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

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

RD_FCC collaboration meeting
15-16 Dec 2021

SAPIENZA
UNIVERSITÀ DI ROMA

Introduction

Case study: τ-identification in the IDEA dual-readout calorimeter (DRC)

- leverage modern machine learning methods based on differentiable deep neural networks
- study performance using only standalone DRC information
- helps in optimizing the detector and design of the readout electronics

Task 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 (e.g. identification of γ and n inside hadronic tau decays) for particle-flow algorithms

Detecting the Tau Lepton

Tau lepton will play a key role in precision measurements and searches of new physics in future high-energy colliders.

Due to the short lifetime ($\tau_\tau \sim 2.9\times10^{-13} s$) it can be detected through its decay products, that can be both leptons or hadrons.

An efficient energy reconstruction is fundamental for the tau identification.

Main τ decays

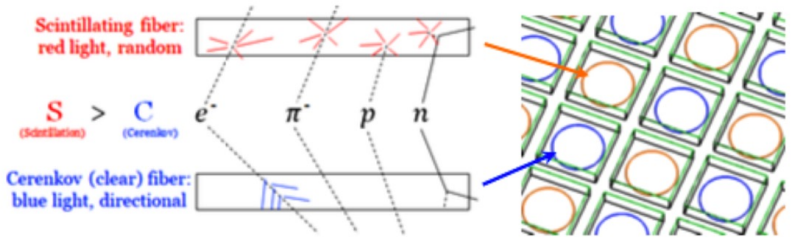
τ

- $\mu\ \nu\nu$
- $e\ \nu\nu$
- $\pi^{\pm}\ \nu$
- $\pi^{\pm}\ \pi^{0}\ \nu$
- $\pi^{\pm}\ \pi^{0}\ \pi^{0}\ \nu$
- $\pi^{\pm}\ \pi^{\pm}\ \pi^{\mp}\ \nu$
- $\pi^{\pm}\ \pi^{\pm}\ \pi^{\mp}\ \pi^{0}\ \nu$

DRC Principle

The electromagnetic fraction f_{em} of an hadronic shower inside the calorimeter is a random variable. Fluctuations are large and affects the final energy resolution.

The Dual Readout Method allows to measure the f_{em} of each event detecting the Scintillation signals (coming from em and hadronic component) and Cherenkov signals (mainly coming from em component).



$$S/E = (h/e)_S + f_{\mathrm{em}}[1-(h/e)_S]$$
$$C/E = (h/e)_C + f_{\mathrm{em}}[1-(h/e)_C]$$

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

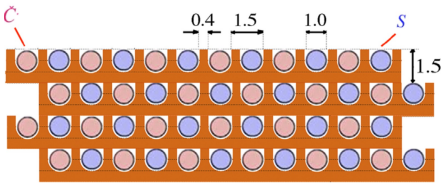
Different patterns of the two signals from different particles - optimal energy reconstruction and, combined with the fine segmentation of the fibres, powerful particle identification.

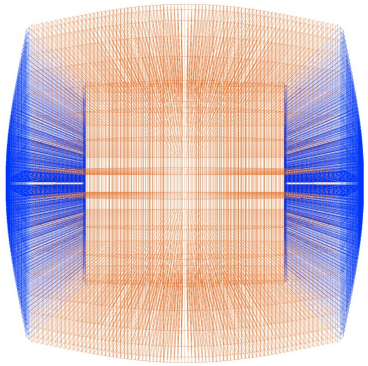
IDEA DRC Simulation

Full Geant4 simulation of the calorimeter geometry:

- Includes B field and solenoid material in front of the calorimeter.
- Fiber-sampling calorimeter: Cu absorber - 1 mm fibers.
- Each fiber read out by a dedicated SiPM.
- A total of 130M fibers, providing excellent granularity and lateral shape sensitivity:

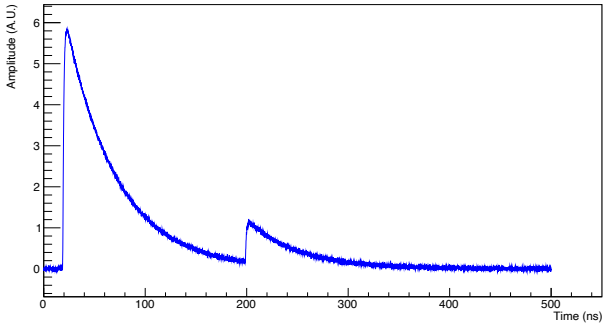
$$\Delta\theta, \Delta\phi = \sim 0.035°$$





Parametrised simulation of SiPM readout and signal processing:

- Dark counts, crosstalk, afterpulses…



Dataset

e^+e^- collision simulated with Pythia8 and signal events with Geant4.
5000 Events for each decay mode.

