Wids
Project
2025

Predicting Physics Observables at the LHC using Machine Learning on Advanced Collider data

By

Ankit Kumar

25N0292

M.Sc Physics

Mentor: Deependra Sharma

Abstract

In high-energy physics analyses involving heavy-flavor hadrons, separating prompt signals from non-prompt contributions and background noise signals is a critical task. Traditional cut-based approaches often fail to capture complex correlations among topological observables, motivating the use of multivariate techniques.

In this project, we study the classification of candidates into Prompt, Non-Prompt, and Background signal using Boosted Decision Trees (XGBoost) and Neural Networks. The analysis is performed in exclusive transverse momentum (pT) intervals, ensuring pT-dependent modeling and avoiding kinematic bias. Only topological variables are used for training, while kinematic variables are reserved for validation and post-training studies.

This study demonstrates that topology-based machine-learning techniques provide a robust and unbiased approach for heavy-flavor signal classification and establishes a solid foundation for future extensions incorporating particle identification information and application to real experimental data.

Introduction

Ultra-relativistic heavy-ion collisions at modern collider facilities such as the Large Hadron Collider (LHC) and the Relativistic Heavy-Ion Collider (RHIC) provide a unique opportunity to study strongly interacting matter under extreme conditions of temperature and energy density. These experiments aim to recreate, for fleeting moments, the quark–gluon plasma (QGP), a deconfined state of quarks and gluons believed to have existed in the early Universe. Since the QGP is extremely short-lived, its properties cannot be measured directly and must instead be inferred through a variety of experimental observables.

A major experimental challenge in heavy-ion and proton–proton collision analyses is the reliable separation of genuine signal particles from large combinatorial backgrounds, as well as the disentanglement of different production mechanisms. For example, inclusive J/ψ or D^0 yields receive contributions from both prompt production, originating directly at the primary interaction vertex, and non-prompt production, arising from the weak decays of longer-lived beauty hadrons. Traditionally, such separation has relied on template fitting techniques using invariant mass distributions and decay-length-based observables, modeled as a linear combination of expected distributions (“templates”) corresponding to different physical sources. For each class—Prompt, Non-Prompt, and Background—a template is constructed using Monte Carlo simulations or data-driven control samples. These templates represent the expected shape of the observable for each component and are usually normalized to unit area.

The measured data distribution is then expressed as:

$$D(x) = f_{\text{prompt}}\, T_{\text{prompt}}(x) + f_{\text{nonprompt}}\, T_{\text{nonprompt}}(x) + f_{\text{background}}\, T_{\text{background}}(x),$$

where $T_i(x)$are the template shapes and $f_i$are the corresponding fractions, constrained such that $f_{\text{prompt}} + f_{\text{nonprompt}} + f_{\text{background}} = 1$. The fractions $f_i$are determined by fitting the sum of templates to the data using a maximum likelihood or χ^2 minimization procedure.

While effective, these methods are often statistically limited, computationally expensive, and constrained in their ability to perform fine-grained, multidimensional analyses.

In recent years, machine learning (ML) techniques have fundamentally transformed the way data are analyzed in heavy-ion and high-energy physics. By learning complex, nonlinear correlations among multiple kinematic and topological observables, ML algorithms provide a powerful alternative to conventional cut-based or template-based approaches. ML-based approaches allow for faster, more flexible, and more robust extraction of physics observables, while reducing reliance on model-dependent fitting procedures. As heavy-ion experiments move toward increasingly high-precision measurements and larger datasets, machine learning is expected to play an essential role in advancing our understanding of QCD matter under extreme conditions.

Motivated by these developments, the present project explores the application of machine-learning techniques to heavy-flavor classification, with particular emphasis on topology-based and kinematic observables using XGBoost and Neural Networks. By building upon recent advances in ML-driven analyses, this work aims to demonstrate how modern data-driven methods can enhance signal identification, improve background rejection, and ultimately deepen our insight into heavy-ion collision dynamics.

Background

D^0 mesons are unstable particles created during high ion collisions and decays into:

$$D^0 \rightarrow K + \pi$$

D^0 itself is not seen. Only its decay products are seen. It is inferred by: finding two tracks (K and π), checking if they come from the same point (SV) or that point is displaced from PV and checking if their combined mass matches D^0.

For every real D^0 there may be hundreds of fake combinations, two random tracks that *happen* to cross tracks from different particles, this is background noise.

