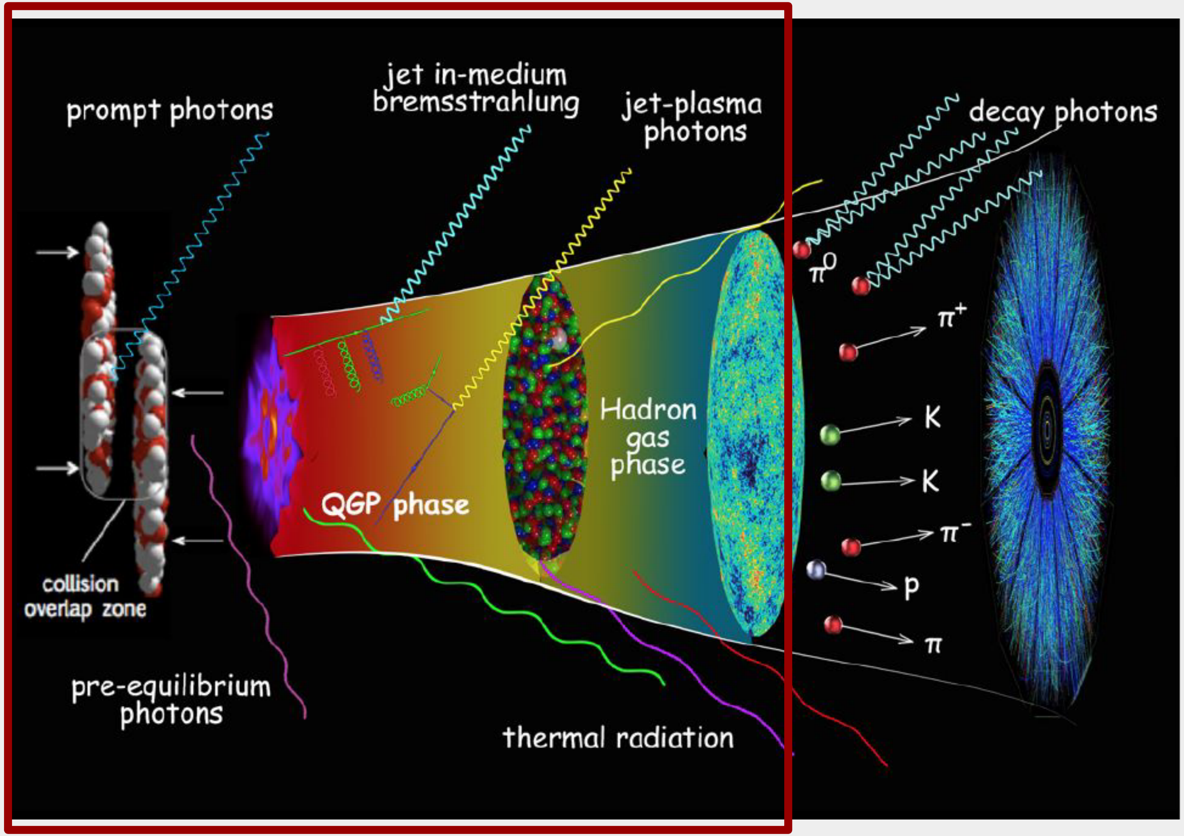
ML-based classification of photons for direct photon measurement in ALICE

CHEP 2024, Kraków
22nd October 2024

Abhishek Nath
for the ALICE collaboration
Heidelberg University

Introduction: Direct Photons



Direct photons

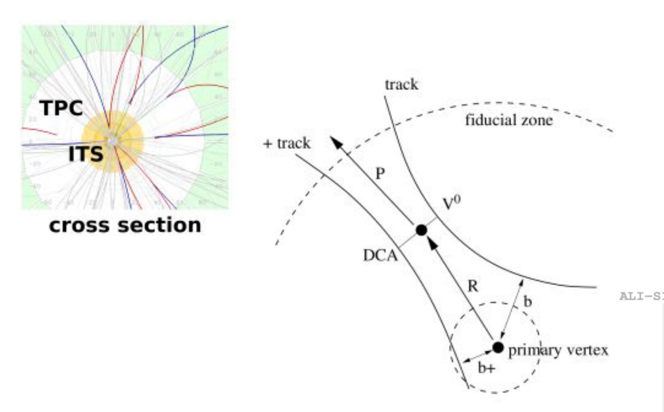
Photons are produced in all stages of collision

