Pandora ML training for HE production

Andy Chappell

26/02/2024   FD sim/reco
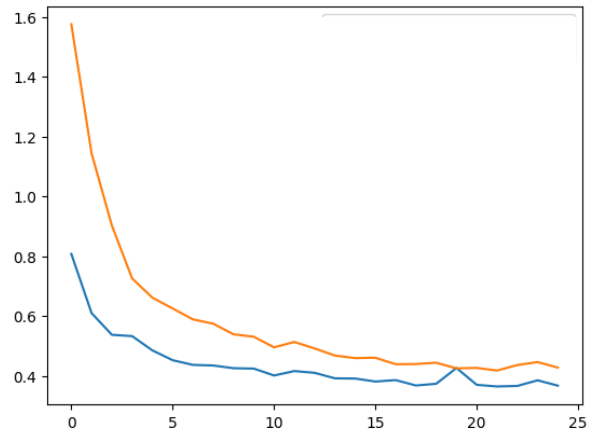
DEEP UNDERGROUND
NEUTRINO EXPERIMENT

WARWICK

Vertex training
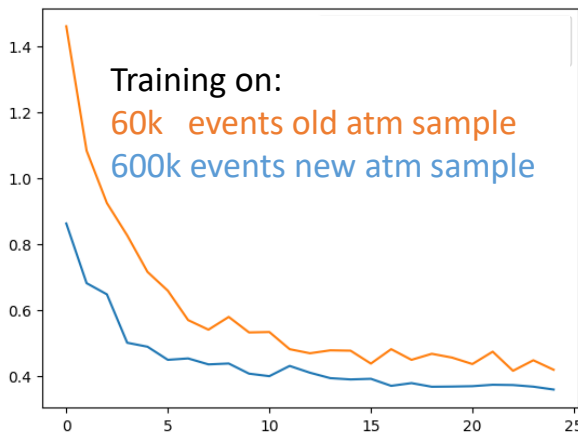(I Cheong Hong)

Training samples

- Samweb dataset:
  - higuera_fardet-hd__fd_mc_he_2023a__mc__hit-reconstructed__prodgenie_atmnu_ max_weighted_randompolicy_dune10kt_1x2x6.fcl__v09_79_00d02__preliminary
  - Trained on the first 14000 files
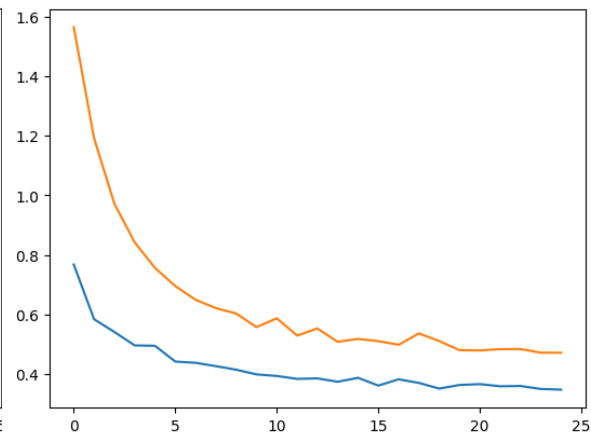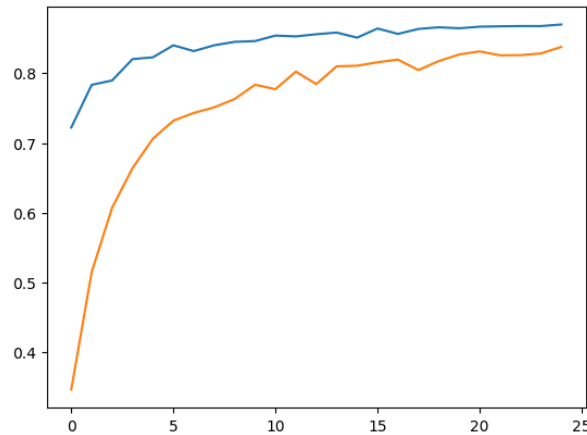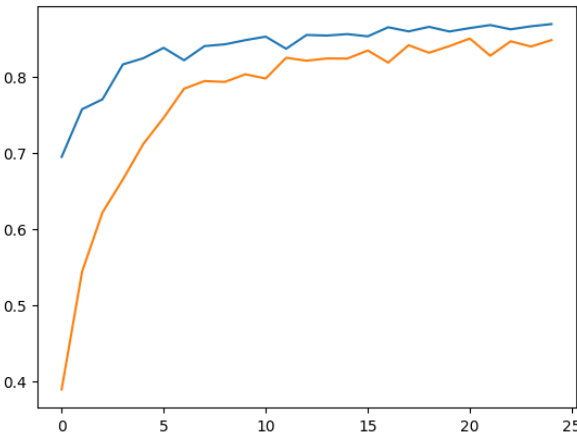 - Approximately 600K events in the training set

