APPLYING MACHINE LEARNING TO

CLUSTER SHAPE ANALYSES

at a 10 TeV Muon Collider

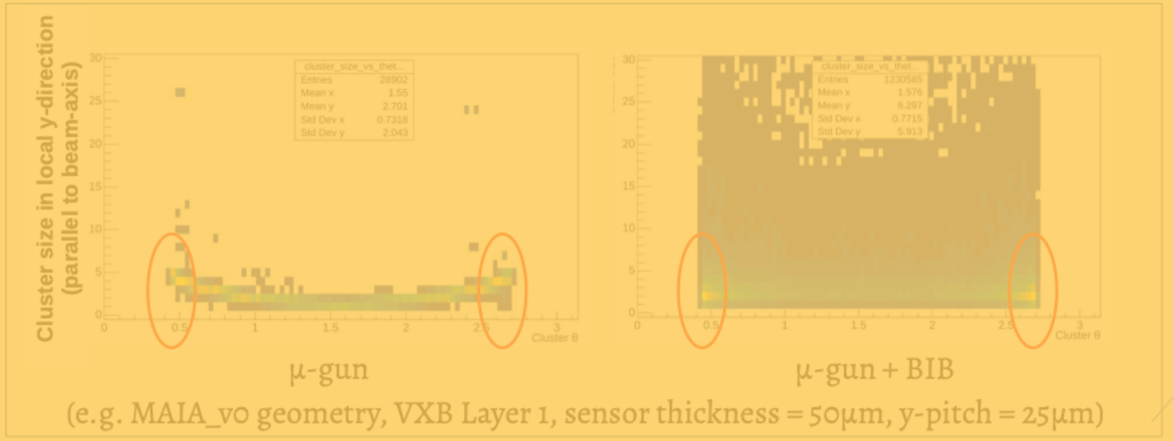
JORDAN ASHLEY

DR. ANGIRA RASTOGI

- ATLAS, Lawrence Berkeley National Laboratory

Cluster shape analysis for BIB rejection

Using correlation between incidence angle and number of pixel hits per cluster to reject long clusters – characteristic of BIB particles from muon decays.



μ-gun μ-gun + BIB

(e.g. MAIA_v0 geometry, VXB Layer 1, sensor thickness = 50μm, y-pitch = 25μm)

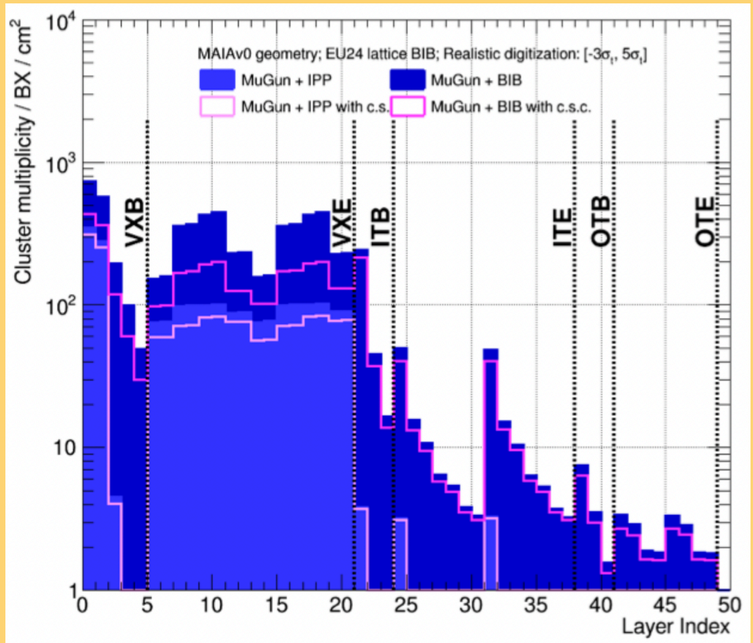
$\tan\theta = 50\mu m/y$
$\tan(0.5) \sim 0.5 = 50\mu m/y$
$y = 100\mu m$
y/pitch ~ 4 pixels

BIB particles either have very short clusters at same angles as signal (due to low-pT particles) or excessively long clusters. In both cases, we can reject them to clean the tracking environment.

DR. ANGIRA RASTOGI

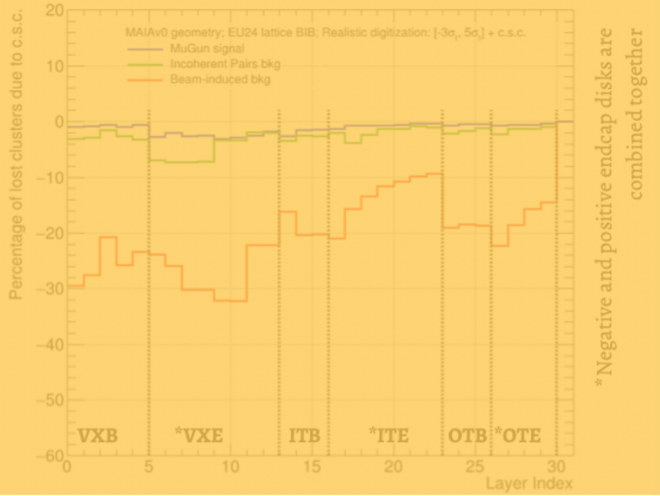
- ATLAS, Lawrence Berkeley National Laboratory

Cluster shape cuts w/ MAIA_v0



(Preliminary simple cluster shape cuts)

Detailed selection in digi_steer_MAIAv0.py



With less than 5% loss of prompt signal clusters, we can cut down BIB clusters upto 20-30% from each layer of various subdetectors!

DR. ANGIRA RASTOGI

- ATLAS, Lawrence Berkeley National Laboratory

ML CAN DO BETTER!

JORDAN ASHLEY, UNDERGRAD

I KNOW NOTHING ABOUT ML!

BDT INPUT VARIABLES

CLUSTER VARIABLES

- Position: x, y, and z, incident angle
- Size: local x and y
- Total size (number of pixels)
- Energy deposited
- Time of arrival
- Calculated from pixel hits:
  - RMS for x and y
  - Skewness in x and y
  - Aspect ratio

PIXEL HIT VARIABLES

- Top 9 pixels in order of energy deposit:
  - Energy deposited
  - Time of arrival

Grand total:

32 input variables

PREVENTING OVERTRAINING

ADAPTIVE BOOSTING

- Lots of trees
- Each tree is shallow
- Additional weight given to misclassified clusters

TMVA VARIABLE WEIGHTING

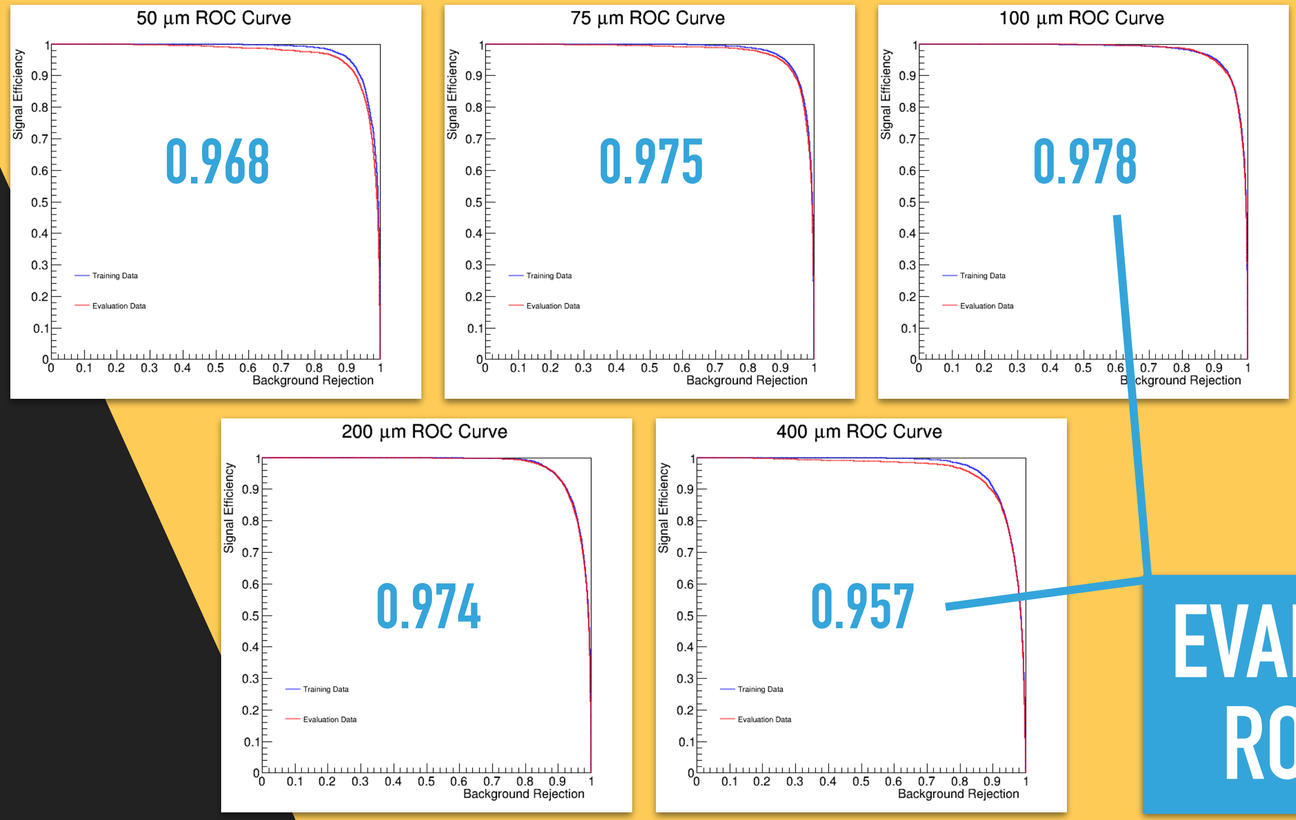
- Calculates correlation coefficients
- Assigns preliminary weights

+ LARGE NUMBER OF CLUSTERS

OPTIMIZING PERFORMANCE

HOW DO WE KNOW WHEN WE'VE GOT IT RIGHT?

