Optimizing of STAR sTGC track finding using boosted decision trees

DNP meeting 2021

Youqi Song (Yale) for the STAR Collaboration

10/14/2021

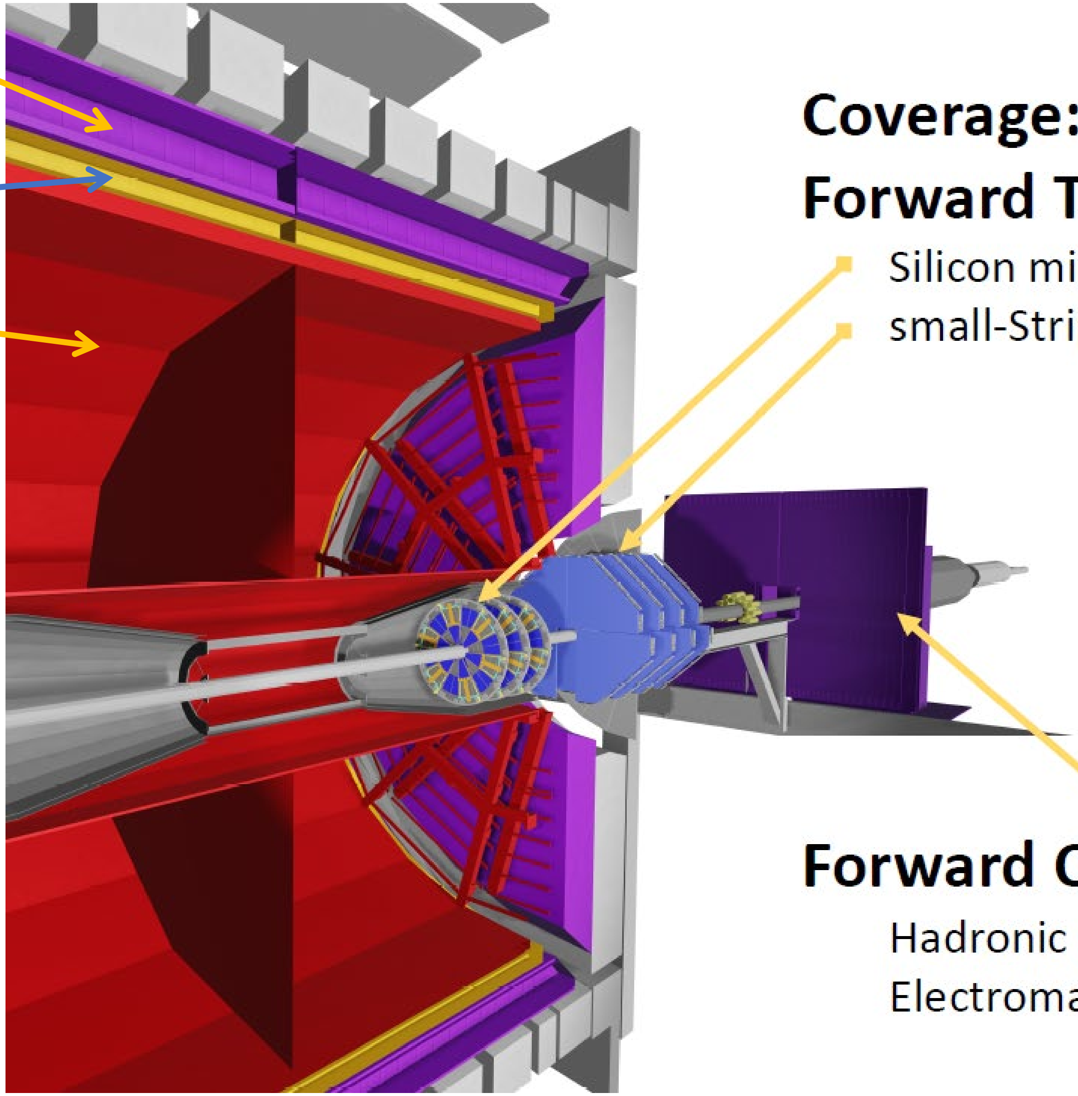
Supported in part by U.S. DEPARTMENT OF ENERGY | Office of Science

STAR forward upgrade: forward detector

Barrel electromagnetic calorimeter

Time of flight

Time projection chamber (TPC)
(2m of outer radius)



Coverage: $2.5 < \eta < 4.0$

Forward Tracking System

- Silicon microstrip sensors
- small-Strip Thin Gap Chambers (**sTGC**)

