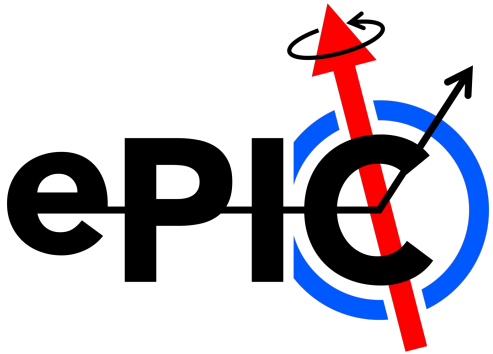
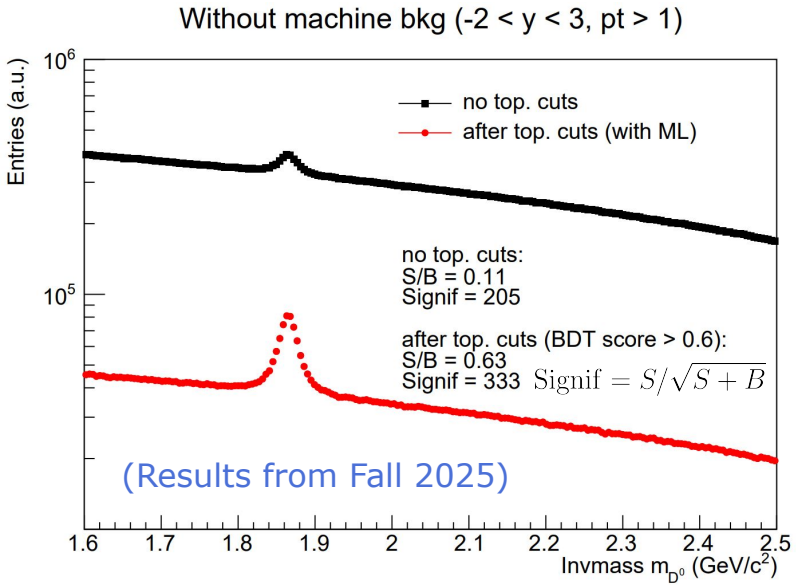
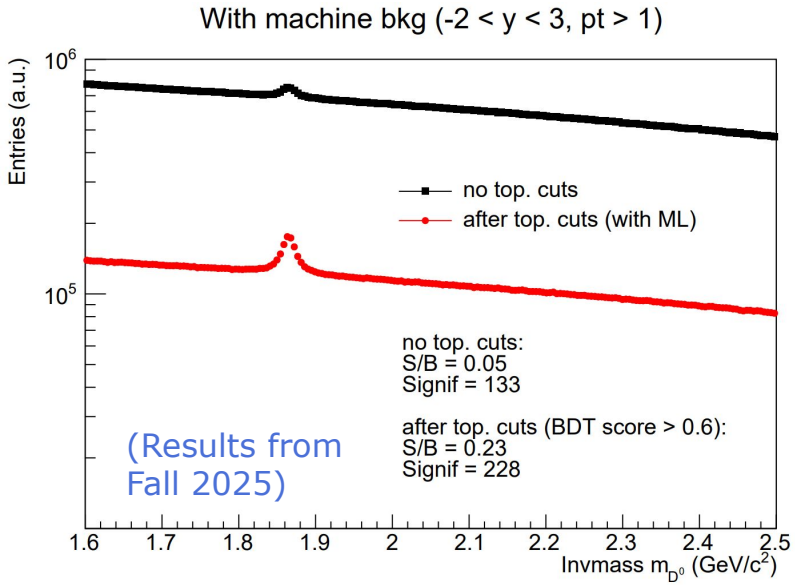
Optimization of boosted decision tree methodology for D0 analysis at the EIC

Connie Yang, Deepa Thomas
ePIC Jets and HF WG Meeting
May 5, 2026



Study of machine-background effects





Fall 20225: Applied existing BDT framework to analyze the impacts of machine-background effects (synchrotron radiation and beam particle interactions)

Spring 2026: Investigate ways to improve BDT framework (for samples without machine-background effects)

Improvements to methodology

Task I: Train BDT separately for different y and pT slices

Task II: Examine the impacts of various feature selections

Task III: Tune BDT hyperparameters

Task I:
Analysis in various y and pT

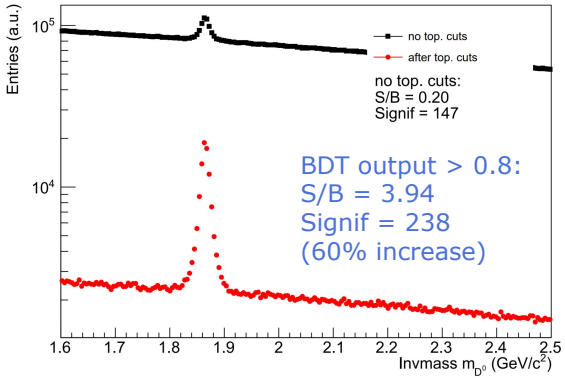
D0 invariant mass in various y and pT slices