Pass 1: Network training

Training on:
60k   events old atm sample
600k events new atm sample

Pass 1: Confusion matrix

- For U view in pass1
  - other confusion matrices are similar
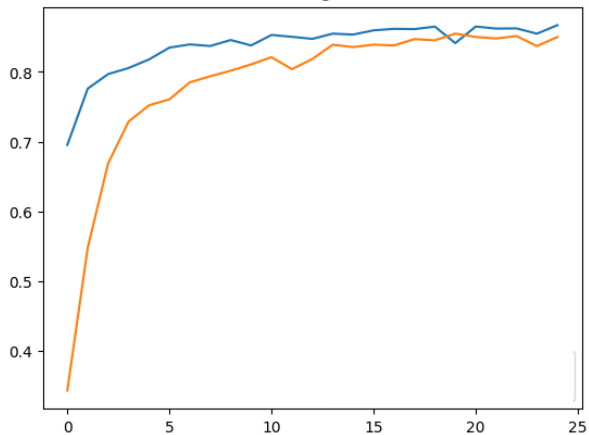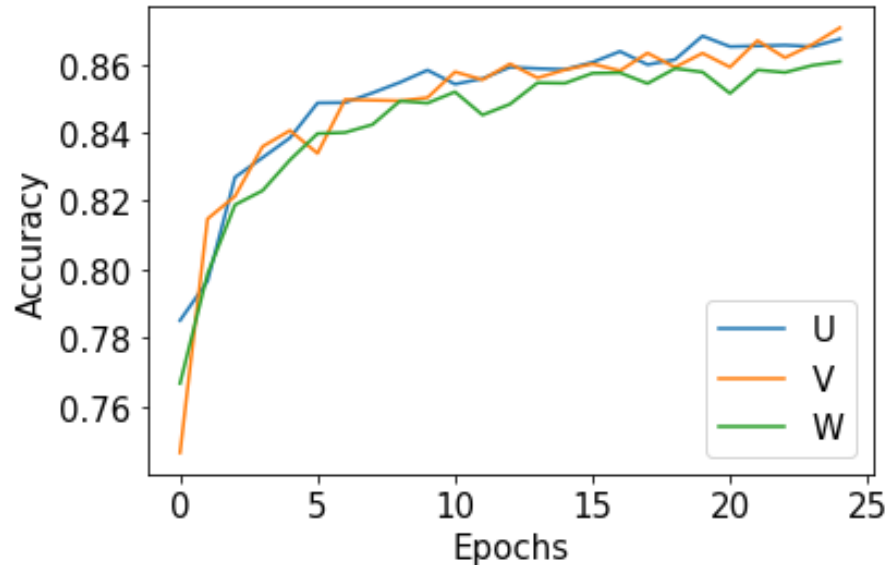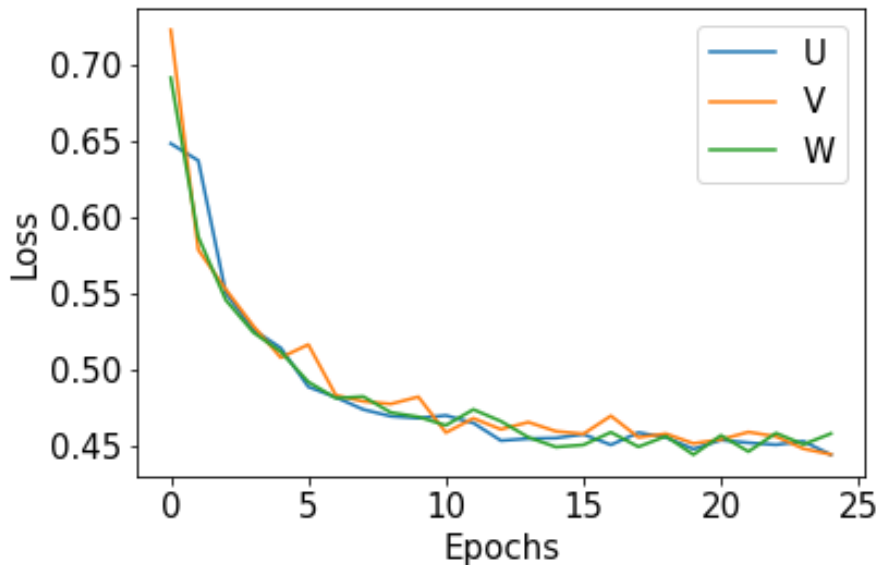
Pass 2: Network training