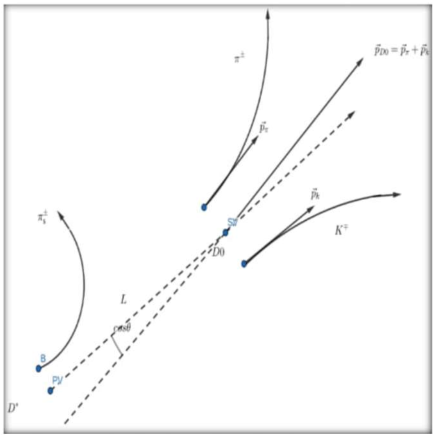
D^0 that is detected at the primary vertex (the point where collision takes place) is termed as prompt signal. Apart from D^o , a bottom quark is produced at the PV, which forms a B meson. The B meson decays as:

$$B \rightarrow D^0 + X$$

When this D^o decays, a non-prompt signal is detected.

In this project, we will be using the following important topological variables (with column names in dataset) for the classification model training:

1. Cosine of pointing angle (fCpaD0) : Related to reconstructed momentum vector. The sum of momentum vec of daughters points towards PV
2. Transverse pointing (fCpaXYD0): Related to reconstructed momentum vector in XY plane
3. Distance from the PV (fDecayLengthXYD0)
4. Product of Daughter tracks' impact parameters (fImpactParameterProductD0)
5. Impact parameter of soft pion (from D*) (fImpactSoftPi)
6. Maximum normalized impact parameter difference (fMaxNormalisedDeltaIPD0)



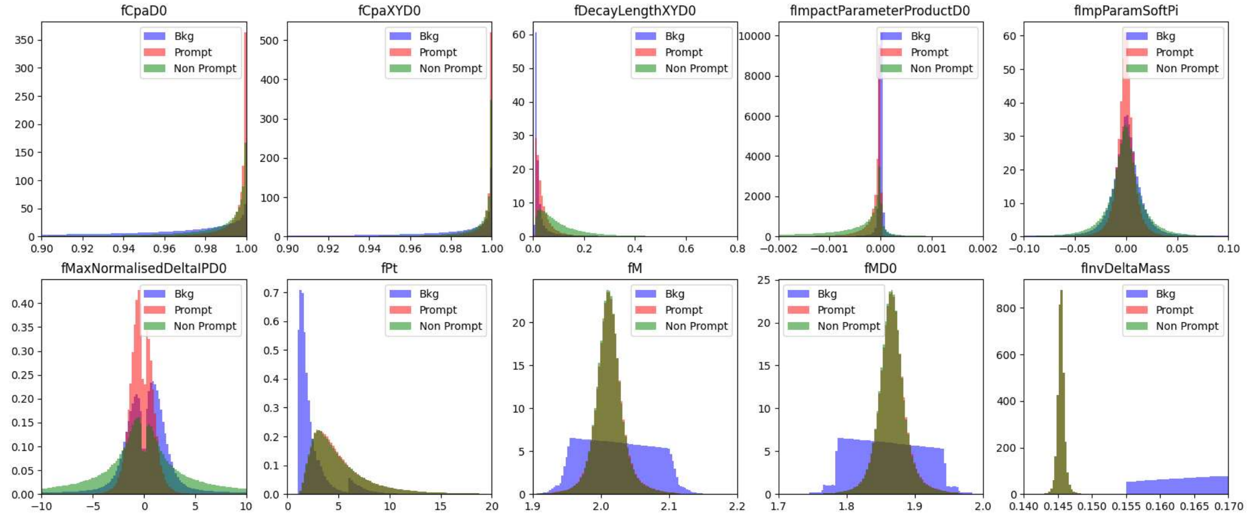
The following kinematic variables (with column names in dataset) useful for binning and verification of the three classes:

1. Transverse momentum of D⁰ (fPt)
2. Invariant mass of reconstructed candidate (fM)

$$M = \sqrt{E^2 - |\vec{p}|^2}$$

3. Invariant mass with D⁰ mass hypothesis (fMD0)
4. Invariant mass difference (D*) (fInvDeltaMass)

$$\Delta M = M(K\pi\pi_{\text{soft}}) - M(K\pi)$$



Workflow

1. Dataset Description

Three ROOT files are used in this analysis:

- **Prompt_DstarToD0Pi.root**
- **Nonprompt_DstarToD0Pi.root**
- **Bkg_DstarToD0Pi.root**

Each file contains reconstructed D* candidates stored in a common TTree structure. The datasets correspond to:

Class	Description
Prompt	D mesons produced directly in the primary interaction
Non-Prompt	D mesons from weak decays of B hadrons
Background	Combinatorial background candidates

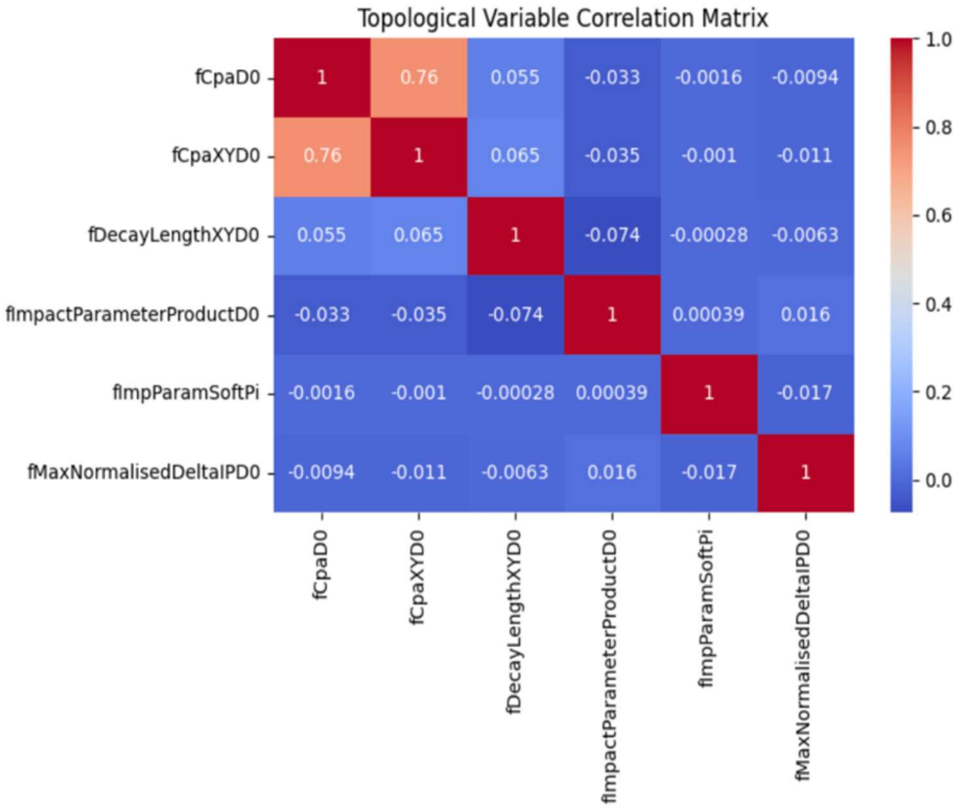
The ROOT files are accessed using the **uproot** Python library, enabling ROOT-independent data analysis. Further these are converted into Pandas data frames.

2. Feature Selection

3.1 Topological Variables (Used for Training)

Only topology-related observables are used as inputs to the BDT to ensure that the classifier learns genuine decay topology rather than kinematic differences. These variables encode information about decay geometry, pointing angles, and impact parameters, which are known to be powerful discriminants between signal and background.

3.2 Kinematic Variables (Used Only for Plotting)



3. pT Binning Strategy

To study the pT dependence of the classification performance and avoid kinematic bias, the analysis is performed in exclusive pT intervals defined by:[1, 1.5, 2, 3, 4, 6, 8, 10, 12, 16, 24] GeV/c. Here, we have chosen [4,6] GeV/c bin.

4. Train–Test Split

The binned dataset is first divided into **training** and **testing** samples using a **70% / 30% split**:

- Training fraction: 0.7
- Testing fraction: 0.3

The number of instances for each class must be equal in the training set in order to avoid any bias towards the majority class. To ensure this, the trained data is oversampled for whichever class with lesser instances. We have avoided under sampling so as not to lose data.

5. XGBoost Model

Boosted Decision Trees (BDT)

Boosted Decision Trees are chosen due to their:

- Ability to capture non-linear correlations
- Robust performance with limited feature sets
- Long-standing validation in heavy-flavor analyses

The **XGBoost** implementation is used, configured for **multi-class classification** with three output classes (Background, Prompt, Non-Prompt).

Hyperparameter Tuning Strategy

Random search method is used for the tuning of the hyperparameters.

