Track/Shower Discrimination for protoDUNE using BDT approach

Mousam Rai

26th August 2020 / ProtoDUNE Meeting/ Supervisor – Dr John Marshall

DEEP UNDERGROUND NEUTRINO EXPERIMENT

WARWICK
THE UNIVERSITY OF WARWICK

Roadmap

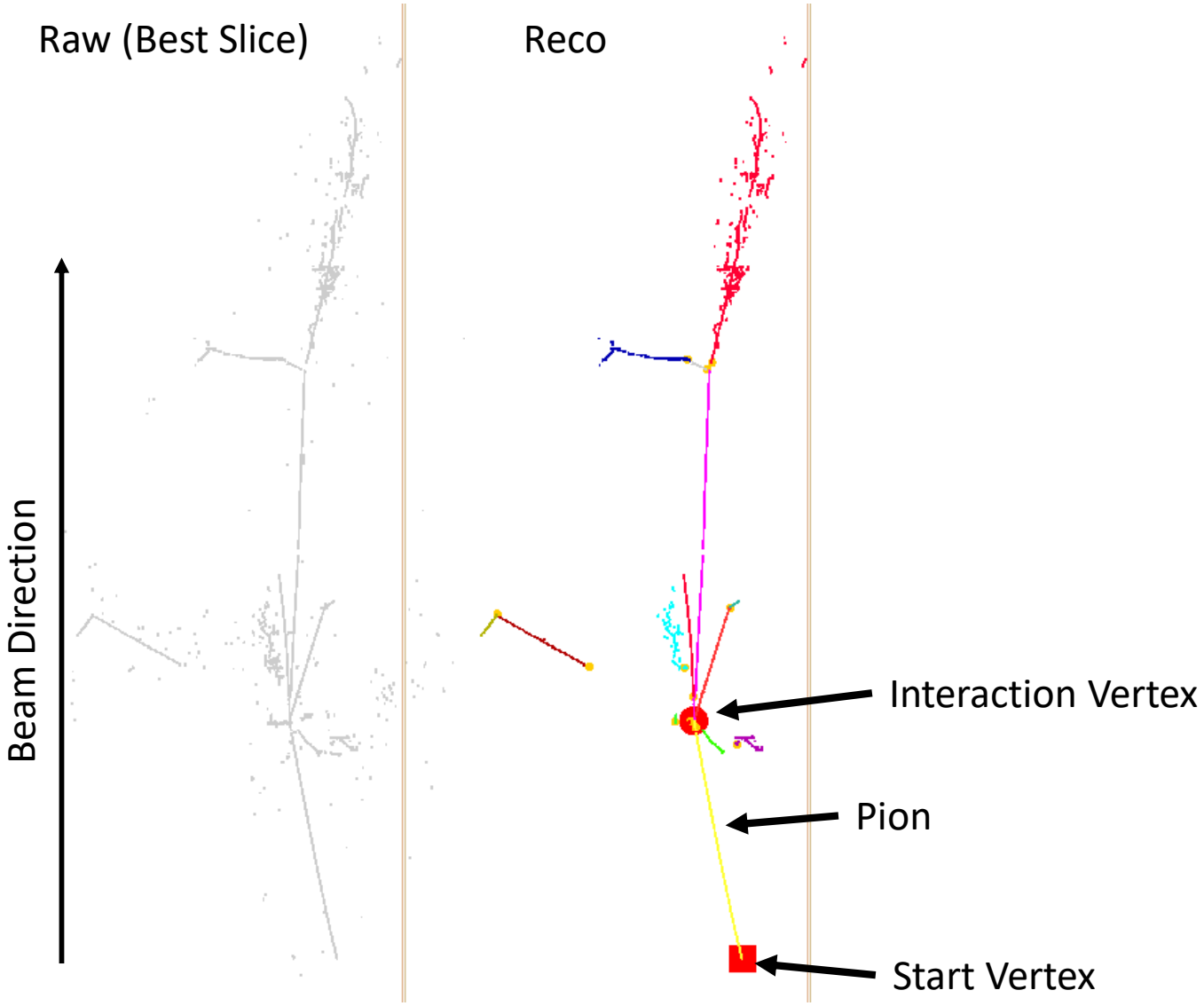
- Terminology
- Small note on vertexing
- BDT variables
- BDT setup
- Brief study on tunable parameters
- Performance plots and numbers
- Summary

Terminology

		True Condition	
	Total PFOs	True Tracks (TT)	True Showers (TS)
Predicted Condition	Predicted Tracks (PT)	True Positive (TP)	False Positive (FP)
	Predicted Showers (PS)	False Negative (FN)	True Negative (TN)

- 2x2 confusion matrix
- Sensitivity/True Positive Rate (TPR) = $\frac{TP}{TT}$
- Specificity/True Negative Rate (TNR) = $\frac{TN}{TS}$
- Overall Accuracy = $\frac{TP+TN}{Total\ PFOs}$
- Trying to maximize Overall Accuracy

Small note on vertexing



- Incoming pion has 2 vertex –
  - Interaction Vertex
  - Start vertex
- Use Interaction Vertex as a basis for BDT variables (more on these on the next slide)

BDT Variables

- 13 variables
- Topological – Hit spread information
- Calorimetric – charge-based information
- Hierarchy – hierarchy-based information

MicroBooNE		DUNE FD
Topological	**Calorimetric**	**Hierarchy**
length	charge1	nAllDaughter
diff	charge2	nHits3DDaughterTotal
gap		daughterParentNhitsRatio
rms		
diffAngle		
pca1		
pca2		
vertexDistance		

BDT Setup

- Training
  - MCC12 samples WITH space charge – prepared by Andy Chappell. (Thanks!)
  - Use tuned BDT parameters (in next few slides)
  - Completeness and Purity Cuts (in next few slides)
  - Fiducial volume cuts : 30 cm in both x-direction, 20 cm in both y-direction, 100 cm in both z-direction
- Testing
  - MCC12 samples WITH space charge – prepared by Andy Chappell. (Thanks again!)
  - No completeness or purity cuts
  - Fiducial Volume cuts : see above