- Change hyperparameter(s)
- Run BDT → get ROC
- Iterate ad nauseam
- Compare ROC AUCs



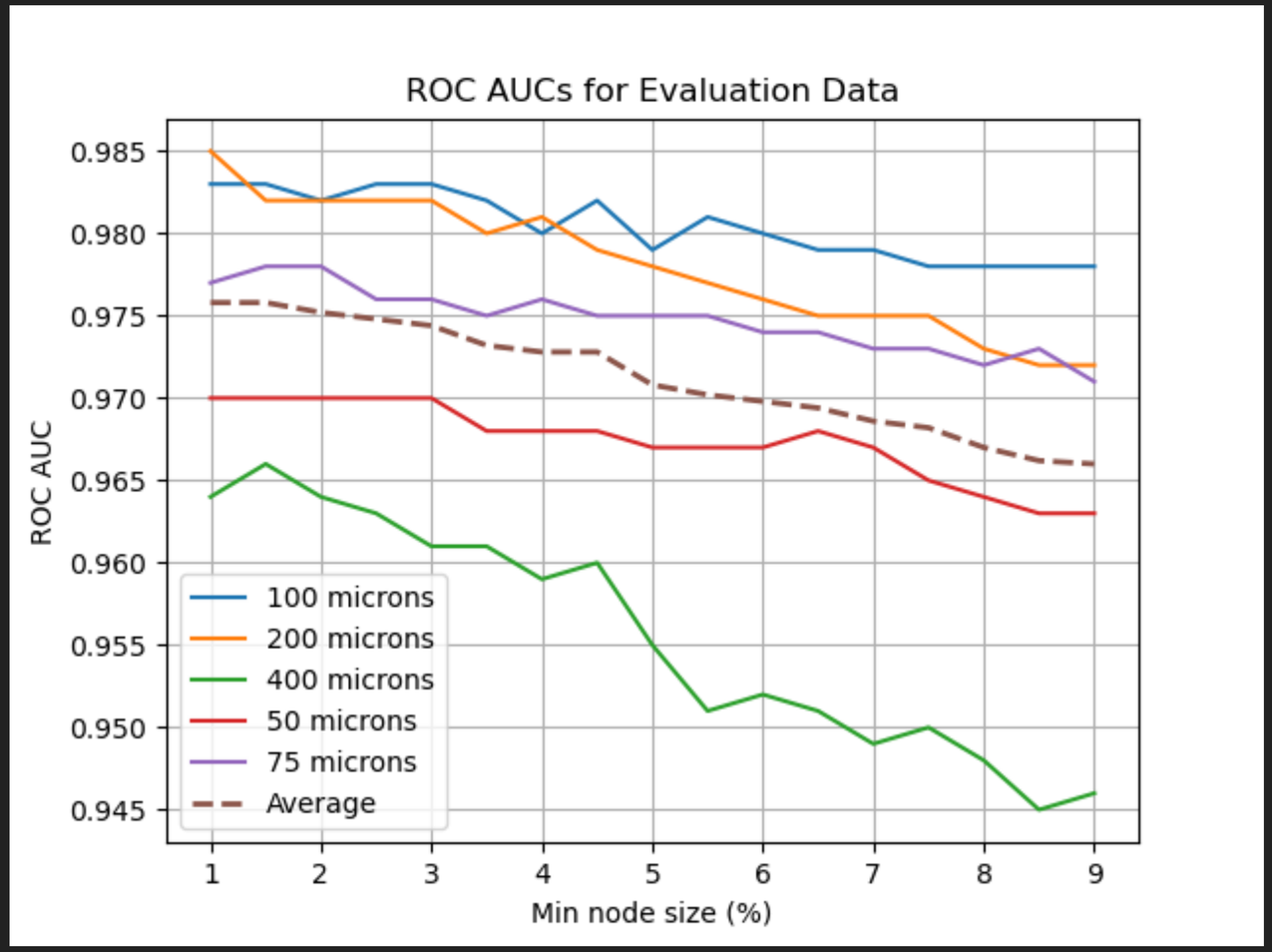
EVALUATION ROC AUC

OPTIMIZING PERFORMANCE

800 trees, 50/50 train/test split, 0.5 adaptive boost beta, max depth of 3

MINIMUM NODE SIZE

How big can each leaf be? Specified as % of available data for each set.



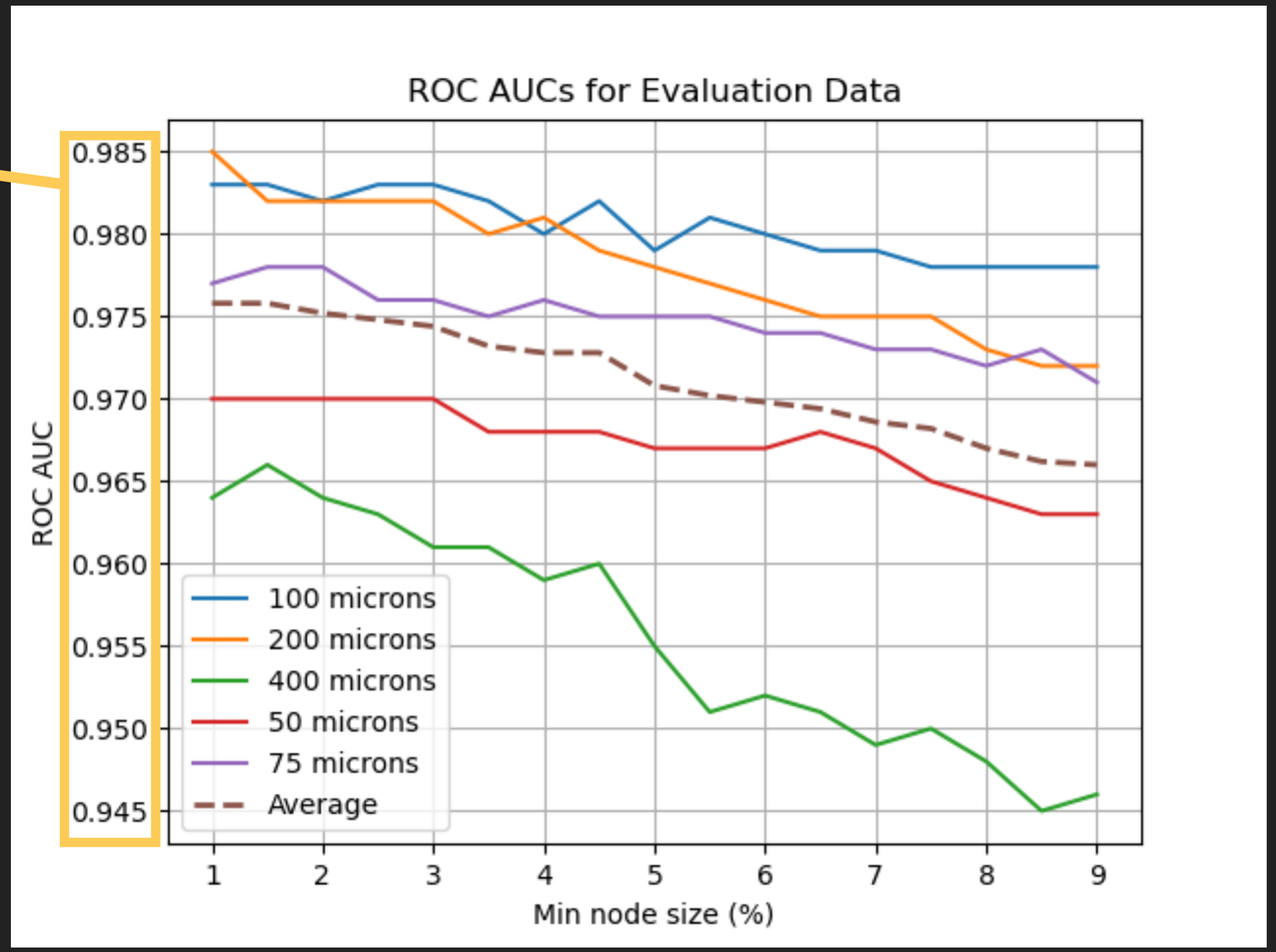
OPTIMAL: 1.5%

- Min signal clusters: 4536
- Min background clusters: 29693

OPTIMIZING PERFORMANCE

WHAT ABOUT PERFORMANCE ERROR?

- Very small range of AUC values
- Basic iteration costs lots of time + computing resources

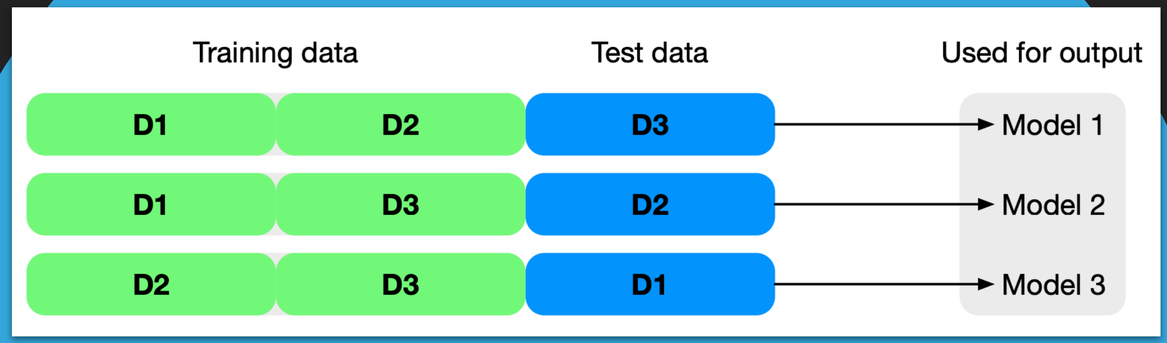


K-FOLD CROSS VALIDATION

THERE'S A TMVA VALIDATION PROCEDURE FOR THAT!

- Current project stage
- Runs BDT on each fold
- Train/test split → K-1 train "folds" (chunks of data) and 1 test fold
- ROC AUC mean and standard deviation automatically calculated

