Channel tagging with scikit-learn: Updates since September 2019

Erin Conley

January 8, 2020

DUNE Low Energy Working Group Meeting

Outline

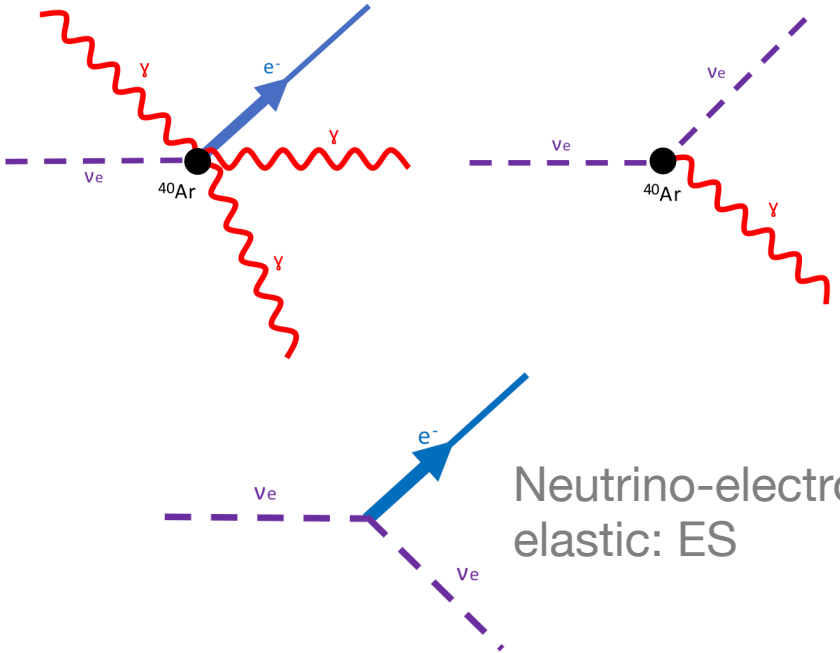
- Introduction
- Results from September 2019
- Current studies and preliminary results:
  - Tagging neutrino energy levels
  - Two different E_{reco} pre-selection cuts
  - Comparing the two pre-selection cuts
- Takeaways and next steps

Introduction

- Motivation: use scikit-learn to tag/distinguish three different supernova interactions
- ADA-boosted decision trees, SN spectrum-weighted ν_eCC, ES, and NC samples produce multi-classification algorithm:
  - ν_eCC: MCC11 MARLEY (clean)
  - ES: A. Roeth's supernova sample
  - NC: MARLEY sample using Pekka's calculations

ν_e-^{40}Ar charged current inelastic: ν_eCC

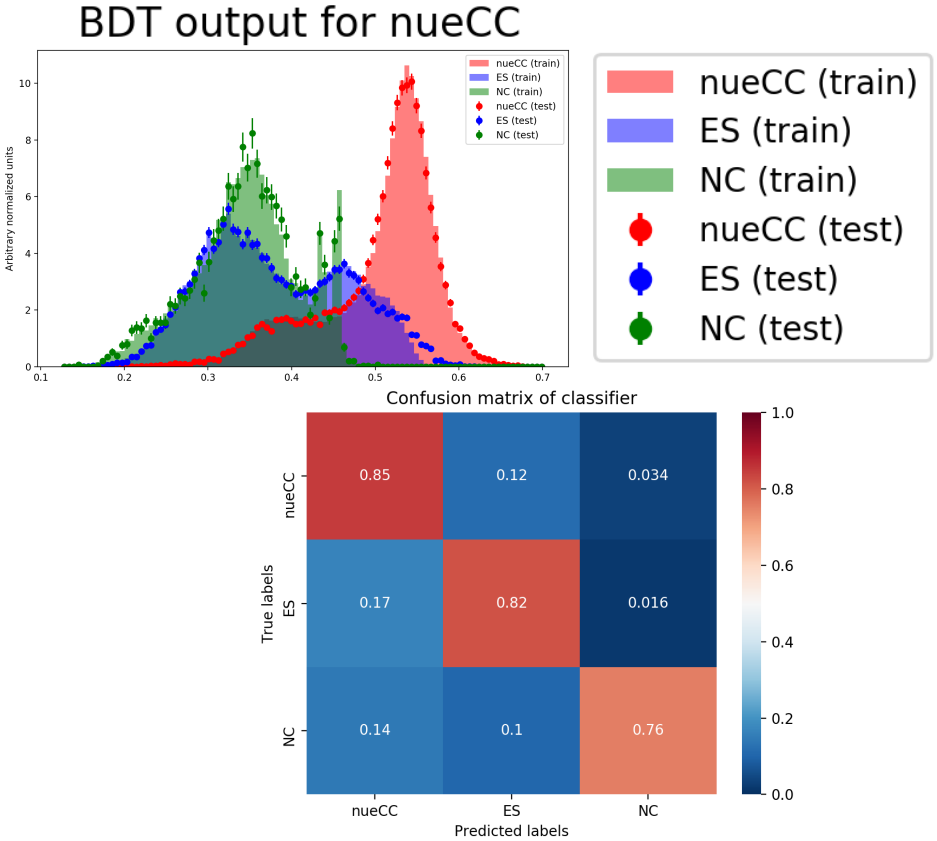
ν-^{40}Ar neutral current inelastic: NC



Neutrino-electron elastic: ES

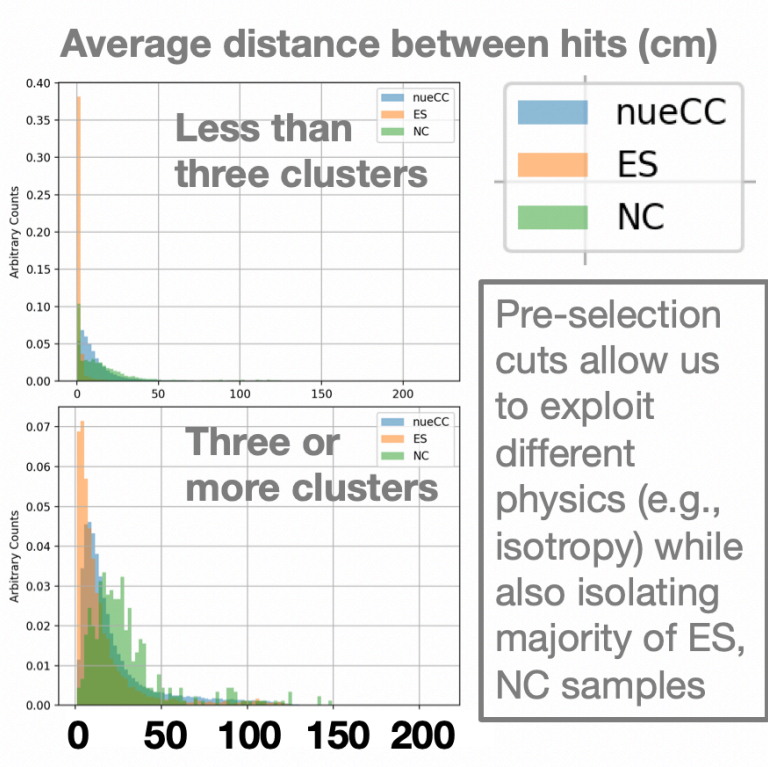
Results from September 2019

- Right-hand results for no pre-selection cuts on SN-weighted samples
  - Presented at [September 2019 collaboration meeting]
  - Total purity: 81%
- Drawbacks to this method: did not exploit physics at different energy levels; NC model is largely unknown



Current studies/algorithms

- Tagging neutrino energy levels (what is DUNE's tagging capability?)
- Current focus is on pre-selection cuts (previously explored cutting on reconstructed clusters)
- Exploring different pre-selection cuts for E_{reco}:
  - E_{reco} cut to isolate the NC signal (two-channel tagging above)
  - E_{reco} cut optimized for NC signal (three-channel tagging)



Shown in September 2019 slides

Tagging Neutrino Energy Levels

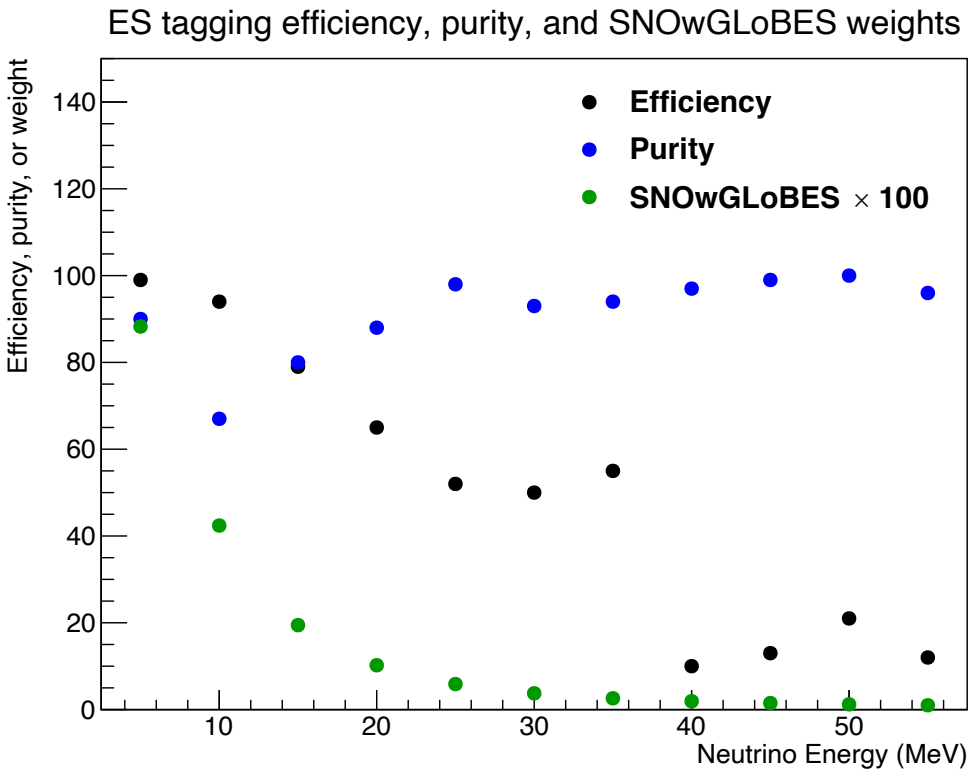
- Defined the energy levels as 5, 10, 15, …, 55 MeV with widths of 5 MeV, e.g., the 5 MeV level is defined as events with neutrino energies in the [2.5, 7.5) MeV range
  - This definition was made due to the MARLEY NC simulation
  - Required events to have at least one reconstructed track
- Produced three-channel classification algorithms for all levels except 5 and 25 MeV (ν_eCC vs ES)
  - Current reconstruction techniques cannot reconstruct 5 MeV NC events; also SNOwGLoBES predicts zero 5 MeV NC events
  - NC simulation did not include 25 MeV (Pekka's table)

BDT Tagging Results

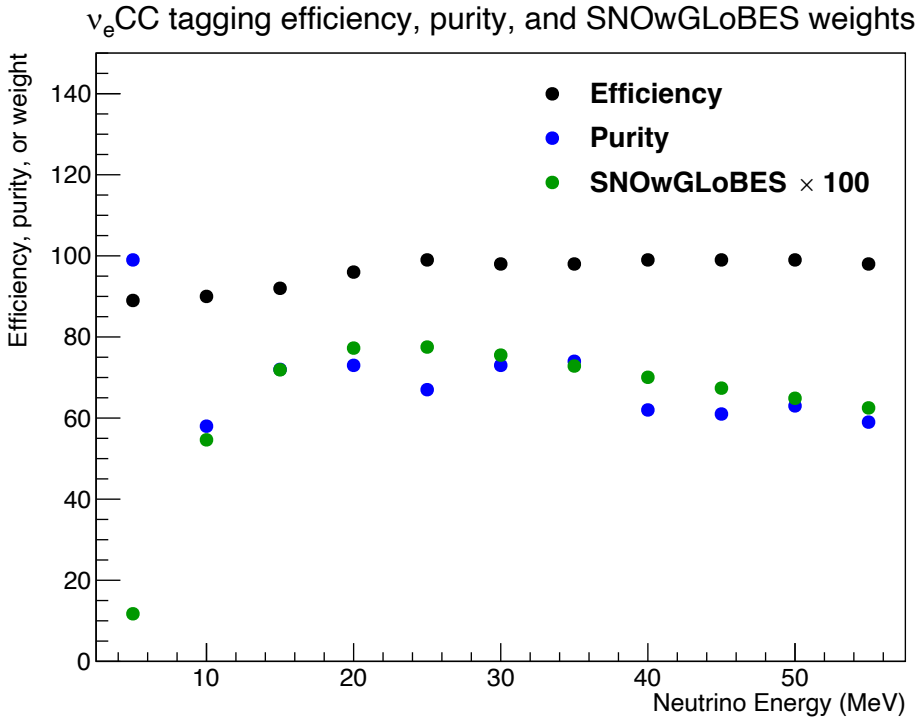
Neutrino Energy (MeV)	# Tagging Variables	ν_eCC purity (efficiency)	ES purity (efficiency)	NC purity (efficiency)	Total purity
5	6	99% (89%)	90% (99%)	N/A	94%
10	7	58% (90%)	67% (94%)	95% (7%)	80.6%
15	8	72% (92%)	80% (79%)	92% (68%)	79.8%
20	10	73% (96%)	88% (65%)	91% (89%)	82.8%
25	6	67% (99%)	98% (52%)	N/A	75.3%
30	10	73% (98%)	93% (50%)	87% (97%)	81.6%
35	10	74% (98%)	94% (55%)	87% (98%)	83%
40	8	62% (99%)	97% (10%)	76% (100%)	85.6%
45	8	61% (99%)	99% (13%)	78% (99%)	86%
50	6	63% (99%)	100% (21%)	78% (100%)	83.5%
55	5	59% (98%)	96% (12%)	81% (99%)	87.6%

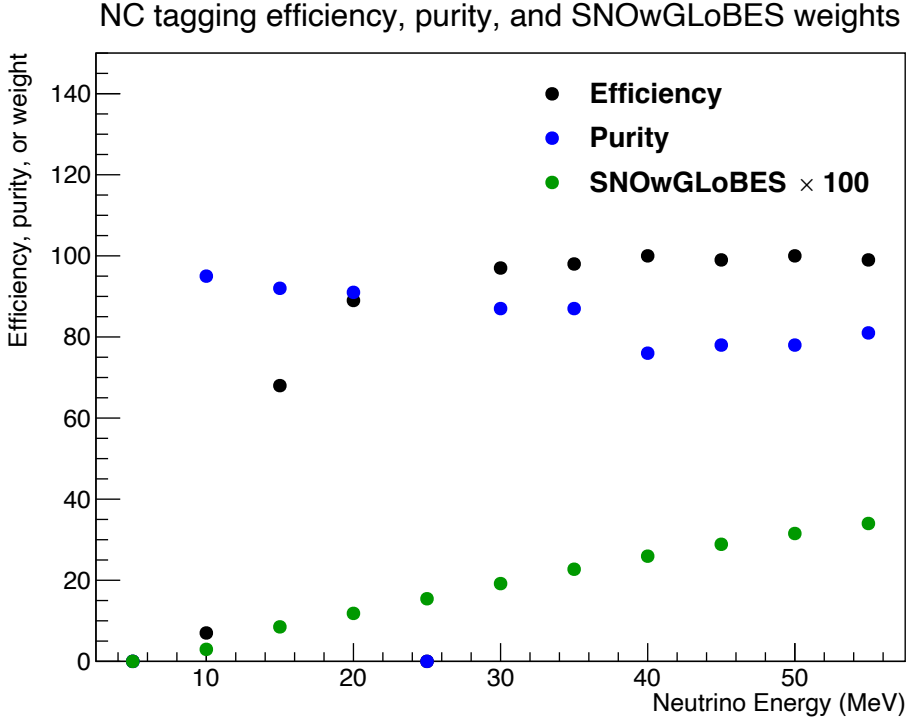
ES Tagging Capability

- Tracking ES efficiency, purity, and SNOwGLoBES weights versus E_ν
- As the weight decreases, so does the ES efficiency
  - Very few events above 35 MeV → weights are low → no tagging capability
- Why does purity increase? ν_eCC and NC events are not misclassified as ES!



