Test towards a new training of the track/shower BTD algorithm in Pandora

Mattia Sotgia*, Alice Campani** (University of Genoa and INFN), Angela Fava (FNAL)

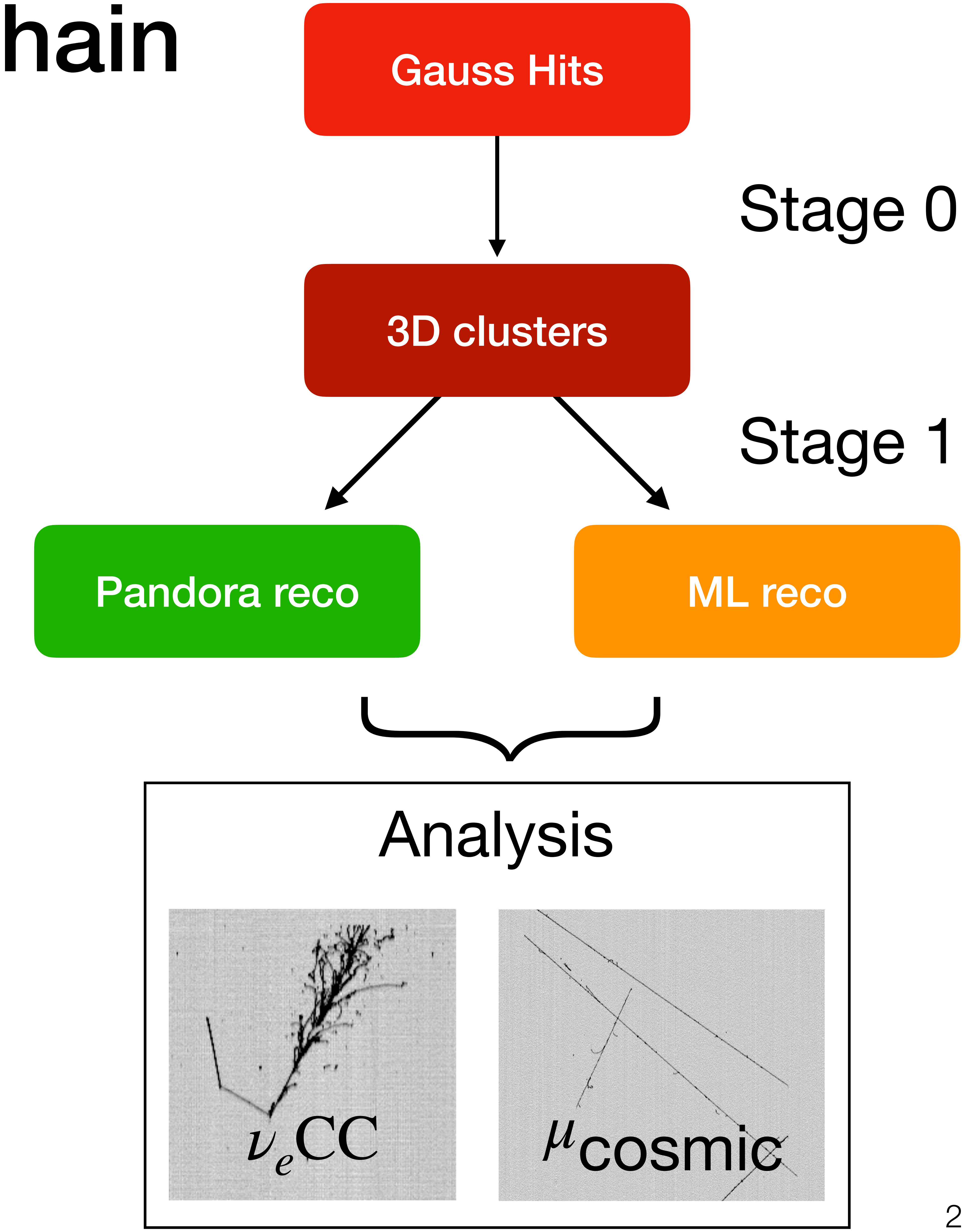
Midterm internship report
(Aug. 23rd, 2024)

Fermilab

Event reconstruction in LArTPCs: ICARUS event reconstruction chain

ICARUS analysis is performed trough a chain of subsequent algorithms, performing all the steps from clustering of hits (portions of waveforms with a signal) in 3D to reconstructing the event hierarchy.

From the **Hits** of the single wires, the **2D reconstruction** for each wire plane is performed, and then the **3D Clusters** are made. These then are used as inputs for two different reconstruction algorithms.



Gauss Hits

Stage 0

3D clusters

Stage 1

Pandora reco

ML reco

Analysis

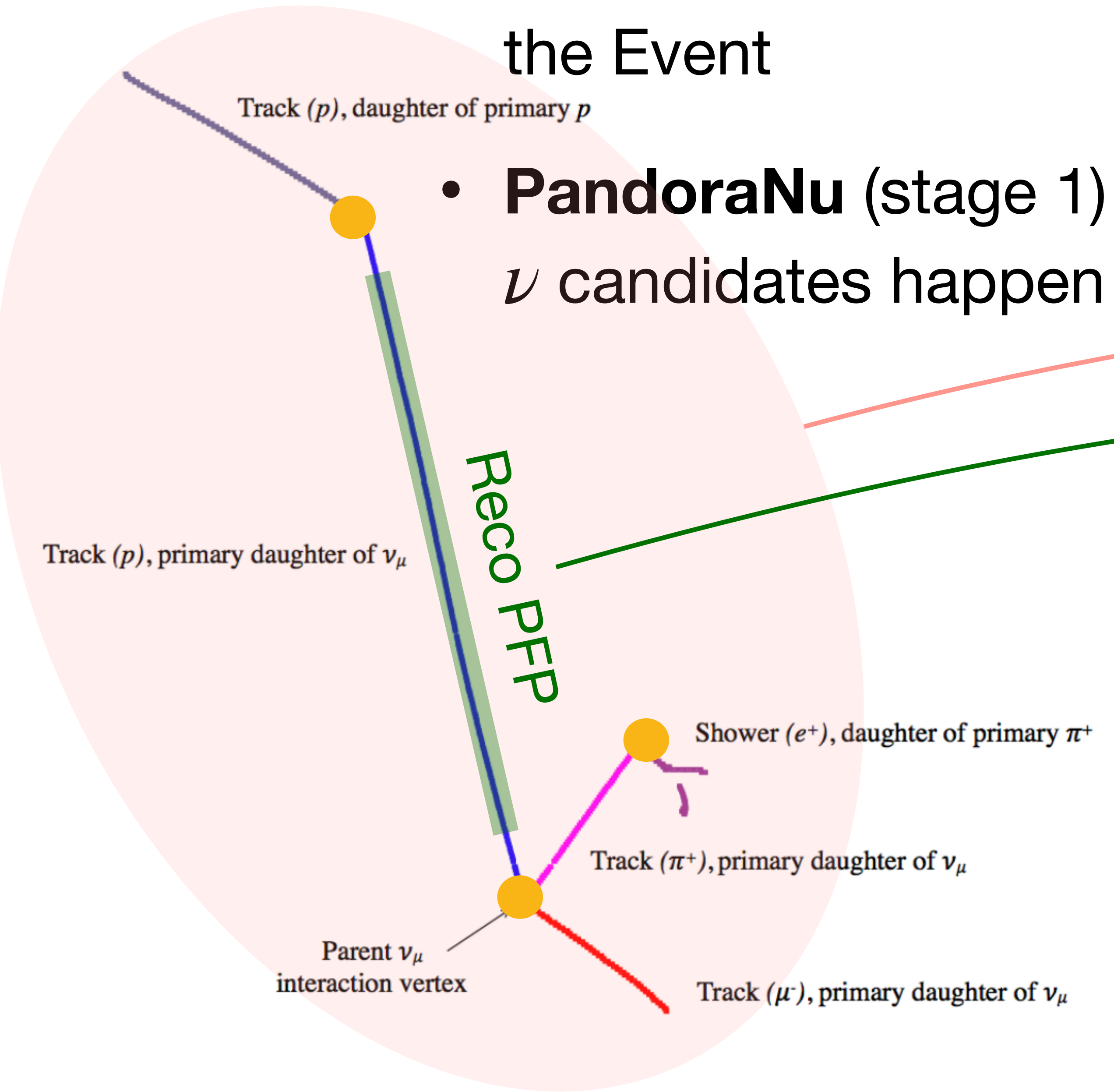
Event reconstruction in LArTPCs: ICARUS event reconstruction chain