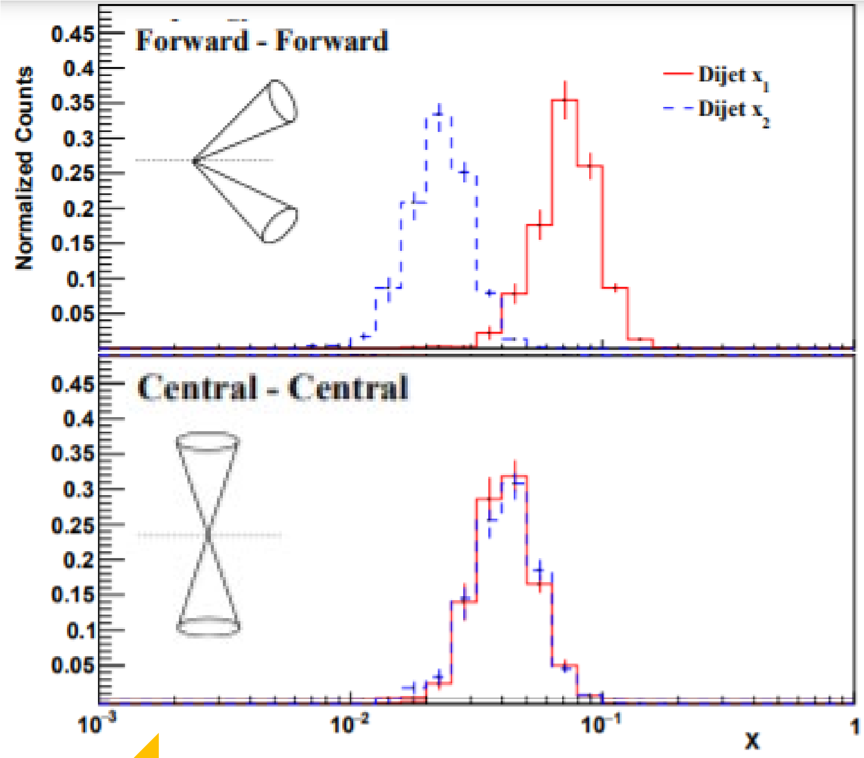
Forward Calorimetry System

- Hadronic Calorimeter
- Electromagnetic Calorimeter

STAR forward upgrade: physics goals

https://arxiv.org/pdf/1906.02740.pdf

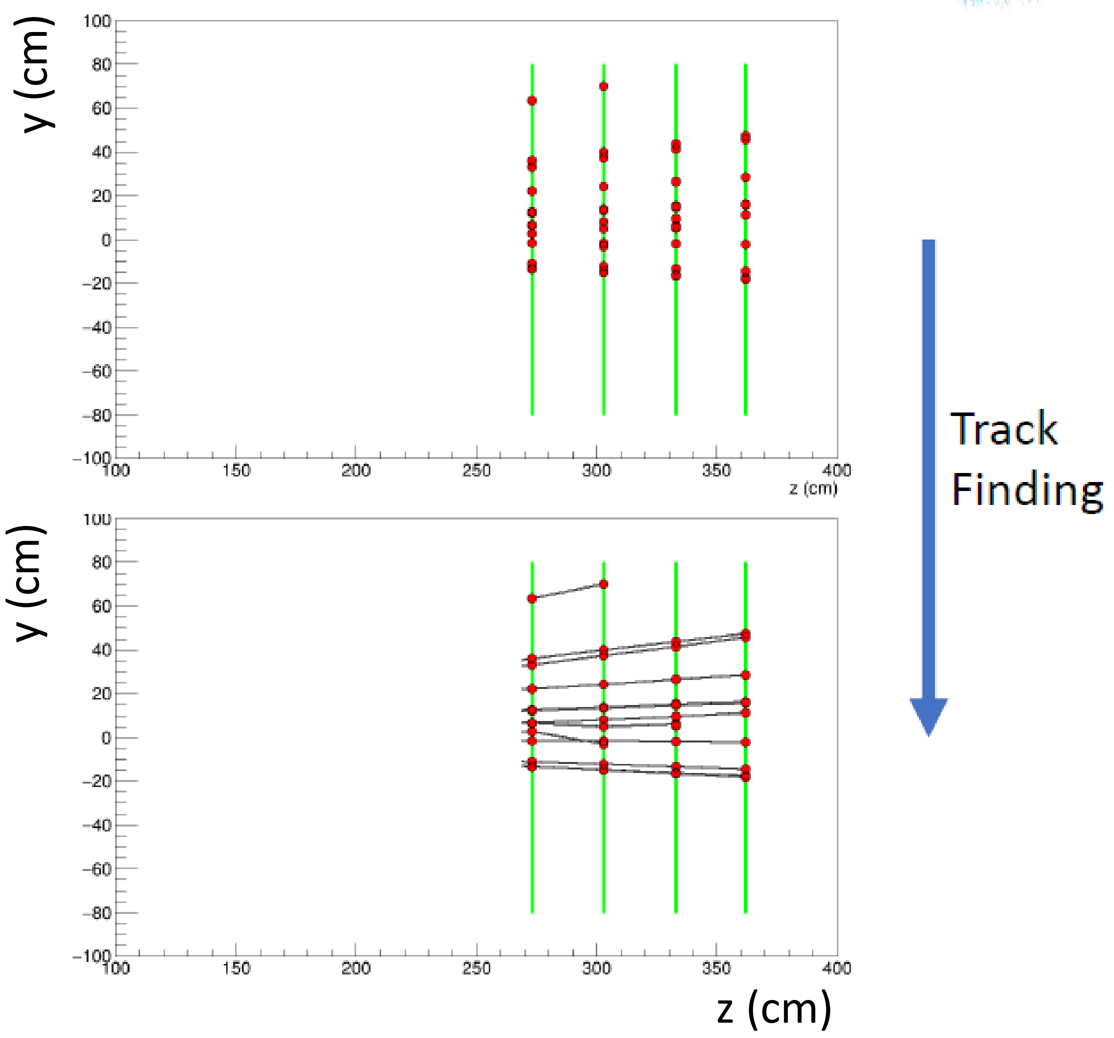
- What are the nuclear parton distribution functions (nPDF) at low-x?
  - Need to go to forward rapidity to access the low-x region
- What is the spatial transverse distributions of nucleons and gluons?
- How saturated is the initial state of the nucleus?



https://www.bnl.gov/eic/

Forward track finding

- Initial goal: find tracks from hits on sTGC.
- 1. Generate events (PYTHIA 8 pp Drell-Yan events at $\sqrt{s}$ = 510 GeV, 50,000 events) and obtain hits.
- 2. Consider hit pairs. Feed them into a model that removes fake pairs.
- 3. Consider hit pairs that passed the first model. Make them into hit triplets. Feed into a second model which removes fake triplets.
- 4. Run track fitting to reconstruct tracks from the hit triplets that pass the selection.



Hit observables

- Hit pair (crit2) observables, calculated for each pair of hits that are:
  - On adjacent tracker layers (1-2, 2-3 or 3-4)
- Similar for hit triplet (crit3) observables

Crit2 delta phi

$$\Delta\phi = \phi_A - \phi_B$$

Crit2 delta rho

$$\rho = \sqrt{x^2 + y^2}$$

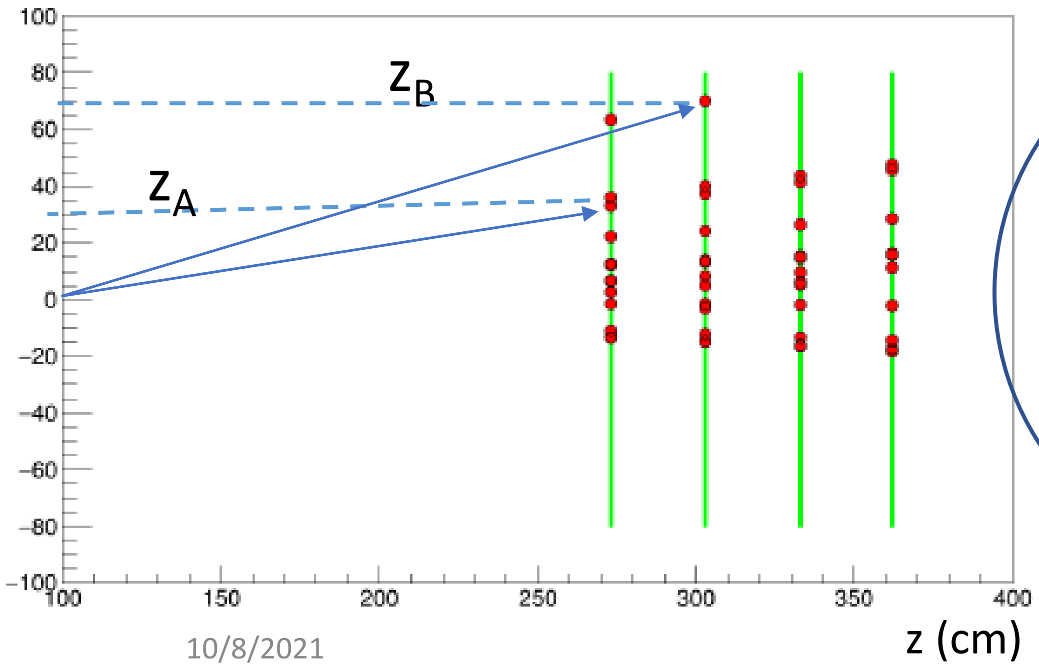
$$\Delta\rho = \rho_A - \rho_B$$

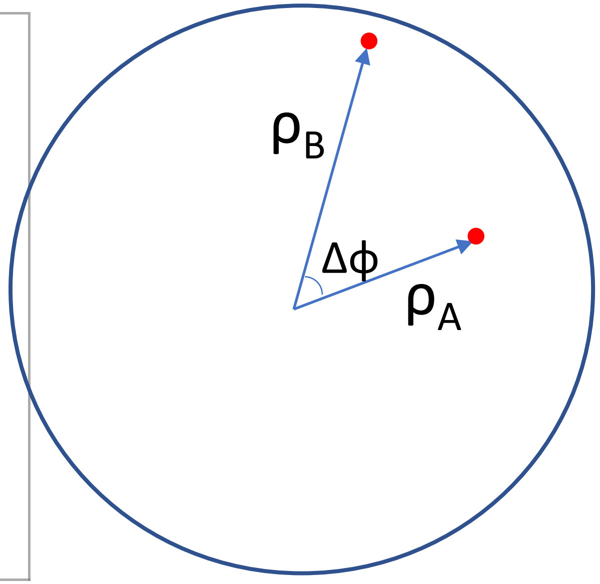
Crit2 RZ ratio

$$\left(\frac{\Delta R}{\Delta z}\right)^2 = \frac{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2}{(\Delta z)^2}$$

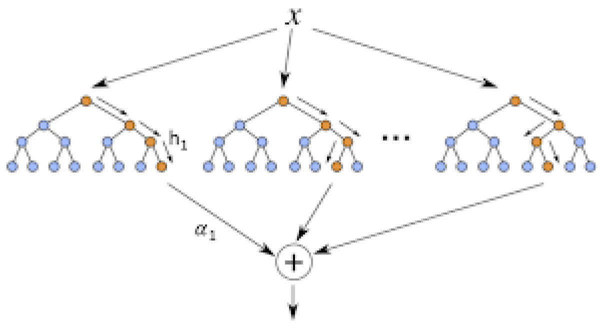
Crit2 straight track ratio

$$\frac{\rho_A * z_B}{\rho_B * z_A}$$





Model construction



- Construct boosted decision trees (800 trees, 3 layers) that filter away the fake hit pairs and triplets that are used in steps 2 and 3.
- Use a machine learning method: adaptive boosting "AdaBoost", implemented with scikit-learn (Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.)

  - 1. Generate events (PYTHIA 8 pp Drell-Yan events at $\sqrt{s}$ = 510 GeV, 50,000 events) and obtain hits.
  - 2. Consider hit pairs. Feed them into a model that removes fake pairs based on hit pair observable values.
  - 3. Consider hit pairs that passed the first model and make them into hit triplets. Feed them to a second model which removes fake triplets based on hit triplet observable values.