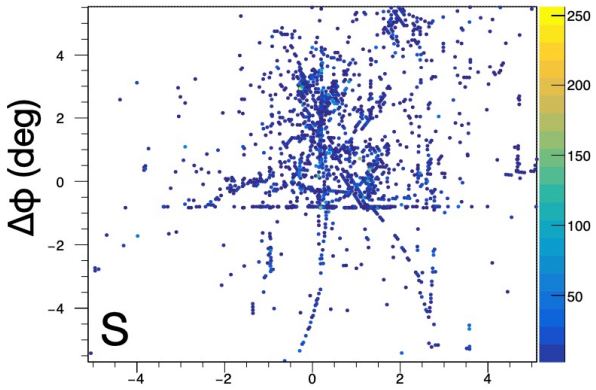
Information available for each event are:

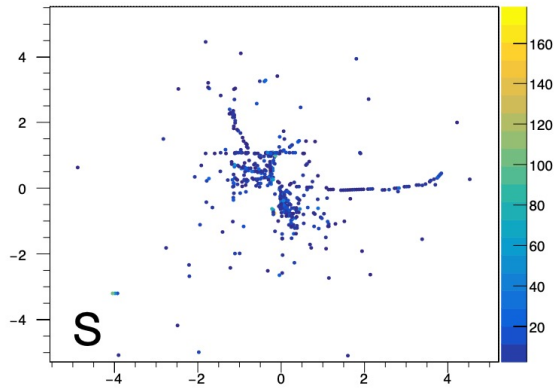
- **Geometric**: θ and ϕ of each fired fiber
- **Fiber Type**: S or C
- **Energetic**: number of photoelectrons in the fibers (used only in preliminary tests)
- **SiPM data**: Integral and Peak of the signal, Time of Arrival, Time over Threshold, Time of Peak
- **Channel Decay Label**

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

Events examples (Full Granularity)

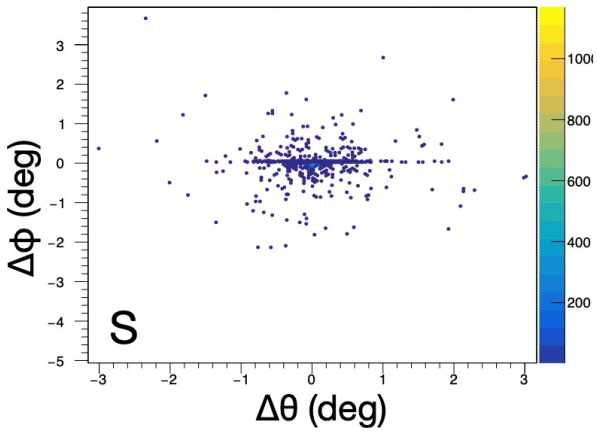
Z→qq Jets

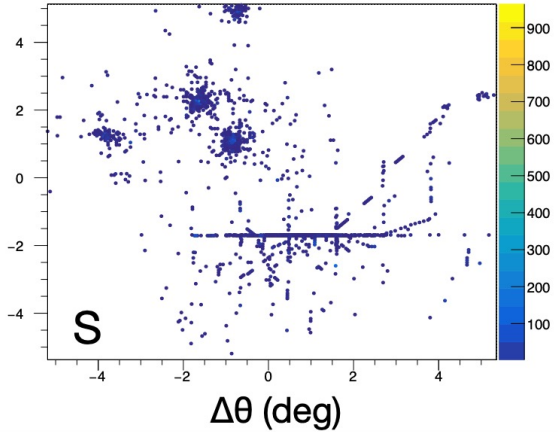




τ →πν

τ →evv





τ →ππ⁰π⁰ν

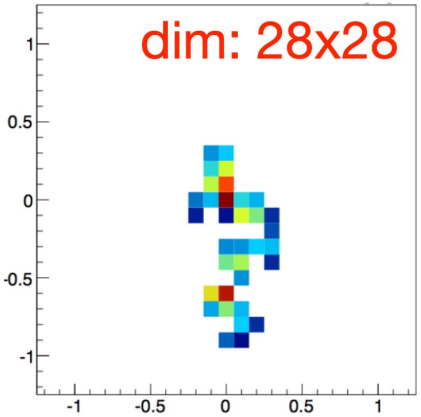
Data Representation

Image-based: energy deposition on each fiber is seen as the pixel intensity of an image of the event in fixed-shape mesh.

+ Can be used very advanced Convolutional Neural Network developed for image identification.

- Unclear how to incorporate additional information of the fibers.

- Very sparse representation: jets/tau decays have O(10) to O(100) particles → more than 90% of the pixels are blank.

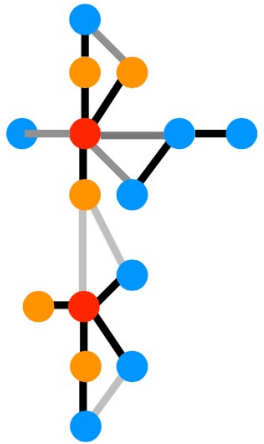


Point Cloud: Unordered sets of entities distributed irregularly in space, analogous to the point cloud representation of 3D shapes.

+ Clouds allow rich internal structures.

+ Easy to incorporate additional information of the fibers.