ICARUS implements the Pandora-based reconstruction algorithm

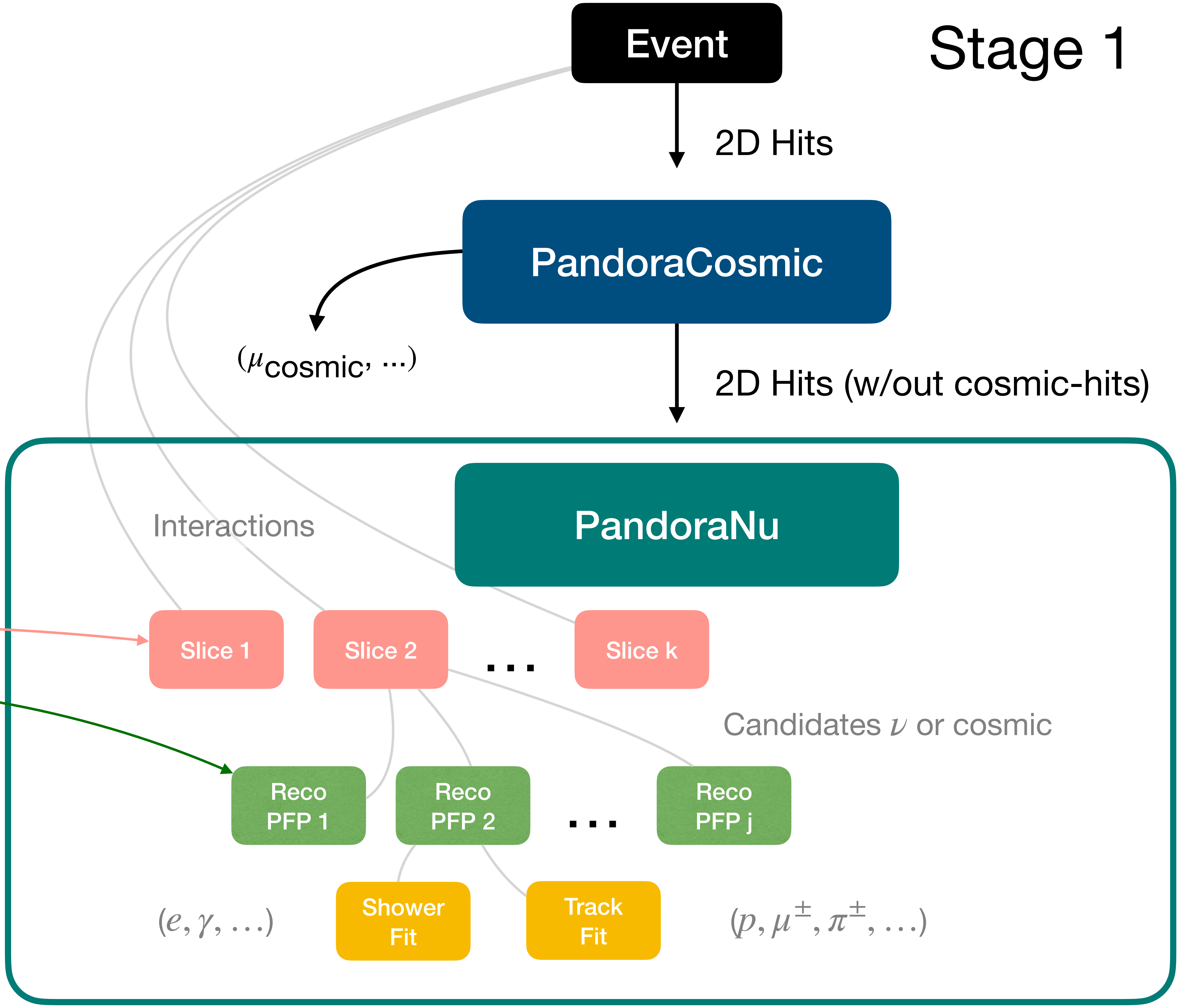
- Based on *clusters, slices* (reconstructed interactions, i.e. groups of particles linked with the same interaction) and pattern recognition

There are two main stages of the reconstruction

- **PandoraCosmic** (stage 0) where the cosmic-like hits are clustered from the 2D hits and are separated from the Event
- **PandoraNu** (stage 1) where the reconstruction of the ν candidates happen


Track (p), daughter of primary p
Track (p), primary daughter of ν_μ
Reco PFP
Shower (e^+), daughter of primary π^+
Track (π^+), primary daughter of ν_μ
Parent ν_μ interaction vertex
Track (μ^-), primary daughter of ν_μ

Boosted Decision Trees (BDTs) are used **1.** in candidate ν/cosmic selection, **2.** in finding the true interaction vertex and **3.** in the track/shower discrimination



Stage 1

Event

2D Hits

PandoraCosmic

$(\mu_{\text{cosmic}}, \ldots)$

2D Hits (w/out cosmic-hits)

Interactions

PandoraNu

Slice 1

Slice 2

. . .

Slice k

Candidates ν or cosmic

Reco PFP 1

Reco PFP 2

. . .

Reco PFP j

$(e, \gamma, \ldots)$

Shower Fit

Track Fit

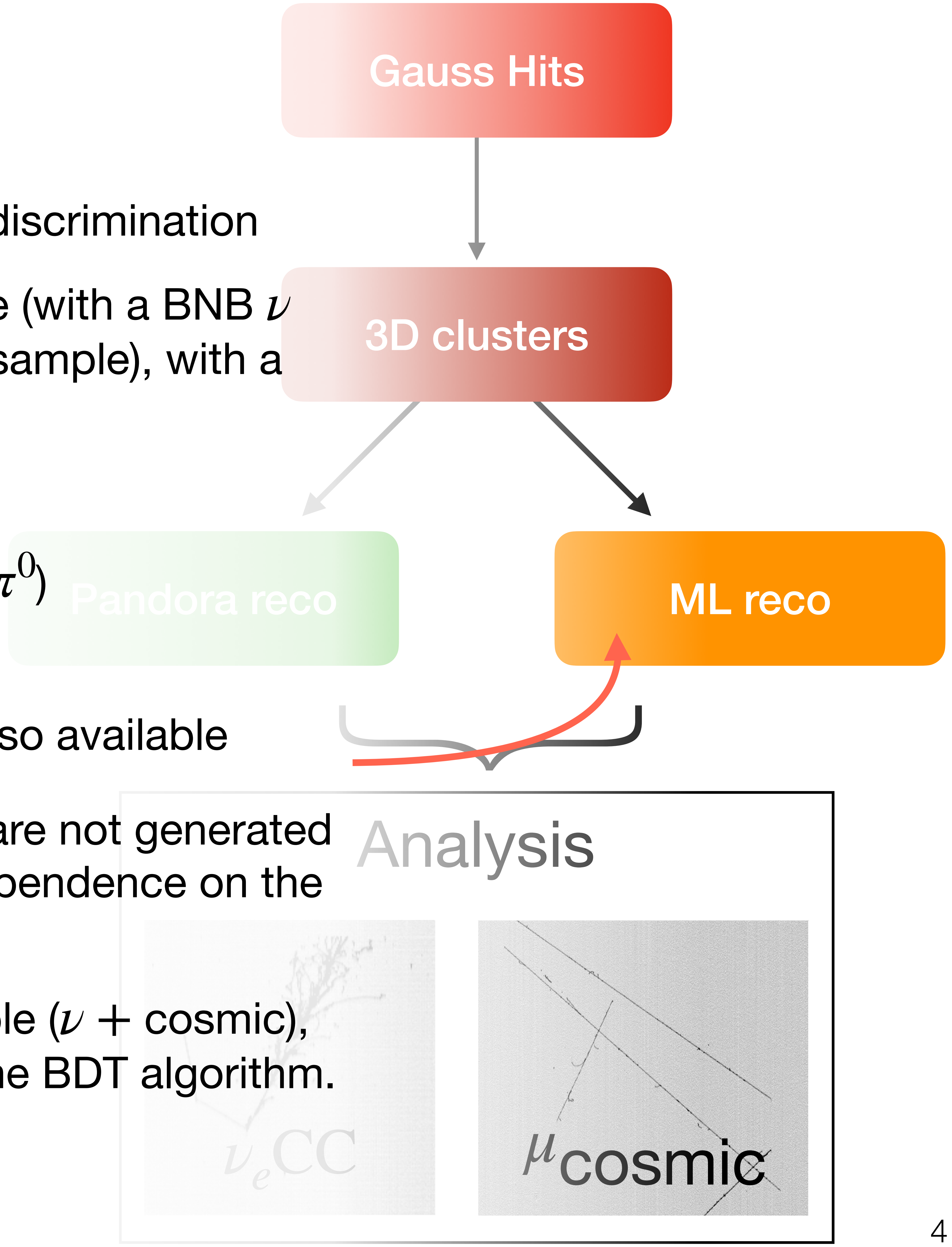
$(p, \mu^{\pm}, \pi^{\pm}, \ldots)$

Track/shower discrimination BDT: testing a new dataset

- The third BTD algorithm is the one responsible for the shower/track discrimination
- The BDT algorithm was last trained on a previous version of the code (with a BNB ν-only sample). This outperformed the precedent training (SBND MC sample), with a classification efficiency of ~80 %.
  - Good performance with track like particles $(p, \pi^{\pm}, \mu^{-})$
  - A slight decrease in performance for shower like particles (e^{-}, γ, π^{0})

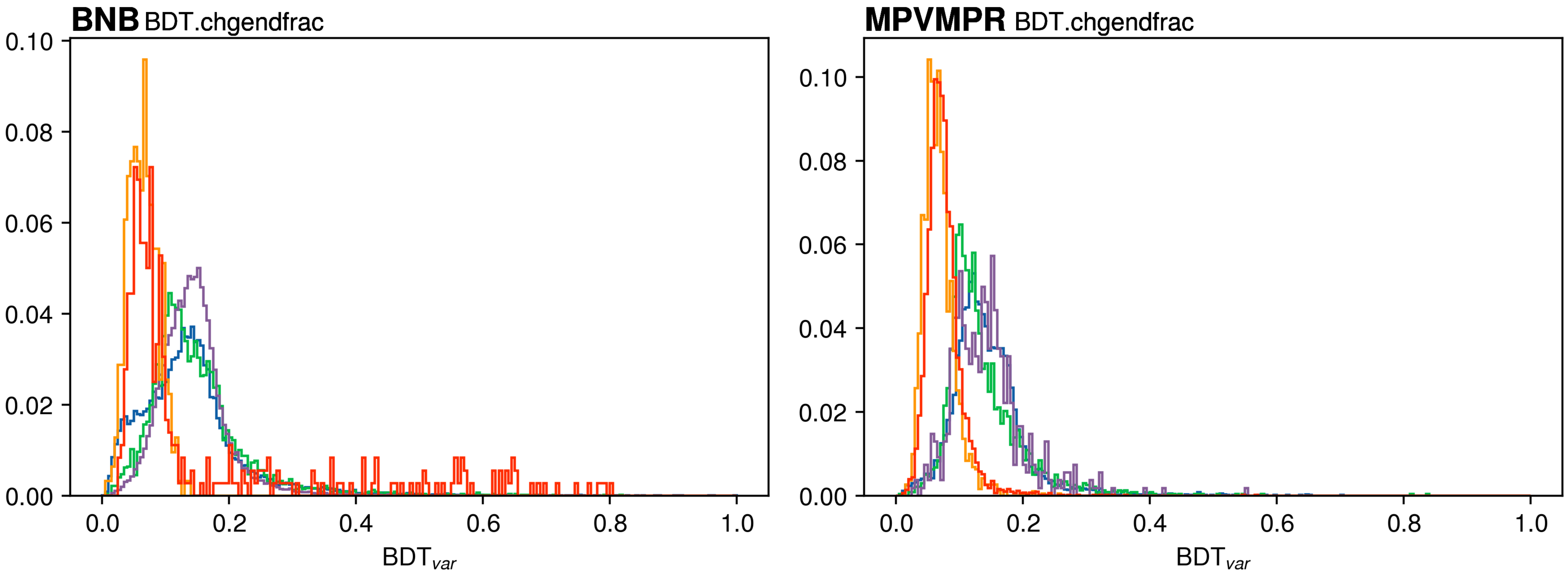
See SBN-doc-34318-v2 for further details

- A new MC dataset from the ICARUS ML working group was made also available
  - **Less biased** than a BNB/NuMI beam simulation, since particles are not generated from the BNB beam but with a uniform energy distribution (no dependence on the signal model)
- In this talk we compare two samples, BNB (ν-only) and the ML sample (ν + cosmic), with the aim of finding the most suitable to perform a re-training of the BDT algorithm.