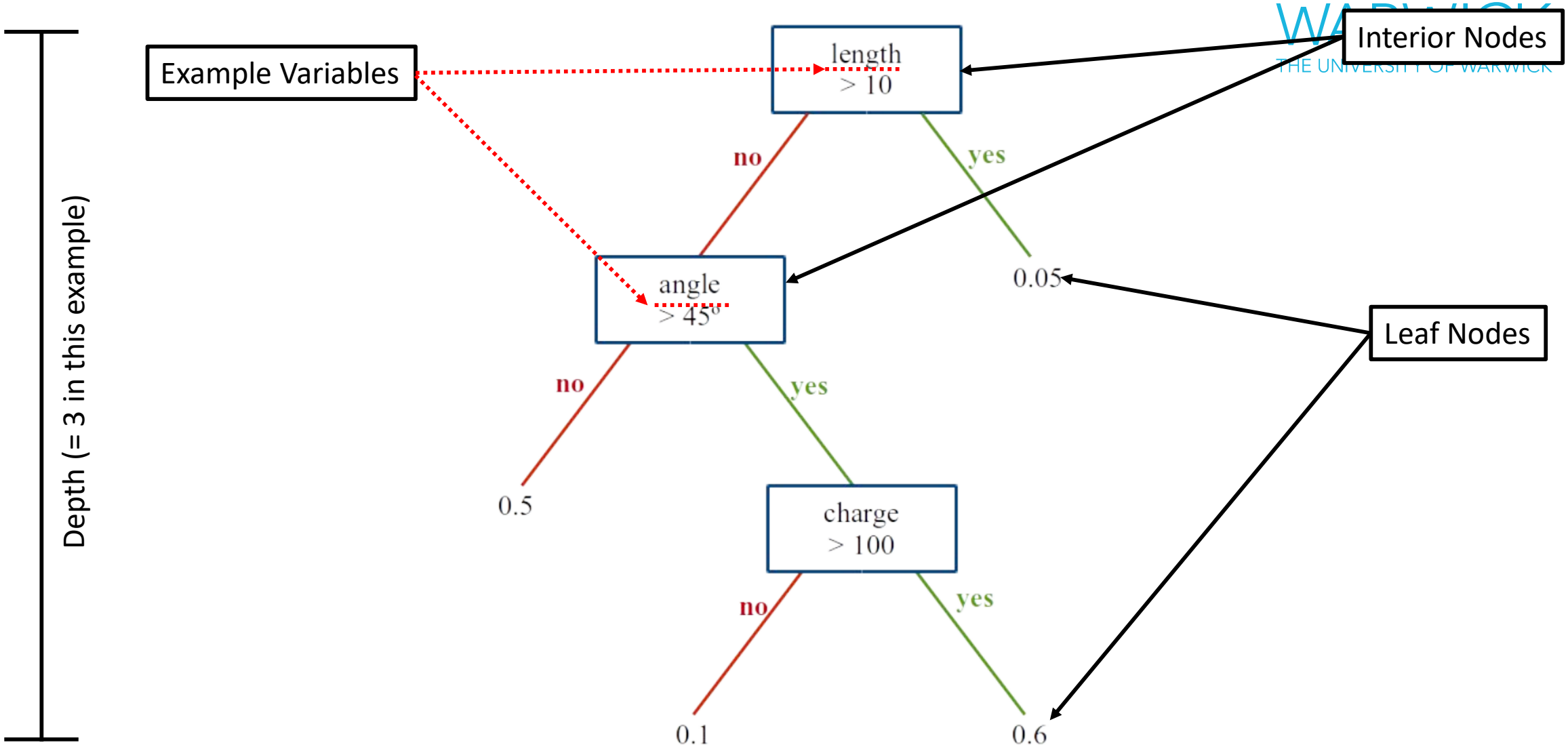
Energy (GeV)	Total #Event Files	Train (40% Total)	Test (60% Total)
1	196	78	118
2	49	20	29
3	396	158	238
6	394	158	236
7	304	122	182
Total #Event Files	1339	536	803
Total #PFOs (before fiducial cuts)	229,099	95,229	133,870
Total #PFOs (after fiducial cuts)	86,479	36,154	50,325

WARWICK
THE UNIVERSITY OF WARWICK

Brief study on tunable parameters

- We can try to tune certain parameters to increase BDT performance
- Focused on Ntrees, MaxDepth, MinNodeSize, and AdaBoostBeta
- Also investigate completeness and purity cuts used in the training samples

Example of a tree



MaxDepth

- Max depth of the decision tree allowed
- In general, the greater the max depth, the better the performance
- But takes longer to train and test

NTrees

- Number of trees in the forest
- As with MaxDepth, in general, more is better
- But also increases training and testing times

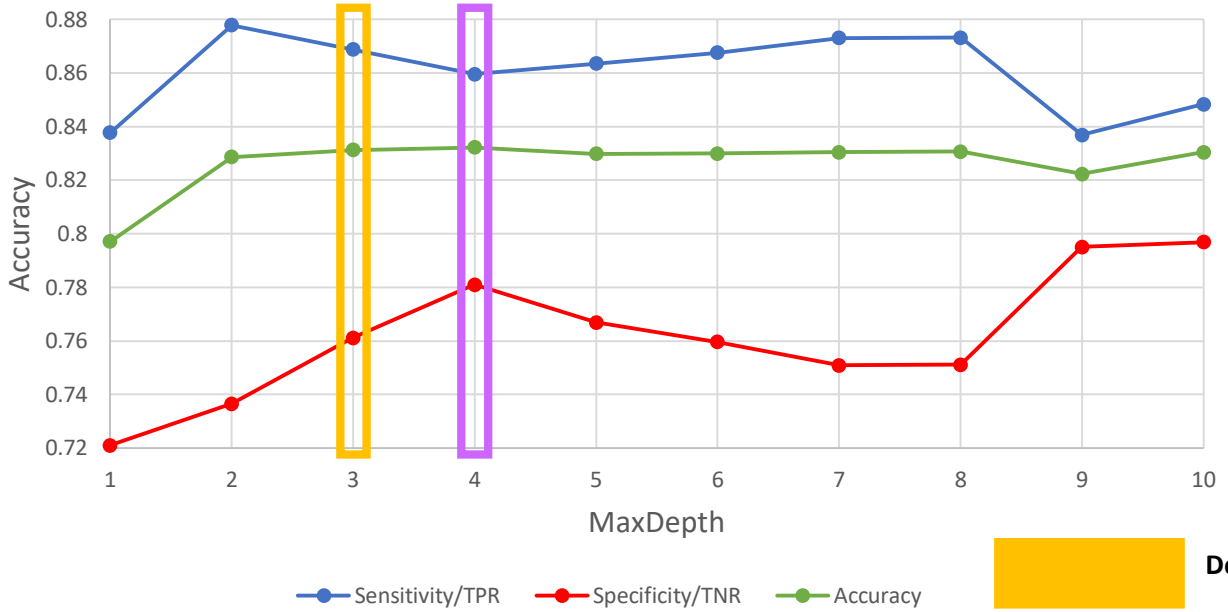
MinNodeSize

- Minimum percentage of training events required in a leaf node
- For example, TMVA user guide has a default value of 5%
- But other sources online recommends something like ~1%
- Just need to tune it for your use case