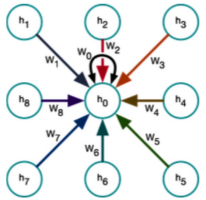
- The architecture of the neural network must be carefully designed to fully exploit the potential of this representation.
→ Dynamic Graph CNN



Edge Convolution Operation

Regular convolution operations cannot be applied on point clouds:

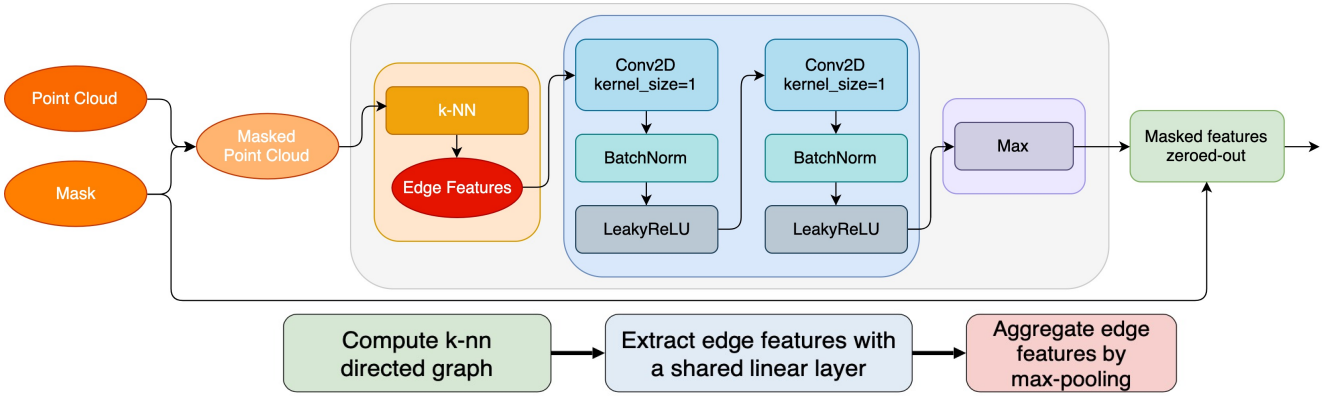
- points distribution is usually irregular (unlike uniform grids of the pixels in an image)
- they're not invariant under permutation of the points



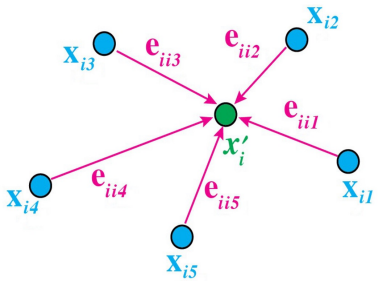
CNN-conv

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

Solution: using the points of the cloud to generate **graphs** with **vertices** (the point themselves) and **edges** (connections between each point to its k-nearest neighboring points) → **regular distribution** for each point.



EDGE-conv



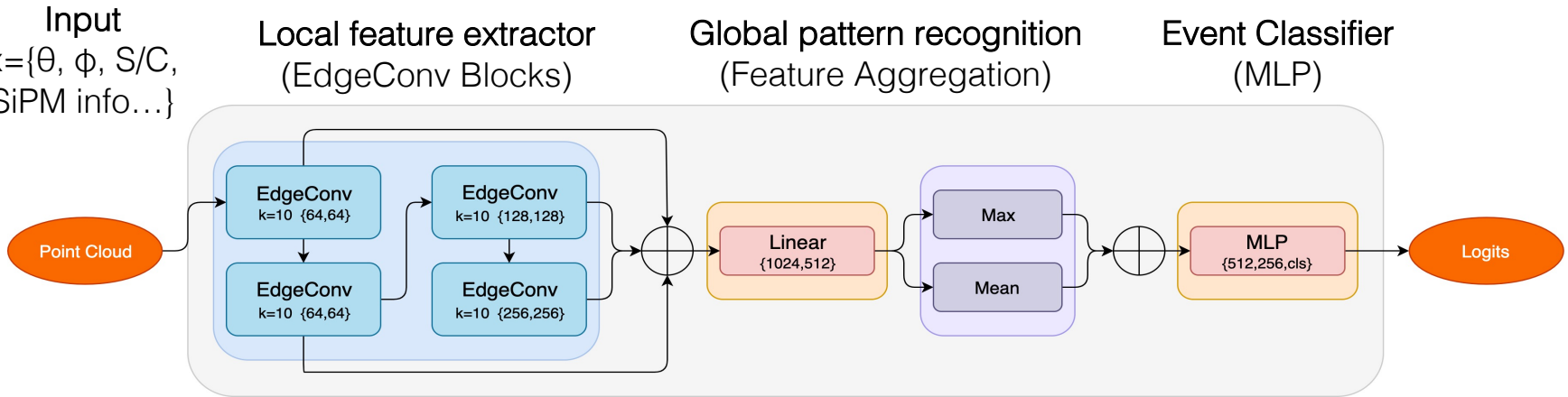
$$e_{i,i_j} = h_\Theta(\mathbf{x}_i, \mathbf{x}_{i_j})$$

$$\mathbf{x}'_i = \square_{j=1}^{k} h_\Theta(\mathbf{x}_i, \mathbf{x}_{i_j} - \mathbf{x}_i)$$

DGCNN

DGCNN is the neural network that stacks several EdgeConv blocks, creating high-level graphs and capturing complex patterns.

The resulting architecture of the DGCNN shows the different steps:



The number of input fibers is fixed and treated as model hyper parameter, discarding those with lowest signals or adding zero valued vectors in case of events with lower active fibers.

