTHIA 8.316.

procedure FINALTWO(`event`, $p^+, p^-, q_{+\text{end}}, q_{-\text{end}}$, species, $m, \vec{p}_\perp$)
  if $p^+p^- \leq 0$ **then**
    return failure ▷ *Fails if last hadron used more energy than available*
  if $\text{flav}_{+\text{end}}$ and $\text{flav}_{-\text{end}}$ are both diquarks **then**
    return failure ▷ *Cannot join two diquarks*
  $\vec{p}_{\perp,\text{other}} \leftarrow \vec{p}_{\perp,+\text{end}} + \vec{p}_{\perp,-\text{end}}$
  repeat
    $\text{species}_{\text{other}} \leftarrow$ COMBINE($\text{flav}_{+\text{end}}, \text{flav}_{-\text{end}}$)
    $m_{\text{other}} \leftarrow$ MSEL($\text{species}_{\text{other}}$)
  until COMBINE succeeds
  $m_{\perp,\text{other}} \leftarrow \sqrt{m_{\text{other}}^2 + |\vec{p}_{\perp,\text{other}}|^2}$
  if $\sqrt{p^+p^-} < m_\perp + m_{\perp,\text{other}}$ **then**
    return failure ▷ *Not enough energy to create final two hadrons*
  $p_z, p_{z,\text{other}}, E, E_{\text{other}} \leftarrow$ values calculated to use remaining string energy and have final two hadrons on-shell
  append hadron with $(\text{species}, m, \vec{p}_\perp, p_z, E)$ to `event`
  append hadron with $(\text{species}_{\text{other}}, m_{\text{other}}, \vec{p}_{\perp,\text{other}}, p_{z,\text{other}}, E_{\text{other}})$ to `event`
  return `event`

of the final two hadrons will be different from the hadrochemistry of regular hadrons. These deviations in the kinematics and hadrochemistry of the final two hadrons comprise the violation of Lorentz covariance introduced by this handling of energy conservation in PYTHIA's string fragmentation algorithm.

A simple picture of the Lorentz covariance of the kinematics of hadrons along the string can be obtained from the shape of the rapidity plateau. Figure 3.2 shows the rapidity plateaus (dN/dy distributions) for $d\bar{d}$ hadronisation across PYTHIA 6.428 (which uses the JETSET tune), PYTHIA 8.186 (which uses the Hoeth tune), and PYTHIA 8.315 (which uses the JETSET tune) [3]. These results are for a rather long 200 GeV string, for which a very flat central region would be expected in the rapidity plateau. Indeed, the plateau in PYTHIA 6.428 is entirely flat in the middle, with the "ears" on either side explainable by endpoint effects, since Lorentz covariance only applies away from the endpoints.

On the other hand, the plateaus for both PYTHIA 8 versions show a significant dip in the middle, deviating from flatness by around 5%. To factor out effects of different PYTHIA versions being coded differently, Figure (((REF))) shows the same rapidity plateaus for these three tunes, all within PYTHIA 8.315. Notably, the plateau for the JETSET tune is not as flat. This is a result of a bug in how hadron kinematics are calculated in PYTHIA 8 [4] [47]. It is, however, mostly flat, and there is still a much larger dip for the Hoeth and Monash tunes. Figure (((REF))) shows how the rapidity plateau for the Monash tune varies across 50, 100, 200 and 500 GeV strings.

[2] In general, the plots shown in this thesis are generated using at least 1,000,000 events. Specific numbers are not included for brevity, but the number of events is selected such that the curve is smooth and uncertainties are negligible. As a result, uncertainties are also generally not included in plots.

[3] PYTHIA 8.315 is used here rather than PYTHIA 8.316 because the change to `stopMass` in PYTHIA 8.316 provides a small improvement in the flatness of the rapidity plateau.

[4] This bug was introduced when PYTHIA was ported to C++ from FORTRAN, between versions 6 and 8. It was discovered by Stephen Mrenna a few weeks before this thesis was submitted [47], and there was not enough time in this project to properly investigate the implications.

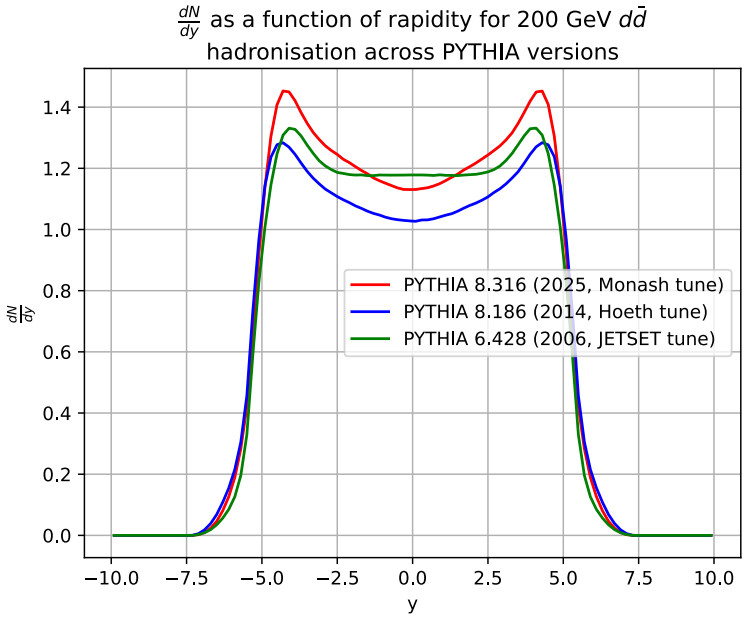


Figure 3.2: Plots of the rapidity plateau $\frac{dN}{dy}$ as a function of rapidity y, using default settings of different PYTHIA versions. Events were generated for $d\bar{d}$ hadronisation at 200 GeV, with 10,000,000 events [2]. PYTHIA 6.428 (released in 2006) is the final version written in FORTRAN, and uses the default JETSET tune. PYTHIA 8.186 (released in 2014) is the final version of PYTHIA 8.1 and written in C++. It uses the Hoeth tune. Finally, PYTHIA 8.315 (released in 2025) is the second-most-current release and uses the Monash tune. The rapidity plateau in PYTHIA 6.428 is flat in the middle, with bumps on either end due to endpoint effects. Both PYTHIA 8 versions have plateaus with dips in the middle, with deviations from a flat plateau by around 5%. (((TODO: quantify this further?))) Evidently, the Lorentz invariance of hadron density per unit rapidity has been violated for over a decade, at least since the release of PYTHIA 8.1.