Standard approach vs machine learning

- Standard cut-based approach
  - 0.999 < Crit2_RZRatio < 1.131
  - -6.1 < Crit2_DeltaRho < 9.9
  - 0 < Crit2_DeltaPhi < 13.4
  - 0.81 < Crit2_StraightTrackRatio < 1.35
- Machine learning approach

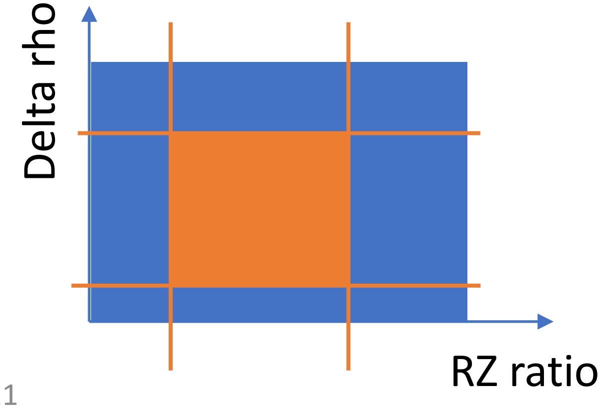
```
|--- Crit2_DeltaPhi <= 0.80
|   |--- Crit2_RZRatio <= 1.02
|   |   |--- Crit2_StraightTrackRatio <= 0.98
|   |   |   |--- class: 0.0
|   |   |--- Crit2_StraightTrackRatio >  0.98
|   |   |   |--- class: 1.0
|   |--- Crit2_RZRatio >  1.02
|   |   |--- Crit2_RZRatio <= 1.11
|   |   |   |--- class: 0.0
|   |   |--- Crit2_RZRatio >  1.11
|   |   |   |--- class: 0.0
|--- Crit2_DeltaPhi >  0.80
|   |--- Crit2_RZRatio <= 1.03
|   |   |--- Crit2_DeltaPhi <= 3.02
|   |   |   |--- class: 0.0
|   |   |--- Crit2_DeltaPhi >  3.02
|   |   |   |--- class: 0.0
|   |--- Crit2_RZRatio >  1.03
|   |   |--- Crit2_RZRatio <= 1.11
|   |   |   |--- class: 0.0
|   |   |--- Crit2_RZRatio >  1.11
|   |   |   |--- class: 0.0
```

800 of these, each with different weight

Standard approach vs machine learning

- Standard cut-based approach
  - 0.999 < Crit2_RZRatio < 1.131
  - -6.1 < Crit2_DeltaRho < 9.9
  - 0 < Crit2_DeltaPhi < 13.4
  - 0.81 < Crit2_StraightTrackRatio < 1.35

Checks the value of each observable SEPARATELY

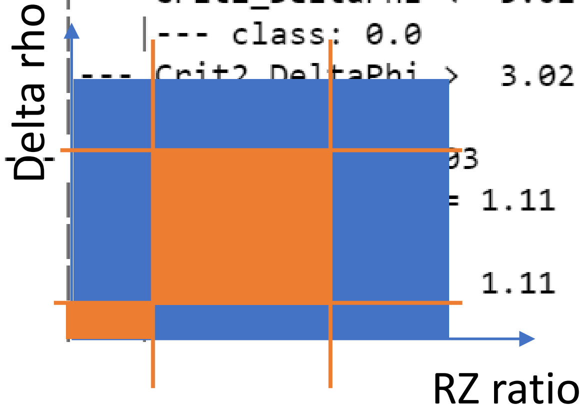


- Machine learning approach

```
|--- Crit2_DeltaPhi <= 0.80
|   |--- Crit2_RZRatio <= 1.02
|   |   |--- Crit2_StraightTrackRatio <= 0.98
|   |   |   |--- class: 0.0
|   |   |--- Crit2_StraightTrackRatio >  0.98
|   |   |   |--- class: 1.0
|   |--- Crit2_RZRatio >  1.02
|   |   |--- Crit2_RZRatio <= 1.11
|   |   |   |--- class: 0.0
```

Assigns a score to each hit combination based on ALL of its observable values

```
|   |   |--- Crit2_DeltaPhi <= 3.02
|   |   |   |--- class: 0.0
|   |   |--- Crit2_DeltaPhi >  3.02
```



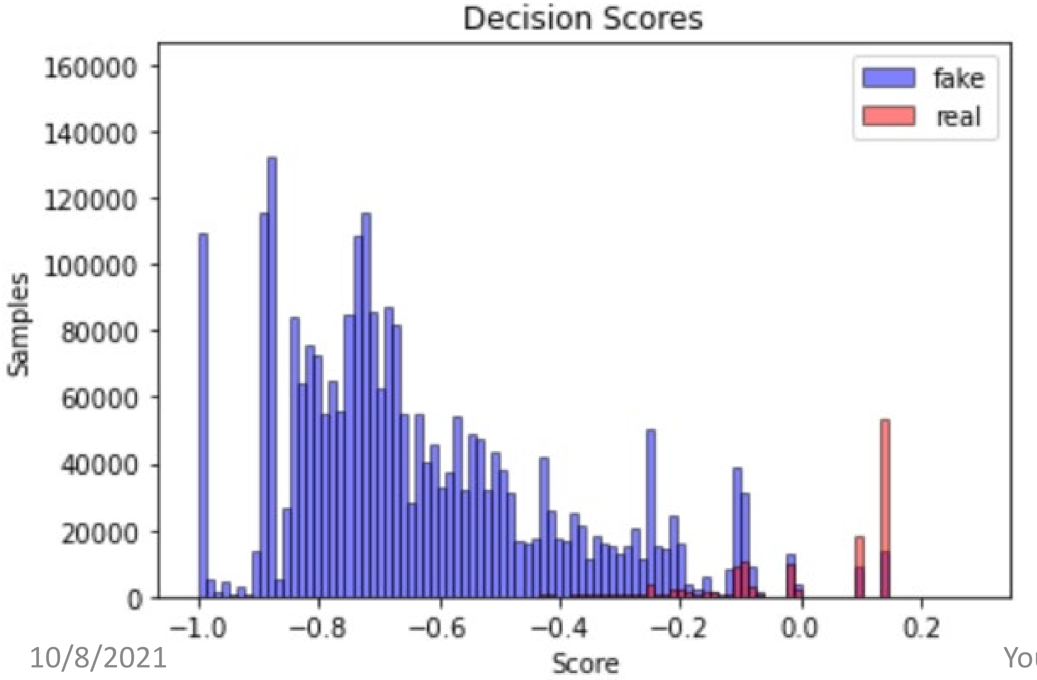
Testing for hit pairs

	precision	recall
fake	0.98	0.99
real	purity=0.76	eff=0.56

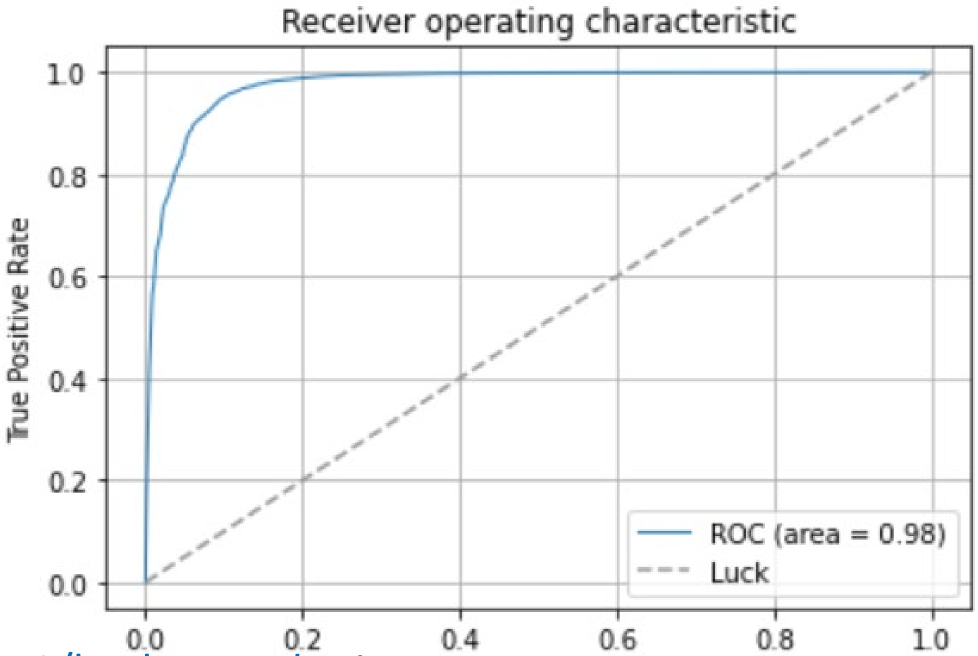
background rej = 1 / fpr
fpr = 1 – fake recall
→ fake recall = 1 – 1 / rej
→ rej = 1 / (1 – fake recall)

$$\text{purity} = \frac{\text{\# real hit pairs identified as real}}{\text{\# all hit pairs identified as real}}$$

$$\text{eff} = \frac{\text{\# real hit pairs identified as real}}{\text{\# real hit pairs}}$$



efficiency=



1/background rej =

Testing for hit triplets

	precision	recall
fake	0.85	0.74
real	purity=0.85	eff= 0.92

purity = $\frac{\text{\# real hit triplets identified as real}}{\text{\# all hit triplets identified as real}}$

