IDENTIFICATION OF TAU AND ITS DECAY MODES USING MACHINE LEARNING (W/ THE DUAL-READOUT CALORIMETER)

Stefano Giagu, Matteo di Filippo, Luca Torresi
Sapienza Università di Roma and INFN

On behalf of the dual-readout calorimeter detector concept group

2021 International Workshop on the high energy Circular Electron-Positron Collider - 11.11.2021

INTRODUCTION

- ongoing work aiming at maximising the physics potential of future collider experiments
  - **case study:** τ-identification in the IDEA dual-readout calorimeter (DRC) concept
    - leverage modern machine learning methods based on differentiable deep neural networks
    - study performance using only standalone DRC information
    - helps in optimising the detector and design of the readout electronics
  - **tasks studied:**
    - classification of τ-decays and separation from QCD jets based on Graph Neural Networks (DGCNN)
    - bayesian-DGCNN for robust estimation of NN predictions
    - DGCNN-based object detection (eg identification of γ and n inside hadronic tau decays) for particle-flow algorithms
- part of a common effort in the DRC group to implement from start ML/DL methods in the design and development of the detector (see also Sanghyun Ko’s talk at this workshop)

DRC PRINCIPLE

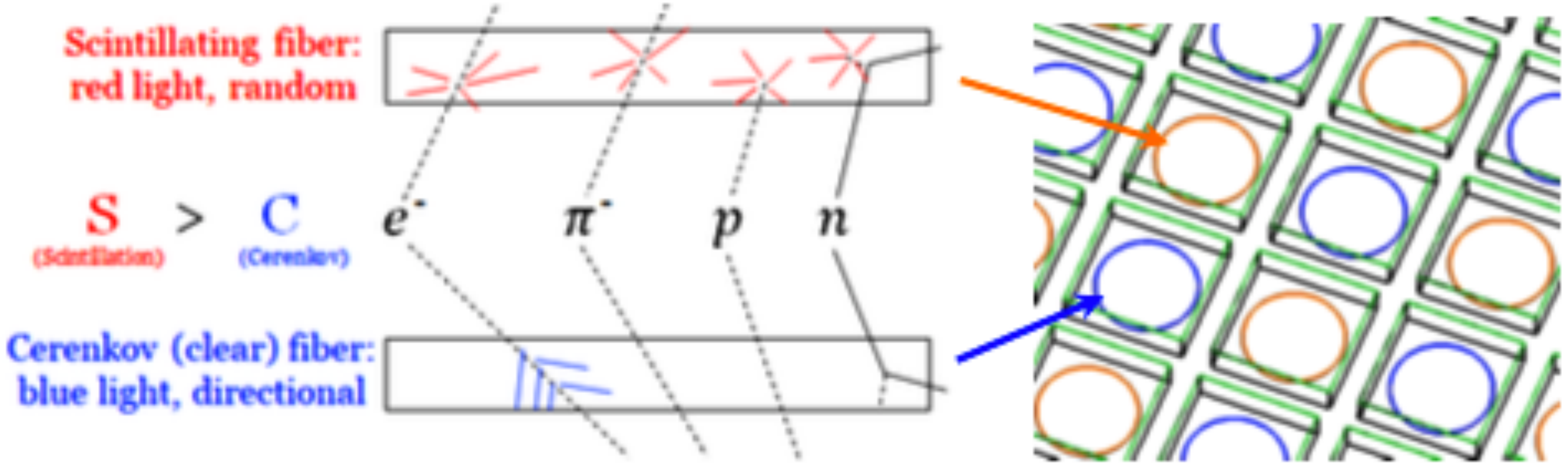
correct shower energy event by event for non-compensation by measuring the EM fraction in hadronic shower by sampling with two readouts of different e/h response: Cherenkov (C) mostly sensitive to the em shower component, Scintillation (S) sensitive to all

$$S = E[f_{em} + (h/e)_s(1 - f_{em})]$$

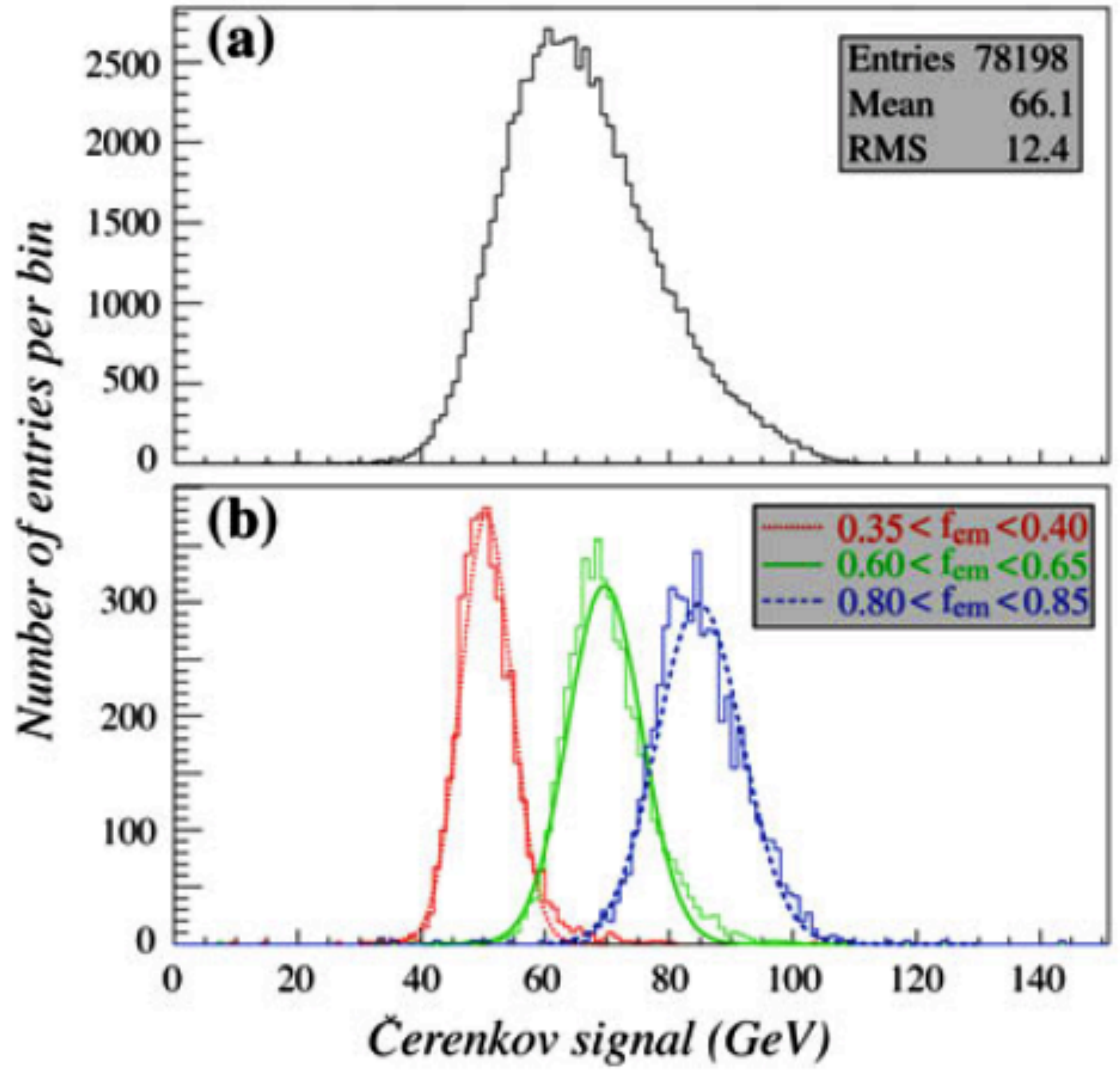
$$C = E[f_{em} + (h/e)_c(1 - f_{em})]$$

two equations in two unknowns: f_{em} and E

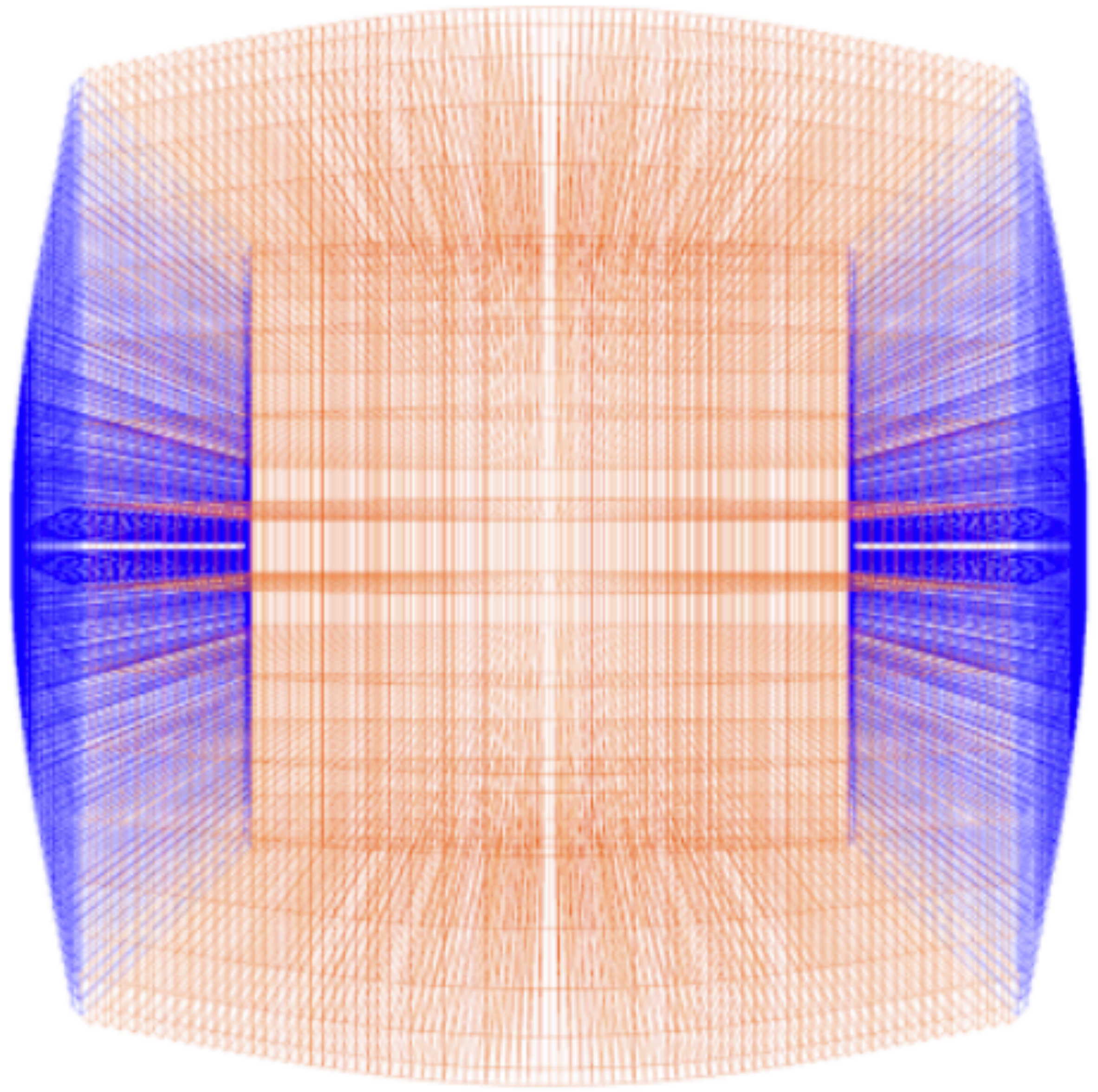
$$E = \frac{S - \chi C}{1 - \chi} \qquad \frac{1 - (h/e)_S}{1 - (h/e)_C} = \chi$$

