Tau-Lepton Identification and Decay Mode Classification using Graph Neural Networks

Robert McNulty

Supervisors:

Dr. Nikolaos Rompotis, Prof. Monica D'Onofrio

With contributions from: *Dr. Joseph Carmignani*

Introduction

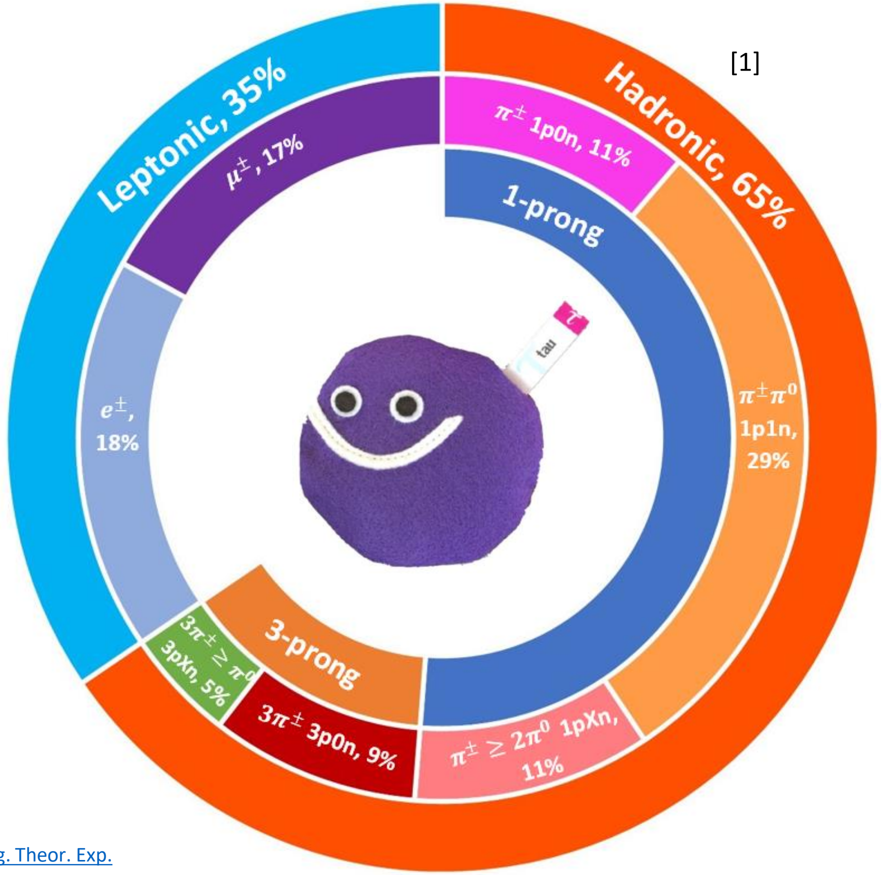
- 2nd Year LIV.INNO Student
- Dual Funded PhD – 50/50 split between LIV.INNO and ARO (Industry Placement)
- Physics Project (ATLAS): Tau-Lepton Identification and Decay Mode Classification using Graph Neural Networks
- Industry Project (ARO/UoL): Use of AI models for predictions in different eye-related disease progressions

TauJetGraphs

GNN Developed by Dr. Joseph Carmignani

Tau-Leptons and QCD Dijets

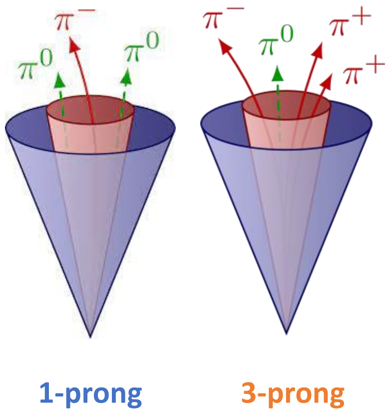
- Leptonic Tau decays, τ_{lep}, (35%) produce $e^{\pm}$, $\mu^{\pm}$ and corresponding ν's
- Hadronic Tau decays, τ_{had}, (65%) produce 1 or 3 $\pi^{\pm}$ (1- & 3-prong decays) and maybe a few π^0



[1]

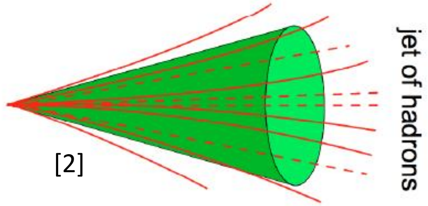
τ_{had} Decay Cones:

- Highly collimated – narrow cone
- Small cross-section
- Low multiplicity



Dijet Production & Cone

- Main background source of fake τ_{had} are jets from QCD
- Shower shape can mimic/drown-out the shower shape of π's from τ_{had}
- Fragment into multiple hadrons (high multiplicity)
- High production cross-section for dijets
- Wider cone area



[2]

[1] R.L. Workman *et al.* (Particle Data Group), Prog. Theor. Exp. Phys. **2022**, 083C01 (2022) and 2023 update
[2] Joern Mahlstedt and the ATLAS collaboration 2014 *J. Phys.: Conf. Ser.* **513** 012021, **DOI** 10.1088/1742-6596/513/1/012021

Motivations and Goals

Motivations:

- BR for τ_{had} almost twice as much as τ_{lep}
- Identification (ID) important for several areas of research, such as:
  - $H \rightarrow \tau\tau$ production CERN-EP-2021-217
  - Di-Higgs searches with $b\bar{b}\tau^{+}\tau^{-}$

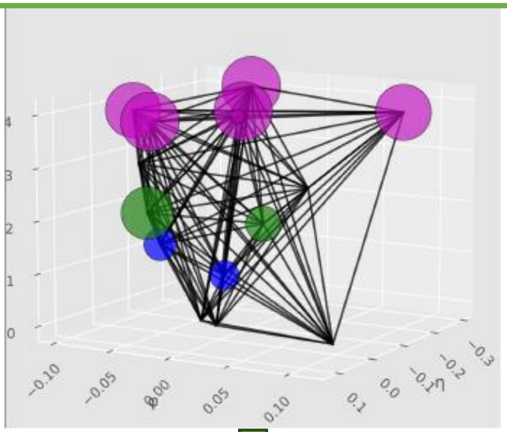
Goals:

- To study further the unification of the Decay Mode Classification (DMC) and ID Neural Network into a single Graph Neural Network (GNN) algorithm
- It should handle τ_{had}-candidates with 1 & 3 tracks
- The final classifier should also be able to classify 5 decay modes (1p0n, 1p1n, 1pXn, 3p0n, 3pXn) and background QCD jets (dijets)

TauJetGraphs – Graph & Model Structure

- Each node has 74 attributes – both local (physics object) and global (jet) variables
- For each layer, nodes that are within a predefined distance, $\Delta\mathrm{R} = 0.4$, are connected by an edge
- Edges are then added between nodes across layers, constructing a 3D graph

TauJetGraphs NN Model Pipeline*



BatchNorm Layer

ARMA Conv Layers

Global Mean Pooling Layers

FC Linear Layer

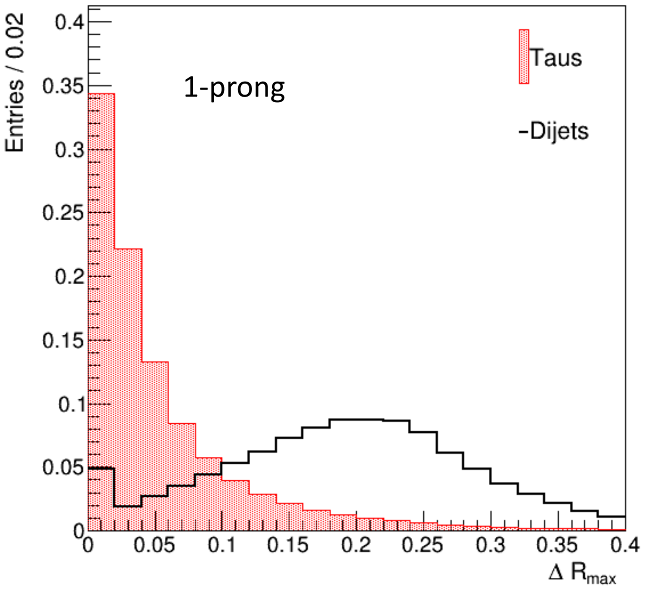
Output Score

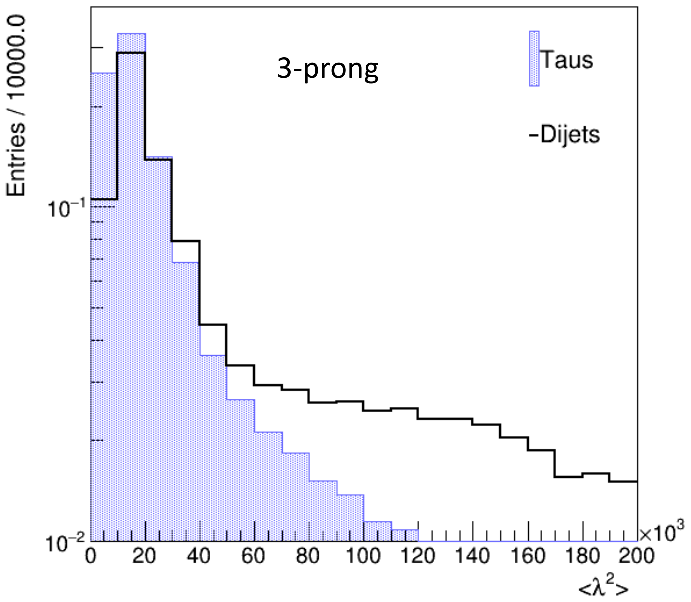
Layer	Description
0	$h^{\pm}$ candidates (from τ_{had} tracks)
1	π^0 candidates (from τ_{had} decay)
2	γ-energy deposits in EM Calorimeters (originating from π^0)
3	e^-e^+ tracks (from $\gamma \rightarrow e^+e^-$)
4	Energy deposits, E_T, in the Calorimeter Layers

