JOINT INSTITUTE
FOR NUCLEAR RESEARCH

Update on di-electron analysis: Machine learning study

Sudhir Pandurang Rode, Itzhak Tserruya

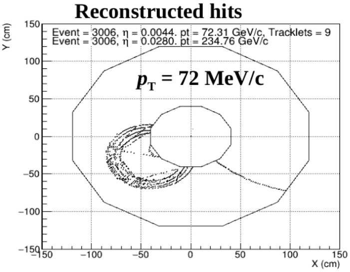
March 19, 2024

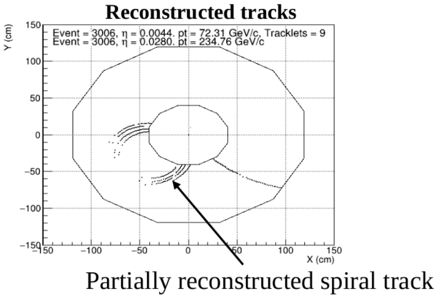
MPD Cross-PWG meeting

Content

- Quick recap of the analysis so far
- Machine learning approach for improving the $e^{\pm}$ PID efficiency
  - Training of the MC sample
  - Performance validation
  - Implementation in the dilepton analysis
- Next steps

Quick recap





- With current track reconstruction algorithm, low p_{T} tracks are not reconstructed properly even though full hit information is available in the detector for tracks that enter the TPC ($p_{\mathrm{T}} > \approx 30$ MeV/c).
- Question is, in an ideal detector, what would be the maximum possible benefit in the combinatorial background (CB) reduction, if we were to detect these tracks.
- As per our principle study, potentially, there is about 5-8 factor improvement possible in CB rejection.

Quick recap: Analysis strategy

$\Rightarrow$ Three electron pools:

$\rightarrow$ Pool-1 for fully reconstructed tracks[1] in fiducial area ($|\eta| < 0.3$)

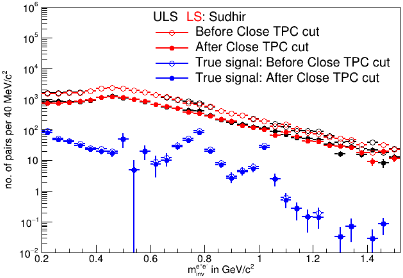
$\rightarrow$ Pool-2 for fully reconstructed tracks in veto area $0.3 < |\eta| < 1.0$.

$\rightarrow$ Pool-3 with tracks reconstructed in the TPC only.

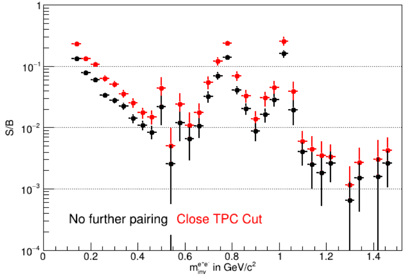
- Step 1 - No further pairing (NFP): Tracks belonging to fully reconstructed π^0 Dalitz are tagged and not used for further pairing.
- Step 2 - Close TPC cut (CTC): Track from Pool-1 in an event is paired with tracks from Pool-3 in the same event and both tracks are removed as a potential Dalitz pair if they have $M_{\text{inv}} < 80$ MeV$/c^2$ and opening angle < 10 degrees (this cut is opening angle dependent).
- Step 3 - Rest of the tracks with $p_{\text{T}} > 200$ MeV from Pool-1 are paired among themselves to build ULS and LS pair spectra.

[1] TOF matched tracks identified in the TPC and TOF

Quick recap: Dielectron cocktail[3]



Request 25 → 36M events



Mass region: 0.2 to 1.5 GeV/c →

Steps	Sig	LS	S/B	[2]BFE $= \frac{S^2}{S+2B}$
Before CTC	644.5	26285.2	0.024	7.8
After CTC	575.9	13317.7	0.043	12.2

- Due to limited satistics, signal is not U-L, but it is true reconstructed di-electron pairs.
- Close TPC cut approach improves S/B ratio by $\approx 75-80\%$ → CB rejection by factor 2.
- Still significant improvement possible by improving the recognition of low p_T tracks.

[2] Background free equivalent - signal with same relative error as in background free situation

[3] TPC+TOF analysis

Quick recap

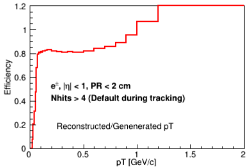
- Trying to understand the origin of remaining background after close TPC cut.