ν_eCC and NC tagging capabilities





Some E_ν Tagging Takeaways

- We don't have tagging capability for ES events above 40 MeV (not a lot of events)
- We don't have tagging capability for 10 MeV NC events, but we could theoretically tag NC events above 30 MeV
- We have pretty good tagging capabilities for the ν_eCC channel over the entire energy range!
- I wish we had a simulation for 25 MeV NC events…
- CAVEAT: I doubt this study tells us anything about DUNE's actual tagging capability since we have no way to accurately reconstruct E_ν for NC events…

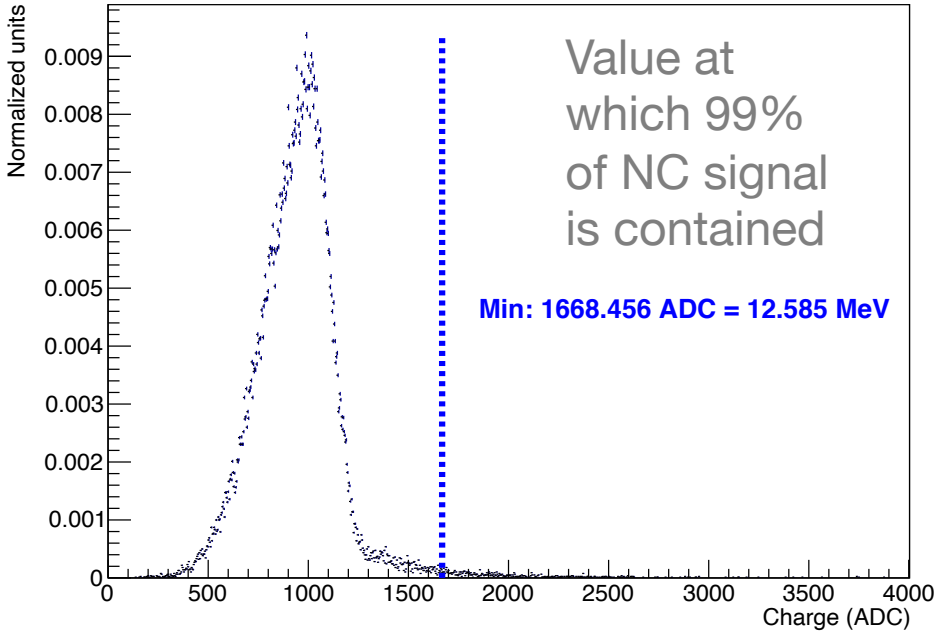
Revisiting the E_{reco} pre-selection cut

- Initial studies with E_{reco} pre-selection cuts: 2200 ADC
  - 2200 ADC ≈ 16.7 MeV
  - Chosen (by eye) to isolate the NC signal
  - Worked okay for NC/ν_eCC, but not for ES
- Re-examine the E_{reco} cut:
  - Isolate the NC signal
  - Optimize using $S/\sqrt{S+B}$

Two methods for E_{reco} cut

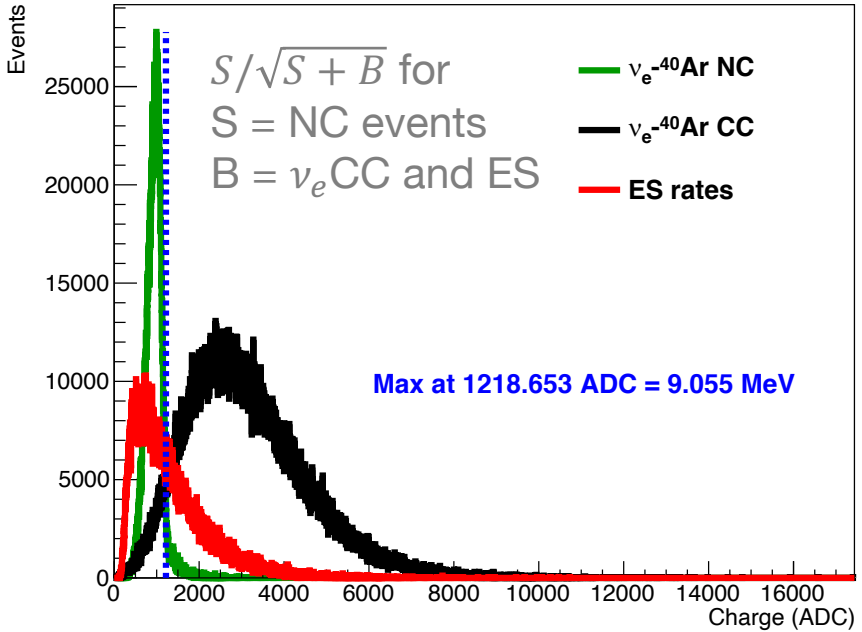
"Isolating the NC signal"

Normalized NC Charge Distribution



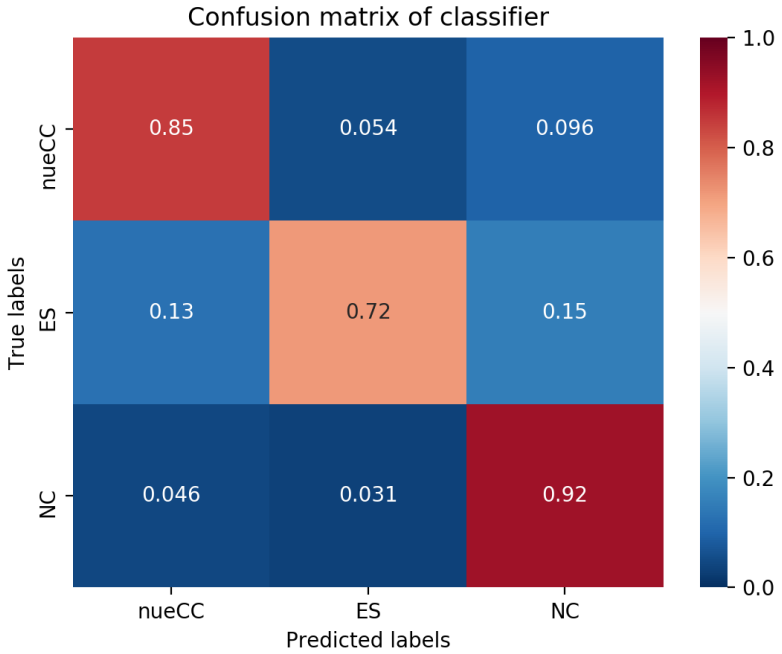
"Optimized cut"

Drift-corrected charge distributions

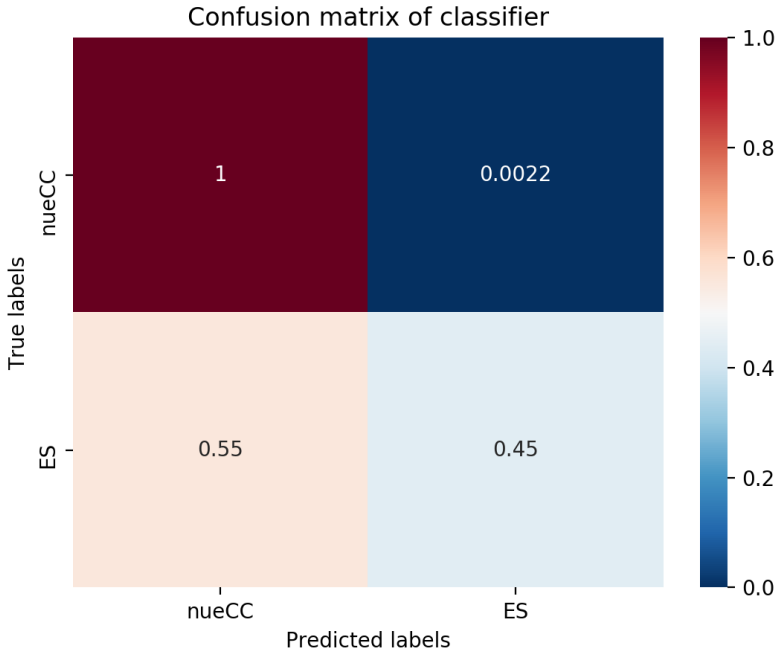


Results for isolating the NC signal

Below 12.585 MeV



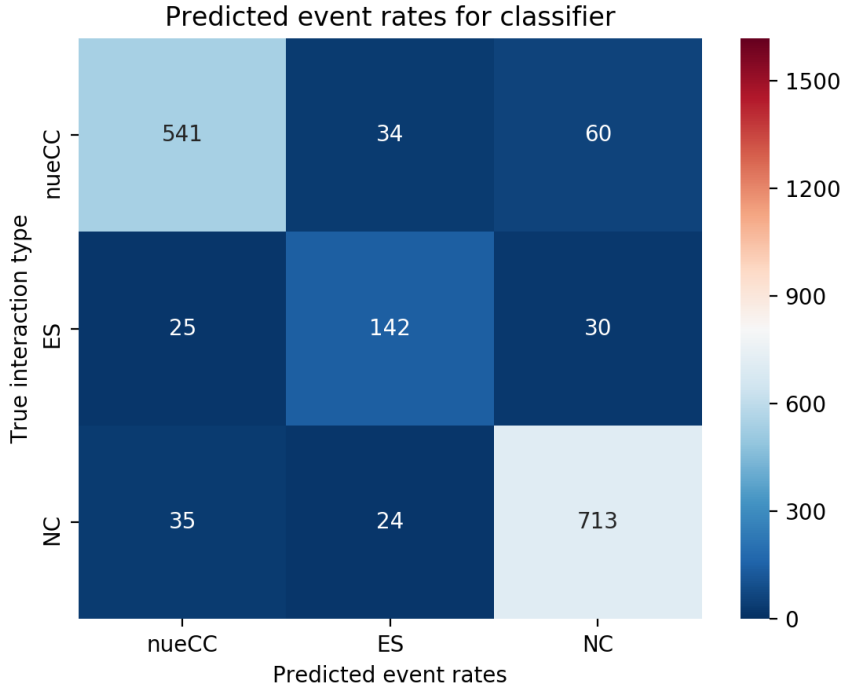
Above 12.585 MeV



Because we isolated the NC signal, above the cut is a two-channel problem

Determining predicted event rates

- Using confusion matrix and predicted SNOwGLoBES event rates, calculate predictions for tagged and misclassified events
  - Number of events we might lose to misclassification
- Due to rounding/fractional predicted events, sklearn predictions don't always sum to SNOwGLoBES predictions



Predicted event rates below 12.585 MeV (isolating the NC signal)

Predicted event rates: isolating NC signal

	Predicted ν_eCC	Predicted ES	Predicted NC
True ν_eCC	4486	42	60
True ES	88	193	30
True NC	35	24	713

Produced these rates by adding up the rates from the 2 algorithms (above and below energy cut)

The predicted ν_eCC signal would be composed of 4486 actual ν_eCC events with contamination from 88 ES and 35 NC

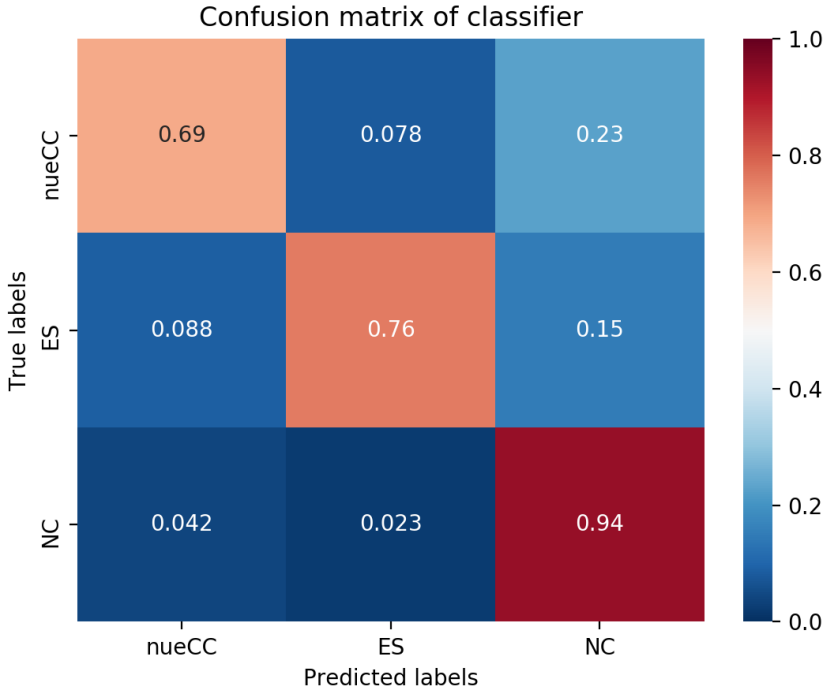
4591 total ν_eCC
314 total ES
783 total NC

Takeaways: Isolating the NC signal

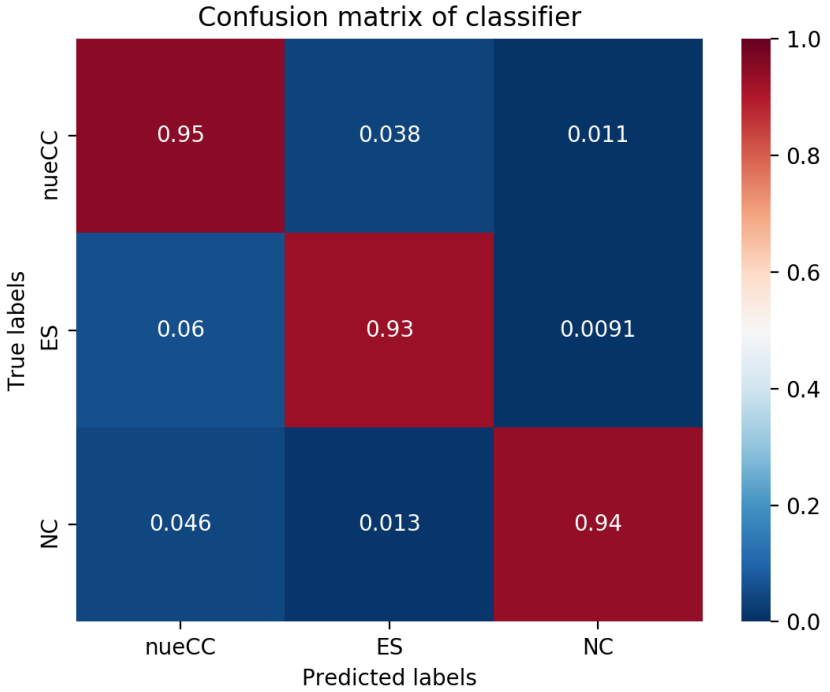
- Predicted ν_eCC signal has relatively small contamination from ES and NC signals (~2%), and only ~2% of the total ν_eCC events are misclassified!
- ~38% of ES signal is misclassified as ν_eCC/NC (pointing measurement will suffer…)
- By making this type of cut, we risk throwing away high-energy NC events with no attempt to look for them…

Results for optimized cut

Below 9.055 MeV



Above 9.055 MeV



Predicted event rates: optimized cut

	Predicted ν_eCC	Predicted ES	Predicted NC
True ν_eCC	4297	183	108
True ES	22	268	21
True NC	32	17	731

Produced these rates by adding up the rates from the 2 algorithms (above and below energy cut)

The predicted ν_eCC signal would be composed of 4297 actual ν_eCC events with contamination from 22 ES and 32 NC

4591 total ν_eCC
314 total ES
783 total NC

Takeaways: Optimized cut

- Predicted ν_eCC signal has less contamination from ES, NC (~1%), but ~6% of the ν_eCC signal is misclassified
- ~13.8% of the ES signal is misclassified, but now the ES signal is swamped by ν_eCC contamination (pointing measurement would suffer…)
  - The ν_eCC contamination is also worse for the NC signal

Comparing the two E_{reco} cuts

12.585 MeV (isolate NC signal)

Predicted signal	Purity (below)	Efficiency (below)	Purity (above)	Efficiency (above)
ν_eCC	83%	85%	64%	100%
ES	89%	72%	100%	45%
NC	79%	92%	N/A	N/A

Total purity below: 83%
Total purity above: 77.5%

9.055 MeV (optimized E_{reco} cut)

Predicted signal	Purity (below)	Efficiency (below)	Purity (above)	Efficiency (above)
ν_eCC	84%	69%	90%	95%
ES	88%	76%	93%	95%
NC	94%	71%	94%	98%

Total purity below: 79.5%
Total purity above: 94.1%

Comparing the two E_{reco} cuts

4591 total ν_eCC
314 total ES
783 total NC

12.585 MeV (isolate NC signal)

Predicted signal	# correctly tagged events	Contam. from…	Contam. (percent)	Percent of misclassified events
ν_eCC	4486	88 ES 35 NC	2.67	2.2
ES	193	42 ν_eCC 24 NC	25.5	37.9
NC	713	60 ν_eCC 30 ES	11.2	7.6

Low ν_eCC contamination
Might throw away the high-energy NC signal
High ES misclassification

9.055 MeV (optimized E_{reco} cut)

Predicted signal	# correctly tagged events	Contam. from…	Contam. (percent)	Percent of misclassified events
ν_eCC	4297	22 ES 32 NC	1.24	6.3
ES	268	183 ν_eCC 17 NC	42.7	13.8
NC	731	108 ν_eCC 21 ES	15	6.3

Low ν_eCC contamination, ES misclassification
High ES/NC contamination
High ν_eCC misclassification

Takeaways

- Tagging neutrino energy levels was interesting and highlighted current DUNE reconstruction and tagging capabilities, but obviously comes with some caveats…
  - Probably won't be able to precisely reconstruct E_ν for NC events
  - We can't make predictions about DUNE event rates, contamination, misclassification
- The two E_{reco} cuts have different benefits and drawbacks
- Overall, ES signal suffers the most while ν_eCC/NC signals are pretty good!