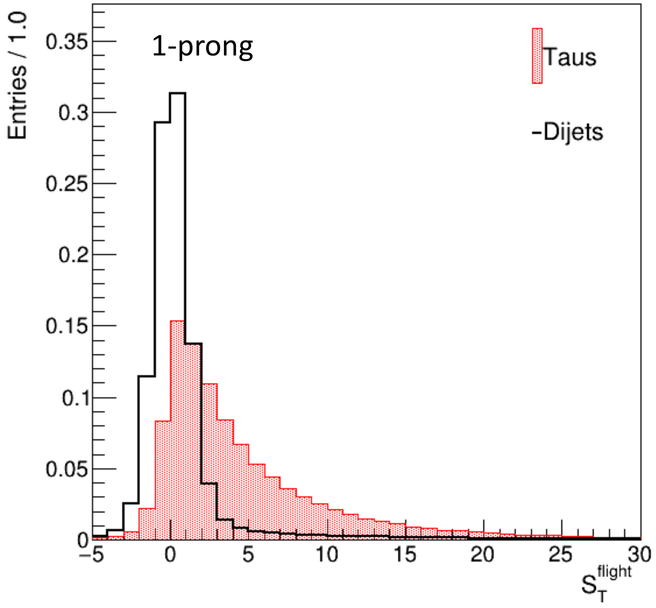
- Each node represents a reconstructed physics object within a given layer of the graph, i.e. each node in Layer 0 is a reconstructed $h^{\pm}$ candidate, and so on

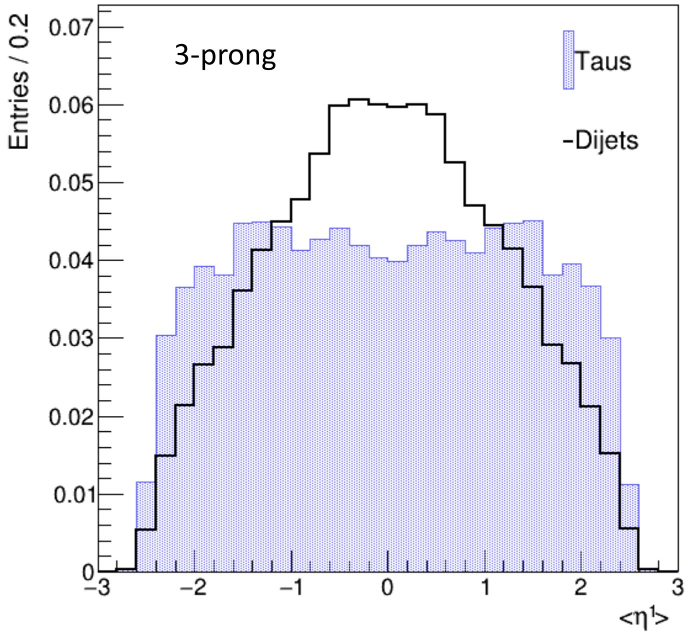
* $\mathbf{p_T}$ can also be represented as the **size of a node**

Some Example Distributions





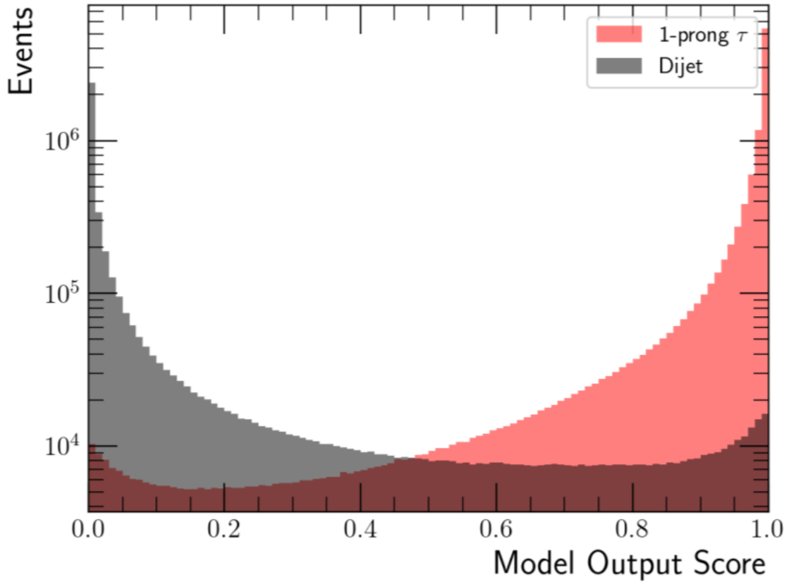




Results (Identification)

- 3-prong performance shows similar performance to 1-prong
- Model shows good general separation between true and fake τ_{had} candidates

1-prong GNN Output Scores

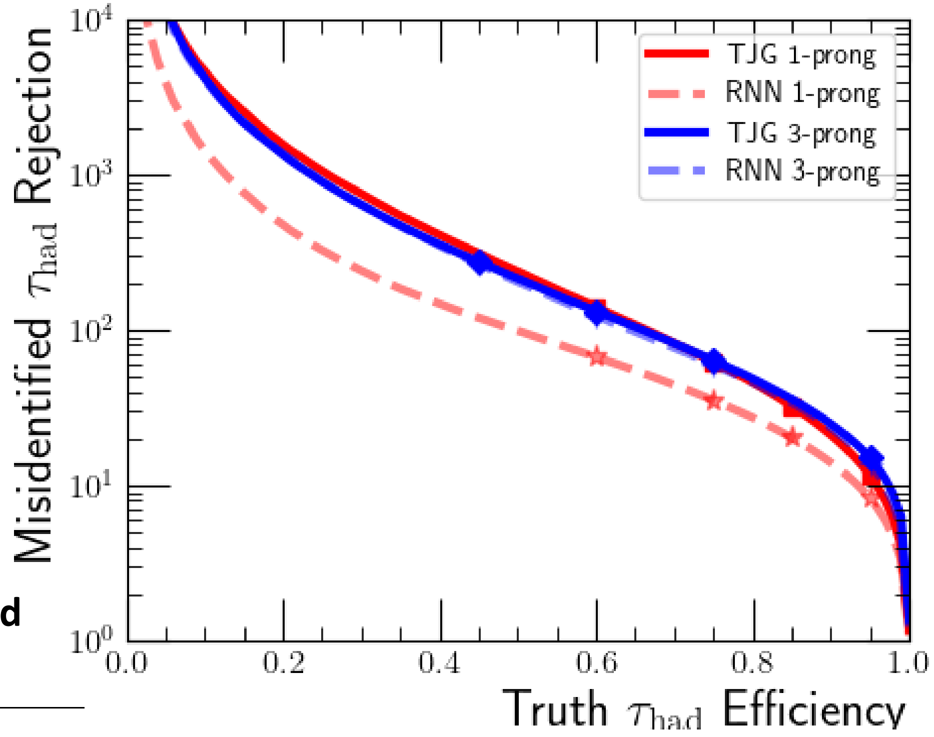


- Rejection = inverse of background selection efficiency
- Receiver Operating Characteristic (ROC) Curve displays the rejection of misidentified background samples as a function of the signal efficiency

At 60% Efficiency

- 1-prong rejection improves in GNN by order of 10
- 3-prong rejection has some but no significant improvement

TauID GNN vs RNN ROC Curve



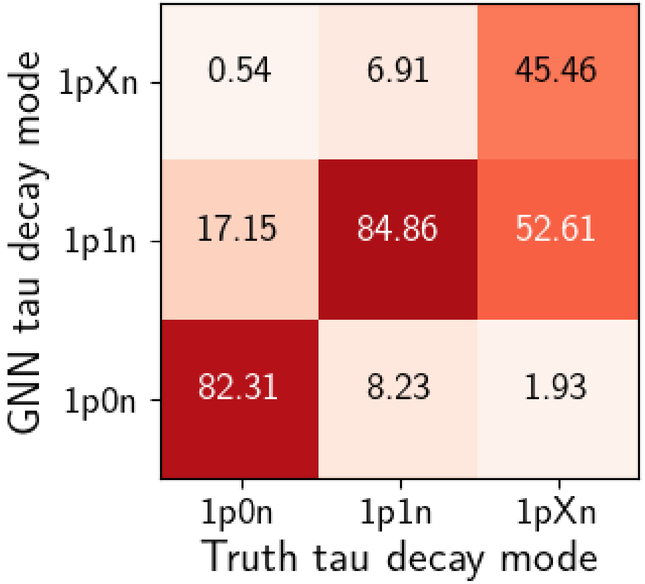
TauID GNN Background Rejection at Specified Efficiencies

	1-prong				3-prong			
Efficiency	60 %	75 %	85 %	95 %	45 %	60 %	75 %	95 %
Rejection	140.96	62.24	32.40	11.62	275.47	132.81	63.95	15.06