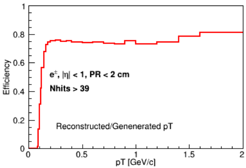
Total reconstructed tracks after close TPC cut:	**1.69268e+06**
Below: Only Conversion and π^0 Dalitz sources are considered --	
a. Track has Partner with pT < 35 MeV ($\lvert\eta\rvert$ < 2.5):	**419595 (~25%)**
b. Track has Partner inside TPC i.e. 35 < pT < 100 MeV ($\lvert\eta\rvert$ < 2.5):	**580428 (~34%)**
c. Track has Partner with pT > 110 MeV ($\lvert\eta\rvert$ < 2.5):	**266075 (~16%)**
Track is hadron:	**102041 (~6%)**
Rest (Signal (η, etc), conversion, π^0 Dalitz whose partner outside TPC, ...)	**324536 (~19%)**

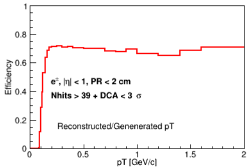
- Is **b.** reflecting inefficiency of the current tracking algorithm for low p_T tracks? Need expert help to improve the low-p_T tracking reconstruction.
- Additional and independent venue:
  - Improve the overall eid efficiency using Machine Learning techniques (both TPC Only and TPC+TOF+ECal) → Will help in improving the signal as well as S/B.

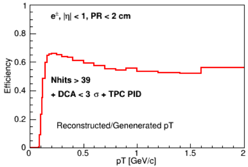
- (a.) is lost but (b.) is still recoverable → requires expert to look into algorithm.
- This study suggests that along with improving efficiency of low p_T track reconstruction, overall improvement in PID efficiency is also going to help in enhancing the S/B, signal significance and background free equivalent signal.

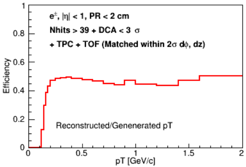
Quick recap

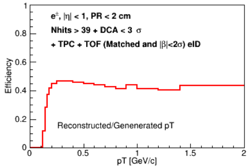


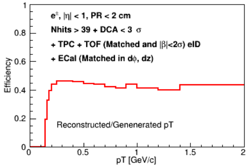


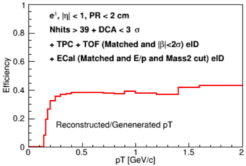












- Significant drop in efficiency due to 1D cuts.
- Improvement in the efficiency → better S/B, signal significance and background free equivalent signal.

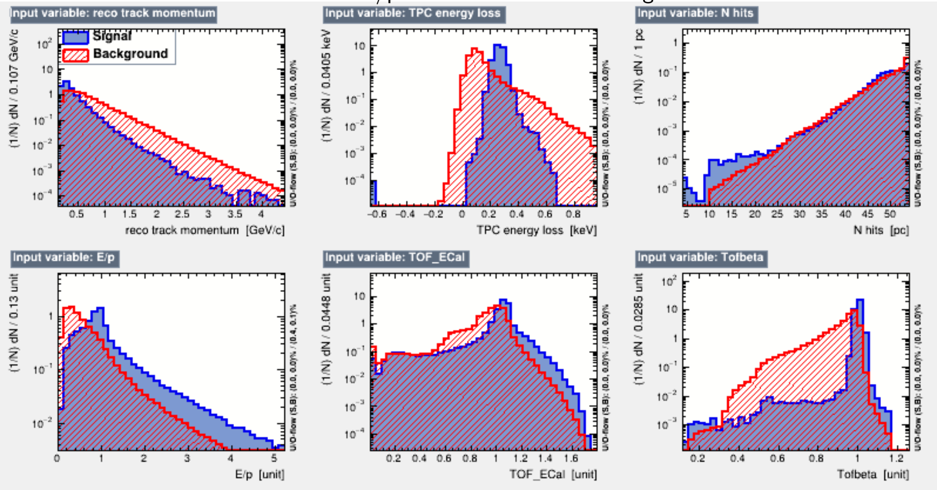
Details

- Machine learning approach can help in improving the particle identification efficiency $\rightarrow$ S/B and significance.
- All charged tracks with DCA $< 3\sigma$ and matched in TOF ($< 2\sigma$ of dϕ and dz) and ECal ($< 3\sigma$ of dϕ and dz) $\rightarrow e^{\pm}$ (Signal) and Rest (Background).
- Two sample: One sample for training and overtraining test: Actual proportion of Signal (568K) and Background (94M) $\rightarrow$ divided into two subsamples, second sample is for performance validation.
- For Training (50%): Actual proportion of Signal (284K) and Background (47M).
- For Overtraining test (50%): Actual proportion of Signal (284K) and Background (47M).
- The Kolmogorov Smirnov test provides a p-value[4] equal to the statistical probability that two samples are drawn from the same distribution.

[4] The smaller the p, the greater the overtraining. Since the training and testing samples will never be identical, a very small degree of overtraining may be unavoidable. As a rule of thumb, it is recommended to try to reduce overtraining if $p < 0.01$, especially if the separation is visibly poorer for the testing samples than for the training samples.