Future work

- However, NC model is largely unknown and carries large uncertainties – we want to study this effect!
- A list of future work:
  - Determine which E_{reco} cut we would like to go with (maybe include both for now?)
  - Study how NC model uncertainty affects the tagging results
  - Write up results in technical note
  - Improving reconstruction algorithms or tagging parameters
  - Try different reconstruction algorithms (e.g., SpacePointSolver)
  - Adding radiological backgrounds to the simulations

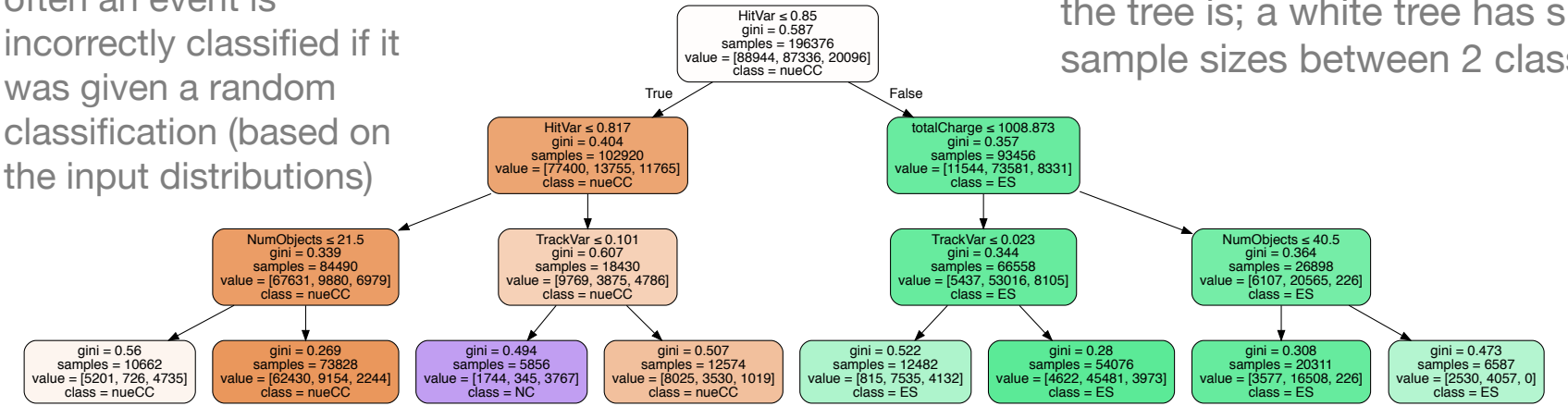
Backup slides

Decision Trees in sklearn

- Non-parametric learning method; “learns” by creating simple decision rules based on features from input data (more information here)
  - ADA-boosted trees apply weights to training samples over many iterations
- Understand function output, diagnose issues by looking at trees

Tree “gini” quantifies how often an event is incorrectly classified if it was given a random classification (based on the input distributions)

Color-scale tells you about how pure the tree is; a white tree has similar sample sizes between 2 classes



ν_eCC
ES
NC

Classifier Output: Decision Function

- Decision function outputs "confidence scores" related to classifier probability that an event is ν_eCC, ES, or NC
- Three values for each class (ν_eCC, ES, NC)
  - Largest score corresponds to classifier final prediction
  - Zero score: event is definitely not that class (according to decision tree)

ν_eCC	ES	NC
[[0.54530473	0.42676481	0.02793046]
[0.34143502	0.47044198	0.188123]
[0.57281998	0.42718002	0.]
...		
[0.35454589	0.25241753	0.39303658]
[0.55565677	0.44434323	0.]
[0.331612	0.32329672	0.34509128]]

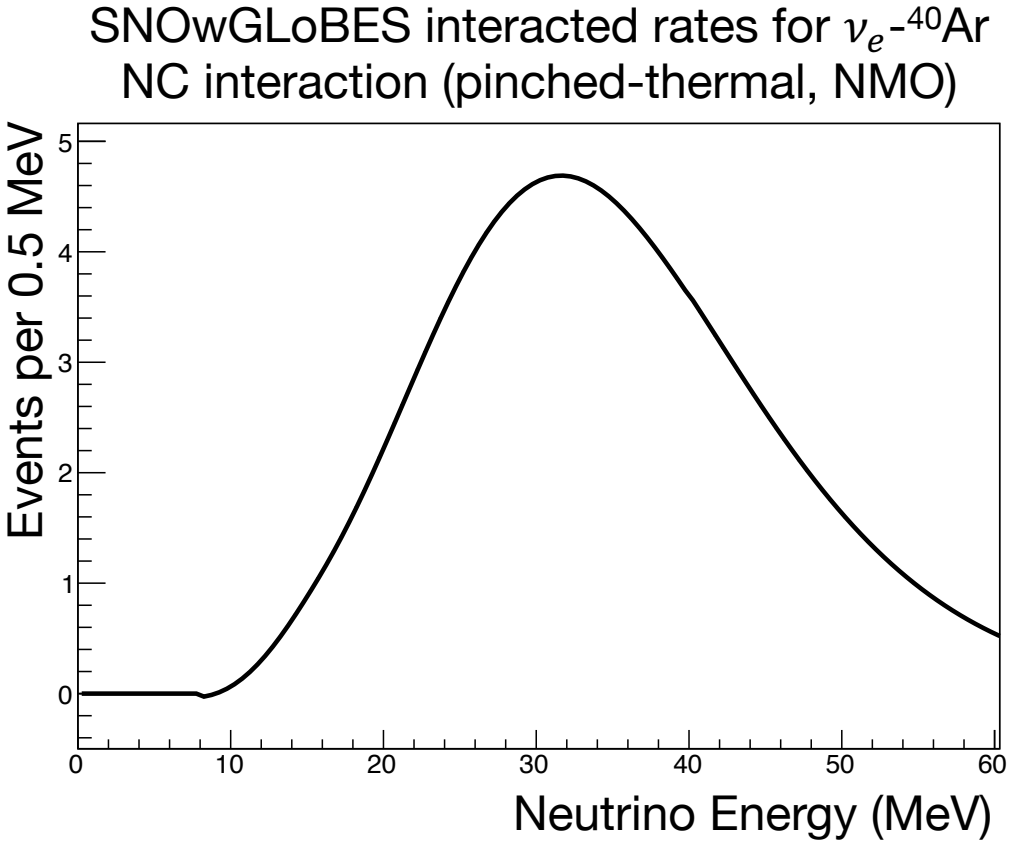
Example output of a decision function over the training sample
Each line represents the decision function output for one event (which could be ν_eCC, ES, or NC)

Tagging Neutrino Energy Levels: Additional Information

- Calculated class weights from SNOwGLOBES using pinched-thermal NMO flux and interacted rates (since we're looking at neutrino energy)
- Requirements:
  - All events must contain at least one reconstructed track
  - All tagging variables must have at least 0.01 importance

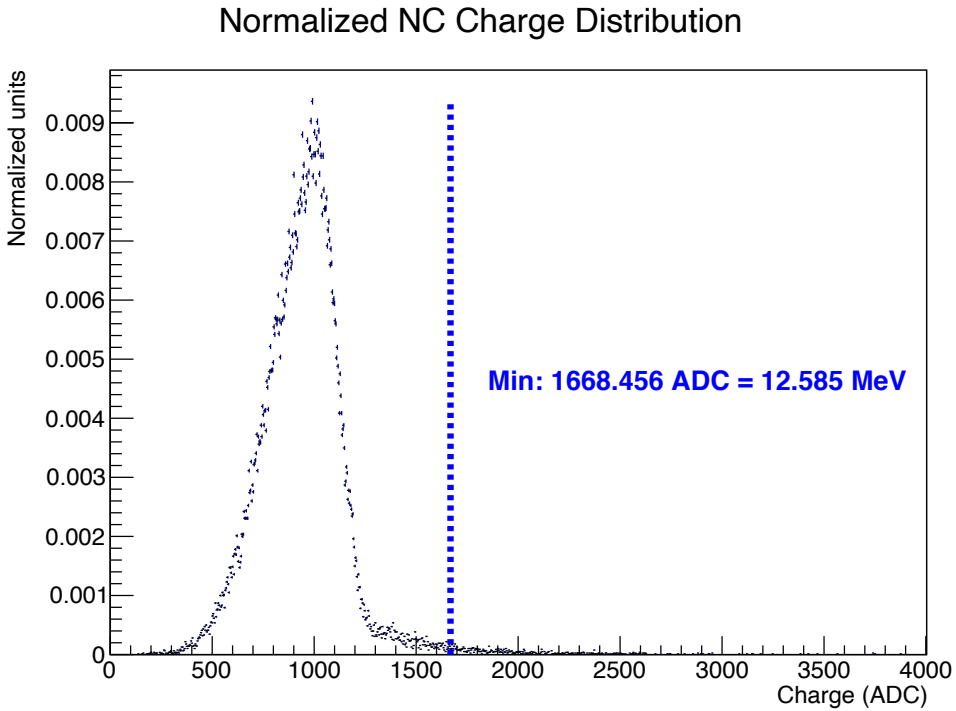
5 MeV NC events

- Interacted rates for NC events “turn on” around 10 MeV – thus we don’t have 5 MeV NC events
- Current LArSoft reconstruction algorithms don’t have the capability to reconstruct 5 MeV NC events



E_{reco} cut: Isolating the NC rates

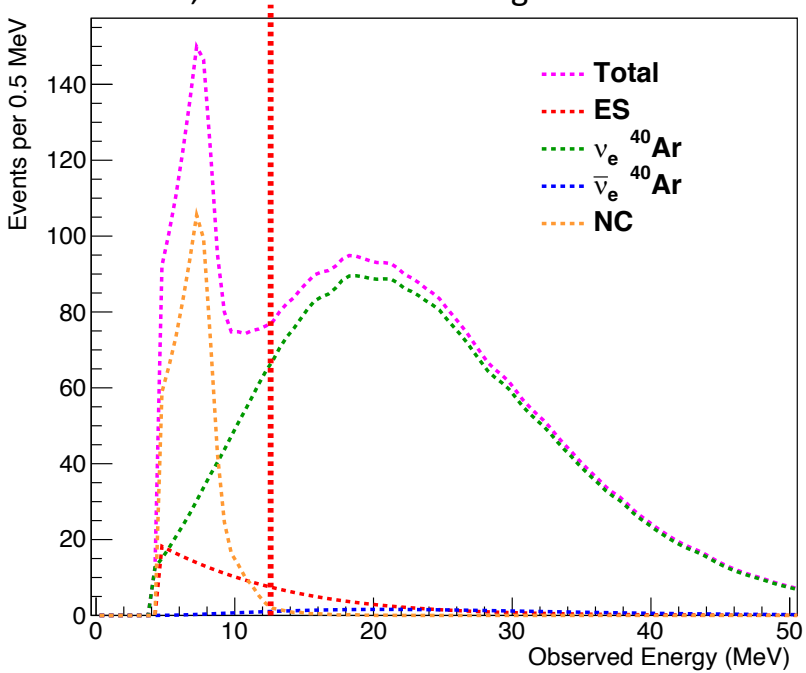
- Normalized NC charge distribution to 1.0
- Found charge value at which 99% of the NC events are contained
  - Converted to MeV using calibration constants
- Lower charge cut than 2200 ADC! (Increases ES sample in the 2-channel problem)



Updated SNOwGLoBES weights: isolated NC rates

- Weights below the cut:
  - ν_eCC: 637.324/1618.7 = 0.394
  - ES: 199.012/1618.7 = 0.123
  - NC: 774.781/1618.7 = 0.479
- Weights above the cut:
  - ν_eCC: 3954.04/4156.64 = 0.951
  - ES: 115.323/4156.64 = 0.0277
  - NC: 8.79454/4156.64 = 0.00213
- Right: smeared rates used to produce the weights; red line corresponds to cut at $E_{\mathrm{reco}} = 12.585$ MeV

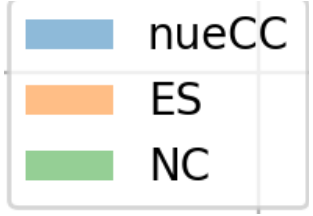
Smeared event rates for pinched-thermal flux, normal mass ordering