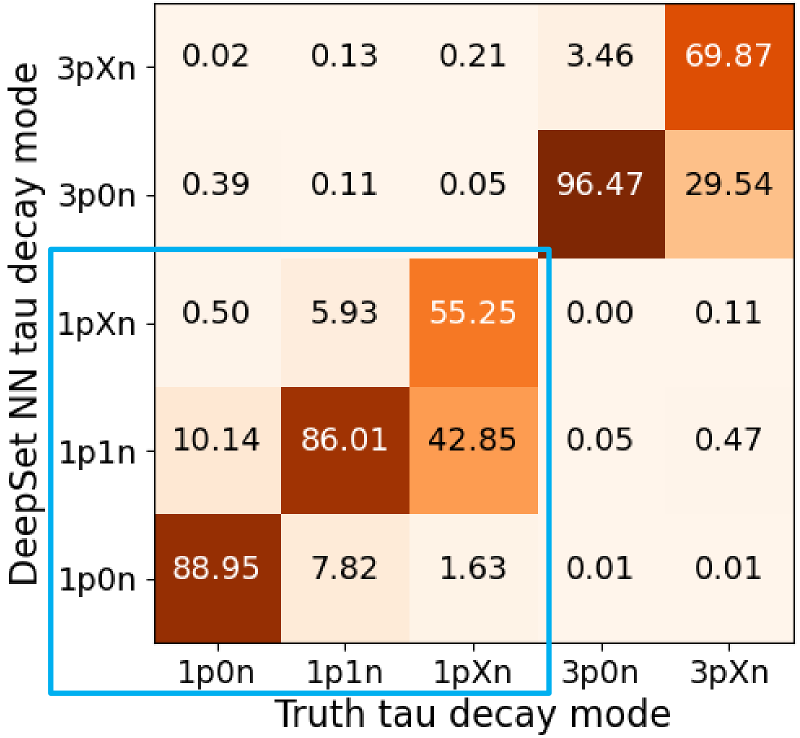
Results (Decay Mode Classification)

- Results shown are for 1-prong at 75% Efficiency
- Efficiency matrix: each column sums to 100%
- Similar performance for 1-prong candidates at same efficiency
- **Efficiency (Recall)** = $\frac{tp}{tp+fn}$
  - *tp* = true positive (signal sample identified as signal)
  - *fn* = false negative (signal sample identified as background)

GNN (1-prong)



Current Method - DSNN



[ATL-PHYS-PUB-2022-044]

Industry Placement and Upcoming Project

Work with ARO and New Project

ARO Technology. Only better.

- Role with ARO is Research Developer.
- This has required working with ARO's clients on their projects.
- As such, I have received training e.g. in the use of the following:
  - PRTG Dashboards
  - REDCap Databases
  - Enovacom Integration Engine
- N.B. the project involving Enovacom has since moved onto using Apache Airflow software.

Upcoming Project

with Dr. Philip Burgess and Prof. Yalin Zheng

- Use of an AI model for prediction of progression of Age-related Macular Degeneration[1]. [3, 4]
- Validation of AI model for progression of diabetic retinopathy[2] to treatment.

[1] Age-related macular degeneration (AMD) is a common condition that affects the middle part of your vision. More information can be found here on the NHS Website.

[2] Diabetic retinopathy is a complication of diabetes, caused by high blood sugar levels damaging the back of the eye (retina). This can cause blindness if left untreated. More information can be found here on the NHS Website.

[3] Bridge J, Harding S, Zheng Y. Development and validation of a novel prognostic model for predicting AMD progression using longitudinal fundus images. *BMJ Open Ophthalmology* 2020;**5**:e000569. doi: 10.1136/bmjophth-2020-000569

[4] Bridge, J., Harding, S., Zheng, Y. (2021). End-to-End Deep Learning Vector Autoregressive Prognostic Models to Predict Disease Progression with Uneven Time Intervals. In: Papież, B.W., Yaqub, M., Jiao, J., Namburete, A.I.L., Noble, J.A. (eds) Medical Image Understanding and Analysis. MIUA 2021. Lecture Notes in Computer Science(), vol 12722. Springer, Cham. https://doi.org/10.1007/978-3-030-80432-9_38

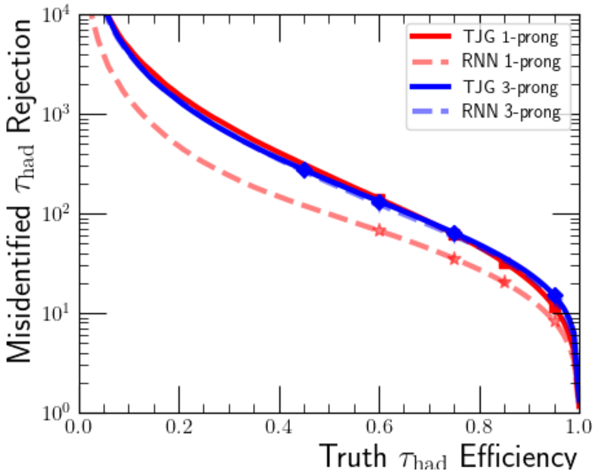
Summary

Summary and Future Steps

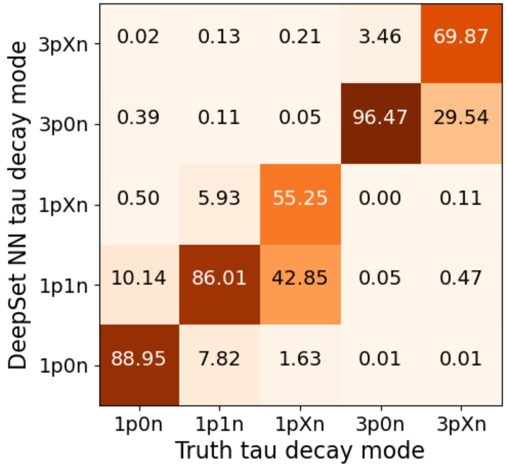
- TauJetGraphs progressing nicely:
  - Showing improvements over current methods
  - Further analysis currently taking place (see Mehul's 1st Year talk, as well as Monica's and Jordy's ATLAS talks)
- Currently writing up TauJetGraphs work for thesis
- Work at ARO proving to be good work/industrial experience
- Project on the use of AI models for predictions in different eye-related disease progressions to start soon

	1-prong				3-prong			
Efficiency	60 %	75 %	85 %	95 %	45 %	60 %	75 %	95 %
Rejection	140.96	62.24	32.40	11.62	275.47	132.81	63.95	15.06

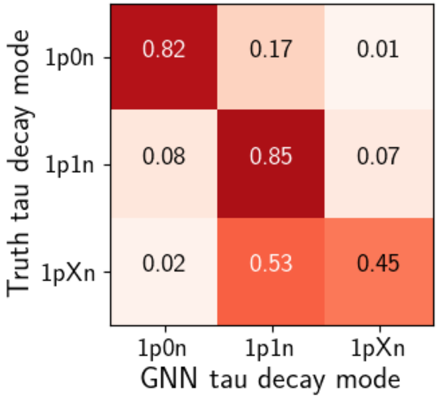
GNN vs RNN



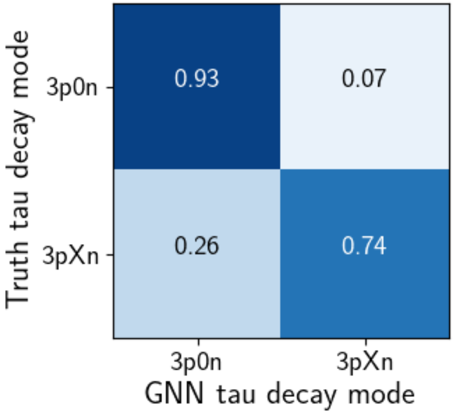
DSNN DMC



GNN DMC (1-prong)



GNN DMC (3-prong)



Thank you

Backup

Backup: Input Variables

Inputs used for TauID RNN, ATL-PHYS-PUB-2019-033,
also used in TauJetGraphs GNN

	Observable	1-prong	3-prong
Track inputs	$p_{\mathrm{T}}^{\text{seed jet}}$	•	•
	$p_{\mathrm{T}}^{\text{track}}$	•	•
	$\Delta\eta^{\text{track}}$	•	•
	$\Delta\phi^{\text{track}}$	•	•
	$\lvert d_0^{\text{track}}\rvert$	•	•
	$\lvert z_0^{\text{track}} \sin\theta\rvert$	•	•
	$N_{\text{IBL hits}}$	•	•
	$N_{\text{Pixel hits}}$	•	•
	$N_{\text{SCT hits}}$	•	•

	Observable	1-prong	3-prong
Cluster inputs	$p_{\mathrm{T}}^{\text{jet seed}}$	•	•
	$E_{\mathrm{T}}^{\text{cluster}}$	•	•
	$\Delta\eta^{\text{cluster}}$	•	•
	$\Delta\phi^{\text{cluster}}$	•	•
	λ_{cluster}	•	•
	$\langle\lambda^2_{\text{cluster}}\rangle$	•	•
	$\langle r^2_{\text{cluster}}\rangle$	•	•

	Observable	1-prong	3-prong
High-level inputs	$p_{\mathrm{T}}^{\text{uncalibrated}}$	•	•
	f_{cent}	•	•
	$f^{-1}_{\text{leadtrack}}$	•	•
	$\Delta R_{\max}$	•	•
	$\lvert S_{\text{leadtrack}}\rvert$	•	
	$S_{\mathrm{T}}^{\text{flight}}$		•
	$f_{\text{iso}}^{\text{track}}$	•	•
	$f_{\text{track}}^{\text{EM}}$	•	•
	$p_{\mathrm{T}}^{\text{EM+track}}/p_{\mathrm{T}}$	•	•
	$m^{\text{EM+track}}$	•	•
	m^{track}		•