Comparing the BNB MC and ML MPVMPR MC datasets
Charge end fraction

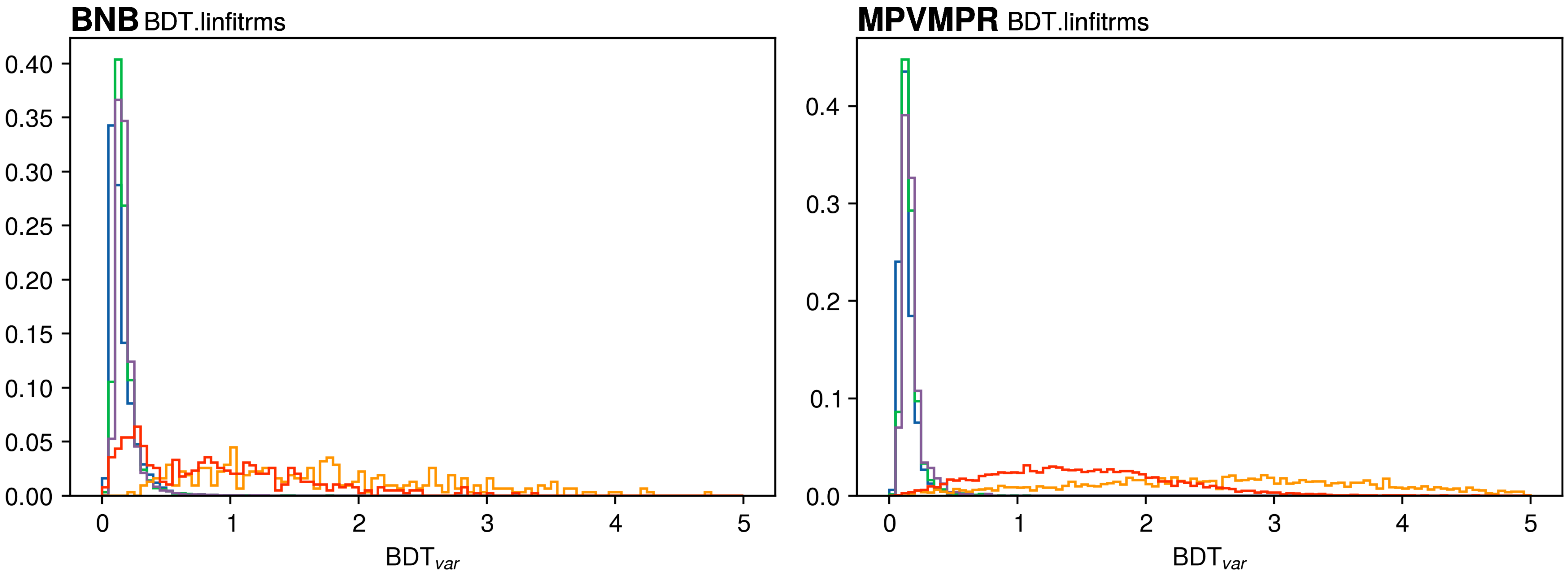
- The distributions are more peaked in the ML sample than those in the BNB sample, though the separation is actually similar.
- The overall shape is more 'Gaussian-like', as no weird features are visible in the ML data sample.



Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles

Comparing the BNB MC and ML MPVMPR MC datasets Linear fit RMS

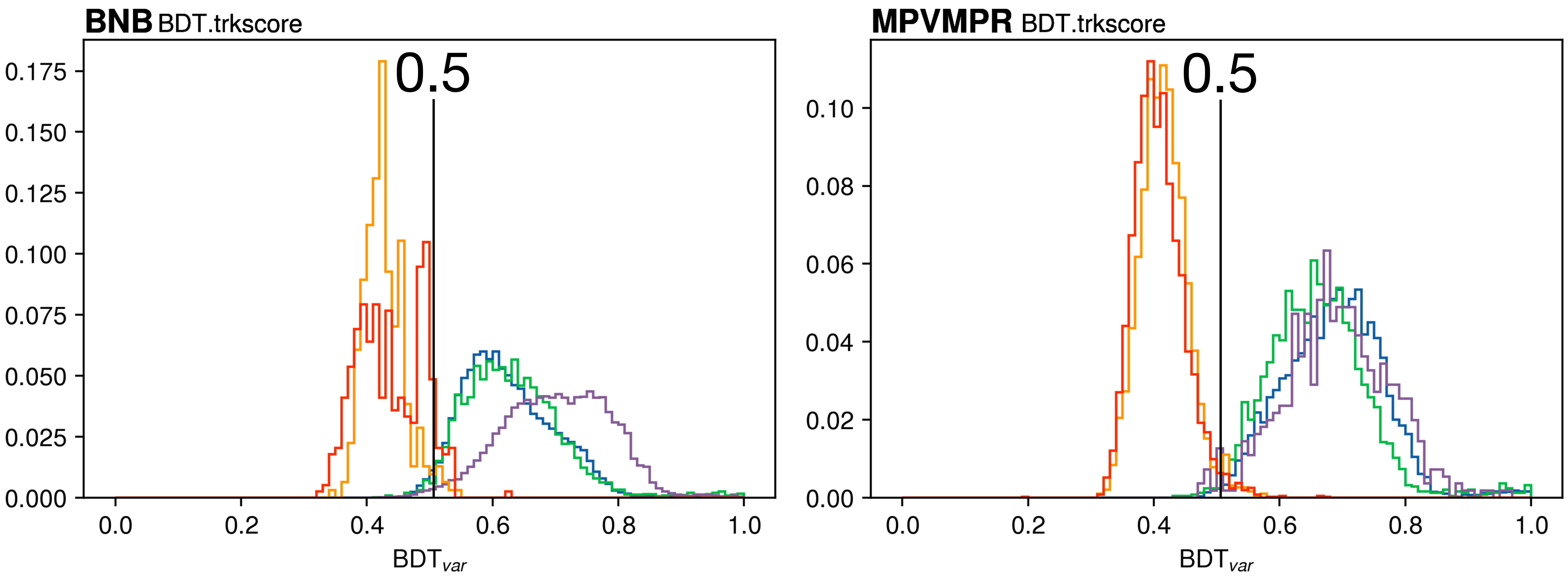
- A greater number of shower like events in the MPVMPR sample
- Fewer contaminations of shower like particles in the area populated mostly by track like particles



Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles

Comparing the BNB MC and ML MPVMPR MC datasets BDT Track Score

- Showing greater separation (although this variable is actually the output of the previous training on both samples and needs to be evaluated after the re-training)
- Overall with the current training of the algorithm the MLM MPVMPR sample is promising



Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles

Looking forward:
Next steps

- The ML MPVMPR sample has shown **greater discrimination power** in some BDT variables
- The ML MPVMPR sample has also a better balance between track-like and shower-like particles contribution
  - The track/shower ratio is lower than the BNB MC sample

$$\text{ratio}_{\text{MPVMPR}} = \frac{\text{\#track-like}}{\text{\#shower-like}}\bigg|_{\text{MPVMPR}} \simeq \frac{9052}{6978} \simeq 1.3$$

$$\text{instead of ratio}_{\text{BNB}} = \frac{\text{\#track-like}}{\text{\#shower-like}}\bigg|_{\text{BNB}} \simeq \frac{64972}{704} \simeq 92.3$$

  - The MPVMPR sample analyzed consist of 325535 events, which is overall lower in respect to the event count of the BNB MC
  - In the view of the training a new MC sample with larger statistic has been produced. The sample contains roughly 200 000 tracks and 150 000 showers per cryostat

Looking forward: Next steps

- A new training will be performed with all the BDT variables, with the ML MPVMPR sample → Testing to be done on the BNB sample
- Some of the variables showed a less than acceptable discrimination power
  - Vertex distance
  - Conicalness
  - Concentration
  - Halo total ratio
  - Linear fit length

A new training will be also performed **without** these 5 variables

Thank you for the attention!

Test towards a new training of the track/shower BTD algorithm in Pandora

Mattia Sotgia, Alice Campani (University of Genoa and INFN), Angela Fava (FNAL)
ICARUS ML Working Group meeting
(Aug. 21st, 2024)

*msotgia@ge.infn.it, **acampani@ge.infn.it

Fermilab

Backup Slides

Backup 1:
Definition of hit purity and completeness

Compare MC particles and reconstructed PFPs (Particle Flow Particles, Pandora Objects)

Definitions

Matched hits $\equiv \text{hits}_{\text{MC particle}} \cap \text{hits}_{\text{reco pfp}}$.