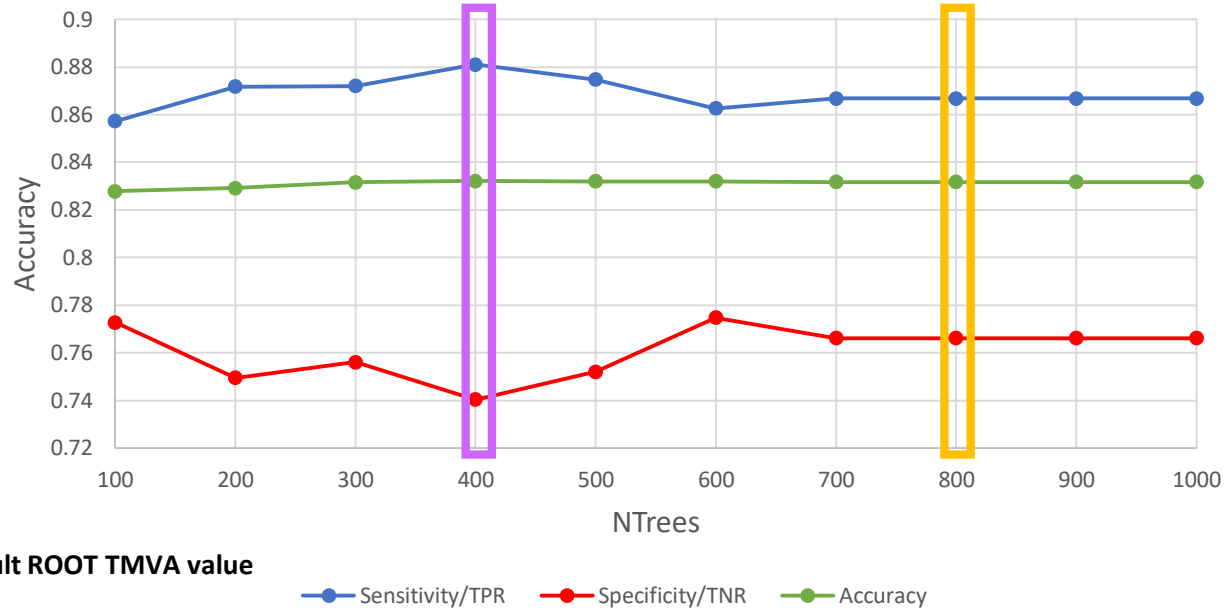
AdaBoostBeta

- Learning rate if using AdaBoost as boosting type
- Based on exponential loss function
- More on CERN TMVA User Guide (Section 7.1)
- Generally, smaller is better but also slower