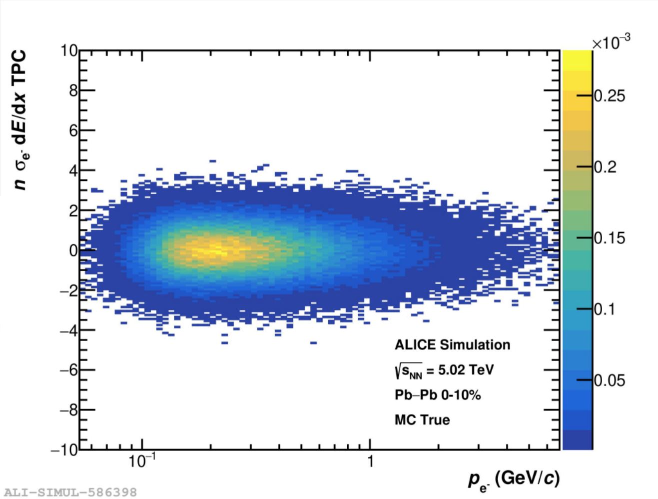
Direct photons (**signal**) is indistinguishable from Decay photon (**background**) and hence traditionally statistical approach is used to subtract background from inclusive photons

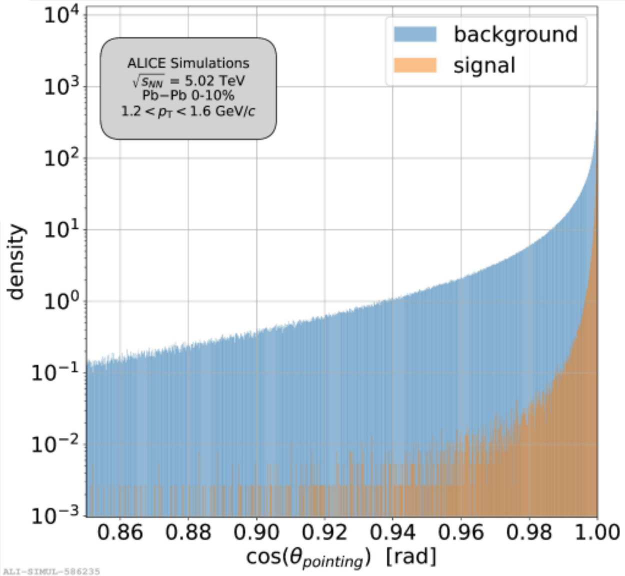
Plot from Chun Shen

Standard analysis: Photon Conversion Method

Standard analysis uses these features and implements rectangular cuts in a single or 2D feature space (**Standard cut**)