Variable	Description
$p_T(\tau_{had})$	p_T of the τ_{had} (using calorimeter based $\tau_{had\text{-}vis}$ energy scale)
p_T(object)	p_T of the object
$\Delta\phi(\text{object}, \tau_{had})$	Distance between the object and τ_{had} in ϕ
$\Delta\eta(\text{object}, \tau_{had})$	Distance between the object and τ_{had} in η
$\Delta\phi$(object, trackECal)	Distance between the object and the extrapolation of highest-p_T τ_{had} track to EM calorimeter in ϕ
$\Delta\eta$(object, trackECal)	Distance between the object and the extrapolation of highest-p_T τ_{had} track to EM calorimeter in η

Physics object kinematic variables

$\langle\eta^1\rangle$	First moment in η in cluster shower axis
$\log(\langle r^2\rangle)$	Second moment in the radial distance of cluster cells from the shower axis
$\Delta\theta$	Distance in θ between the EM shower axis and the vector pointing from the primary vertex to the centre of the shower
$\log(\lambda_{centre})$	Distance of the cluster shower centre from the calorimeter front face measured along the shower axis
$\langle\lambda^2\rangle$	Mean distance of a cell from the shower centre along the shower axis
$\log(\langle\rho^2\rangle)$	Second moment in the cluster energy density, where $\rho = E^{cluster}/V^{cluster}$
f_{core}	Sum of energy fractions in the most energetic cells per sampling
f_{core}^{EM1}	Same as f_{core} but only consider EM1
$N_{pos,EM1}$	Number of cells with positive energy in EM1
$N_{pos,EM2}$	Number of cells with positive energy in EM2
E_{EM1}	Energy in the EM1 layer
E_{EM2}	Energy in the EM2 layer
$\langle\eta^1_{EM1}\rangle$ w.r.t. cluster	First moment in η in EM1 with respect to the cluster
$\langle\eta^1_{EM2}\rangle$ w.r.t. cluster	First moment in η in EM2 with respect to the cluster
$\log(\langle\eta^2_{EM1}\rangle)$ w.r.t. cluster	Second moment in η in EM1 with respect to the cluster
$\log(\langle\eta^2_{EM2}\rangle)$ w.r.t. cluster	Second moment in η in EM2 with respect to the cluster

Variables used in Decay Mode Classification DSNN, ATL-PHYS-PUB-2022-044, also utilised in TauJetGraphs GNN

Neutral pion cluster variables

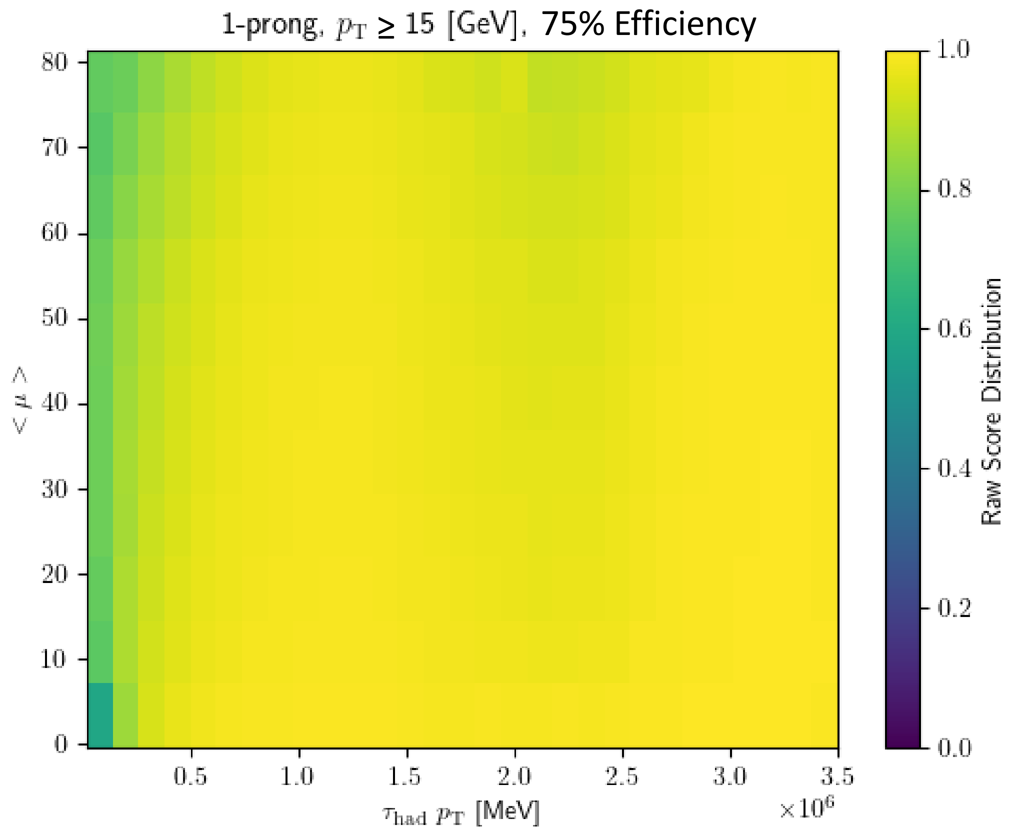
Backup: Score Bias

Accounting for Bias (1)

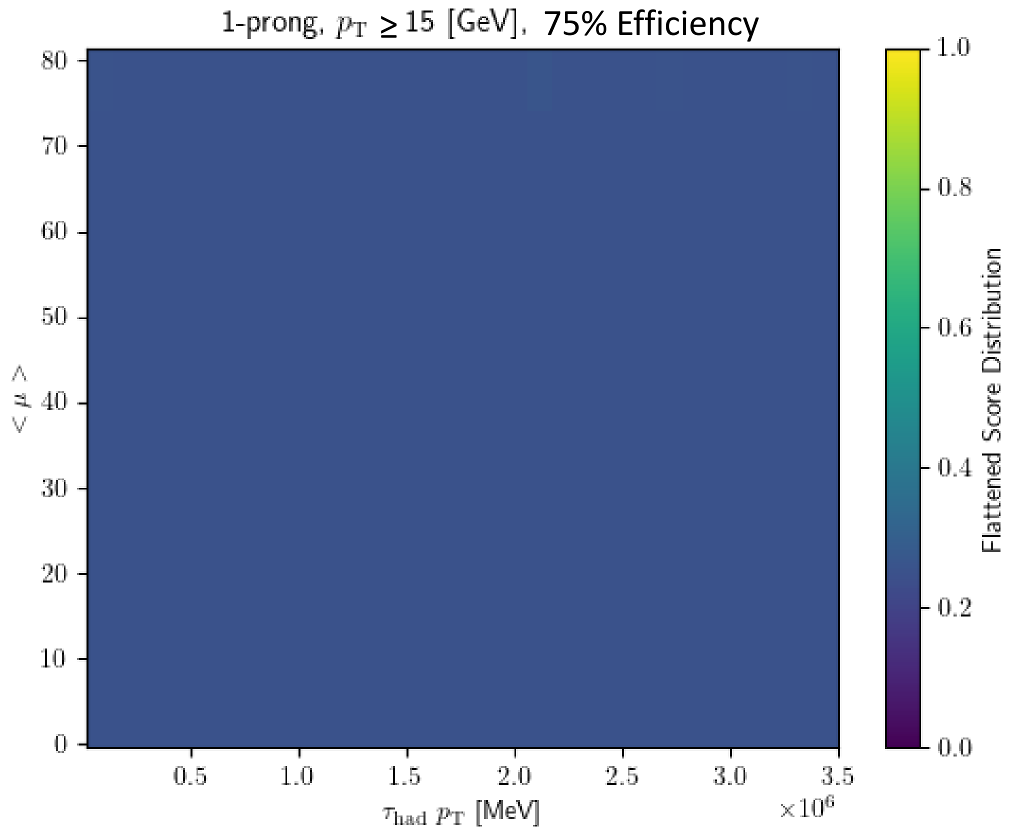
- Performance is typically measured through efficiency, purity, and background rejection
- Observations have shown that classification methods typically favour high-p_T events for a higher score than for those with a lower p_T
- This bias is accounted for by a transformation on the output scores for a given efficiency:
  - A 2D histogram of transverse momentum, p_T, and the average number of interactions per bunch-crossing, μ, is created from the signal dataset (slide 17)
  - The top percentage (given by the desired efficiency) of samples in each bin is taken, and the score threshold of this bin is determined by the sample with the lowest score in this selection.
  - The scores of each sample is then transformed w.r.t. the desired efficiency.
  - Background samples are then treated in a similar fashion, however the bin threshold scores from the signal histogram are used as the threshold for the background histogram
    - Samples which pass this selection are labelled as signal in the predictions for the given efficiency (referred to as “False Positives”)
    - From here, the background rejection for the given efficiency is determined

Accounting for Bias (2)