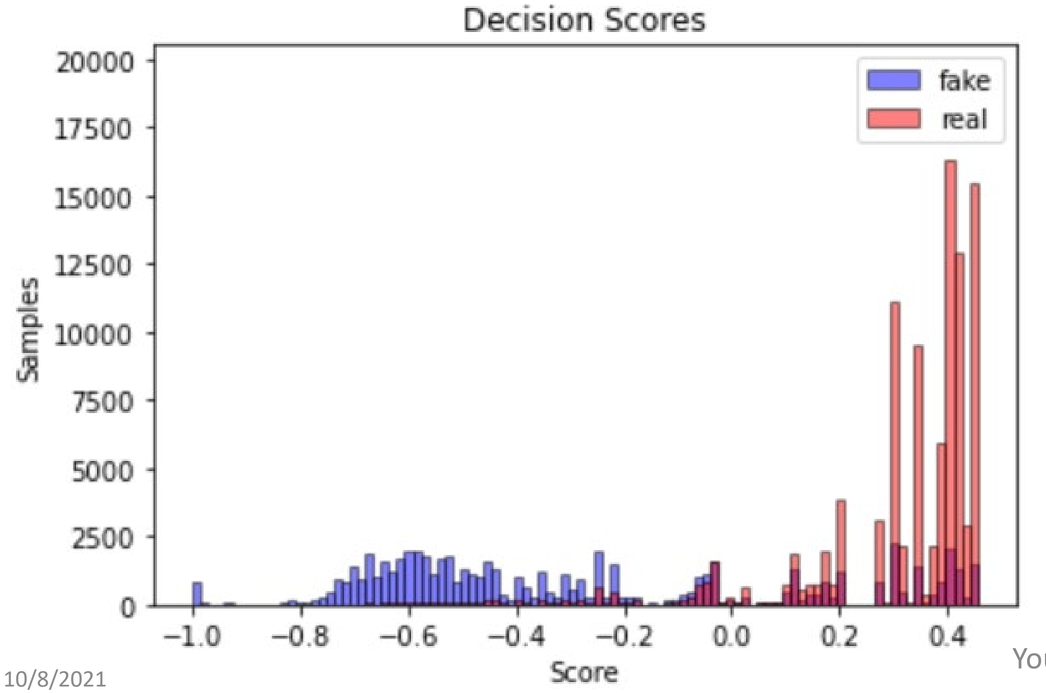
eff = $\frac{\text{\# real hit triplets identified as real}}{\text{\# real hit triplets}}$

background rej = 1 / fpr

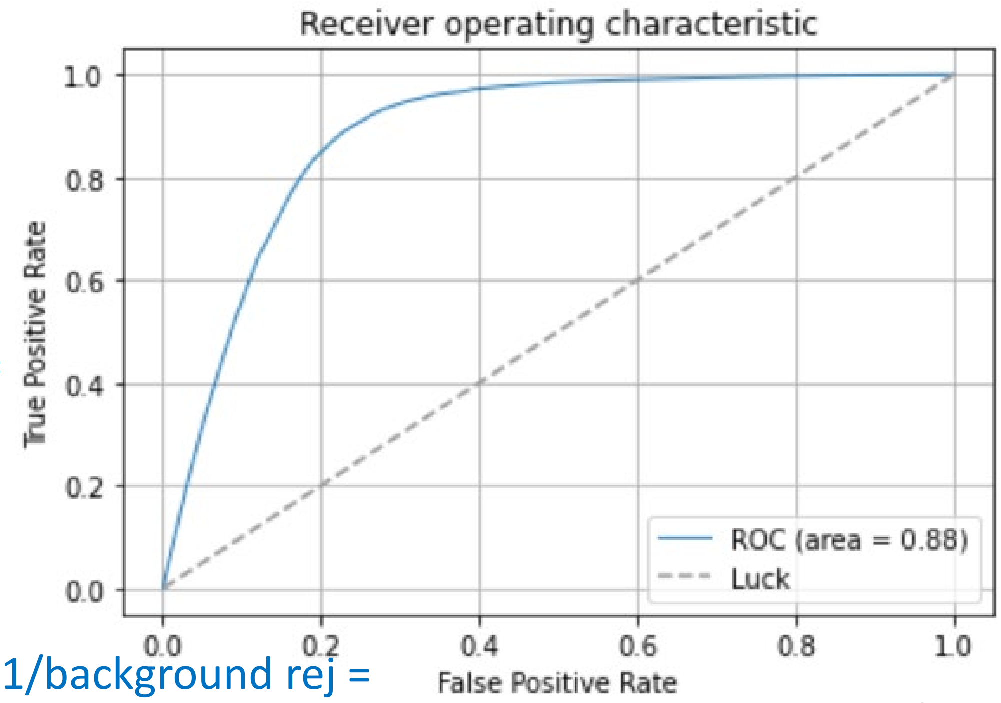
fpr = 1 – fake recall

→ fake recall = 1 – 1 / rej

→ rej = 1 / (1 – fake recall)