mid-y

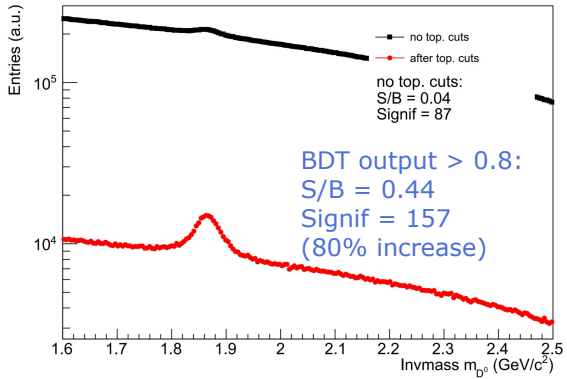
forward y

lower pT

-1 < y < 1, 1 < pT < 2

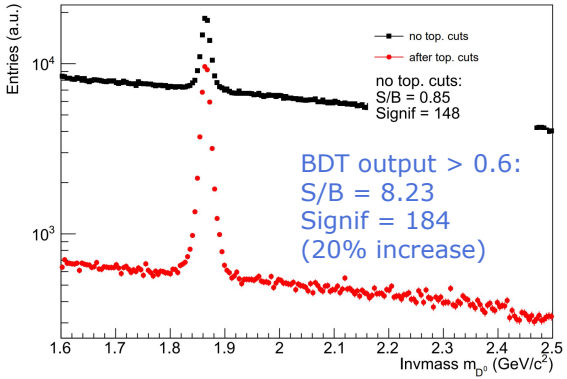


1 < y < 3, 1 < pT < 2

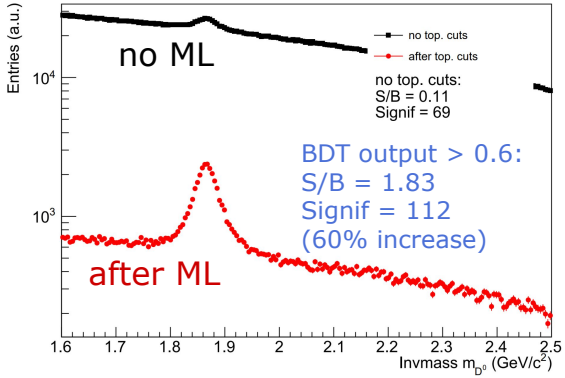


higher pT

-1 < y < 1, 2 < pT < 3



1 < y < 3, 2 < pT < 3

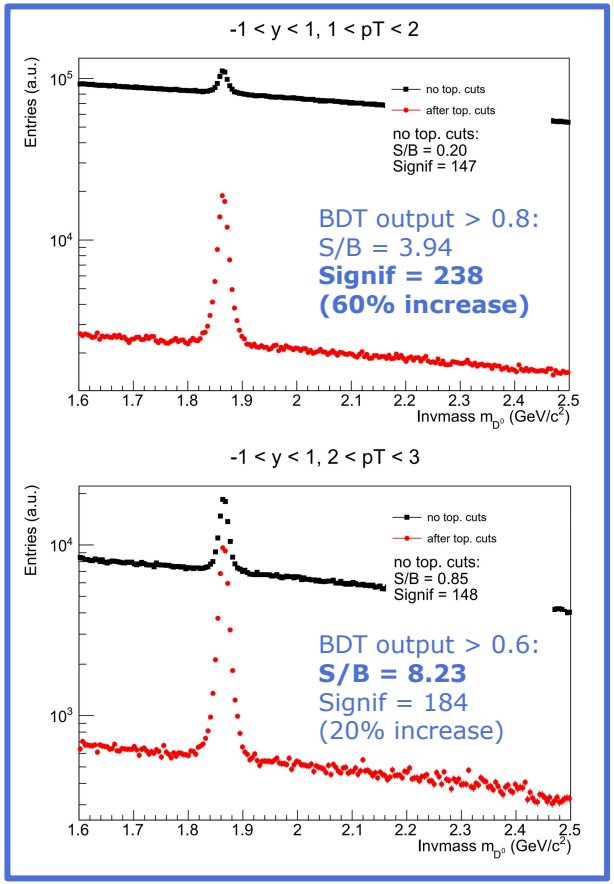


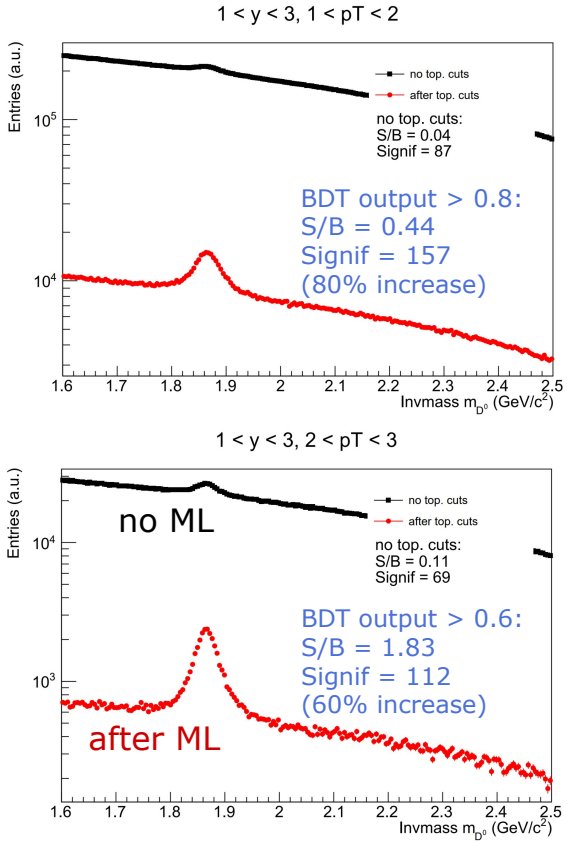
Same features and hyperparameters as preliminary analyses.

Trained separate BDT for each slice.

Optimized each BDT threshold by maximizing Signif.