Signal sample score distribution before transformation

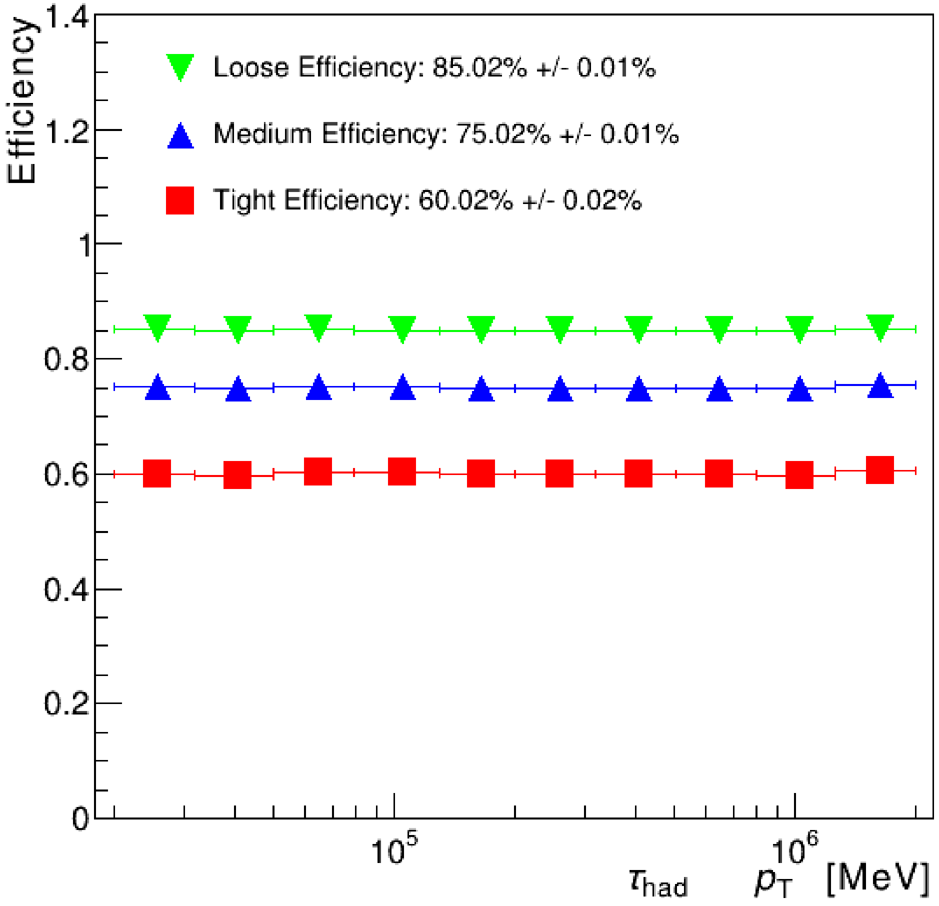


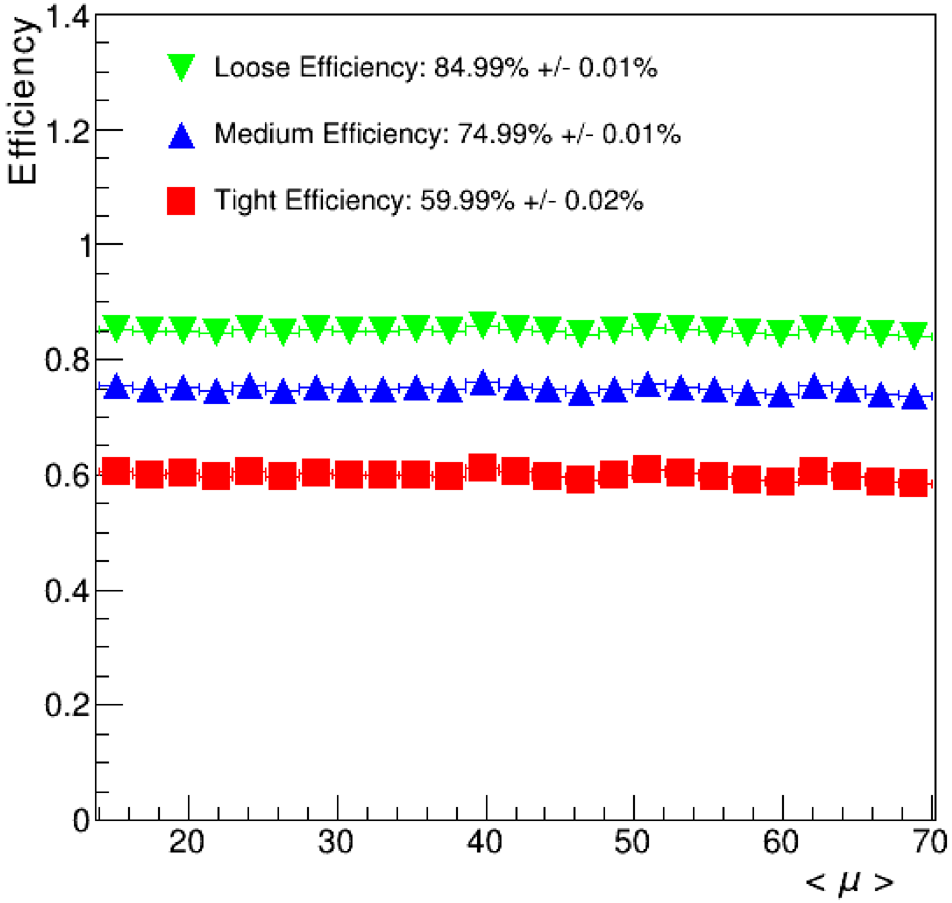
Signal sample score distribution after transformation



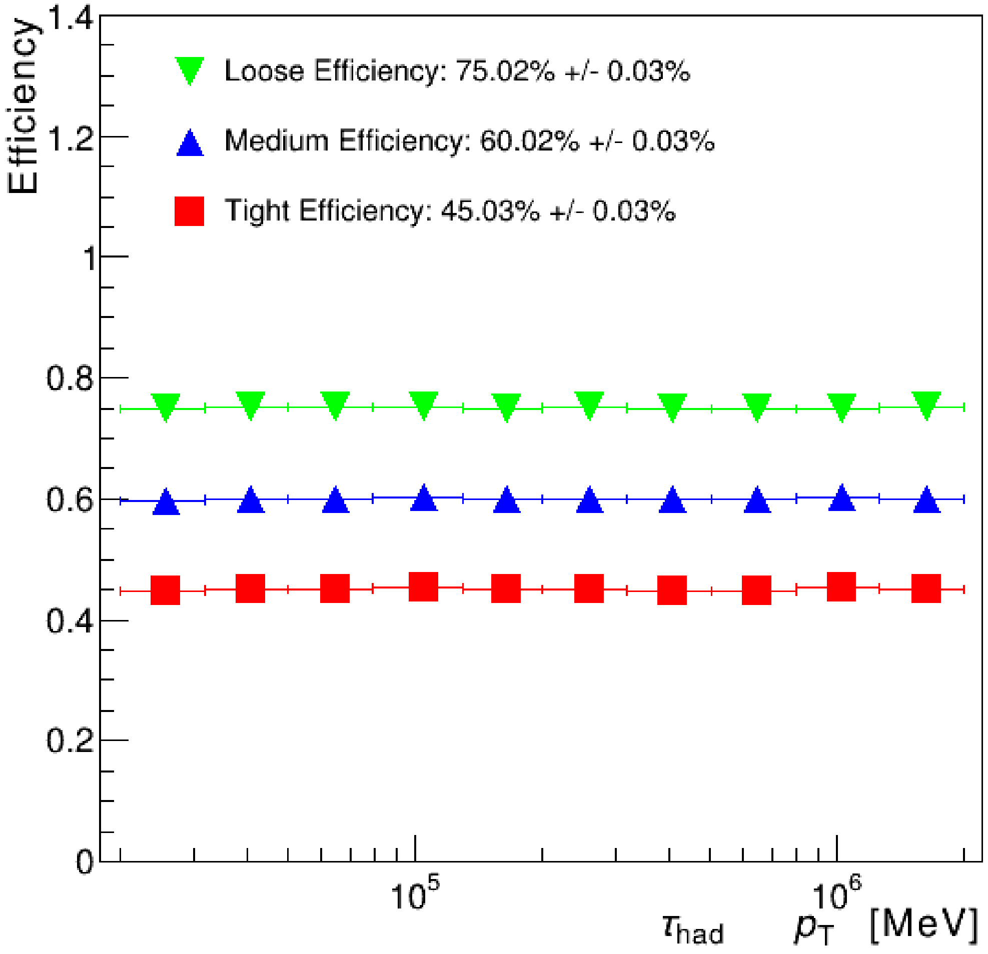
Backup: TauJetGraphs Signal Efficiency and Background Rejection