- Pass 2 loss functions and per class accuracies
  - a little surprised to see the collection plane yield slightly lower accuracy, but overall performance remains high
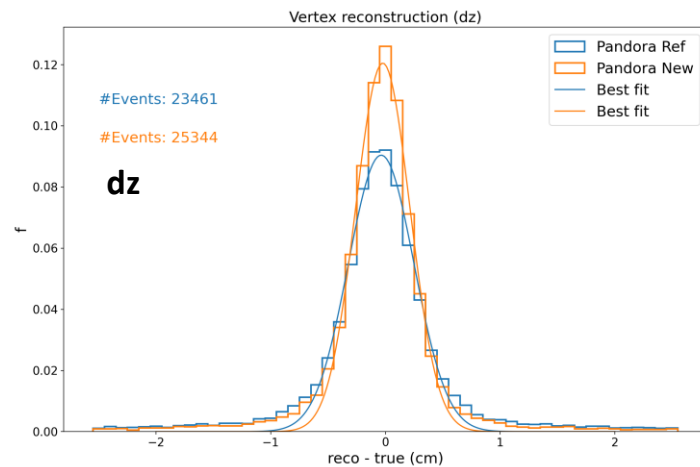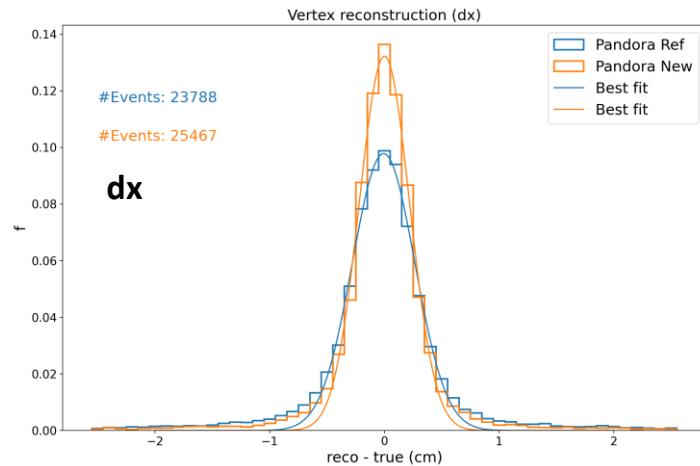
Vertex
reconstruction
performance