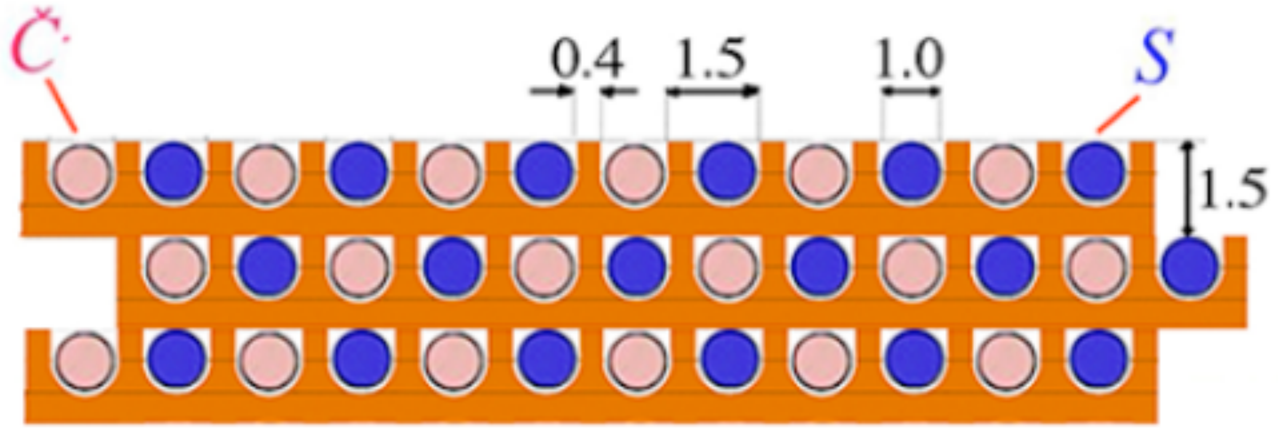


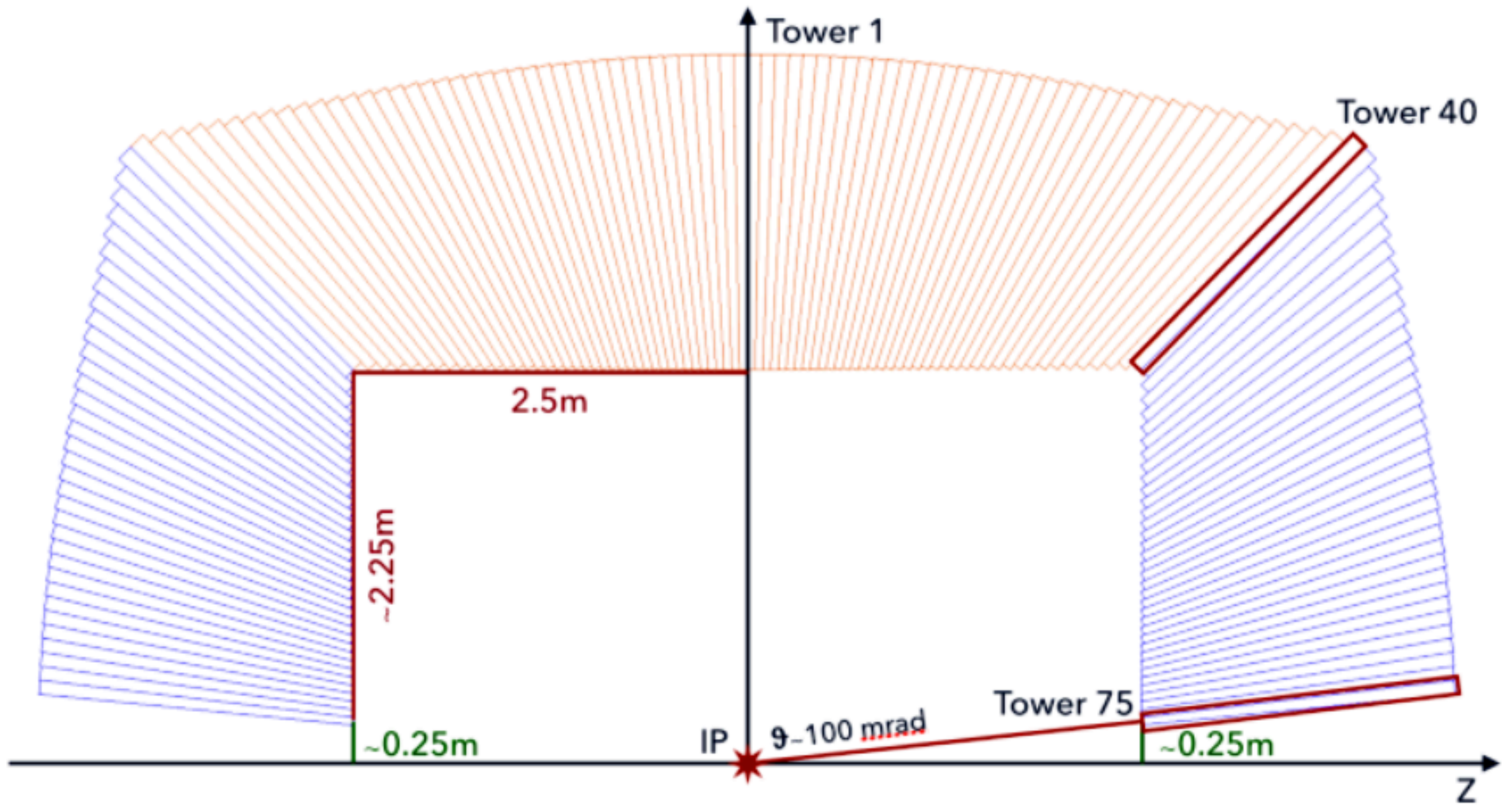
different patterns of S vs C light from different particles, combined with the fine segmentation provided by the fibres can be leveraged for powerful particle identification …



IDEA DRC SIMULATION





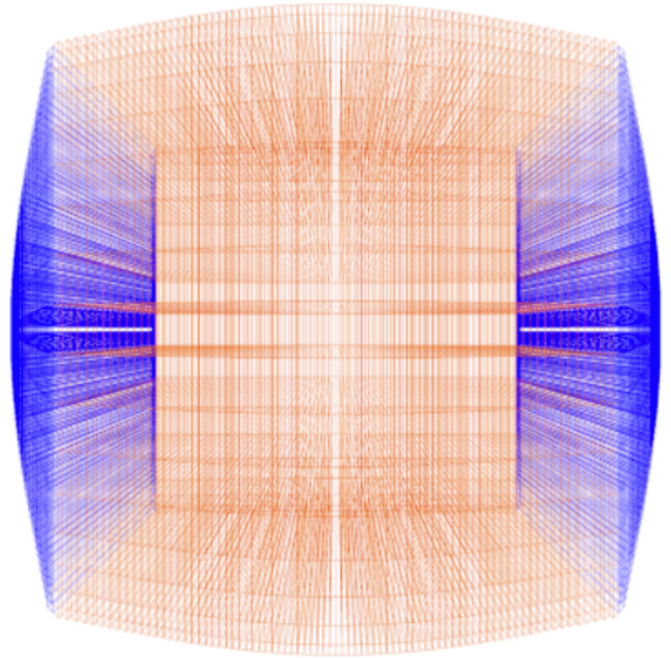


- full G4 simulation of the calorimeter geometry:
  - includes B field and solenoid material in front of the calorimeter
  - fiber-sampling calorimeter: Cu absorber, 1mm fibres, 1.5mm pitch
  - read out of each single fibre via SiPM
  - 130 M channels, excellent granularity and lateral shape sensitivity:

$$\Delta\theta, \Delta\phi = \sim 0.035^\circ$$

- parametrised simulation of SiPM readout and signal processing
  - dark counts, crosstalk, afterpulses, saturation, noise, ...

DATASETS



- Pythia8 $e^+e^- \rightarrow Z \rightarrow \tau\tau$ and qq at Z pole
- 5000 events for each decay mode

Decay	Label 8-class
$\tau^- \rightarrow e^- \nu_e \nu_\tau$	0
$\tau^- \rightarrow \pi^- \nu_\tau$	1
$\tau^- \rightarrow \pi^0 \pi^- \nu_\tau$	2
$\tau^- \rightarrow \pi^0 \pi^0 \pi^- \nu_\tau$	3
$\tau^- \rightarrow \pi^- \pi^- \pi^+ \nu_\tau$	4
$\tau^- \rightarrow \pi^0 \pi^- \pi^- \pi^+ \nu_\tau$	5
$\tau^- \rightarrow \mu^- \nu_\mu \nu_\tau$	6
$Z \rightarrow q\bar{q} \rightarrow jet\, jet$	7

- Information available for each fibre:
  - geometrical quantities: $\Delta\theta$, $\Delta\phi$ wrt the tau/jet cluster center
  - energetic quantities: # of photo-electrons in fibres and energy (scintillation and Cherenkov)
  - SiPM information (1 SiPM per fibre): Integral and Peak of the SiPM output, Time of Arrival, Time over Threshold, Time of Peak
- Labels:
  - fiber type (scintillating or cherenkov)
  - decay type label

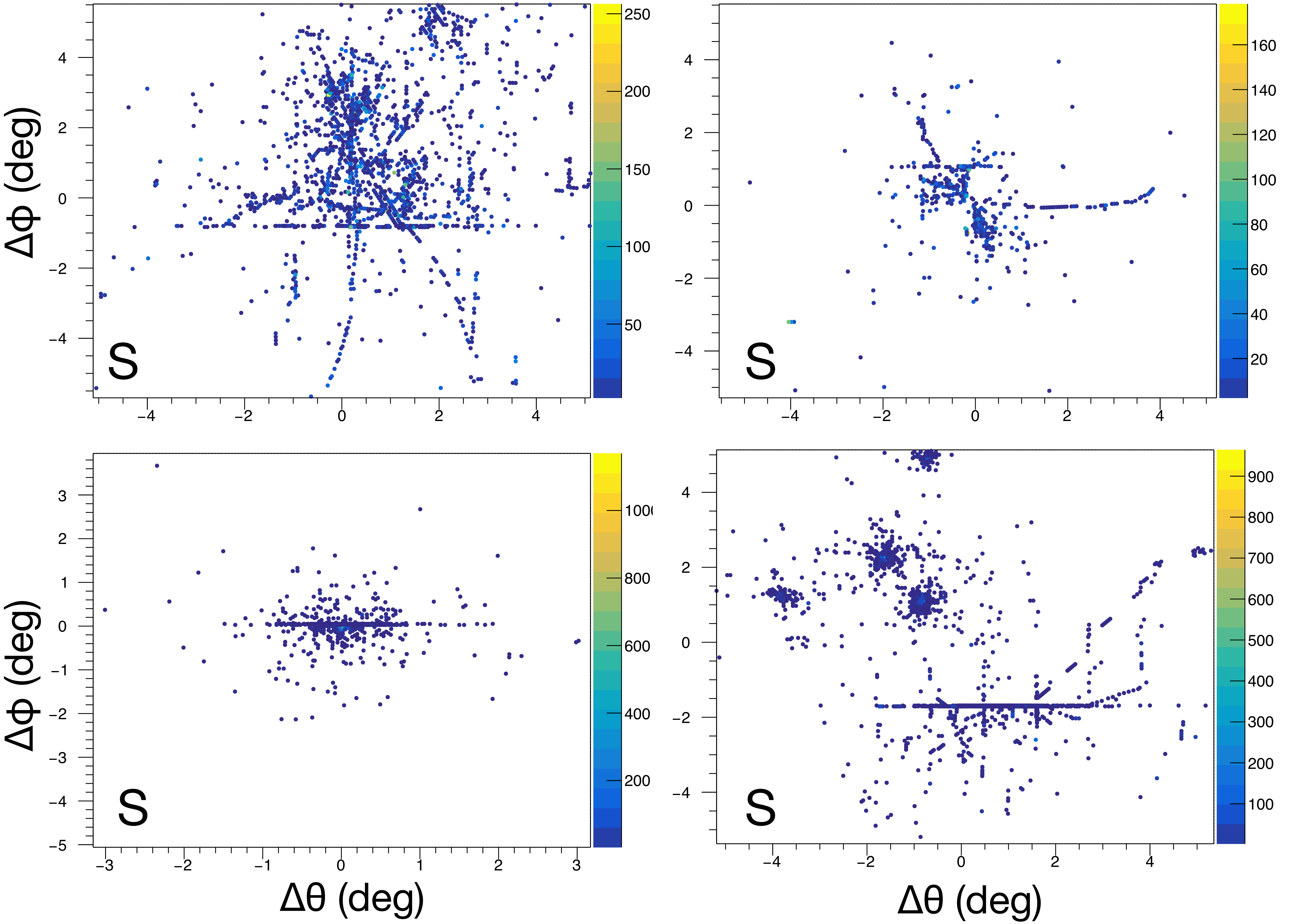
EXAMPLES OF EVENTS WITH FULL GRANULARITY