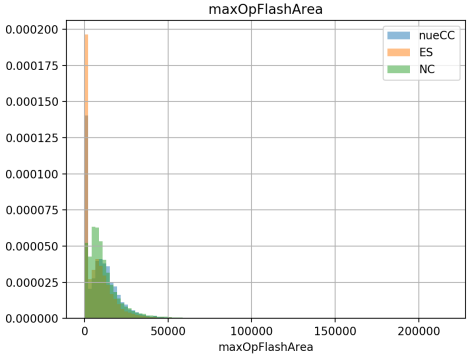


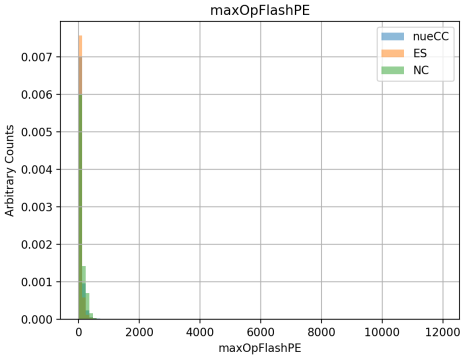
Results for isolating the NC signal

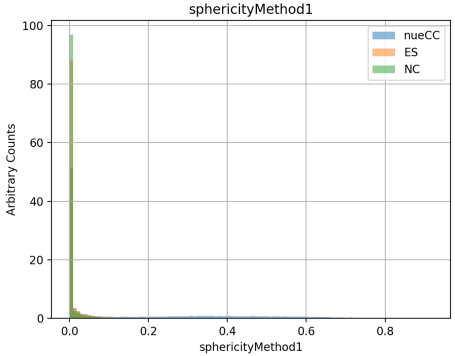
12.585 MEV CUT

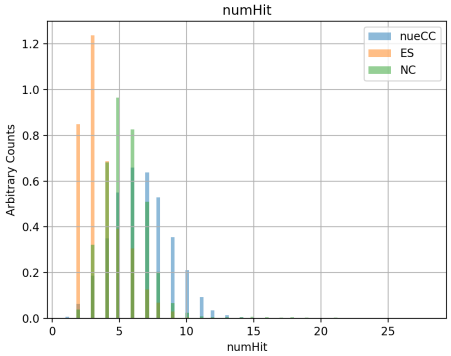
Variables: below 12.585 MeV

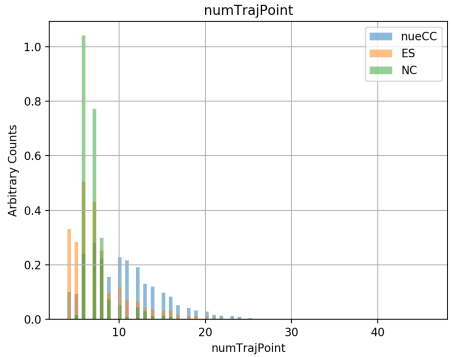


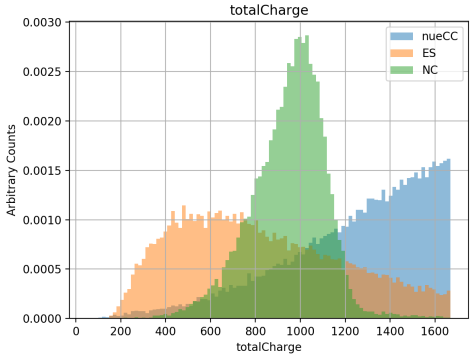




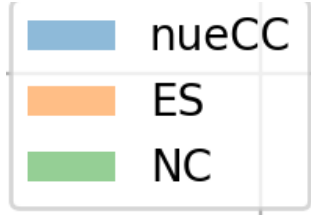


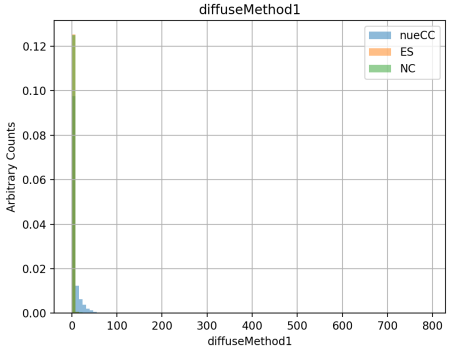


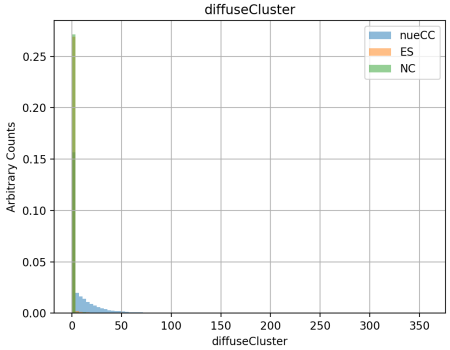


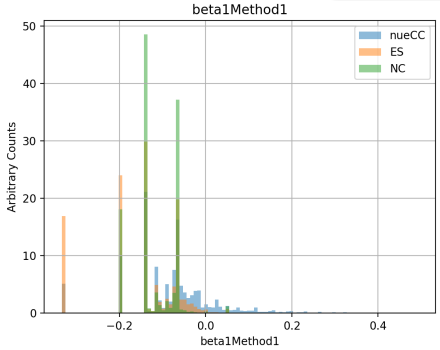


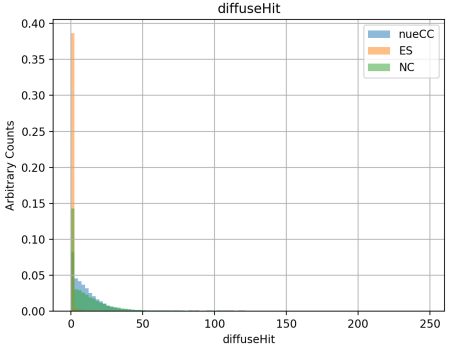
Variables: below 12.585 MeV

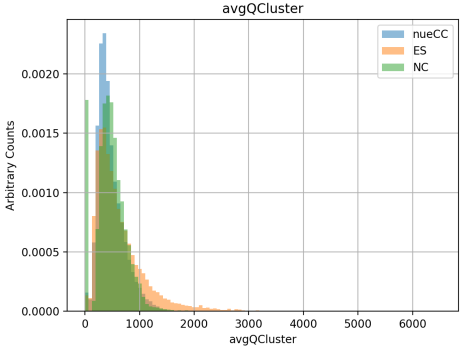


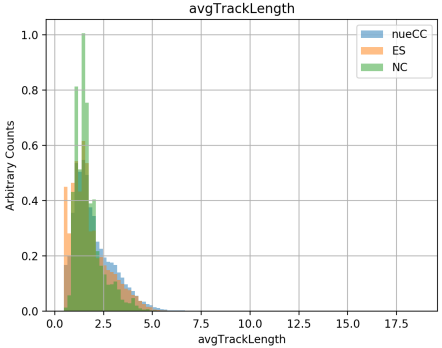




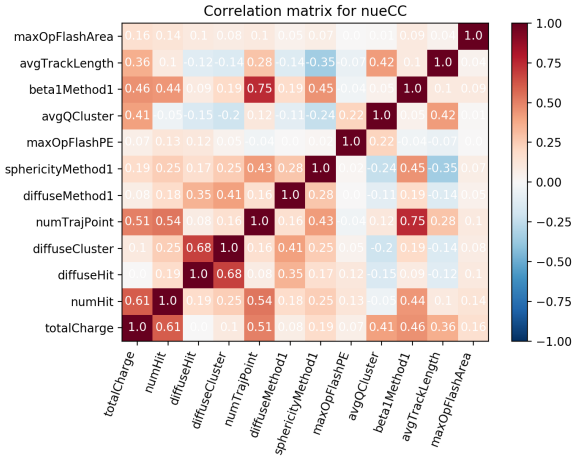


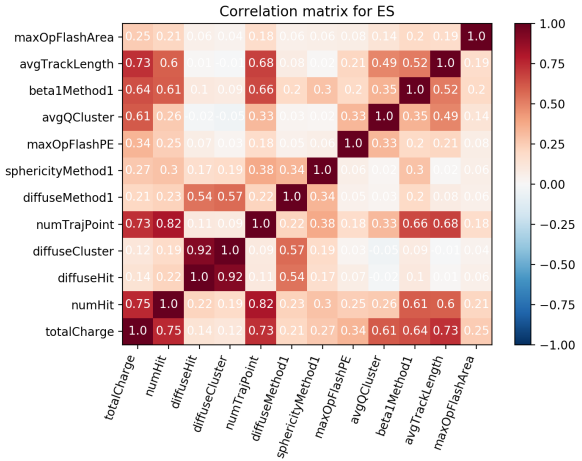


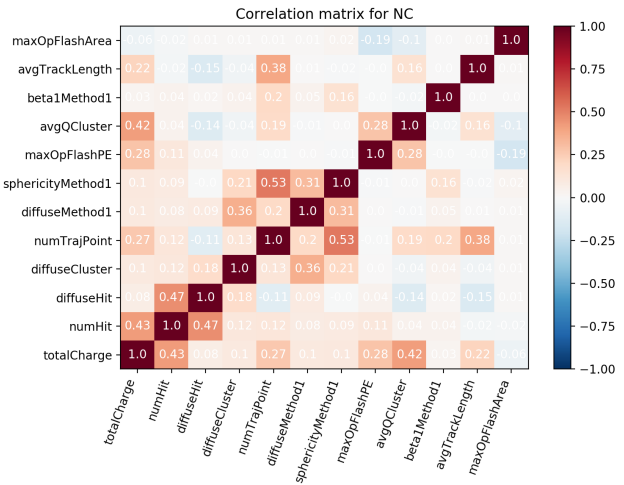




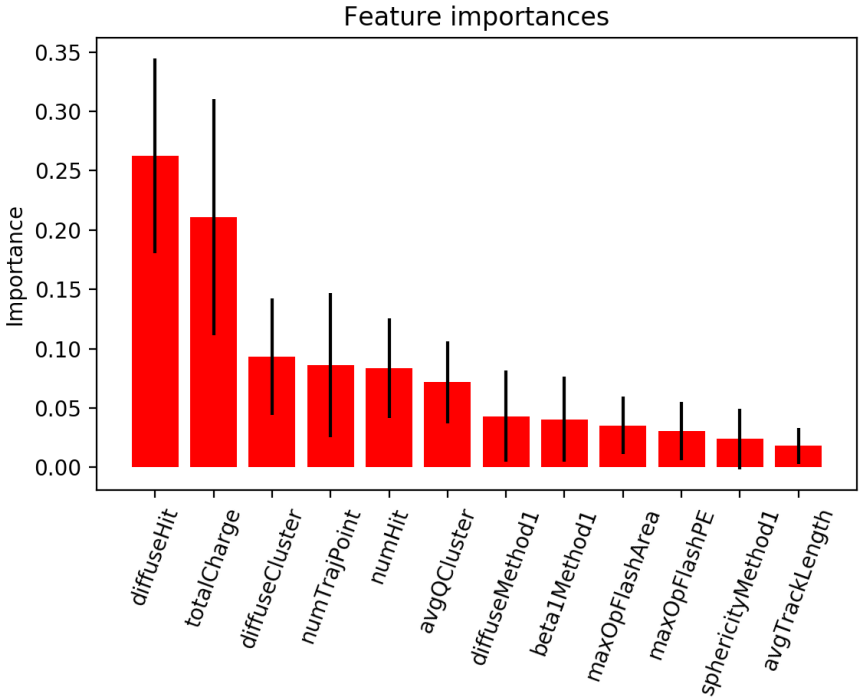
Correlation matrices: below 12.585 MeV





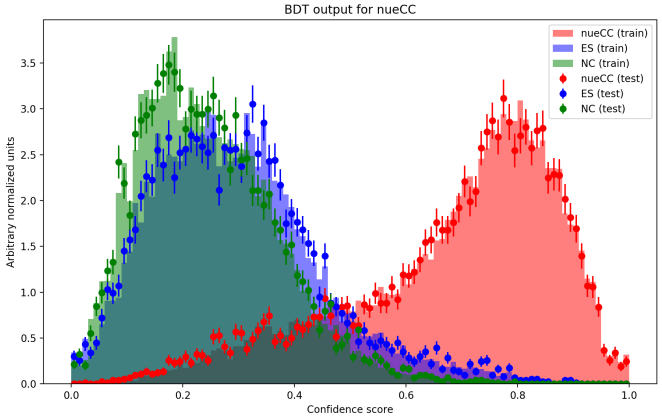


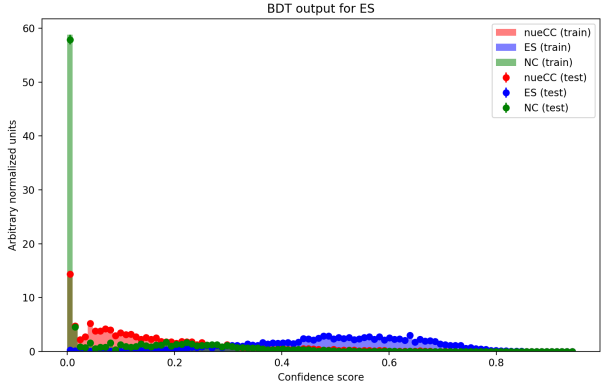
BDT statistics/results: below 12.585 MeV



- Overall purity: 83%
- ν_eCC efficiency (purity): 85% (83%)
- ES efficiency (purity): 72% (89%)
- NC efficiency (purity): 92% (79%)

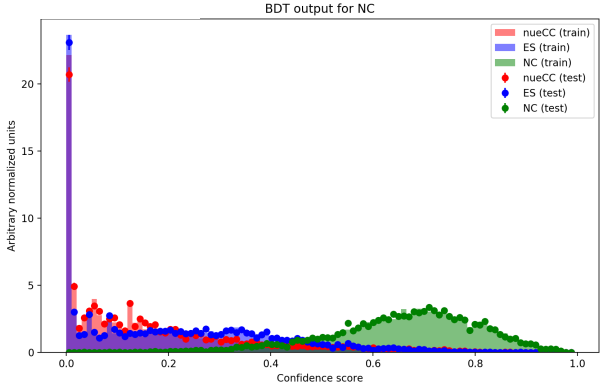
BDT outputs: below 12.585 MeV



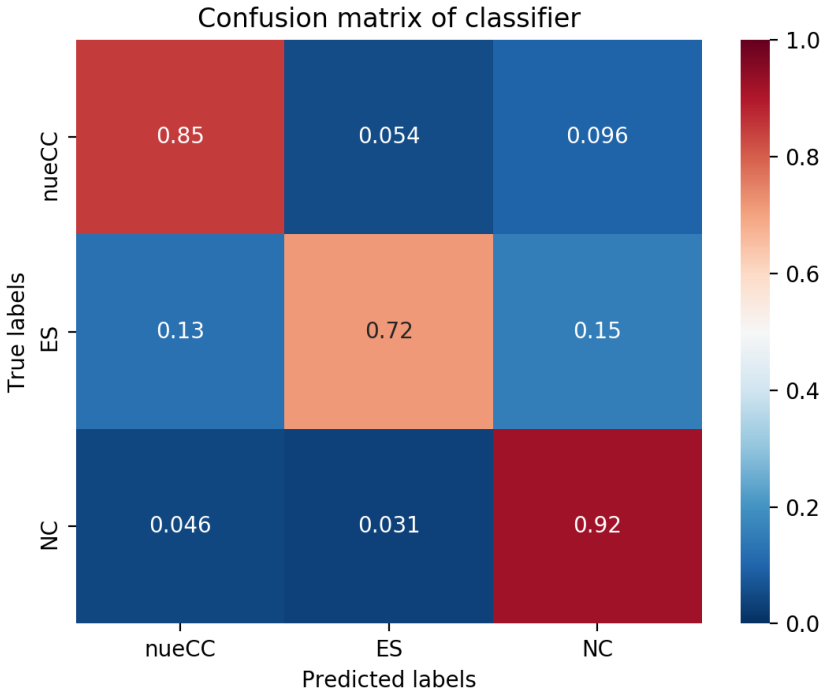


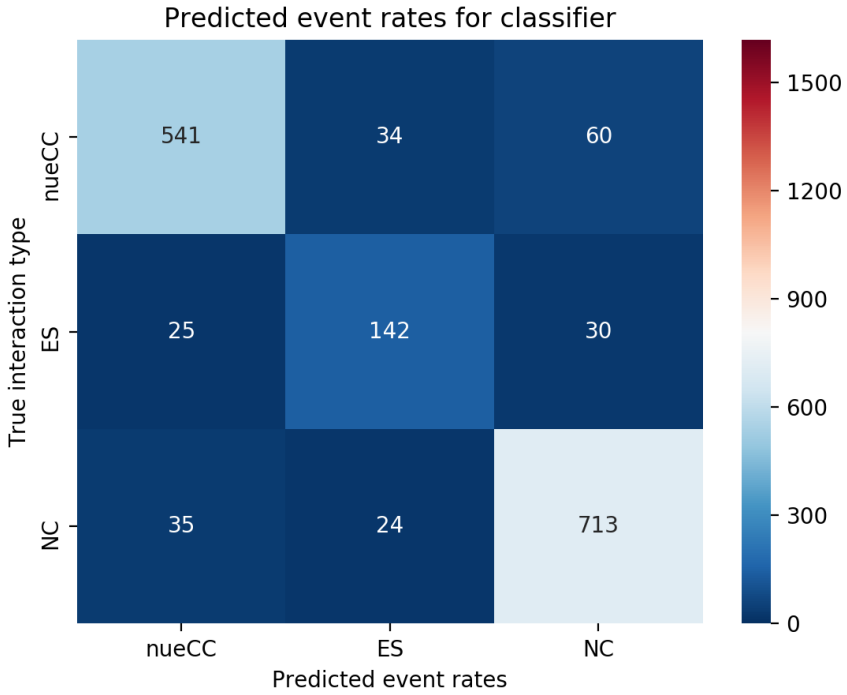
Observations from output distributions:

- Many NC events rarely classified as ES (vice versa)
- The distribution don't look overtrained by eye
- ES tends to bleed into nueCC, NC more – will have the worst efficiency among the three channels