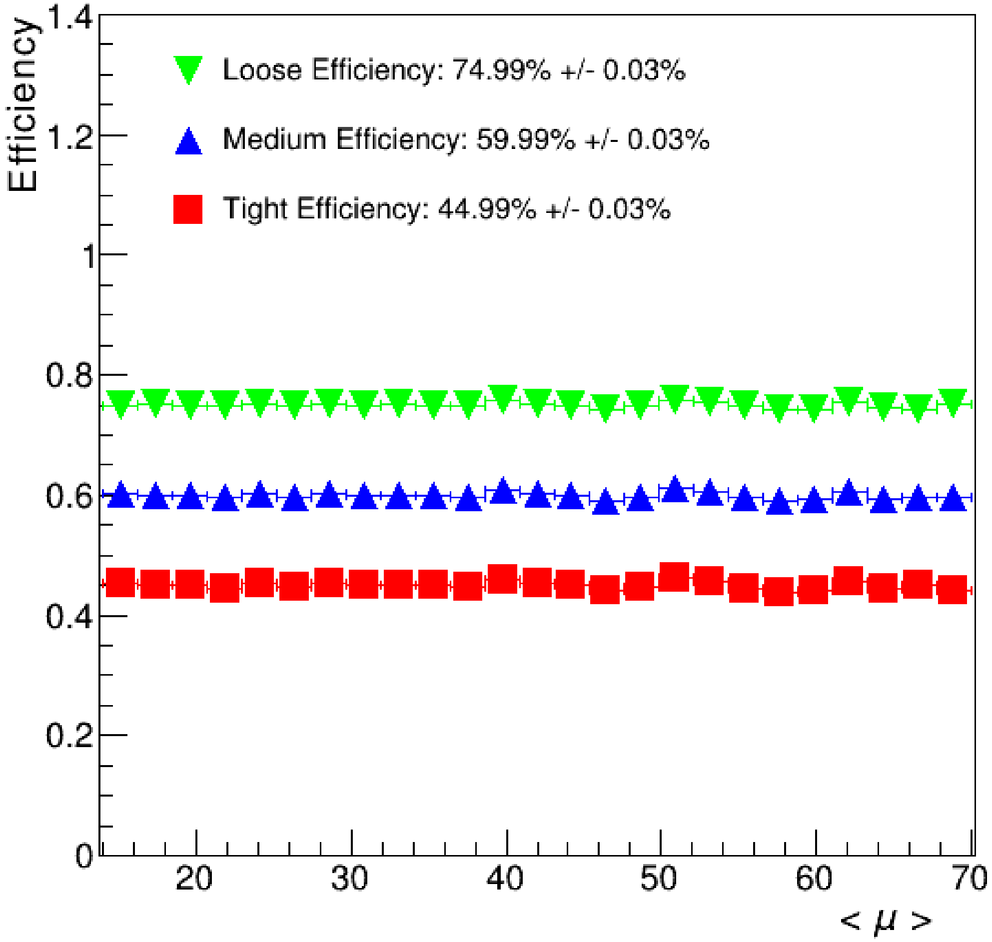
GNN Efficiency Plots (1-prong)



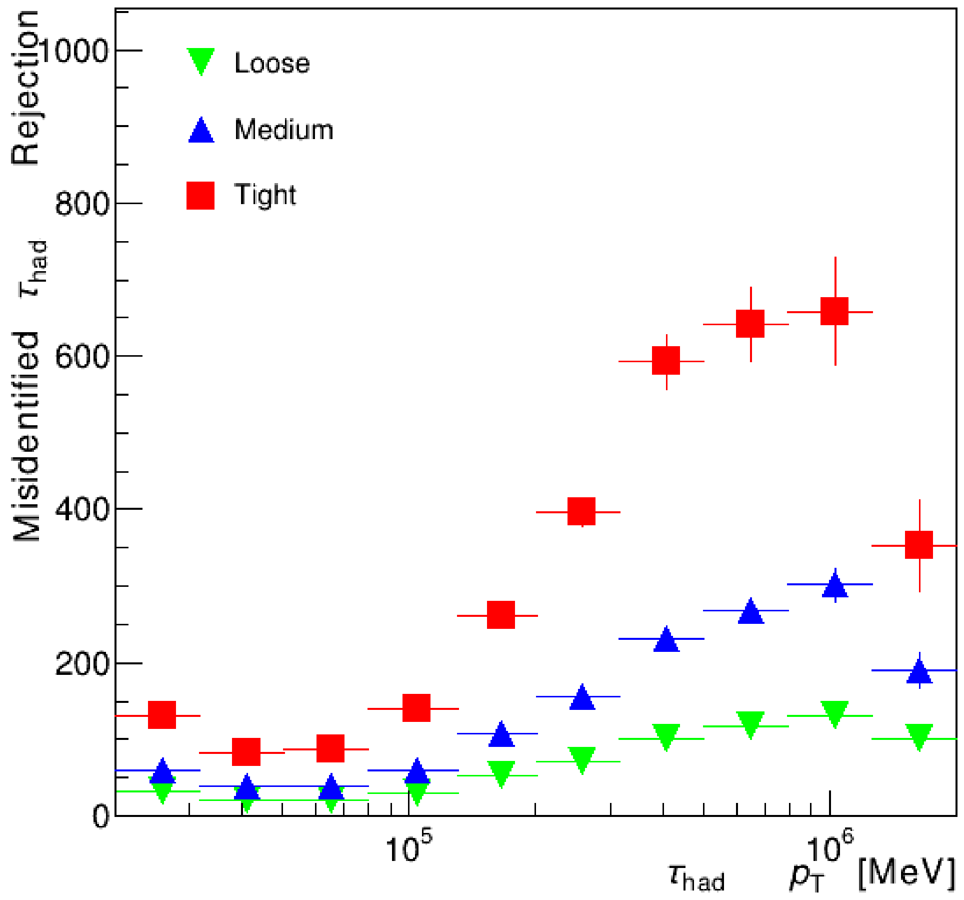


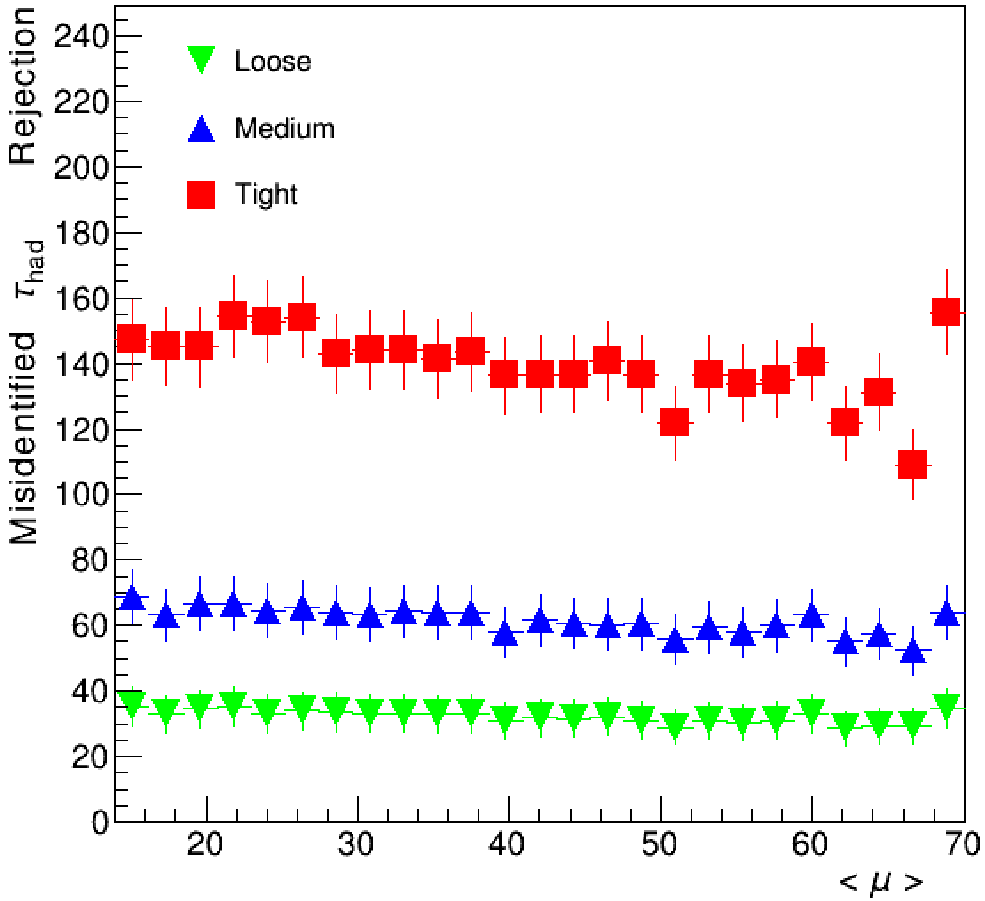
GNN Efficiency Plots (3-prong)