Z→qq Jets

τ →πν

τ →eνν

τ →ππ^{0}π^{0}ν

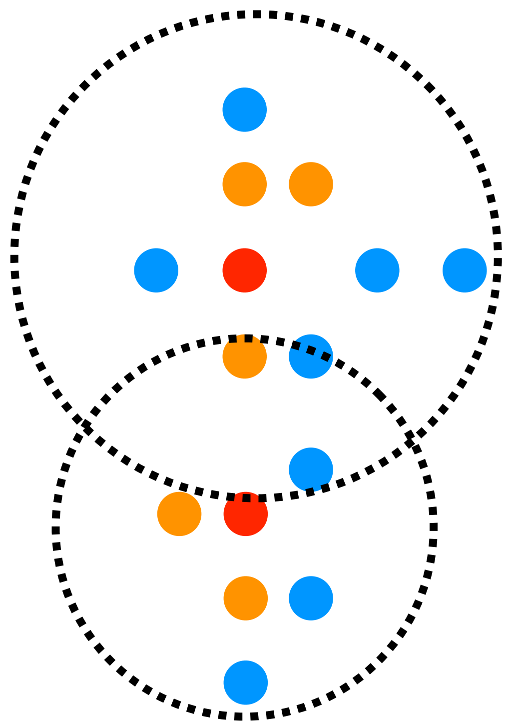


DATA REPRESENTATION

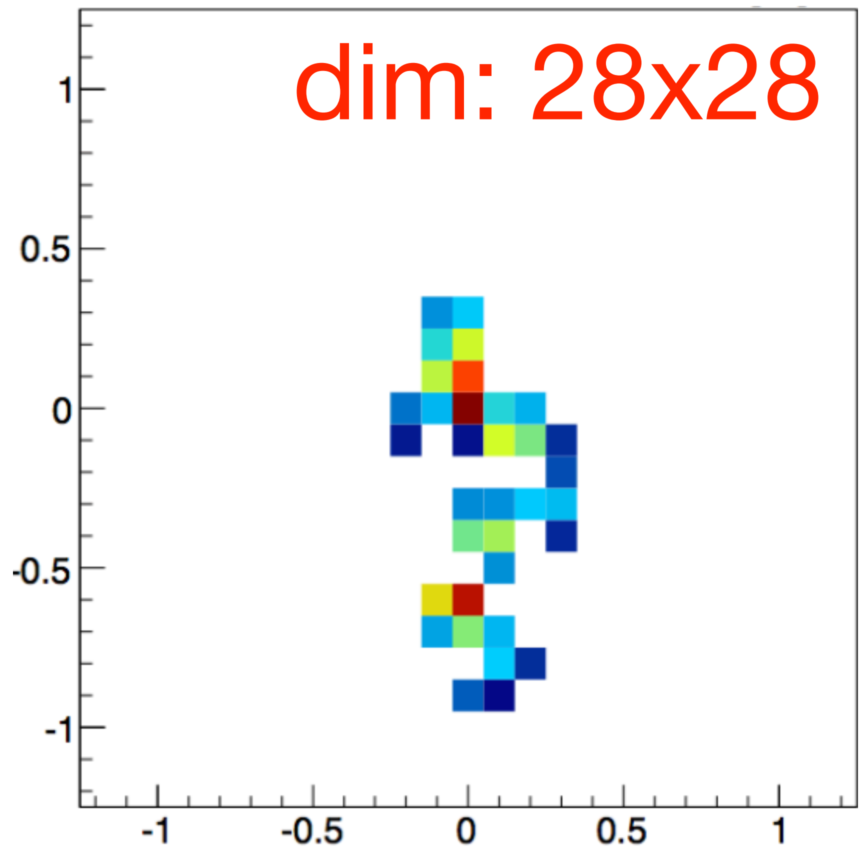
- Image-based: treating the energy deposition on each fiber as the pixel intensity creates an image of the event in fixed-shape mesh
  - natural representation for Convolutional Neural Networks
  - unclear how to incorporate additional information of the fibers
  - very sparse and inefficient representation: jets/tau decays have O(10) to O(100) particles → more than 90% of the pixels are blank

- Point cloud-based: unordered sets of entities distributed irregularly in space, analogous to the point cloud representation of 3D shapes
  - clouds allow rich internal structures
  - easy to incorporate additional information of the fibers (fibre type, energy, time information, …)
  - the architecture of the neural network has to be carefully designed to fully exploit the potential of this representation → Dynamic Graph CNN

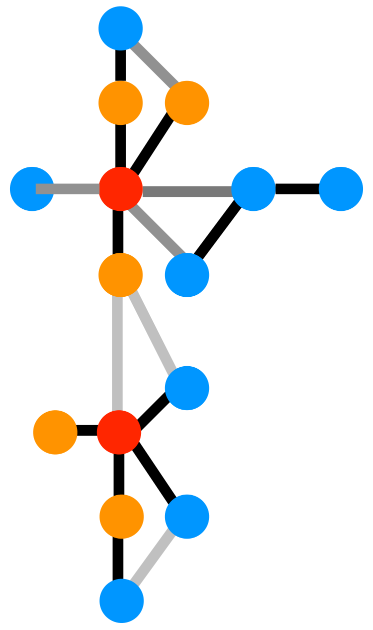
boosted W→qq’



fixed-mesh image


dim: 28x28

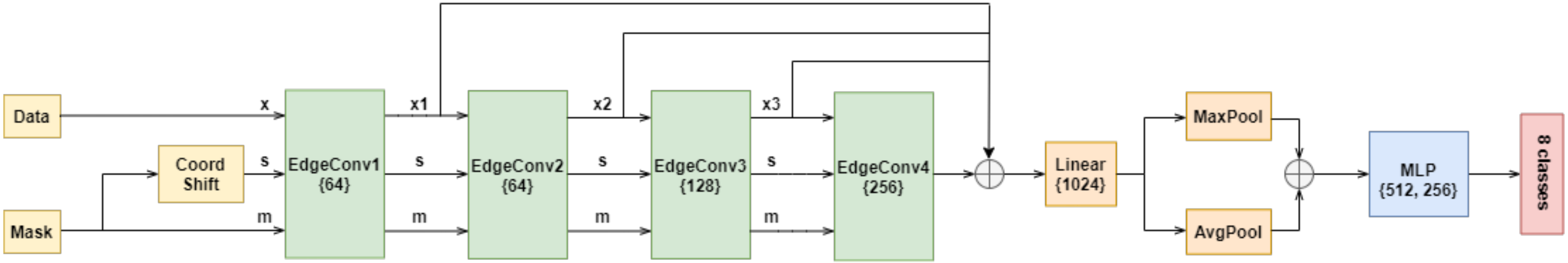
graph



dim: 15x(3+1) + 18

Y. Wang et al., arXiv:1801.07829 [cs.CV]
H. Qu and L. Gouskos, arXiv:1902.08570 [hep-ph]

DGCNN ARCHITECTURE



Input:
x = {θ, ϕ, geometrical features, SiPM features, …}

local feature extractor

feature aggregator

MLP global classifier

- flexible architecture optimised for point cloud inputs able to learn both local (trough the edge convolution) and global (through the feature aggregator) structures
- simplify inclusion of additional features and SiPM signal timing information
- # of input fibres fixed and treated as model hyper parameter, discarding those with lowest signals or adding zero valued vectors in case of events with lower active fibres
- hyper-parameters chosen using a validation set

EDGE CONVOLUTION

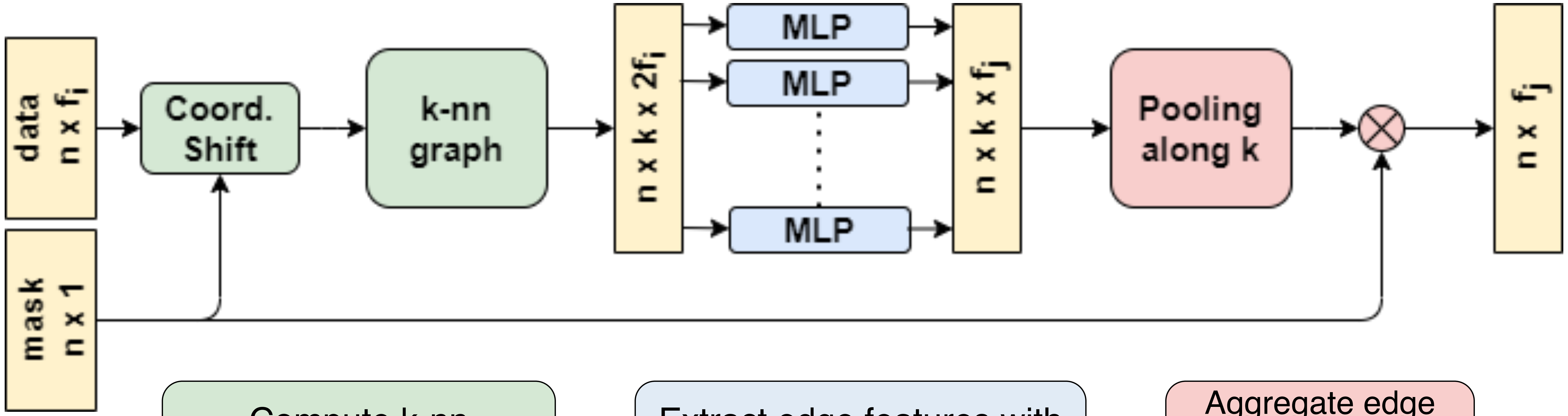
Regular convolution operations cannot be applied on point clouds:

- points distribution is usually irregular (unlike uniform grids of the pixels in an image)
- they’re not invariant under permutation of the points

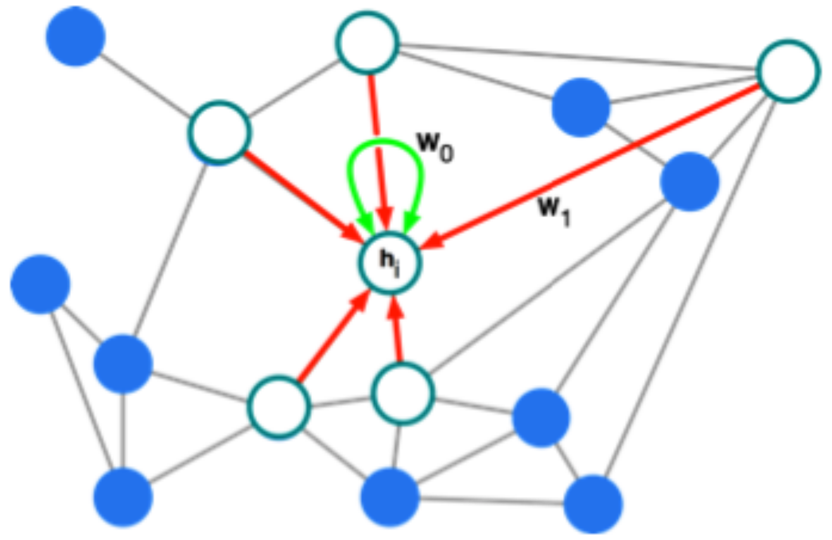
CNN-conv

$$C = \sum_{i=0}^{8} h_i w_i$$

A viable solution: **EDGE convolution**: point cloud represented as a graph with **Vertices (**the points themselves) and **Edges** (connections between each point to its k nearest neighbouring points): results in a **regular distribution** for each point, for which is possible to define convolution operations



Compute k-nn directed graph → Extract edge features with a shared linear layer → Aggregate edge features by max-pooling