Parameter Grid:

```
 "max_depth":        [2, 4,6],
"learning_rate":    [0.3,0.5,0.7],
"n_estimators":     [400,600,1000],
"subsample":        [0.7, 0.8,1],
"colsample_bynode": [0.5,0.7, 1.0],
"min_child_weight": [6,7,8],
"colsample_bytree": [0.5,0.7, 1.0],
"reg_alpha":        [0,0.001,0.01],
"reg_lambda":       [0,0.001,0.01],
"gamma":            [0, 0.1, 0.5],
```

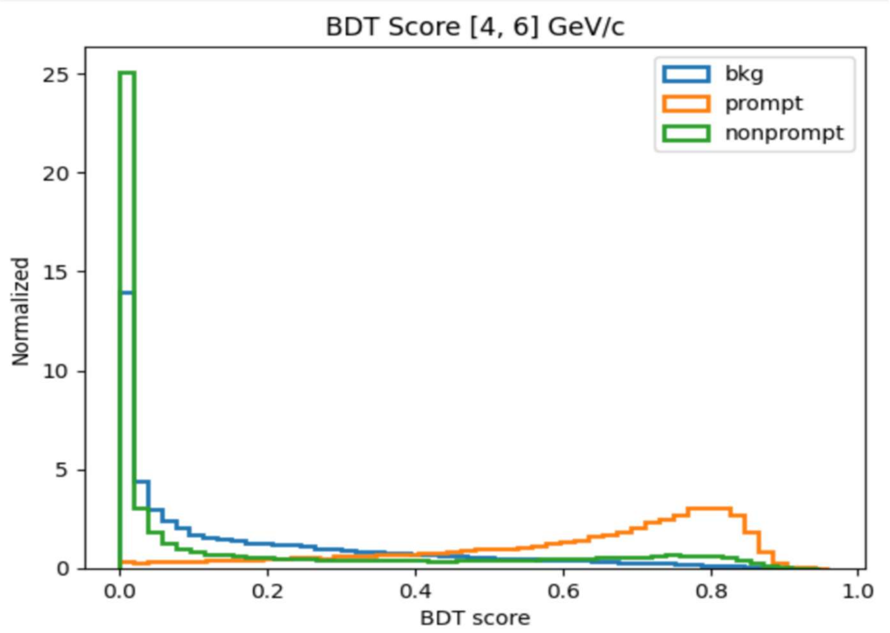
With 5 combinations from the above grid and 8 cross validations, the model with optimum ROC AUC number is choosen. The obtained model has the following hyperparameter

Best Parameters: {'subsample': 0.7, 'reg_lambda': 0, 'reg_alpha': 0, 'n_estimators': 400, 'min_child_weight': 6, 'max_depth': 6, 'learning_rate': 0.3, 'gamma': 0.1, 'colsample_bytree': 1.0, 'colsample_bynode': 0.5}