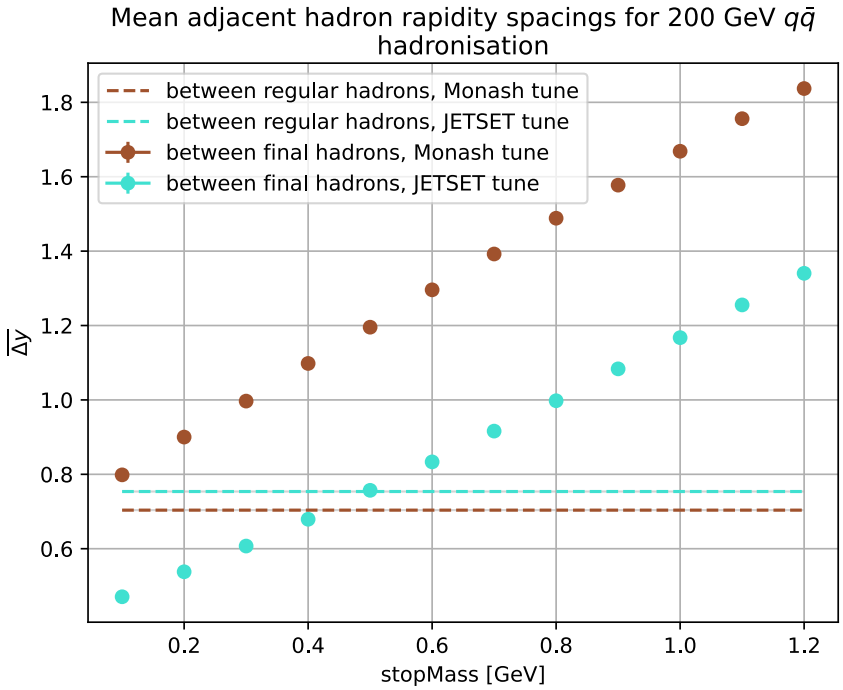


Figure 3.3: stopmass

(((HADRON-SPECIFIC)))

The presence of this dip in the rapidity plateaus of the Monash and Hoeth tunes exposes the violation of Lorentz covariance in the PYTHIA string fragmentation algorithm, and disproves the assertion that the values of the final two parameters (`stopMass`, `stopNewFlav`, and `stopSmear`) do not need to be adjusted between tunes. As mentioned earlier, `stopMass` was changed from 0.8 to 1.0 in the move from FORTRAN to C++. The rapidity plateau for a `stopMass` of 0.8 in the Monash tune is also shown in Figure (((REF))), and this change is evidently not enough to explain the violation of Lorentz covariance.

The reason for this dip can be understood in the following way. A useful measure for the longitudinal kinematics of the final two hadrons (especially in terms of how they affect the rapidity plateau) is the rapidity difference Δy_{final} between them. When the final two hadrons are created, the remaining string energy W_{rem} is used up to create the masses of the final hadrons, and the kinetic energy of the final hadrons. If W_{rem} is on average larger, then Δy_{final} would also be larger, since there would be more kinetic energy available to create a larger rapidity spacing. On the other hand, if W_{rem} is on average smaller, then Δy_{final} would also be smaller.

This behaviour is visible in Figure 3.3 which shows how, for the Monash and JETSET tunes, the mean final rapidity difference $\overline{\Delta y}_{\text{final}}$ varies for differing values of `stopMass` (the parameter with the most direct impact on the mean of W_{rem}). As expected, $\overline{\Delta y}_{\text{final}}$ increases as `stopMass` increases and more kinetic energy is available for the final two. The horizontal dotted lines show the mean rapidity differences between regular hadrons along the string, $\overline{\Delta y}_{\text{reg}}$. The mean rapidity differences for both regular and final hadrons vary with the tune. This is because the kinematics of both regular and final hadrons are dependent on the various PYTHIA parameters that control fragmentation functions and relative quark/hadron flavour/species probabilities. The dependence is complicated, and

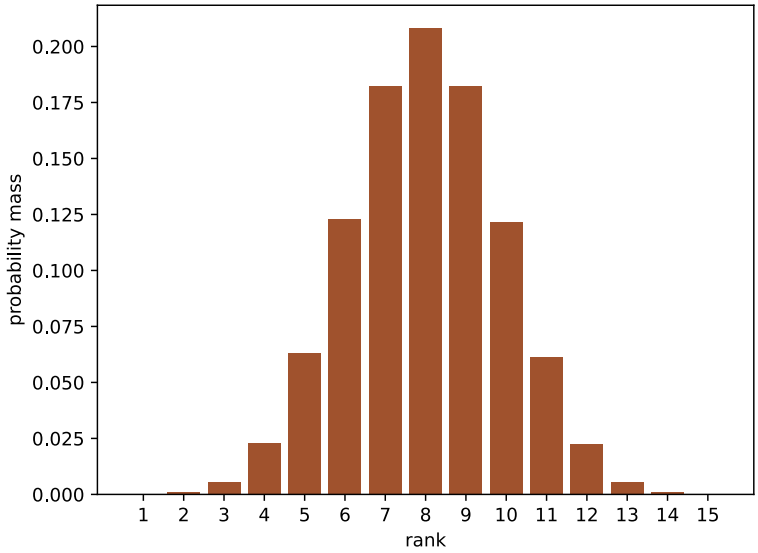


Figure 3.4: A histogram showing the probability distribution of the rank of the final two hadrons in PYTHIA $d\bar{d}$ hadronisation events. Only events that result in 15 primary hadrons are shown, such that the distribution is isolated. Since hadrons are fragmented from either end with equal probability, the majority of final two pairs are produced in the centre of the string.

an analytic expression is not necessary here.