Tau Decay Identification Task

Classification Task:

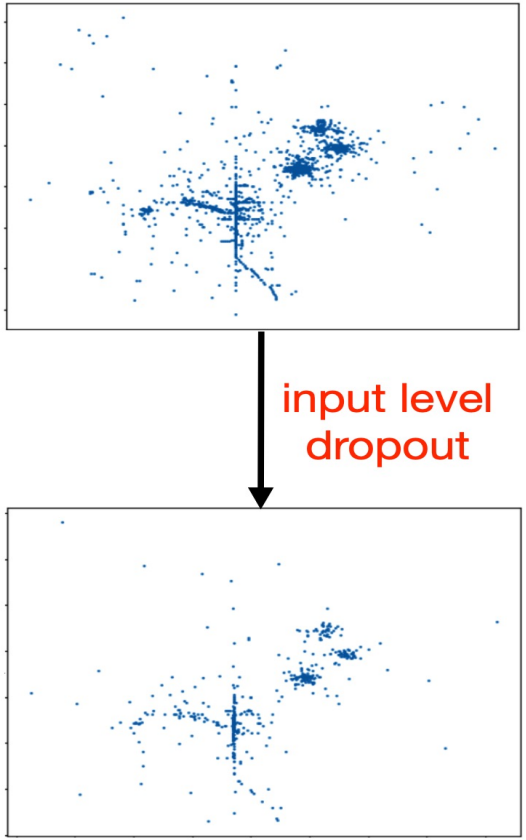
- 8-classes: 7 tau decays + QCD jets
- training/validation/test sets: 22k/6k/7k events (balanced among classes)

Data-preprocessing:

- simple geometrical clustering, no specific selection or fiducial volume applied
- saved fibers signal around the clusters

Data augmentation/regularization: overfitting and memorization for the DNN model controlled using:

- *at input level:* some of the fired fibers are switched off
- *in the neural network layers:* some of the parameters of the last MLP block are randomly zeroed during the training phase

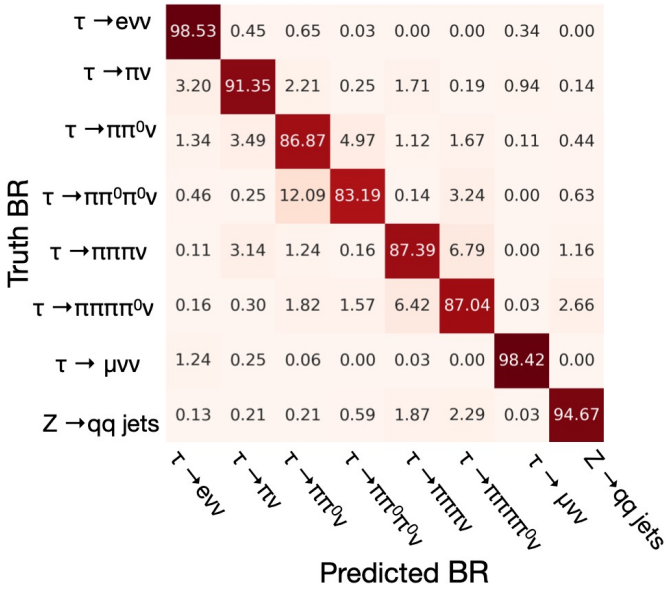


Tau Decay Identification Performance

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

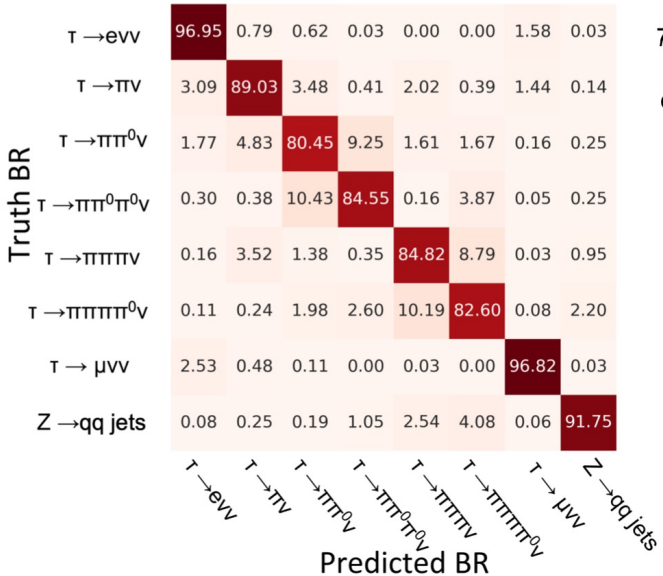
using **coordinates, type** of fibre, and **#of photo-electrons** in each fibre