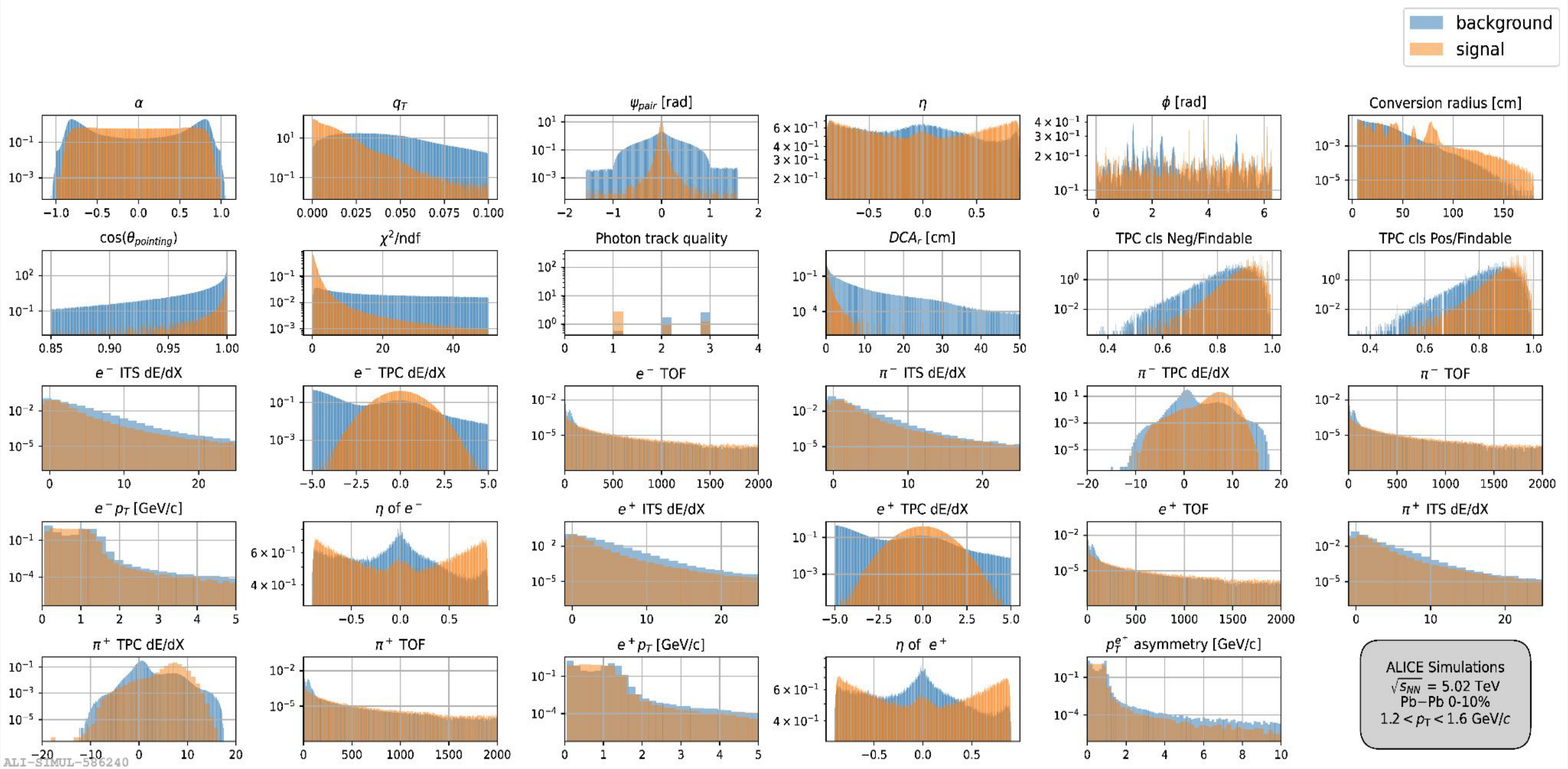




This leaves room for ML to improve sample purity and efficiency

θ_{pointing} :
angle between the two lines labelled P and R

Feature space



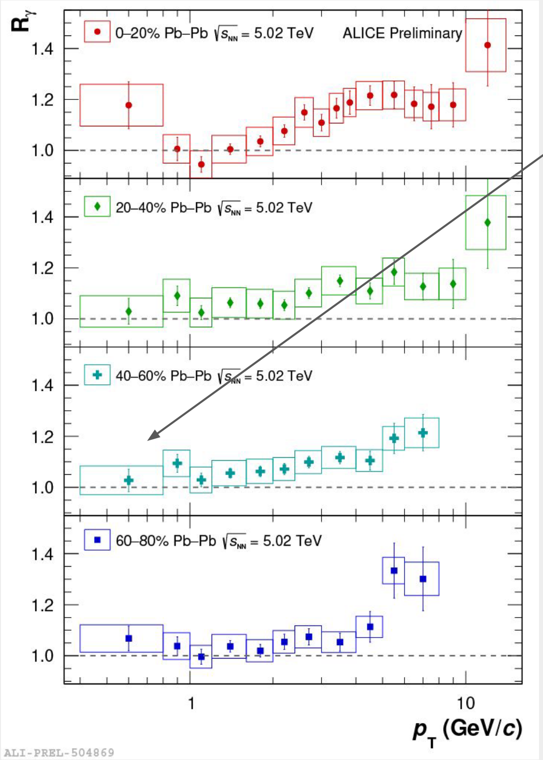
Some differ more, others less

Motivation

$\gamma_{direct} = \gamma_{inc}(1\text{-}1/R_\gamma)$

Where R_γ
$= \gamma_{inclusive}/\gamma_{decay}$
$= \gamma_{inc}\pi^0_{meas}/\gamma_{decay}\pi^0_{sim}$

Direct photon signal if: $R_\gamma > 1$



Upper limits (90% CL) are given where γ_{direct} is consistent with 0

Reduction of systematic uncertainties is essential for obtaining a significant result.

Can ML algorithms bring an improvement?

Photon training setup

Signal (γ) => **kind = 0**
Background => **kind != 0**

test_size=0.3,

kind : target variable defined based on MC truth, classifying photons according to their sources.

Training in p_T bins and centrality 0-10%

Signal = Primary photons (kind 0)
Background = Combinatorial unassociated pairs (kind = 1)
Secondary photons (kind = 5),
Dalitz decay (kind = 3) etc

PRODUCTION NAME	GENERATOR / DESCRIPTION	N_TrueGamma 0-10%	N_Event 0-10%
LHC 20e3a	HIJING MB	**0.2M**	**~3M**
LHC 20e3b	HIJING 0-10	**~5M**	**~55M**
LHC 24a1	HIJING + custom flat p_T	**~16.5M**	**~21M**