For the example on the side Matched hits$_j \rightarrow 6$ and Matched hits$_k \rightarrow 2$

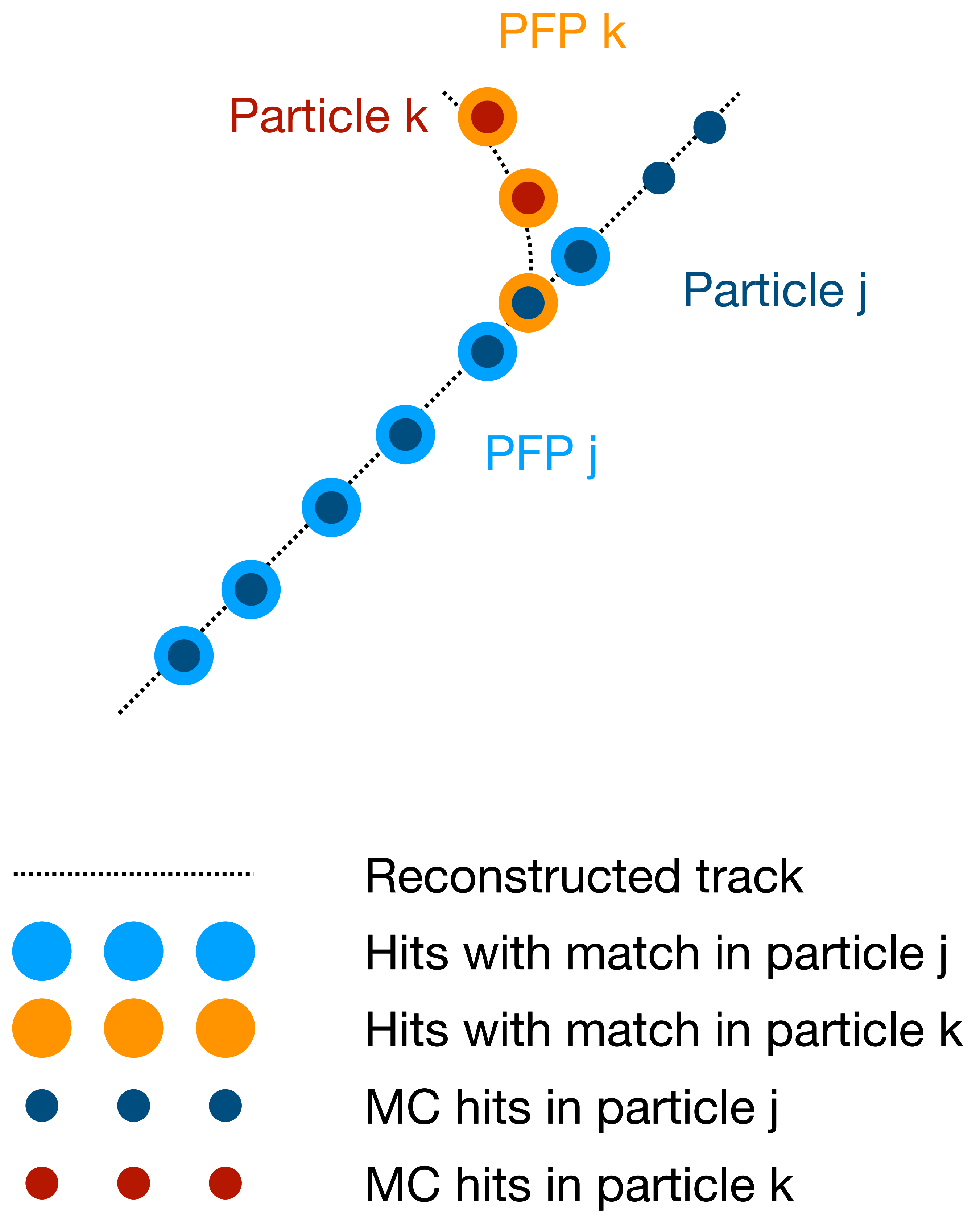
Purity $\equiv \dfrac{\text{hits}_{\text{MC particle}} \cap \text{hits}_{\text{reco pfp}}}{\text{hits}_{\text{reco pfp}}}$.

For the example on the side Purity$_j \rightarrow \dfrac{6}{9} \simeq 67\,\%$ and Purity$_k \rightarrow \dfrac{2}{9} \simeq 22\,\%$

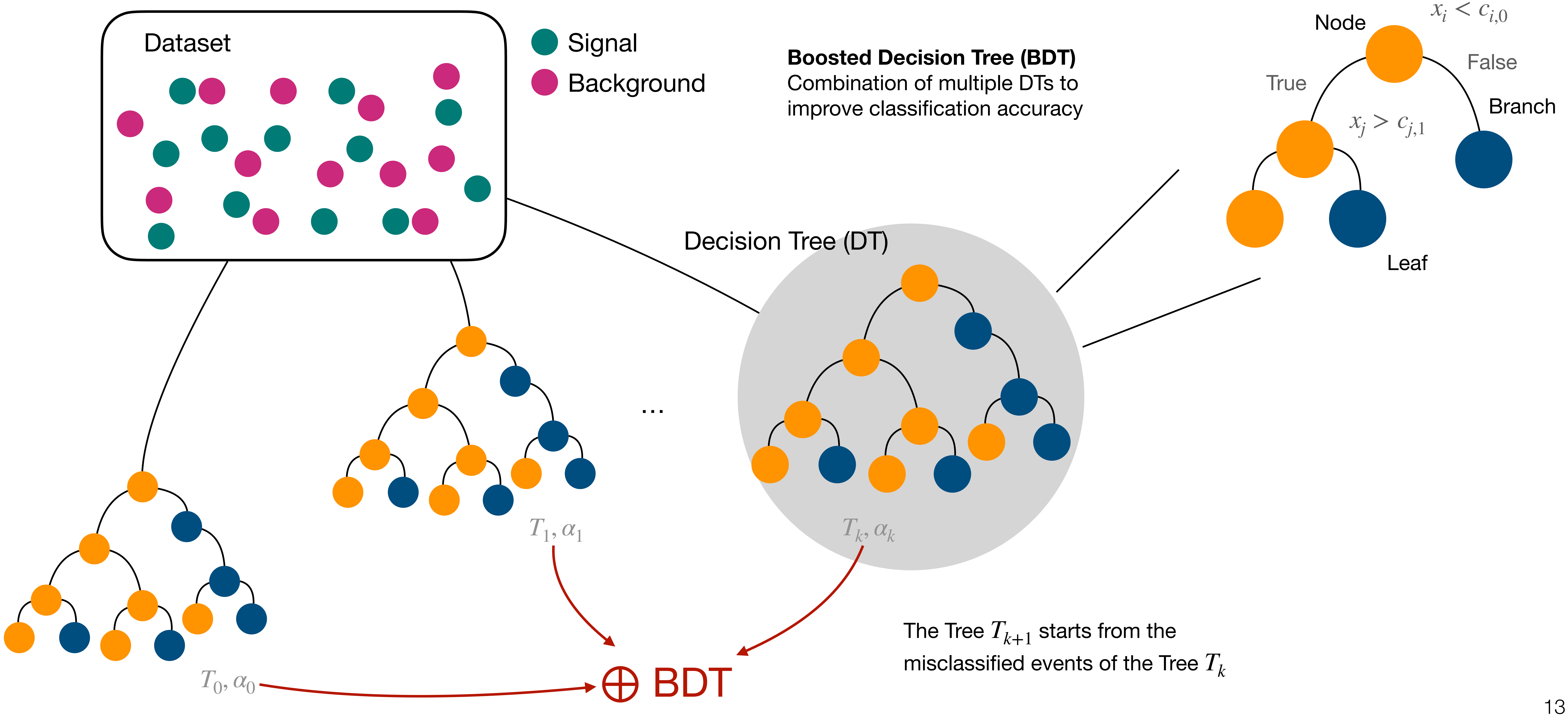
Completeness $\equiv \dfrac{\text{hits}_{\text{MC particle}} \cap \text{hits}_{\text{reco pfp}}}{\text{hits}_{\text{MC particle}}}$.

For the example on the side Completeness$_j \rightarrow \dfrac{6}{9} \simeq 67\,\%$ and

Completeness$_k \rightarrow \dfrac{2}{2} \simeq 100\,\%$



Backup 2:
The Boosted Decision Tree algorithm



Boosted Decision Tree (BDT)
Combination of multiple DTs to improve classification accuracy

The Tree T_{k+1} starts from the misclassified events of the Tree T_k

BDT variables:
charge variables and cone charge variables

The current version of the BDT track/shower algorithm implements 13 variables (hyper parameters) to perform the cuts of the decision tree

All the BDT charge variables are computed on the Hits of the induction 1 wire plane. The other make use of the full 3D information from the reconstructed PFFs.

The first two are the ‘charge-based variables’.

1. **Charge end fraction (BDT.chendfrac)**, defined as the ratio of the deposited charge in the last 10% of the PFP hits, over the total deposited charge. Tracks are expected to have a more uniform charge distribution than showers. For this variable the expected values are in the range [0, 1]. Smaller values mean a less uniform charge distribution trough the length of the pfp.
2. **Charge fraction spread (BDT.chfracspread)**, defined as the ratio of the variance of the deposited charge of the hits to the deposited charge mean value. Showers are expected to have a more spread variety of charge related to the hits. It is a ratio but it is not normalized (i.e. the range is not in [0, 1]). The binning is chosen to be in [0, 2.5]. Tracks are expected in < 1, whereas showers are expected in > 1.

BDT variables:
charge variables and cone charge variables

The last update of the BDT algorithm introduced three new variables, called ‘cone charge variables’

Defining the chargeCore (the hits inside the 20% of the direction of the primary eigenvector) and chargeHalo (hits beyond the 20% threshold)

3. **Concentration (BDT.concentration)**, defined as $\dfrac{\text{chargeCon}}{\text{chargeCore} + \text{chargeHalo}}$

   Values are expected in the range [0, 100]

4. **Halo total ratio (BDT.halototratio)**, defined as $\dfrac{\text{chargeCore}}{\text{chargeCore} + \text{chargeHalo}}$, where chargeCon is the sum of the Hits inside the cone.
   It being a ratio, the values are expected in the range [0, 1].