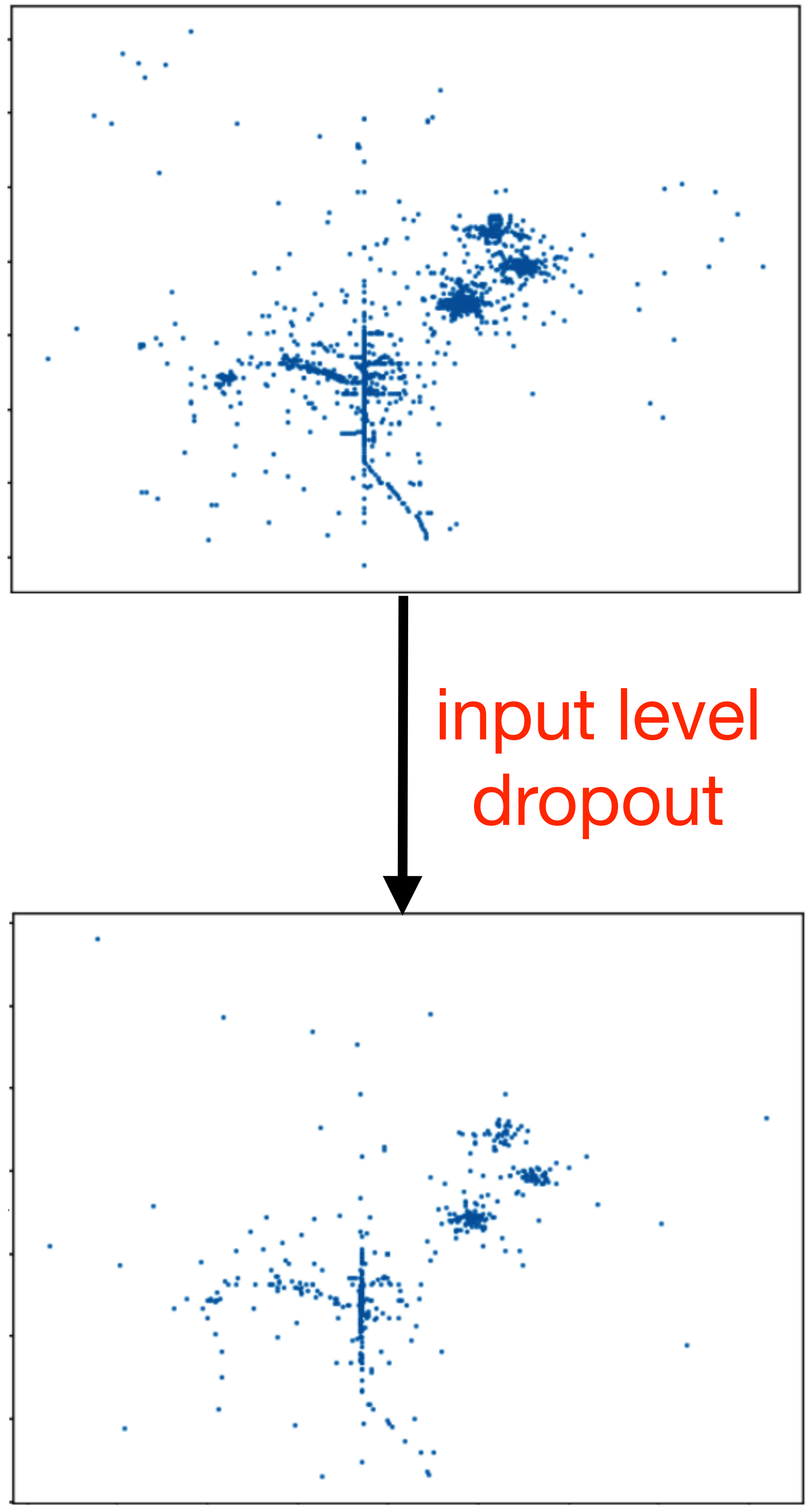


EDGE-conv

$$EC_i = \sum_{q \in N_p} f(h_i, h_q, w_{iq})$$

τ DECAY IDENTIFICATION WITH DGCNN

- Classification task:
  - 8-classes: 7 tau decays + QCD jets
  - training/validation/test sets: 22k/6k/7k events (balanced among classes)
- Data-preprocessing:
  - simple geometrical clustering, no specific selection or fiducial volume applied
  - saved fibres signal around the clusters ($\sqrt{\Delta\theta^2+\Delta\phi^2}) < 1$)
- DGCNN inputs:
  - jet/tau representation: 2D point-cloud of fibres coordinates
  - fiber type (S, C), #photo-electrons, SiPM's: Integral and Peak of the SiPM output, ToT, ToA, ToP (in different combinations)
- Data augmentation/regularisation: overfitting and memorisation for the DNN model controlled using dropout
  - at input level: some of the fired fibres are switched off
  - in the neural network layers: some of the parameters of the last MLP block are randomly zeroed during the training phase
  - better generalisation obtained leveraging both methods

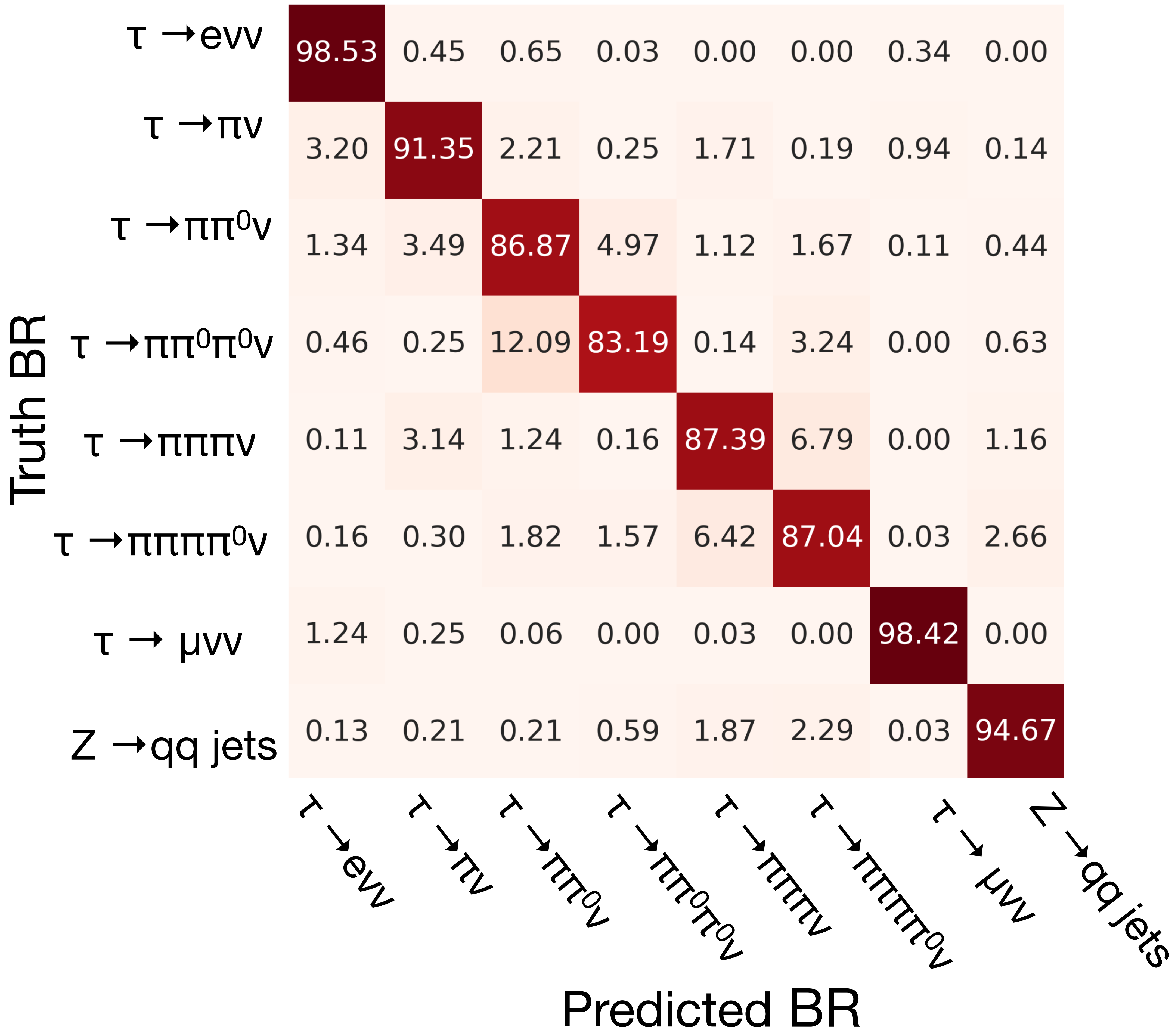


input level dropout

RESULTS (input features: fibers coordinates, type (S, C), w/ & w/o #p.e.)

using **coordinates, type** of fibre, and **#of photo-electrons** in each fibre

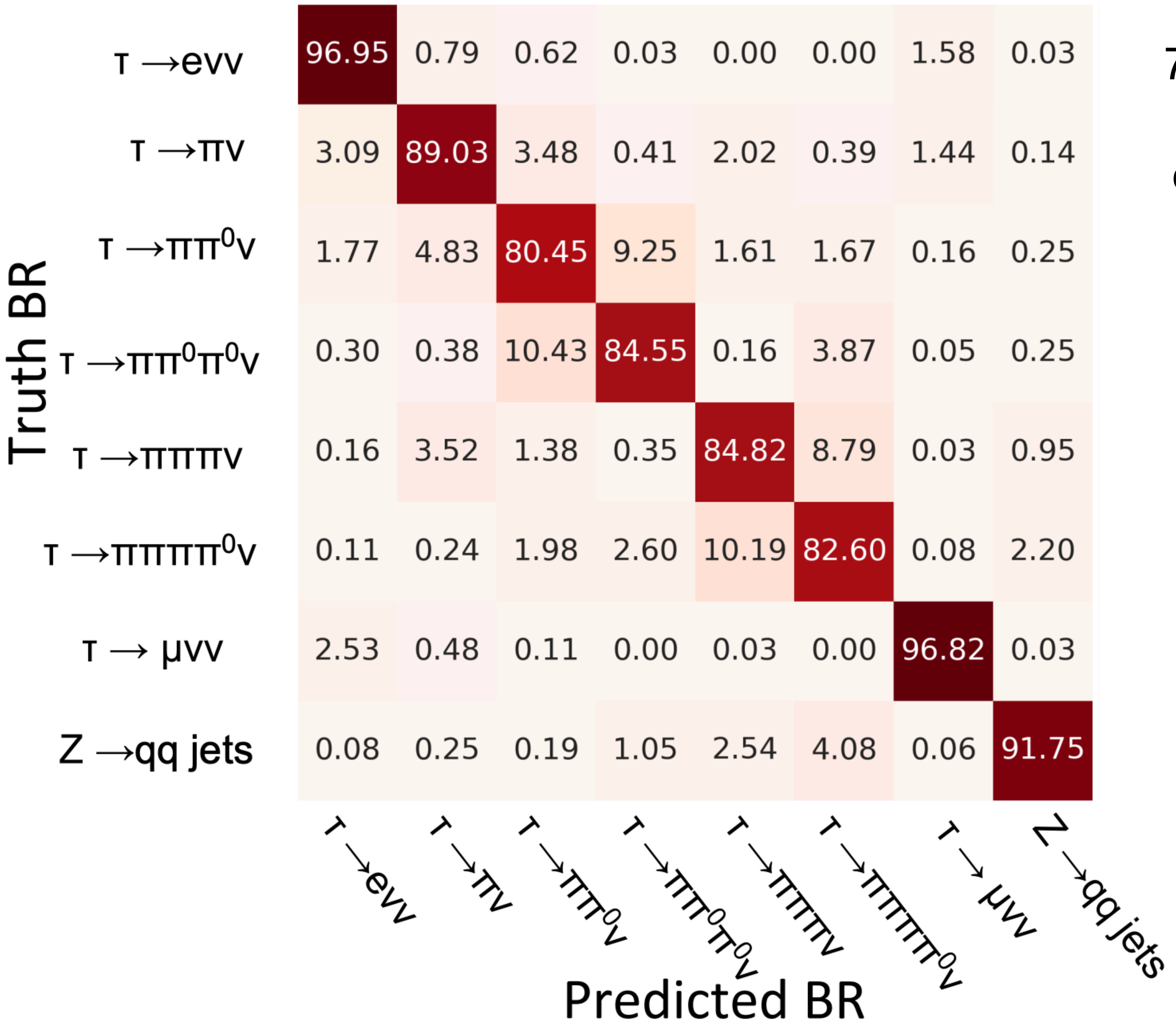
average accuracy:
90.8%



	$\tau \rightarrow e\nu\nu$	$\tau \rightarrow \pi\nu$	$\tau \rightarrow \pi\pi^0\nu$	$\tau \rightarrow \pi\pi^0\pi^0\nu$	$\tau \rightarrow \pi\pi\pi\nu$	$\tau \rightarrow \pi\pi\pi\pi^0\nu$	$\tau \rightarrow \mu\nu\nu$	$Z \rightarrow qq$ jets
$\tau \rightarrow e\nu\nu$	98.53	0.45	0.65	0.03	0.00	0.00	0.34	0.00
$\tau \rightarrow \pi\nu$	3.20	91.35	2.21	0.25	1.71	0.19	0.94	0.14
$\tau \rightarrow \pi\pi^0\nu$	1.34	3.49	86.87	4.97	1.12	1.67	0.11	0.44
$\tau \rightarrow \pi\pi^0\pi^0\nu$	0.46	0.25	12.09	83.19	0.14	3.24	0.00	0.63
$\tau \rightarrow \pi\pi\pi\nu$	0.11	3.14	1.24	0.16	87.39	6.79	0.00	1.16
$\tau \rightarrow \pi\pi\pi\pi^0\nu$	0.16	0.30	1.82	1.57	6.42	87.04	0.03	2.66
$\tau \rightarrow \mu\nu\nu$	1.24	0.25	0.06	0.00	0.03	0.00	98.42	0.00
$Z \rightarrow qq$ jets	0.13	0.21	0.21	0.59	1.87	2.29	0.03	94.67