Notably, the value of `stopMass` required to match the mean rapidity spacing between the final two and the rest of the string is much higher for the JETSET tune than it is for the Monash tune. Without a decrease in `stopMass` moving from the JETSET to the Monash tune (indeed, `stopMass` was actually increased), the mean rapidity spacing of the final two hadrons is much larger than the mean rapidity spacing of regular hadrons. This has the effect of "pushing out" the final two in rapidity space.

If the rank of the final two hadrons along the string was truly uniformly smeared along the string, as is intended by the random fragmentation from the left or right, then the effects of this "pushing out" would also be uniformly smeared. However, this random fragmentation implies the number of string breaks that are done from the left (or right) is binomially distributed, and therefore the rank of the final two hadrons is on average in the centre of the string. Figure 3.4 shows a histogram of the ranks of the final two hadrons in 200 GeV $d\bar{d}$ hadronisation in PYTHIA 8.316, limited specifically to events with 14 hadrons. An overall symmetrical Gaussian distribution is visible.

The dip in the rapidity plateau occurs because of these two effects. The final two hadrons are on average produced in the middle of the string, and they are on average spaced further apart in rapidity space than the rest of hadrons along the string, causing a relative underproduction of hadrons in rapidity space at the centre of the string. It is also worth noting that even in PYTHIA 6, where the rapidity plateau is flat, the shape of the distribution of Δy is different for regular and final hadrons. Figures (((REF))) and (((REF))) show how Δy is distributed for final and regular hadrons in the JETSET and Monash tunes respectively. In both tunes, there is a dip at $\Delta y = 0$ and the overall shape is obviously not the same. This indicates a violation of Lorentz covariance in hadron kinematics beyond what is visible in the rapidity plateau.

We also note here the large deviations in hadrochemistry of the final two hadrons.

Figure (((REF))) shows the percentage differences in the proportions of hadron species produced in the final two compared to regular hadrons. Hadrons are ordered by mass along the horizontal axis, and antiparticles are counted as the same species as particles (since they have identical masses, and string fragmentation is symmetric under charge conjugation). The dependence of the deviations (or their direction) on mass is complex. It would be expected that the bias introduced by refragmenting the string when there is insufficient energy to create the final hadrons would lead to a bias towards lighter hadrons. This is somewhat visible — π^0 is overproduced by (((VALUE))) and Δ^0 is underproduced by (((VALUE))). However, other biases are introduced by the diquark joining failures of FinalTwo as well as effects of `stopNewFlav`. There is also an additional bias introduced since quark flavours are reselected in Algorithm 1 when combine fails, but quark flavours are never reselected in Algorithm 3. As a whole, these deviations in hadrochemistry can be expressed by the sum of squared errors, given by

$$\text{SSE} = \sum_{\{i\}} (p_{i,\text{final}} - p_{i,\text{reg}})^2, \tag{3.2}$$

where $\{i\}$ runs over all hadron species, $p_{i,\text{final}}$ is the proportion of hadron species i in the final two hadrons, and $p_{i,\text{reg}}$ is the proportion of hadron species i in regular hadrons.

The values of the sum of squared errors for the JETSET and Monash tunes are shown in Figure (((REF))), alongside the average number of failures per generated event. Evidently, for both tunes, the failure rate of `finalTwo` is high, and there are notable deviations in hadronic chemistry.

In conclusion, the algorithmic implementation of the Lund model in PYTHIA implements energy conservation by manifestly breaking Lorentz covariance. This violation of Lorentz covariance did not lead to a non-flat rapidity plateau in earlier versions of PYTHIA because in the JETSET tune the rapidity spacing was similar between final and regular hadrons. However, in the Hoeth tune (and later the Monash tune), the value of `stopMass` required for a flat plateau became much smaller, and since this value was not actually adjusted, a dip was introduced to the rapidity plateau, where it remained for over a decade. Aside from the deviation in kinematics, the hadrochemistry of the final hadrons is also distributed differently than regular hadrons across all tunes because of the high failure rate of `finalTwo`. This failure rate also causes difficulties in statistical analysis and machine learning using PYTHIA (((CITATIONS))).

As a whole, the violation of Lorentz covariance in PYTHIA hadronisation is quite severe, and represents quite a large deviation from the theoretical predictions of the Lund model. As an immediate solution, `stopMass` was reverted to 0.8 in PYTHIA 8.316, which provides a slightly flatter rapidity plateau. The PYTHIA online manual was also updated to reflect this new understanding. The rest of this thesis explores techinques for restoring or preserving Lorentz covariance in PYTHIA, through tuning the W_{min} parameters, introducing a new parameter, and redesigning the fragmentation model.

Chapter 4

Tuning Lightcone Scaling in PYTHIA

4.1 Restoring Lightcone Scaling by Tuning Parameters

Having established the extent and causes of the breaking of Lorentz covariance in PYTHIA hadronisation, we can now move on to exploring solutions to this problem. It would be easiest if we could at the very least tune for a flat rapidity plateau (and ideally restore Lorentz covariance) using the three parameters that control when the final two hadrons are created — `stopMass`, `stopNewFlav`, and `stopSmear`. Figure (((REF))) shows that, in the Monash tune, a flat rapidity plateau can be obtained by decreasing `stopMass` to approximately (((VALUE))), and that the deviation in Δy increases linearly with `stopMass`. Figure (((REF))) shows the rapidity plateau for 200 GeV $d\bar{d}$ hadronisation for varying values of `stopMass`, and the effect on the flatness is clear, with a sweet spot shown at the value of (((VALUE))) where the Δy means are equal across final and regular hadrons.