5. **Conicalness (BDT.conicalness)**, defined as $\sqrt{\dfrac{\text{chargeConEnd}}{\text{chargeConStart}}} \Big/ \dfrac{\text{totalChargeEnd}}{\text{totalChargeStart}}$

   Its values are expected in the range [0, 600].

BDT variables:
linear and geometrical variables

There are also the linear variables

6. **Linear fit length (BDT.linfitlen)**, defined as the length of the reco particle. The long tracks ($\mu, \pi^{\pm}$) can be some ~1 m, protons are usually shorter and showers are smaller.
7. **Linear fit difference (BDT.linfitdiff)**, defined as the difference in linearity variation, between the end and the start point. This is expected to be quite small, in the range [0, 0.15] [arb. U.] both for showers and tracks.
8. **Linear fit gap length (BDT.linfitgaplen)**, defined as the gap between the hits on the linear fit. Tracks are expected to have smaller gap length. The common gap length is in the centimeters, so the range is [0, 0.5] cm.
9. **Linear fit RMS (BDT.linfitrms)**, defined as the RMS of the fit. Tracks are expected to have smaller RMS. The binning is in [0, 5], tracks are expected in [0, ~1], and showers are expected in [~1, ~5].

BDT variables: linear and geometrical variables

And also the geometrical parameters

10. **Distance from vertex (to BDT.vtxdist)**, defined as the distance from the reconstructed vertex and its closest hit. This is usually very short for tracks and normally the distance of an electromagnetic shower from the reconstructed interaction vertex is greater. The range is chosen [0, 200] cm to account for events which were otherwise not included.

12. **PCA2 ratio (BDT.pca2ratio)**, defined as the ratio of the eigenvalue v_2 over the eigenvalue v_1 obtained from the Principal Component Analysis (PCA) algorithm, describing the orientation of the hits in space.
13. **PCA3 ratio (BDT.pca3ratio)**, defines as the ration of the third eigenvalue over the first. It is expected to be a good variable, being shower more tridimensional than tracks, since this value highlight the 3D aspect of the cluster, along with the PCA2 ratio. The chosen range is to get all the events plotted.
14. **Opening angle difference (BDT.openanglediff)**, defined as $\tan^{-1}\left(\sqrt{\text{PCA2}}\sin\theta\right)$, where θ is the angle between the two eigenvectors.
The chosen range is in [0, ~35] deg, but most events are to be expected in the [0, 20] deg range.

The datasets:
MC fractional population of shower- and track-like particles

The simulations were made for the BNB dataset and for the MPVMPR dataset with the particle composition shown on the top right

Two datasets:

- **BNB MC** created for the study on Central Value (CV) systematics for Neutrino 2024 icaruspro_production_v09_89_01_01_2024A_ICARUS_MC_CV_Sys_2024A_MC_CV_Sys_flatcaf

- **MPVMPR MC** samples, produced by the ICARUS ML WG acampani_training_caf_default_v09_89_01_01_mpvmpr

BNB sample is ν-only, whereas ML MPVMPR (Multi Particle Vertex, Multi Particle Rain) is ν + cosmic.

A cut (wellRecoCut) is applied to only select **well reconstructed** particles, which have *hit completeness* and *purity* above 80% (more of their definition is in backup). This avoids biasing the result of the comparison with other mis-reconstruction effects, such as clustering issues, track-splitting, …

Shower-like

Track-like

noSpillCut

	BNB (total 213512)		**MPVMPR (total 66331)**	
	Fraction	(events)	Fraction	(events)
protons	0.527	(112594)	0.429	(28455)
charged_pi	0.121	(25925)	0.223	(14822)
muons	0.237	(50707)	0.018	(1221)
electrons	0.003	(610)	0.092	(6074)
photons	0.111	(23676)	0.238	(15759)

wellRecoCut

	BNB (total 65676)		**MPVMPR (total 16030)**	
	Fraction	(events)	Fraction	(events)
protons	0.391	(25704)	0.377	(6050)
charged_pi	0.073	(4807)	0.153	(2450)
muons	0.525	(34461)	0.034	(552)
electrons	0.005	(313)	0.119	(1911)
photons	0.006	(391)	0.316	(5067)

📚 SBN-doc-34318-v2

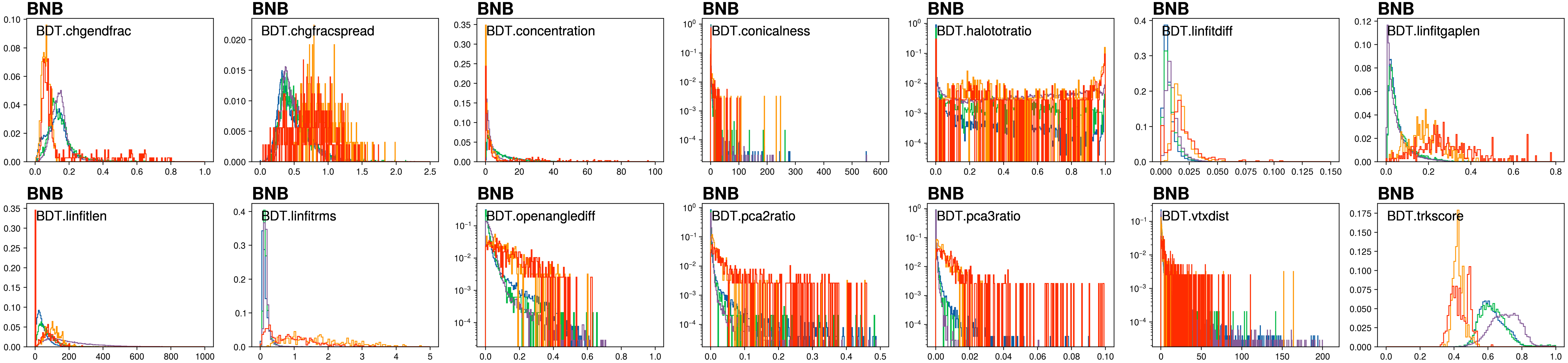
📚 Neutrino 2024

📚 ML sample update SBN-doc-35469-v1

BNB data:

Plot of BDT$_{var}$ comparing the distributions for shower/track like particles

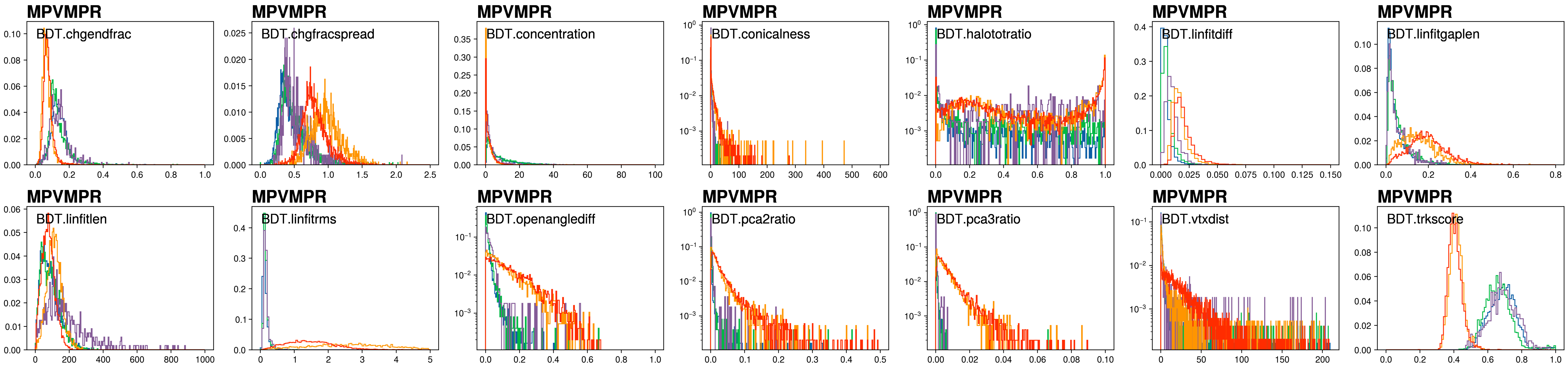
Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles



BNB data sample (samweb definition icaruspro_production_v09_89_01_01_2024A_ICARUS_MC_CV_Sys_2024A_MC_CV_Sys_flatcaf)

MPVMPR data:
Plot of BDT$_{var}$ comparing the distributions for shower/track like particles