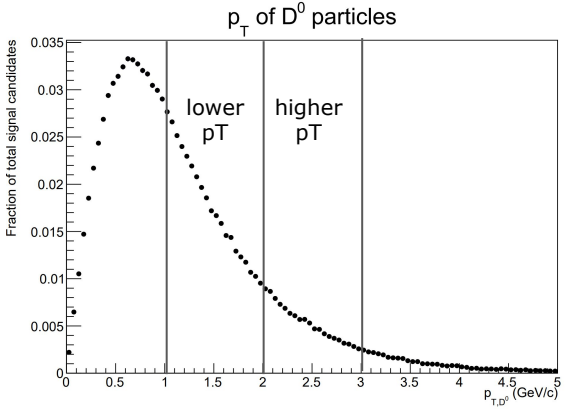
D0 invariant mass: lower pT vs. higher pT



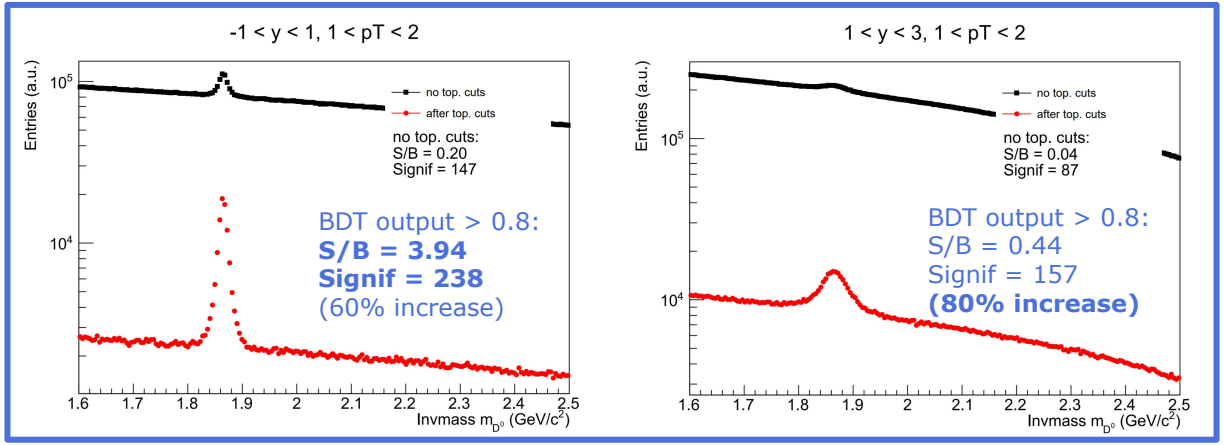


At lower pT, S/B is lower but Signif is higher.

- Worse reconstruction
- Better statistics
- Greater improvement in Signif



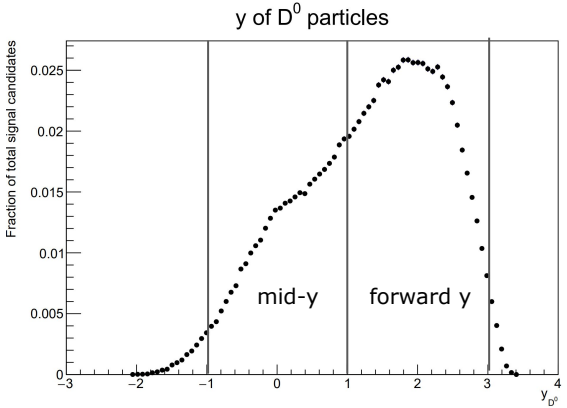
D0 invariant mass: mid-y vs. forward y



At forward y, results are worse but improvement in Signif is greater.

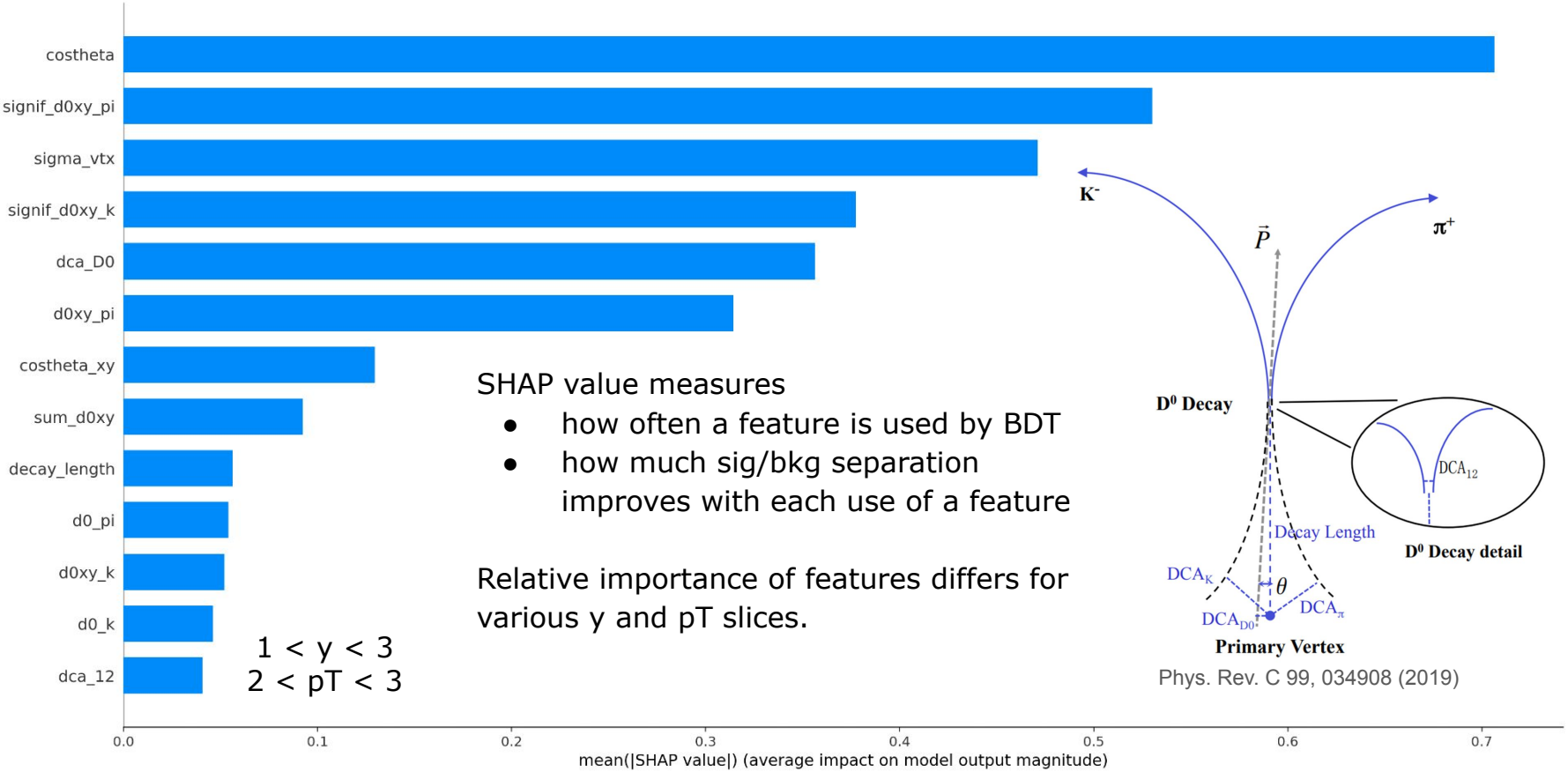
- Large background
- Worse reconstruction

