Track vs. Shower Identification Improvements using ML in Pandora

Mousam Rai

6th Jan 2020 / DUNE FD Sim-Reco Meeting / Supervisor – Dr John Marshall

DEEP UNDERGROUND NEUTRINO EXPERIMENT

WARWICK
THE UNIVERSITY OF WARWICK

WARWICK
THE UNIVERSITY OF WARWICK

Roadmap for this presentation

- Aim
- Variables
- Current and Proposed approach to track/shower ID in Pandora for DUNE FD
- Performance plots for proposed implementation
- Summary/Future Works

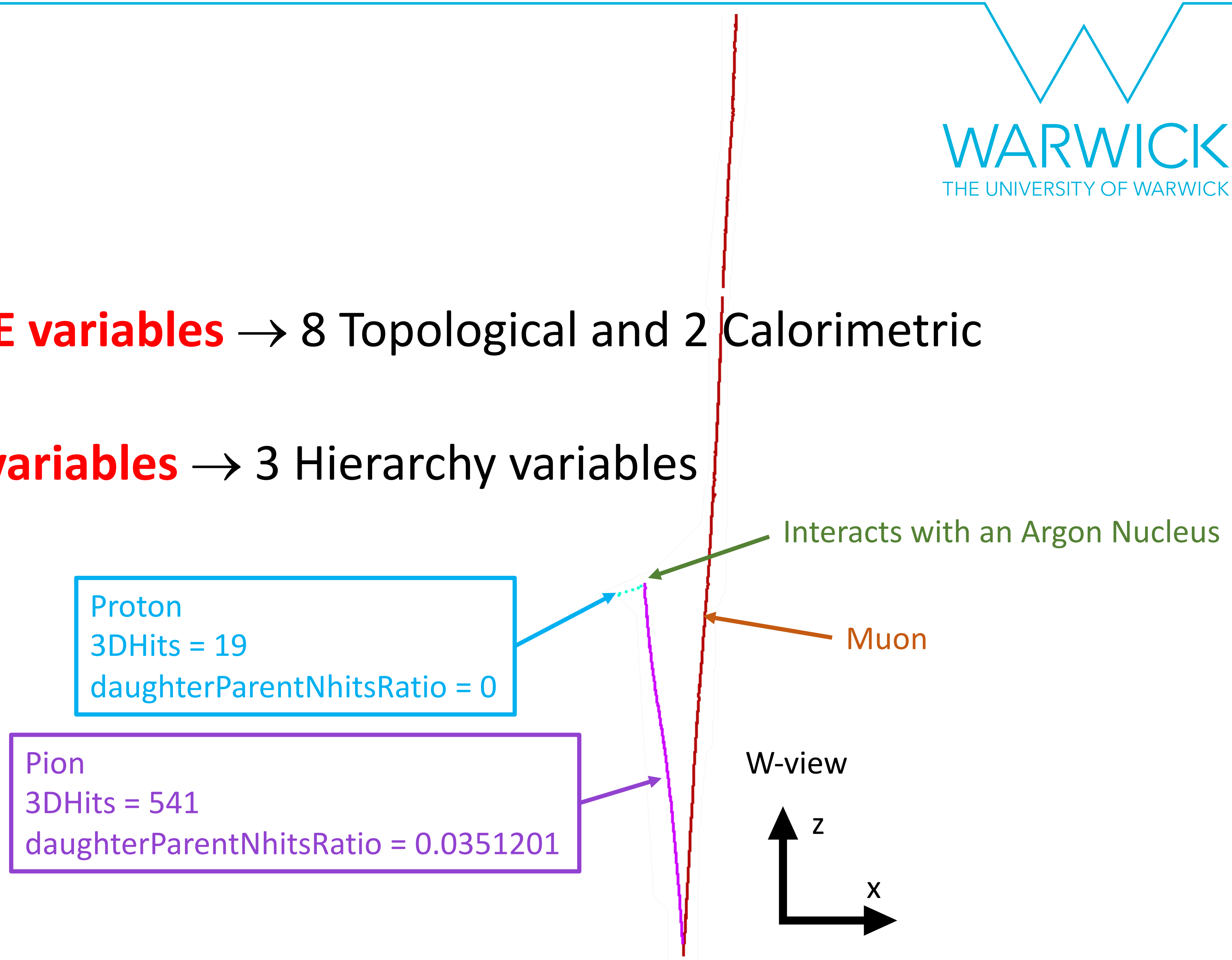
WARWICK
THE UNIVERSITY OF WARWICK

Aim

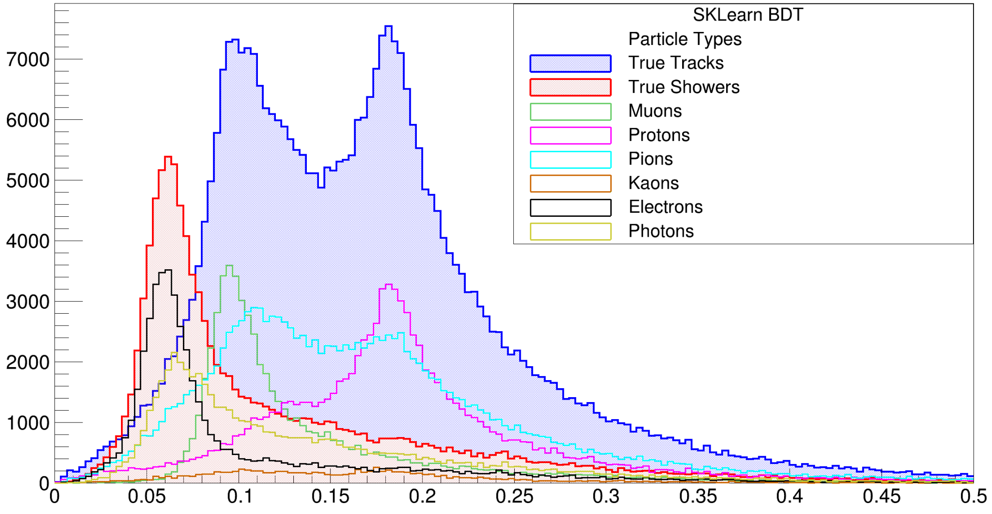
- Take particles reconstructed by Pandora and tag them as “track-like” or “shower-like”

Variables

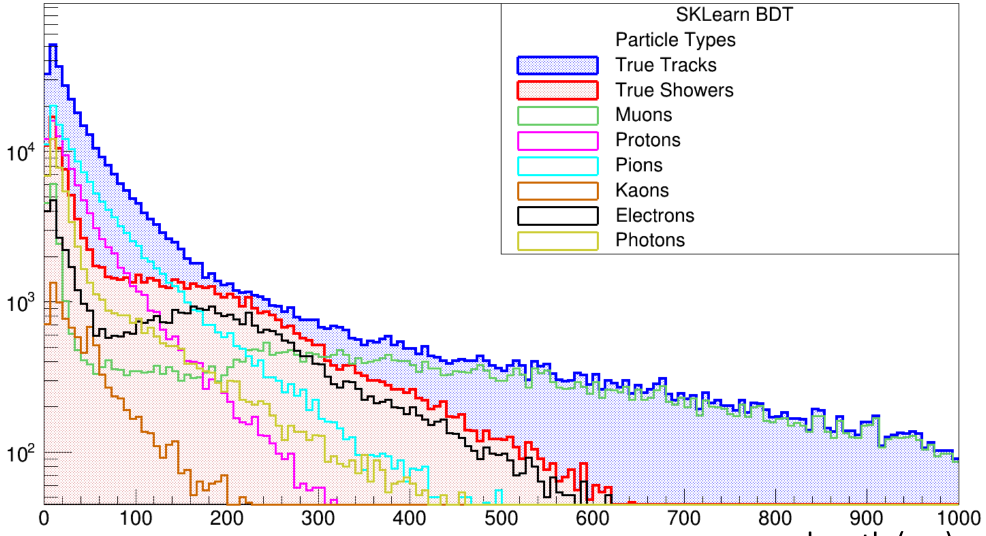
- **MicroBooNE variables** → 8 Topological and 2 Calorimetric variables
- **Additional variables** → 3 Hierarchy variables