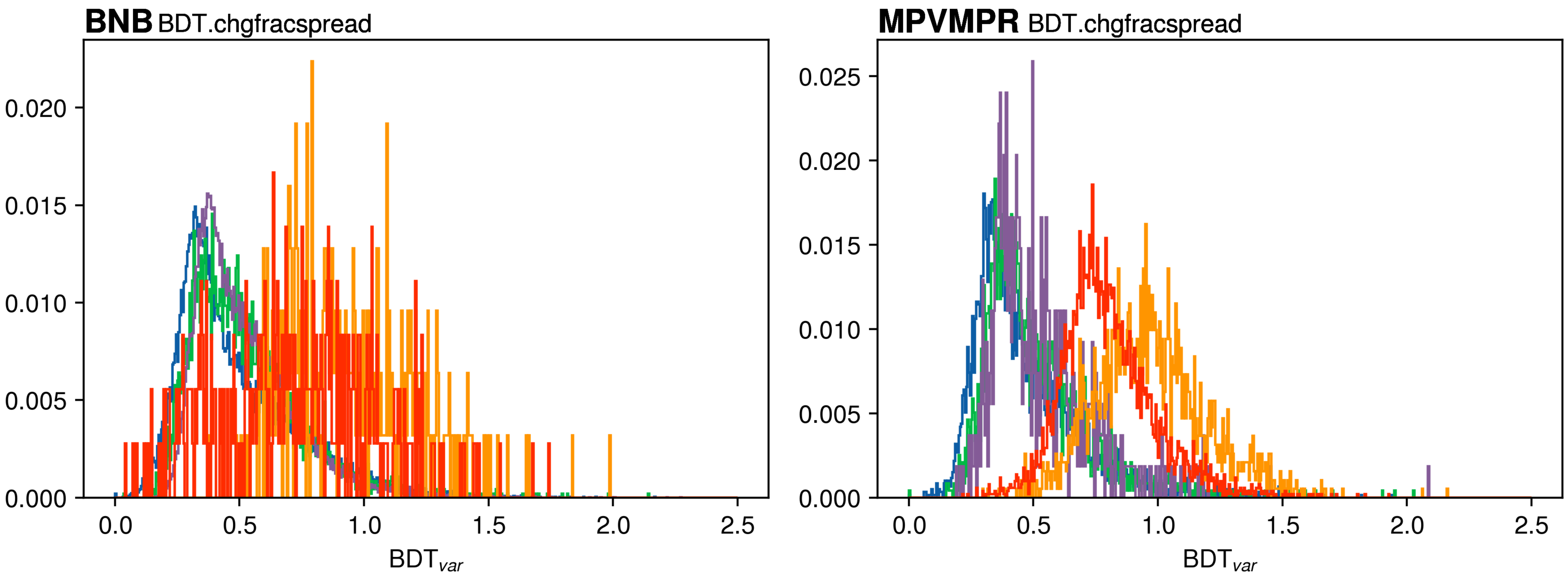
Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles



MPVMPR data sample (samweb definition acampani_training_caf_default_v09_89_01_01_mpvmpr)

Comparing the BNB MC and ML MPVMPR MC datasets Charge fraction spread

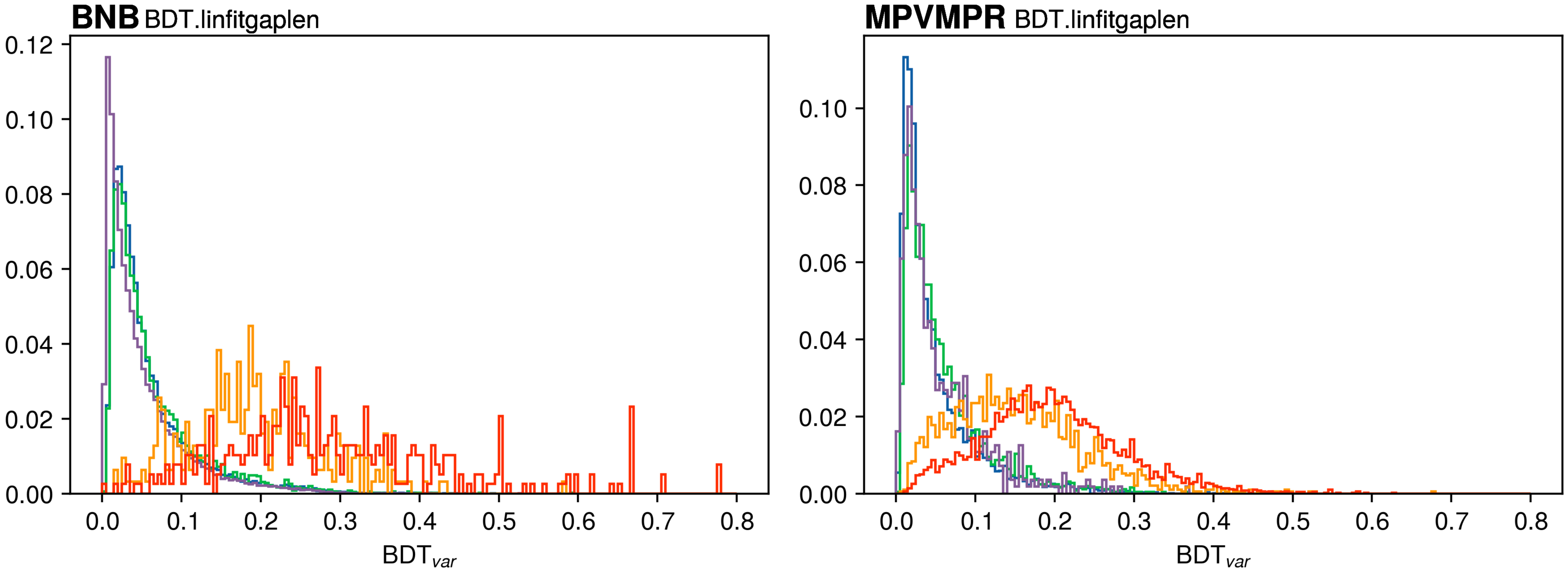
- With a larger statistics in the ML sample a better separation between the two population arises



Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles

Comparing the BNB MC and ML MPVMPR MC datasets Linear fit gap length

- With a larger statistics in the ML sample a better separation between the two population arises



Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles

Comparing the BNB MC and ML MPVMPR MC datasets PCA2 ratio and PCA3 ratio

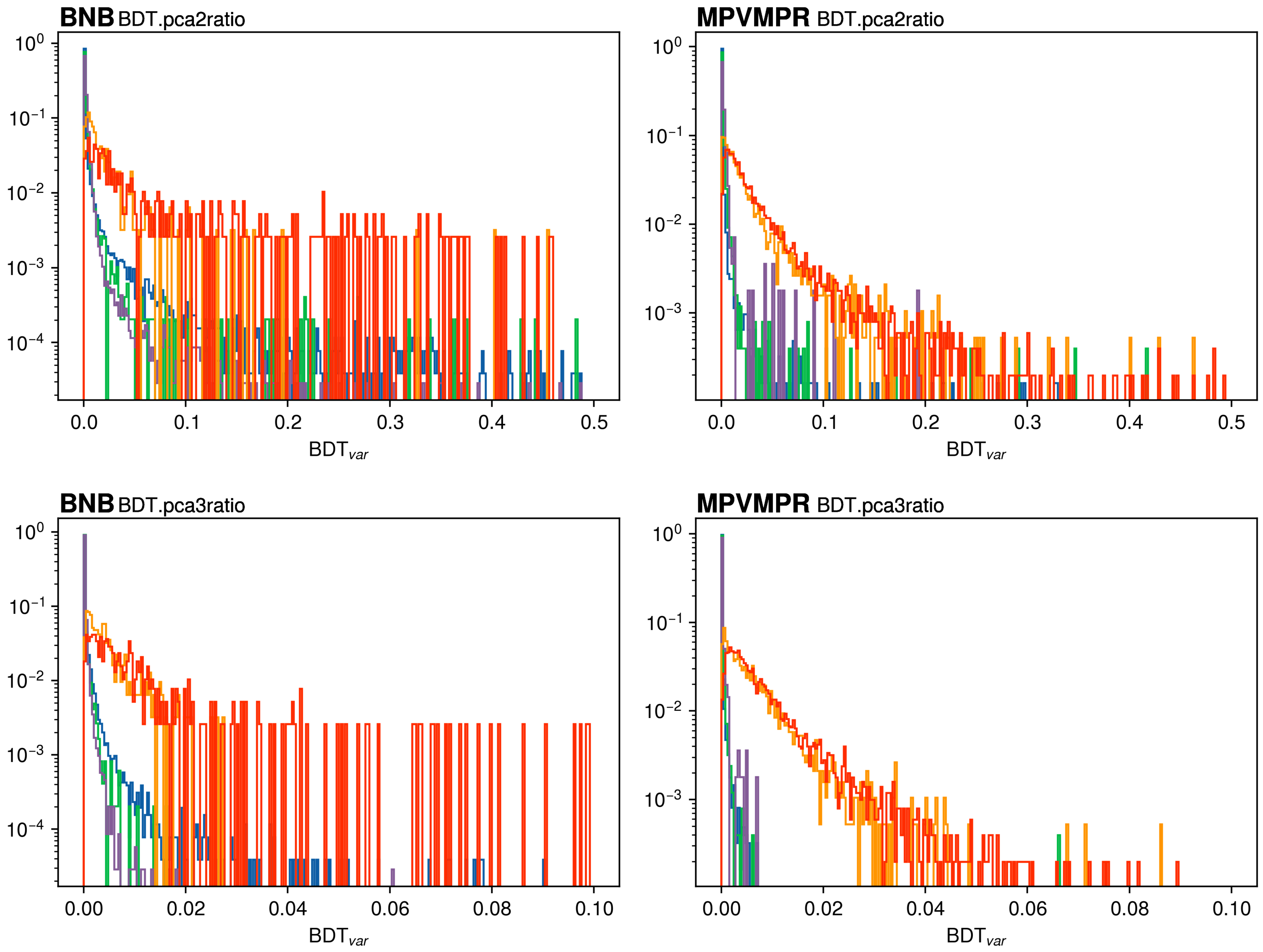
- No major difference is evident, but overall the shape has changed a bit, especially for the PCA3 ratio

- Plots are in logY scale.

- Still interesting for the possibility of a joint cut alongside another variable.