➔ Greatest impact in forward y and lower pT.



Task II:
Examining impacts of feature selection

Relative importance of topological features



SHAP value measures

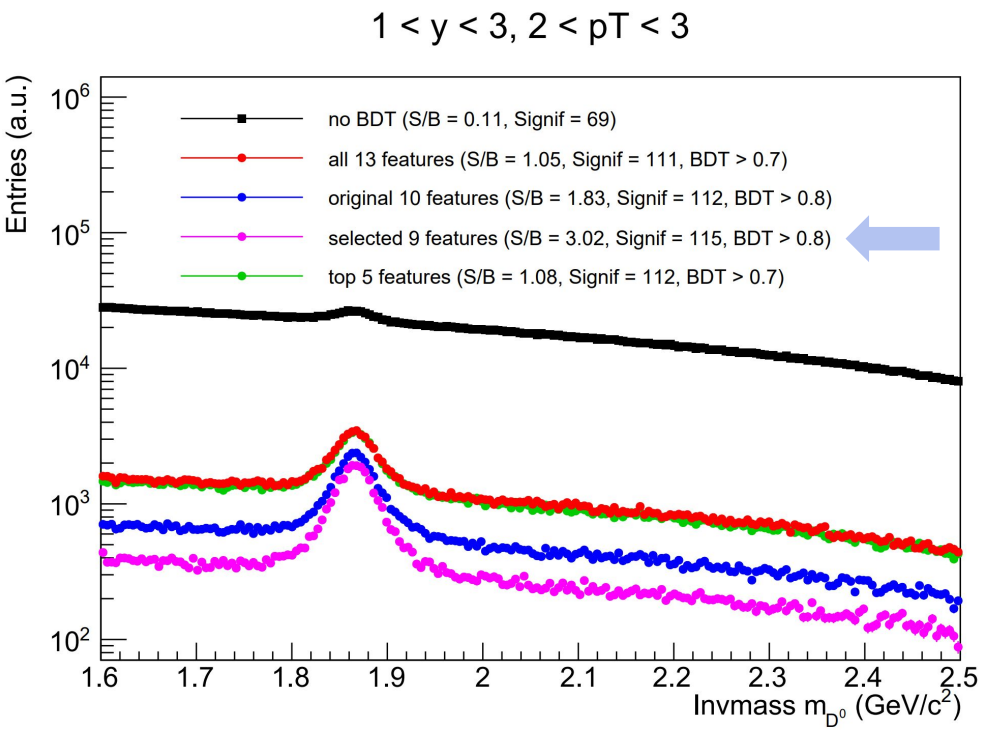
- how often a feature is used by BDT
- how much sig/bkg separation improves with each use of a feature

Relative importance of features differs for various y and pT slices.

1 < y < 3
2 < pT < 3

Phys. Rev. C 99, 034908 (2019)

Results for various feature combinations



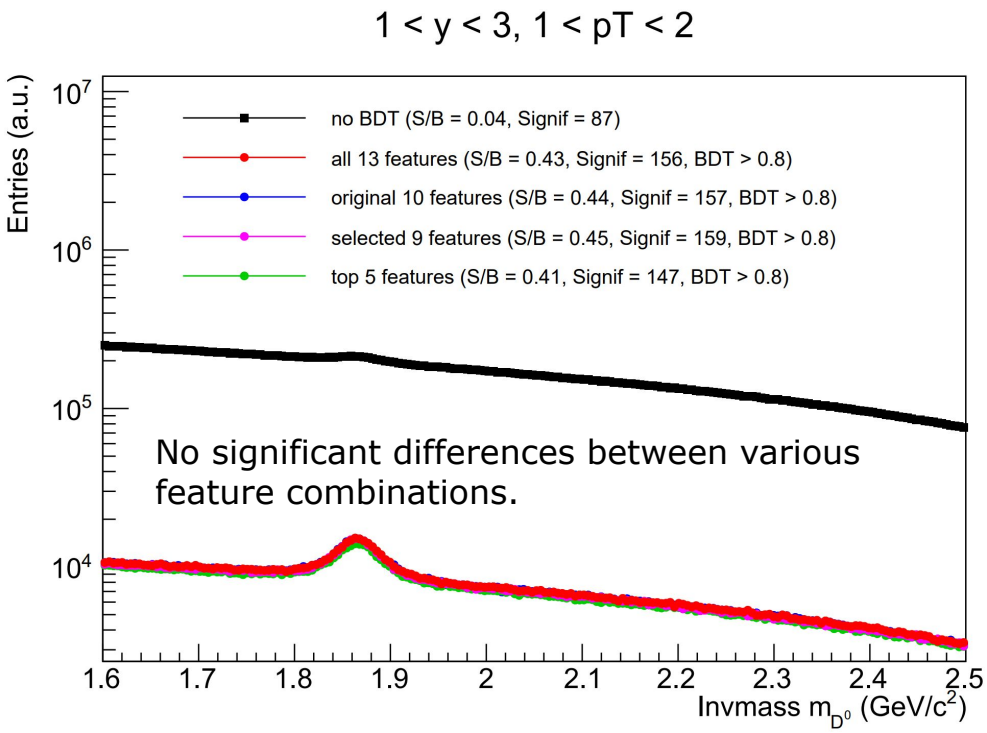
All 13 features:

top 5
- costheta
- sigma_vtx
- signif_d0xy_pi
- signif_d0xy_k
- dca_D0
- costheta_xy
- decay_length
- d0_pi
- d0_k

selected 9 (best combination)
- d0xy_pi
- d0xy_k
- sum_d0xy
- dca_12

Adding more features does not necessarily improve BDT.
- Additional features may have low importance.
- Too many features can lead to overfitting.

Results for various feature combinations



Impact of feature selection varies in different y and pT slices.

- BDT is less sensitive to feature selection at low pT due to larger statistics.
- Feature selection is more important for computational efficiency and interpretability.

Task III: Tuning BDT hyperparameters

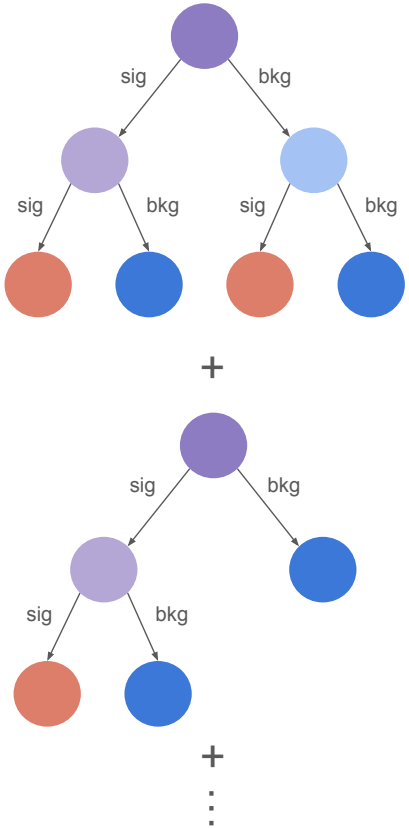
BDT hyperparameters examined

- n_estimators: number of trees
- max_depth: max number of iterations per tree
- learning_rate: size of each boosting step

Basic hyperparameters

- min_child_weight: min number of candidates required in each node
  - Default: 1
- min_split_loss: min improvement required to split a node
  - Default: 0

Regulators to help prevent overfitting



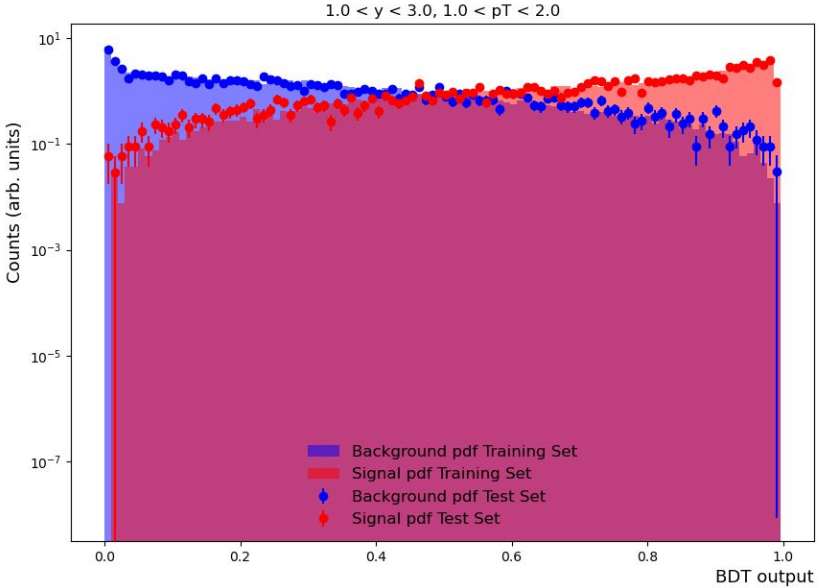
Optuna tests combinations of hyperparameters to find the best values.

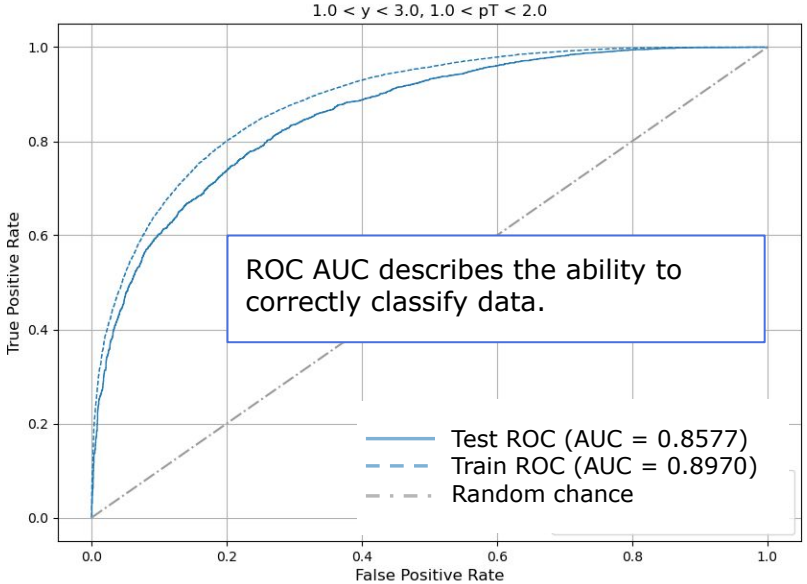
BDT results: basic hyperparameter optimization

n_estimators:	(100, 500)	→	480
max_depth:	(1, 3)	→	3
learning_rate:	(0.01, 0.1)	→	0.067
min_child_weight:	default	→	1
min_split_loss:	default	→	0

BDT is well-trained and stable.