using only **coordinates and type** for each fibre

average accuracy:
88.3%



	$\tau \rightarrow e\nu\nu$	$\tau \rightarrow \pi\nu$	$\tau \rightarrow \pi\pi^0\nu$	$\tau \rightarrow \pi\pi^0\pi^0\nu$	$\tau \rightarrow \pi\pi\pi\nu$	$\tau \rightarrow \pi\pi\pi\pi^0\nu$	$\tau \rightarrow \mu\nu\nu$	$Z \rightarrow qq$ jets
$\tau \rightarrow e\nu\nu$	96.95	0.79	0.62	0.03	0.00	0.00	1.58	0.03
$\tau \rightarrow \pi\nu$	3.09	89.03	3.48	0.41	2.02	0.39	1.44	0.14
$\tau \rightarrow \pi\pi^0\nu$	1.77	4.83	80.45	9.25	1.61	1.67	0.16	0.25
$\tau \rightarrow \pi\pi^0\pi^0\nu$	0.30	0.38	10.43	84.55	0.16	3.87	0.05	0.25
$\tau \rightarrow \pi\pi\pi\nu$	0.16	3.52	1.38	0.35	84.82	8.79	0.03	0.95
$\tau \rightarrow \pi\pi\pi\pi^0\nu$	0.11	0.24	1.98	2.60	10.19	82.60	0.08	2.20
$\tau \rightarrow \mu\nu\nu$	2.53	0.48	0.11	0.00	0.03	0.00	96.82	0.03
$Z \rightarrow qq$ jets	0.08	0.25	0.19	1.05	2.54	4.08	0.06	91.75

73.7% using only the coordinates

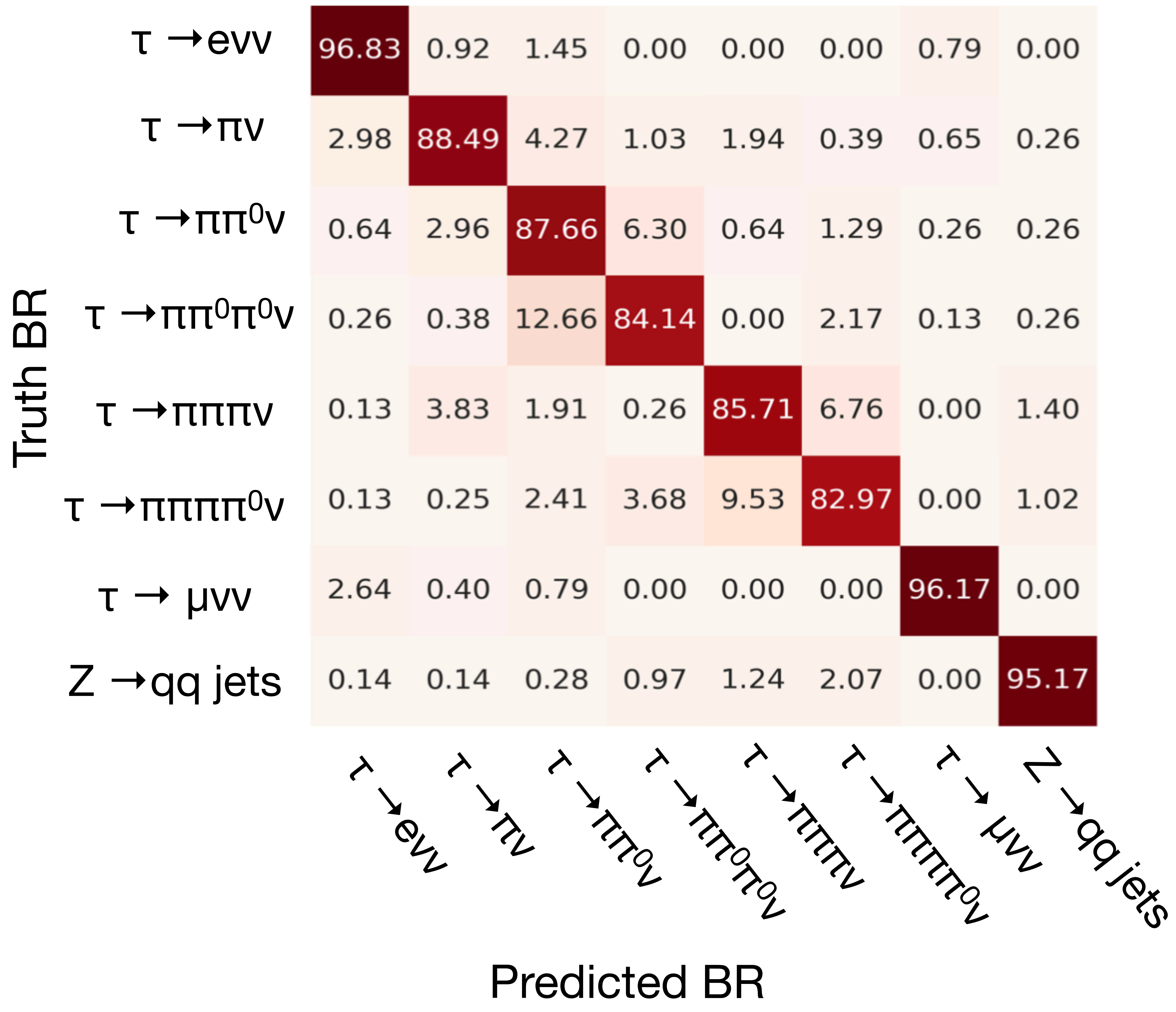
uncertainty on accuracies ~3÷5%

double-readout geometry alone allows excellent tau identification

RESULTS (input features: fibers coordinates, type (S, C), SiPM information)

using only geometrical and Integral/Peak of the signal

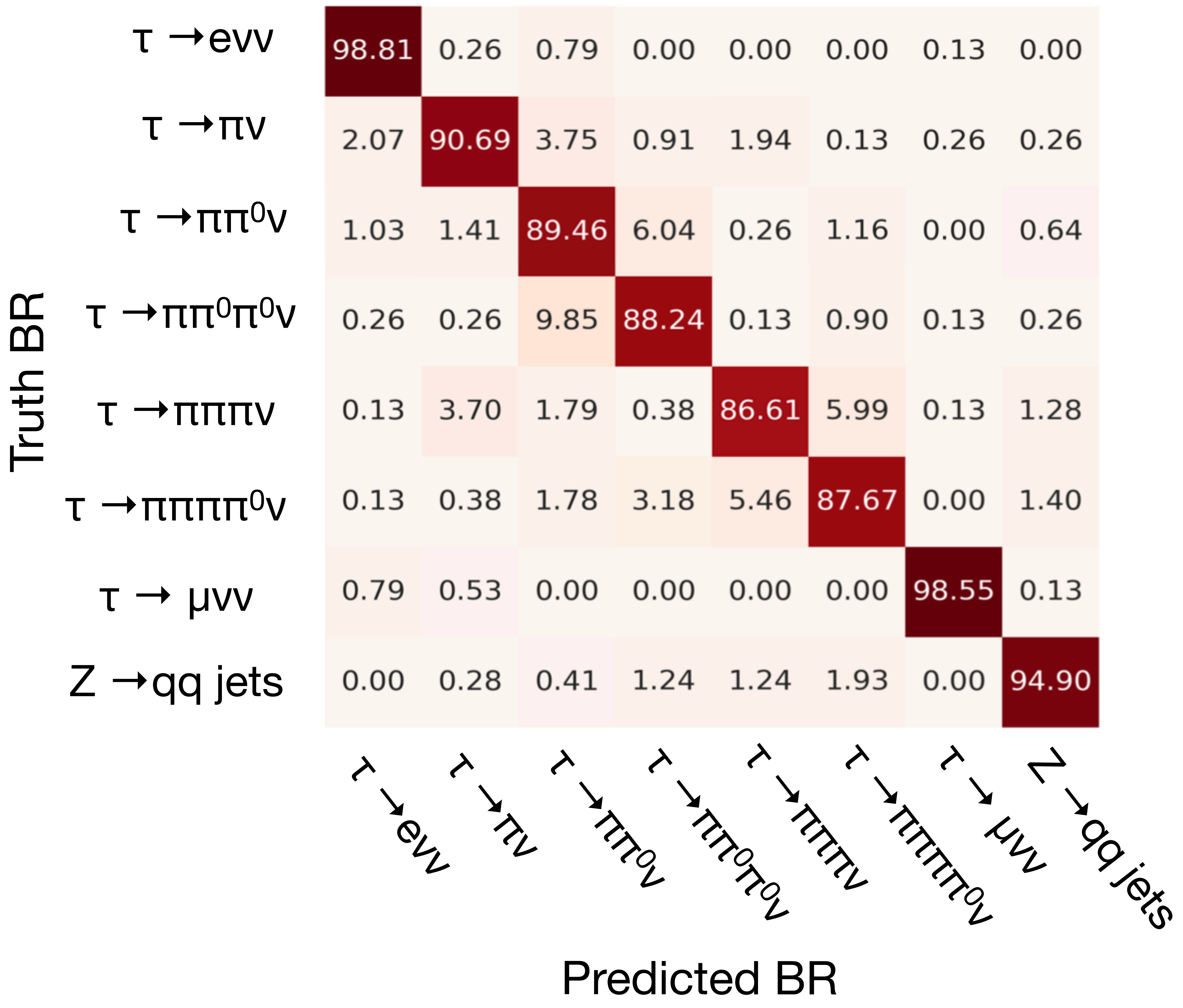
average accuracy: **88.8%**



Truth BR \ Predicted BR	$\tau \to e\nu\nu$	$\tau \to \pi\nu$	$\tau \to \pi\pi^0\nu$	$\tau \to \pi\pi^0\pi^0\nu$	$\tau \to \pi\pi\pi\nu$	$\tau \to \pi\pi\pi\pi^0\nu$	$\tau \to \mu\nu\nu$	$Z \to qq$ jets
$\tau \to e\nu\nu$	96.83	0.92	1.45	0.00	0.00	0.00	0.79	0.00
$\tau \to \pi\nu$	2.98	88.49	4.27	1.03	1.94	0.39	0.65	0.26
$\tau \to \pi\pi^0\nu$	0.64	2.96	87.66	6.30	0.64	1.29	0.26	0.26
$\tau \to \pi\pi^0\pi^0\nu$	0.26	0.38	12.66	84.14	0.00	2.17	0.13	0.26
$\tau \to \pi\pi\pi\nu$	0.13	3.83	1.91	0.26	85.71	6.76	0.00	1.40
$\tau \to \pi\pi\pi\pi^0\nu$	0.13	0.25	2.41	3.68	9.53	82.97	0.00	1.02
$\tau \to \mu\nu\nu$	2.64	0.40	0.79	0.00	0.00	0.00	96.17	0.00
$Z \to qq$ jets	0.14	0.14	0.28	0.97	1.24	2.07	0.00	95.17