eff=

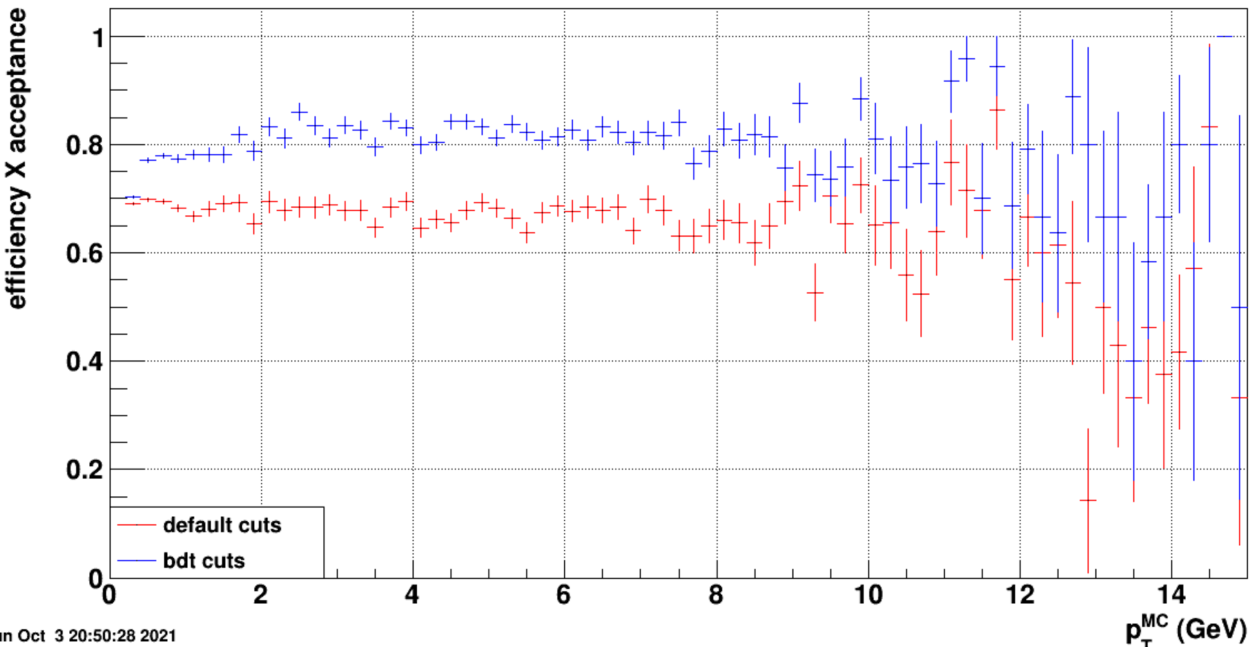


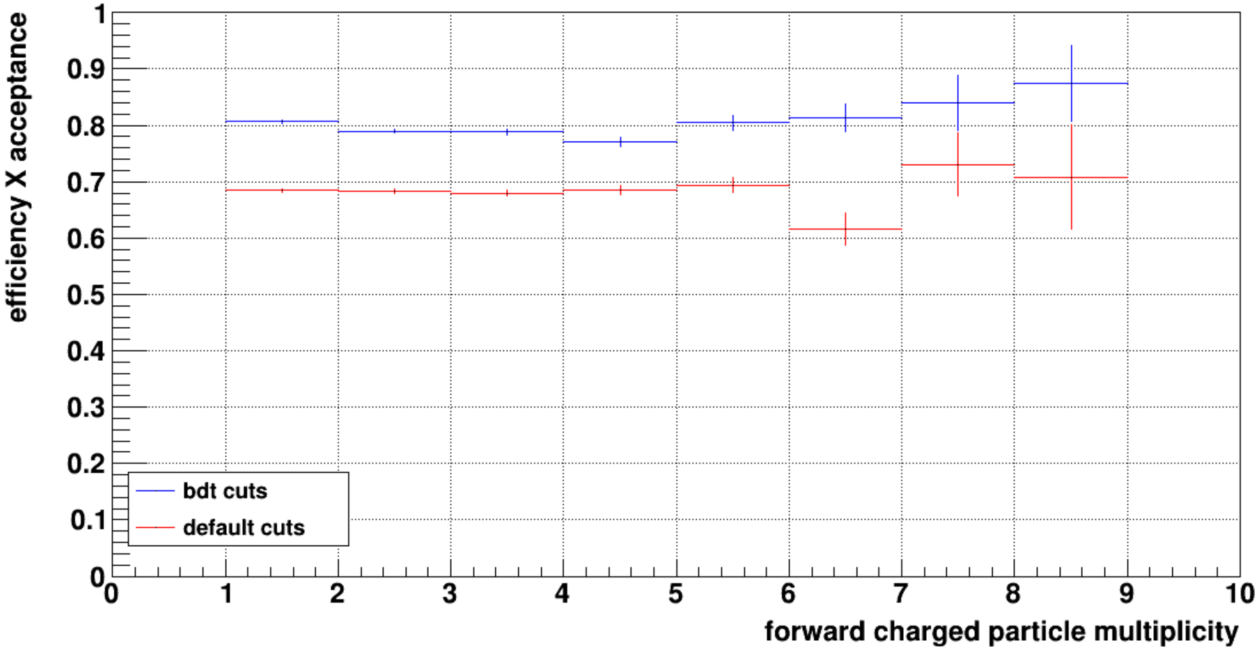
1/background rej =

Results

- PYTHIA 8 pp Drell-Yan at $\sqrt{s}$ = 510 GeV
- Selection: $2.5 < \eta < 4$, $\phi > 0$, $p_T > 0.2$ GeV

$$\textbf{efficiency X acceptance} = \frac{\text{\# reconstructed tracks within the selection}}{\text{\# MC truth tracks within the selection}}$$





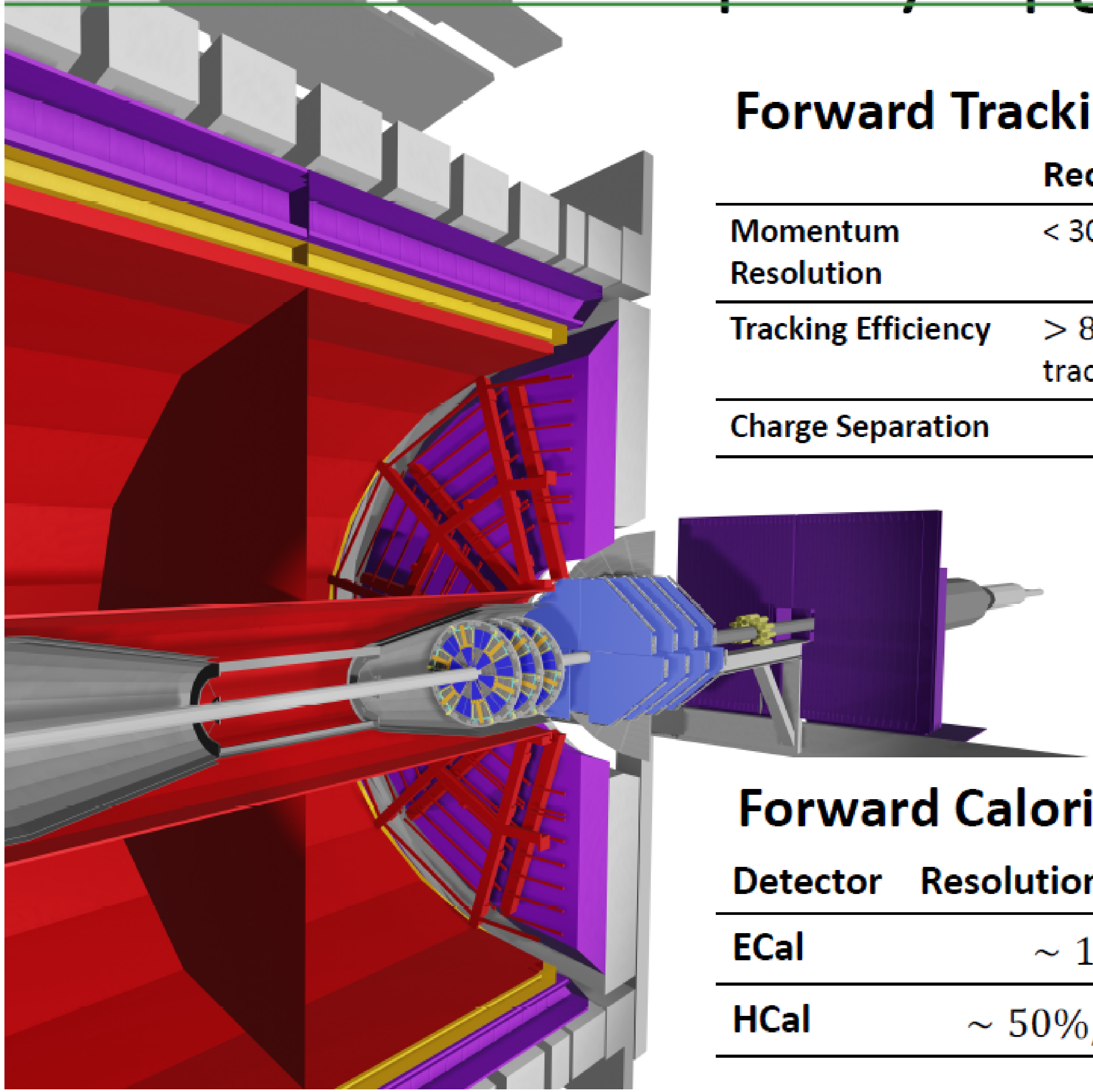
purity: (82.6% ± 0.6%) -> (93.7% ± 0.7%) (see plots in backup)

Summary

- STAR forward upgrade will allow us to make measurements at low x, and give us new insights of nucleus structure
- Using machine learning technique, we can reconstruct forward tracks from hits with improved efficiency and purity.

Backup

STAR Forward Rapidity Upgrade



Forward Tracking System

	Requirement	Motivation
Momentum Resolution	< 30%	A+A goals
Tracking Efficiency	> 80% @ 100 tracks / event	A+A goals
Charge Separation	–	p+p / p+A goals

Forward Calorimeter System

Detector	Resolution p+p and p+A	Resolution A+A
ECal	$\sim 10\%/\sqrt{E}$	$\sim 20\%/\sqrt{E}$
HCal	$\sim 50\%/\sqrt{E} + 10\%$	–

Hit triplet observables

- Hit triplet (crit3) observables, calculated for each triplet of hits that are:
  - On adjacent tracker layers (1-2-3 or 2-3-4)

Criteria 2DAngle :

$$\Delta x_1 = x_B - x_A$$
$$\Delta y_1 = y_B - y_A$$

$$\Delta x_2 = x_C - x_B$$
$$\Delta y_2 = y_C - y_B$$

$$\cos^2(\theta) = \frac{(\Delta x_1 \Delta x_2 + \Delta y_1 \Delta y_2)^2}{(\Delta x_1^2 + \Delta y_1^2)(\Delta x_2^2 + \Delta y_2^2)}$$