average accuracy: **90.8%**



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

average accuracy: **88.3%**



73.7% using only the coordinates

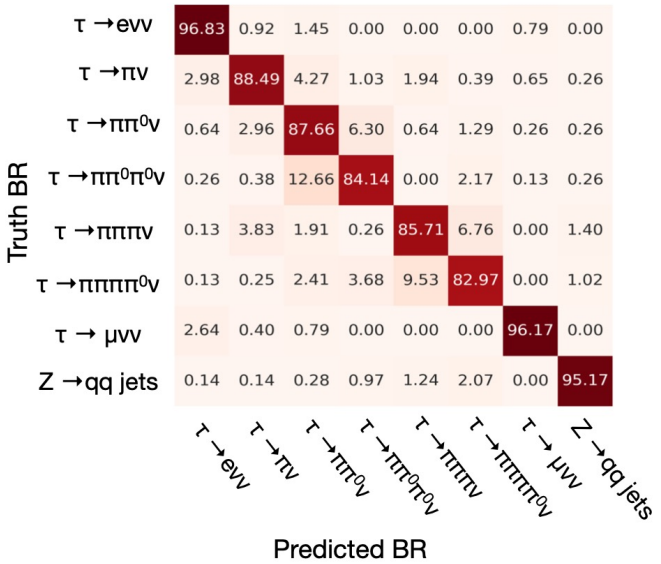
double-readout geometry alone allows excellent tau identification

Tau Decay Identification Performance

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

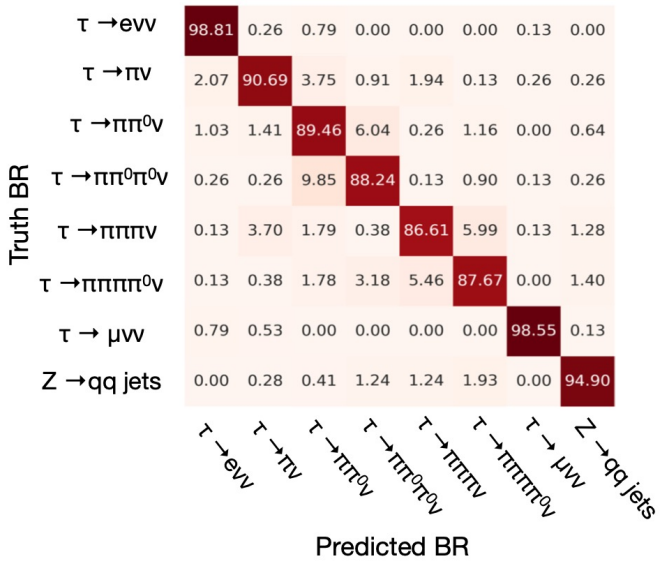
using only geometrical and Integral/Peak of the signal

average accuracy: **88.8%**



adding also SiPM timing information

average accuracy: **90.8%**



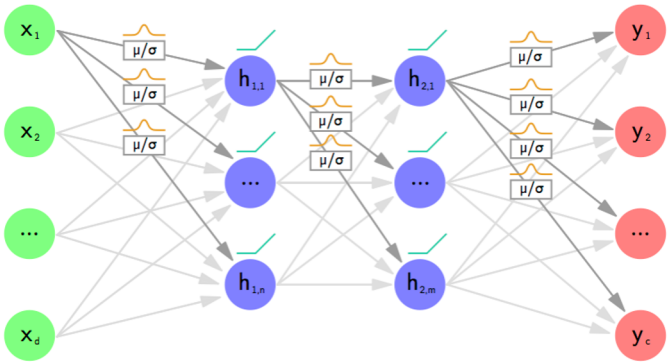
comparable identification performance with input from SiPM emulation

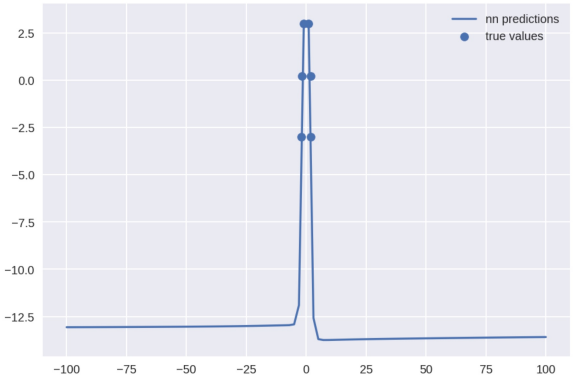
Uncertainty in the classification

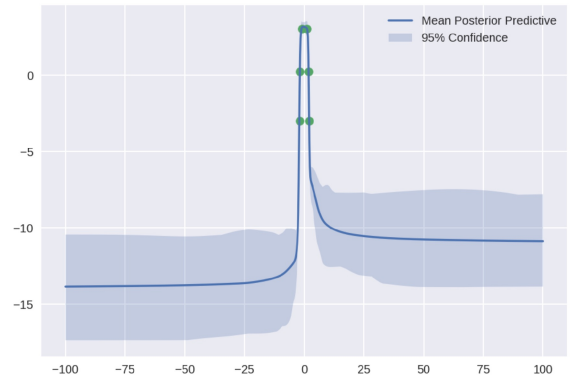
Neural networks based on point values for weights suffer of **overconfidence** when analyzing new data.