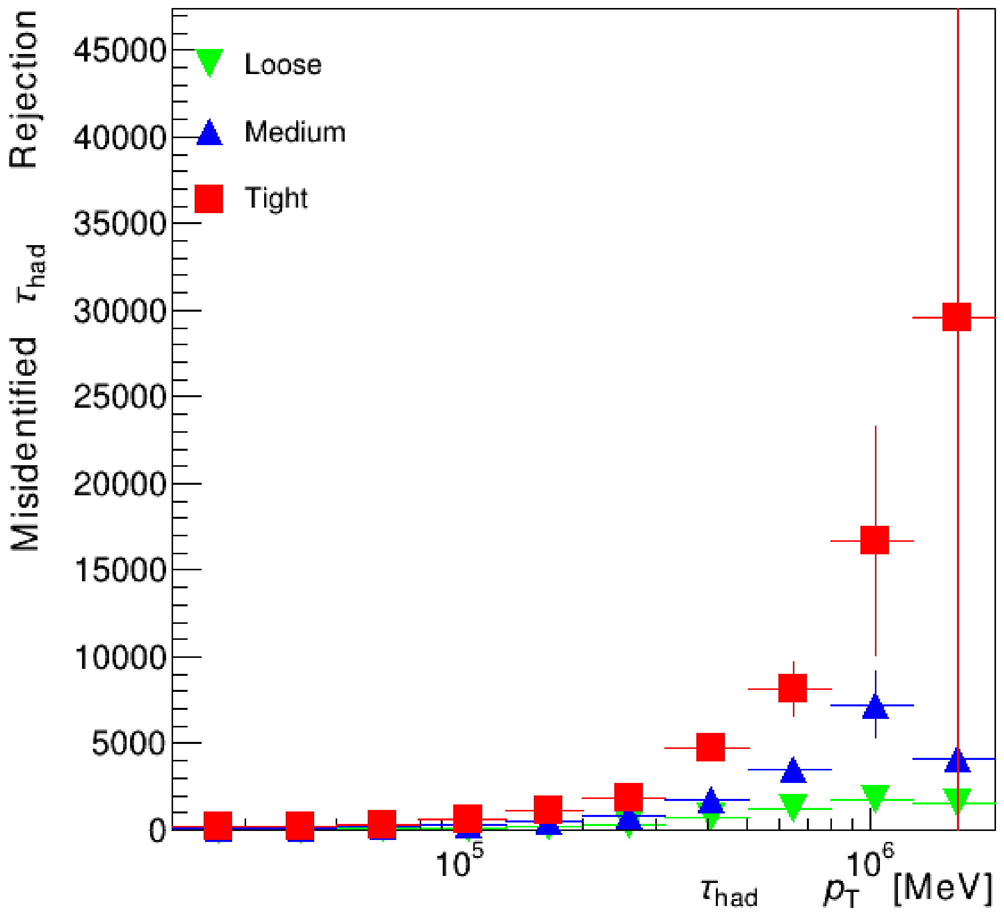


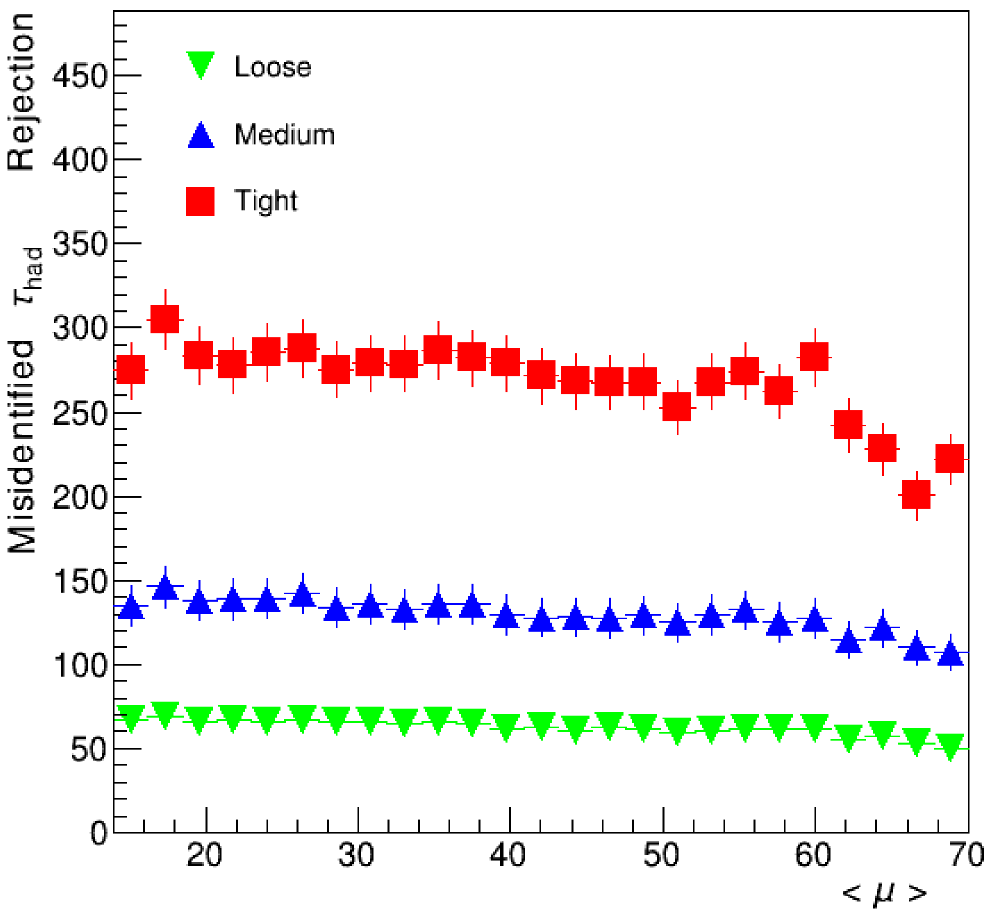
GNN Rejection (1-prong)





GNN Rejection (3-prong)