The issue with simply turning down `stopMass` (or reducing W_{min} in general) is that the bias towards lighter hadrons in the final two hadrochemistry is significantly increased for smaller W_{min}, and in the extreme case of `stopMass` = 0.05 failures are common enough to cause noticeable slowdowns and pions make up (((VALUE)))% of all final hadrons. Figure (((REF))) shows how the hadrochemistry SSE varies with `stopMass` — worse, there is no value that brings it close to zero.

Similar to `stopMass`, `stopNewFlav` affects the mean of W_{min}, but with a more favourable effect on hadronic chemistry since this parameter accounts somewhat for the mass of the final hadrons that will be produced in its effect on W_{min}. Figure (((REF))) shows how the hadrochemistry SSE varies with `stopNewFlav`. (((COMMENT))).

Since `stopSmear` primarly affects the spread of W_{min}, its main utility is in aligning the shape of the Δy distribution between the final and regular hadrons. Figure (((REF))) shows, however, how these distributions are inherently shaped differently beyond just a dilation. Through experimentation, it was found that the values (((VALUES))) give a good improvement on the flatness of the rapidity plateau without a significant worsening of hadrochemistry SSE, as well in Figure (((REF))) which compares the resulting rapidity plateau and SSE with the PYTHIA 8.316 defaults.

In general, for the Monash tune, and a wide range of potential tunes, there is no way

to restore Lorentz covariance using the existing `finalTwo` parameters in PYTHIA. Even a flat rapidity plateau cannot be obtained without extreme violations in Lorentz covariance of hadrochemistry along the string. Fundamentally, the issue is that there is only one degree of freedom — the mean of W_{min} and hence Δy — with which to adjust for a flat rapidity plateau. Increasing the mean improves the hadronic chemistry but worsens the dip in the rapidity plateau, and vice versa.

The next section describes a new parameter, `probRevertFinal`, which provides some additional freedom in parameter space to adjust for Lorentz covariance.

4.2 The probRevertFinal parameter

Algorithm 4 The PYTHIA 8.316 algorithm for $q\bar{q}$ hadronisation, modified to implement `probRevertFinal`

procedure FRAGMENT(E_{CM}, flav$_{+\text{end}}$, flav$_{-\text{end}}$)
  repeat
    initialise event record
    loop
      fromPos ← true or false with equal probability ▷ $\pm$ *and* $\mp$ *reflect this selection*
      select transverse momenta for string break
      repeat
        select flavour for string break
        select species and mass for new hadron
        select z fraction for new hadron
      until COMBINE succeeds
      if $\sqrt{p^+p^-} \leq$ WMIN(flav$_{+\text{end}}$, flav$_{-\text{end}}$, flav$_{+\text{break}}$, flav$_{-\text{pbreak}}$) **then**
        break
      calculate energy and longitudinal momentum for new hadron
      if $E < 0$ **then**
        break
      append new hadron with (species, m, $\vec{p}_\perp$, p_z, E) to `event`
    if FINALTWO(`event`, p^+, p^-, $q_{+\text{end}}$, $q_{-\text{end}}$, species, m, $\vec{p}_\perp$) fails **then**
      Revert previous string break
      Repeat FINALTWO with COMBINECONDITIONAL
  until FINALTWO succeeds
  return `event`

The next option for restoring Lorentz covariance would ideally provide a way to improve hadrochemistry without a significant affect on the rapidity plateau. One way to do this is to reduce the bias in the final two hadrochemistry when there is insufficient energy to create the final two hadrons. By default, PYTHIA refragments the entire string in this case. Instead of refragmenting from the start, we could instead undo the previous string break from the end we are fragmenting from, and then try to create the final two hadrons again. The parameter `probRevertFinal` was implemented in PYTHIA 8.316, which gives a probability that the final string break will be reverted when `finalTwo` fails because of insufficient string mass. Algorithm 4 shows how the FRAGMENT procedure is adjusted to

implement this parameter — changes are highlighted in red, and some lines are collapsed for brevity.

Algorithm 5 High level description of the algorithm to combine quark flavours with conditional spin flipping

```
procedure CombineConditional(flav_1, flav_2, flav_old, spin_old)
    condProbPrev ← probability of selecting spin_old from flav_1 and flav_old
    condProbCurrent ← probability of selecting spin_old from flav_1 and flav_2
    if Random(0, 1) > condProbCurrent/condProbPrev then
        select hadron with opposite light/heavy spin from spin_old
    else
        select hadron with same light/heavy spin as spin_old
    return selected hadron
```

When one of the final hadrons is reselected from its new constitutent flavours after the string break is reverted, the selection between a light and heavy (pseudoscalar vs. scalar, or vector vs. axial vector) hadron spin is flipped with a probability Pr(flip), given by

$$Pr(\text{flip}) = \frac{Pr(\text{previous heavy spin}|\text{current flavour})}{\Pr(\text{previous heavy spin}|\text{previous flavour})}. \tag{4.1}$$

Figure (((REF))) illustrates how this equation has the effect of (as much as possible) ensuring the "same" random number is used to select both spins. Algorithm 5 has high level pseudocode of the CombineConditional procedure which takes in the old flavour and spin and selects a hadron species accordingly.