Criteria 3DAngle:

$$\Delta x_1 = x_B - x_A$$
$$\Delta y_1 = y_B - y_A$$
$$\Delta z_1 = z_B - z_A$$

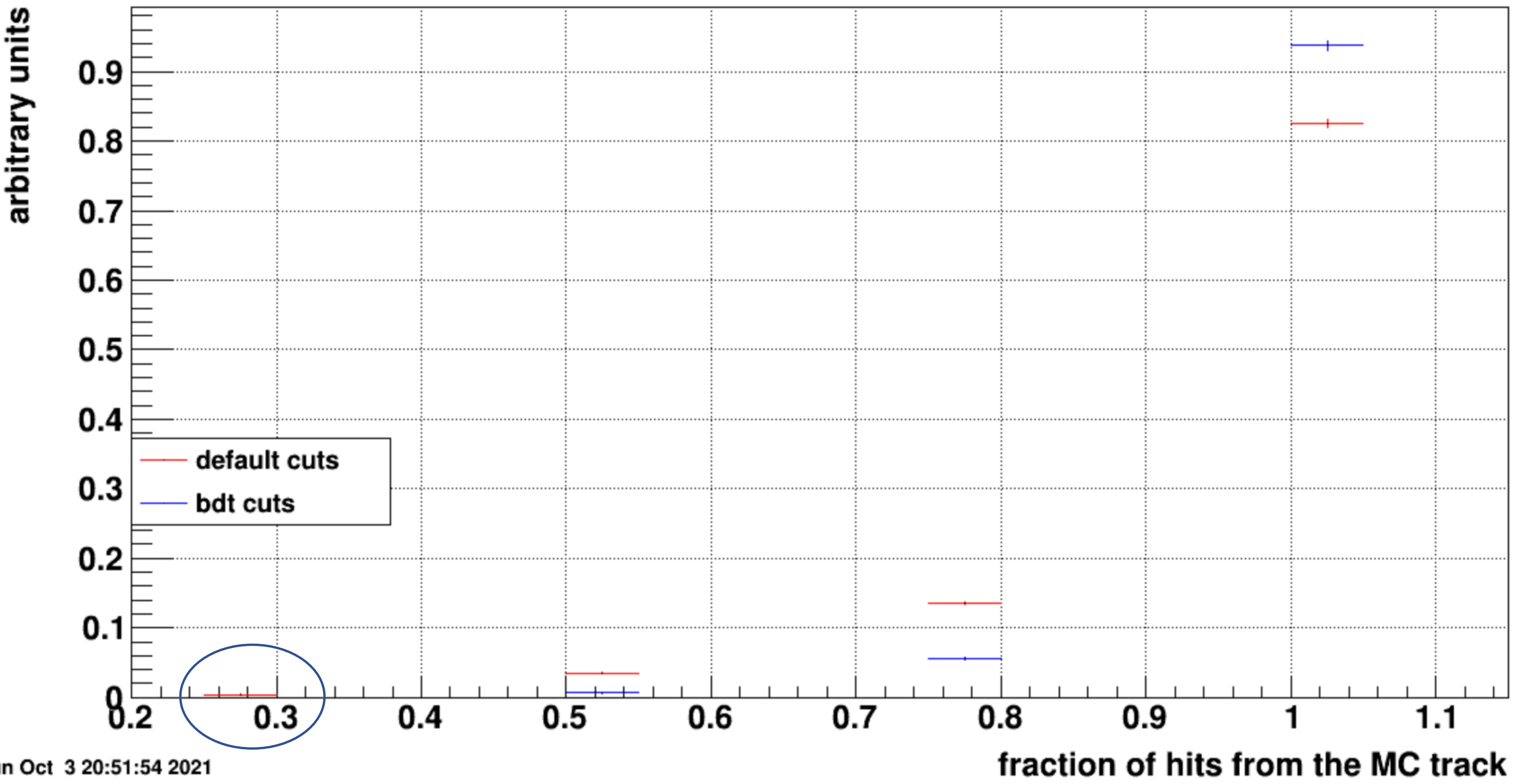
$$\Delta x_2 = x_C - x_B$$
$$\Delta y_2 = y_C - y_B$$
$$\Delta z_2 = z_c - z_B$$

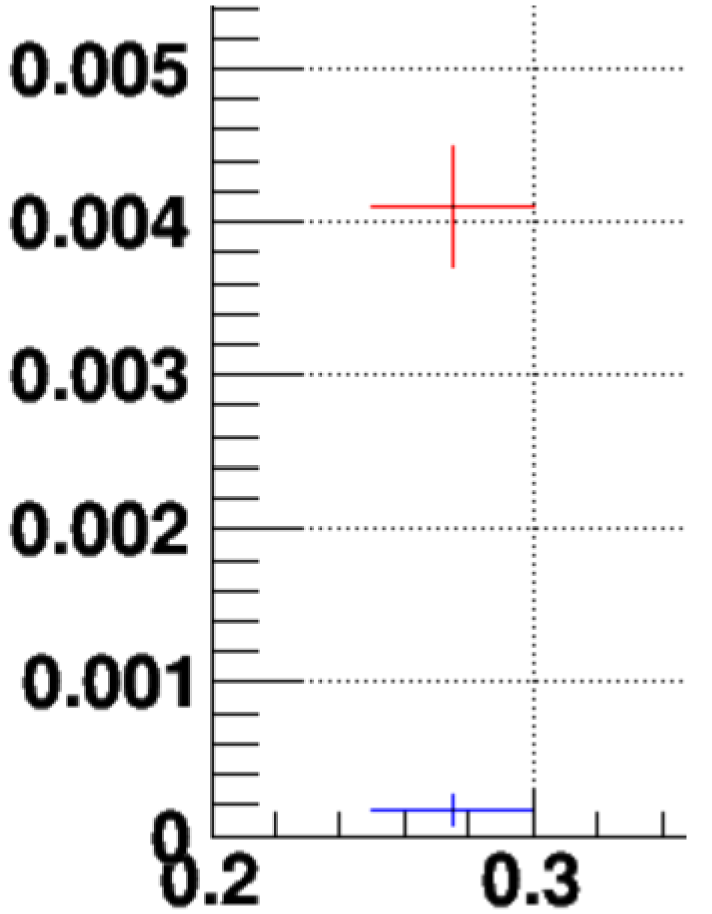
$$\cos^2(\theta) = \frac{(\Delta x_1 \Delta x_2 + \Delta y_1 \Delta y_2 + \Delta z_1 \Delta z_2)^2}{(\Delta x_1^2 + \Delta y_1^2 + \Delta z_1^2)(\Delta x_2^2 + \Delta y_2^2 + \Delta z_2^2)}$$

Criteria ChangeRZRatio :

$$\Delta RZ = \left(\frac{\Delta R}{\Delta z}\right)_{BA}^2 - \left(\frac{\Delta R}{\Delta z}\right)_{BC}^2$$