2D plots are work in progress

Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles

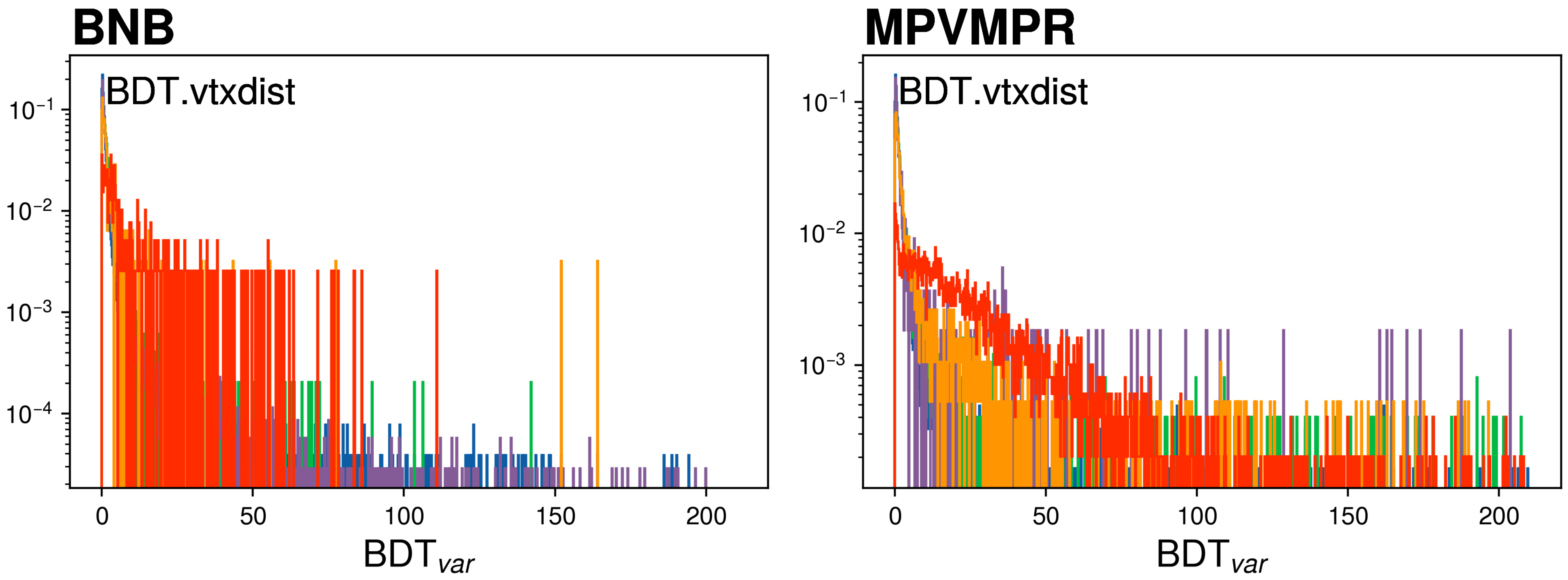


Looking forward: Next steps

- A new training will be performed with all the BDT variables, with the ML MPVMPR sample

→ Training to be done on the BNB sample

- Some of the variables showed a less than acceptable discrimination power
  - **Vertex distance**
  - Conicalness
  - Concentration
  - Halo total ratio
  - Linear fit length



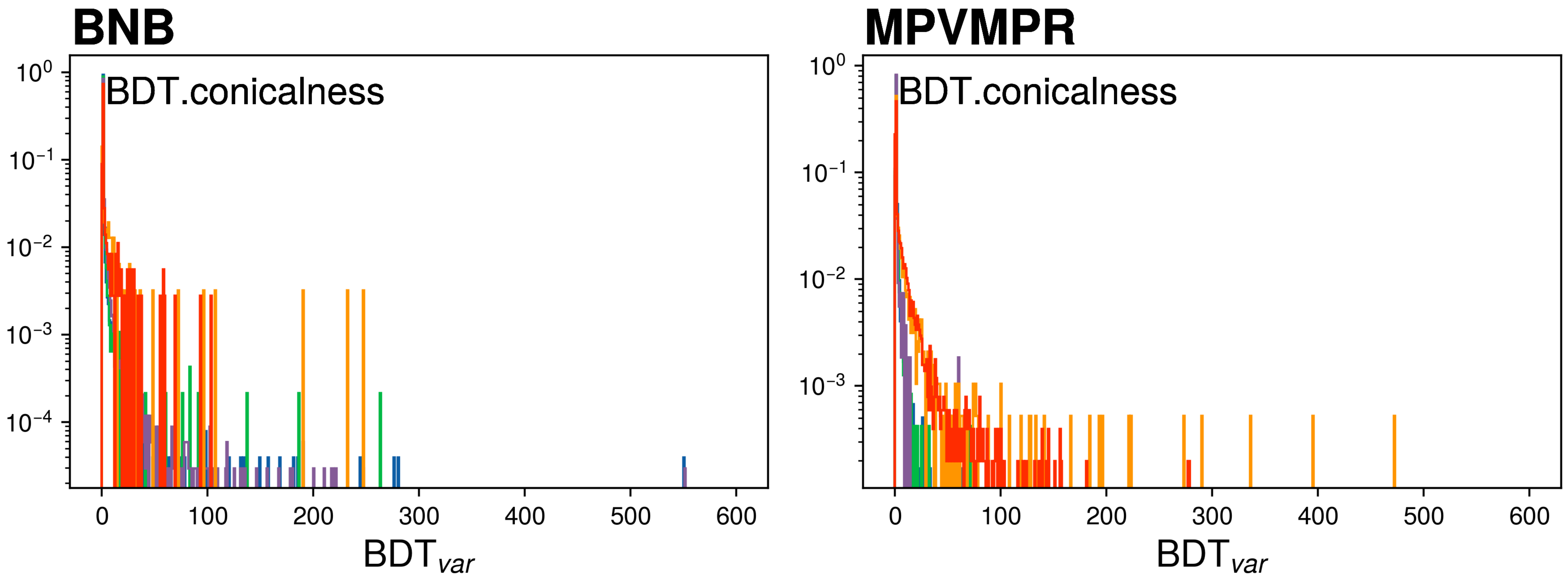
Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles

Looking forward: Next steps

- A new training will be performed with all the BDT variables, with the ML MPVMPR sample

→ Training to be done on the BNB sample

- Some of the variables showed a less than acceptable discrimination power
  - Vertex distance
  - **Conicalness**
  - Concentration
  - Halo total ratio
  - Linear fit length



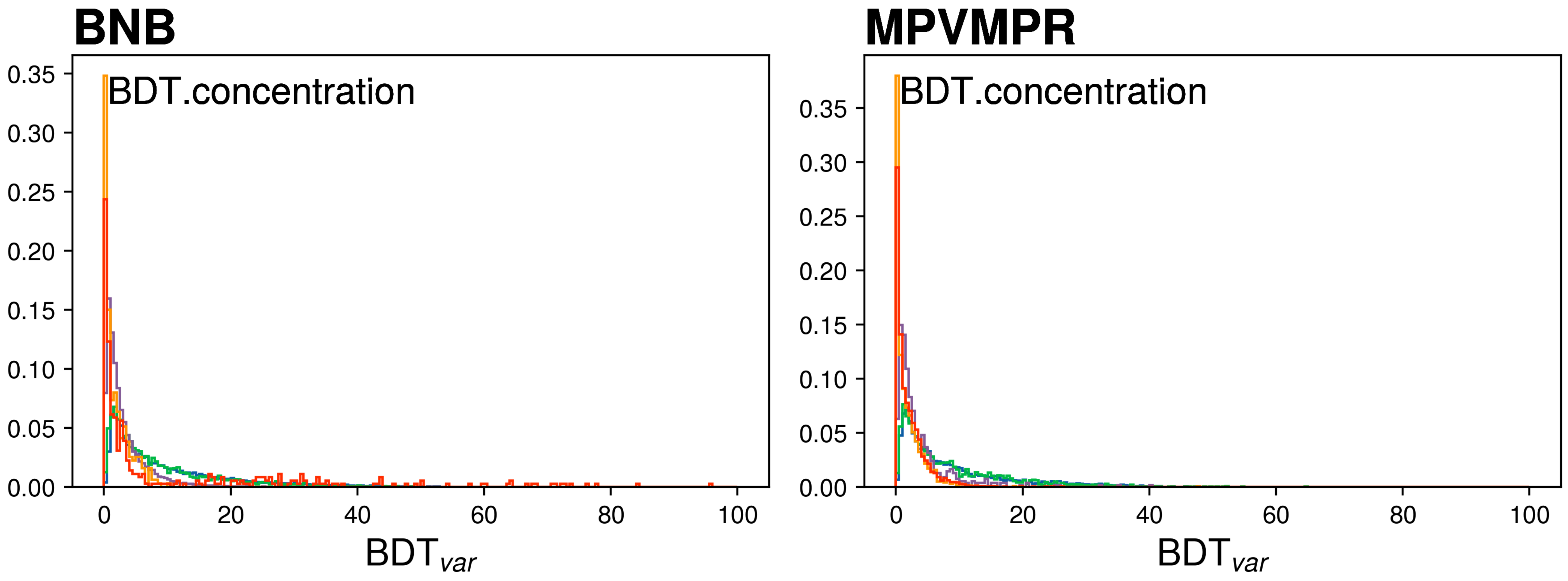
Comparing track like (p, $\pi^\pm$, μ^-) and shower like (e^-, γ) particles

Looking forward: Next steps

- A new training will be performed with all the BDT variables, with the ML MPVMPR sample

Training to be done on the BNB sample

- Some of the variables showed a less than acceptable discrimination power
  - Vertex distance
  - Conicalness
  - **Concentration**
  - Halo total ratio
  - Linear fit length