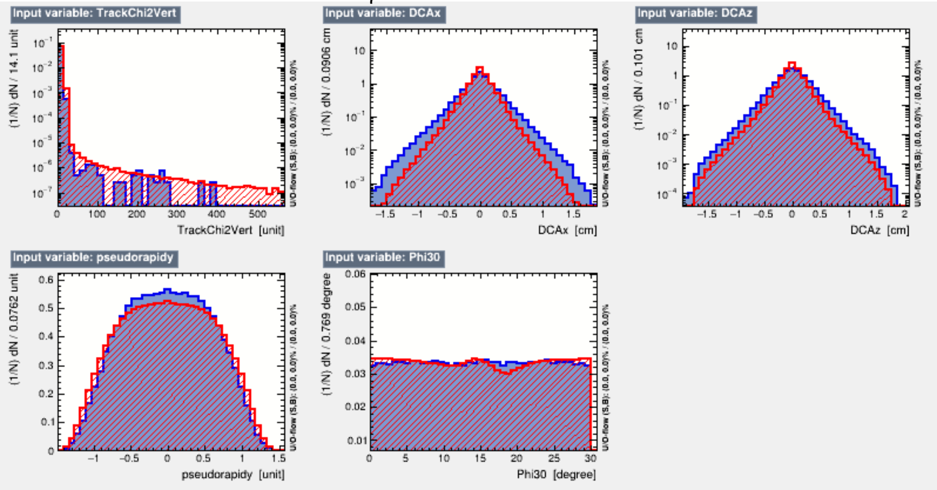
Input variables

- Momentum
- dEdX
- No of Hits
- E/p
- Time of flight in the ECal
- Time of flight in the TOF

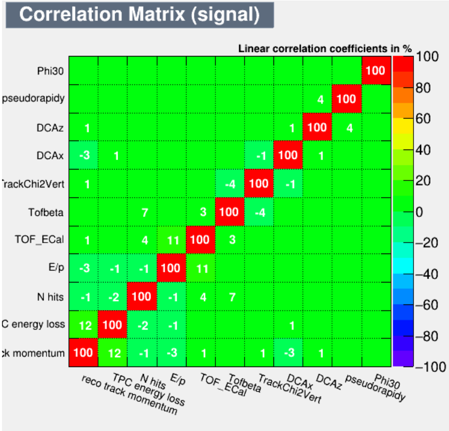


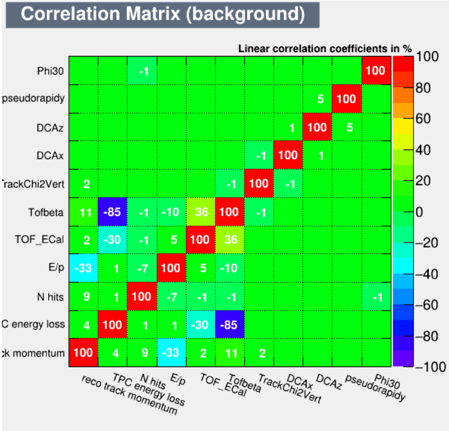
Input variables

- Track chi2 to vertex
- DCAx
- DCAz
- η
- Azimuthal angle, ϕ



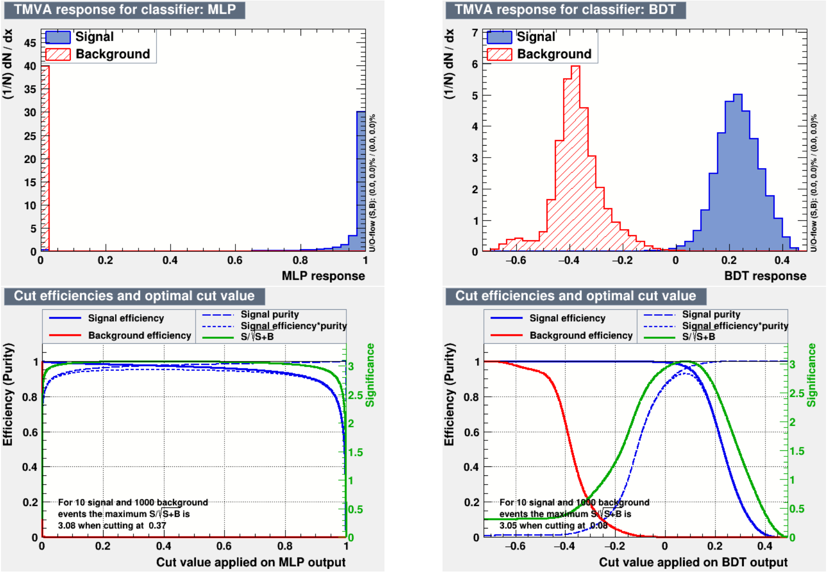
Correlation matrices: $e^{\pm}$ (Signal) and Rest (Bkg)



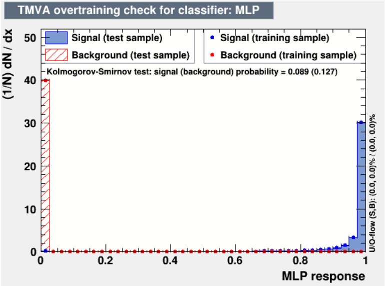


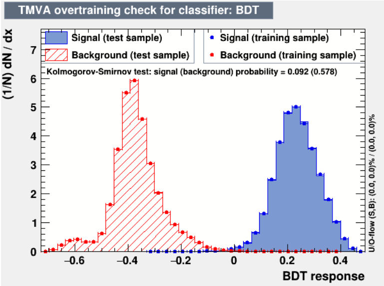
- Almost all variables for signal are independent.
- In case of background, there is correlation among some variables, for instance, dEdx and Tofbeta.