Testing samples

- Samweb dataset:
  - higuera_fardet-hd__fd_mc_he_2023a__mc__hit-reconstructed__prodgenie_atmnu_ max_weighted_randompolicy_dune10kt_1x2x6.fcl__v09_79_00d02__preliminary
  - Tested on events from 500 files after the first 15000 training files
  - For vertexing only events with a true vertex within the fiducial volume are considered (~32,000)
  - <u>The reference performance uses the previously trained DL vertexing model</u>

Vertex deltas

- Improved performance in all dimensions over previous network
  - About 10% more precisely reconstructed vertices
  - Improved resolution
  - No longer biased in Y
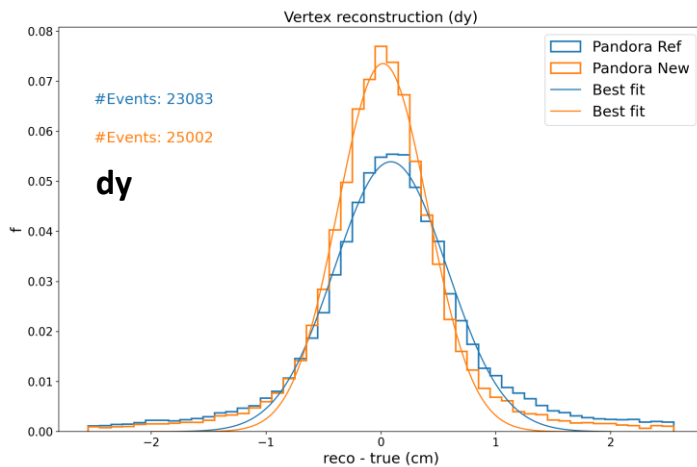
Vertex deltas

- Larger training set yields notably improved performance

- About 10% of events have no reconstructed vertex
  - Typically implies very low visible hit count
 - Needs further investigation to confirm

- **Ref**: previous DL vertex training
- **New**: latest   DL vertex training
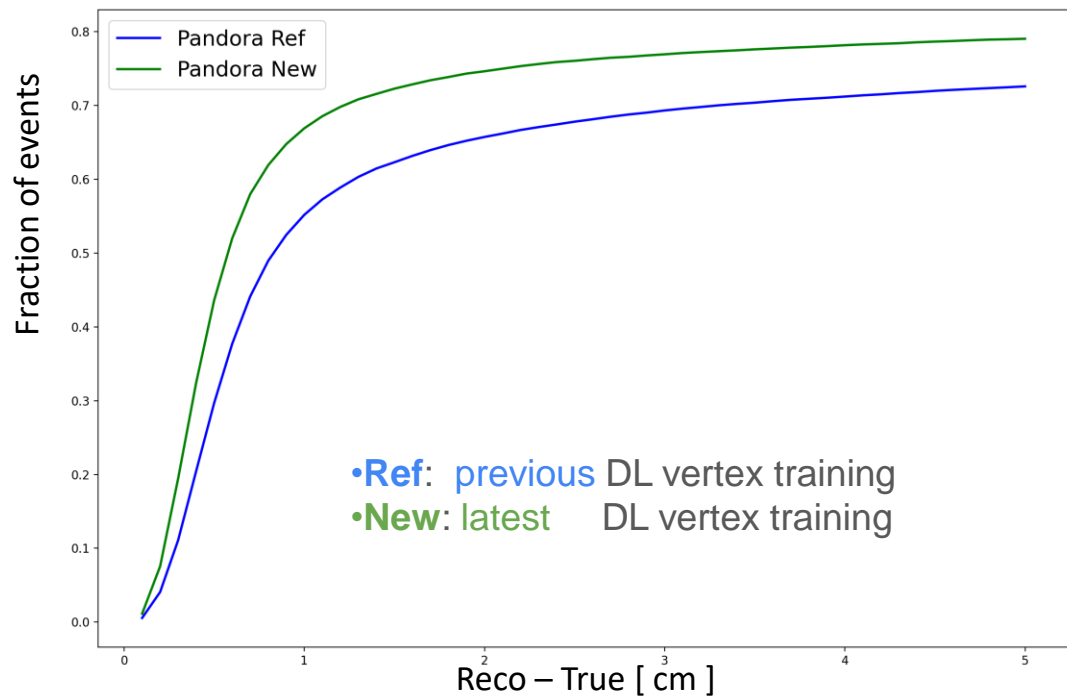
PFP
characterization
training