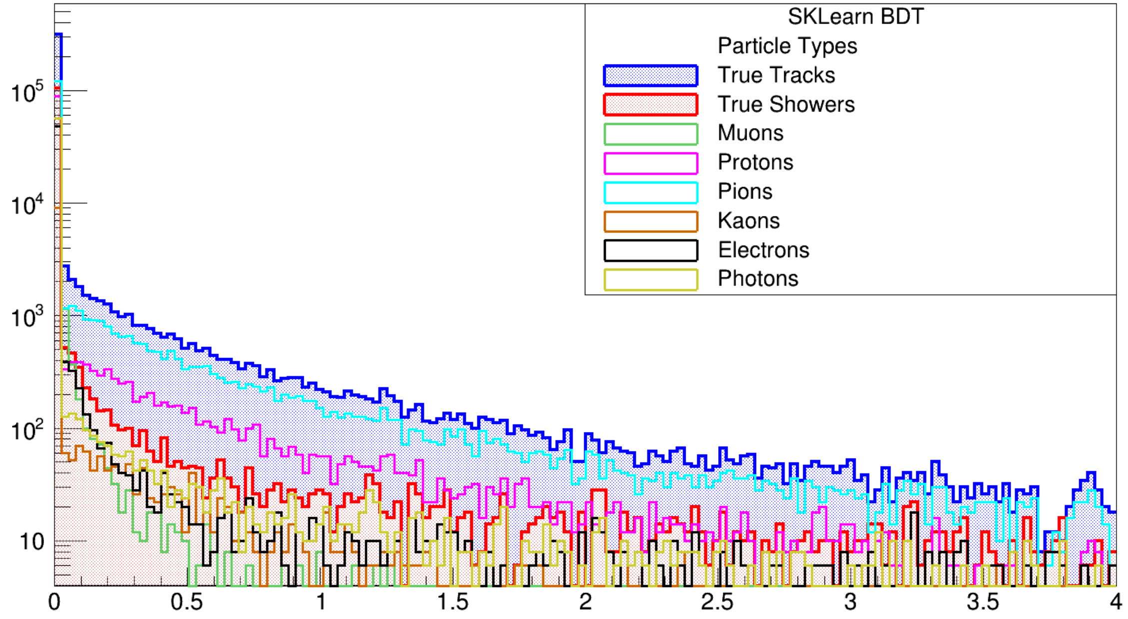
Distributions for selected variables



charge2 – Ratio of charge in the last 10% of the PFO and the mean charge in the collection plane