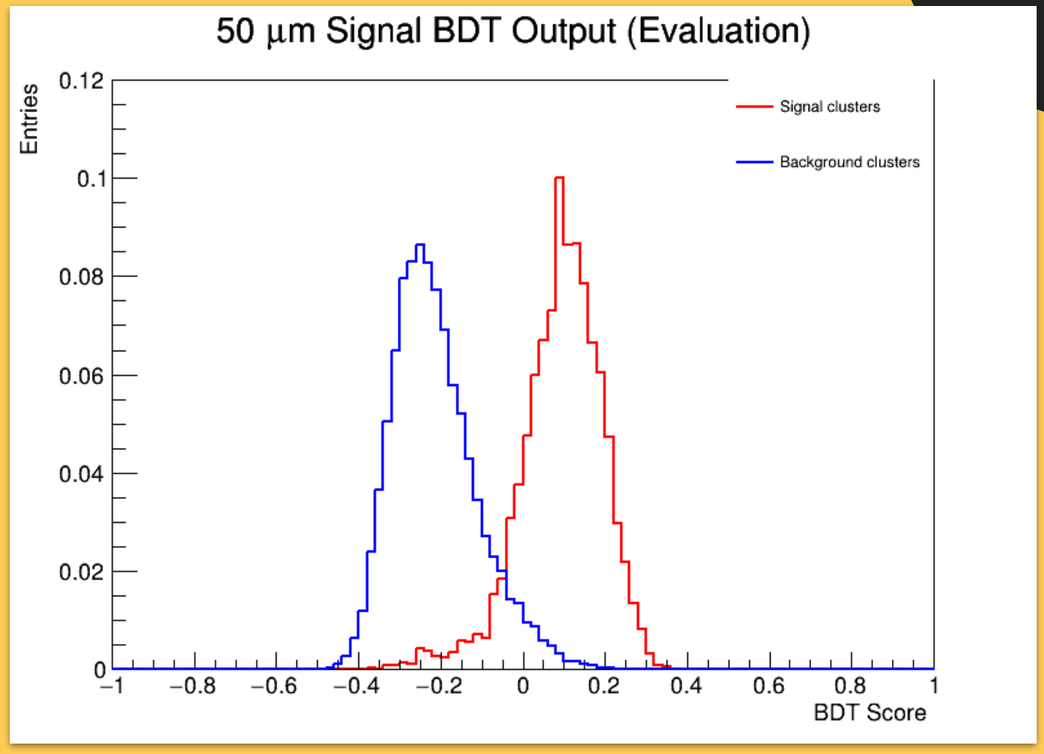
Example from TMVA Users Guide



$K = 3$

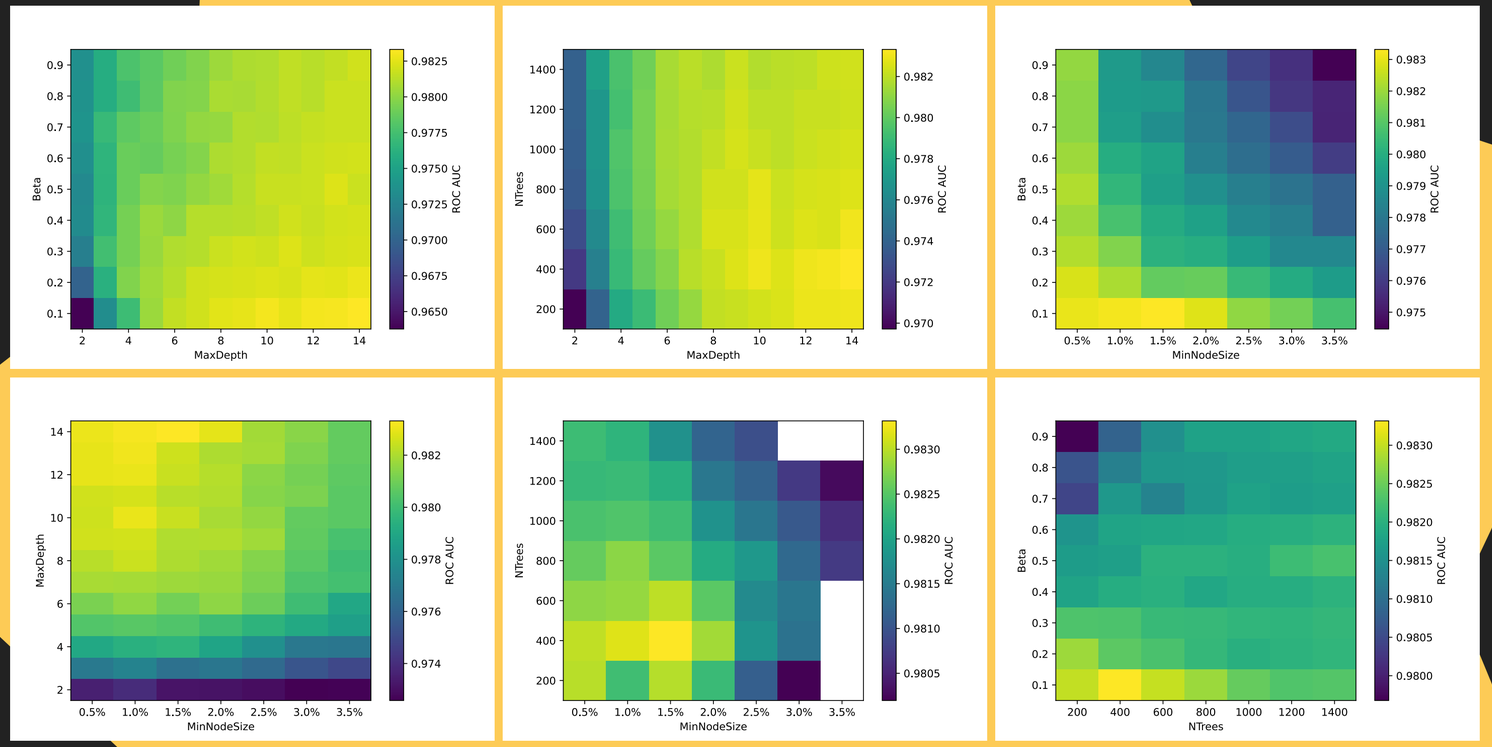
CURRENT STATUS

- Looking good!