- Good ROC AUC
- Similar results for training and test data

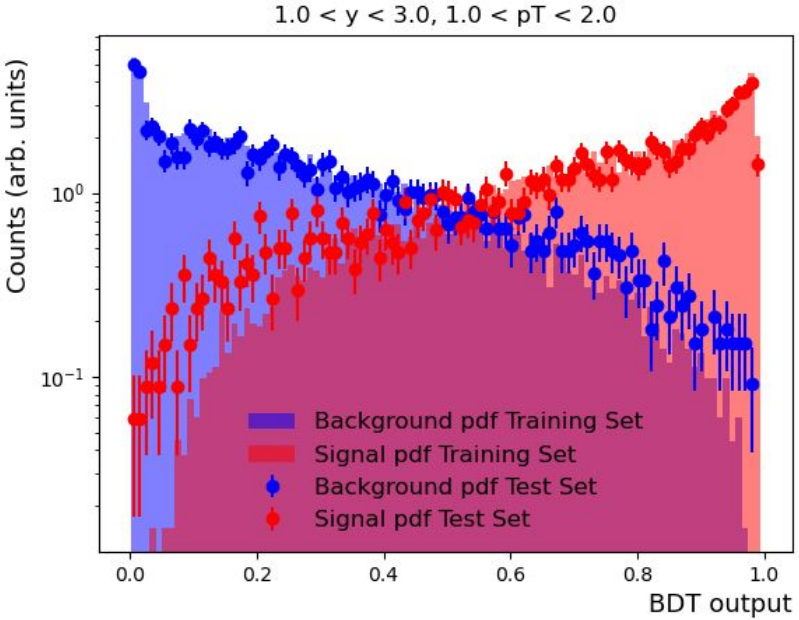


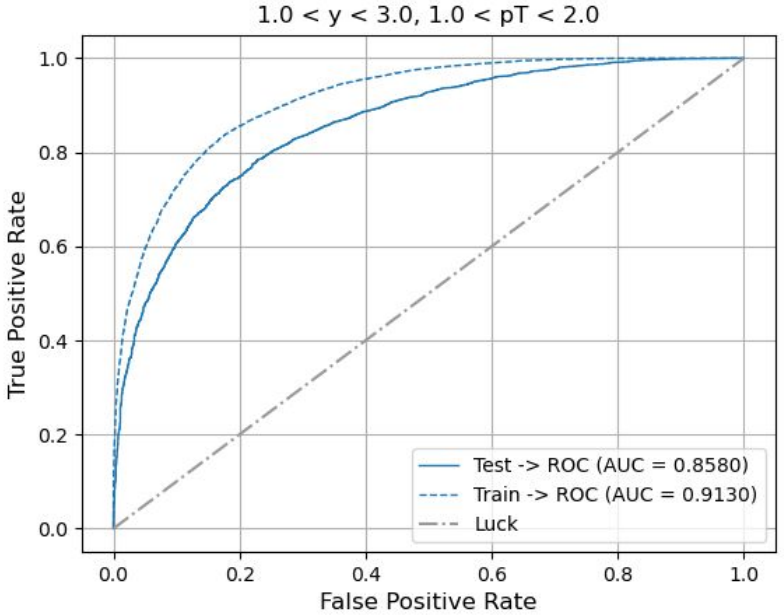


BDT results: increased hyperparameter flexibility

n_estimators:	(100, 800)	→	347
max_depth:	**(1, 8)**	→	**6**
learning_rate:	(0.01, 0.3)	→	0.034
min_child_weight:	default	→	1
min_split_loss:	default	→	0

- No significant improvement in test ROC AUC
- Overfitting due to large tree depth