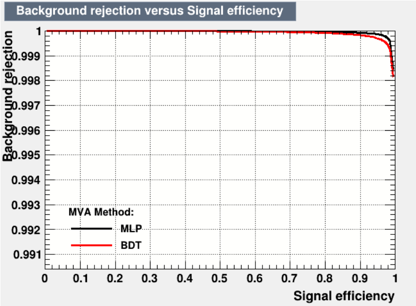
Response with Prior DCA 3σ cut; All $e^{\pm}$ (Signal) and Rest (Bkg)



Response with Prior DCA 3σ cut; All $e^{\pm}$ (Signal) and Rest (Bkg)

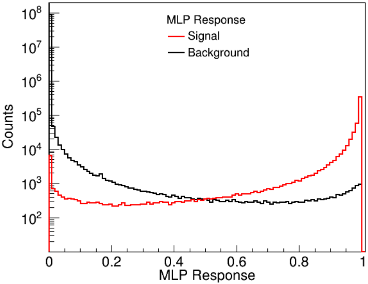


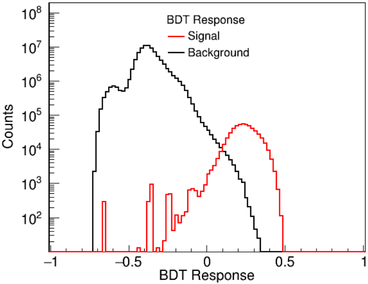




Performance validation using test sample

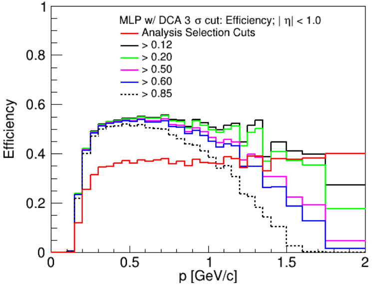
Response with Prior DCA 3σ cut; All $e^{\pm}$ (Signal) and Rest (Bkg)

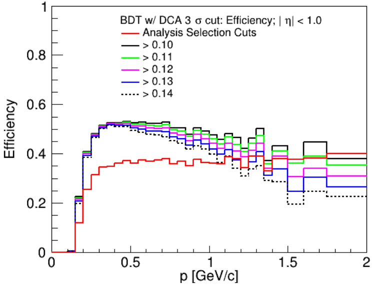




- Response for actual proportion of signal and background in the test sample.
- Clear separation between signal and background by both classifiers.

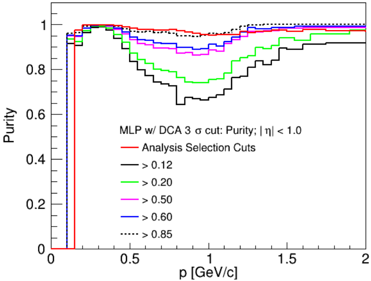
Efficiency: Primary $e^{\pm}$

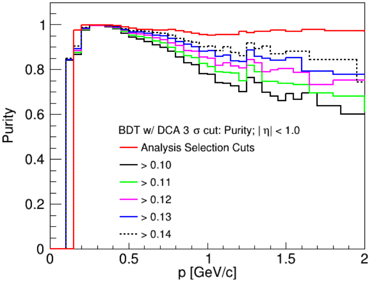




- Denominator: All generated $e^{\pm}$ tracks (PR < 2 cm).
- Numerator: $+$ Response cut.

Purity; All $e^{\pm}$ (Signal) and Rest (Bkg)



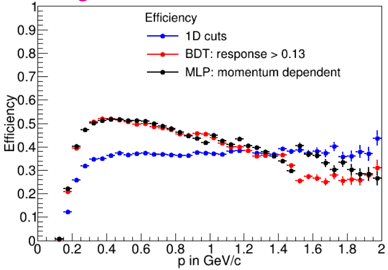


- Denominator: All tracks with $\mathrm{DCA} < 3\sigma$ matched in TOF and ECAL within Response cut.
- Numerator:All $e^{\pm}$ tracks with $\mathrm{DCA} < 3\sigma$ matched in TOF and ECAL within Response cut.
- With momentum dependent selection of response, purity as good as 1D cuts (analysis selection cuts) and better efficiency can be achieved.