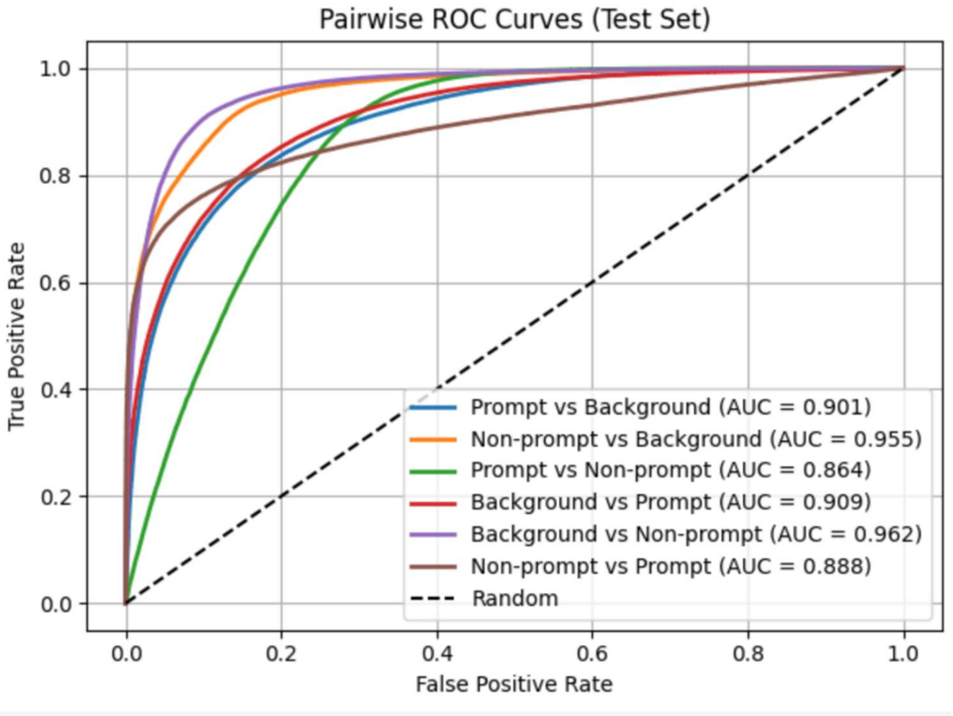
Best Accuracy: 0.920464188024964

Performance Evaluation

BDT Score Distributions



ROC Curves



6. Neural Networks Model

In addition to Boosted Decision Trees (BDTs), a Neural Network (NN)–based classifier was developed as an independent multivariate approach to cross-check the robustness of the results. Neural networks are capable of learning complex non-linear correlations among input variables and provide a complementary perspective to tree-based methods. The NN study was performed using the **same input variables, pT binning strategy, and train–test split** as the BDT analysis to ensure a fair comparison.

Standard Scaling

Unlike BDTs, neural networks are sensitive to the scale of input variables. Therefore, all input features were standardized using **z-score normalization** (zero mean and unit variance). The normalization parameters were derived **only from the training dataset** and subsequently applied to the test dataset to avoid information leakage.

Network Architecture

- An input layer with six nodes (one per topological variable)
- Two hidden layers with ReLU activation
- A softmax output layer with three nodes corresponding to Background, Prompt, and Non-Prompt classes

Training Procedure

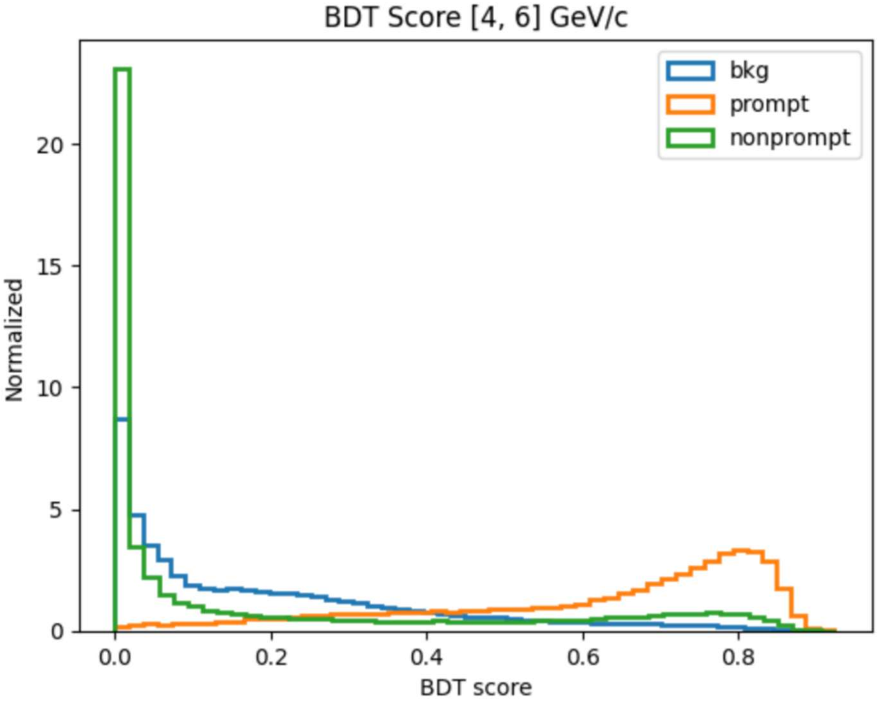
The neural network was trained independently in each pT bin, following the same pT binning scheme used for the BDT analysis. Training was performed using the Adam optimizer with categorical cross-entropy loss in sparse form.