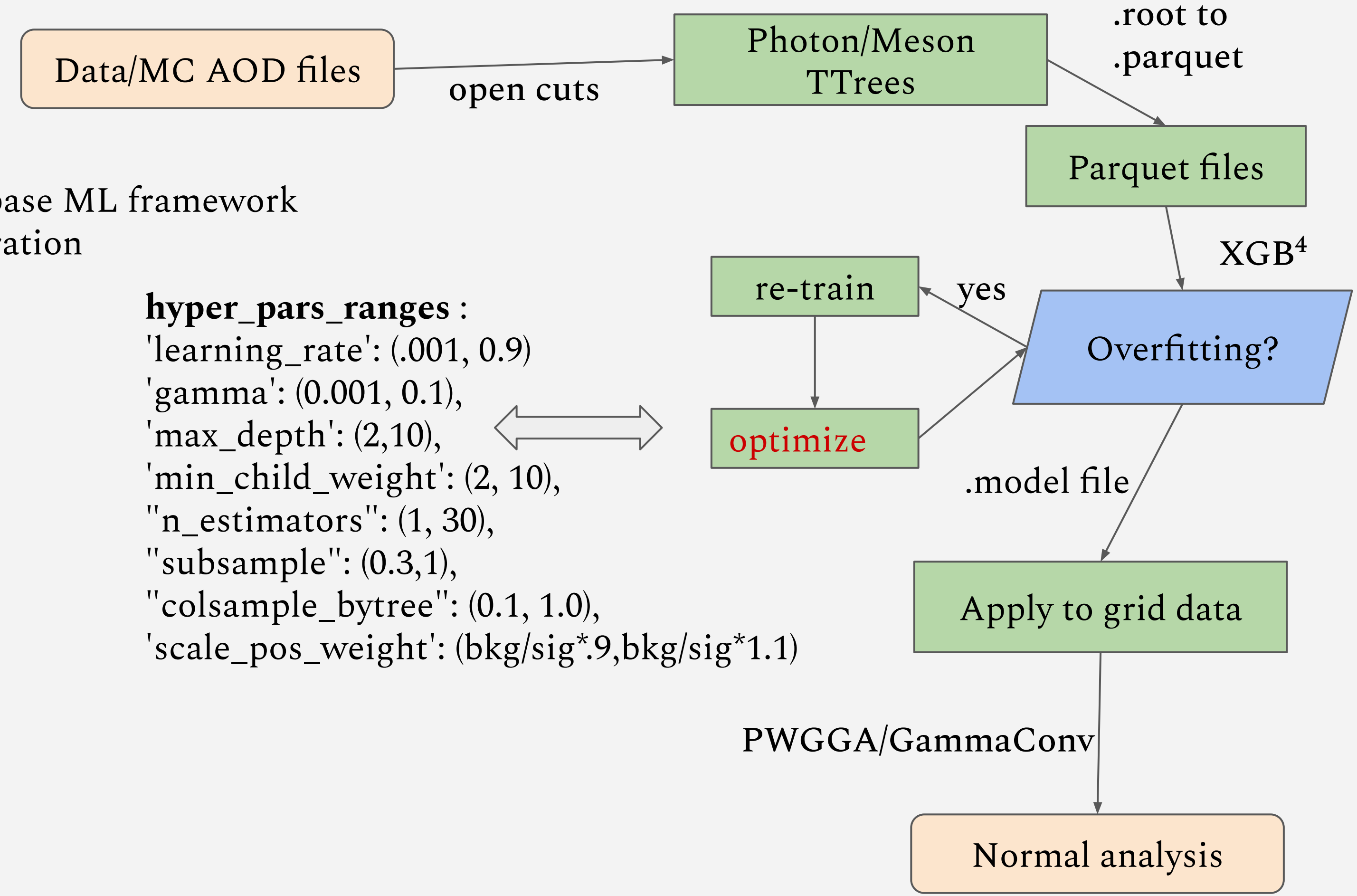
Train+Test on MC only

Workflow details with XGboost

scikit-learn[1], optuna[2] for base ML framework
hipe4ml[3] for ROOT integration

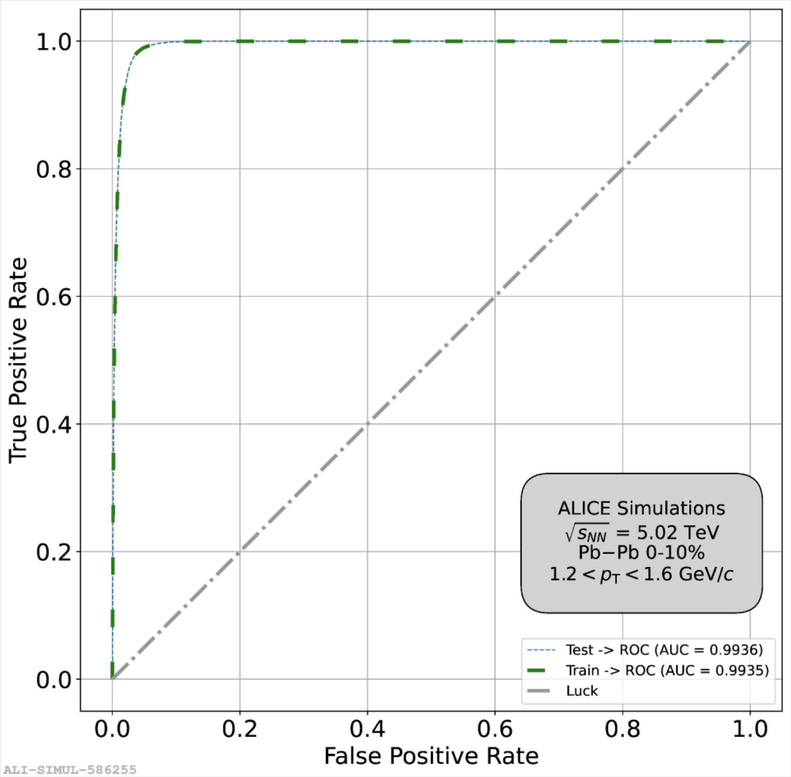
all are open sourced

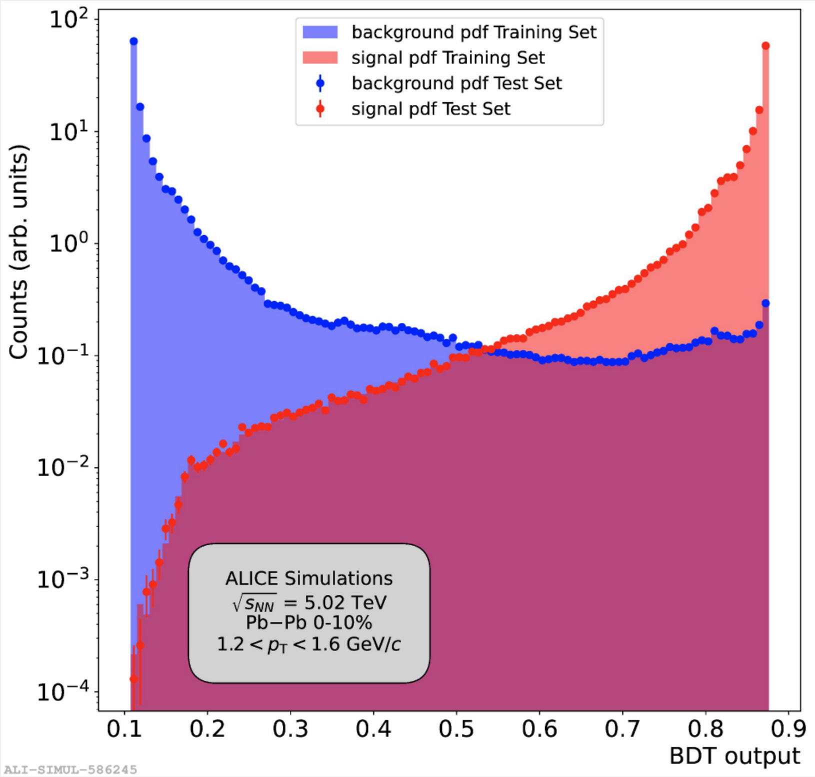
hyper_pars_ranges :
'learning_rate': (.001, 0.9)
'gamma': (0.001, 0.1),
'max_depth': (2,10),
'min_child_weight': (2, 10),
"n_estimators": (1, 30),
"subsample": (0.3,1),
"colsample_bytree": (0.1, 1.0),
'scale_pos_weight': (bkg/sig*.9,bkg/sig*1.1)



1. https://scikit-learn.org/stable/
2. https://optuna.org/
3. https://hipe4ml.github.io/
4. https://xgboost.readthedocs.io