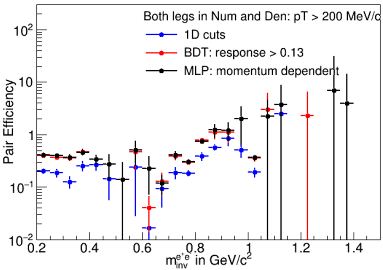
Implementation of Machine learning results in pair analysis: $\approx$ 13M events

Efficiencies and Purity: $\approx$ 13M events

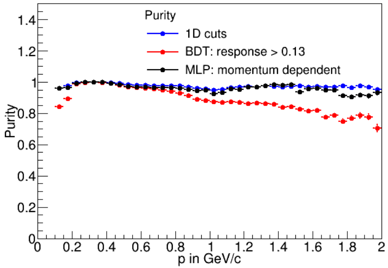
Total single electron reconstruction efficiency



Total dielectron pair reconstruction efficiency



Electron purity



- MLP is performing better at higher momenta.
- Significant improvement in both single as well as pair efficiency.
- Purity with MLP matches with the 1D cuts.
- BDT: response > 0.13.
- MLP: momentum dependent, for $p < 1.0$, response > 0.85, $1.0 < p < 1.15$, response > 0.7, $1.15 < p < 1.25$, response > 0.6, $1.25 < p < 1.5$, response > 0.5, $1.5 < p < 1.75$, response > 0.2 upto $p > 1.75$, response > 0.12 $\rightarrow$ smoothening required.

Analysis strategy (slightly updated) - Reminder

$\Rightarrow$ Three electron pools:

$\rightarrow$ Pool-1 for fully reconstructed tracks[5] in fiducial area ($|\eta| < 0.3$)

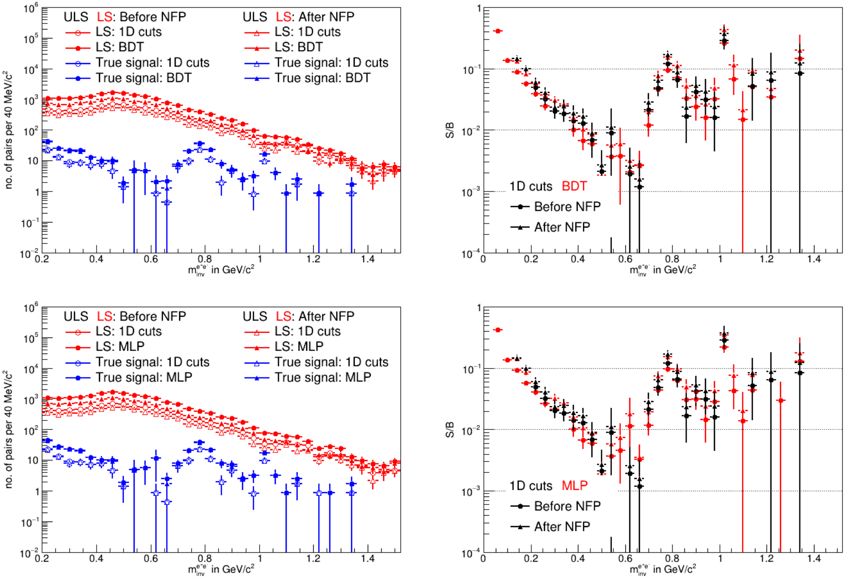
$\rightarrow$ Pool-2 for fully reconstructed tracks in veto area $0.3 < |\eta| < 1.0$.

$\rightarrow$ Pool-3 with tracks not matched/identified in the TOF.

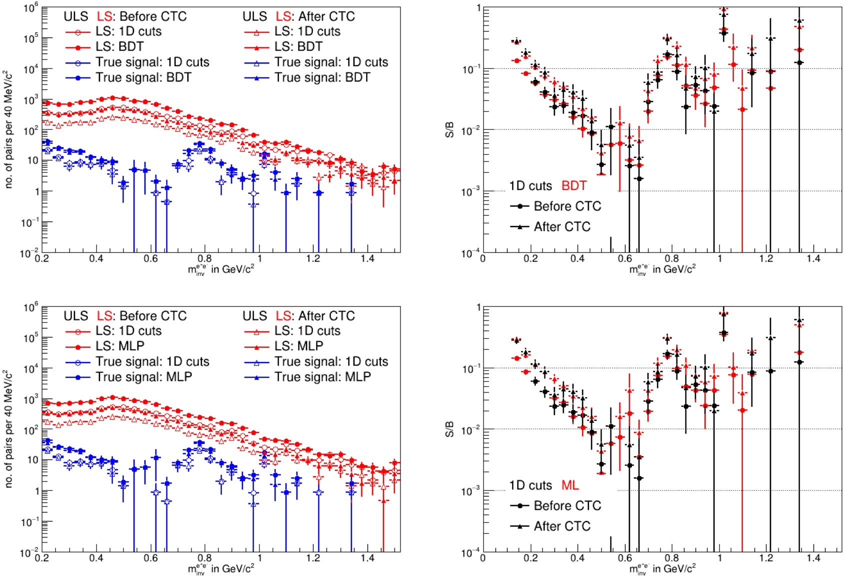
- Step 1 - No further pairing (NFP): Tracks belonging to fully reconstructed π^0 Dalitz are tagged and not used for further pairing.
- Step 2 - Close TPC cut (CTC): Track from Pool-1 in an event is paired with tracks from Pool-3 in the same event and both tracks are removed as a potential Dalitz pair if they have $M_{\mathrm{inv}} < 80$ MeV$/c^2$ and opening angle < 10 degrees (No opening angle dependent selection).
- Step 3 - Rest of the tracks with $p_{\mathrm{T}} > 200$ MeV from Pool-1 are paired among themselves to build ULS and LS pair spectra.