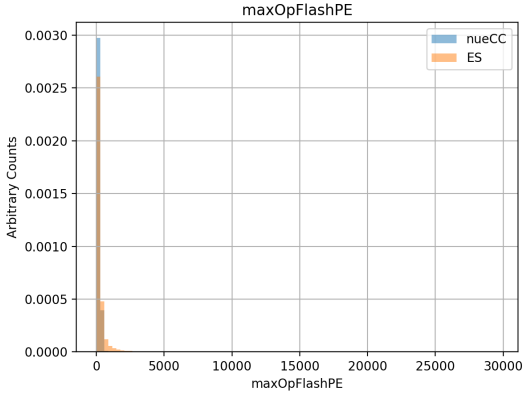


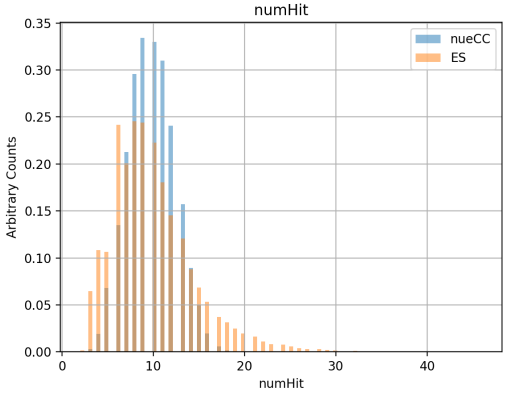
Confusion matrix and predicted events: below 12.585 MeV

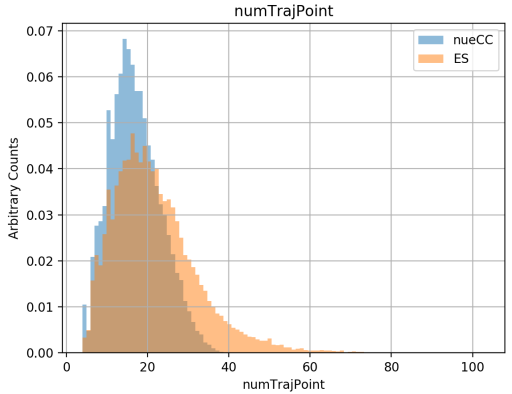


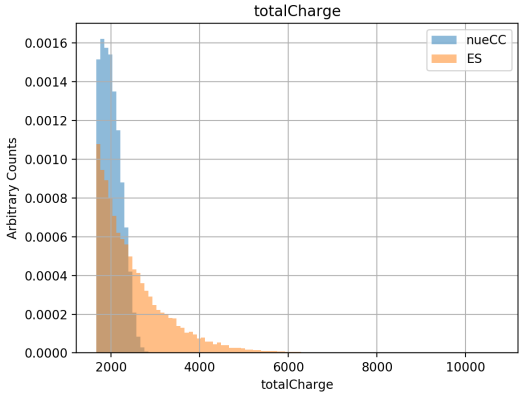


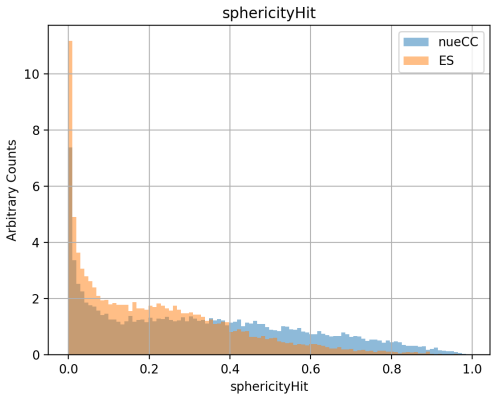
Variables: above 12.585 MeV



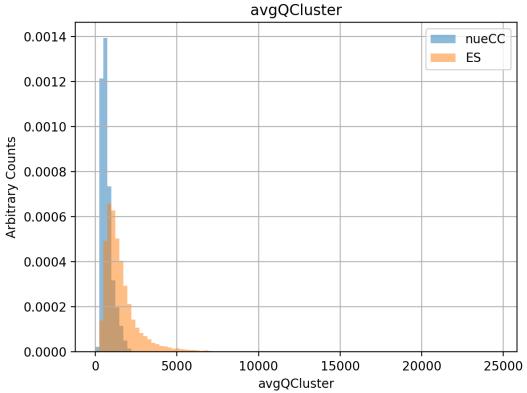


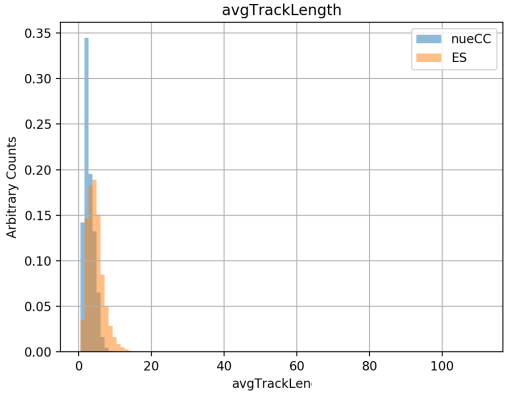


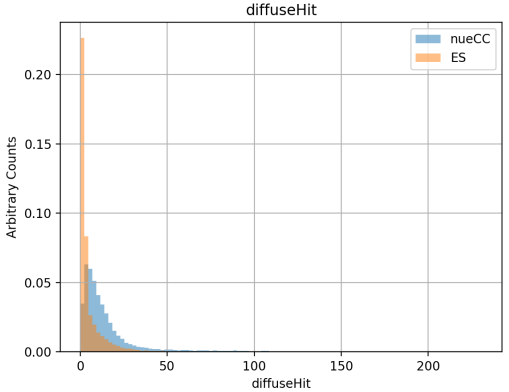


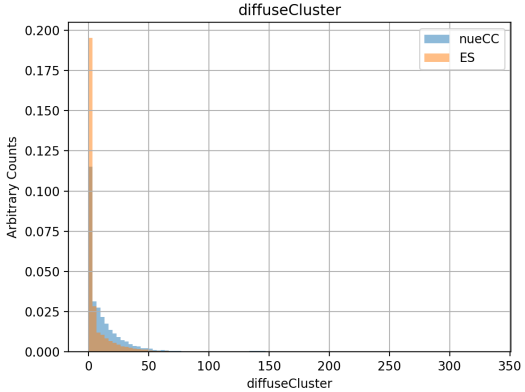


Variables: above 12.585 MeV

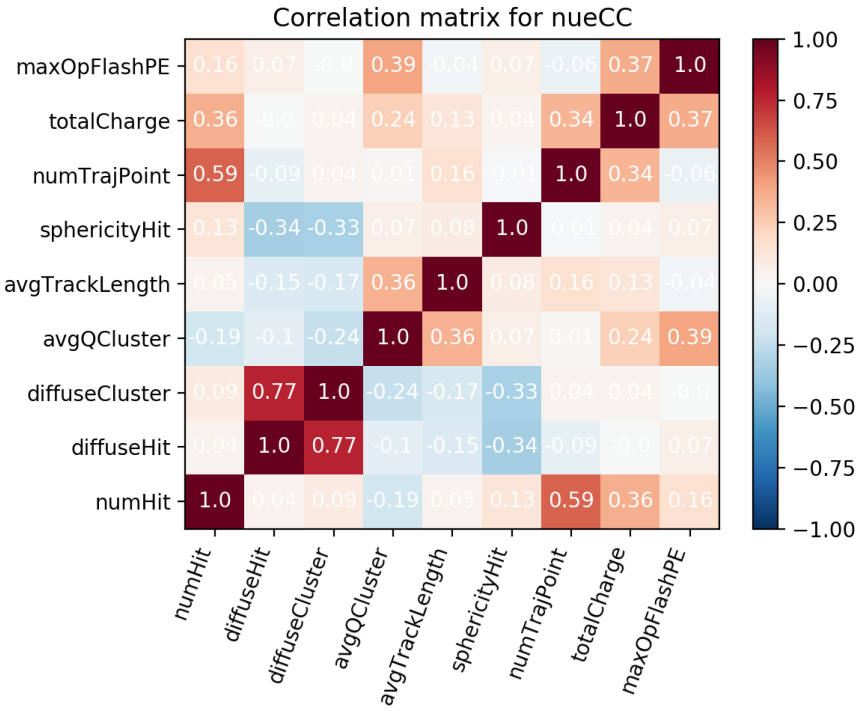


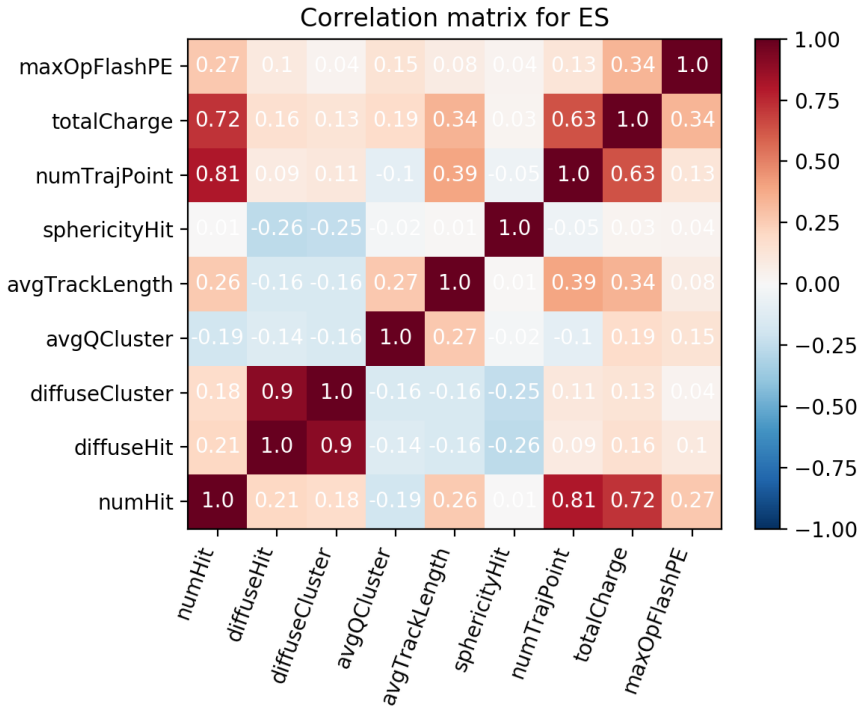




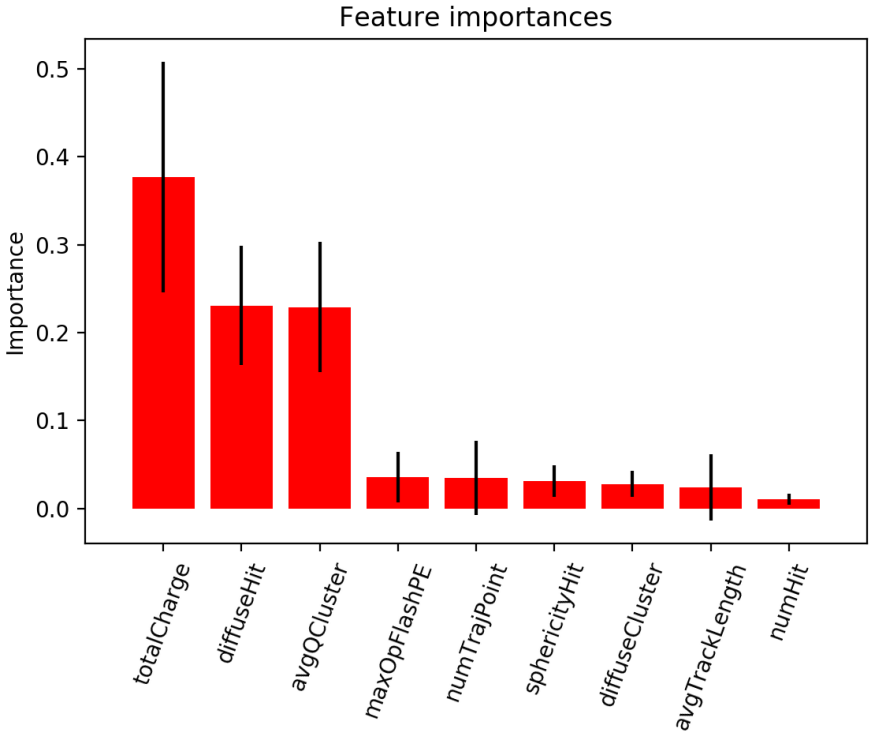


Correlation matrices: above 12.585 MeV



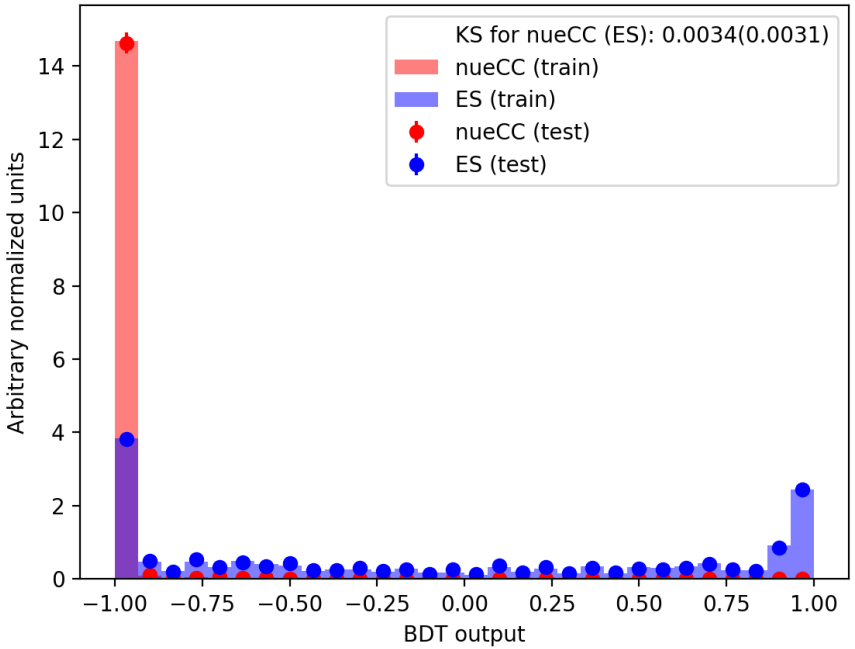


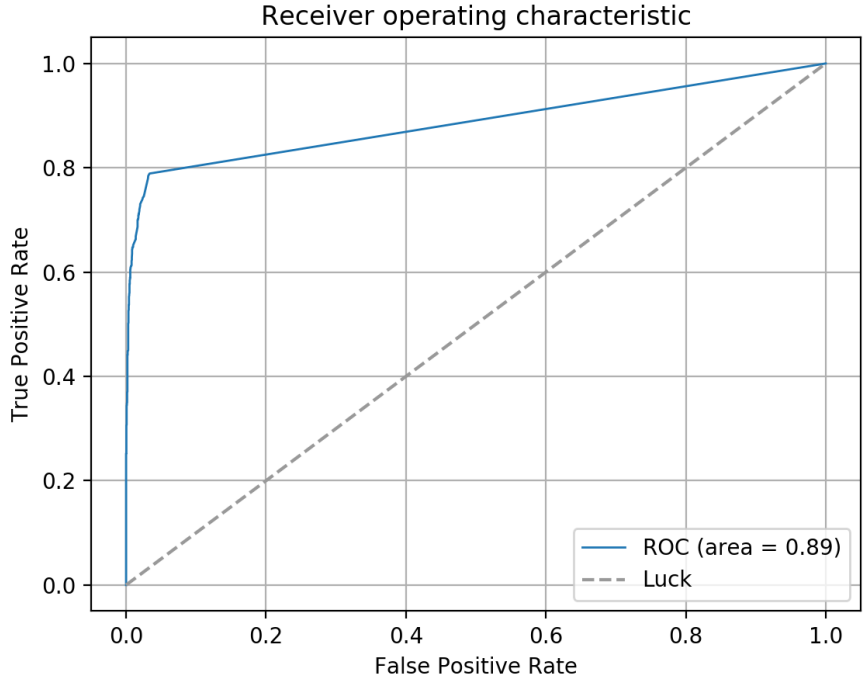
BDT statistics/results: above 12.585 MeV



- Overall purity: 77.5%
- ν_eCC efficiency (purity): 100% (64%)
- ES efficiency (purity): 45% (100%)