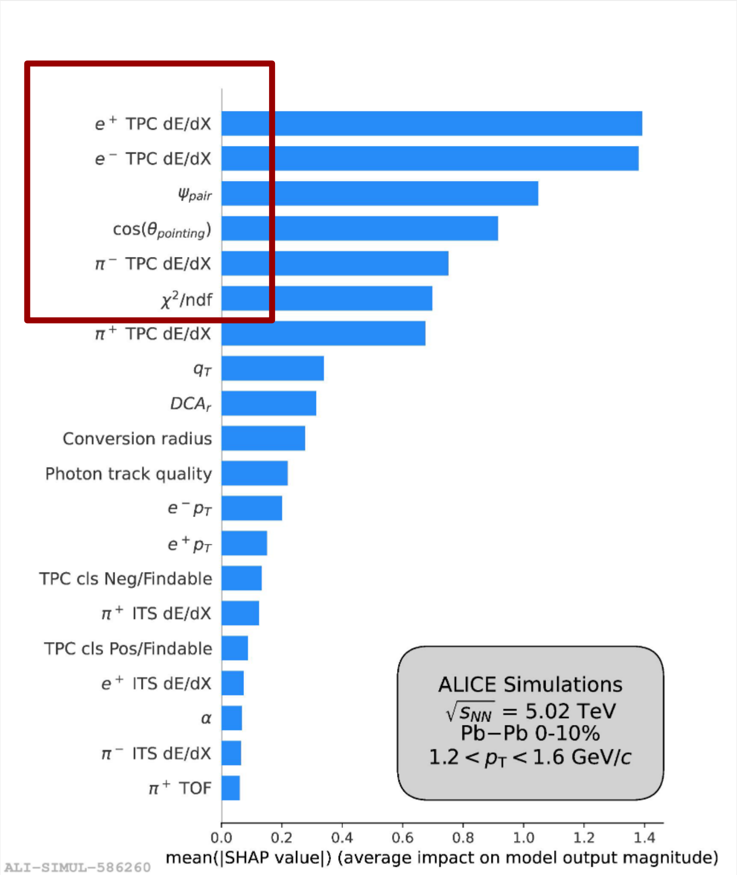
Photon training





Various XGB models based on BDT output score cuts are tested to find optimal efficiency and purity combination