[5] TOF and ECal matched tracks identified in the TPC, TOF and ECal

Cocktail after No further pairing (NFP) using BDT & MLP



Cocktail after Close TPC Cut (CTC)[6] using BDT & MLP



Comparison of results using 1D cuts, BDT and MLP

Following values are estimated in the invariant mass between 0.2 to 1.5 GeV/c →

	1D cuts				BDT				MLP			
	S	B	S/B	BFE ($\frac{S^2}{S+2B}$)	S	B	S/B	BFE ($\frac{S^2}{S+2B}$)	S	B	S/B	BFE ($\frac{S^2}{S+2B}$)
Before NFP	155	7627	0.020	1.6	296	17753	0.017	2.5	313	17780	0.018	2.7
After NFP	152	5791	0.026	2.0	288	11363	0.025	3.6	303	11278	0.027	4.0
After CTC	129	2776	0.047	2.9	251	5101	0.049	6.0	266	5053	0.053	6.8

- At no further pairing step, S/B ratio remains similar for all three cases.
- Background free equivalent signal seems to have improved.
- After Close TPC cut, hint of improvement in the S/B ratio using MLP and BDT classifers.

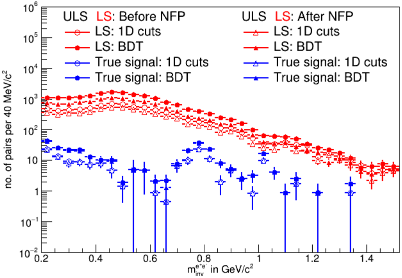
Conclusions and Next steps

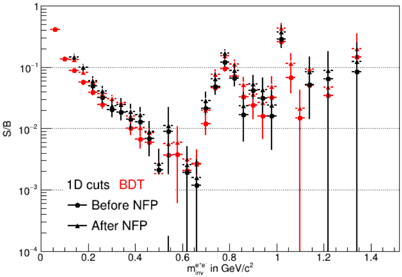
- Machine learning seems to be improving the PID efficiency.
- Enhancement in the background free equivalent signal, keeping S/B unchanged after no further pairing.
- Hint of improvement in the S/B after close TPC cut.
- Extend training to TPC only as well as TPC + ECal samples to further improve the S/B and significance.
- Optimise response cut for best efficiency and purity.
- Momentum differential training of the MC sample.

Special thanks to Igor Rufanov for the discussions.

BACK-UP

Cocktail after No further pairing (NFP) using BDT

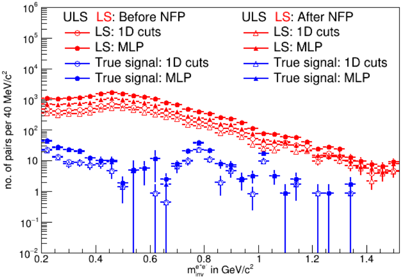


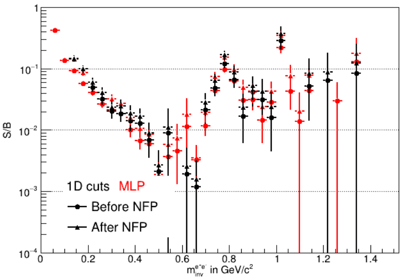


Mass region: 0.2 to 1.5 GeV/c $\rightarrow$

Steps	Sig	Err	LS	Err	S/B	Err	$\frac{S}{\sqrt{S+B}}$	$\frac{S^2}{S+2B}$
1D Cuts before NFP	155.0	12.5	7626.8	87.3	0.0203	0.0017	1.76	1.56
1D Cuts after NFP	151.7	12.3	5791.3	76.1	0.0262	0.0022	1.97	1.96
BDT before NFP	296.2	17.2	17752.7	133.2	0.0167	0.001	2.2	2.45
BDT after NFP	287.9	16.9	11362.6	106.6	0.0253	0.0015	2.67	3.6

Cocktail after No further pairing (NFP) using MLP

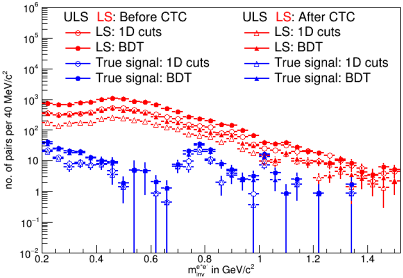


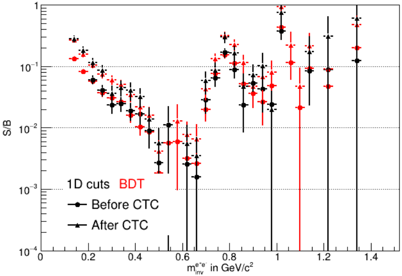


Mass region: 0.2 to 1.5 GeV/c $\rightarrow$

Steps	Sig	Err	LS	Err	S/B	Err	$\frac{S}{\sqrt{S+B}}$	$\frac{S^2}{S+2B}$
1D Cuts before NFP	155.0	12.5	7626.8	87.3	0.0203	0.0017	1.76	1.56
1D Cuts after NFP	151.7	12.3	5791.3	76.1	0.0262	0.0022	1.97	1.96
BDT before NFP	296.2	17.2	17752.7	133.2	0.0167	0.001	2.2	2.45
BDT after NFP	287.9	16.9	11362.6	106.6	0.0253	0.0015	2.67	3.6
MLP before NFP	313.1	17.7	17780.1	133.3	0.0176	0.0010	2.3	2.73
MLP after NFP	303.2	17.4	11277.5	106.2	0.0269	0.0016	2.82	4.02