Figure (((REF))) shows how the rapidity plateau and hadrochemistry SSE vary as `probRevertFinal` varies from 0 to 1. As expected, increasing `probRevertFinal` decreases the SSE, and while there is some worsening of the rapidity plateau (since reverting the final break on average increases the energy available for the final two), the effect is quite mild. Through experimentation with all four parameters, the combination (((VALUES))) was found to give a good improvement in the rapidity plateau with a lessened effect on SSE compared to the same values with `probRevertFinal = 0`, as shown in Figure (((REF))).

4.3 Limitations

The `probRevertFinal` parameter successfully provides more parameter space and allows for Lorentz covariance to be restored further. However, the rapidity plateau in Figure (((REF))) is not as flat as that of the JETSET tune in Figure (((REF))), and the hadrochemistry SSE is enough for noticeable distortions in particle spectra across the string. The shape of the Δy distribution is also different between regular and final hadrons. Finally, the manifest breaking of Lorentz covariance in the structure of the algorithm means that restoring Lorentz covariance by obtaining a flat rapidity plateau and a negligible hadrochemistry SSE is impossible without using a different model to conserve energy in string fragmentation.

One might assume the thesis would end here by describing the impact of this research and the possibilities for further investigations. The attuned reader, though, would realise that this feels like a narratively unsatisfying ending to the paper. The reader can rest assured that the humble author is goated. In the next section, we describe a new model

of string fragmentation that enforces energy conservation in a Lorentz covariant way, restoring Lorentz covariance to the PYTHIA hadronisation algorithm.

Chapter 5

The Accordion Model of String Fragmentation

5.1 The Accordion Model

To motivate this new model, it is useful to take a step back to the basics of the Lund model — what properties we expect string fragmentation to have, what these say about how the algorithm can be structured, and the nature of the actual physical process of hadronsation. The Lund model describes $q\bar{q}$ string fragmentation in terms of causally independent string breaks via the Schwinger mechanism on a string worldsheet. This property combined with the possibility of considering string breaking iteratively and left-right symmetry imply the form of the Lund symmetric fragmentation function in Equation (2.20). In the limit where the remaining string mass is infinite, the longitudinal momenta (or the Lorentz invariant rapidity spacings) of hadrons can be selected according to this fragmentation function, and this provides a Lorentz covariant fragmentation.

More fundamentally, hadronisation is the evolution of an incoming partonic $q\bar{q}$ state into an outgoing state consisting of a set of hadrons, subject to the constraint of energy conservation and the requirement that the evolution into the final state is Lorentz covariant away from the endpoints. Although the breakup vertices are causally independent, they are also correlated in the enforcement of energy conservation, and Lorentz covariance indicates there can be no "special treatment" of any hadrons. As such, the effect of the energy conservation constraint on the fragmentation process should be global. A Lorentz covariant model for string fragmentation should therefore enforce energy conservation in a way that is both global and Lorentz covariant. Physically, this would correspond to the quantum entanglement of the breakup vertices, alongside their causal disconnection. As a phenomenological model, we cannot expect to accurately model quantum entangelement effects, but this at least provides a justification for the model exhibiting global behaviour. At least metaphorically, the partonic state evolves "simultaneously" to the final state, in the sense that while nothing that occurs at any point along the string has any causal or time-ordered influence on any other point (so any ordering is theoretically appropriate), there are also global considerations that forbid a strictly iterative model.

The accordion model of string fragmentation preserves Lorentz covariance by allowing energy conservation to be violated and sampling rapidity spacings for the final two hadrons based on selections of trial z fractions from fragmentation functions, so that the rapidity spacings and particle spectra are identically distributed along the string. Energy conservation is then restored by a global Lorentz covariant rapidity dilation $y \to ky$,

where a solution to k is obtained numerically and negligible corrections are done to ensure energy conservation is exact. Such a rapidity dilation preserves a flat rapidity plateau by stretching or shrinking all rapidity differences like a accordion (hence the name "accordion" model). Lorentz covariance is therefore maintained in fragmentation.

Figure (((REF))) shows a schematic illustration of how the default PYTHIA 8.316 string fragmentation algorithm implements hadronisation, and Figure (((REF))) shows a similar diagram for the accordion model.

5.2 Pseudocode

Algorithm (((REF))) shows the FRAGMENT algorithm used to implement the accordion model of $q\bar{q}$ hadronisation in PYTHIA. The flag `StringFragmentation:accordionModel` can be toggled on to enable this model. In this algorithm, the main fragmentation loop is stopped if there is insufficient energy left in the string to produce a hadron, or if, after producing a hadron, the remaining string energy W_{rem} is less than a W_{min} value determined according to algorithm WMIN. The `StringFragmentation:stopMass` and `StringFragmentation:stopSmear` parameters are still used to set the mean and smearing of W_{min}, but `StringFragmentation:stopNewFlav` is unused since a new string break has not been selected at this stage.

Note that the flavour selection in the string breaking has been adjusted such that the $q\bar{q}$ flavour is selected only once, rather than being re-selected every time COMBINE fails. The previous behaviour violated Lorentz covariance, because the flavour re-selection introduced a bias towards hadrons that did not sometimes fail in COMBINE [1].

Upon stopping the fragmentation loop, the final two hadrons are created according to Algorithm (((REF))). Firstly, a new string break is selected according to the Schwinger mechanism if necessary. The transverse momenta of the final two hadrons can then be calculated. Next, the hadron species and masses are selected according to the quark flavours, in the same manner as the rest of the hadrons from the fragmentation loop.

The transverse masses can now be calculated, and these in turn are used to select trial z fractions for each hadron according to the ZFRAG procedure. Then, equation (2.18) is used to calculate required rapidity spacings between the final two hadrons and their neighbours. Finally, the new hadrons and hadron jets created from the positive and negative string ends are boosted in order to obtain these rapidity spacings.

