Predicting Physics Observables from LHC Data

A Study of ROOT-Based Data Structures, Heavy-Flavor Physics,
and Multivariate Classification Techniques

Aryan Gupta
Roll No.: 24B1810
Indian Institute of Technology Bombay

January 7, 2026

Abstract

High-energy physics experiments at the Large Hadron Collider (LHC) produce extremely large and complex datasets that require specialized data formats and analysis techniques. This report presents a systematic study of ROOT-based datasets used in heavy-flavor physics, with particular emphasis on the reconstruction and classification of $D^{*+} \to D^0\pi^+$ candidates. The structure of analysis-level and event-level ROOT files is examined using both C++ ROOT and Python-based tools such as `uproot`. The physics motivation behind relativistic kinematics, heavy-quark production, and decay topology is reviewed in detail. Finally, the role of topological observables and multivariate machine-learning classifiers in separating prompt, non-prompt, and background candidates is discussed from both a physical and statistical perspective.

Contents

1 Introduction

Modern collider experiments operate at energy scales where relativistic effects dominate and quantum field theories are required to describe particle interactions. The Large Hadron Collider (LHC) produces proton–proton collisions at center-of-mass energies of several tera-electronvolts, resulting in the creation of a vast number of particles per second. To handle this data volume efficiently, the high-energy physics community relies on the ROOT framework, which provides optimized storage, compression, and analysis capabilities.

The goal of this work is to understand how physics observables can be extracted from LHC data by first analyzing the internal structure of ROOT files and then connecting this structure to the underlying particle physics. Special attention is given to heavy-flavor mesons, whose displaced decay topologies make them ideal candidates for multivariate classification methods.

2 Understanding ROOT Data Structures

2.1 Analysis-Level ROOT Files

Analysis-level ROOT files typically store aggregated results rather than event-wise information. Such files are organized into task-based directories, each corresponding to a specific step in the reconstruction or analysis workflow. The stored objects are predominantly histograms (TH1 and TH2), representing distributions of reconstructed quantities such as impact parameters and momentum components.

The absence of `TTree` objects in these files implies that the data cannot be directly interpreted as a table of events. Instead, each histogram represents a projection of the data after applying specific selections and reconstruction algorithms.

2.2 Event-Level ROOT Files and TTrees

In contrast, event-level or candidate-level datasets are stored using `TTree` objects. A TTree can be viewed as a columnar table where each entry corresponds to one reconstructed candidate and each branch corresponds to a physical or derived variable.

The columnar nature of TTrees allows efficient access to individual variables without loading the entire dataset into memory. This design is particularly well suited for large-scale machine-learning applications, where selective feature loading is essential.

2.3 Python-Based Inspection Using `uproot`

The `uproot` library enables ROOT files to be accessed directly in Python without requiring the ROOT C++ environment. Using `uproot`, TTrees can be converted into pandas DataFrames, allowing seamless integration with modern data-science and machine-learning workflows. For the dataset considered here, the resulting table contains hundreds of thousands of candidates and several dozen features per candidate.

3 Relativistic Kinematics in High-Energy Physics

3.1 Motivation for High Energies

The spatial resolution of a probe is limited by its de Broglie wavelength,

$$\lambda = \frac{h}{p}. \tag{1}$$

To probe distances on the order of a femtometer, momenta in the GeV range are required. High energies also enable the creation of massive particles through the relation

$$E = mc^2. \tag{2}$$

3.2 Four-Vectors and Lorentz Invariance

Relativistic dynamics is naturally formulated in terms of four-vectors. The energy–momentum four-vector is defined as

$$p^\mu = \left(\frac{E}{c}, \vec{p}\right), \tag{3}$$

with invariant magnitude

$$p_\mu p^\mu = m^2 c^2. \tag{4}$$

This leads to the familiar relativistic energy–momentum relation

$$E^2 = p^2 c^2 + m^2 c^4. \tag{5}$$

3.3 Transverse Momentum and Rapidity

In collider experiments, the beam direction defines a preferred axis. The transverse momentum,

$$p_T = \sqrt{p_x^2 + p_y^2}, \tag{6}$$

is invariant under boosts along the beam direction and therefore plays a central role in data analysis.

Rapidity is defined as

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right), \tag{7}$$

and has the useful property of being additive under Lorentz boosts.

4 Heavy Quarks and Heavy-Flavor Mesons

4.1 Charm and Bottom Quarks

Charm and bottom quarks are significantly heavier than the QCD scale Λ_{QCD}. They are produced predominantly in hard partonic scatterings during the early stages of proton–proton collisions. Because of their large masses, heavy quarks provide a clean testing ground for perturbative QCD calculations.

4.2 Prompt and Non-Prompt Production

Heavy-flavor mesons are classified according to their production mechanism. Prompt mesons originate directly from the hadronization of heavy quarks produced at the primary vertex. Non-prompt mesons arise from the decay of longer-lived parent hadrons, such as B mesons, and therefore exhibit displaced decay vertices.

This distinction forms the basis for topological separation techniques.

5 Experimental Reconstruction of D^* Mesons

The D^{*+} meson is reconstructed through the decay chain

$$D^{*+} \to D^0 \pi^+_{\mathrm{soft}}, \quad D^0 \to K^- \pi^+. \tag{8}$$

The invariant mass of the parent particle is reconstructed from the four-momenta of its decay products:

$$m^2 = \left(\sum_i E_i \right)^2 - \left| \sum_i \vec{p}_i \right|^2. \tag{9}$$

Signal candidates form a narrow peak in the invariant-mass distribution, while random combinations produce a smooth combinatorial background.

6 Decay Topology and Discriminating Variables

Particles with finite lifetimes decay at measurable distances from the primary vertex. This displacement is quantified using several topological observables.

6.1 Pointing Angle

The cosine of the pointing angle is defined as

$$\cos\theta = \frac{\vec{L}\cdot\vec{p}}{|\vec{L}||\vec{p}|}, \tag{10}$$

where $\vec{L}$ connects the primary and secondary vertices. Values close to unity indicate a decay consistent with the reconstructed momentum direction.

6.2 Decay Length and Impact Parameters

The transverse decay length,

$$L_{xy} = |\vec{r}_{\mathrm{SV}} - \vec{r}_{\mathrm{PV}}|_{xy}, \tag{11}$$

is particularly effective in distinguishing prompt and non-prompt decays.

Impact parameters quantify how strongly daughter tracks deviate from the primary vertex and provide complementary information.

7 Machine Learning Classification

Multivariate classifiers are employed to separate prompt, non-prompt, and background candidates. Only topological variables are used as input features to avoid biasing physics observables such as invariant mass or transverse momentum.

The classifier outputs a continuous score interpreted as the probability of a candidate being prompt. Receiver Operating Characteristic (ROC) curves are used to evaluate performance and to select a physics-motivated working point balancing efficiency and background rejection.

8 Conclusion

This report presented a comprehensive overview of how physics observables can be extracted from LHC data by combining a detailed understanding of ROOT data structures with relativistic kinematics, heavy-flavor physics, and machine-learning techniques. Topological observables derived from decay geometry provide strong discrimination power

and form the foundation of modern heavy-flavor analyses. The methods discussed here closely mirror those used in contemporary experimental analyses at the LHC.