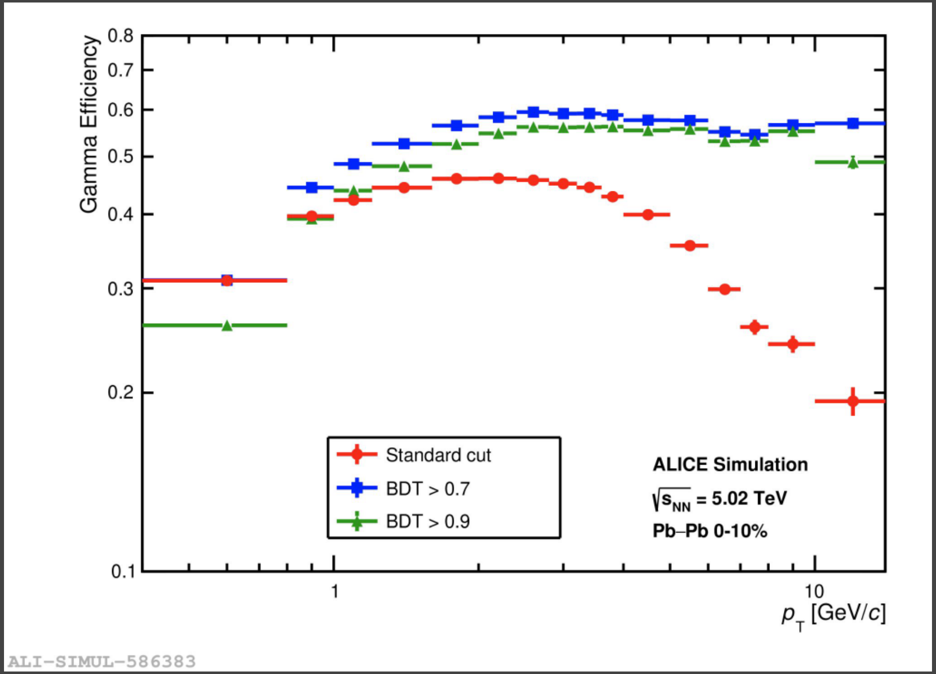
Feature importance

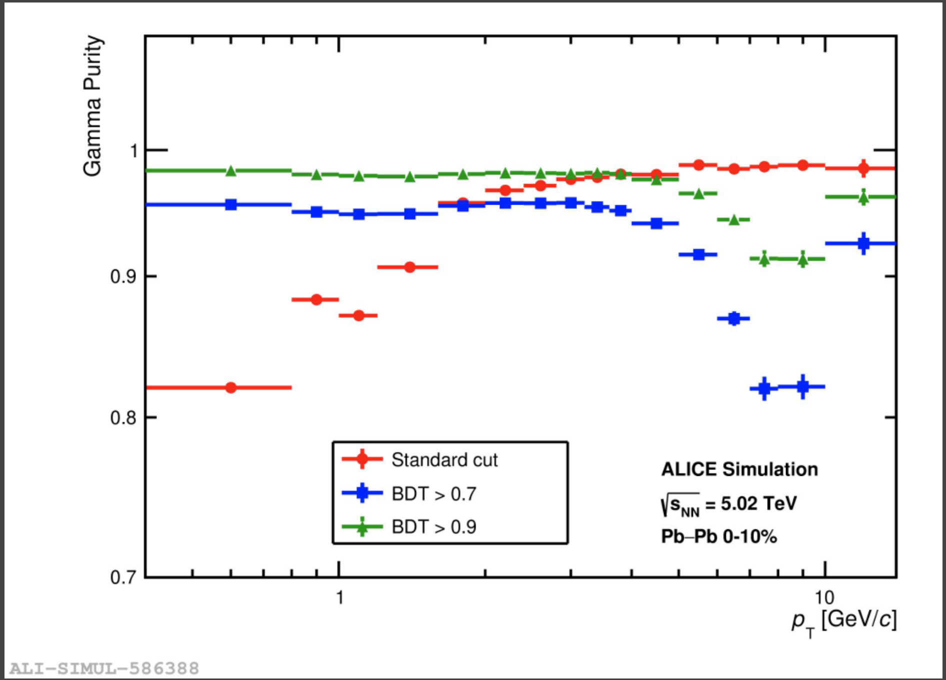


Importance as such not surprising, but are there correlations?

Results

Efficiency & purity across BDT

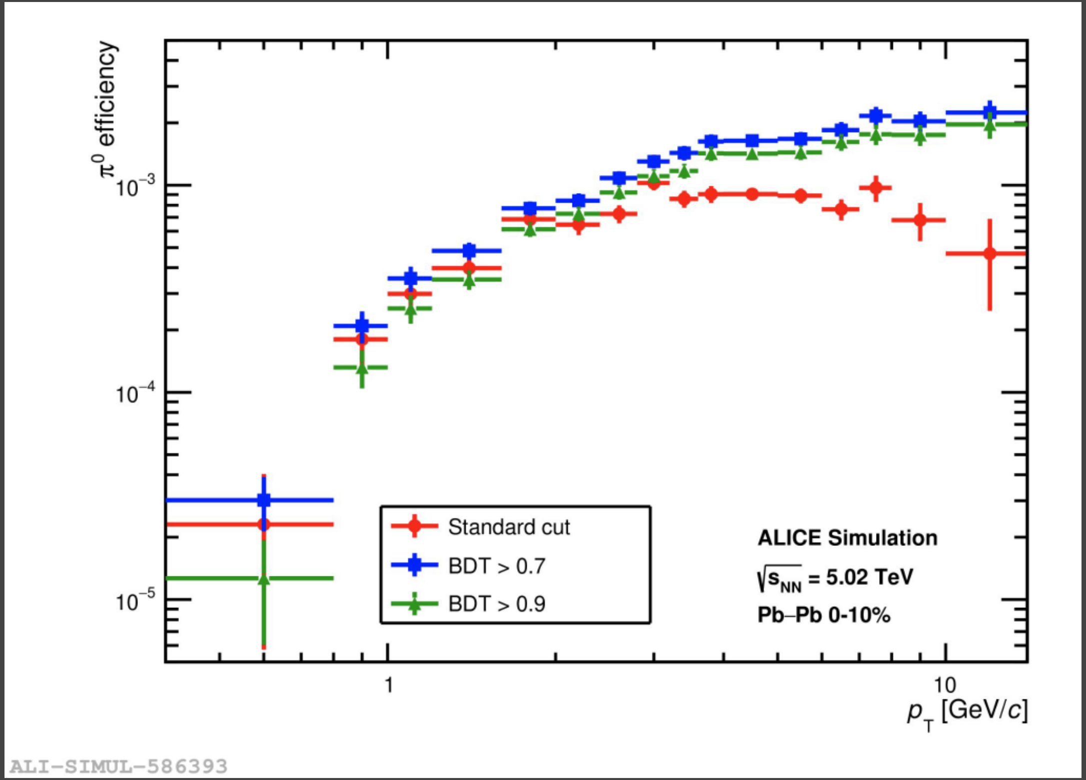




higher purity and efficiency achieved for p_T < 4 GeV/c compared to cut based analysis

MC only

Pion efficiency comparison across BDT



Pion efficiency using the converted photons selected by the XGB model.