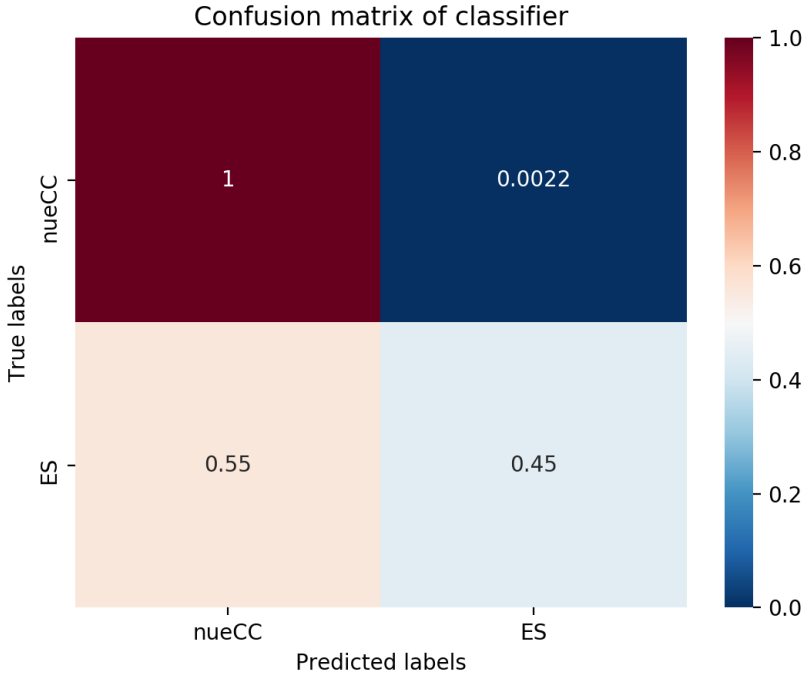
BDT output: above 12.585 MeV

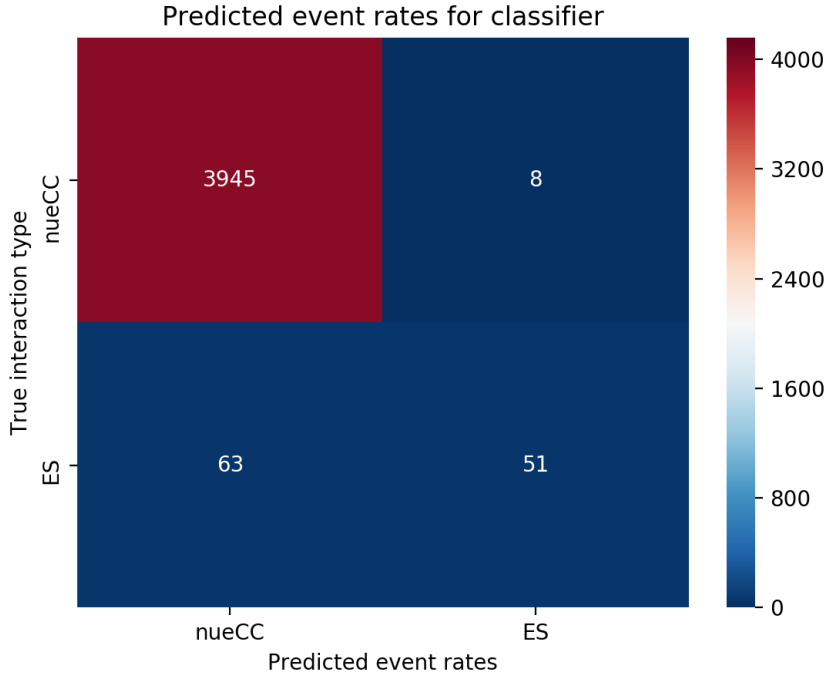




Note: the “notch” indicates tagging issues with at least one channel

Confusion matrix and predicted events above 12.585 MeV





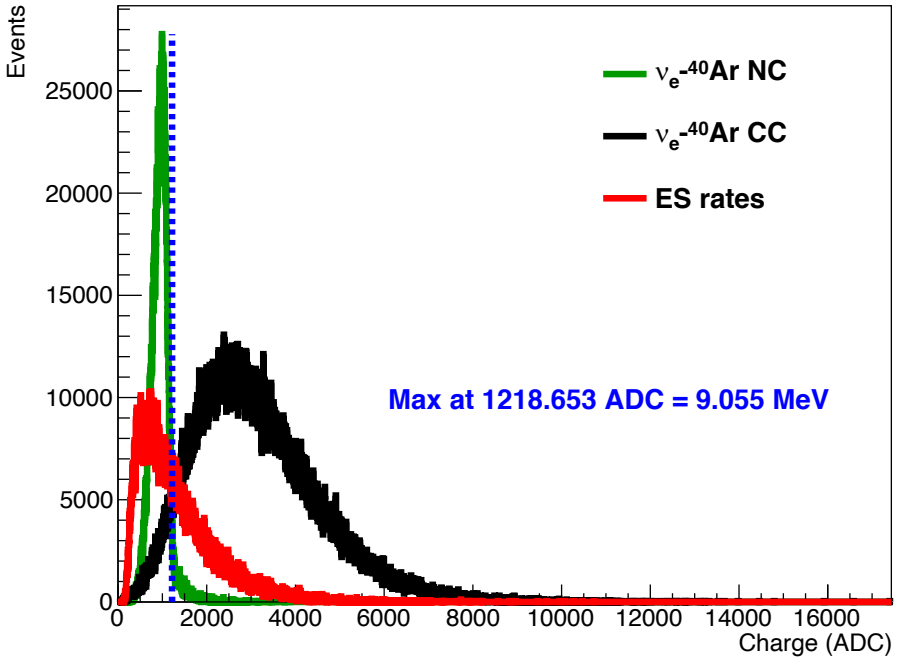
Study to optimize the E_{reco} cut

Significance study to determine optimal energy cut

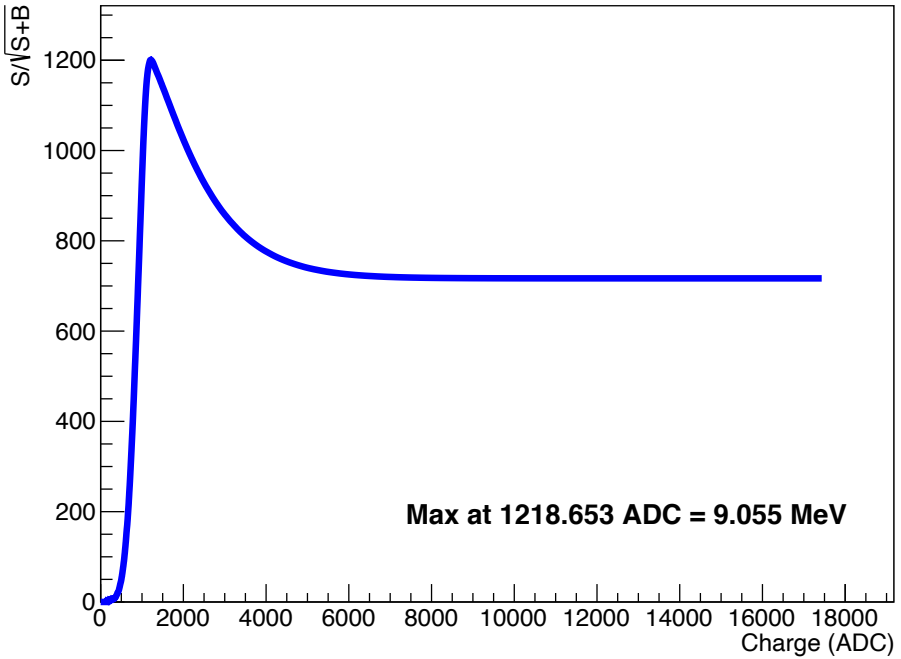
- Per Dan's suggestion: choose optimized cut based on significance metric, $S/\sqrt{S+B}$
  - S: NC events
  - B: ν_eCC and ES events
- This will isolate most of the NC signal, but there will be a small fraction of "high-energy" NC events that we will want to tag

LArSoft simulations

Drift-corrected charge distributions



$S/\sqrt{S+B}$ versus charge

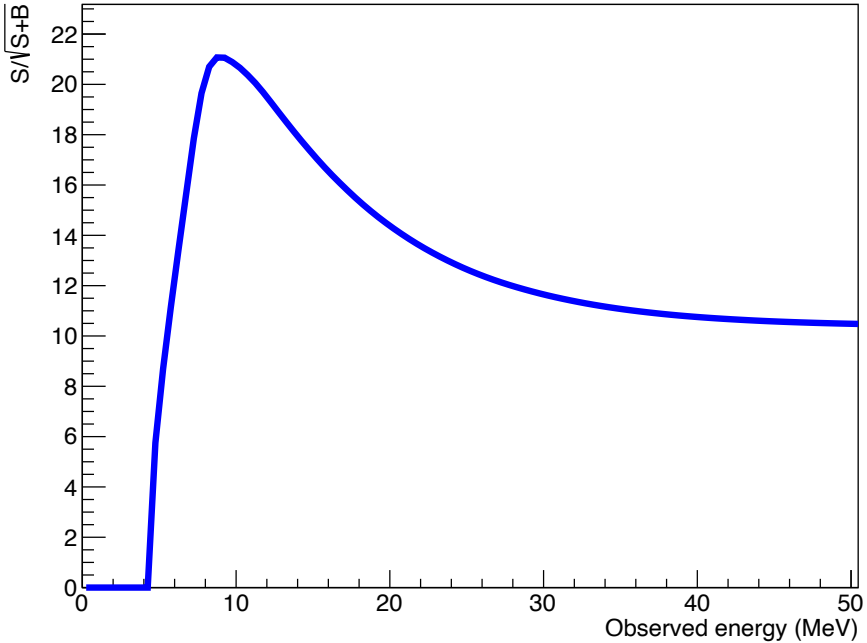


Small NC fraction above the cut; much more ES signal above the cut, which might help with tagging efficiency

Significance for SNOwGLoBES rates

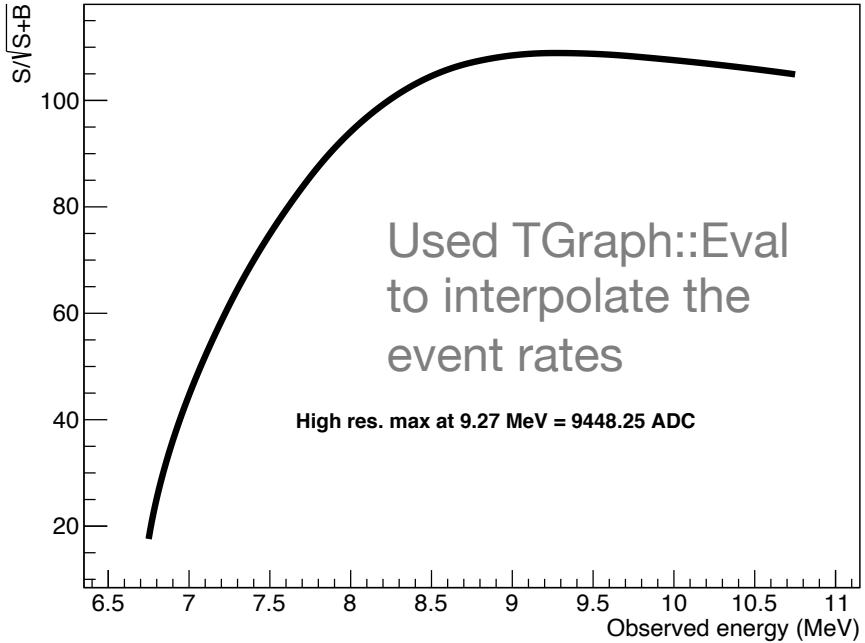
0.5 MeV bin width

$S/\sqrt{S+B}$ versus observed energy



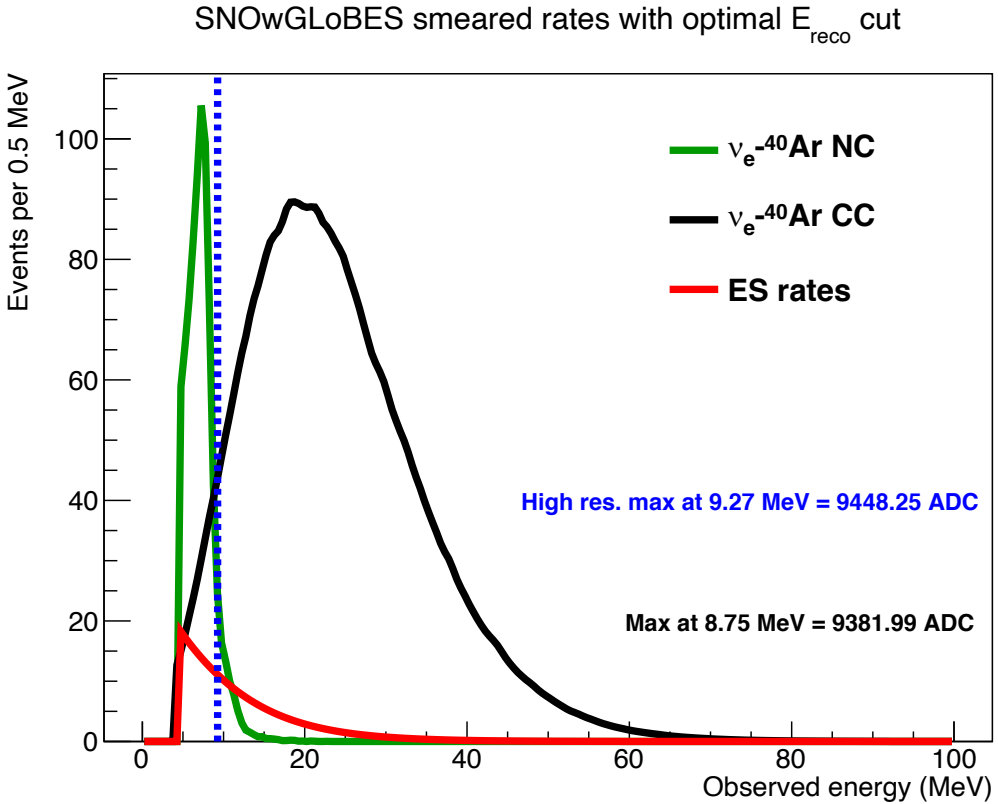
0.01 MeV bin width

$S/\sqrt{S+B}$ high res version



SNOwGLoBES rates with cut

- SNOwGLoBES cut agrees with LArSoft simulations (9.055 MeV vs 9.27 MeV)
- Same conclusions as LArSoft study: small “high-energy” NC tail, more ES included above the cut



Takeaways from significance study

- Will use energy cut determined from LArSoft simulations, but it's nice to see that SNOwGLoBES and LArSoft ~agrees
- Optimized energy cut: $9.055\ \mathrm{MeV} \approx 1218.65\ \mathrm{ADC}$
- Three-channel classification algorithms for both sides of the cut

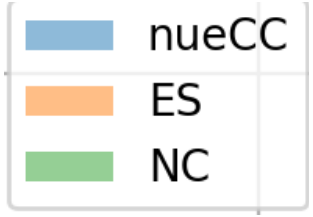
New weights from SNOwGLoBES smeared rates for NMO pinched-thermal flux

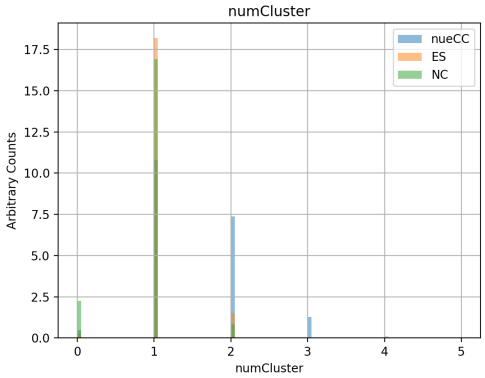
- Weights below the cut:
  - ν_eCC: 258.214/1087.97 = 0.237
  - ES: 133.381/1087.97 = 0.123
  - NC: 694.328/1087.97 = 0.638
- Weights above the cut:
  - ν_eCC: 4333.15/4687.37 = 0.924
  - ES: 180.954/4687.37 = 0.039
  - NC: 89.2473/4687.37 = 0.019