Training samples

- Samweb dataset:
  - higuera_fardet-hd__fd_mc_he_2023a__mc__hit-reconstructed__prodgenie_atmnu_ max_weighted_randompolicy_dune10kt_1x2x6.fcl__v09_79_00d02__preliminary
  - Tested on events from 1000 files after the first 14000 training files
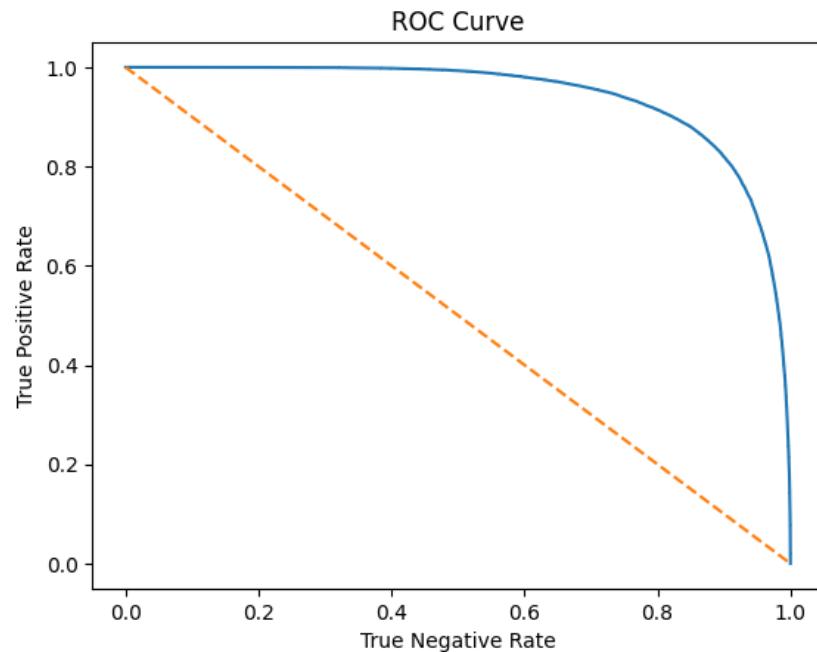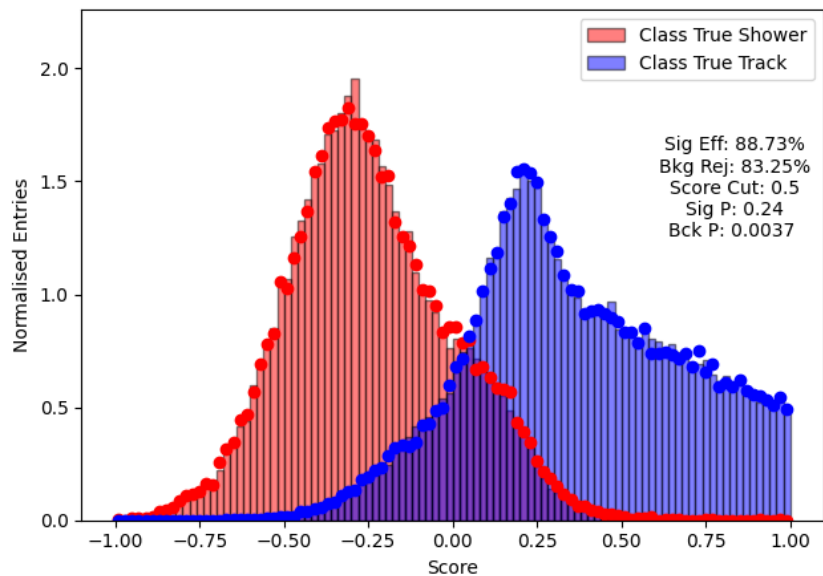
Feature correlations