Significant improvement over large p_T range [from 2 GeV/c and above]

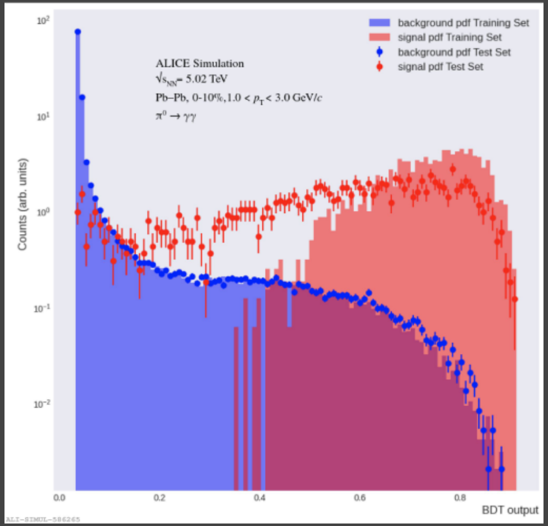
MC only

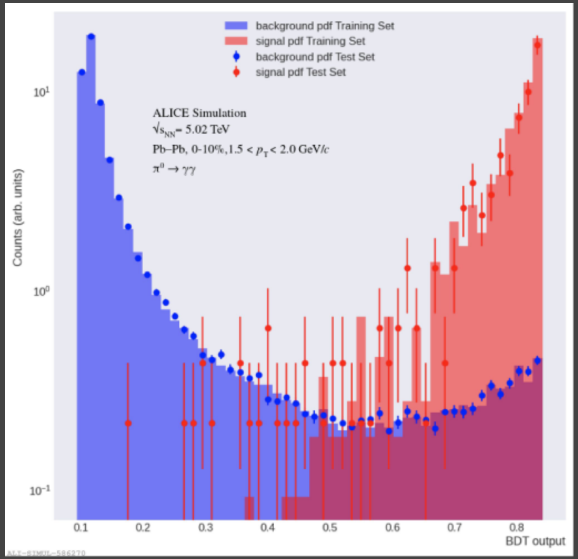
XGB on neutral meson

p_T GeV/c	0-0.5	0.5-1.0	1.0-1.5	.1.5-2.0	2.0-3.0	3.0-5.0	5.0-10.0
Signal (MC)	**43**	**1396**	**1037**	**1620**	**1026**	**574**	**181**
Background (data)	**167977**	**174189**	**174189**	**543711**	**60676**	**7659**	**658**
factor	**~3948**	**~1203**	**~168**	**~335**	**~59**	**~13**	**~6**

Large imbalance in the signal (π^0) with respect to background ($\gamma\gamma$ pairs).

The imbalance comes mainly because of lack of MC





After retraining with class-weights

MC only

Summary and outlook

- Standard analysis suggest purity and efficiency to be one of the top contributor for systematic uncertainties
- An improved efficiency and purity combination is achieved for converted photons at $p_T < 4$ GeV/c which will reduce the error estimates for direct photons in similar p_T range. Work on this is ongoing
- A significant increase in meson efficiency is also found from 2 GeV/c onwards
- The current focus is to look for higher efficiency and purity at high p_T region using XGB models at photon level and at meson level.
  Results for 30-50% centrality are also being explored.
  Beside XGB, incorporating other models in the analysis are also being explored

Backups

Features description:

$q_T = p_e x \sin(\theta_{v0,e})$:
Transverse projection of daughter particle ($e^{+/-}$) momentum to mother particle candidate (V_0) a.k.a. secondary vertex

η = Pseudorapidity of V^0 and tracks

$\psi_{pair} = \arcsin(\Delta\theta/\xi_{pair})$:
Angle between the plane of the electron and positron pair and the bending plane of the magnetic field

$\Delta\theta = \theta_{e-} - \theta_{e+}$ (polar)
$\xi_{pair} = \arccos[(\vec{p}_{e-} \cdot \vec{p}_{e+})/(|p_{e-}|.|p_{e+}|)]$

ϕ = angle between the x (left) and y-axis (up), counts clockwise along the beam direction of V^0

Features description:

Conversion Radius :
line connecting primary vertex to secondary vertex (V^0)

cos($\theta_{pointing}$) :
angle between the total momentum of the track pair and the line connecting primary and secondary vertex

χ^2/ndf :
associated with reconstruction algorithm, based on the Kalman filter (χ^2, no. of degrees of freedom == ndf)

Photon track quality :

both tracks are TPC only	==	1
one track is TPC only	==	2
both track have more than 1 ITS cluster	==	3

Features description:

TPC [Pos/Neg]/Findable cluster :
associated TPC clusters over the total number of theoretically findable clusters of a [pos/neg] track

ITS/TPC dEdX = nσ of the ITS/TPC signal
TOF = nσ of TOF signal

$\boldsymbol{\alpha = (p_L^{e+} - p_L^{e-})/(p^{e+} + p^{e-})}$ **:**
longitudinal momentum asymmetry between the secondary tracks

DCA_r :
distance of closest approach between V^0 and primary vertex (radially)

P^{e+} asymmetry = p^{e+}/p **:**
ratio of positive track momentum to net track pair momentum