Bayesian learning: introduce probability distributions over the weights, providing an estimate on the confidence for the final prediction, **predicting distributions instead of point values.**

The algorithm uses a variational approximation $q(\boldsymbol{w}|\theta)$ to the true posterior $P(\boldsymbol{w}|\mathcal{D})$ to predict an estimate of the expected value $\mathbb{E}_{P(\mathbf{w}|\mathcal{D})}\left[P(\mathbf{y}|\mathbf{x},\mathbf{w})\right]$



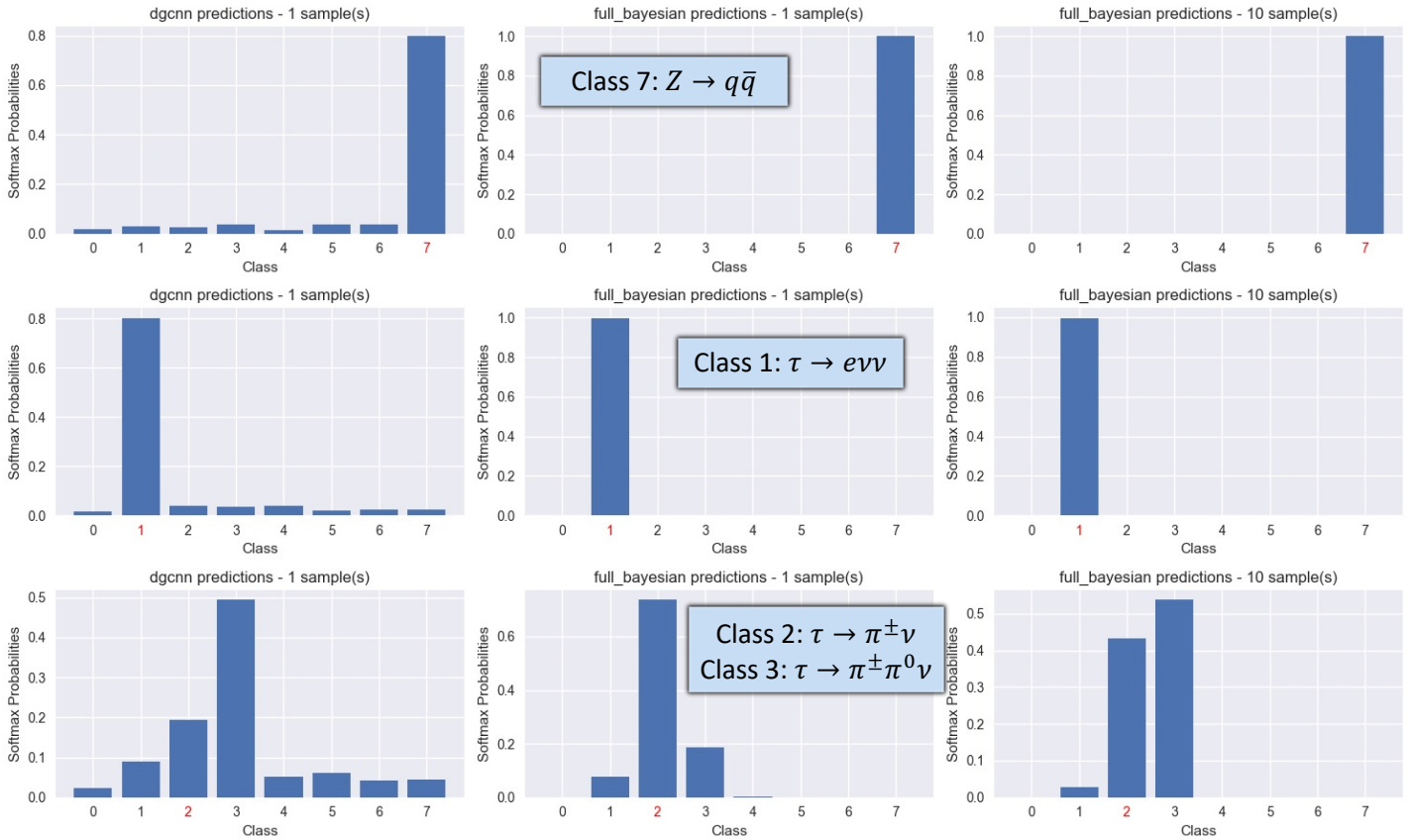




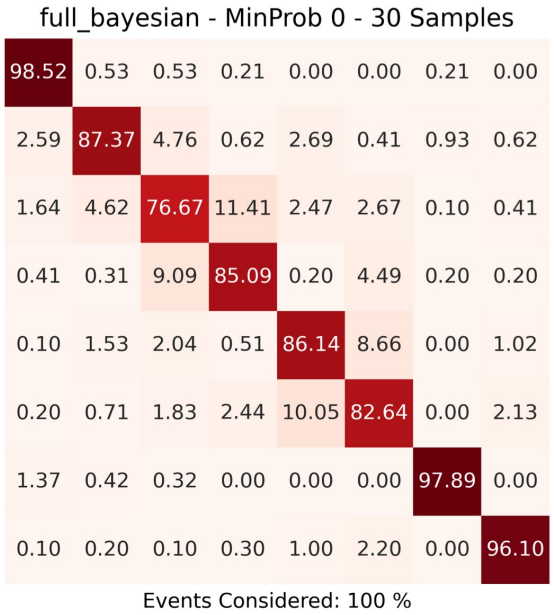
B-DGCNN Predictions

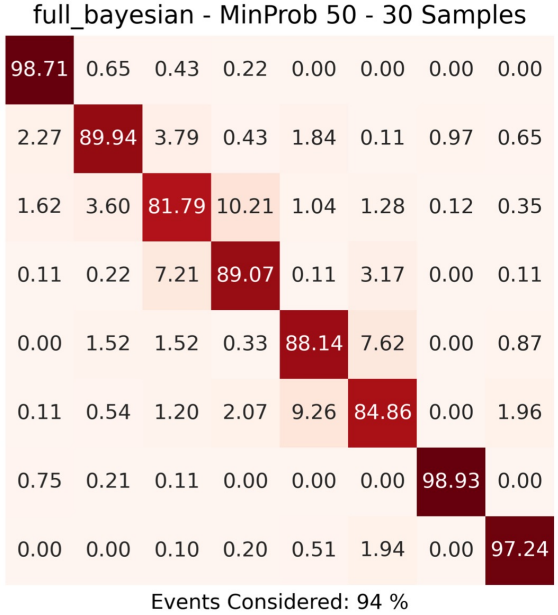
Comparison of the same 3 random events evaluated by:
DGCNN
B-DGCNN (1 sample)
B-DGCNN (10 samples)