Hyper parameter investigations
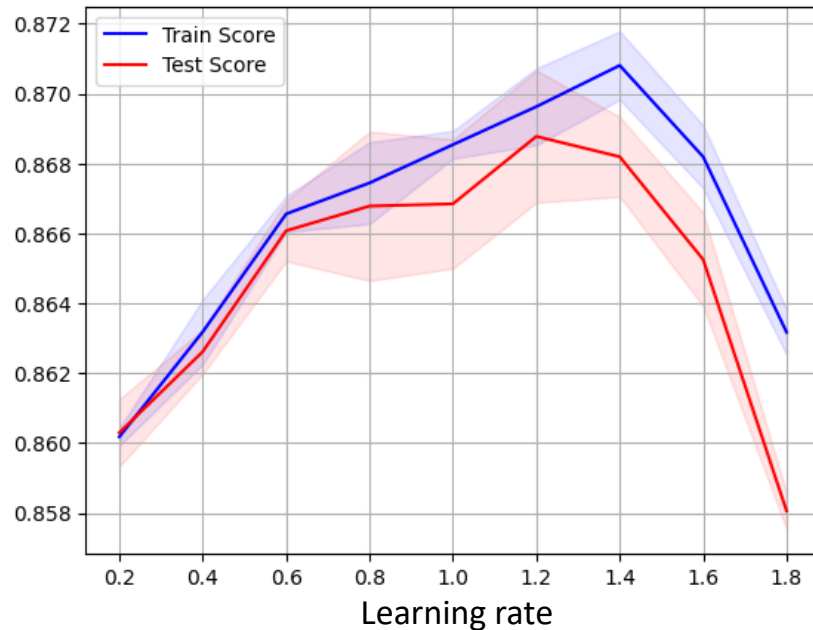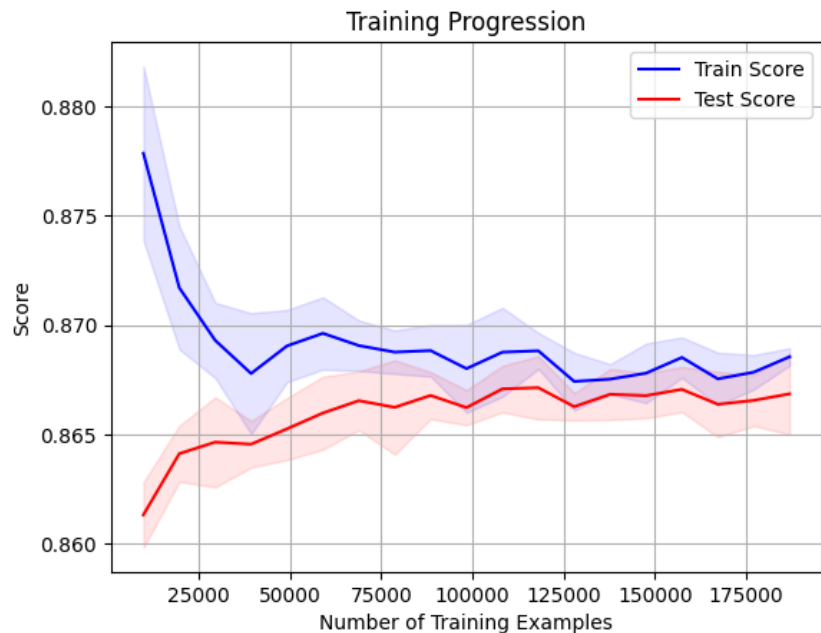
- ~190,000 PFPs in each of the training and test samples for final training
- Learning rate of 1.2