CURRENT STATUS

- Extracted ROC AUC data from k-fold procedure results



https://github.com/j-s-ashley/beta

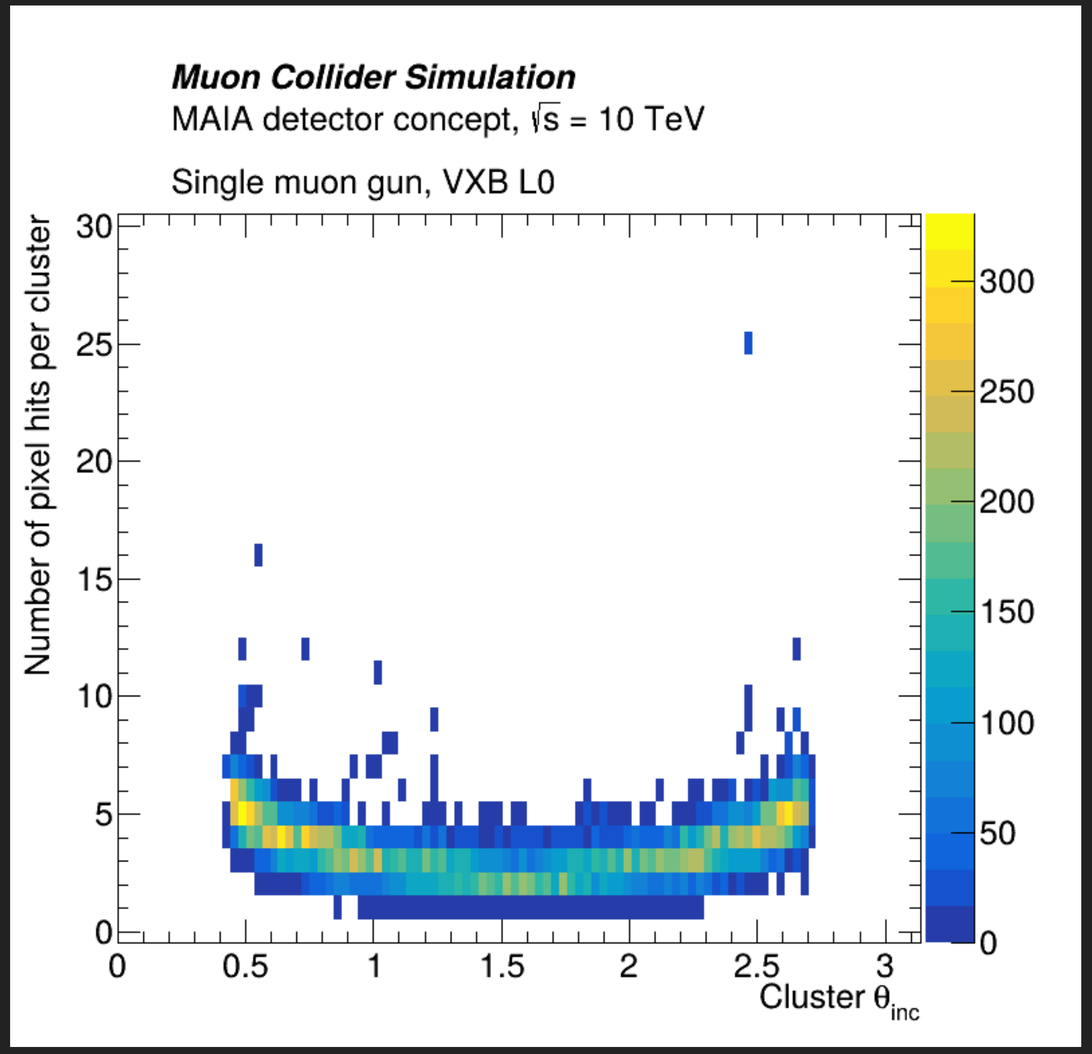
ADDITIONAL INFO

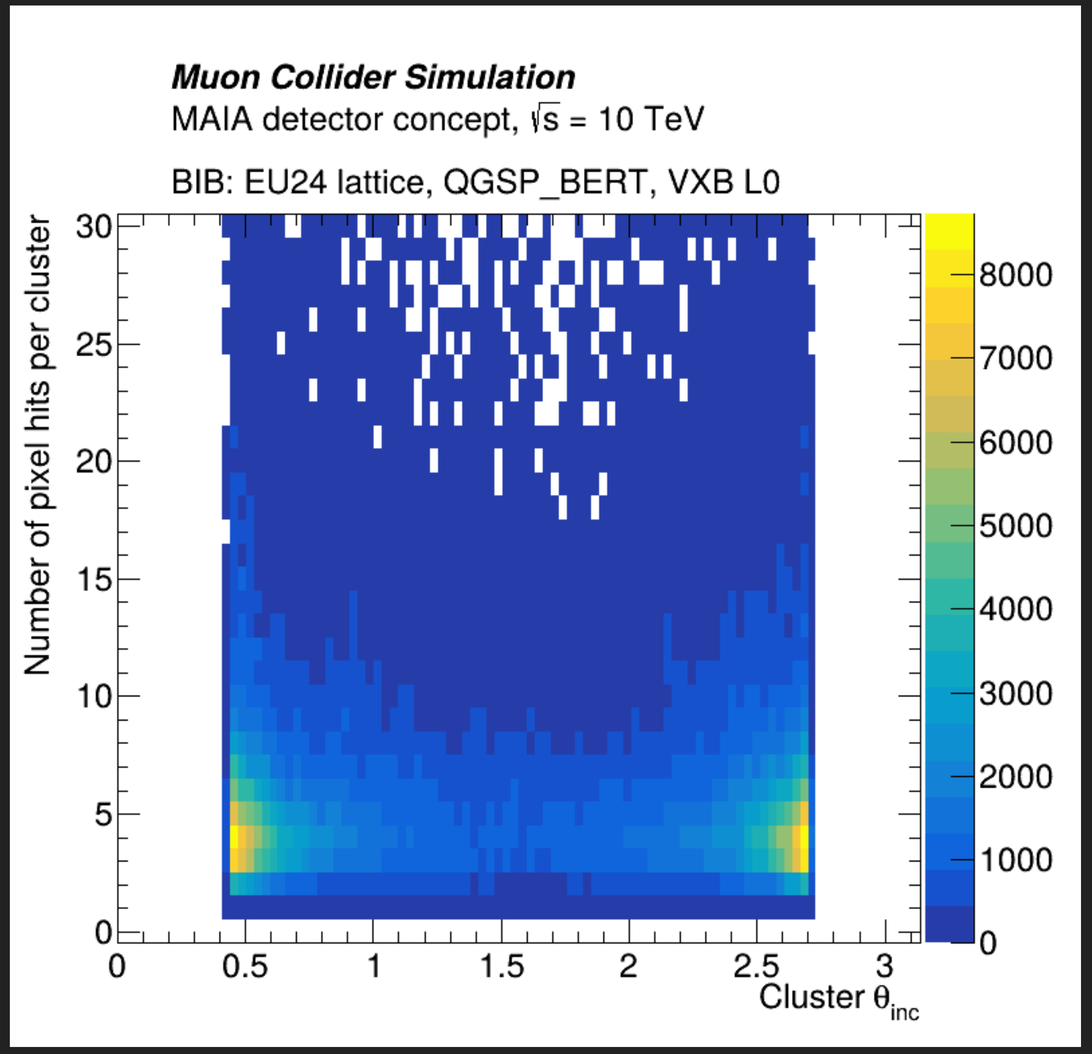
CLUSTER INCIDENT ANGLE VS SIZE

SIGNAL

Cluster size in y (parallel to beam axis), incident angle from interaction point

BACKGROUND



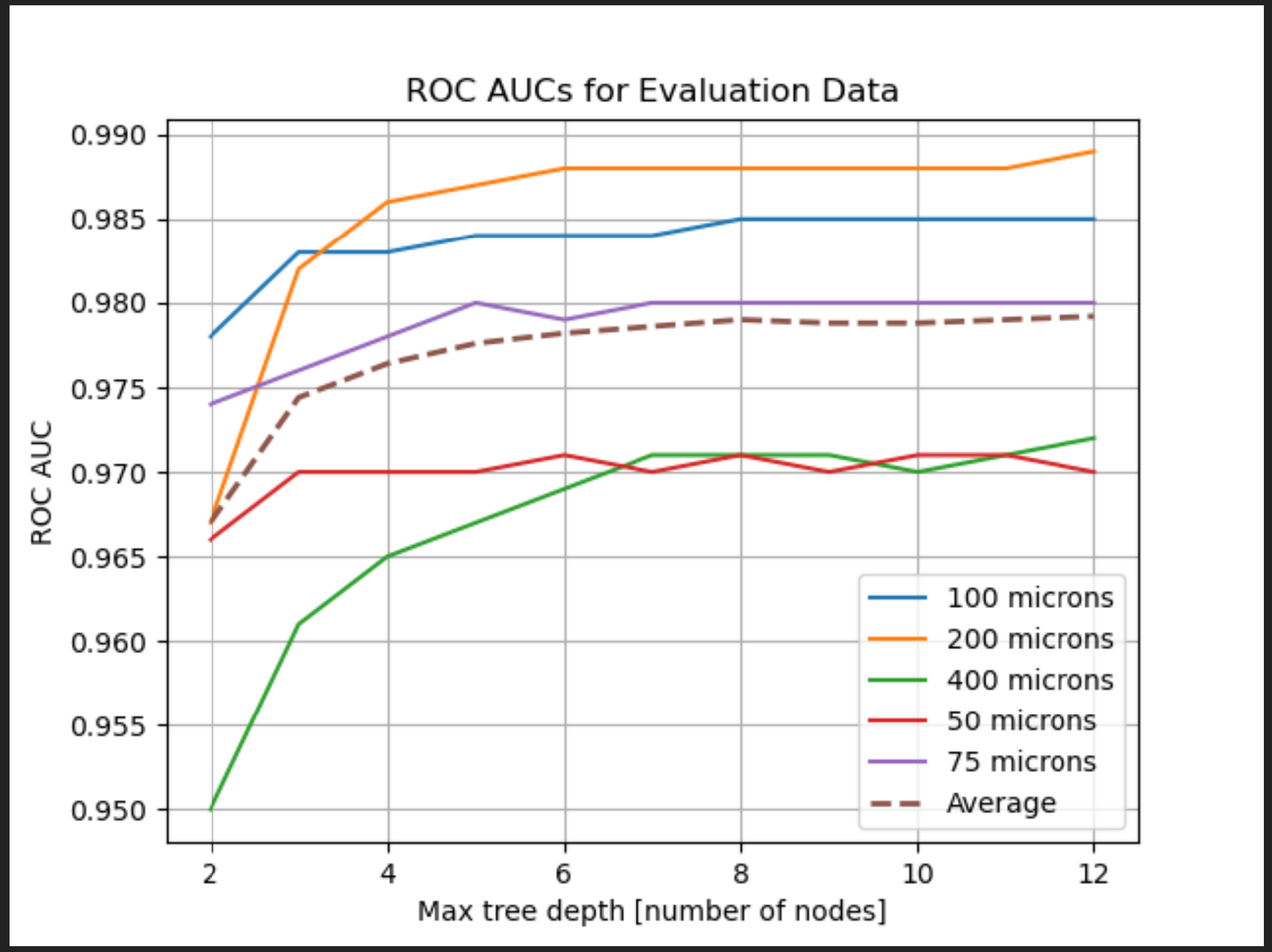


TMVA HYPERPARAMETERS

800 trees, 3% min node size, 50/50 train/test split, 0.5 adaptive boost beta

MAXIMUM DEPTH

How large can each tree grow? Specified as # of nodes (leaves).



OPTIMAL: 8

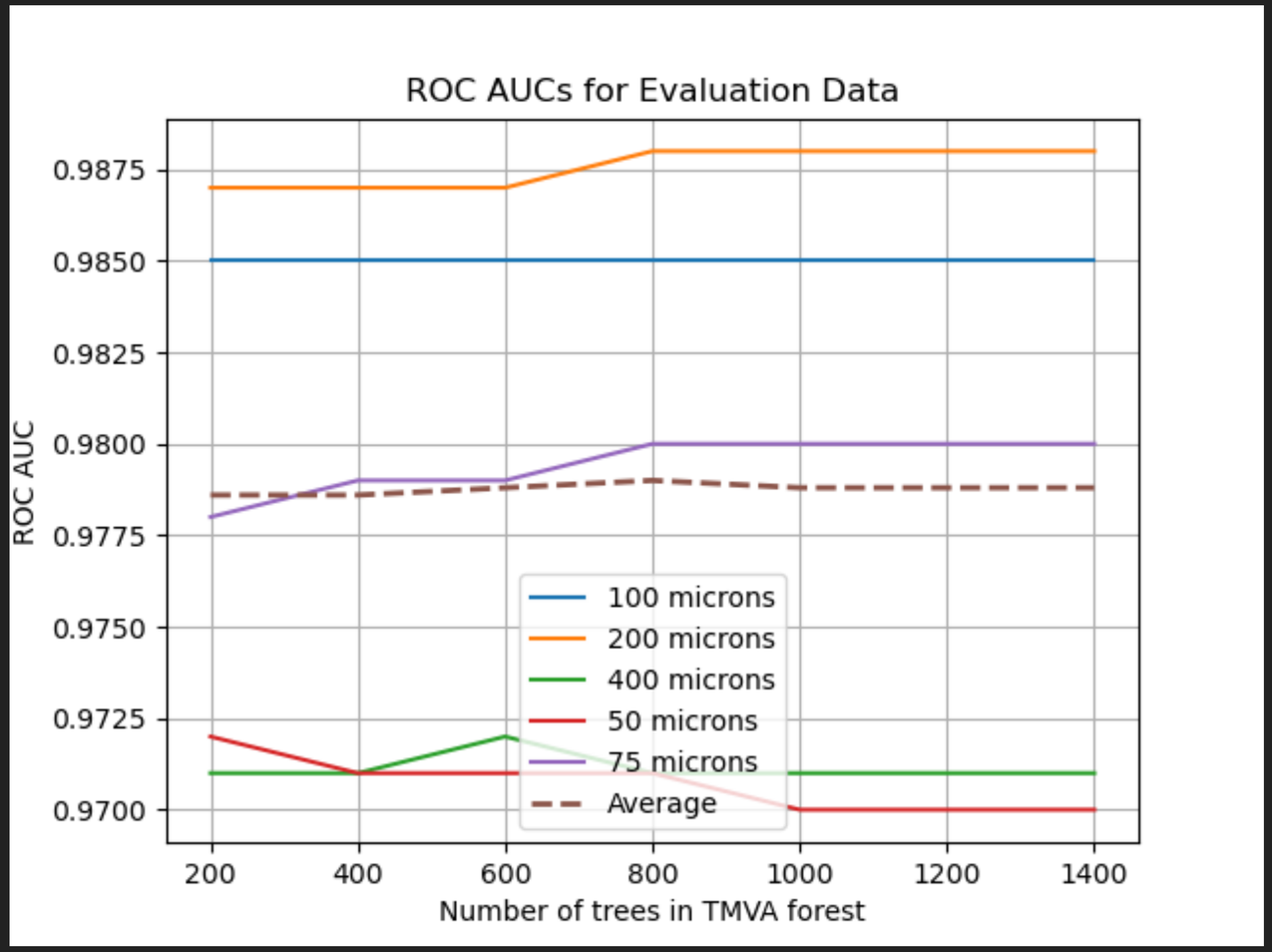
Smallest number of nodes within stable range

TMVA HYPERPARAMETERS

3% min node size, 50/50 train/test split, 0.5 adaptive boost beta, max depth of 8

NUMBER OF TREES

How many trees can be in our forest?



OPTIMAL: 800

Accidentally did this one first, so these results are biased.

VARIABLE RANKS

200 trees, 5% min node size, 50/50 train/test split, 0.5 adaptive boost beta, max depth of 3 (These are the parameters we've been using to date.)