Improved purity





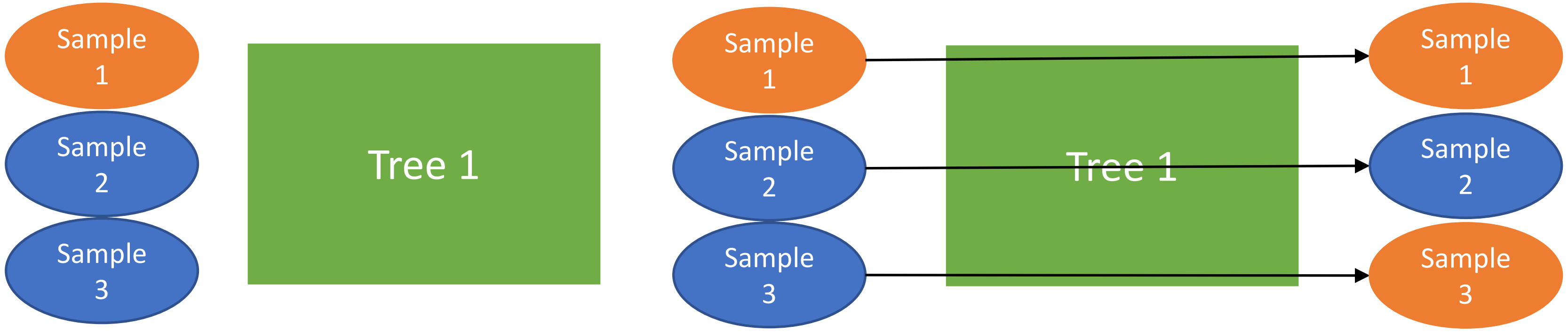
Model construction

Initialize

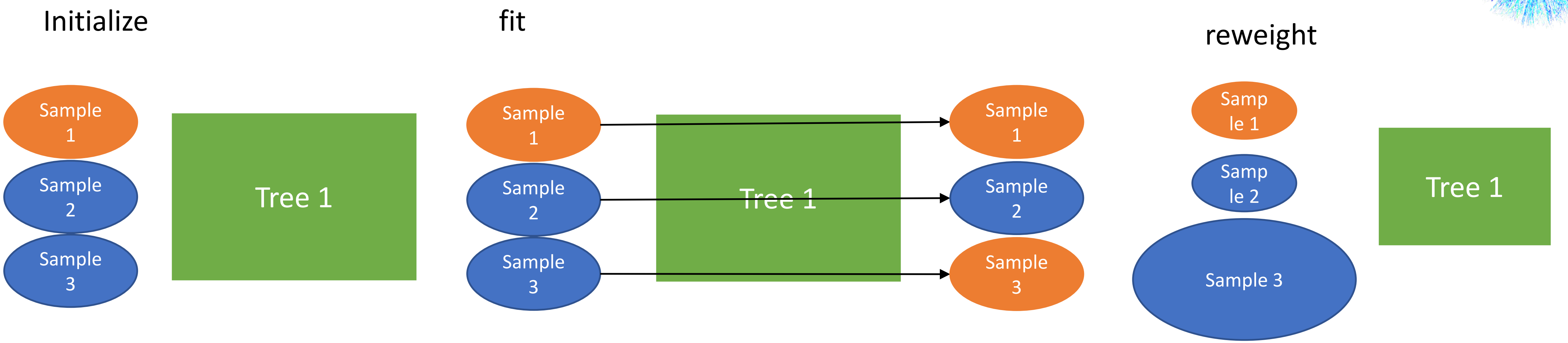
Model construction

Initialize

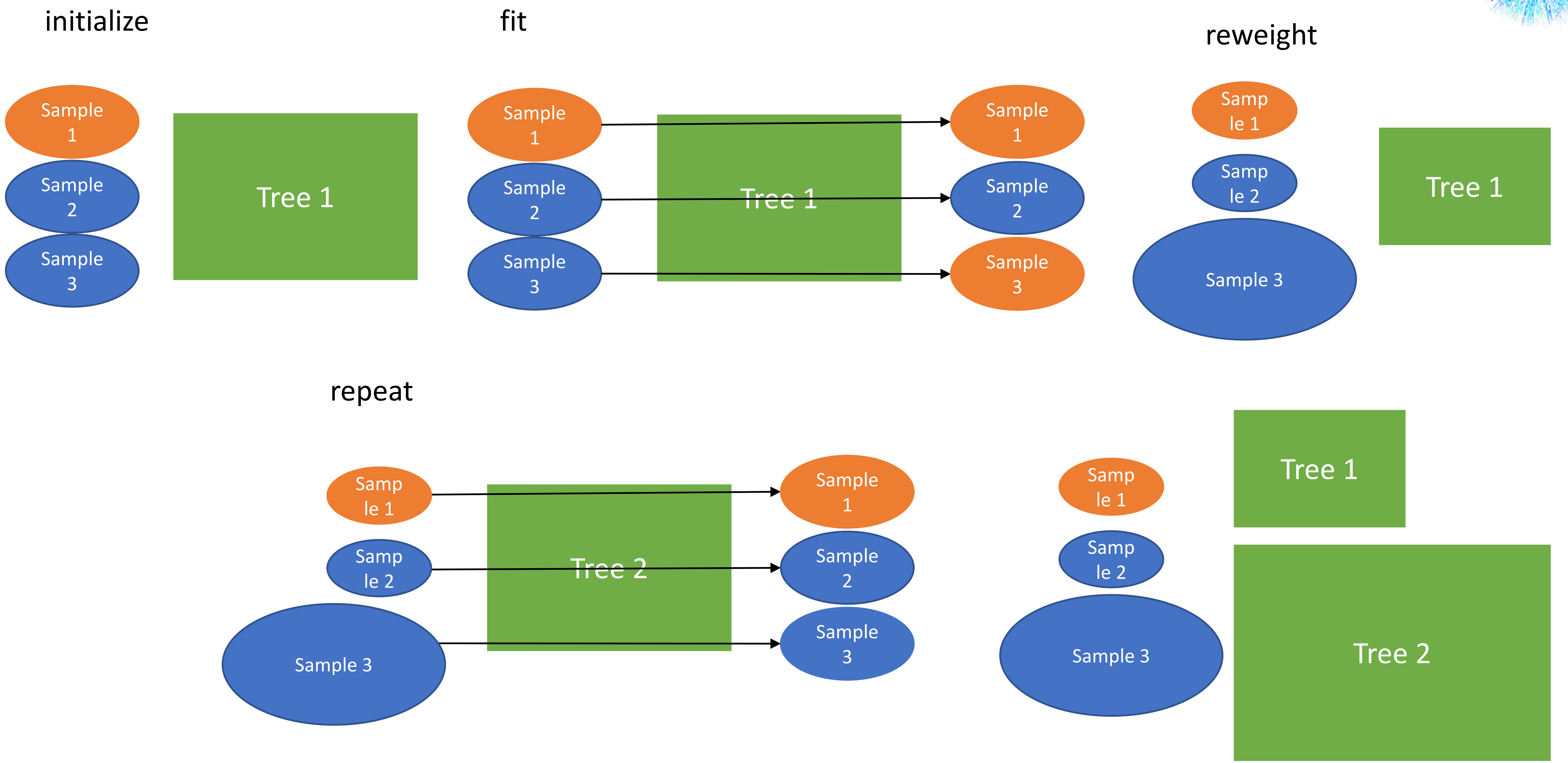
fit



Model construction



Model construction



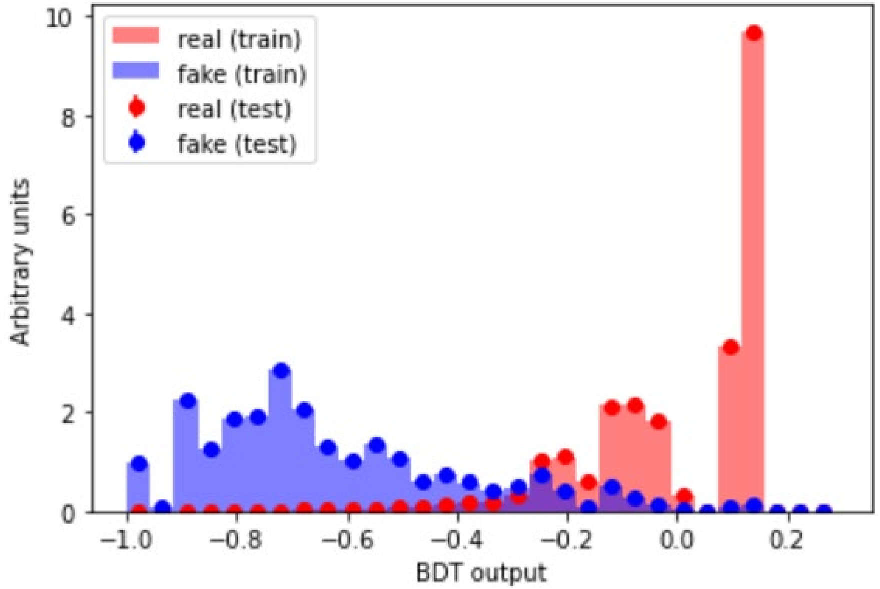
The samples and trees

- Training samples: 5.7 million of hit pairs
- Trees: 800 of 3-layer estimators

entry	subentry	Crit2_RZRatio	Crit2_DeltaRho	Crit2_DeltaPhi	Crit2_StraightTrackRatio
0	0	1.021419	-4.738697	0.801099	0.691984
	1	1.022302	-4.839012	0.601002	0.688314
	2	1.304533	-17.746450	19.210548	0.409136
	4	1.594602	-26.755651	22.318817	0.318865
	6	1.001245	0.990740	1.653889	0.973371
...	...	...	...	...	...
49499	4	1.007454	2.785313	0.434769	0.998189
	5	1.299637	-15.502758	12.340158	0.688496
	6	1.299396	-15.427925	12.452241	0.689373
	7	1.006945	2.691032	0.081525	0.996621
	8	1.006917	2.640335	0.675205	0.995380

```
|--- Crit2_DeltaPhi <= 0.80
|   |--- Crit2_RZRatio <= 1.02
|   |   |--- Crit2_StraightTrackRatio <= 0.98
|   |   |   |--- class: 0.0
|   |   |--- Crit2_StraightTrackRatio >  0.98
|   |   |   |--- class: 1.0
|   |--- Crit2_RZRatio >  1.02
|   |   |--- Crit2_RZRatio <= 1.11
|   |   |   |--- class: 0.0
|   |   |--- Crit2_RZRatio >  1.11
|   |   |   |--- class: 0.0
|--- Crit2_DeltaPhi >  0.80
|   |--- Crit2_RZRatio <= 1.03
|   |   |--- Crit2_DeltaPhi <= 3.02
|   |   |   |--- class: 0.0
|   |   |--- Crit2_DeltaPhi >  3.02
|   |   |   |--- class: 0.0
|   |--- Crit2_RZRatio >  1.03
|   |   |--- Crit2_RZRatio <= 1.11
|   |   |   |--- class: 0.0
|   |   |--- Crit2_RZRatio >  1.11
|   |   |   |--- class: 0.0
```