length – 3D length of the PFO



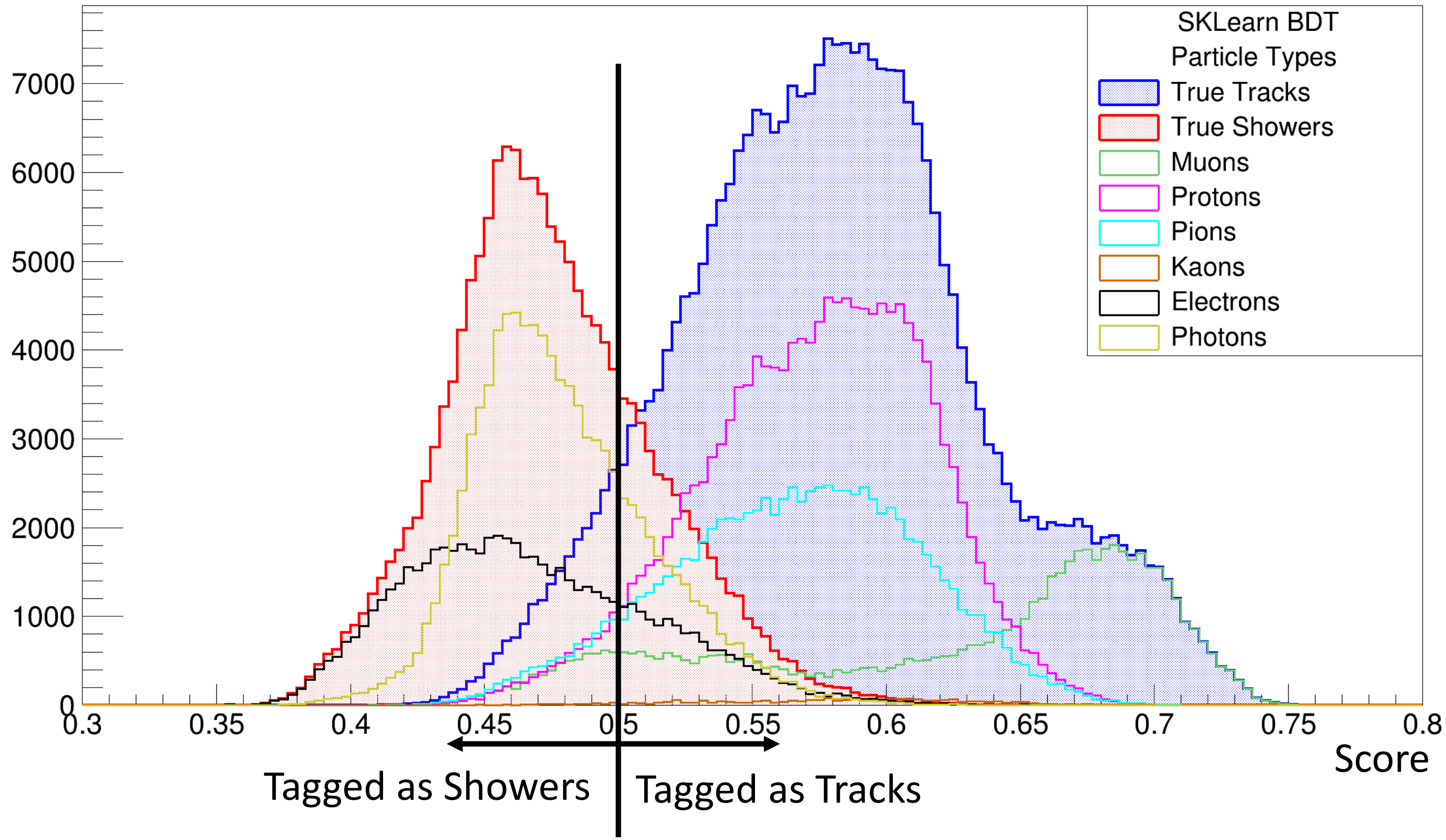
daughterParentNhitsRatio – 3D hits ratio between all downstream daughter pfos and parent pfo.

WARWICK
THE UNIVERSITY OF WARWICK

Current and Proposed approach to track/shower ID in Pandora

- Current Implementation
  - Basic cut flow approach
- MicroBooNE $\rightarrow$ Support Vector Machine approach
- Looking to implement similar ML approach for DUNE FD
- Proposed Implementation
  - Boosted Decision Tree approach using SciKit-Learn which Pandora supports
  - 13 variables
  - Training $\rightarrow$ 50% numu and 50% nue DUNE FD 1X2X6 MCC11 samples, completeness and purity $\geq$ 80%, fiducial volume cuts
  - Testing $\rightarrow$ 50% numu and 50% nue DUNE FD 1X2X6 MCC11 samples, no completeness and purity cuts, no fiducial volume cuts