MaxDepth vs Accuracy



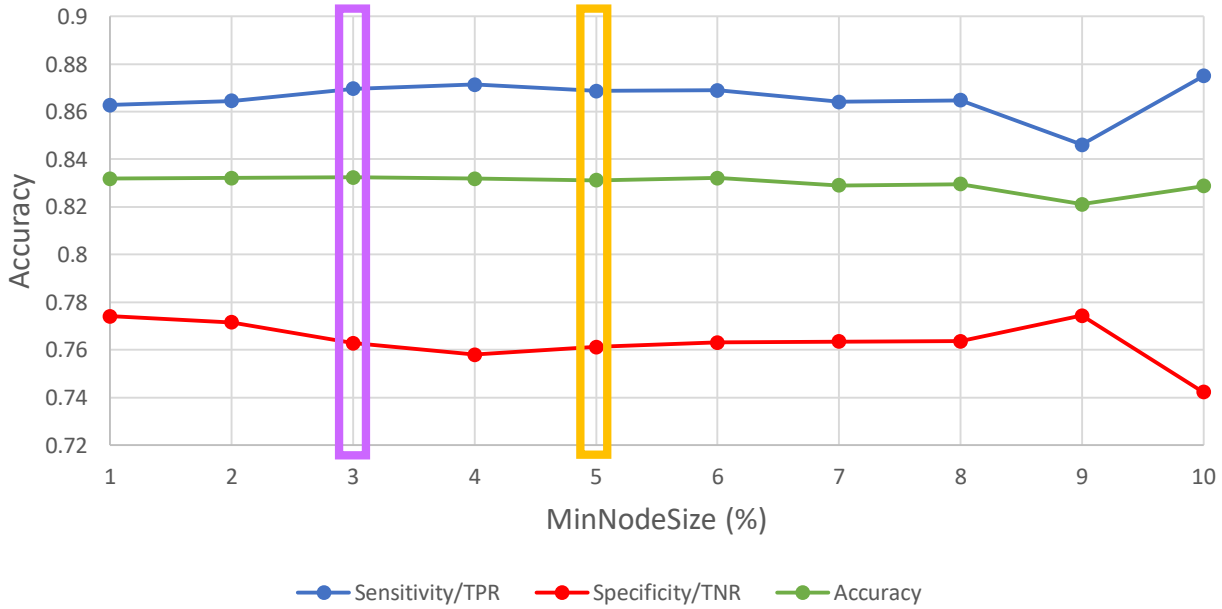
NTrees vs Accuracy



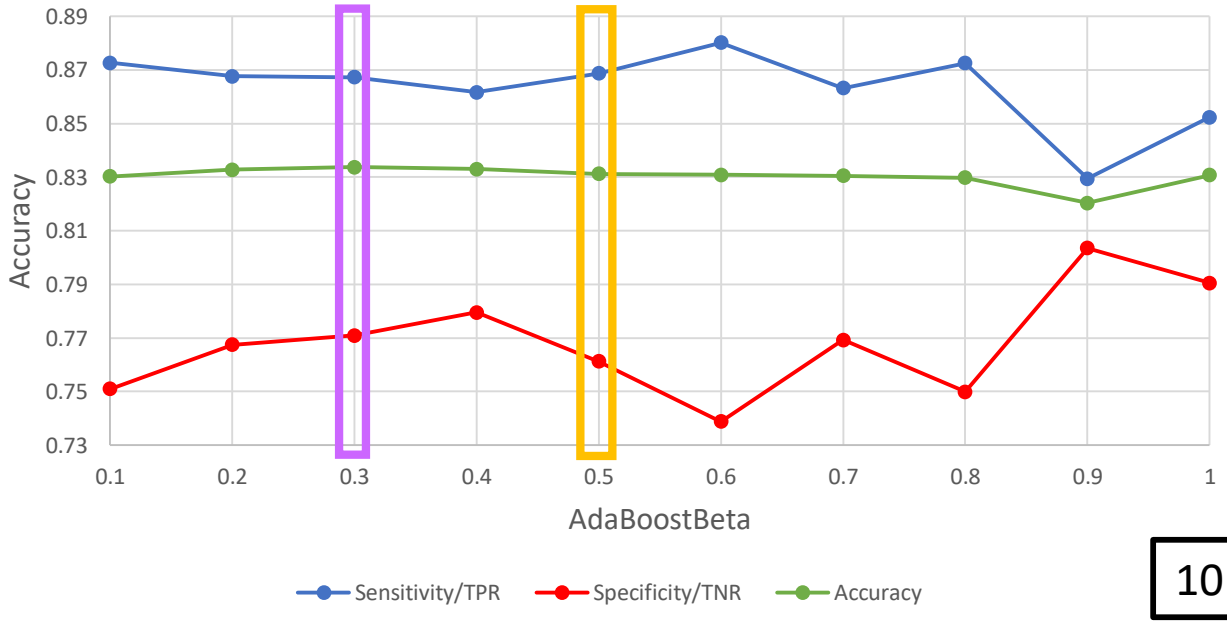
Default ROOT TMVA value

Best performance value

MinNodeSize vs Accuracy

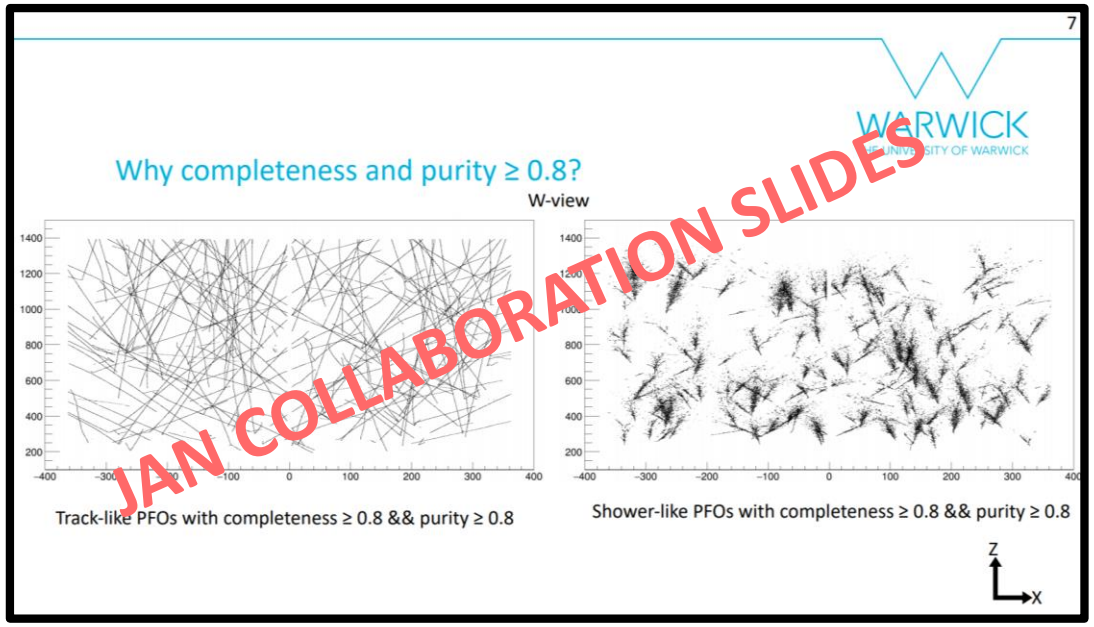


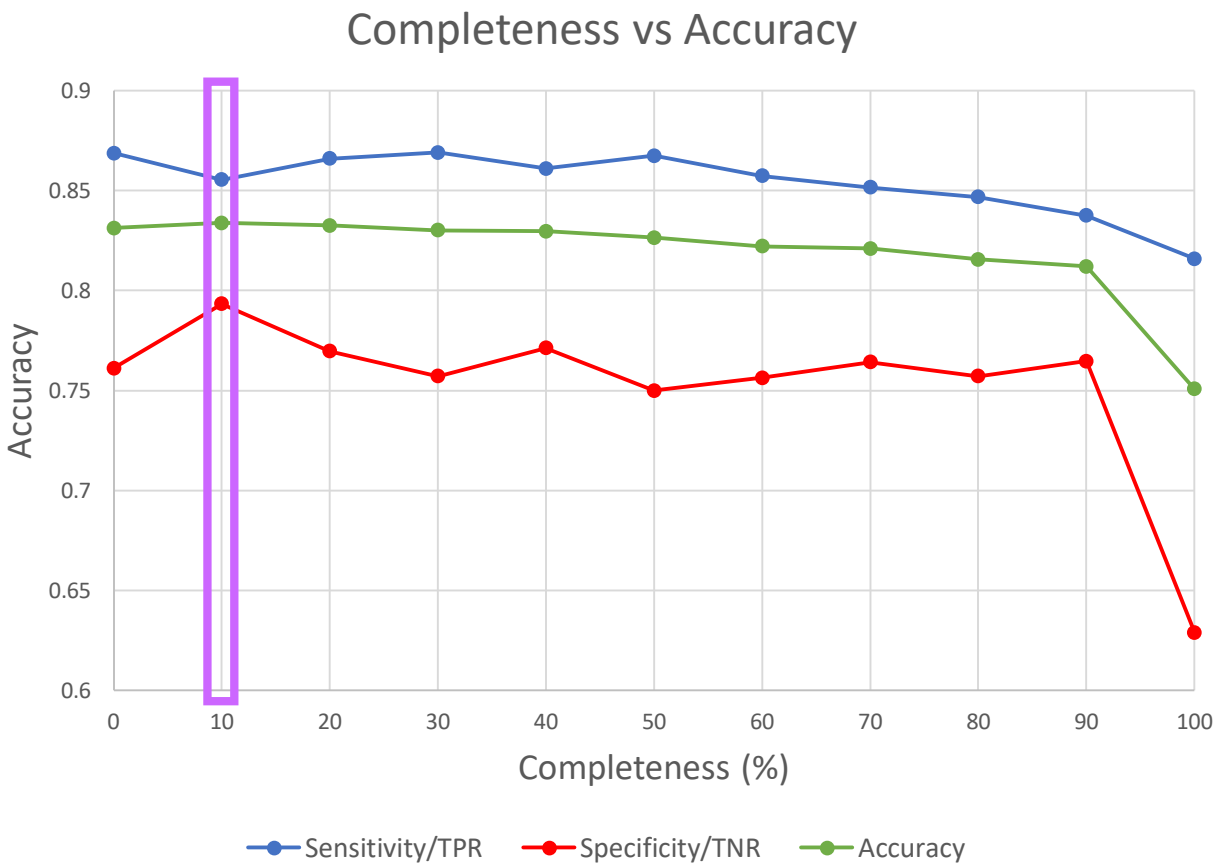
AdaBoostBeta vs Accuracy

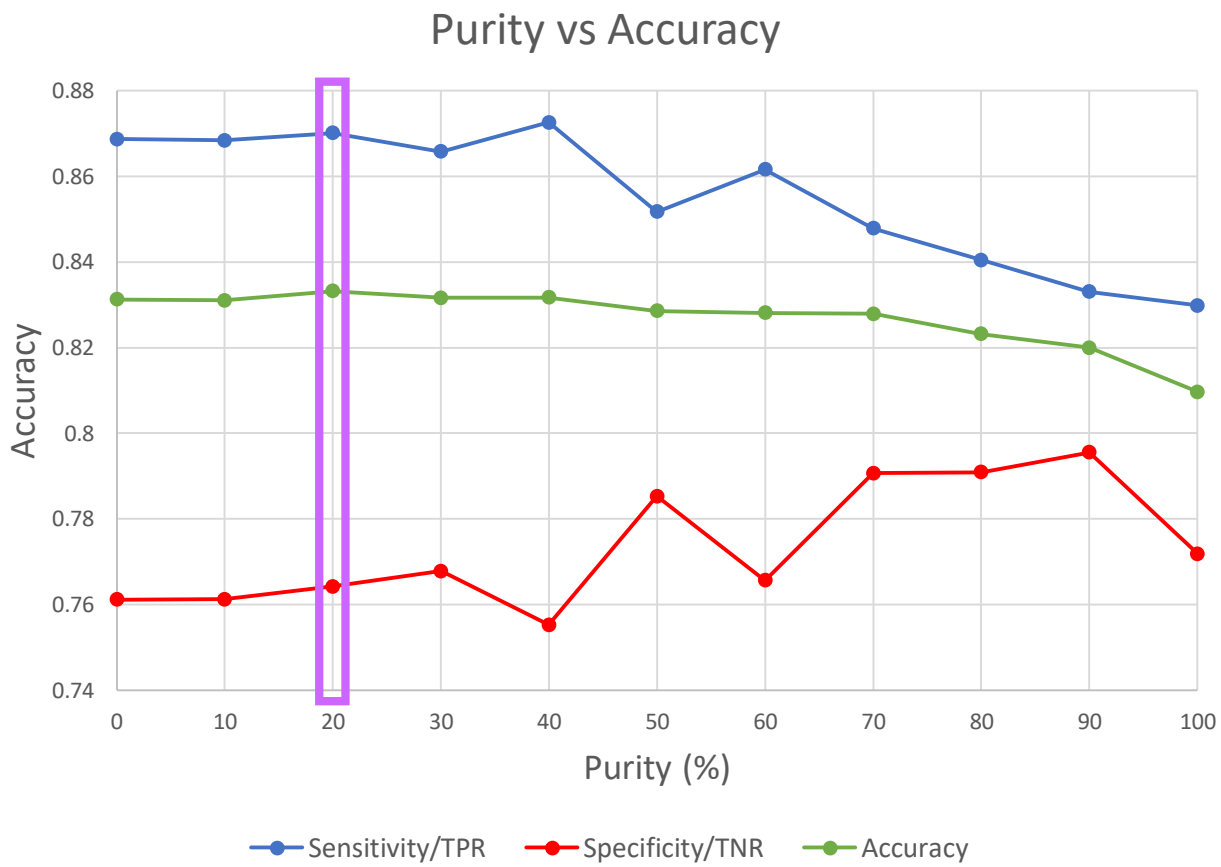


Tuning Completeness and Purity cuts

- Lesson learnt from working on DUNE FD mc events
- Initially thought training on high completeness and purity PFOs was better but on further study, it turns out, it isn't.
- Mainly because there are a lot of low completeness and purity PFOs and have very sparse hit spread so they don't look anything like high completeness and purity PFOs so what you are training on isn't representative of the actual topologies present in your event.
- Using completeness and purity cuts ONLY during training