adding also SiPM timing information

average accuracy: **90.8%**

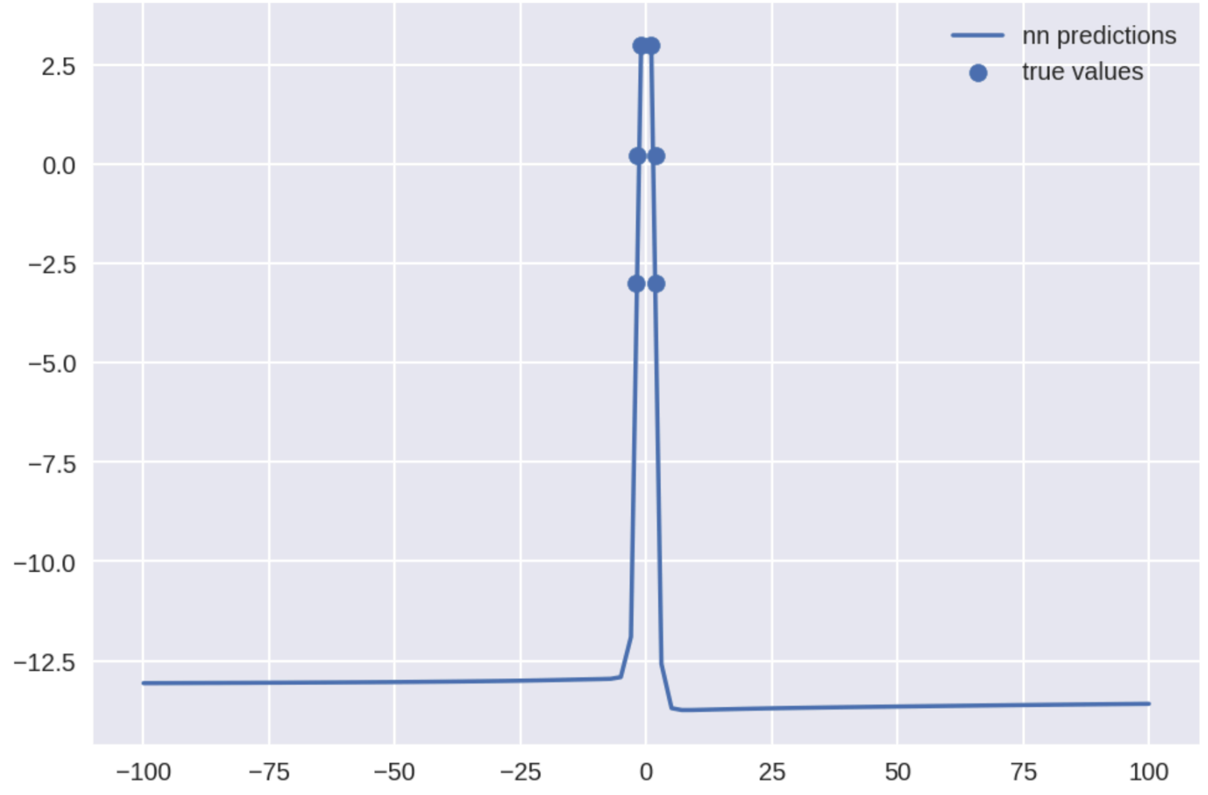


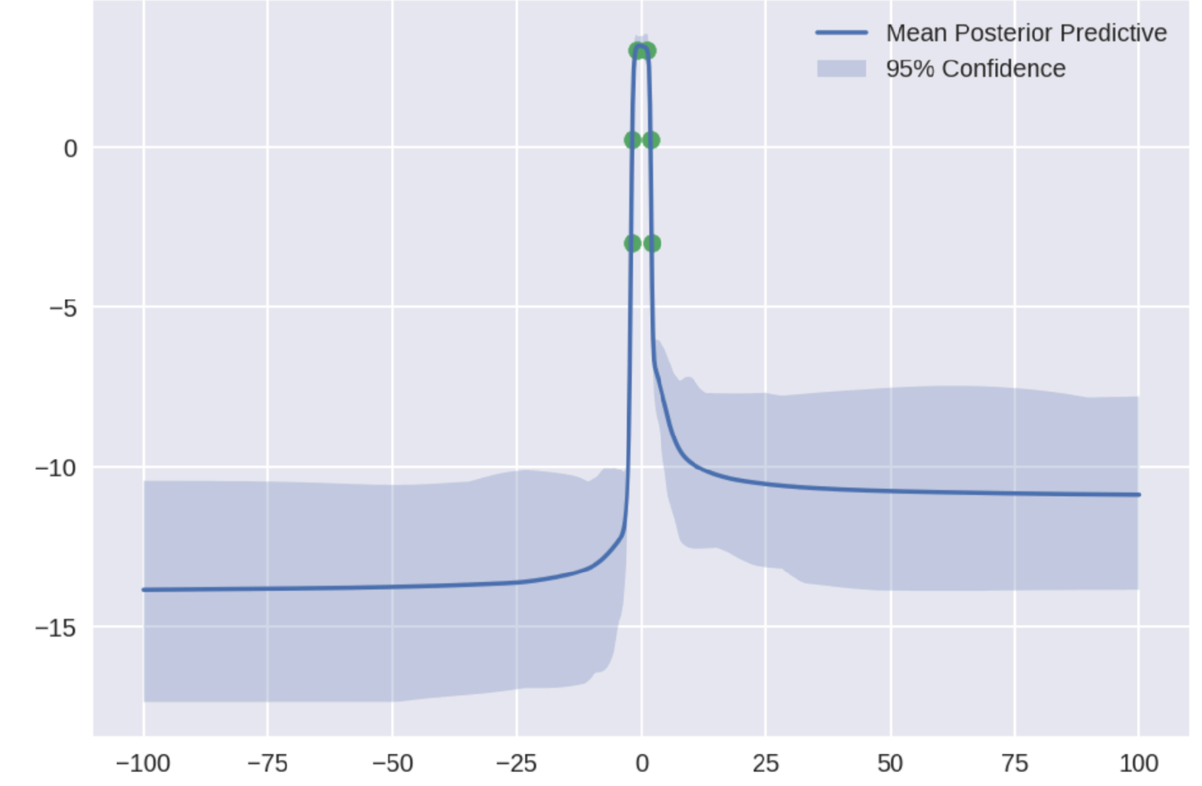
Truth BR \ Predicted BR	$\tau \to e\nu\nu$	$\tau \to \pi\nu$	$\tau \to \pi\pi^0\nu$	$\tau \to \pi\pi^0\pi^0\nu$	$\tau \to \pi\pi\pi\nu$	$\tau \to \pi\pi\pi\pi^0\nu$	$\tau \to \mu\nu\nu$	$Z \to qq$ jets
$\tau \to e\nu\nu$	98.81	0.26	0.79	0.00	0.00	0.00	0.13	0.00
$\tau \to \pi\nu$	2.07	90.69	3.75	0.91	1.94	0.13	0.26	0.26
$\tau \to \pi\pi^0\nu$	1.03	1.41	89.46	6.04	0.26	1.16	0.00	0.64
$\tau \to \pi\pi^0\pi^0\nu$	0.26	0.26	9.85	88.24	0.13	0.90	0.13	0.26
$\tau \to \pi\pi\pi\nu$	0.13	3.70	1.79	0.38	86.61	5.99	0.13	1.28
$\tau \to \pi\pi\pi\pi^0\nu$	0.13	0.38	1.78	3.18	5.46	87.67	0.00	1.40
$\tau \to \mu\nu\nu$	0.79	0.53	0.00	0.00	0.00	0.00	98.55	0.13
$Z \to qq$ jets	0.00	0.28	0.41	1.24	1.24	1.93	0.00	94.90

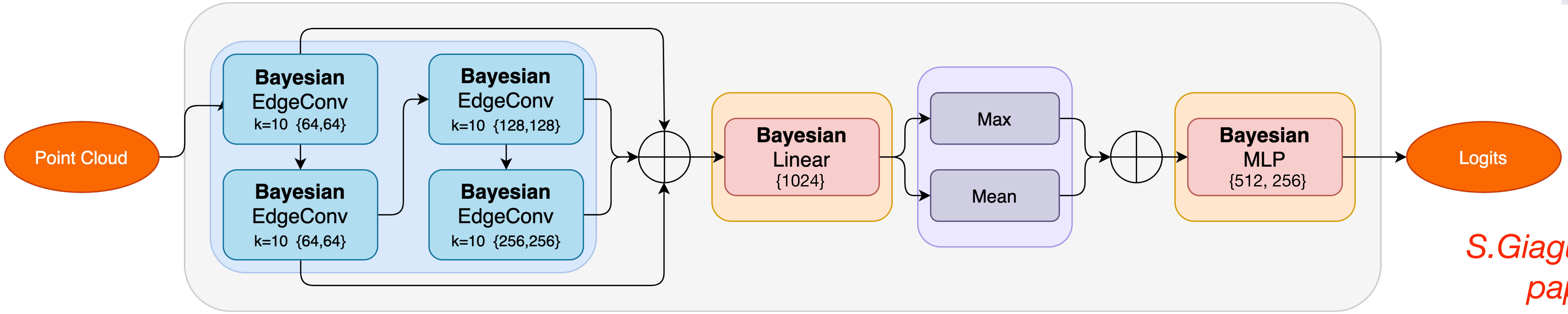
comparable identification performance with input from SiPM emulation

UNCERTAINTY IN THE CLASSIFICATION: BAYESIAN-DGCNN

- Neural networks based on point values for weights may suffer of overconfidence when analysing new data especially concerning generalization in regions without examples in the training set
- Bayesian neural networks solve the problem by introducing probability distributions over the weights and **predicting distributions instead of point values**
  - a Bayesian-NN learns a variational approximation of the true posterior distribution P(w|D), and predict an estimate of the expected value $E_{P(w|D)}[P(y|x,w)]$ → since the weights are random variables, each predictions is a random variable too
  - allows to measure uncertainty, identify outliers in the input, regularise the whole model
- Designed and implemented in pytorch a full Bayesian version of a DGCNN (leveraging the *Bayes by Backprop* algorithm (https://arxiv.org/abs/1505.05424)