Score distribution and ROC curve
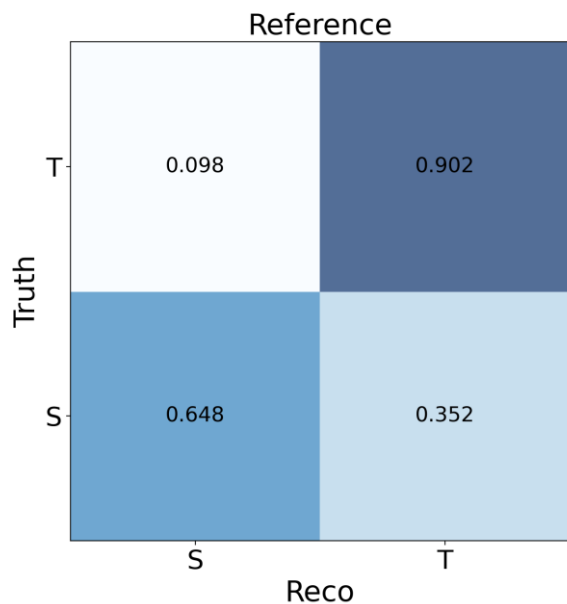
PFP characterization
reconstruction performance

Testing samples

- Samweb dataset:
  - higuera_fardet-hd__fd_mc_he_2023a__mc__hit-reconstructed__prodgenie_atmnu_max_weighted_randompolicy_dune10kt_1x2x6.fcl__v09_79_00d02__preliminary
  - Tested on events from 500 files after the first 15000 training files (same as vertexing)
  - The reference performance uses the previously (LBL) trained BDT

Score distribution and ROC curve
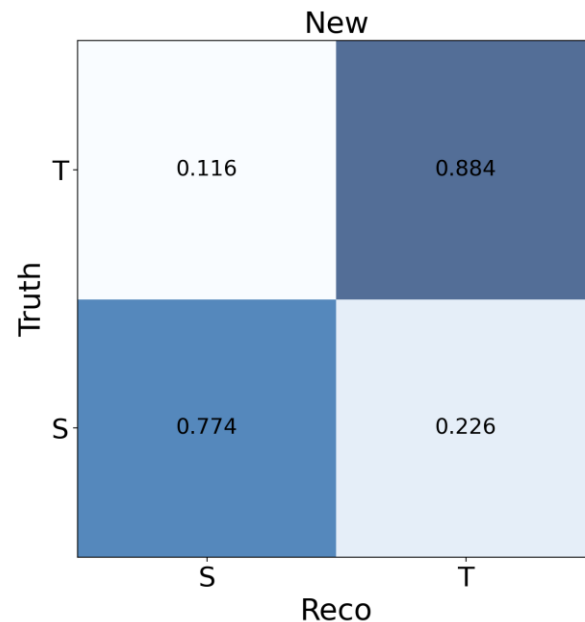
- Similar track characterization performance
- Much improved shower characterization performance

Summary and next steps

- Substantial improvements in vertexing and PFP characterization performance

- Have requested that new model files be added to DUNE StashCache (pending)

- Feature branches exist for dunereco and dunesw to use the updated models
  - dunesw: https://github.com/AndyChappell/dunesw/tree/feature/atmos_prod_phase2
  - dunereco: https://github.com/AndyChappell/dunereco/tree/feature/atmos_prod_phase2
  - Introduces two new fcl files for reco2:
    - standard_reco2_atmos_dune10kt_1x2x6.fcl: Long-term fcl file
    - reco2_atmos_dune10kt_1x2x6_geov5.fcl: What we should use for this production to accommodate the fact that the head of dunesw develop now uses a v6 geometry, while the production has been run with v5