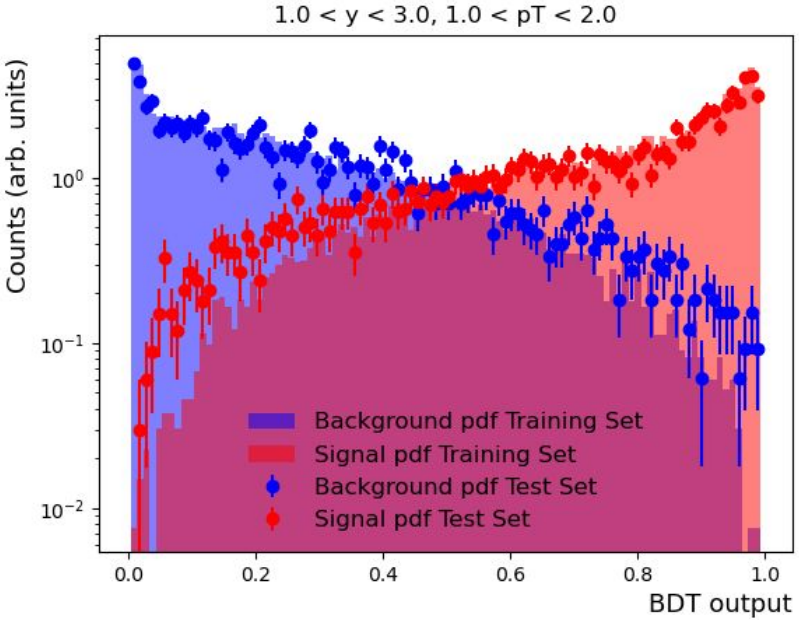


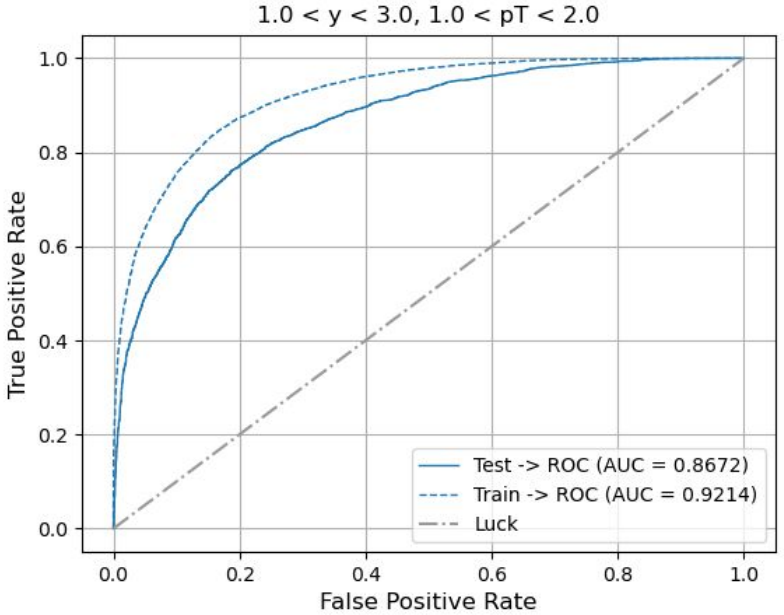
BDT results: with additional hyperparameters

n_estimators:	**(100, 800)**	→	**728**
max_depth:	**(1, 8)**	→	**8**
learning_rate:	(0.01, 0.3)	→	0.011
min_child_weight:	(1, 10)	→	10
min_split_loss:	(0, 5)	→	1

Slight improvement in test ROC AUC, but still overfits.

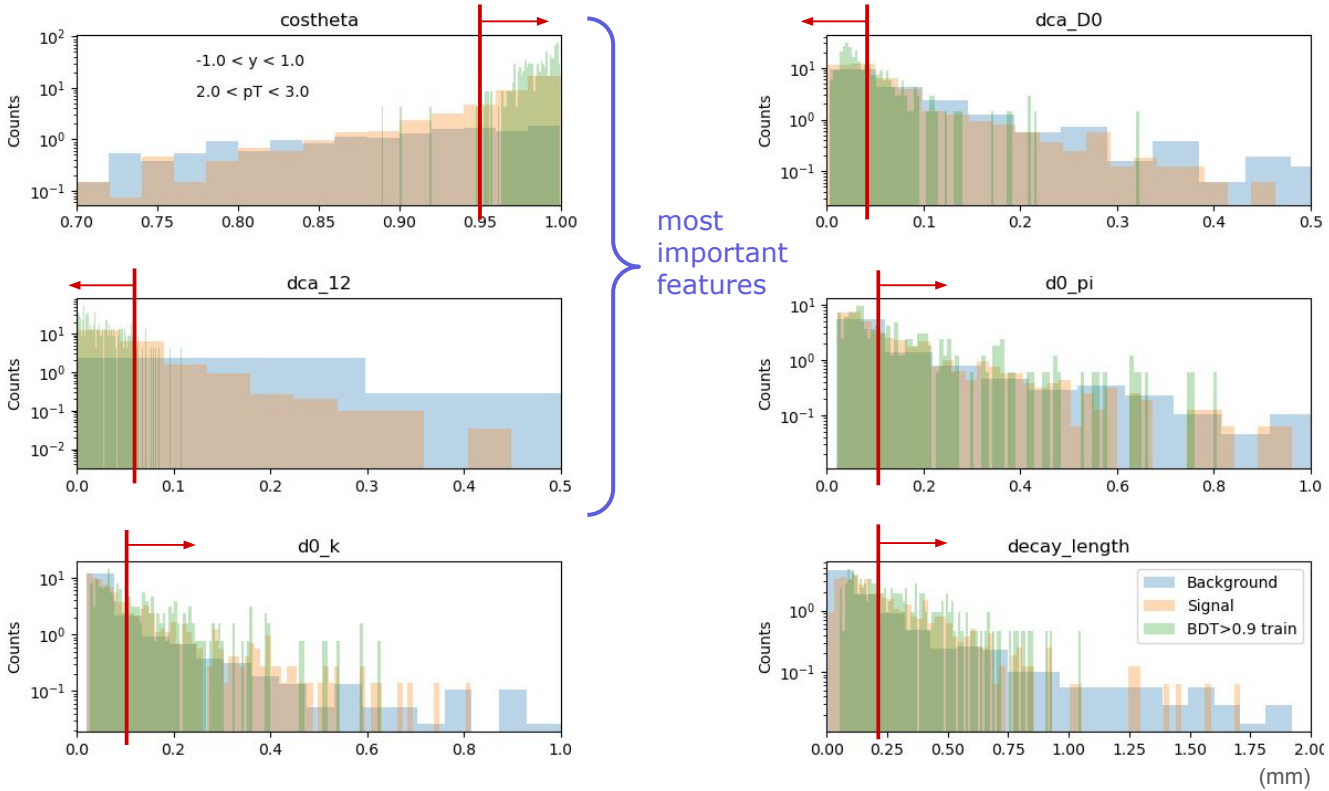
➔ Limit the ranges of n_estimators and max_depth, or increase values of regulators.





