Predicting Physics Observables by Analysing LHC Data

Nissim Sahoo
Roll No.: 24B1818

Mentor: Deependra Sharma

WiDS, Analytics Club, IIT Bombay

February 1, 2026

Abstract

This report documents the analysis workflow for predicting physics observables from LHC data. It covers ROOT file inspection, reconstruction of heavy-flavor mesons, particle identification, and the implementation of a physics-aware machine learning classifier.

1 Introduction

High-energy physics experiments at the Large Hadron Collider (LHC) generate large-scale datasets containing reconstructed particle candidates. These data are stored using the ROOT framework and analysed to extract physical observables such as production mechanisms and decay properties.

The goal of this project is to build a physics-informed machine learning pipeline to distinguish prompt, non-prompt, and background D^* meson candidates.

2 ROOT Data Structure

2.1 Aggregated Analysis Output

The file `AnalysisResults67.root` contains task-wise directories holding histogram objects (`TH1`, `TH2`). No `TTree` objects are present, indicating that this file is not suitable for candidate-level machine learning.

2.2 Candidate-Level ROOT Files

The files

- `Prompt_DstarToD0Pi.root`
- `NonPrompt_DstarToD0Pi.root`
- `Background_DstarToD0Pi.root`

each contain a single `TTree` named `treeMLDstar`, where each row corresponds to a reconstructed D^* candidate.

3 Physics Background

3.1 Relativistic Kinematics

- Energy-momentum relation:
$$E^2 = p^2c^2 + m^2c^4$$

- Transverse momentum:
$$p_T = \sqrt{p_x^2 + p_y^2}$$

- Pseudorapidity:
$$\eta = -\ln \tan \frac{\theta}{2}$$

3.2 Heavy-Flavor Mesons

- $D^{*+} \to D^0\pi^+_{\text{soft}}$
- $D^0 \to K^-\pi^+$
- Prompt production: direct charm
- Non-prompt production: $B \to D^* + X$

4 Experimental Reconstruction

4.1 Invariant Mass

$$m^2 = (E_1 + E_2)^2 - (\vec{p}_1 + \vec{p}_2)^2$$

4.2 Topological Variables

- Decay length L_{xy}
- Impact parameters
- Pointing angle:
$$\cos\theta_p = \frac{\vec{p}_D \cdot \vec{L}}{|\vec{p}_D||\vec{L}|}$$

5 Particle Identification (PID)

Particle identification exploits detector responses:

- TPC:
$$n\sigma_{\text{TPC}} = \frac{(dE/dx)_{\text{meas}} - (dE/dx)_{\text{exp}}}{\sigma_{dE/dx}}$$

- TOF:
$$n\sigma_{\text{TOF}} = \frac{t_{\text{meas}} - t_{\text{exp}}}{\sigma_t}$$

PID variables are included as ML features, while invariant mass and p_T are excluded to avoid sculpting physical distributions.

6 Machine Learning Framework

6.1 Feature Vector

$$x = (\text{topology},\ \text{PID},\ \text{geometry})$$

Target labels:

$$y \in \{0 = \text{prompt},\ 1 = \text{non-prompt},\ 2 = \text{background}\}$$

6.2 Softmax Classifier

$$p_i(x) = \frac{e^{z_i}}{\sum_{j=0}^{2} e^{z_j}}$$

6.3 Cross-Entropy Loss

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{n=1}^{N} \log p_{y_n}(x_n)$$

7 Physics-Aware Loss

Background candidates should not be classified as prompt. A physics penalty is introduced:

$$\mathcal{L}_{\text{phys}} = \frac{1}{N_{\text{bkg}}}\sum_{n:\, y_n=2} p_0(x_n)$$

Total loss:

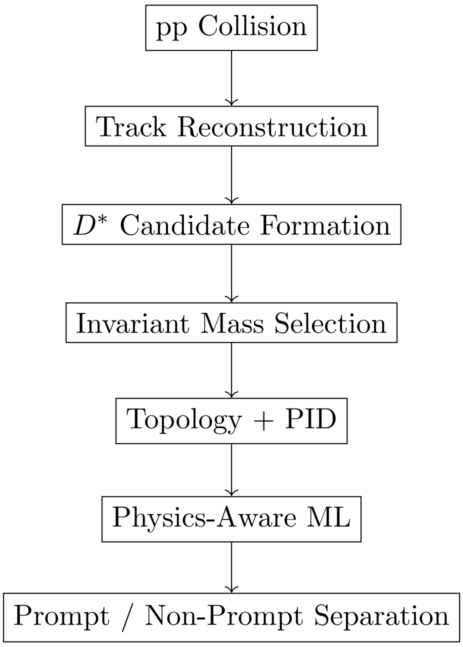
$$\boxed{\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}}$$

8 Model Evaluation

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

ROC curves are constructed by varying the classifier threshold, and performance is summarized using the Area Under Curve (AUC).

9 Analysis Workflow



10 Conclusion

A complete physics-informed machine learning workflow was developed to analyse LHC D^* meson data. The integration of PID variables and a physics-aware loss ensures robust classification while preserving physical consistency.