At this point, the event contains a set of hadrons with rapidity spacings and species selections that are identically distributed, assuming that (2.18) gives an accurate distribution of rapidities. This turns out to not be the case, for reasons explained in Section 5.4, but the distributions are approximately similar enough to provide a flat rapidity plateau, especially with a good selection of `stopMass` and `stopSmear`.

A global rapidity $y \to ky$ is then applied to restore energy conservation in a Lorentz covariant manner. First, the event record is boosted to the centre-of-mass frame. Then, k is selected to solve the equation

$$\sum_i \sqrt{m_i^2 + p_{x,i}^2 + p_{y,i}^2 + p_{z,i}^2} = E_{CM}, \tag{5.1}$$

where i runs over all hadrons and E_{CM} is the centre-of-mass energy. This equation does not in general have closed-form solutions, so a numerical solution is obtained using the

[1] Failures of COMBINE are intended effects to produce the hadron species distributions enforced by the `StringFlav` parameters.

`brent` function in PYTHIA, which implements Brent's method [48] to solve equation (5.1).

After applying the $y \to ky$ rescaling, the system is boosted again to the centre-of-mass frame. Since k is calculated numerically, energy conservation will not be exact, and to make it exact, the (negligible) difference in energy is subtracted (or added) to all hadrons proportionally to their longitudinal momentum. Before these corrections, the accordion rescaling is by default repeated twice [2] (((CHECK))) to minimise the impact of the corrections on Lorentz covariance.

5.3 Results

Having developed the accordion model and elaborated on its algorithmic implementation in PYTHIA, we can now move on to the results. An ideal selection of `stopMass` and `stopSmear` to match distributions of hadron multiplicity with the existing tunes and obtain a flat rapidirty plateau was determined through trial and error. Figure (((REF))) shows how the accordion model performs compared to PYTHIA 8.316 defaults with the Monash tune for 200 GeV $d\bar{d}$ hadronisation in terms of the rapidity plateau, hadrochemistry SSE, and `finalTwo` failure rate. The results show a flat rapidity plateau away from the endpoints, and a SSE that is negligible compared to PYTHIA defaults, indicating that Lorentz covariance is preserved by the accordion model.

Figure (((REF))) shows the rapidity plateau, SSE, and failure rate for $d\bar{d}$ hadronisation using the accordion model for the Monash tune across string masses of 50, 100, 200, and 500 GeV. The flat rapidity plateau in the central string region is preserved at all these string masses, although for 25 and 50 GeV this central region is vanishingly small and endpoint effects dominate.

Figure (((REF))) show the same results across the JETSET, Monash, and Hoeth tunes. For each tune, a different selection of `stopMass` and `stopSmear` are used. These plots show that the accordion model can preserve Lorentz covariance and attain a flat rapidity plateau with negligible SSE for a variety of fragmentation function and flavour/species selection parameters.

Figure (((REF))) shows the hadron multiplicity distributions for 200 GeV $d\bar{d}$ hadronisation for PYTHIA defaults and the accordion model in the Monash tune. The distributions are closely matched, indicating that there is no deviation in this observable for the selections of stopping parameters used.

Figure (((REF))) shows the rapidity plateaus in 200 GeV $d\bar{d}$ hadronisation for four different species of hadrons (combined with the corresponding antiparticle) — π^0, Δ^0, K^0, and p. Unlike the overall rapidity plateaus, these are not exactly flat, which is again because equation (2.18) does not exactly match the rapidity spacings of regular hadrons. However, compared to Figure (((REF))), they are significantly flatter than the equivalent plots using PYTHIA defaults.

Figure (((REF))) shows the distributions of the rapidity spacing Δy between hadrons adjacent in rank for regular hadrons, and between the final two hadrons and their neighbours. The distributions are significantly more similar than those shown in Figure (((REF))) for default PYTHIA fragmentation, but still do not exactly match, reflecting the inaccuracy of equation (2.18).

[2] This can be adjusted in the code using a constant

It is also worth mentioning the effect of the accordion join on the z fractions, as calculated based on the final set of hadrons. For the default PYTHIA implementation of string fragmentation, the z fractions are identical for regular hadrons, and significantly different for the final hadrons selected before the joining step, as shown in Figure (((REF))). This is expected, since in this algorithm the z fractions are identically selected for all regular hadrons and are not adjusted afterwards — the final two hadrons are then selected with different kinematics, causing the deviation in the final z fractions.

Figure (((REF))) shows the retrospective z fractions of hadrons produced using the accordion model. Unlike the default PYTHIA model, the accordion model leads to different z fractions being selected for hadrons towards the centre of the string, with a significant increase in very small z fractions. This does not necessarily imply a violation of Lorentz covariance. The identical and independent distribution of z fractions is a consequence of Lorentz covariance in the asymptotic limit of infinite remaining string energy and fragmentation solely in one direction, and this property is still approximately maintained in the centre of of the string, despite the odd shape of the distribution. There is no longer a significant deviation in z fractions adjacent to the final two hadrons, and the deviation is limited to the string endpoints, which is more in accordance with the physical properties of the string.

In all cases, the hadronic chemistry SSE is negligible, especially compared to the SSE from PYTHIA defaults. Reflecting this, the number of failures in `finalTwo` is two orders of magnitude lower, and isolated to the case where two diquarks have to be joined. These results show a staggering improvement moving from the default PYTHIA implementation of `finalTwo` to the accordion model of string fragmentation. Lorentz covariance is approximately restored in the Monash tune, with a flat rapidity plateau and negligible SSE. Furthermore, the accordion model can provide a flat rapidity plateau and negligible SSE across different tunes with some adjustments to the stopping parameters, which would ideally be integrated into the tunes themselves. It is apparent that the global rapidity rescaling used to enforce energy conservation in the accordion model is successful in preserving Lorentz covariance — the deviations from Lorentz covariance stem from the differences in Δy along the string shown in Figure (((REF))).