With the bayesian approach, probabilities are better aligned with physics expectations.



B-DGCNN Results





Applying **thresholds on the minimum confidence** that the network must have to keep the prediction.

Minimum Confidence	Events Considered	Test Accuracy
0.5	94.72 %	0.900
0.6	87.57 %	0.925
0.7	79.82 %	0.947
0.8	72.21 %	0.964
0.9	60.52 %	0.981

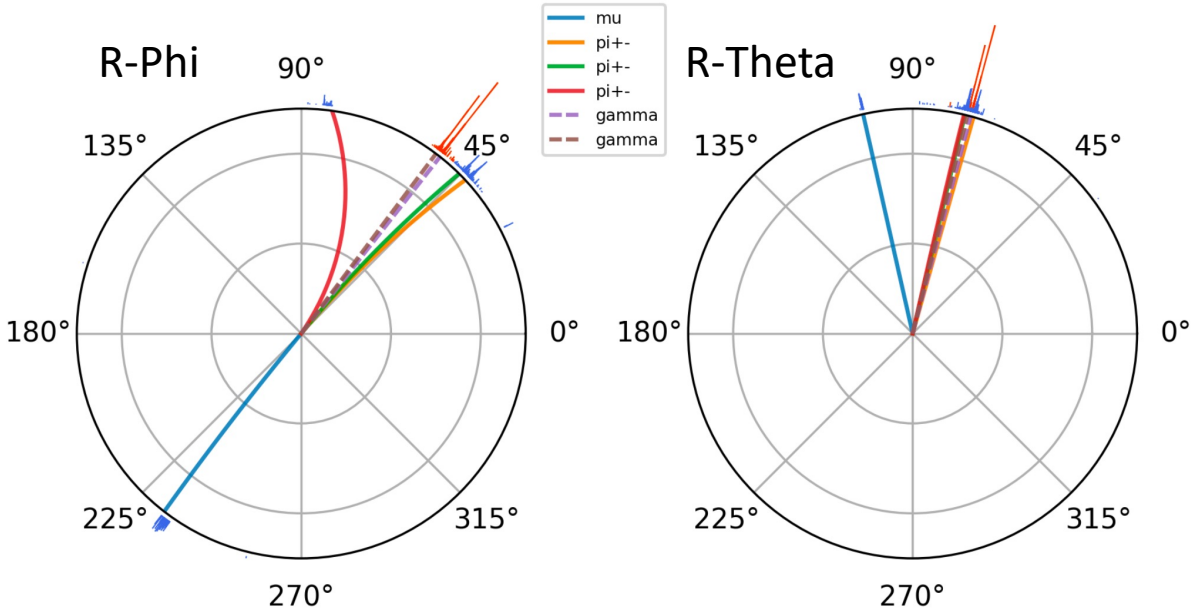
Segmentation

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

Fiber-Particle Label: obtained extrapolating Monte Carlo truth particles from production to the DRC into the IDEA magnetic field.

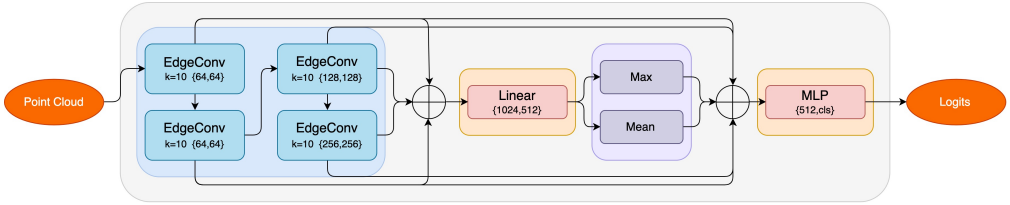
Ongoing study: initial tests only on photons/neutrons VS other particles identification in tau decays.