50 µm

Rank	Variable	Variable Importance
1	Cluster_ArrivalTime	7.814e-02
2	Cluster_z	7.229e-02
3	Incident_Angle	5.577e-02
4	Cluster_Size_x	5.557e-02
5	PixelHits_EnergyDeposited_2	5.425e-02
6	Cluster_Size_y	5.143e-02
7	PixelHits_ArrivalTime_3	4.709e-02
8	PixelHits_ArrivalTime_1	4.289e-02
9	PixelHits_ArrivalTime_2	4.262e-02
10	Cluster_RMS_x	3.927e-02
11	Cluster_x	3.745e-02
12	PixelHits_EnergyDeposited_4	3.731e-02
13	PixelHits_ArrivalTime_0	3.477e-02
14	Cluster_y	3.368e-02
15	Cluster_RMS_y	3.331e-02
16	PixelHits_EnergyDeposited_0	3.266e-02
17	PixelHits_ArrivalTime_5	3.023e-02
18	PixelHits_ArrivalTime_4	2.978e-02
19	PixelHits_EnergyDeposited_7	2.711e-02
20	PixelHits_EnergyDeposited_3	2.504e-02
21	Cluster_EnergyDeposited	2.324e-02
22	PixelHits_EnergyDeposited_1	2.245e-02
23	PixelHits_ArrivalTime_6	2.021e-02
24	Cluster_Skew_x	1.993e-02
25	PixelHits_EnergyDeposited_5	1.616e-02
26	PixelHits_EnergyDeposited_6	1.294e-02
27	Cluster_AspectRatio	1.052e-02
28	Cluster_Size_tot	7.182e-03
29	Cluster_Skew_y	6.687e-03
30	Cluster_Eccentricity	0.000e+00
31	PixelHits_EnergyDeposited_8	0.000e+00
32	PixelHits_ArrivalTime_7	0.000e+00
33	PixelHits_ArrivalTime_8	0.000e+00

75 µm

Rank	Variable	Variable Importance
1	Cluster_ArrivalTime	8.166e-02
2	Cluster_z	7.138e-02
3	Cluster_Size_y	5.725e-02
4	Incident_Angle	5.412e-02
5	PixelHits_ArrivalTime_4	4.639e-02
6	Cluster_Size_x	4.517e-02
7	PixelHits_ArrivalTime_3	4.183e-02
8	Cluster_x	4.107e-02
9	PixelHits_ArrivalTime_6	4.052e-02
10	Cluster_y	3.800e-02
11	PixelHits_ArrivalTime_2	3.754e-02
12	PixelHits_ArrivalTime_0	3.731e-02
13	PixelHits_ArrivalTime_5	3.720e-02
14	PixelHits_EnergyDeposited_2	3.330e-02
15	Cluster_RMS_x	2.927e-02
16	Cluster_Skew_x	2.775e-02
17	Cluster_RMS_y	2.754e-02
18	PixelHits_EnergyDeposited_7	2.571e-02
19	PixelHits_ArrivalTime_1	2.566e-02
20	PixelHits_EnergyDeposited_0	2.516e-02
21	PixelHits_EnergyDeposited_5	2.460e-02
22	PixelHits_ArrivalTime_8	2.446e-02
23	PixelHits_ArrivalTime_7	2.039e-02
24	PixelHits_EnergyDeposited_8	1.775e-02
25	Cluster_EnergyDeposited	1.543e-02
26	PixelHits_EnergyDeposited_3	1.528e-02
27	PixelHits_EnergyDeposited_6	1.450e-02
28	Cluster_Skew_y	1.360e-02
29	PixelHits_EnergyDeposited_1	1.194e-02
30	Cluster_AspectRatio	6.808e-03
31	Cluster_Size_tot	6.567e-03
32	PixelHits_EnergyDeposited_4	4.842e-03
33	Cluster_Eccentricity	0.000e+00

100 µm

Rank	Variable	Variable Importance
1	Cluster_ArrivalTime	8.157e-02
2	Cluster_Size_x	6.670e-02
3	Cluster_z	6.505e-02
4	Incident_Angle	5.303e-02
5	Cluster_Size_y	4.908e-02
6	PixelHits_ArrivalTime_6	4.846e-02
7	PixelHits_ArrivalTime_7	4.235e-02
8	Cluster_y	3.543e-02
9	PixelHits_ArrivalTime_4	3.521e-02
10	PixelHits_EnergyDeposited_4	3.295e-02
11	PixelHits_ArrivalTime_5	3.281e-02
12	PixelHits_ArrivalTime_1	3.265e-02
13	Cluster_x	3.114e-02
14	PixelHits_EnergyDeposited_6	2.911e-02
15	PixelHits_EnergyDeposited_7	2.780e-02
16	PixelHits_ArrivalTime_3	2.698e-02
17	PixelHits_EnergyDeposited_3	2.693e-02
18	PixelHits_EnergyDeposited_0	2.656e-02
19	PixelHits_ArrivalTime_0	2.630e-02
20	Cluster_RMS_x	2.579e-02
21	PixelHits_ArrivalTime_2	2.530e-02
22	PixelHits_ArrivalTime_8	2.512e-02
23	PixelHits_EnergyDeposited_5	2.437e-02
24	Cluster_RMS_y	2.416e-02
25	Cluster_Size_tot	1.852e-02
26	PixelHits_EnergyDeposited_1	1.693e-02
27	PixelHits_EnergyDeposited_8	1.628e-02
28	Cluster_Skew_x	1.323e-02
29	Cluster_EnergyDeposited	1.080e-02
30	PixelHits_EnergyDeposited_2	1.041e-02
31	Cluster_Skew_y	1.006e-02
32	Cluster_AspectRatio	8.949e-03
33	Cluster_Eccentricity	0.000e+00

200 µm

Rank	Variable	Variable Importance
1	Cluster_ArrivalTime	1.147e-01
2	Cluster_Size_y	5.791e-02
3	PixelHits_EnergyDeposited_8	5.548e-02
4	PixelHits_ArrivalTime_8	5.393e-02
5	Cluster_Size_x	5.269e-02
6	Cluster_AspectRatio	4.988e-02
7	Cluster_EnergyDeposited	4.112e-02
8	Cluster_Size_tot	3.762e-02
9	PixelHits_ArrivalTime_4	3.740e-02
10	Cluster_x	3.568e-02
11	Cluster_z	3.547e-02
12	PixelHits_EnergyDeposited_7	3.437e-02
13	PixelHits_ArrivalTime_7	3.250e-02
14	PixelHits_ArrivalTime_6	3.238e-02
15	PixelHits_ArrivalTime_0	3.150e-02
16	Incident_Angle	3.114e-02
17	PixelHits_ArrivalTime_2	2.959e-02
18	Cluster_y	2.781e-02
19	Cluster_RMS_x	2.662e-02
20	PixelHits_ArrivalTime_3	2.549e-02
21	PixelHits_ArrivalTime_5	2.491e-02
22	Cluster_RMS_y	2.075e-02
23	PixelHits_EnergyDeposited_6	1.704e-02
24	Cluster_Skew_x	1.521e-02
25	PixelHits_ArrivalTime_1	1.494e-02
26	PixelHits_EnergyDeposited_3	1.255e-02
27	PixelHits_EnergyDeposited_1	1.201e-02
28	Cluster_Skew_y	1.081e-02
29	PixelHits_EnergyDeposited_2	1.051e-02
30	PixelHits_EnergyDeposited_0	8.543e-03
31	PixelHits_EnergyDeposited_5	8.361e-03
32	PixelHits_EnergyDeposited_4	1.116e-03
33	Cluster_Eccentricity	0.000e+00

400 µm

Rank	Variable	Variable Importance
1	Cluster_ArrivalTime	1.247e-01
2	Cluster_Size_y	6.118e-02
3	Cluster_AspectRatio	5.154e-02
4	Cluster_EnergyDeposited	4.696e-02
5	Cluster_z	4.487e-02
6	Cluster_x	3.785e-02
7	Incident_Angle	3.747e-02
8	Cluster_Size_x	3.594e-02
9	PixelHits_ArrivalTime_7	3.583e-02
10	Cluster_y	3.564e-02
11	Cluster_RMS_x	3.488e-02
12	Cluster_Skew_y	3.445e-02
13	PixelHits_ArrivalTime_5	3.251e-02
14	PixelHits_ArrivalTime_4	3.228e-02
15	PixelHits_ArrivalTime_1	3.099e-02
16	PixelHits_ArrivalTime_2	2.960e-02
17	Cluster_RMS_y	2.911e-02
18	PixelHits_ArrivalTime_3	2.897e-02
19	PixelHits_ArrivalTime_0	2.725e-02
20	PixelHits_EnergyDeposited_6	2.673e-02
21	PixelHits_ArrivalTime_8	2.539e-02
22	PixelHits_ArrivalTime_6	2.467e-02
23	PixelHits_EnergyDeposited_2	2.000e-02
24	PixelHits_EnergyDeposited_0	1.851e-02
25	PixelHits_EnergyDeposited_8	1.754e-02
26	PixelHits_EnergyDeposited_4	1.742e-02
27	PixelHits_EnergyDeposited_1	1.626e-02
28	Cluster_Size_tot	1.373e-02
29	Cluster_Skew_x	1.154e-02
30	PixelHits_EnergyDeposited_3	8.733e-03
31	PixelHits_EnergyDeposited_5	7.446e-03
32	Cluster_Eccentricity	0.000e+00
33	PixelHits_EnergyDeposited_7	0.000e+00

Top ranked variables across range

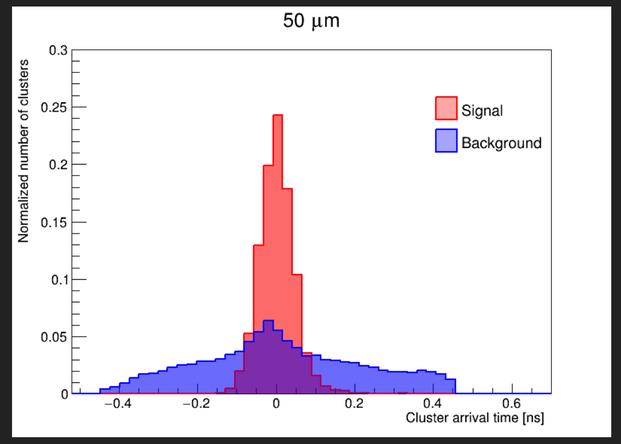
1. Cluster_ArrivalTime
2. Cluster_Size_y
3. Cluster_z
4. Cluster_Size_x
5. PixelHits_ArrivalTime*
6. Incident_Angle
7. PixelHits_EnergyDeposited*

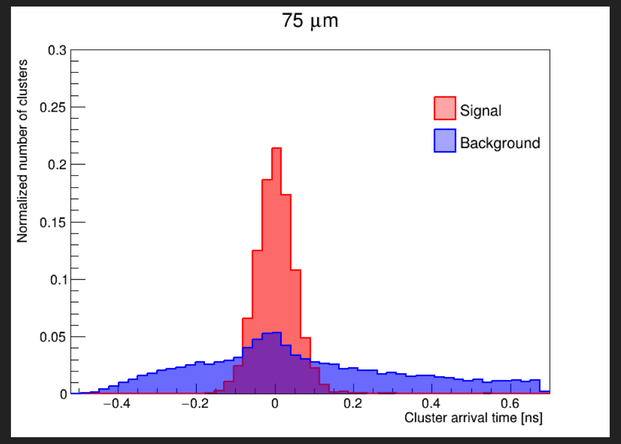
*Calculated by taking the average ranking of the top ranked variable, regardless of pixel number