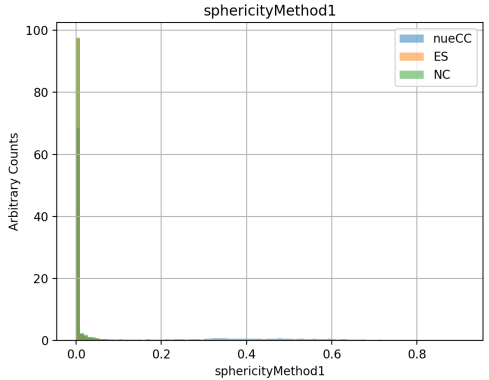
Results for optimized NC cut

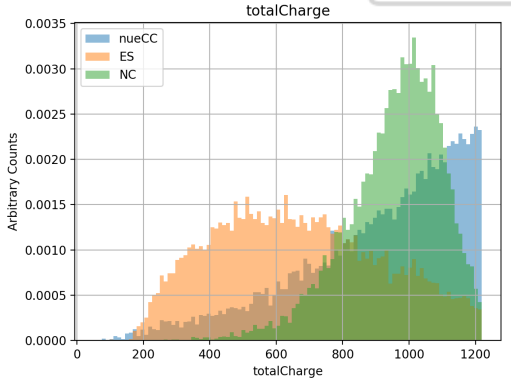
9.055 MEV CUT

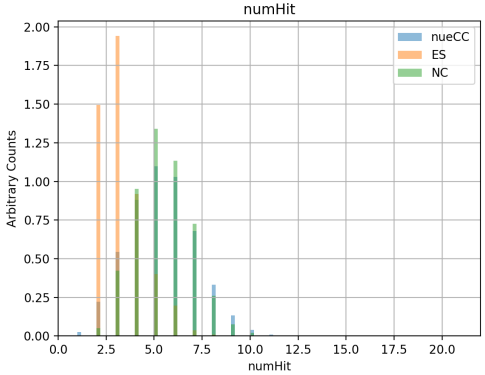
Variables: below 9.055 MeV

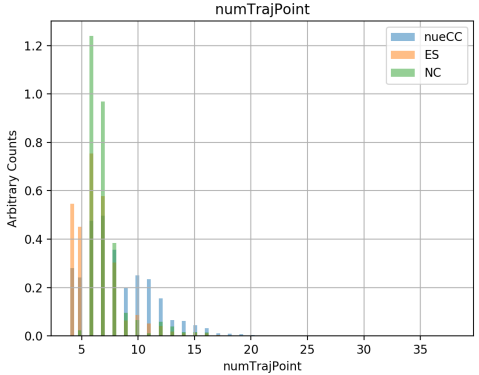


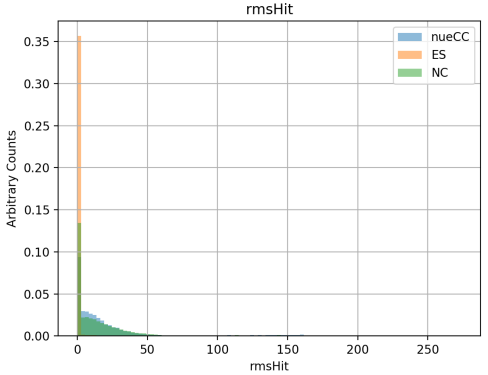




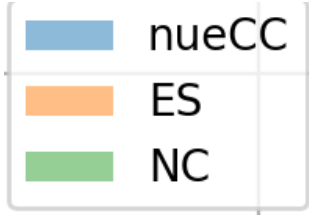


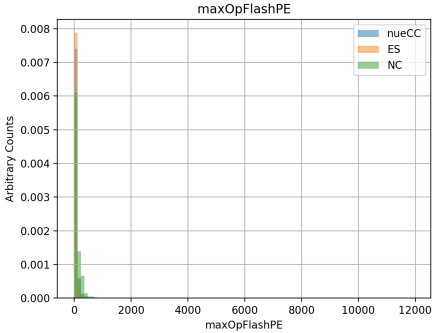


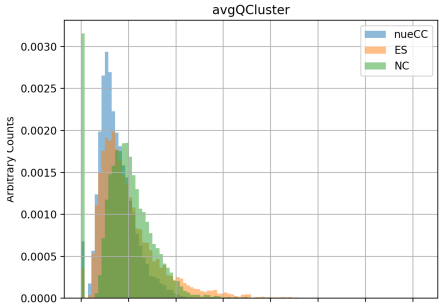


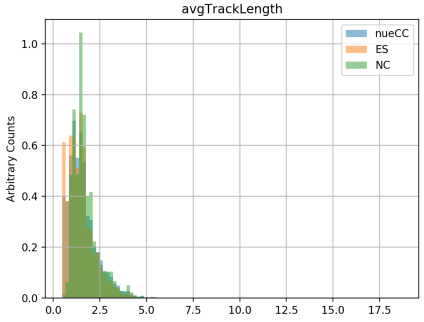


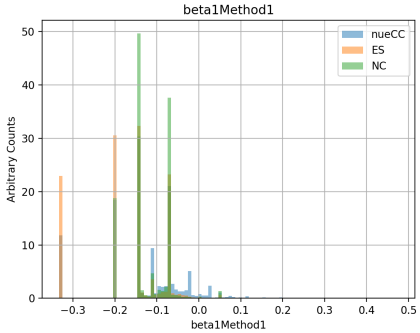
Variables: below 9.055 MeV

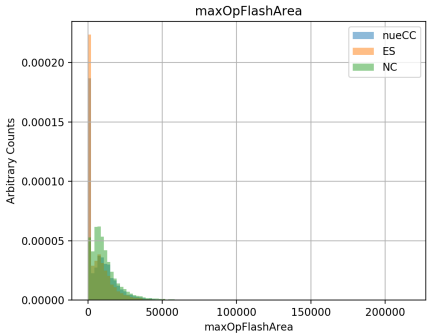


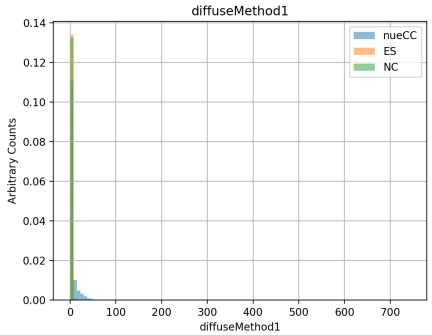


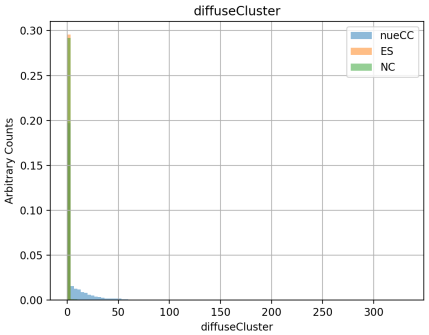




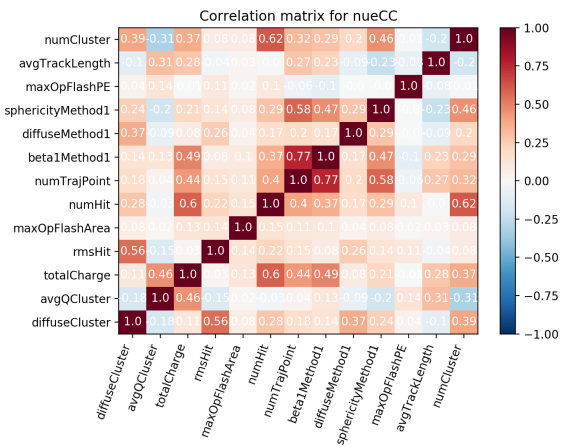


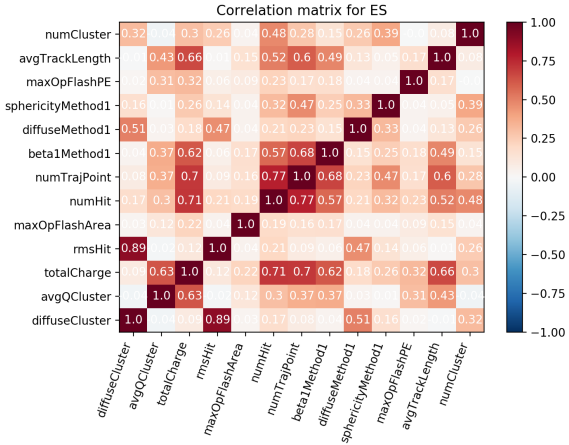


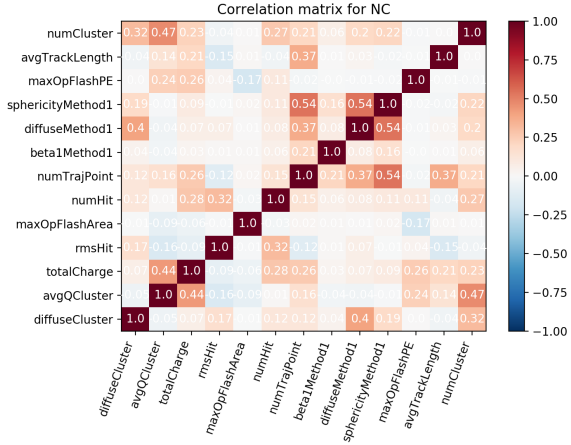




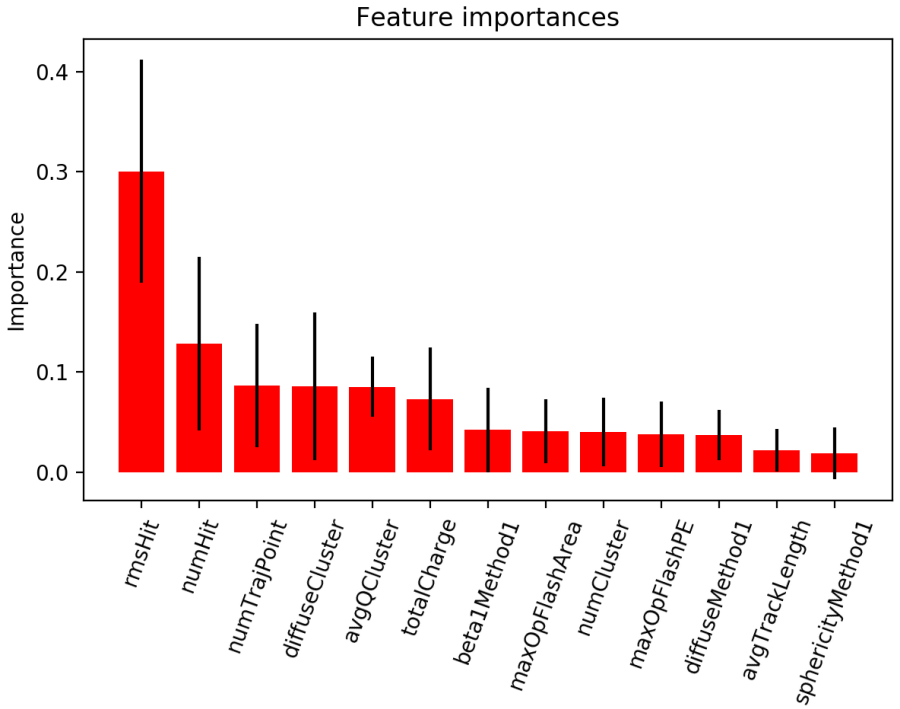
Correlation matrices: Below 9.055 MeV





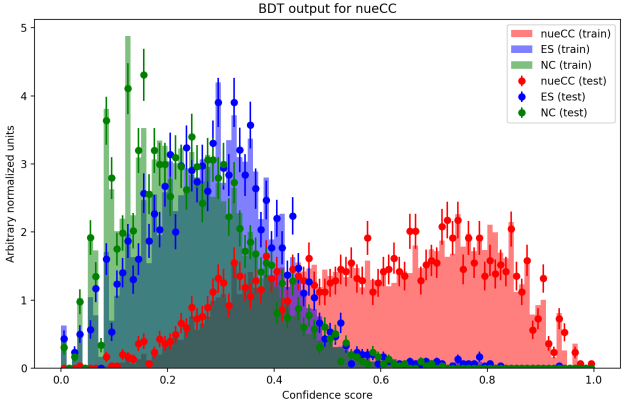


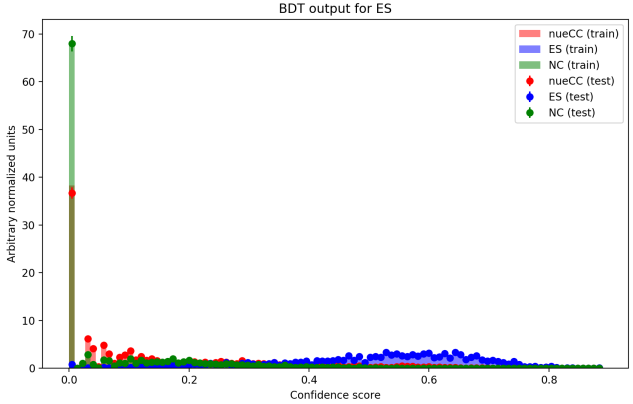
BDT statistics/results: below 9.055 MeV



- Overall purity: 79.5%
- ν_eCC efficiency (purity): 69% (84%)
- ES efficiency (purity): 76% (88%)
- NC efficiency (purity): 94% (71%)

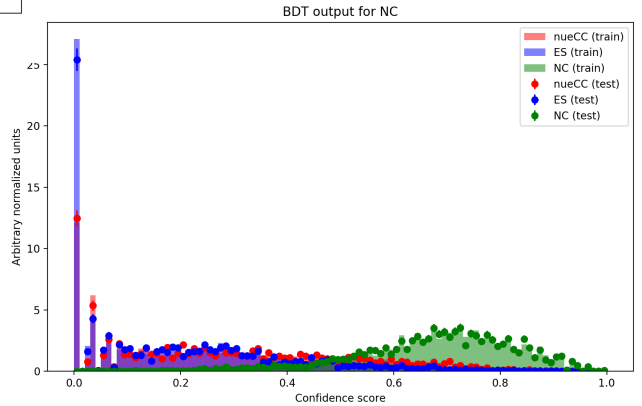
BDT outputs: below 9.055 MeV



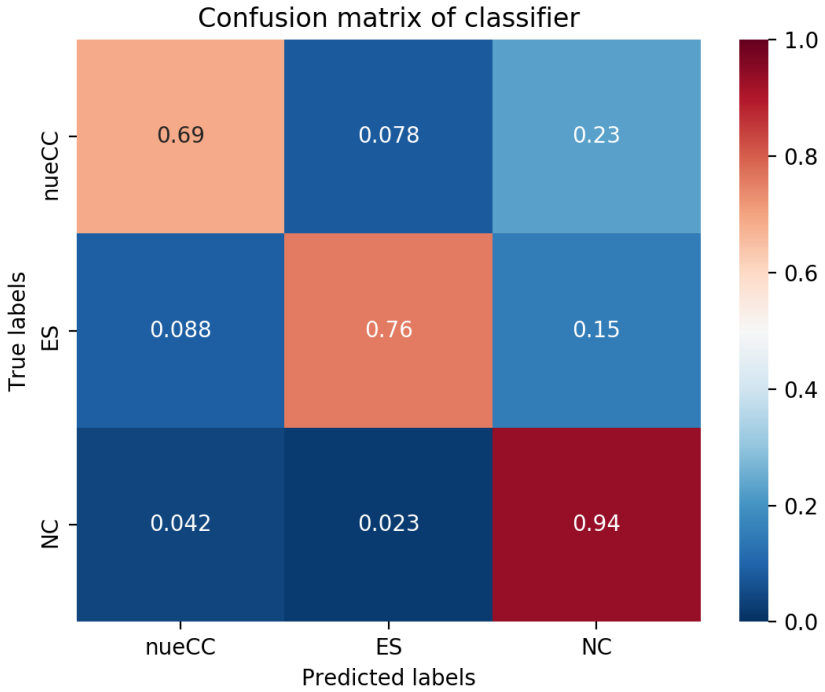


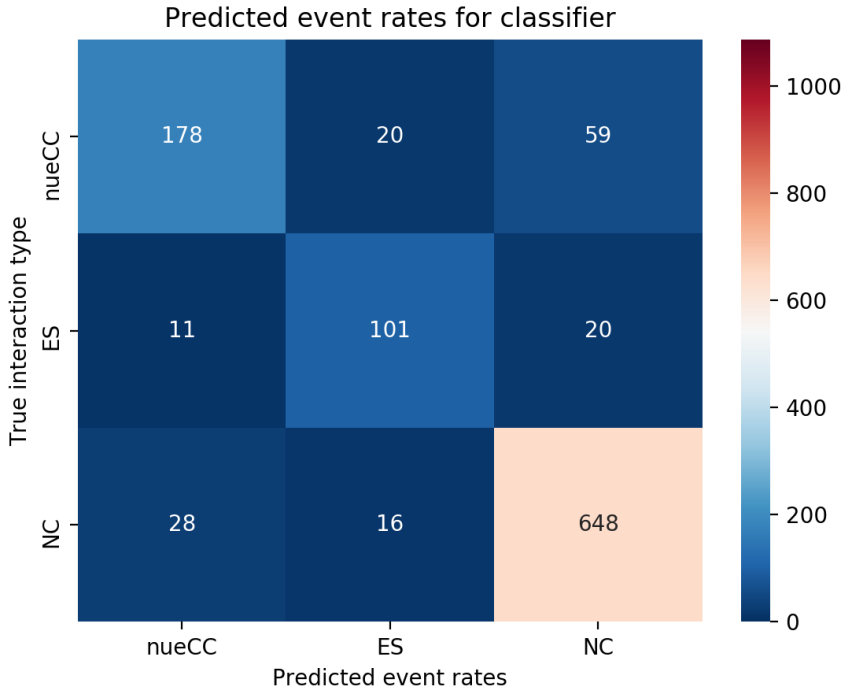
Observations from output distributions:

- Less obvious that the distributions aren't overtrained
- NC rarely misclassified as ES or nueCC

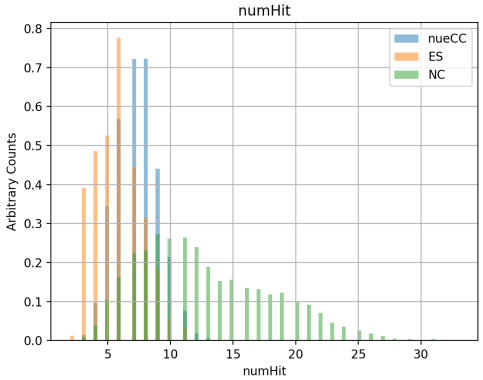


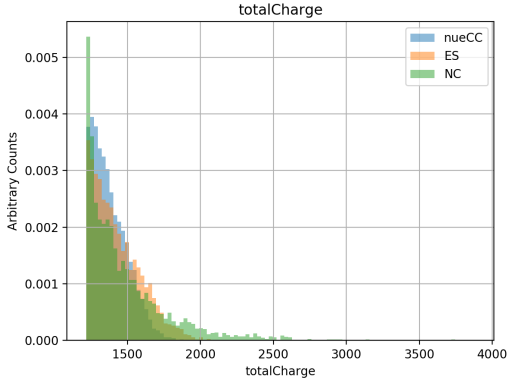
Confusion matrix and predicted events: below 9.055 MeV

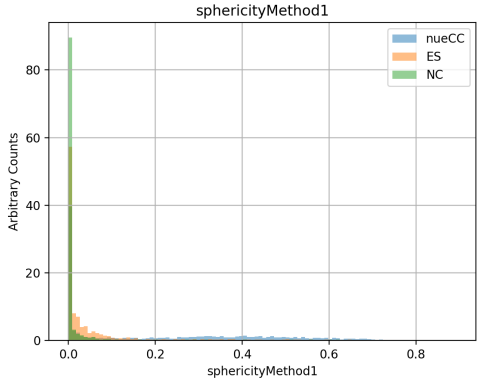


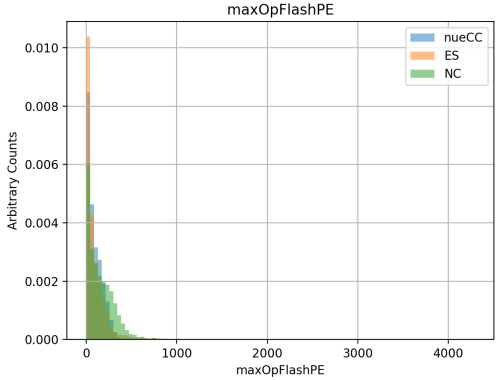


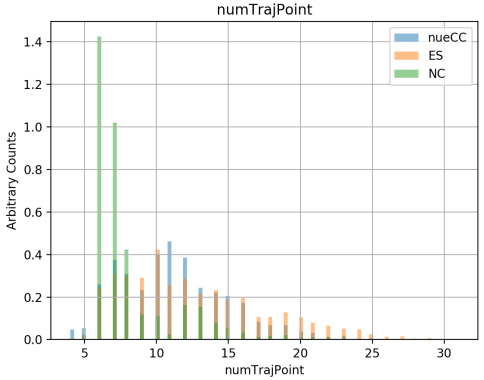
Variables: above 9.055 MeV

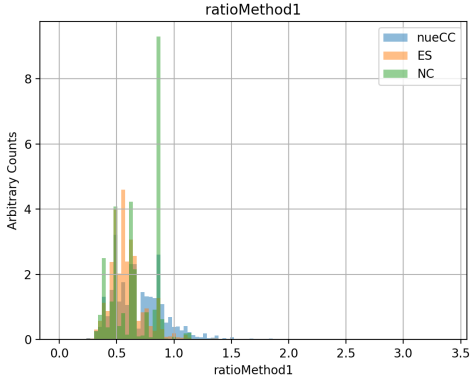




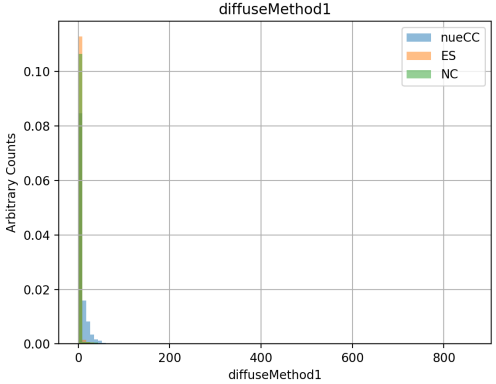


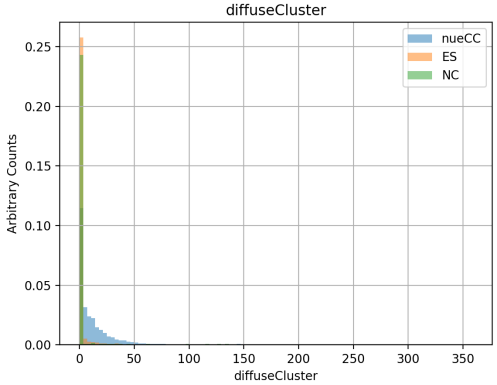


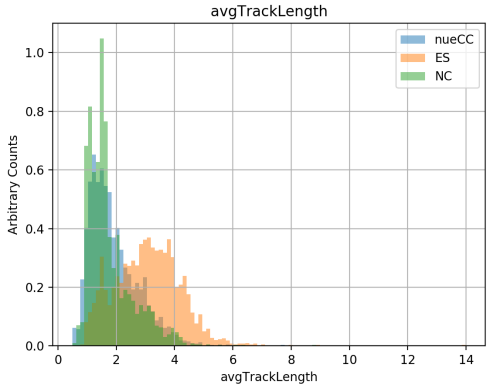


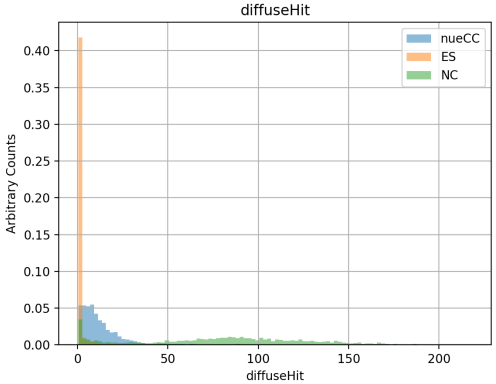


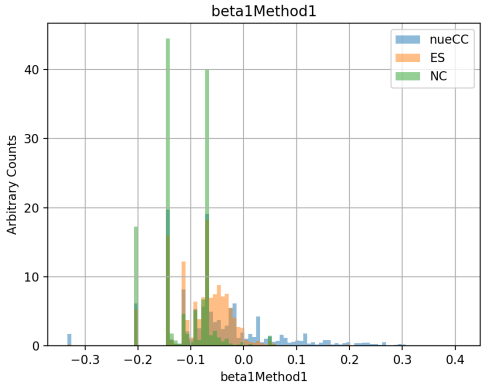
Variables: above 9.055 MeV



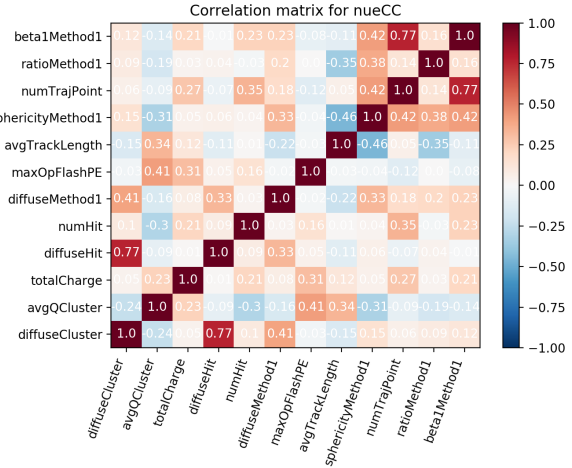


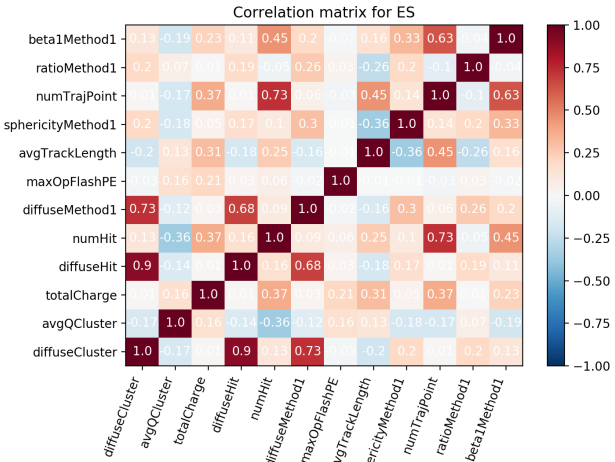


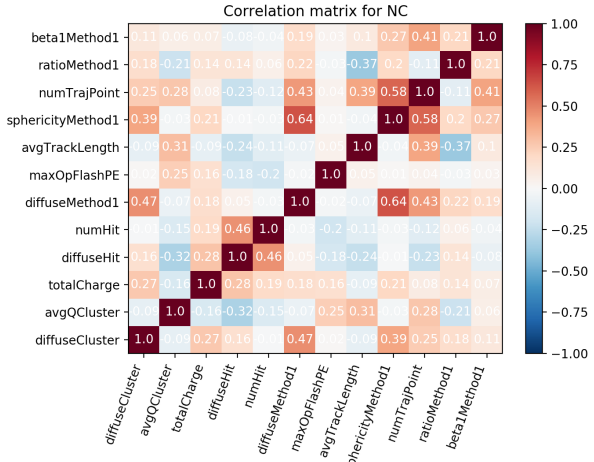




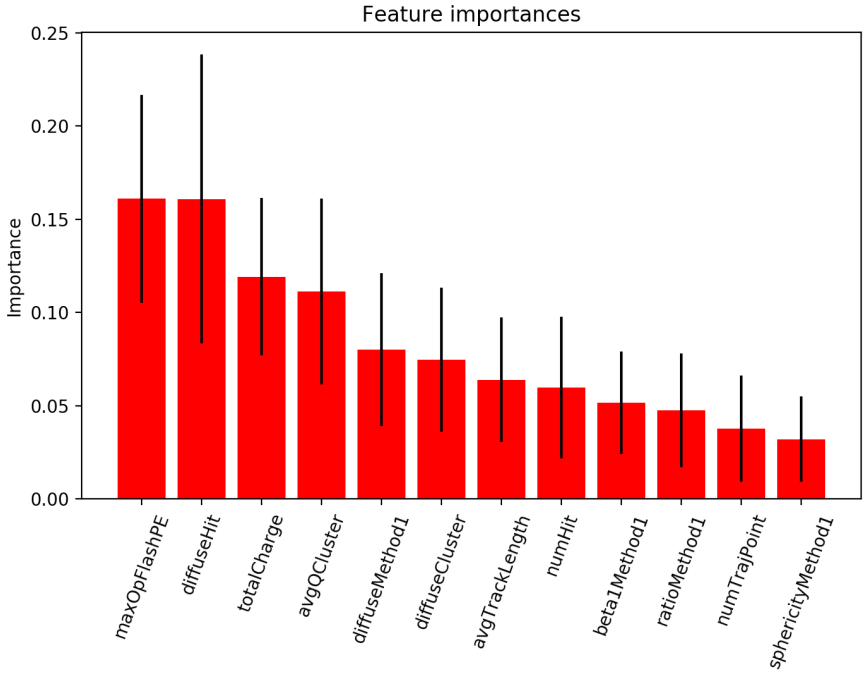
Correlation matrices: Above 9.055 MeV





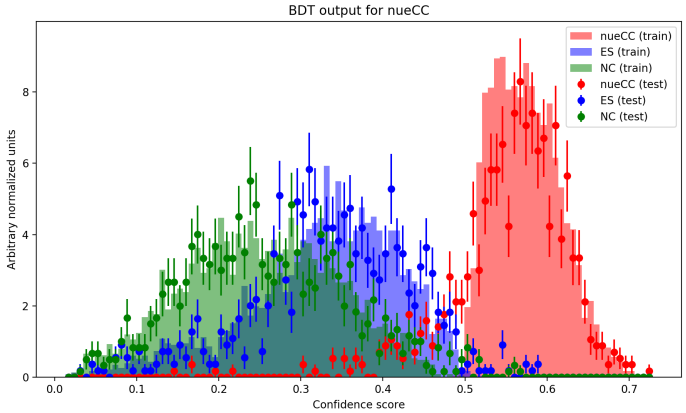


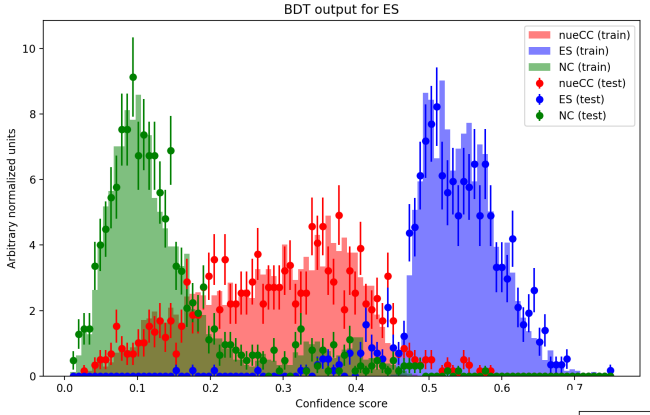
BDT statistics/results: above 9.055 MeV



- Overall purity: 94.1%
- ν_eCC efficiency (purity): 95% (90%)
- ES efficiency (purity): 93% (95%)
- NC efficiency (purity): 94% (98%)

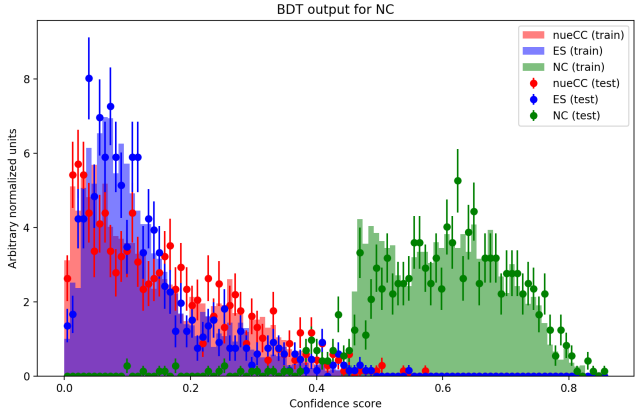
BDT outputs: Above 9.055 MeV





Observations from output distributions:

- While the training/testing samples might not visually agree, the KS statistic claims that the distributions are the same with 95% confidence
- Pretty good separation for all three channels!
- ES bleeds into nueCC and vice versa



Confusion matrix and predicted events Above 9.055 MeV