5.4 Limitations

Despite the success of the accordion model in preserving Lorentz covariance, there are a number of limitations. Firstly, the Lorentz covariance of the current PYTHIA implementation of the accordion model is only approximate, as evidenced by the need to tune `stopMass` and `stopSmear` for a flat plateau, as well as the deivations from flatness in the species-specific rapidity plateaus. This occurs because the combination of a trial zFrag selection and equation (2.18) does not provide a distribution of Δy that is identical to that of regular hadrons. Equation (2.18) is derived in the situation where fragmentation is left-to-right or right-to-left, but not a combination of both. For any two rank-adjacent hadrons produced in fragmentation, there could have been a number of hadrons fragmented from the other end between their creation, which would affect the energies and momenta of those two hadrons and therefore their rapidity spacing. Ideally, an implementation of the accordion model would accurately reproduce the effect of randomly fragmenting from either end on the distribution of Δy in its selection of rapidity spacings for the final two hadrons and their neighbours. This is one potential future extension of the model.

Furthermore, the accordion model has only been implemented for hadronisation of a

single $q\bar{q}$ string, with the popcorn model of baryon production switched off. In an actual PYTHIA event, there are dozens of such processes, which often involve more complicated topologies including gluon kinks, loops, and string junctions. It is not immediately obvious how the property of Lorentz covariance or the global energy conservation of the accordion model would apply to these situations, and further development of analytic models as well as practical implementations is required in order to produce full $p\bar{p}$ collision events using the accordion model.

The tests carried out in the scope of this project to verify the Lorentz covariance of the accordion model are limited. Future studies could compare observables between events generated using the accordion model and LHC data. Particle-particle correlations should also be more thoroughly investigated to ensure Lorentz covariance is preserved beyond just a flat rapidity plateau and negligible SSE. Finally, a thorough analysis of the time complexity and runtime of PYTHIA using the accordion model should be carried out, alongside refactoring the code for better efficiency and performance.

Chapter 6

Summary and Outlook

This project investigated the property of Lorentz covariance in the Lund model of string fragmentation, with a focus on the current implementation in PYTHIA. It was found that in PYTHIA 8.1 and later versions, Lorentz covariance is violated, resulting in a dip in the rapidity plateau and deviations in the hadronic chemistry of the final two hadrons. This violation is inherent in the hadronisation algorithm currently used in PYTHIA, which produces the final two hadrons in a different manner than the rest, and introduces a bias in hadronic chemistry by refragmenting the string when there is insufficient energy to do so.

In the JETSET tune used in earlier versions of PYTHIA, it was possible to tune the `stopMass`, `stopNewFlav`, and `stopSmear` parameters in order to obtain a flat rapidity plateau. However, the changes in fragmentation functions and flavour/species distributions in the newer Hoeth and Monash tunes make tuning for a flat rapidity plateau impossible without significantly worsening the deviations in hadronic chemistry of the final two hadrons.

A new parameter, `probRevertFinal`, was implemented, which provides a probability that the final string break will be undone when there is insufficient energy to create the final two (rather than refragmenting the string). Using this parameter, the effect of tuning for a flat plateau on the final two hadronic chemistry can be somewhat mitigated. However, restoring Lorentz covariance remains impossible even with this expansion to the parameter space.

The accordion model of string fragmentation, which enforces energy conservation by a global, Lorentz covariant dilation in rapidity space, was implemented in PYTHIA and found to approximately preserve Lorentz covariance. It provides a flat rapidity plateau, negligible hadronic chemistry deviations, and a significantly lowered failure rate, across different tunes and string energies.

While the results so far indicate that the accordion model performs well and preserves Lorentz covariance for $q\bar{q}$ hadronisation, full verification of the accordion model will require an expansion of the model to all string topologies. Next, a new tune will have to be created for the accordion model, and finally observables from the model can be compared with LHC data in order to practically test the accordion model. Future research could also involve more thorough testing and checking of particle-particle correlations as well as an analysis of runtime and computational complexity.

Overall, the findings of this project indicate a glaring and longstanding issue in how the Lund model is implemented in the PYTHIA generator, and offer a promising, novel model that manifestly preserves Lorentz covariance. There are numerous new avenues

of investigation opened by this research, including further development, expansion, and testing of the accordion model, as well as an investigation into issues caused by the dip in the rapidity plateau over the past decade.

Bibliography

[1] Particle Data Group collaboration, *Review of particle physics*, *Phys. Rev. D* **110** (2024) 030001.

[2] F. Gross et al., *50 Years of Quantum Chromodynamics*, *Eur. Phys. J. C* **83** (2023) 1125 [2212.11107].

[3] M.D. Schwartz, *Quantum Field Theory and the Standard Model*, Cambridge University Press (3, 2014).

[4] ATLAS collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett. B* **716** (2012) 1 [1207.7214].

[5] CMS collaboration, *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC*, *Phys. Lett. B* **716** (2012) 30 [1207.7235].

[6] R.K. Ellis, W.J. Stirling and B.R. Webber, *QCD and collider physics*, vol. 8, Cambridge University Press (2, 2011), 10.1017/CBO9780511628788.

[7] L. Evans and P. Bryant, *LHC Machine*, *JINST* **3** (2008) S08001.

[8] C. Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, *SciPost Phys. Codeb.* **2022** (2022) 8 [2203.11601].

[9] A. Buckley et al., *General-purpose event generators for LHC physics*, *Phys. Rept.* **504** (2011) 145 [1101.2599].

[10] P. Skands, *Introduction to QCD*, in *Theoretical Advanced Study Institute in Elementary Particle Physics: Searching for New Physics at Small and Large Scales*, pp. 341–420, 2013, DOI [1207.2389].

[11] J.C. Collins and D.E. Soper, *Parton Distribution and Decay Functions*, *Nucl. Phys. B* **194** (1982) 445.