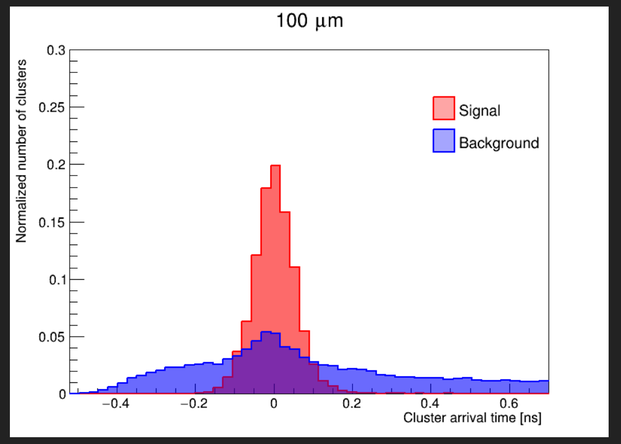
SELECT VARIABLE DISTRIBUTIONS

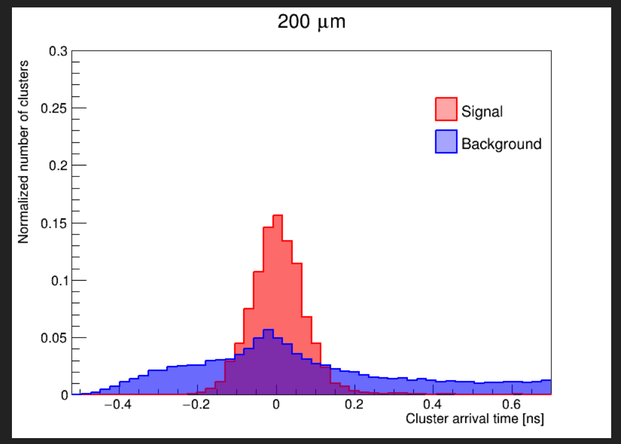
800 trees, 3% min node size, 50/50 train/test split, 0.5 adaptive boost beta, max depth of 8

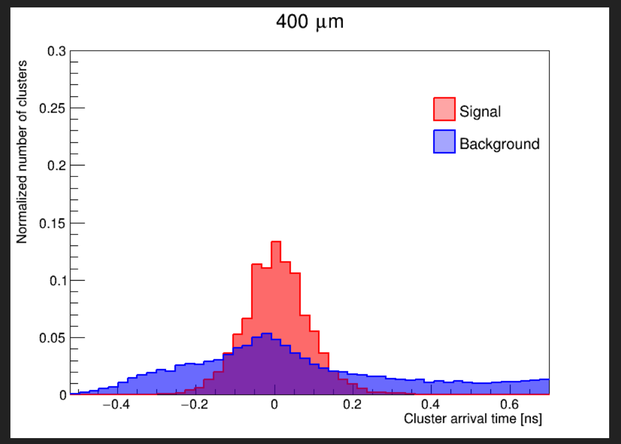
CLUSTER ARRIVAL TIME











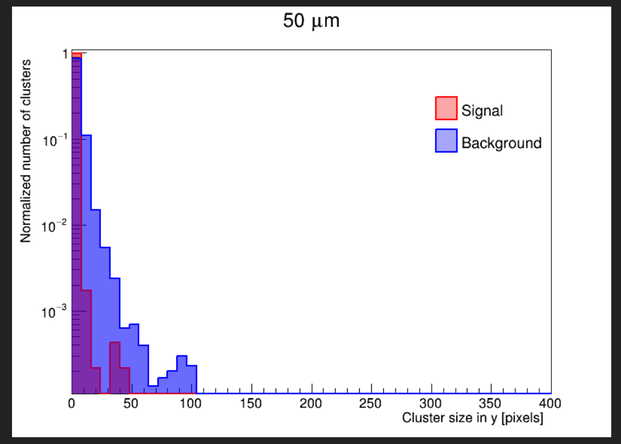
TOP-RANKED VARIABLE ACROSS ALL THICKNESSES

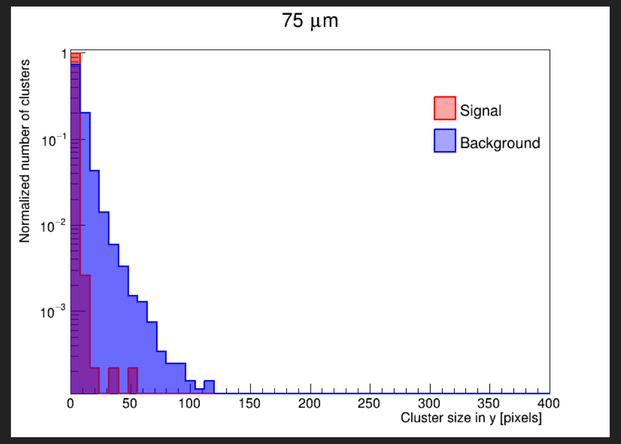
SELECT VARIABLE DISTRIBUTIONS

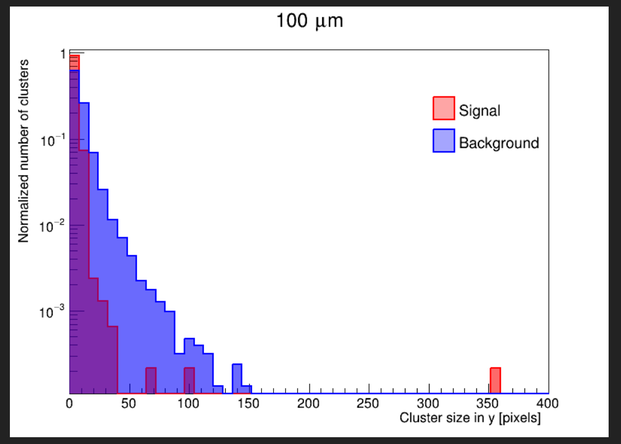
200 trees, 5% min node size, 50/50 train/test split, 0.5 adaptive boost beta, max depth of 3

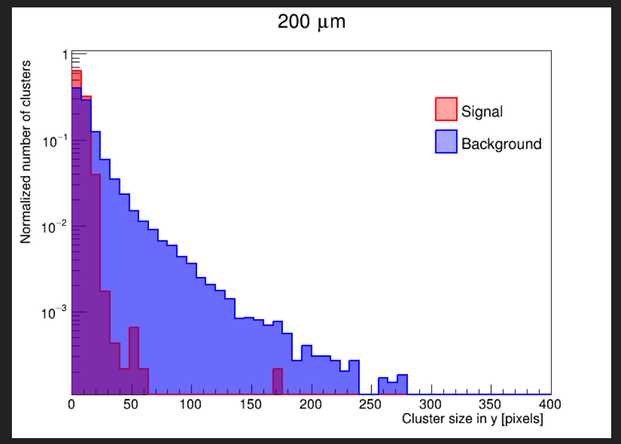
(These are the parameters we've been using to date.)

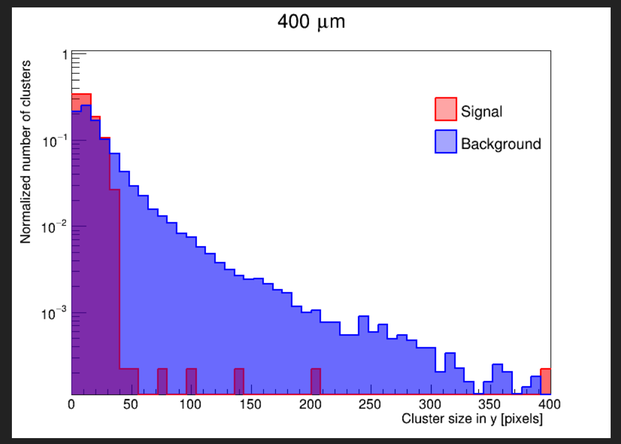
CLUSTER SIZE IN Y







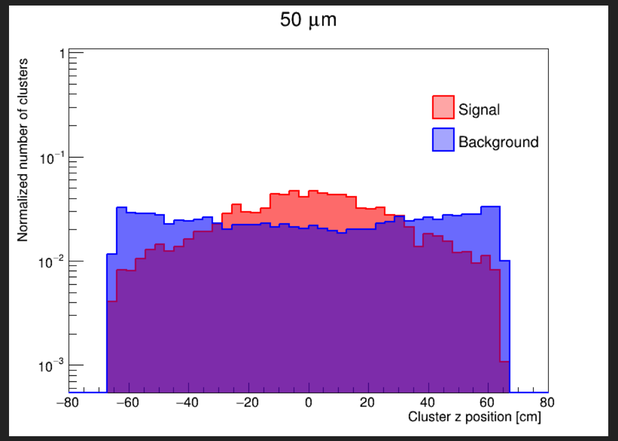


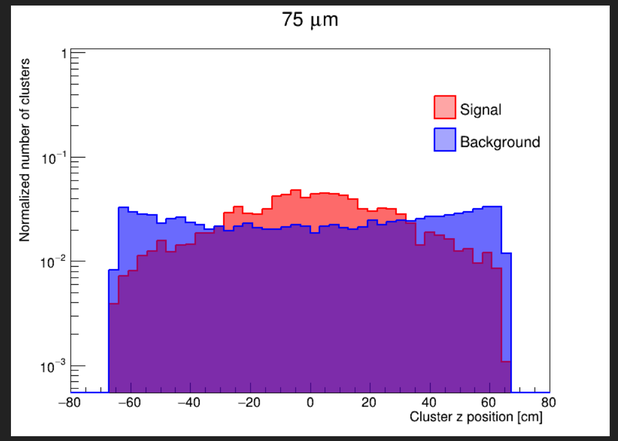


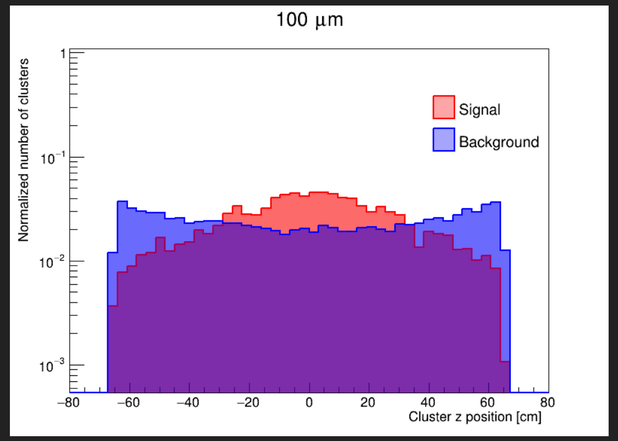
SELECT VARIABLE DISTRIBUTIONS

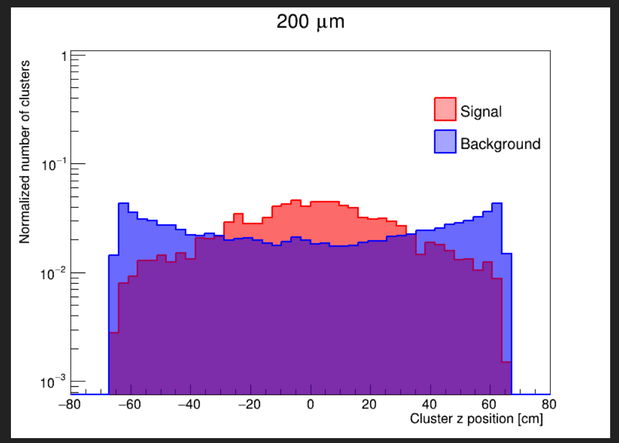
800 trees, 3% min node size, 50/50 train/test split, 0.5 adaptive boost beta, max depth of 8