Testing for hit pairs – different number of layers and trees



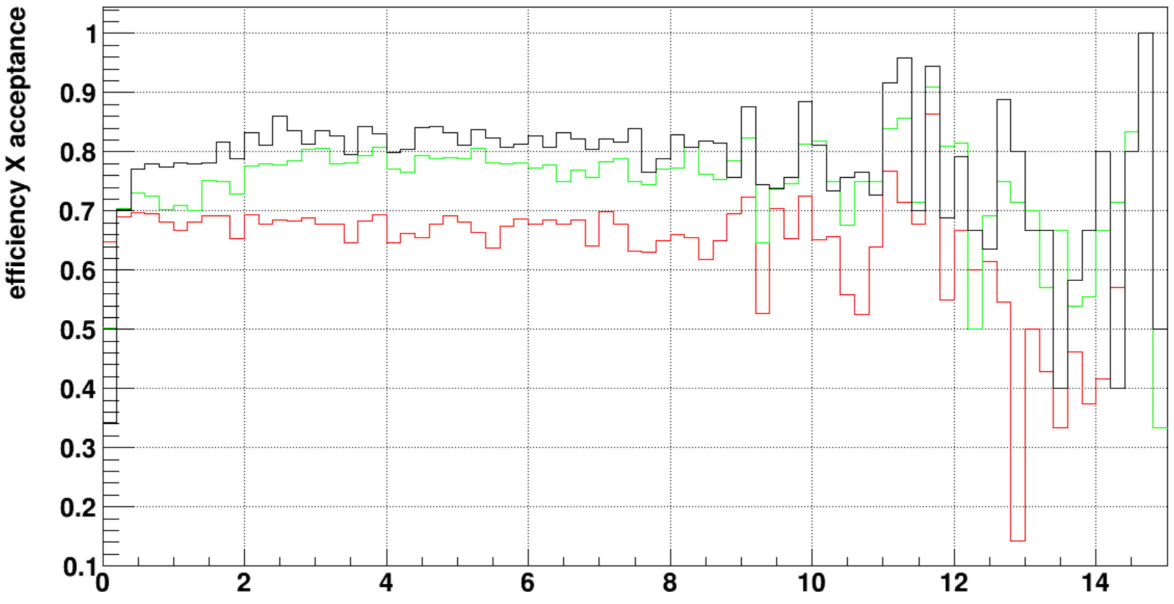
	3 layers, 800 estimators	3 layers, 20 estimators	4 layers, 20 estimators
K-S statistic	0.00065	0.00048	0.00044
p-value	0.39	0.77	0.86

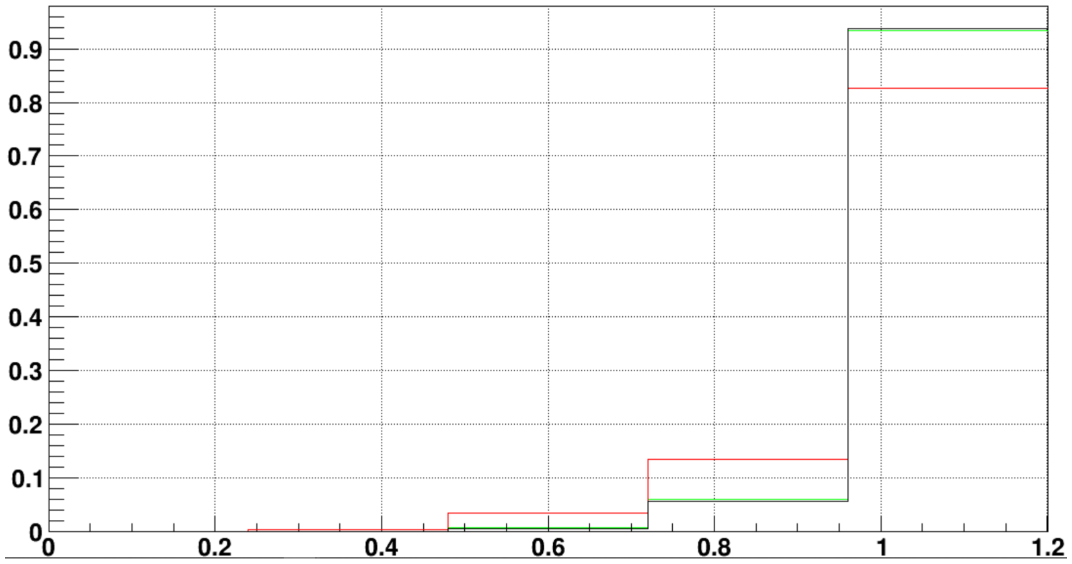
If the KS statistic is small or the p-value is high, then we cannot reject the hypothesis that the distributions of the two samples are the same.

Our BDT is stable against overtraining

Different number of layers and trees

- Drell Yan events
- Black = 3 layers 800 trees (blue error bars in main slides)
- Green = 4 layers 20 trees for crit2, 3 layers 20 trees for crit3
- Red = default geometric cuts (red error bars in main slides)

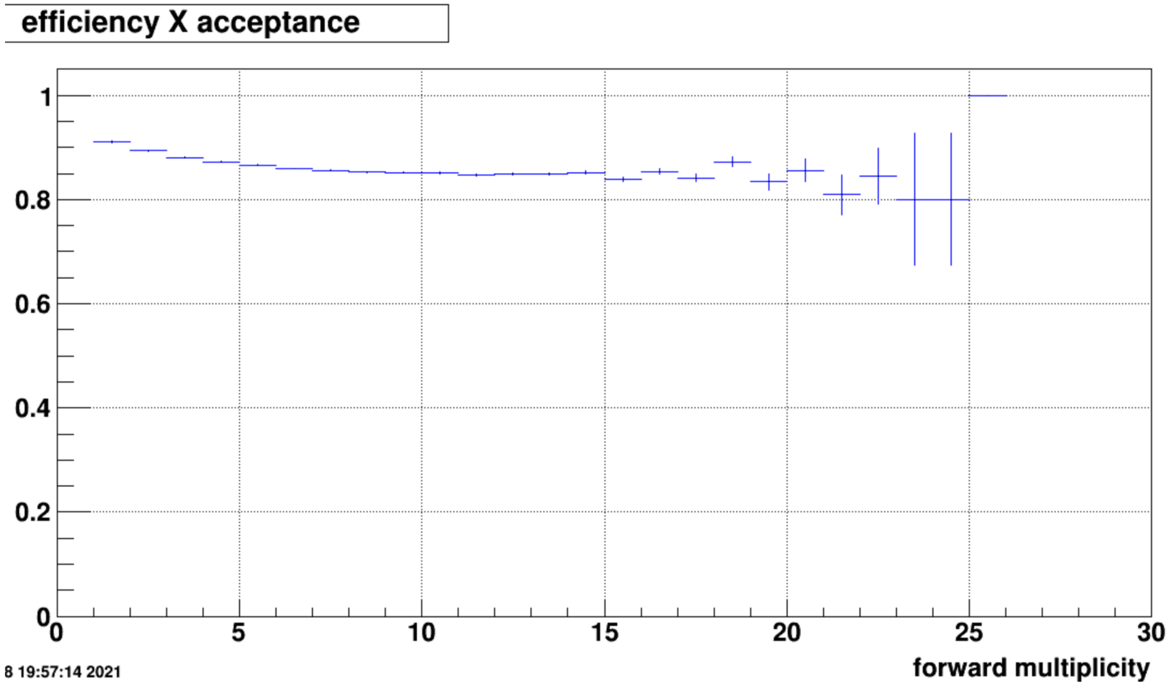




Different number of layers and trees

- 4 layers 20 trees for crit2, 3 layers 20 trees for crit3

Drell Yan events



Minbias events