$$\tau \rightarrow \pi^{\pm}\pi^{\pm}\pi^{\pm}\pi^{0}$$

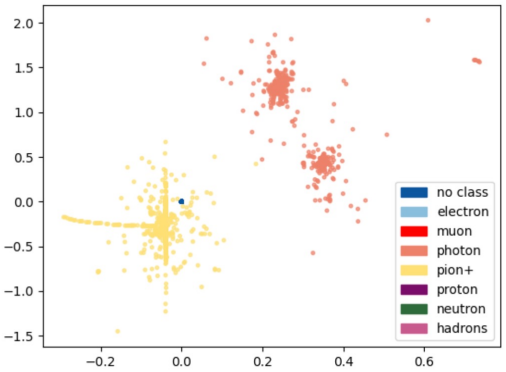


Segmentation Results

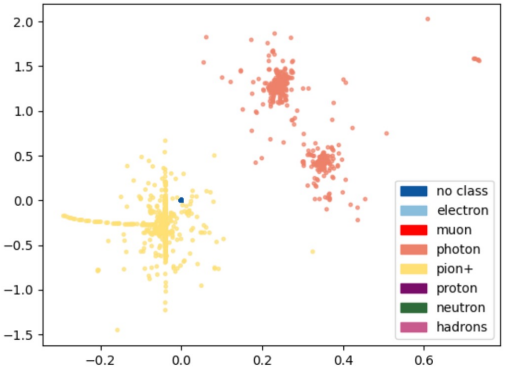
DGCNN architecture similar to the previous one.
Implemented *skip connections* to increase sensitivity to multiscale features.



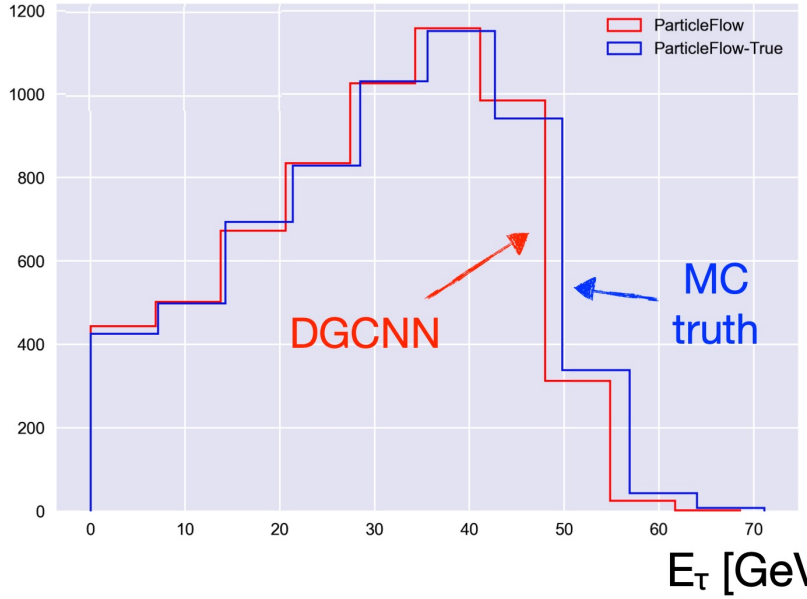
Ground Truth



Reconstructed



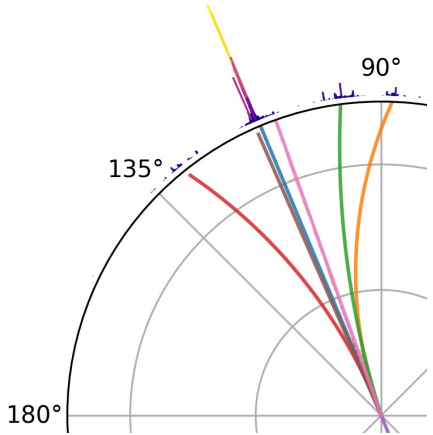
Tau visible energy reconstructed using:
- DRC for photons
- MC Truth for charged particles



Future Implementations

Introduce Drift Chamber information in the input Point Cloud:

- Point Clouds allows an easy implementation of other detector signals
- Including tracking data for the charged particles provides a better reconstruction of the full event and would improve the Neural Networks performances



Improve the Segmentation model for an efficient ParticleFlow algorithm:

- Optimize the hyperparameters tuning
- A new simulation with a MC matching of Particle-Fired Fiber would help the model to find the optimal patterns for the task

Model in FCCee framework software

All the models are designed and implemented using the PyTorch framework.

PyTorch provide a simple way to convert models into the ONNX (Open Neural Network Exchange) format. ONNX is an open format built to represent machine learning models, used also in the Atlas experiment.

Export by tracing: using the `torch.onnx.export()` function, it will execute the model, recording a trace of what operators are used to compute the outputs.

ONNX Runtime: it is possible to instantiate a *ONNX Runtime Session* in a C++ environment and load the converted trained model.

Since the data preprocessing is very simple, this approach can be exploited to implement the full DGCNN algorithm in the FCCee framework software.

Summary

- Developed a new Geometric DL algorithm exploiting data simulated in the IDEA Dual Readout Calorimeter from e^+e^- collisions.
- Accuracies reached in the tau decay channels and QCD jets identification overtake typical performances obtained with conventional techniques (~90% vs 70-80%).
- Implemented a probabilistic content in the neural network that provides probability distributions as prediction outputs, without loss of performances.

- Promising results obtained using the same architecture in the initial studies for segmentation tasks.