SKLearn BDT Distribution



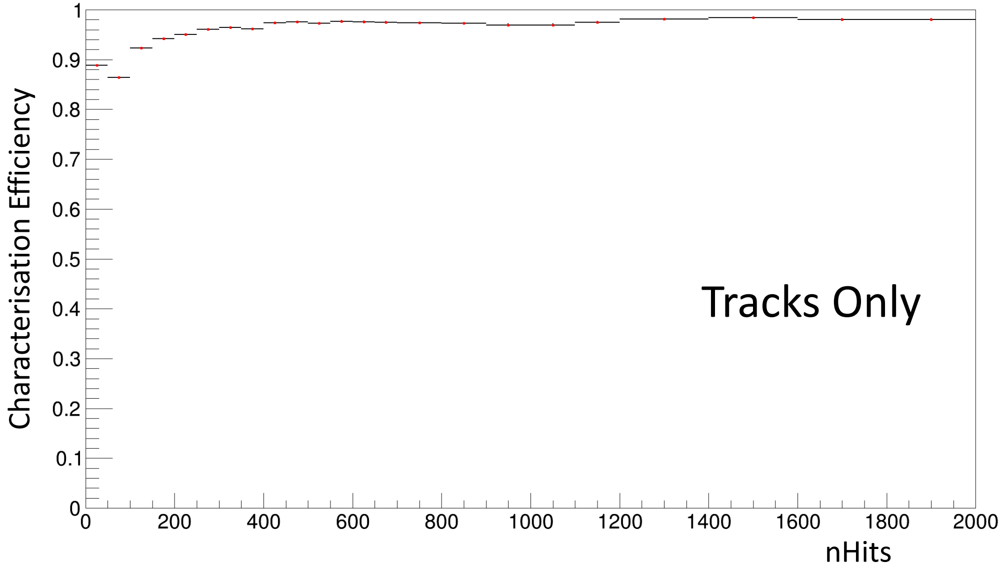
WARWICK
THE UNIVERSITY OF WARWICK

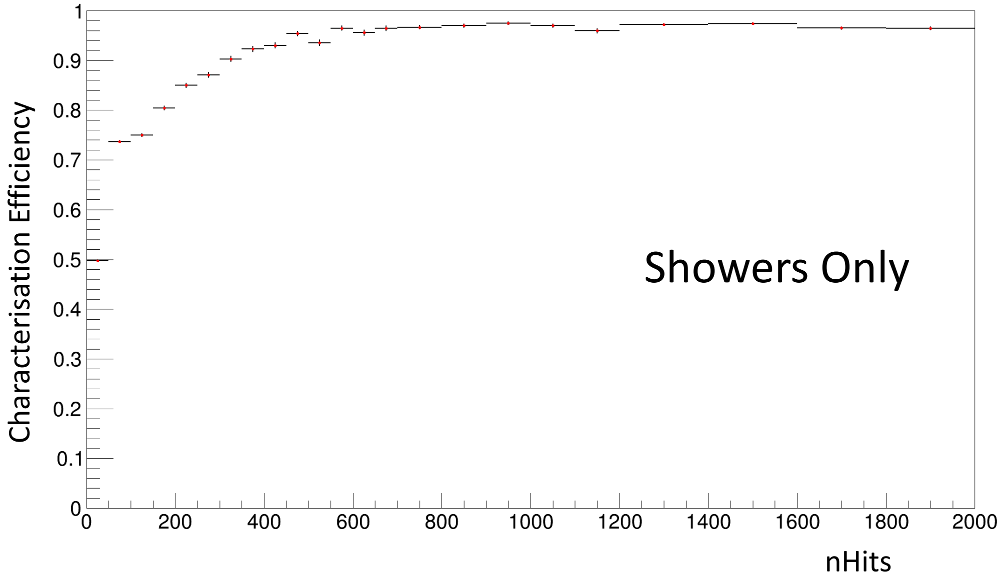
Efficiency Numbers

Key	T = Tracks	S = Showers	TT = True Tracks	TS = True Showers

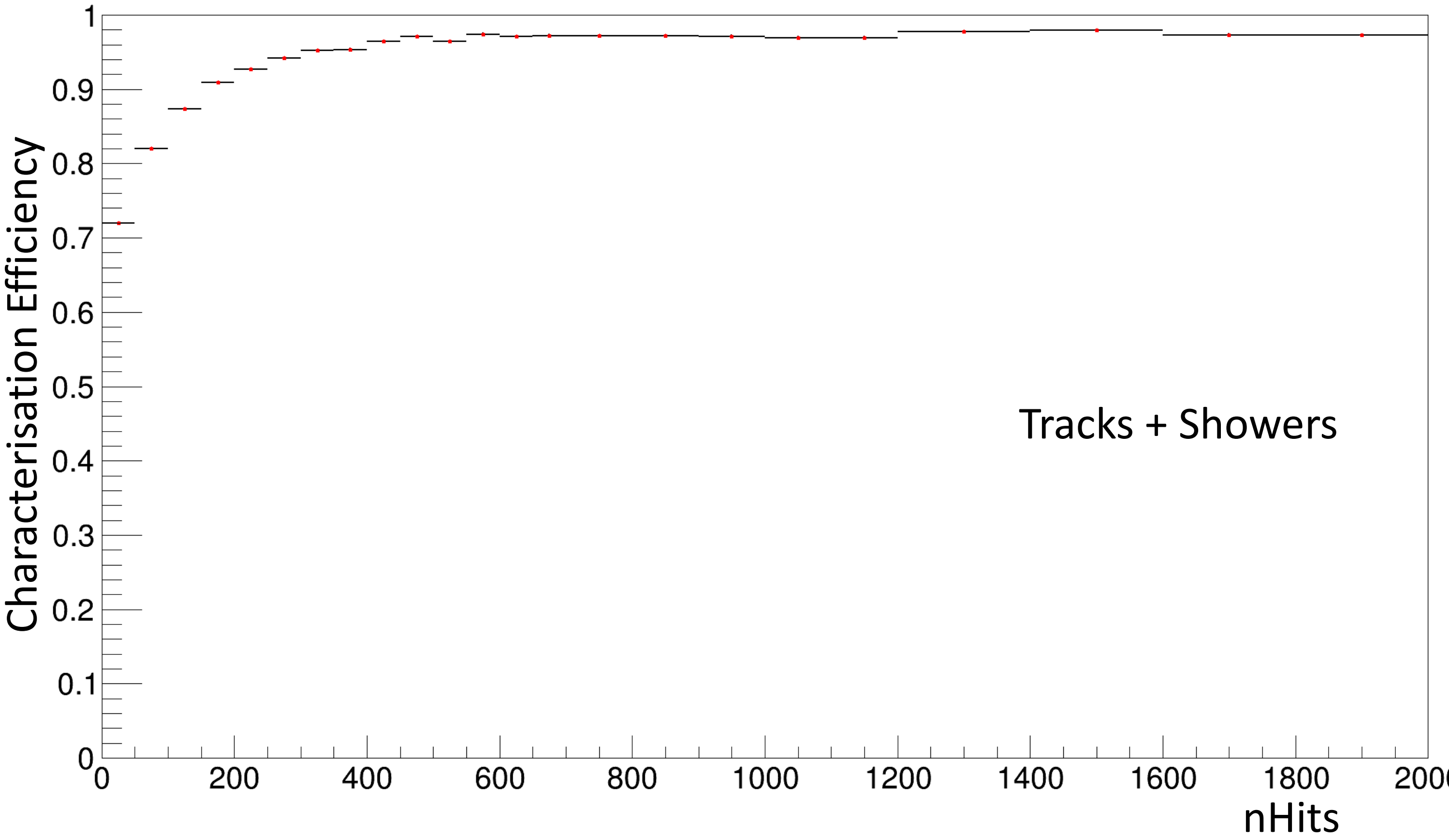
T/S Characterisation Approach	TT as T (#Pfos)	TT as S (#Pfos)	Efficiency (T only)	TS as S (#Pfos)	TS as T (#Pfos)	Efficiency (S only)	Total (#Pfos)	Efficiency (All Pfos)
Cut Based Approach	212900	100588	**0.679 ± 0.0008**	149980	13774	**0.916 ± 0.0007**	477242	**0.760 ± 0.0006**
Root TMVA BDT	283724	29764	**0.905 ± 0.0005**	128639	35115	**0.786 ± 0.0010**	477242	**0.864 ± 0.0005**
SKLearn BDT	290678	22810	**0.927 ± 0.0005**	120746	43008	**0.737 ± 0.0011**	477242	**0.862 ± 0.0005**