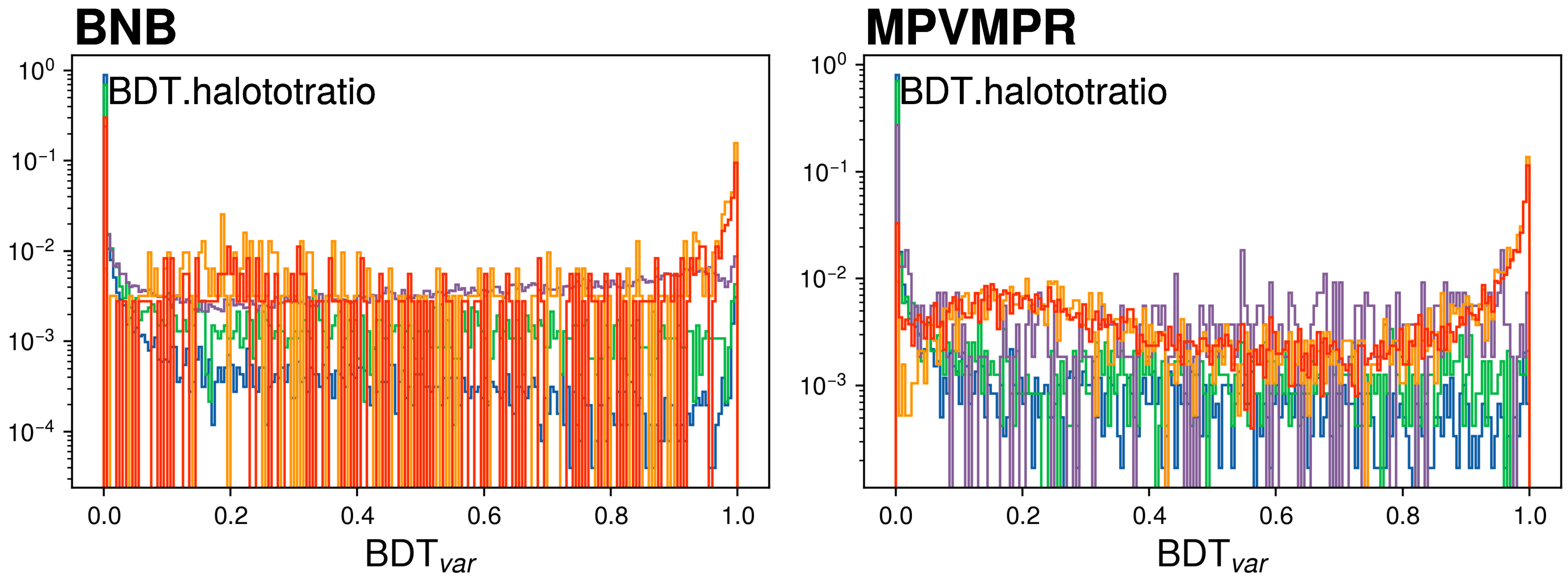
Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles

Looking forward:
Next steps

- A new training will be performed with all the BDT variables, with the ML MPVMPR sample

Training to be done on the BNB sample

- Some of the variables showed a less than acceptable discrimination power
  - Vertex distance
  - Conicalness
  - Concentration
  - **Halo total ratio**
  - Linear fit length



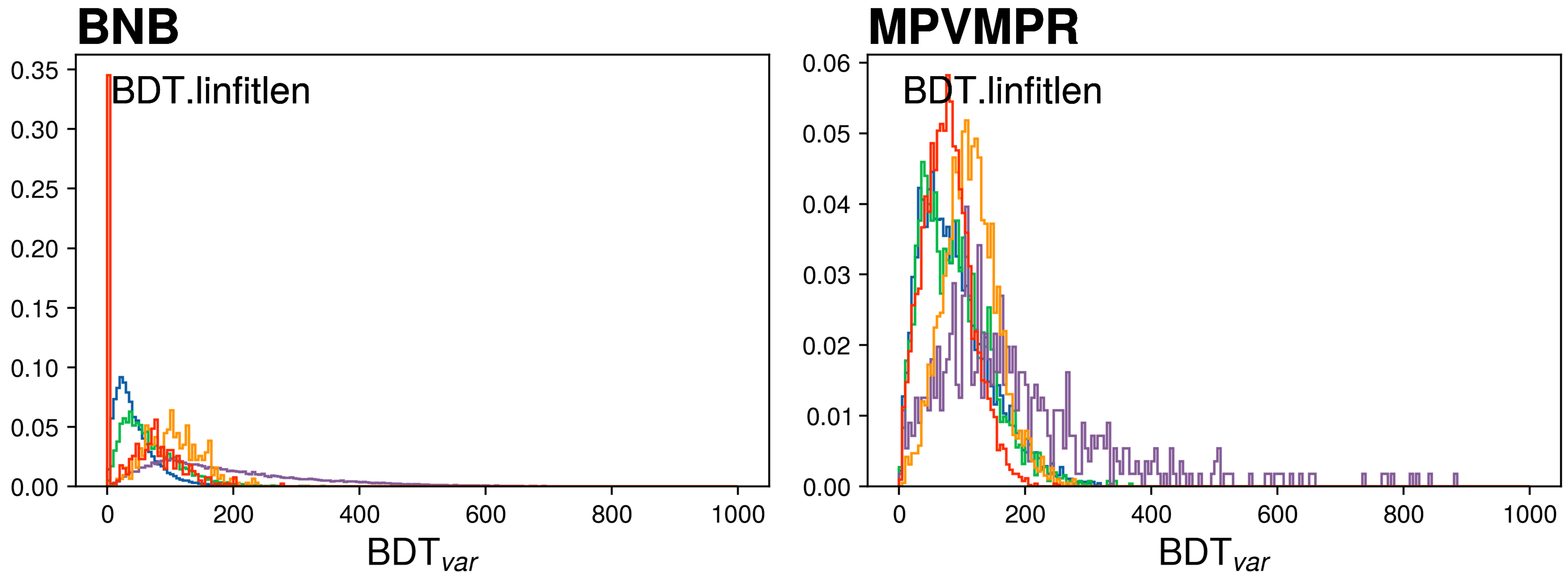
Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles

Looking forward: Next steps

- A new training will be performed with all the BDT variables, with the ML MPVMPR sample

→ Training to be done on the BNB sample

- Some of the variables showed a less than acceptable discrimination power
  - Vertex distance
  - Conicalness
  - Concentration
  - Halo total ratio
  - **Linear fit length**



Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles

Generation of the ML MPVMPR data sample
ML MPVMPR Working Group

The MC sample data is generated with the Multi Particle Vertex Multi Particle Rain module in sbncode/EventGenerator/Multipart/gen_mpvmpr.fcl

The data is in the samweb definition icaruspro_production_2024A_MPVMPR_MC_v09_89_01_01_stage1

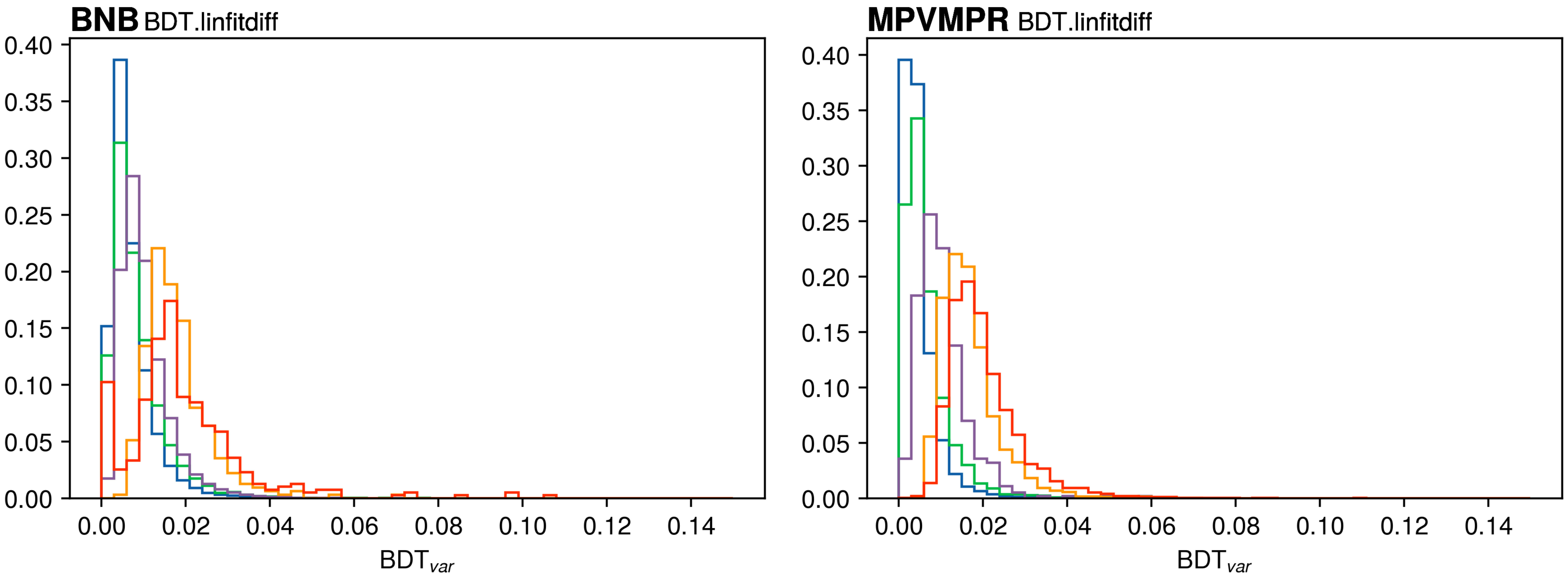
It is generated in three steps

1. One **Multi-Particle Vertex** is generated, random number (with a flat distribution) of particles sampled from a uniform energy distribution
   - The energy range is taken from the expected energies in the BNB
   - The beam spill is set similar, but slightly longer than NuMI, so MPV are generated in the [0, 10] μs range (NuMI is 9.5 μs, whereas BNB is 1.6 μs)
2. A random number (flat distribution in [3, 5]) of single particles sampled from different energy distribution is generated (**rain2**), covering the kind of cosmic we could see.
   - Generated in time (during the beam spill)
   - Generated in a larger volume than the TPC fiducial (+20 cm each direction)
3. A random number (flat distribution in [2, 4]) of single particles sampled from different energy distribution is generated (**rain**), covering the kind of cosmic we could see.
   - Generated out of time (not during the beam spill)
   - Generated in a smaller volume than the TPC fiducial (-20 cm each direction)

Further details in ML sample SBN-doc-35469-v1

Comparing the BNB MC and ML MPVMPR MC datasets
Linear fit difference

- Not so much improvement, it shows in both cases a potential discrimination power



Comparing track like (p, $\pi^{\pm}$, μ^{-}) and shower like (e^{-}, γ) particles