Default ROOT TMVA value

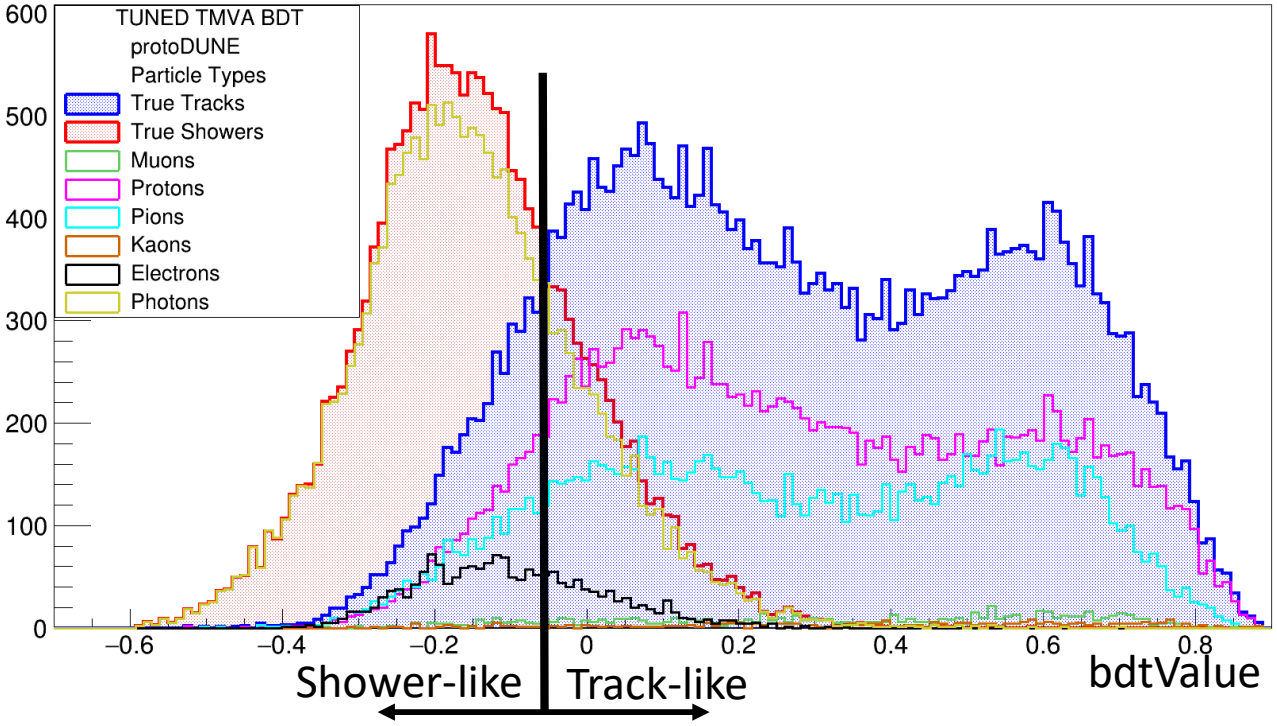
Best performance value

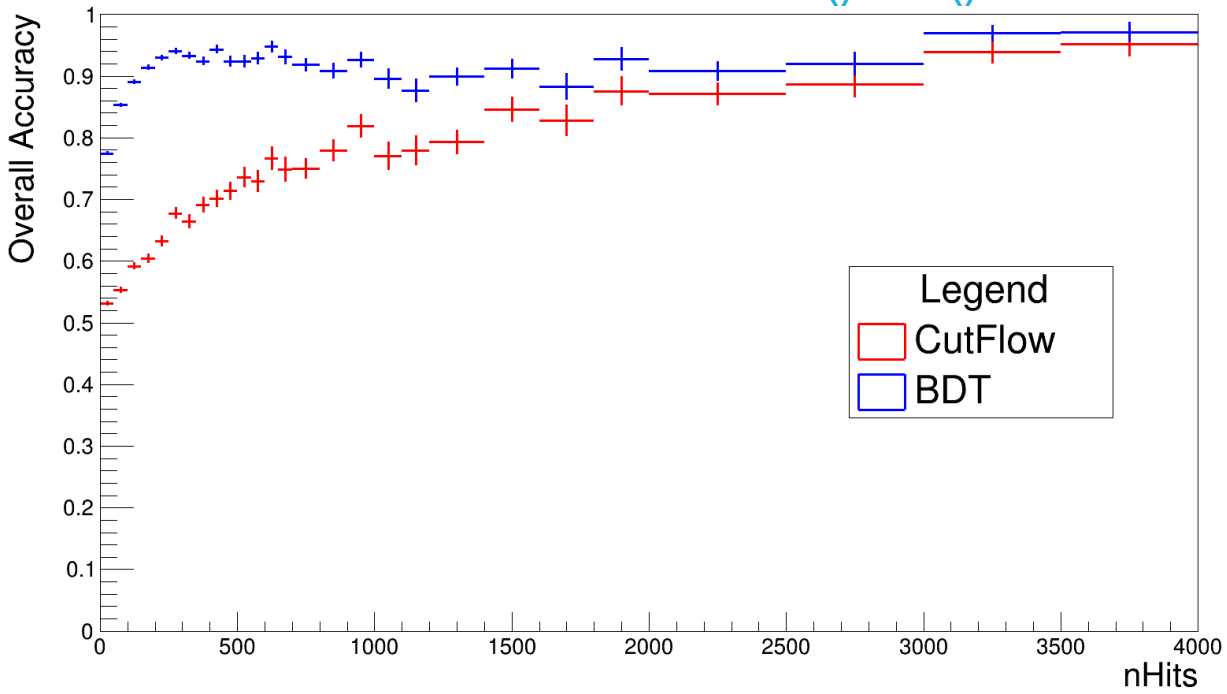
WARWICK
THE UNIVERSITY OF WARWICK

Tuned BDT for protoDUNE

- Take all the best performing parameter values and cuts and train a new BDT

Performance Numbers and Plots





	Sensitivity	Specificity	Accuracy	#PFOs
CutFlow	0.4464	0.9524	0.6224	~50,000
Tuned BDT	0.8718	0.7667	0.8352	~50,000