Cocktail after Close TPC Cut (CTC)[7] using BDT



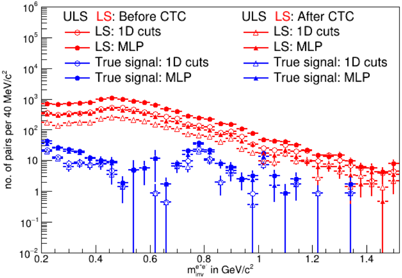


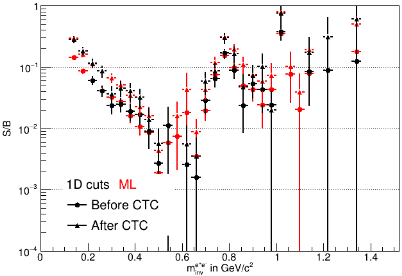
Mass region: 0.2 to 1.5 GeV/c $\rightarrow$

Steps	Sig	Err	LS	Err	S/B	Err	$\frac{S}{\sqrt{S+B}}$	$\frac{S^2}{S+2B}$
1D Cuts before CTC	151.7	12.3	5791.3	76.1	0.0262	0.0022	1.97	1.96
1D Cuts after CTC	129.1	11.4	2776.5	52.7	0.0465	0.0042	2.40	2.93
BDT before CTC	287.9	16.9	11362.6	106.6	0.0253	0.0015	2.67	3.6
BDT after CTC	250.8	15.8	5100.6	71.4	0.0492	0.0032	3.43	6.01

[7] Here, along with TPC only, tracks matched in ECal but not in the TOF are also included.

Cocktail after Close TPC Cut (CTC)[8] using MLP





Mass region: 0.2 to 1.5 GeV/c →

Steps	Sig	Err	LS	Err	S/B	Err	$\frac{S}{\sqrt{S+B}}$	$\frac{S^2}{S+2B}$
1D Cuts before CTC	151.7	12.3	5791.3	76.1	0.0262	0.0022	1.97	1.96
1D Cuts after CTC	129.1	11.4	2776.5	52.7	0.0465	0.0042	2.40	2.93
BDT before CTC	287.9	16.9	11362.6	106.6	0.0253	0.0015	2.67	3.6
BDT after CTC	250.8	15.8	5100.6	71.4	0.0492	0.0032	3.43	6.01
MLP before CTC	303.2	17.4	11277.5	106.2	0.0269	0.0016	2.82	4.02
MLP after CTC	265.6	16.3	5052.6	71.1	0.0526	0.0033	3.64	6.8

[8] Here, along with TPC only, tracks matched in ECal but not in the TOF are also included.

Request 25 $\rightarrow$ 11M events

$\rightarrow$ **Fully reconstructed tracks: Pool 1**

- $|Vz| < 100$ cm.
- DCA x,y,z $< 3\sigma$.
- Nhits > 39
- TPC nSigma -2 to 2 sigma at $p = 0$ and -1 to 2 sigma for $p > 800$ MeV/c2.
- TOF nSigma -2 to 2 sigma
- TOF matching -2 to 2 sigma
- Limiting the eta acceptance of the reconstructed track to 0.3

$\rightarrow$ **Cuts on Partner: Pool 2**

- Same as Pool 1 except in $0.3 < \eta < 1.0$

$\rightarrow$ **Cuts on Partner for Close TPC Cut: Pool 3**

- $|\eta| < 2.5$, Nhits < 10
- DCA < 3.5 sigma
- $|$TPC nSigma$| < 2$ sigma, Those tracks who DO NOT Matched in TOF within 2 Sigma (TPC ONLY).