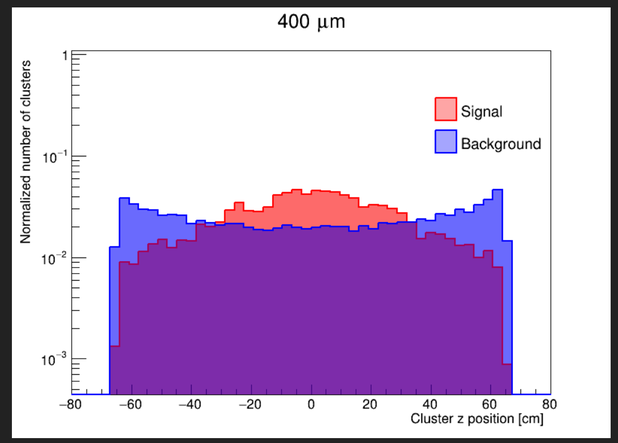
CLUSTER Z POSITION









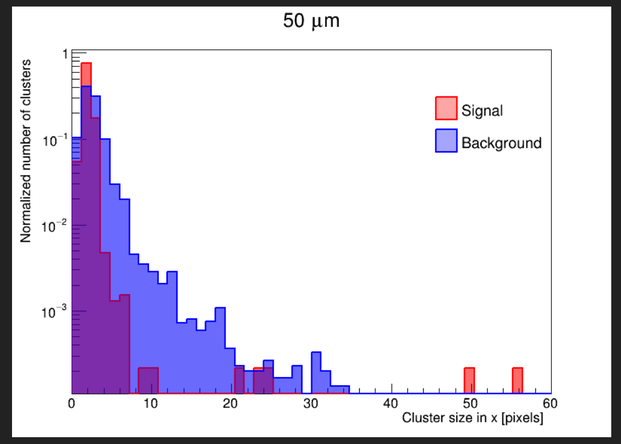


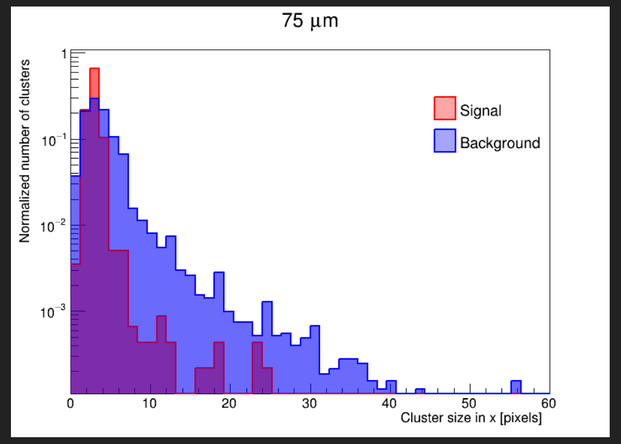
SELECT VARIABLE DISTRIBUTIONS

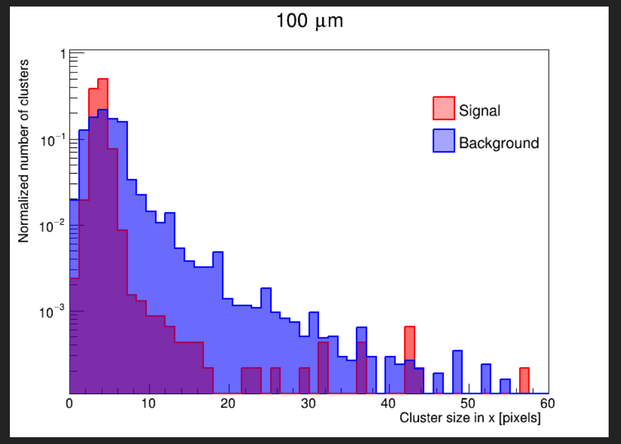
200 trees, 5% min node size, 50/50 train/test split, 0.5 adaptive boost beta, max depth of 3

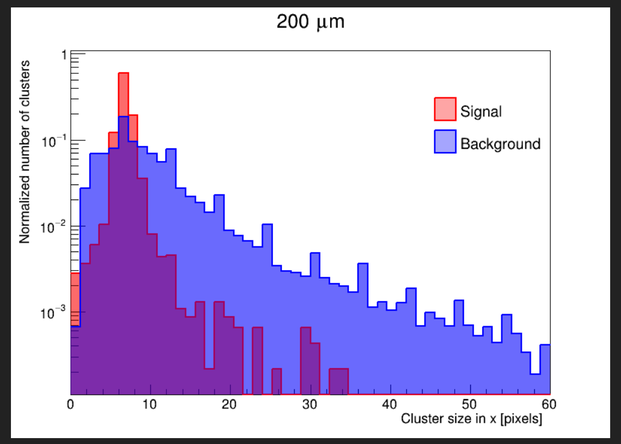
(These are the parameters we've been using to date.)

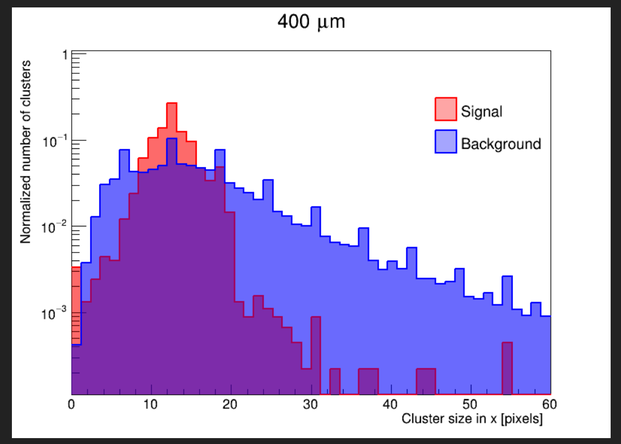
CLUSTER SIZE IN X









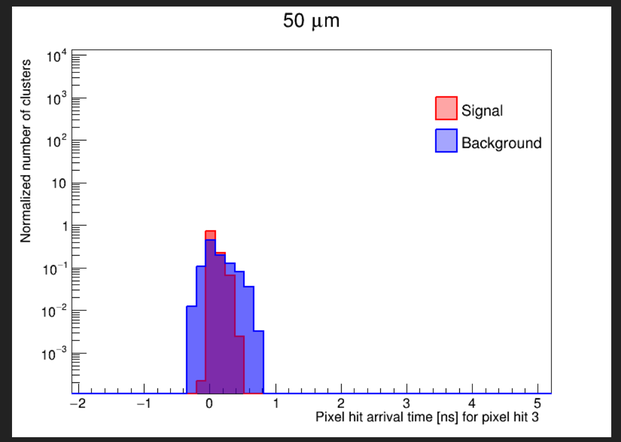


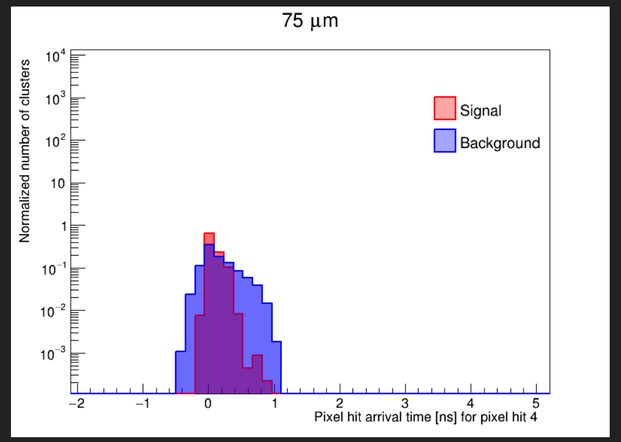
SELECT VARIABLE DISTRIBUTIONS

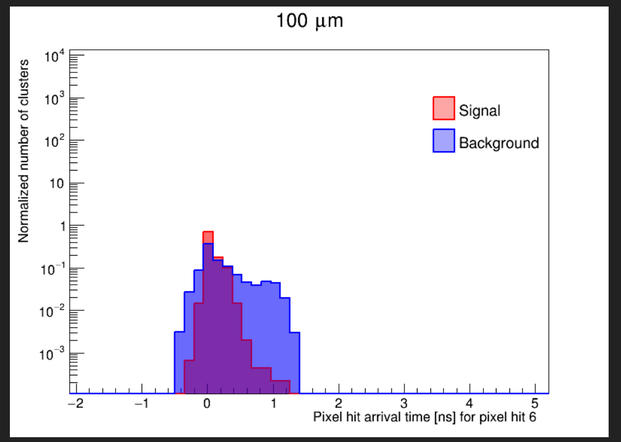
200 trees, 5% min node size, 50/50 train/test split, 0.5 adaptive boost beta, max depth of 3

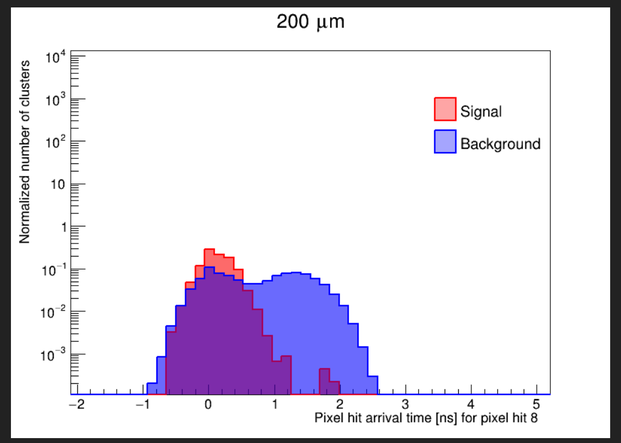
(These are the parameters we've been using to date.)