BDT results vs. box cuts

BDT vs. box cuts: topological selections



most important features

Box cuts from STAR D0 analysis (-1<y<1, 2<pT<3):

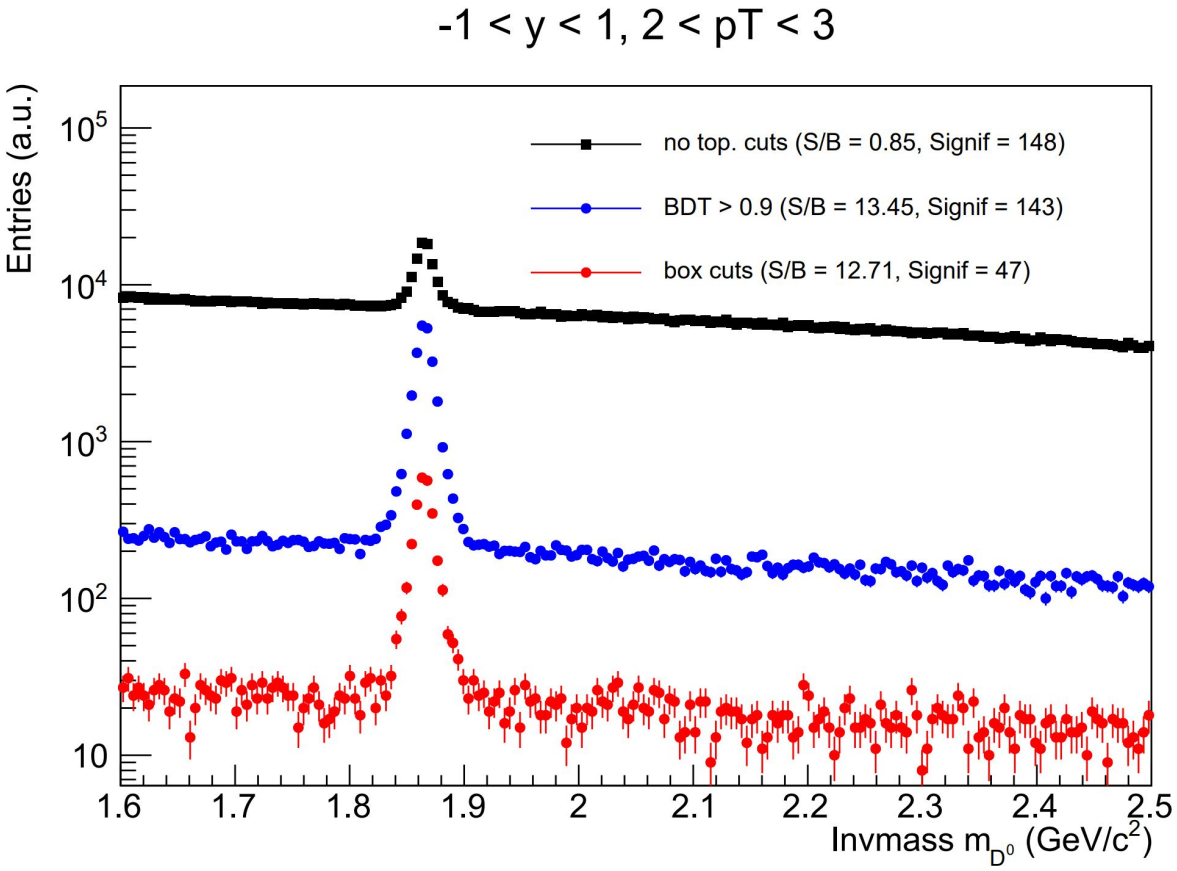
$0-10\%$ | p_T (GeV/c)		(2,3)
Decay Length (μm)	>	232
DCA_{12} (μm)	<	63
$\mathrm{DCA}_{\mathrm{D}^0}$ (μm)	<	40
DCA_{π} (μm)	>	97
$\mathrm{DCA}_{\mathrm{K}}$ (μm)	>	94
$\cos(\theta)$	>	0.95

Phys. Rev. C 99, 034908 (2019)

BDT selections are looser and more flexible than box cuts.

Similar treatments of the most important features.

BDT vs. box cuts: D0 invariant mass



BDT yields higher Signif

- Similar S/B
- Higher statistics

Conclusions

Summary and future directions

- BDT performance varies in different y and pT slices
  - Separately train BDT for each slice since topological distributions differ
- Feature selection may not significantly impact BDT
  - More relevant for physical and practical considerations
- Greater flexibility in hyperparameter tuning can lead to overfitting
  - Restrict max_depth and n_estimators to small values
- BDT is similar to but more efficient than box cuts
  - BDT yields higher significance for similar S/B

Future directions:

- Optimize analysis for data with machine-background effects
- Further explore hyperparameter optimization
  - Additional regulators (max_delta_step, subsample)
  - Different sampling and optimization algorithms in Optuna

References

October ep simulations from ePIC
D0 samples (for signal candidates):
/volatile/eic/EPIC/RECO/25.10.4/epic_craterlake/Bkg_Exactly1SignalPer2usFrame/SIDIS/D0_ABCONV/pythia8.306-1.1/10x100/q2_1/hiDiv/

DIS samples (for background candidates):
/volatile/eic/EPIC/RECO/25.10.4/epic_craterlake/Bkg_1SignalPer2usFrame/DIS/NC/10x100/minQ2=1

Frameworks for sample analysis and BDT from Shyam
Analysis framework: https://github.com/eic/snippets/tree/main/JetsAndHF/SecondaryVertex_Chi2

Machine learning framework: https://github.com/eic/snippets/tree/main/JetsAndHF/ML_HF_Reconstruction
Based on hipe4ml: https://doi.org/10.5281/zenodo.5070131

Acknowledgements

Thank you to

- Deepa Thomas
- Shyam Kumar
- Rongrong Ma
- ePIC Collaboration
- RHIP group