[12] D.J. Gross and F. Wilczek, *Asymptotically Free Gauge Theories - I*, *Phys. Rev. D* **8** (1973) 3633.

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

[14] G.P. Salam, *Elements of QCD for hadron colliders*, in *2009 European School of High-Energy Physics*, 11, 2010 [1011.5131].

[15] B. Andersson, G. Gustafson, G. Ingelman and T. Sjostrand, *Parton Fragmentation and String Dynamics*, *Phys. Rept.* **97** (1983) 31.

[16] G.S. Bali and K. Schilling, *Static quark - anti-quark potential: Scaling behavior and finite size effects in SU(3) lattice gauge theory*, *Phys. Rev. D* **46** (1992) 2636.

[17] B. Andersson, *The Lund Model*, vol. 7, Cambridge University Press (1998), 10.1017/9781009401296.

[18] T. Sjostrand, S. Mrenna and P.Z. Skands, *PYTHIA 6.4 Physics and Manual*, *JHEP* **05** (2006) 026 [hep-ph/0603175].

[19] E. Rutherford, *The scattering of alpha and beta particles by matter and the structure of the atom*, *Phil. Mag. Ser. 6* **21** (1911) 669.

[20] J. Chadwick, *The Existence of a Neutron*, *Proc. Roy. Soc. Lond. A* **136** (1932) 692.

[21] P.A.M. Dirac, *The quantum theory of the electron*, *Proc. Roy. Soc. Lond. A* **117** (1928) 610.

[22] C.D. Anderson, *The Apparent Existence of Easily Deflectable Positives*, *Science* **76** (1932) 238.

[23] M. Thomson, *Modern particle physics*, Cambridge University Press, New York (10, 2013), 10.1017/CBO9781139525367.

[24] Y. Ne'eman, *Derivation of strong interactions from a gauge invariance*, *Nucl. Phys.* **26** (1961) 222.

[25] M. Gell-Mann, *Symmetries of baryons and mesons*, *Phys. Rev.* **125** (1962) 1067.

[26] JADE collaboration, *Observation of Planar Three Jet Events in e+ e- Annihilation and Evidence for Gluon Bremsstrahlung*, *Phys. Lett. B* **91** (1980) 142.

[27] M.E. Peskin and D.V. Schroeder, *An Introduction to quantum field theory*, Addison-Wesley, Reading, USA (1995), 10.1201/9780429503559.

[28] K. Hubner, *Design and construction of the ISR*, in *40th Anniversary of the First Proton-Proton Collisions in the CERN Intersecting Storage Rings (ISR)*, 6, 2012 [1206.3948].

[29] S. Erhan, W.S. Lockman, T. Meyer, J. Rander, P. Schlein, R. Webb et al., *Hyperon production in pp interactions at* $\sqrt{s} = 53$ *and* 62 GeV, *Phys. Lett. B* **85** (1979) 447.

[30] "Design Report Tevatron 1 project." https://lss.fnal.gov/archive/design/fermilab-design-1984-01.pdf, 1984.

[31] D0 collaboration, *Observation of the top quark*, *Phys. Rev. Lett.* **74** (1995) 2632 [hep-ex/9503003].

[32] CDF collaboration, *Observation of top quark production in* $\bar{p}p$ *collisions*, *Phys. Rev. Lett.* **74** (1995) 2626 [hep-ex/9503002].

[33] CERN, "Atlas images gallery." https://home.cern/resources/image/experiments/atlas-images-gallery, 2011.

[34] T. Sjöstrand, *The PYTHIA Event Generator: Past, Present and Future*, *Comput. Phys. Commun.* **246** (2020) 106910 [1907.09874].

[35] T. Sjostrand, *The Lund Monte Carlo for Jet Fragmentation and e+ e- Physics: Jetset Version 6.2*, *Comput. Phys. Commun.* **39** (1986) 347.

[36] G. Corcella, I.G. Knowles, G. Marchesini, S. Moretti, K. Odagiri, P. Richardson et al., *HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)*, *JHEP* **01** (2001) 010 [hep-ph/0011363].

[37] Sherpa collaboration, *Event Generation with Sherpa 2.2*, *SciPost Phys.* **7** (2019) 034 [1905.09127].

[38] P. Skands, S. Carrazza and J. Rojo, *Tuning PYTHIA 8.1: the Monash 2013 Tune*, *Eur. Phys. J. C* **74** (2014) 3024 [1404.5630].

[39] ALICE collaboration, *Enhanced production of multi-strange hadrons in high-multiplicity proton-proton collisions*, *Nature Phys.* **13** (2017) 535 [1606.07424].

[40] ATLAS collaboration, *Measurements of jet cross-section ratios in 13 TeV proton-proton collisions with ATLAS*, *Phys. Rev. D* **110** (2024) 072019 [2405.20206].

[41] J. Altmann and P. Skands, *String junctions revisited*, *JHEP* **07** (2024) 238 [2404.12040].

[42] C. Bierlich, G. Gustafson, L. Lönnblad and A. Tarasov, *Effects of Overlapping Strings in pp Collisions*, *JHEP* **03** (2015) 148 [1412.6259].

[43] C. Bergman. Honours thesis.

[44] J.S. Schwinger, *Gauge Invariance and Mass. 2.*, *Phys. Rev.* **128** (1962) 2425.

[45] P. collaboration, "Pythia 8 online manual." https://pythia.org/manuals/pythia8315/Welcome.html, 2025.

[46] P. Eden, *On energy conservation in Lund string fragmentation*, *JHEP* **05** (2000) 029 [hep-ph/0004132].

[47] S. Mrenna. Private communication.

[48] R.P. Brent, *Algorithms for Minimization without Derivatives*, Prentice-Hall, Englewood Cliffs, New Jersey, 1st ed. (1973).