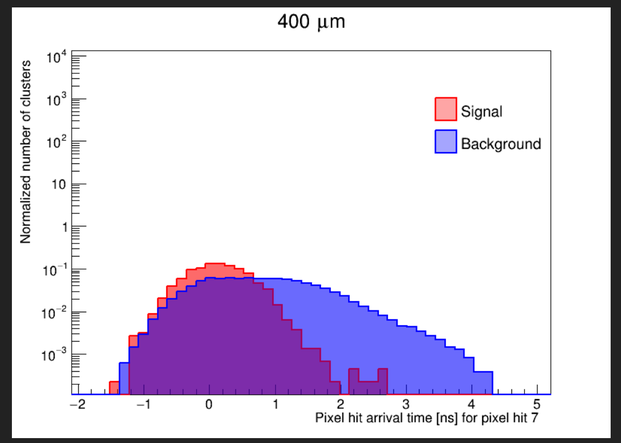
PIXEL HIT ARRIVAL TIME







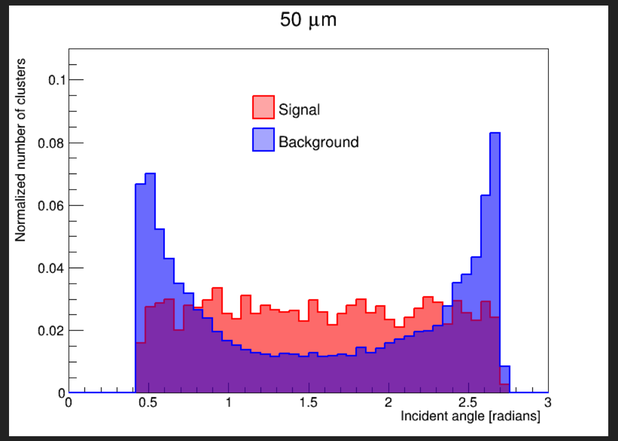


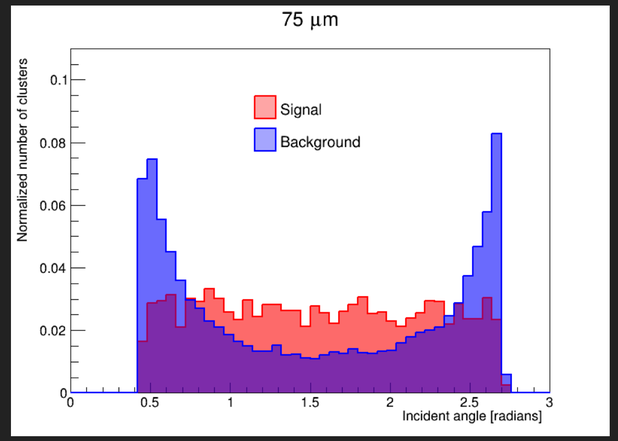


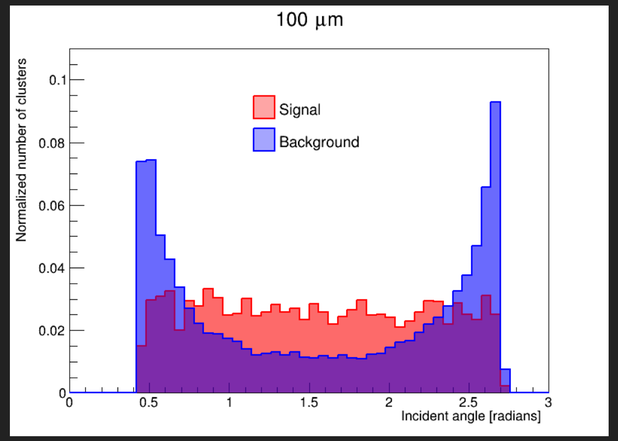
SELECT VARIABLE DISTRIBUTIONS

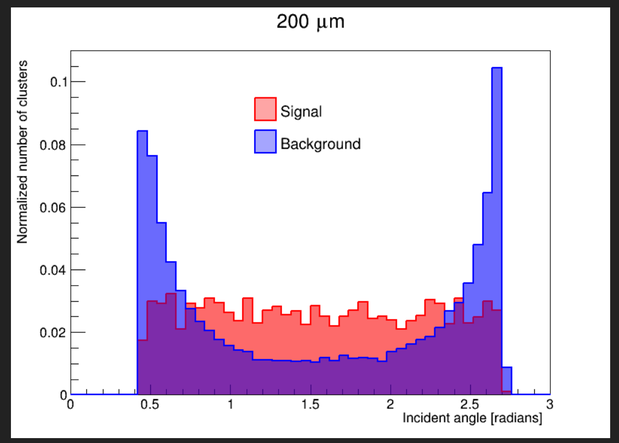
800 trees, 3% min node size, 50/50 train/test split, 0.5 adaptive boost beta, max depth of 8

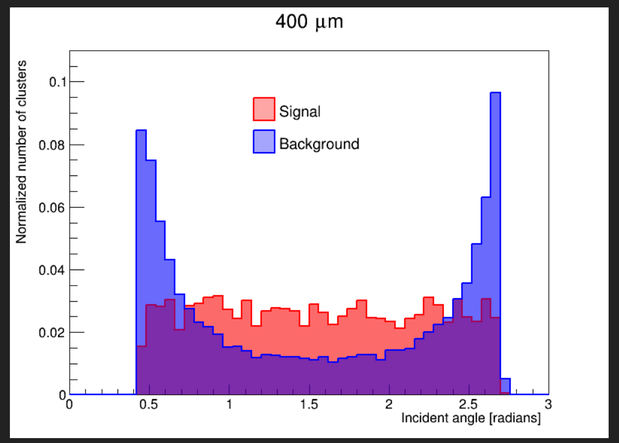
INCIDENT ANGLE









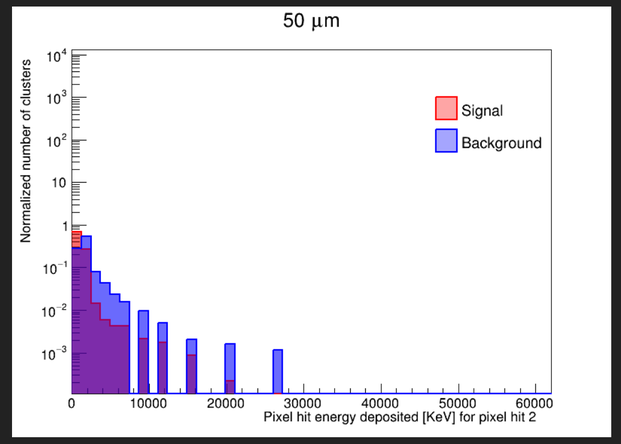


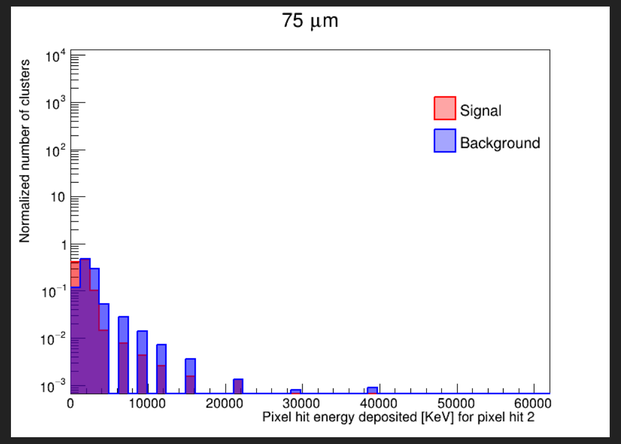
SELECT VARIABLE DISTRIBUTIONS

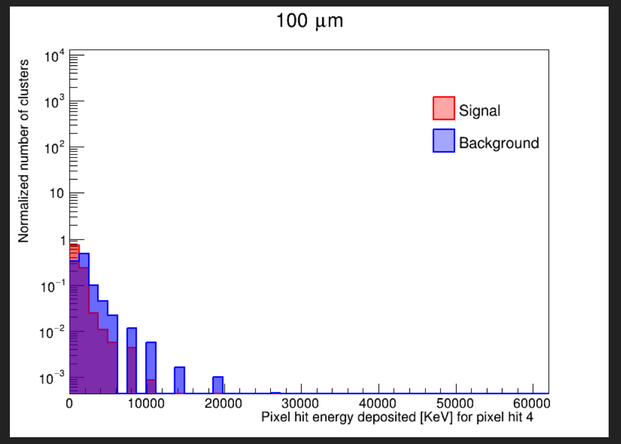
200 trees, 5% min node size, 50/50 train/test split, 0.5 adaptive boost beta, max depth of 3

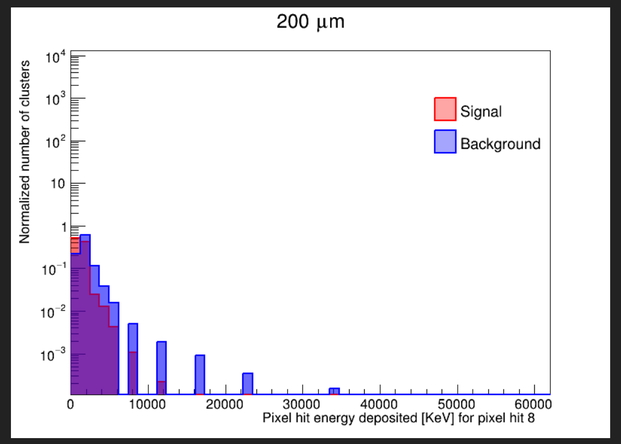
(These are the parameters we've been using to date.)

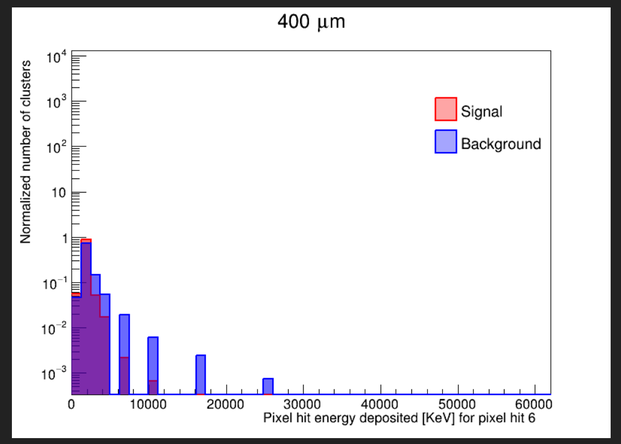
PIXEL HIT ENERGY DEPOSIT











ADAPTIVE BOOSTING

From Nicolas Chanon lecture:

$$y_{\text{Boost}}(x) = \frac{1}{N_{\text{collection}}} \cdot \sum_{i}^{N_{\text{collection}}} \ln(a_i) \cdot h_i(x)$$

$$h_i = \pm 1$$

$$\alpha_m = \beta \times \ln\left(\frac{1 - \text{err}_m}{\text{err}_m}\right)$$

$$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq T_m x_i)}{\sum_{i=1}^{N} w_i}$$

$$w_i \rightarrow w_i \times e^{\alpha_m I(y_i \neq T_m(x_i))}$$

$$I = \begin{cases} 1 & \text{misclassified event} \\ 0 & \text{otherwise} \end{cases}$$