Point Cloud

Bayesian EdgeConv k=10 {64,64}

Bayesian EdgeConv k=10 {64,64}

Bayesian EdgeConv k=10 {128,128}

Bayesian EdgeConv k=10 {256,256}

Bayesian Linear {1024}

Max

Mean

Bayesian MLP {512, 256}

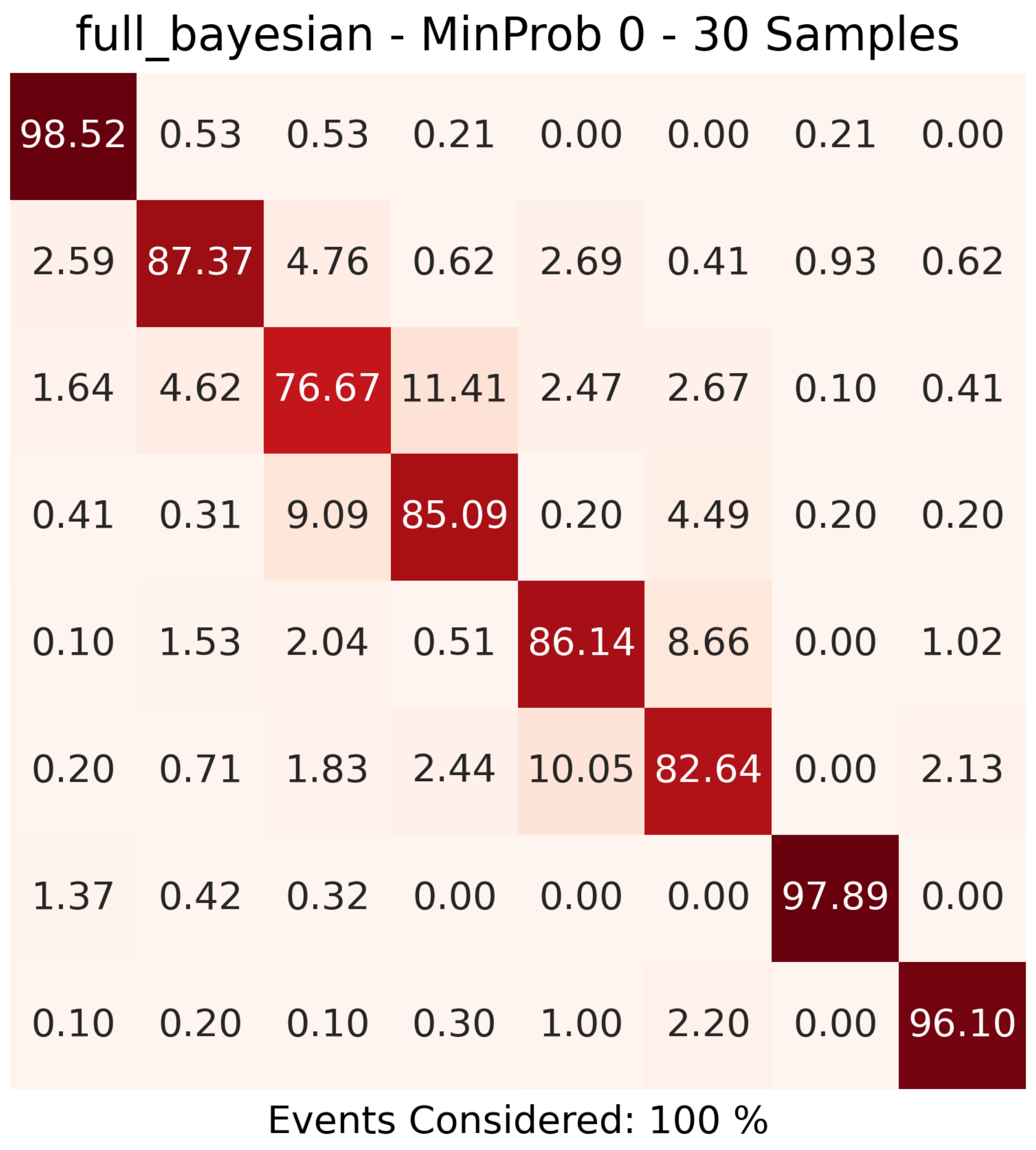
Logits

S.Giagu, M. di Filippo, L.Torresi paper in preparation …

RESULTS B-DCGNN

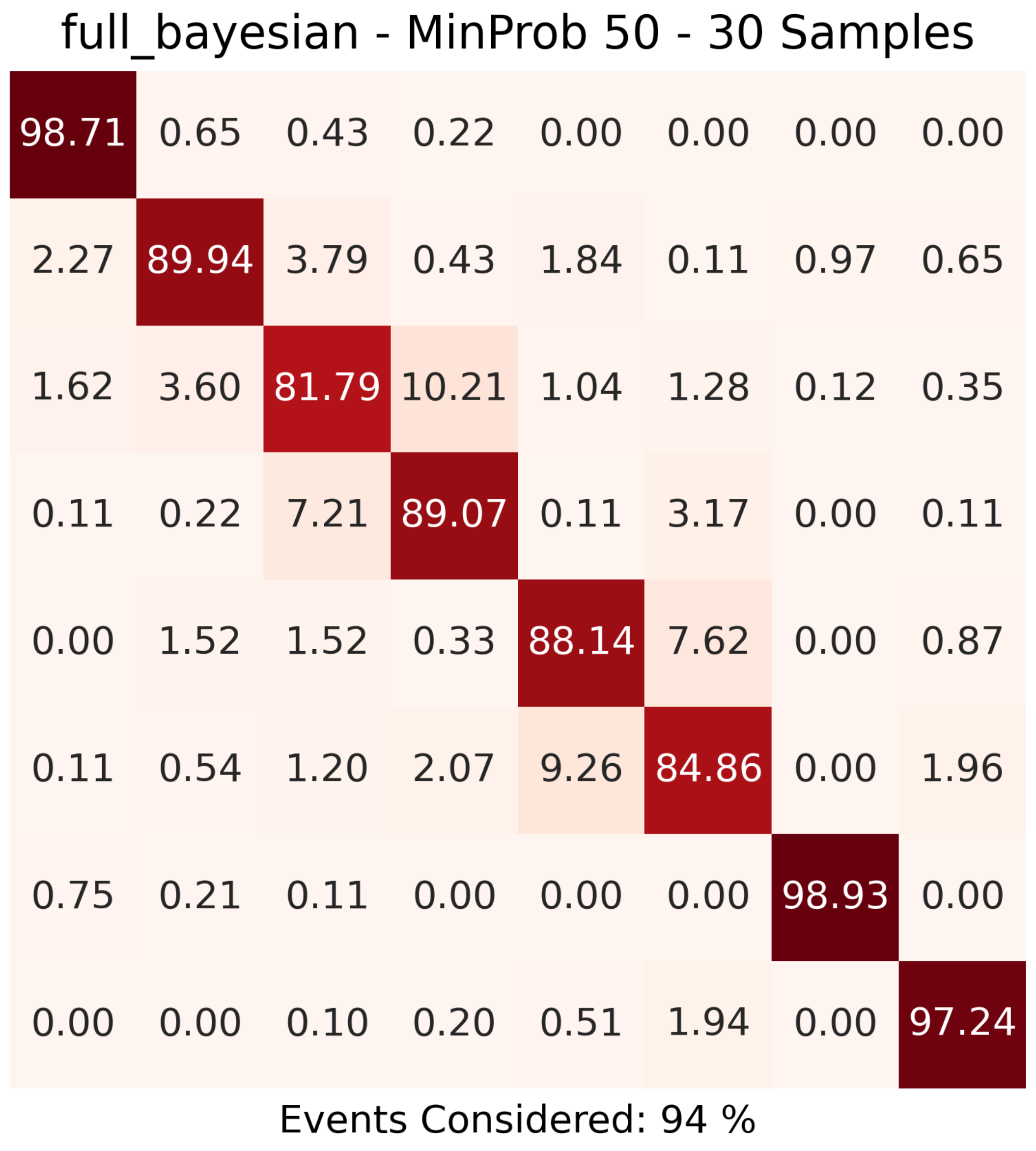
- same performance as with the DGCNN
- class probabilities better aligned with physics expectations

confusion matrix


full_bayesian - MinProb 0 - 30 Samples

98.52	0.53	0.53	0.21	0.00	0.00	0.21	0.00
2.59	87.37	4.76	0.62	2.69	0.41	0.93	0.62
1.64	4.62	76.67	11.41	2.47	2.67	0.10	0.41
0.41	0.31	9.09	85.09	0.20	4.49	0.20	0.20
0.10	1.53	2.04	0.51	86.14	8.66	0.00	1.02
0.20	0.71	1.83	2.44	10.05	82.64	0.00	2.13
1.37	0.42	0.32	0.00	0.00	0.00	97.89	0.00
0.10	0.20	0.10	0.30	1.00	2.20	0.00	96.10

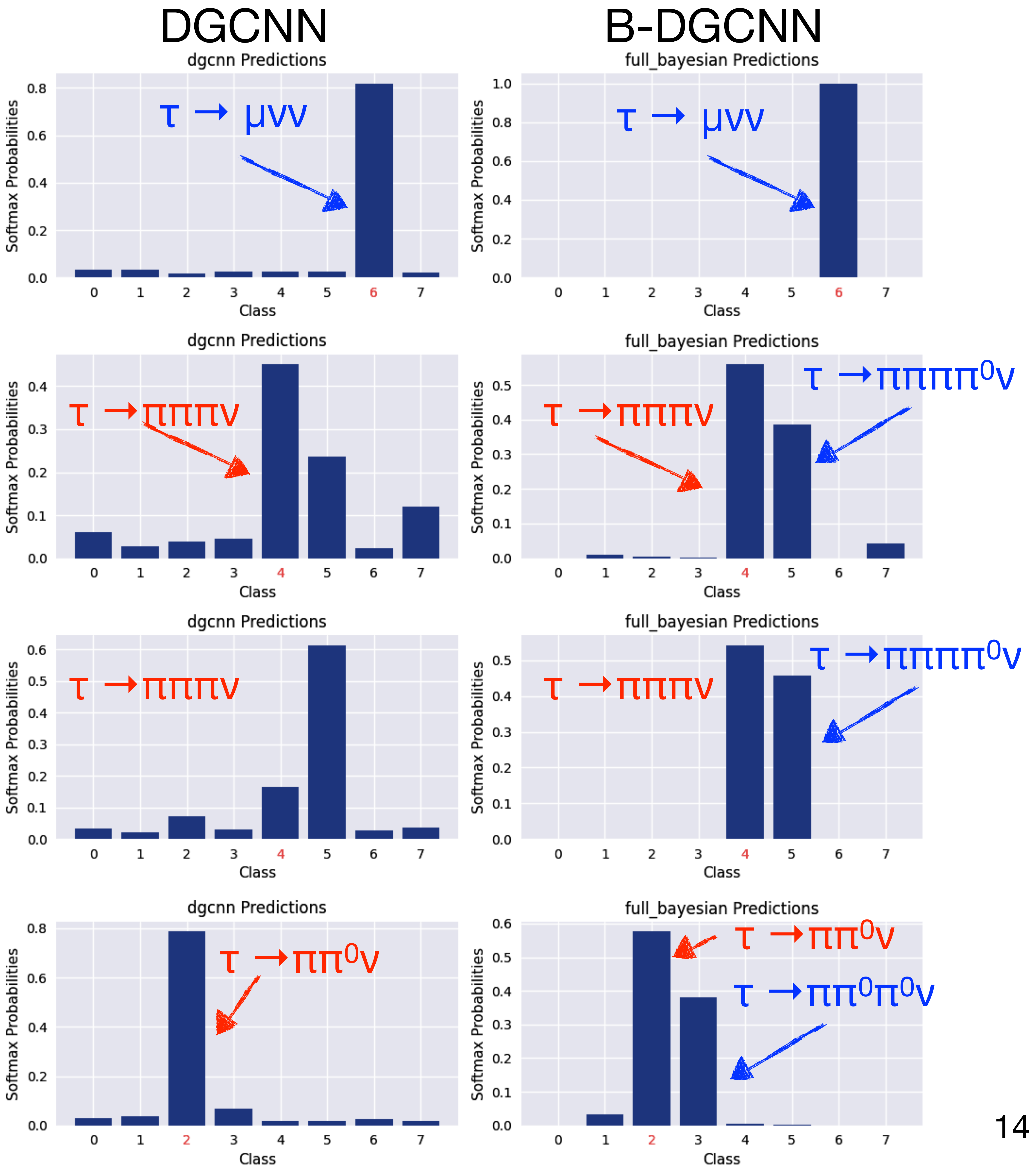
Events Considered: 100 %


full_bayesian - MinProb 50 - 30 Samples

98.71	0.65	0.43	0.22	0.00	0.00	0.00	0.00
2.27	89.94	3.79	0.43	1.84	0.11	0.97	0.65
1.62	3.60	81.79	10.21	1.04	1.28	0.12	0.35
0.11	0.22	7.21	89.07	0.11	3.17	0.00	0.11
0.00	1.52	1.52	0.33	88.14	7.62	0.00	0.87
0.11	0.54	1.20	2.07	9.26	84.86	0.00	1.96
0.75	0.21	0.11	0.00	0.00	0.00	98.93	0.00
0.00	0.00	0.10	0.20	0.51	1.94	0.00	97.24

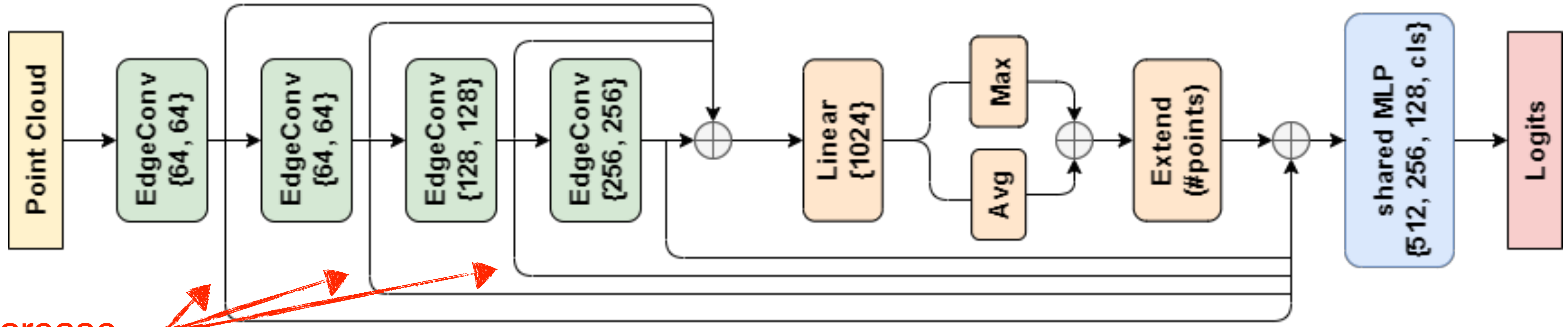
Events Considered: 94 %

distributions of softmax class probabilities sampling the model multiple times



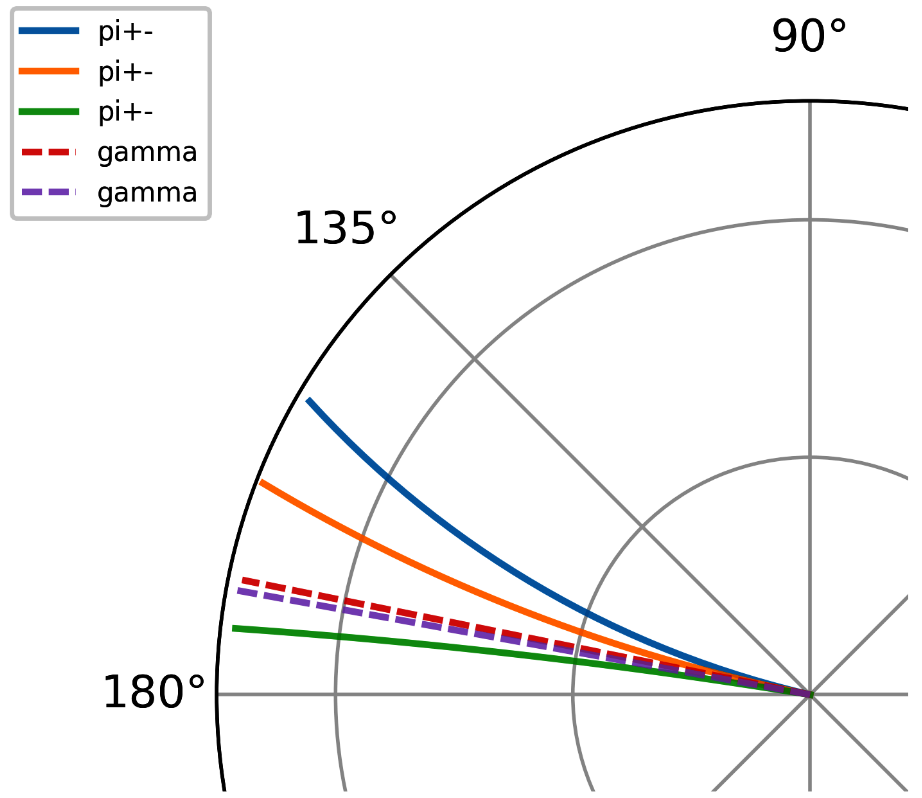
SEGMENTATION

- DGCNN and dual-readout calorimeter high granularity can also be exploited for object (particle) detection inside taus and jets
  - a proto-step for a particle flow algorithm for taus and jets
  - a similar approach as in segmentation in medical imaging (CT, MRI, …)
- DenseNet like modification of the DGCNN architecture for a segmentation task:



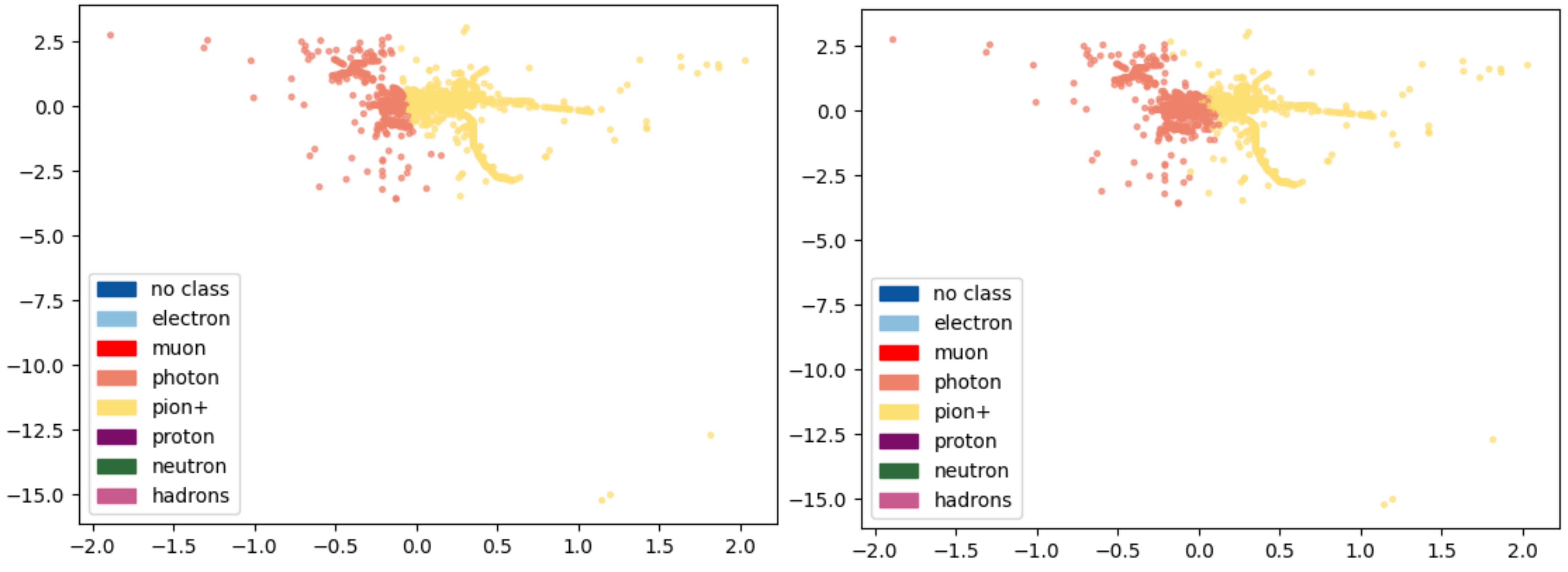
skip connections to increase sensitivity to multiscale features …

- identify the particle associated to the larger energy deposit in each fibre
- label each fibre by extrapolating Monte Carlo truth particles from production to the DRC into the IDEA magnetic field
- train the DGCNN to predict the label associated to each fibre
- **Ongoing study:** initial tests only on photons/neutrons VS other particles identification in tau decays

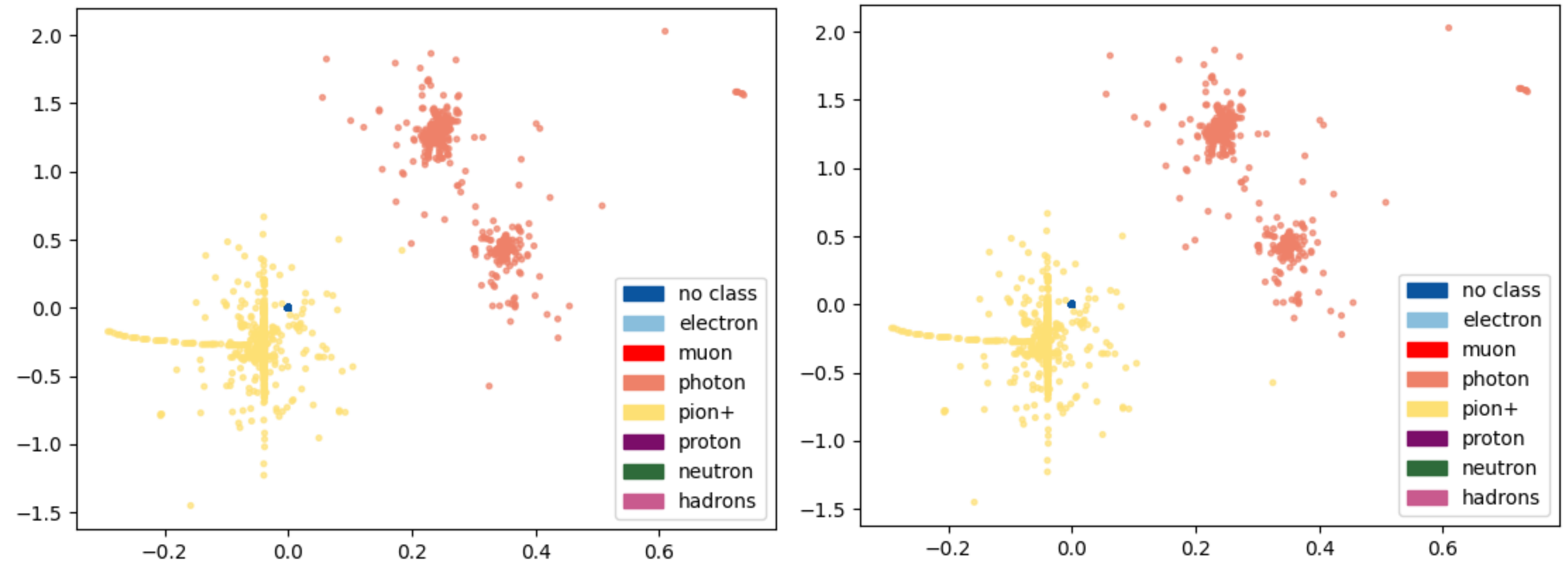


RESULTS SEGMENTATION

Example: segmentation of two $\tau \rightarrow \pi\pi^0\nu_\tau$ events



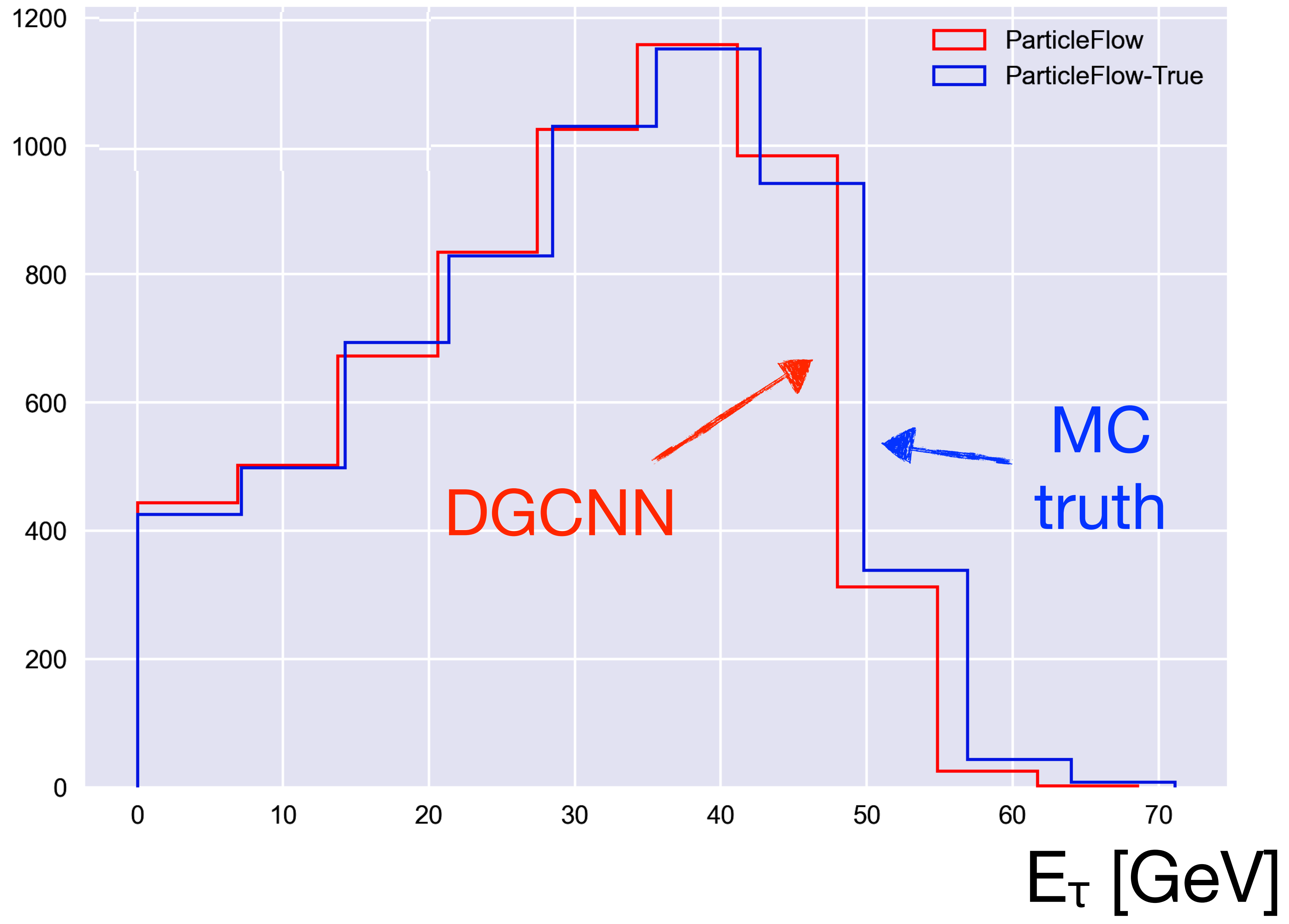
Ground Truth Reconstructed



tau visibile energy reconstructed using:

- DRC for photons
- MC truth for other particles

comparison of the distributions obtained when photons are identified by the DGCNN and when using the MC truth



SUMMARY

Very good performances in tau leptons identification obtained by leveraging geometrical deep learning models (DGCNN) using standalone dual readout calorimeter of the IDEA concept detector

- results based on full GEANT4 simulation of the IDEA detector geometry including B field, solenoid material effects, and parametric simulation of SiPM readout
- 91% average identification accuracy for a 8-class classification of QCD jets and leptonic and hadronic tau decay modes (88% using only geometrical information (fibre positions and types))
- developed a Bayesian-DGCNN for robust estimation of model prediction and uncertainties with comparable performances as the conventional DGCNN
- ongoing: identification of γ and n inside hadronic tau decays and QCD jets for proto particle-flow, promising initial results

extension the developed techniques with the use of the whole IDEA detector will follow soon …