Efficiency vs nHits





nHits – number of reconstructed 3d hits



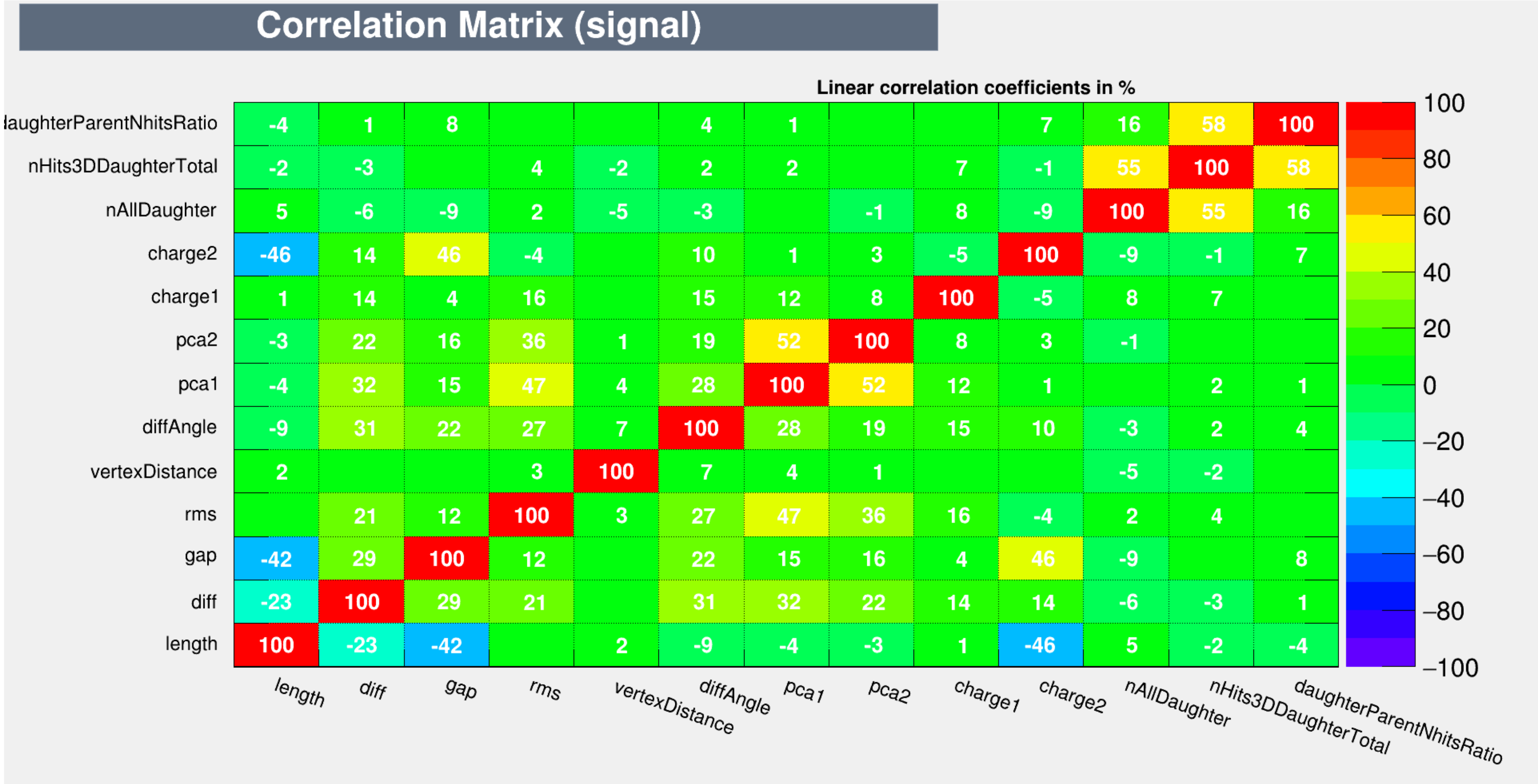
WARWICK
THE UNIVERSITY OF WARWICK

Summary/Future Works

- Cut Flow $\rightarrow$ BDT approach (SKLearn)
- Significant Improvements
- Test on ProtoDUNE MC/data
- Use Andy Chappell's work
- Alan Turing Institute (mid-Jan 2020)
- Any questions or comments are deeply appreciated

BACK UP SLIDES

Correlation Matrix for 13 variables



Definition of the variables

Topological

- length – 3D length of the PFO
- diff – Mean difference between the position of the hits and a straight line, divided by the straight line length
- gap – Average max gap distance, divided by 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 primary 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