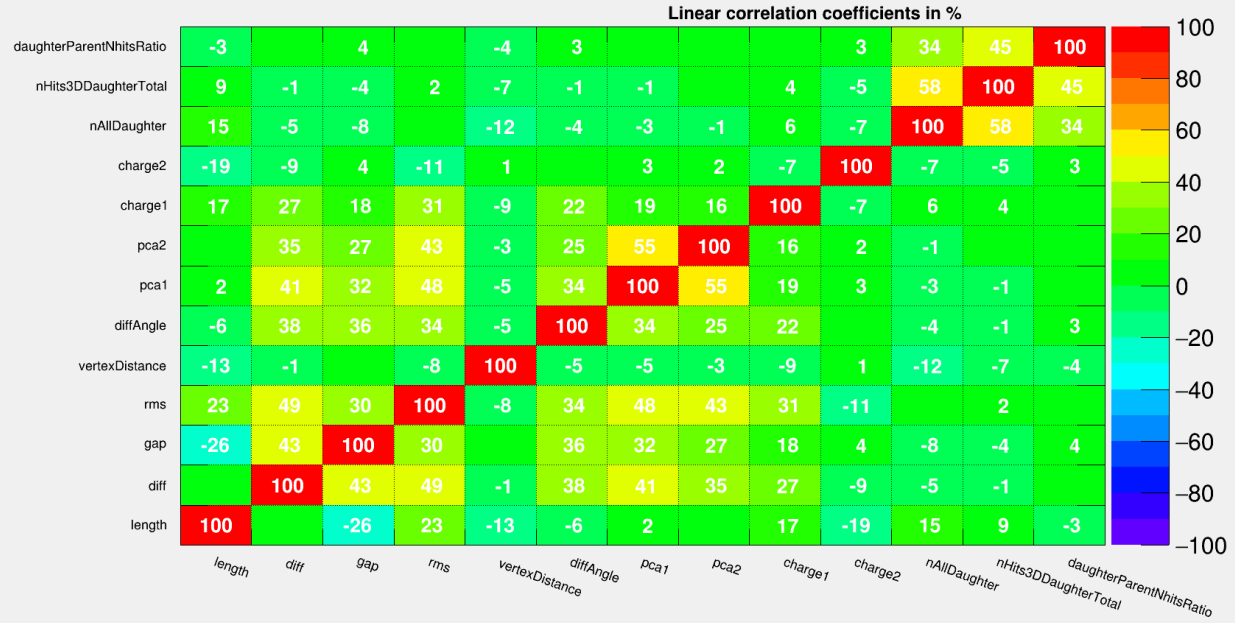
Summary

- ROOT TMVA BDT is significantly outperforming the current cutflow approach (especially at the low nHits region).
- Next steps include implementing the BDT using SKLearn (which is what PANDORA supports), testing and replicating performance numbers and plots, and ideally, work towards its release.
- Any questions and comments are deeply appreciated.

WARWICK
THE UNIVERSITY OF WARWICK

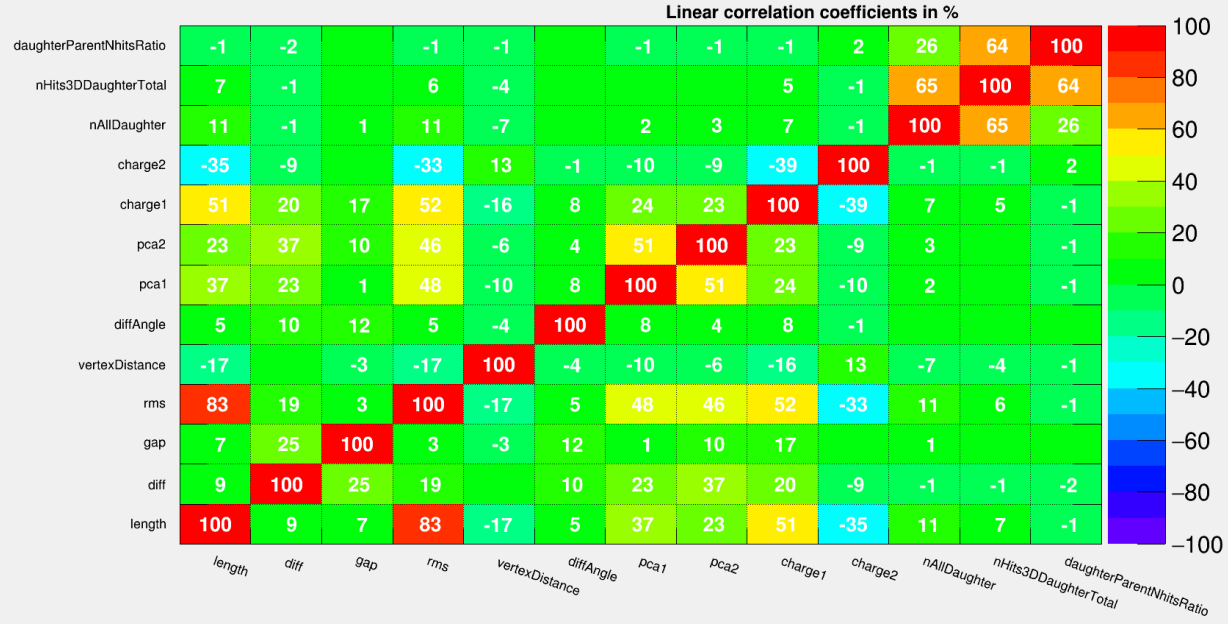
BACK UP SLIDES

Correlation Matrix (signal)