The model was trained only on the training subset and evaluated exclusively on the independent test subset

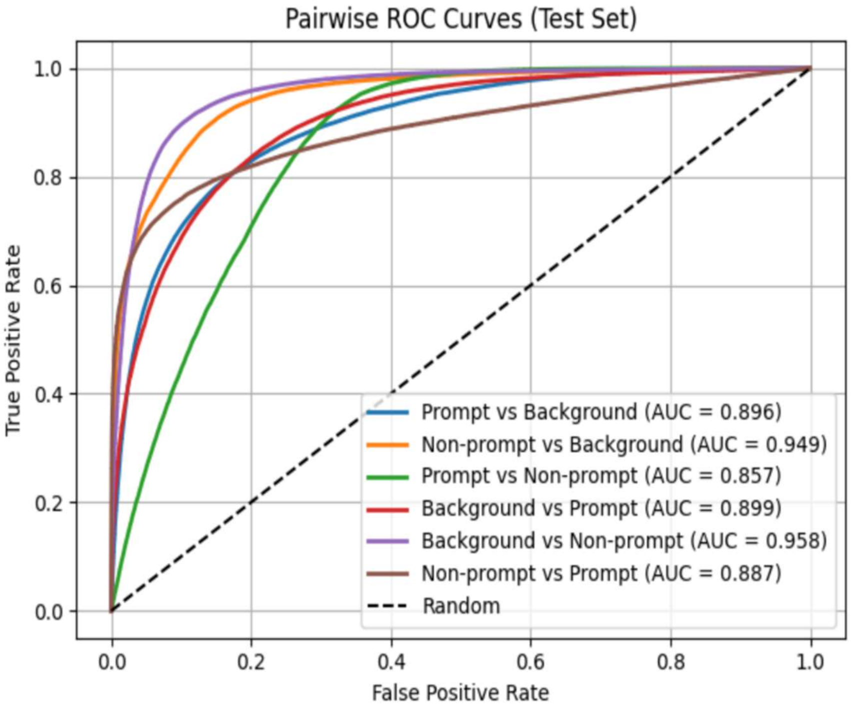
Performance Evaluation

Accuracy: 0.7746   Loss: 0.55

BDT Score Distributions



ROC Curves



Inference

Using an XGBoost Boosted Decision Tree (BDT) trained on the selected topological variables, strong discrimination between signal and background was achieved across the studied pT bins. On the independent test dataset, the pairwise ROC analysis yielded the following representative Area Under the Curve (AUC) values:

- Prompt vs Background: AUC ≈ 0.89–0.90
- Non-Prompt vs Background: AUC ≈ 0.95–0.96
- Prompt vs Non-Prompt: AUC ≈ 0.86–0.88

These results indicate excellent background rejection for both prompt and non-prompt signals, while the separation between prompt and non-prompt candidates remains comparatively weaker. This behavior is physically expected, as both classes correspond to genuine D^0 mesons and differ primarily through subtle displacement-related features.

On the test dataset, the NN classifier showed AUC values of approximately:

- Prompt vs Background: AUC ≈ 0.89
- Non-Prompt vs Background: AUC ≈ 0.94–0.95
- Prompt vs Non-Prompt: AUC ≈ 0.85–0.87

The NN test accuracy was observed to be approximately 0.77–0.78, significantly above the random baseline of ~0.33 for a three-class problem. The close agreement between the NN and BDT performance demonstrates that the observed separation is driven by genuine physical information encoded in the topological variables rather than by model-specific effects.

The ROC curves for both models exhibited smooth behavior without sharp transitions, and consistent performance was observed between training and test datasets, indicating that overtraining and information leakage were successfully controlled.

Conclusion

A robust, pT-differential machine learning framework for heavy-flavor signal classification has been established using topology-based observables. The XGBoost BDT demonstrated strong and stable performance, achieving up to ~90% separation power for Prompt vs Background and ~95% for Non-Prompt vs Background, while maintaining physically realistic separation (~86–88% AUC) between Prompt and Non-Prompt candidates.

The Neural Network classifier provided an independent cross-check of the results, yielding performance consistent with the BDT across all studied pT bins. The agreement between the two models confirms the robustness of the analysis strategy and validates the use of topology-only inputs for multivariate classification.

Overall, this work shows that machine learning techniques can significantly enhance signal extraction in heavy-flavor analyses while maintaining strong physical interpretability and control over biases. The developed framework provides a solid foundation for future extensions, including the incorporation of Particle Identification (PID) variables, application to mixed-class datasets, and eventual deployment on real experimental data. The methodology and results presented here are directly relevant to contemporary heavy-ion and proton–proton collision studies at collider experiments.

References

1. **A short course on Relativistic Heavy Ion Collisions by** A. K. Chaudhuri
2. **Relativistic Kinematics by** Raghunath Sahoo
3. Inclusive, prompt and nonprompt J=ψ identification in proton-proton collisions at the Large Hadron Collider using machine learning by Suraj Prasad ,† Neelkamal Mallick ,‡ and Raghunath Sahoo