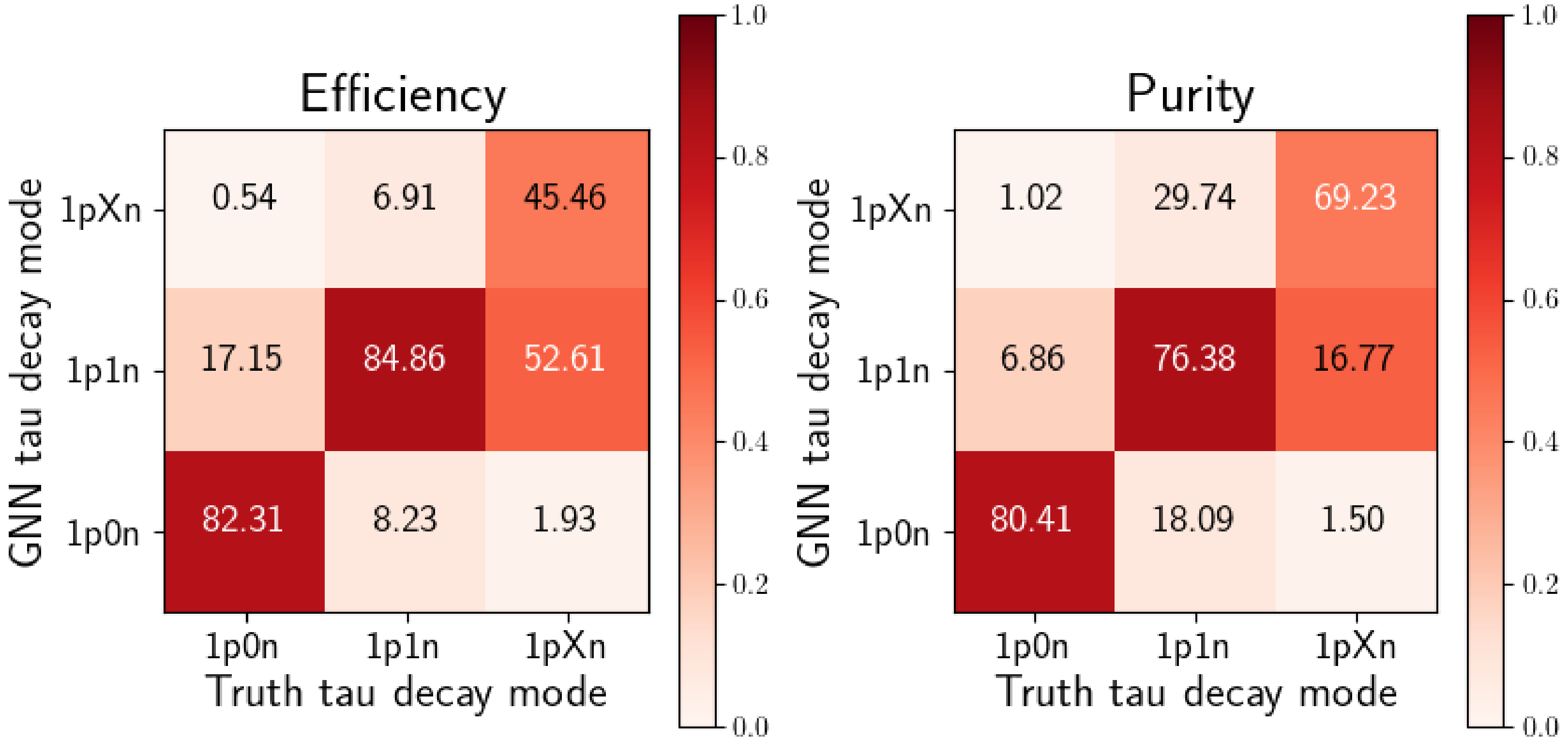




Backup: GNN Confusion Matrices

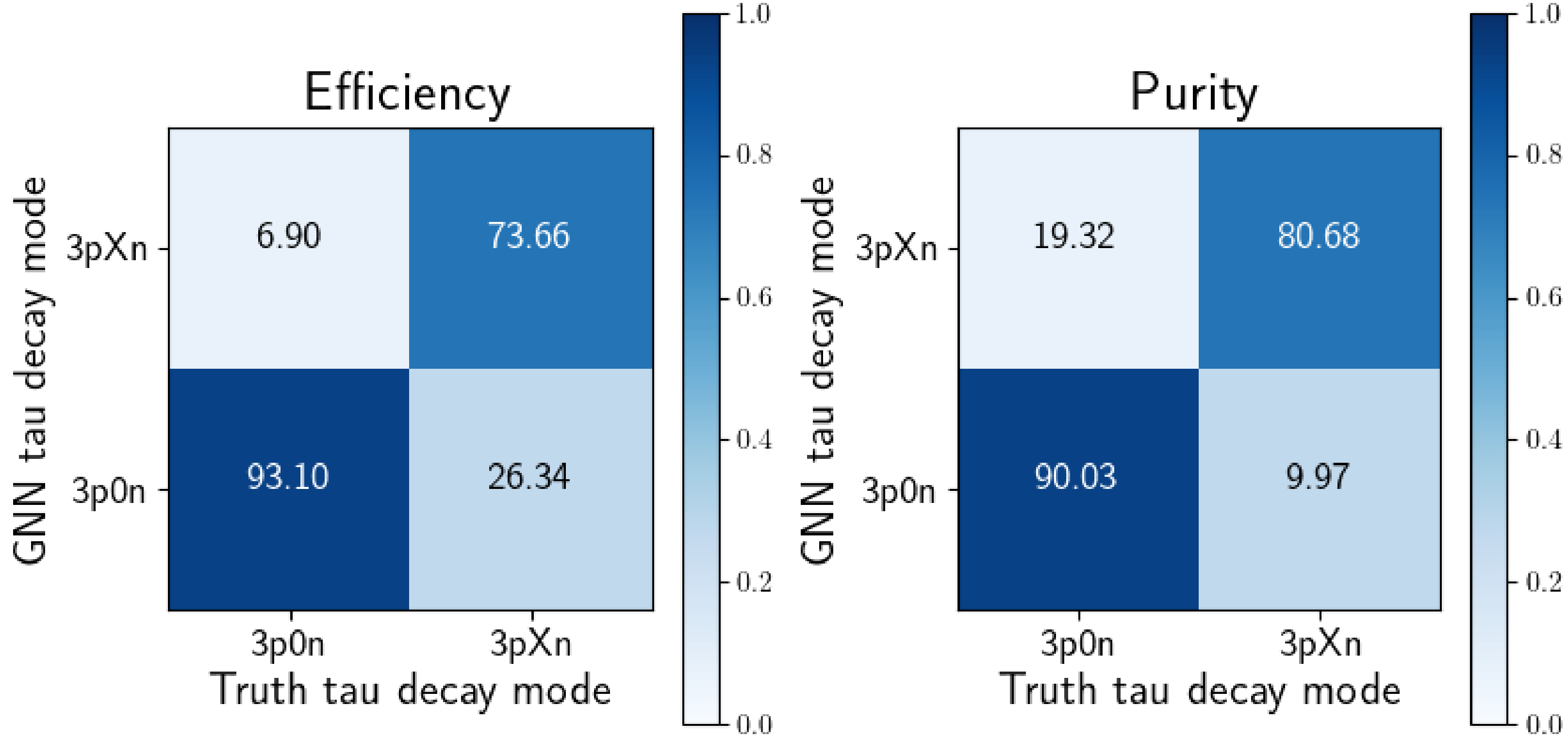
GNN Confusion Matrices (1-prong)

For 75% Signal Efficiency



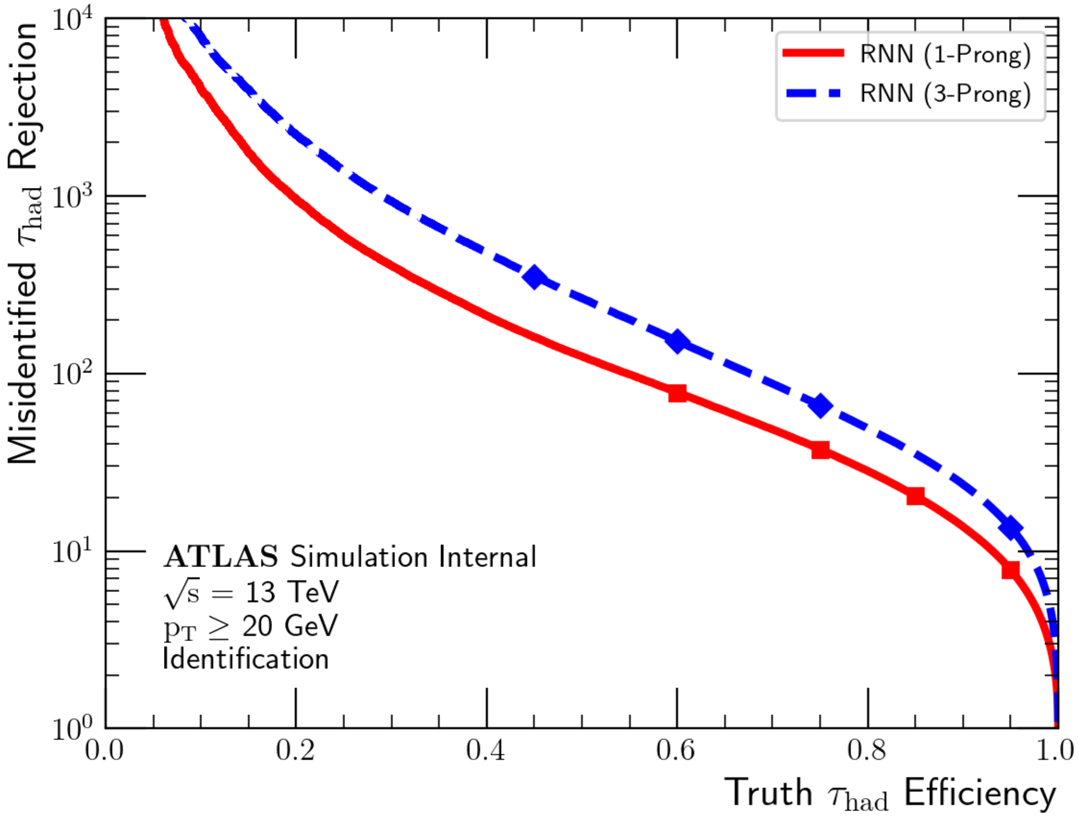
GNN Confusion Matrices (3-prong)

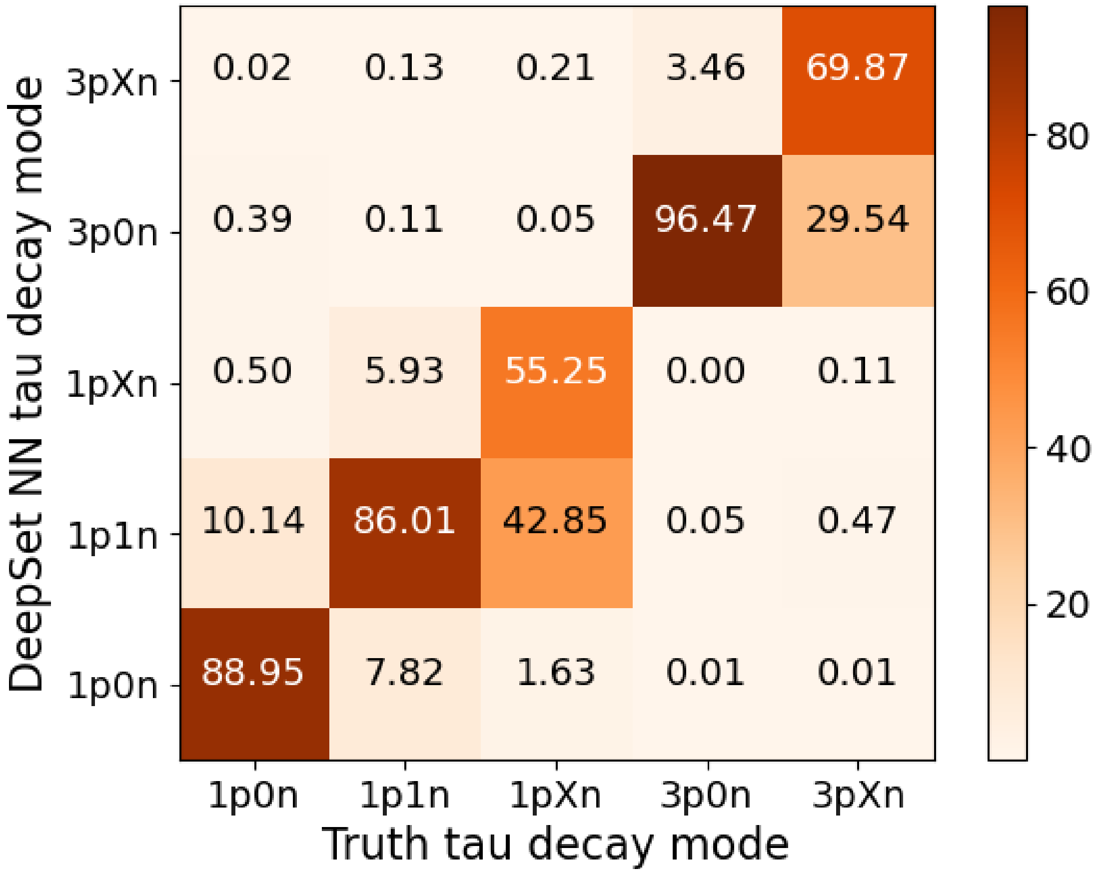
For 60% Signal Efficiency



Backup: Performance of Current Methods

Tau ID RNN and Decay Mode Classifier DSNN





Backup: Definitions and Glossary

Metric Definitions

- **Accuracy** – The fraction of correctly classified samples (if normalised = True)
- **Purity (Precision)** – Purity is the measure of how well a classifier avoids incorrectly labelling a sample as positive. It's calculated as true positives divided by true positives plus false positives:
  - $\frac{tp}{tp+fp}$ where *tp* is true positive and *fp* is false positive
- **Efficiency (Recall)** – Efficiency measures how well a classifier finds all the true positives. It's calculated as true positives divided by true positives plus false negatives:
  - $\frac{tp}{tp+fn}$ where *tp* is true positive and *fn* is false negative
- **Background Rejection –** The inverse of the Background Selection Efficiency, depending on the Signal Selection Efficiency

Glossary

- **ID –** Identification
- **DMC –** Decay Mode Classification
- $\boldsymbol{\tau}_{\mathbf{had}}$ **-** Hadronically decaying τ-lepton
- **RNN –** Recurrent Neural Network
- **DSNN -** DeepSet Neural Network
- **GNN –** Graph Neural Network
- **ROC Curve -** Receiver Operator Characteristic Curve