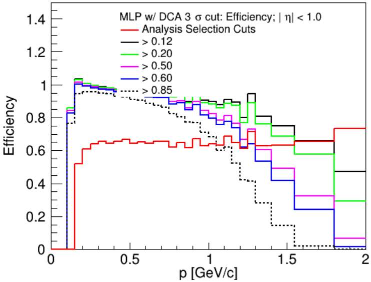
Analysis Selection Cuts vs Machine Learning

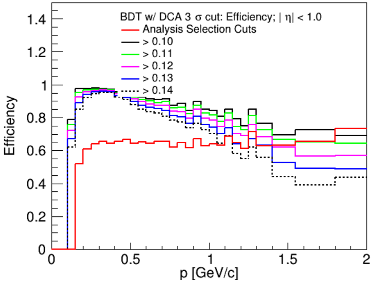
Steps	1D Cuts	Machine Learning
Denominator OR Input Sample	DCA $< 3\sigma$ Tracks matched in TOF and ECAL	DCA $< 3\sigma$ Tracks matched in TOF and ECAL
Numerator/Step 2	1D cuts	Train the model and test

$$\text{Efficiency in ML} = \frac{\text{No of primary } e^{\pm}\text{s after response cut}}{\text{No of } e^{\pm}\text{s in the input sample with DCA} < 3\sigma + |\eta| < 1.0 + \text{PR} < 2.0 \text{ cm}}$$

$$\text{Efficiency in 1D cuts} = \frac{\text{No of primary } e^{\pm}\text{s after selection cuts}}{\text{No of } e^{\pm}\text{s in the input sample with DCA} < 3\sigma + |\eta| < 1.0 + \text{PR} < 2.0 \text{ cm}}$$

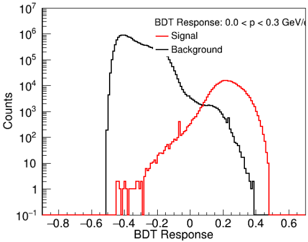
Efficiency: Primary $e^{\pm}$

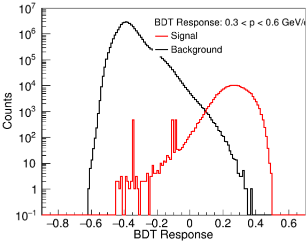


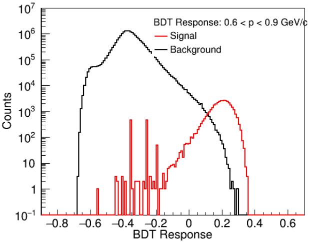


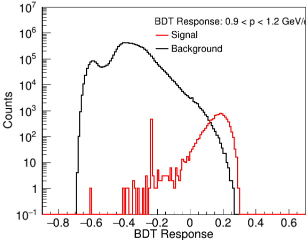
- Denominator: All $e^{\pm}$ tracks (PR $<$ 2 cm) with $\mathrm{DCA} < 3\sigma$ and matched in TOF and ECAL.
- Numerator: $+$ Response cut
- Denominator is same in both 1D cuts and machine learning.
- Benefit is that the inefficiency due to cuts on Nhits, TPC, TOF and ECAL is reduced with negligible comprise on the purity.
- However, the conversion contribution is more here because the Positron efficiency has increased.

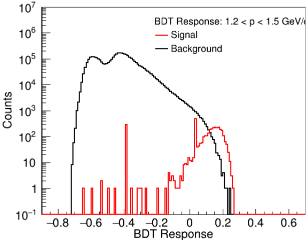
p dependent BDT Response with Prior DCA 3σ cut; All $e^{\pm}$ (Signal) and Rest

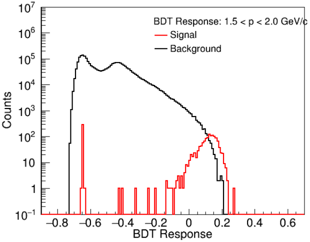




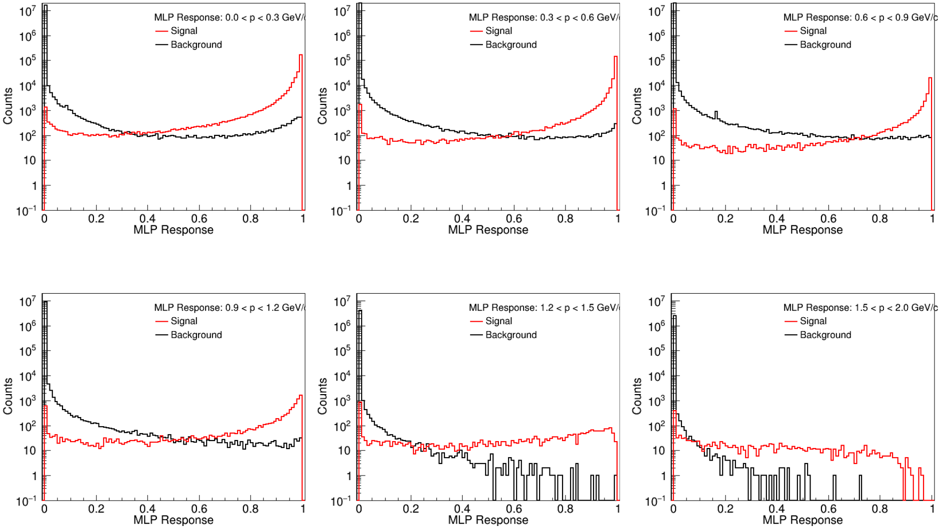








p dependent MLP Response with Prior DCA 3σ cut; All $e^{\pm}$ (Signal) and Rest



p dependent BDT Response with Prior DCA 3σ cut; All $e^{\pm}$ (Signal) and Rest