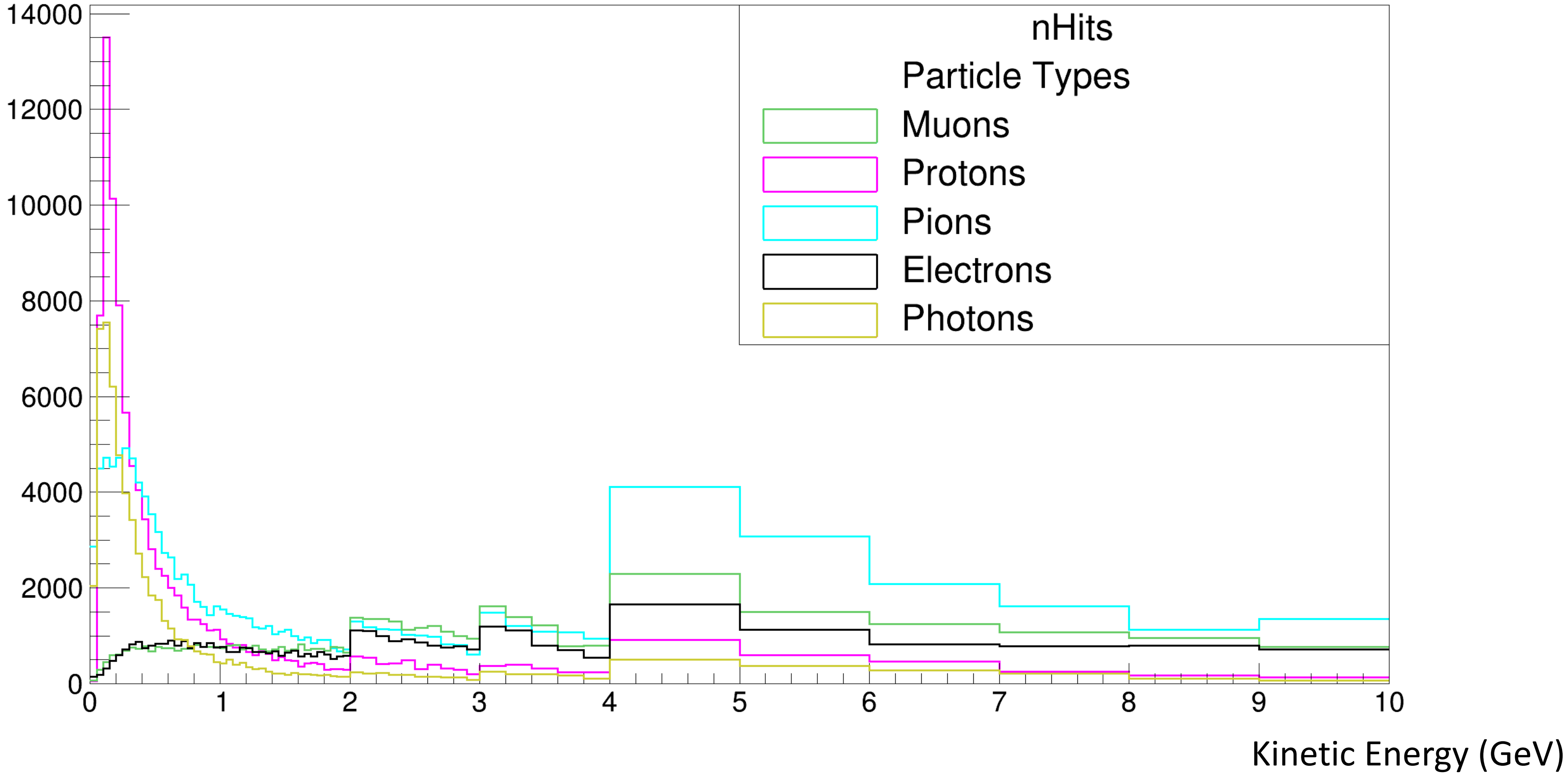
Calorimetric

- charge1 – Ratio between sigmaCharge $((charge - meanCharge)^2)$ and the mean charge in collection plane.
- charge2 – Ratio of charge in the last 10% of the PFO and the mean charge in the collection plane

Hierarchy

- nAllDaughter – total number of all downstream daughter pfos
- nHits3DDaughterTotal – total number of 3D hits in all downstream daughter pfos
- daughterParentNhitsRatio – 3D hits ratio between all downstream daughter pfos and parent pfo.

T/S Distribution for Kinetic Energy



T/S Distribution for nHits