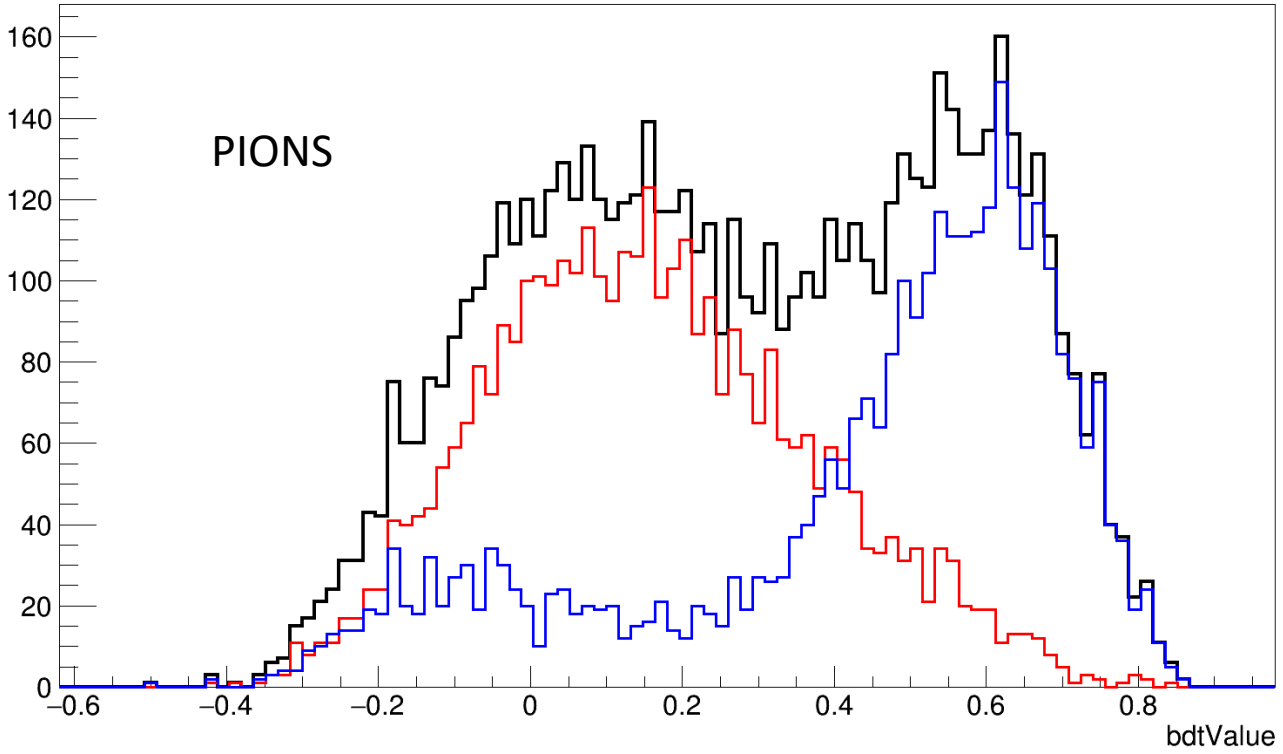


WARWICK
THE UNIVERSITY OF WARWICK

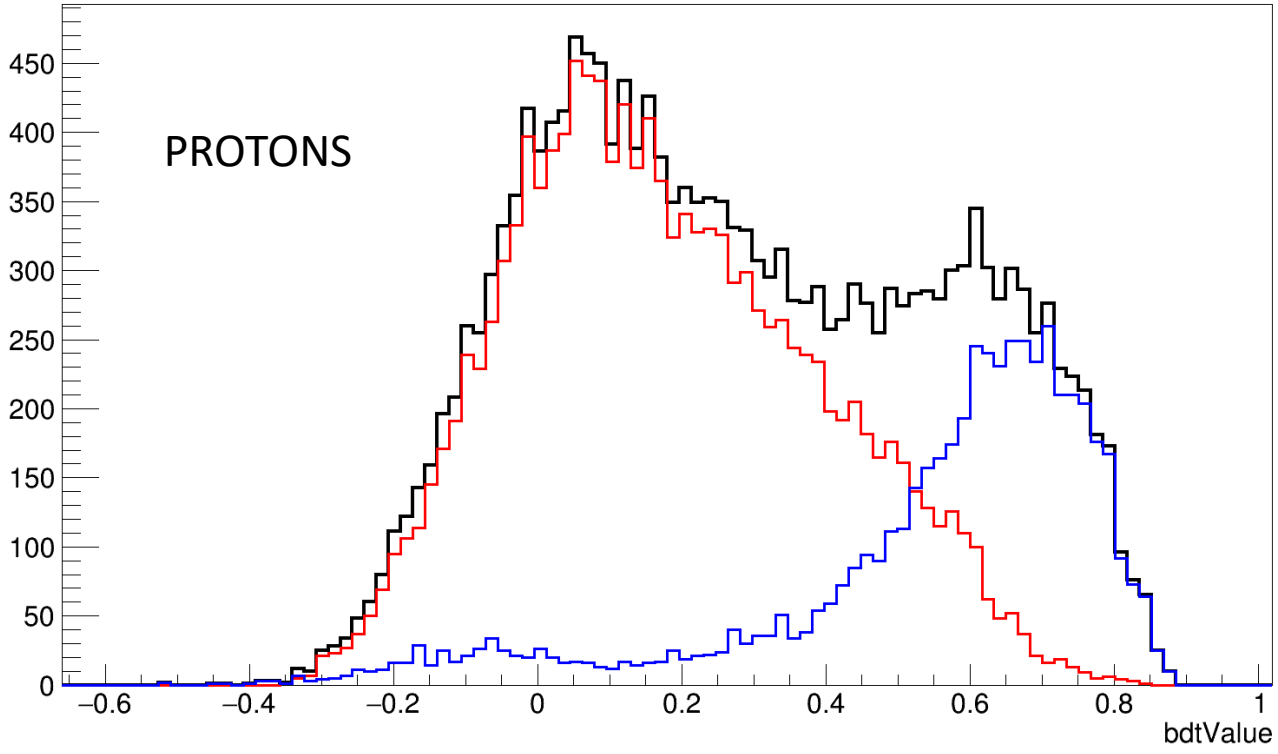
Correlation Matrix (background)



DOUBLE PEAK PION AND PROTON DISTRIBUTION



	ALL PIONS
	PIONS (nHits < 200)
	PIONS (nHits >= 200)



	ALL PROTONS
	PROTONS (nHits < 200)
	PROTONS (nHits >= 200)

Variable	Definition
length	3D length of a PFO
diff	Average mean difference between the position of hits and a straight line divided by the straight line length
gap	Average max gap distance divided by the straight line length
rms	Average root mean square of linear sliding fit divided by straight line length
vertexDistance	Distance between the PFO vertex and the interaction vertex
diffAngle	Difference between the opening and closing angles calculated over 50% of the pfo closest and furthest from the vertex
pca1	Ratio between the second largest and the largest PCA eigenvalue
pca2	Ratio between the third largest and the largest PCA eigenvalue
charge1	Ratio between sigmaCharge (= (charge – meanCharge)^2) and the mean charge in the collection plane
charge2	Ratio of charge in the last 10% of the pfo and the mean charge in the collection plane
nAllDaughter	Number of all downstream daughter PFOs
nHits3DDaughterTotal	Number of 3D hits in all downstream daughter PFOs
daughterParentNhitsRatio	Ratio between the 3D hits in all downstream PFOs and the parent PFO