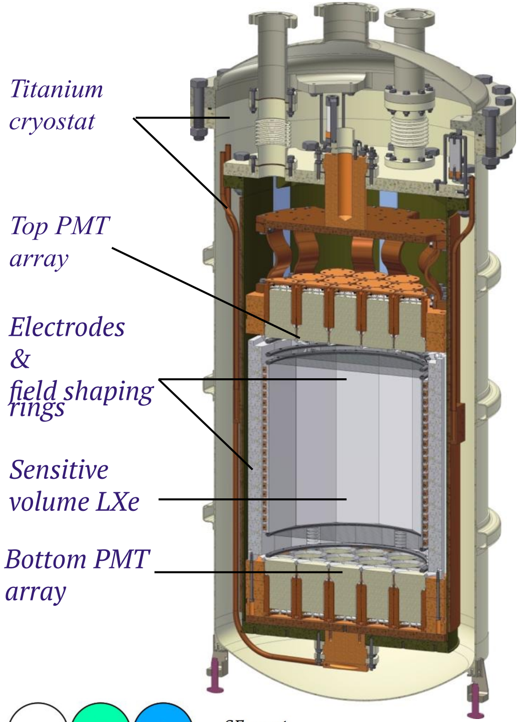
Pointlike events selection in the RED-100 experiment using ML algorithms.

Olga Razuvaeva on behalf of the RED collaboration

NPML
2023

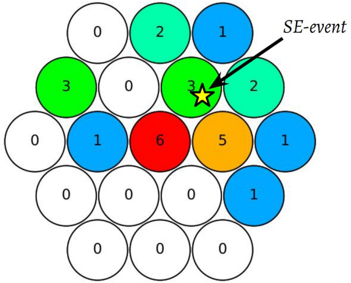
RED-100 experiment

- Two-phase noble gas emission detector
- Dedicated to study coherent elastic neutrino-nucleus (CEvNS) scattering
- Contains ~200 kg of LXe (~ 100 kg in FV)
- 2 arrays of PMTs
- Physical run on Kalinin NPP (Udomlya, Russia)

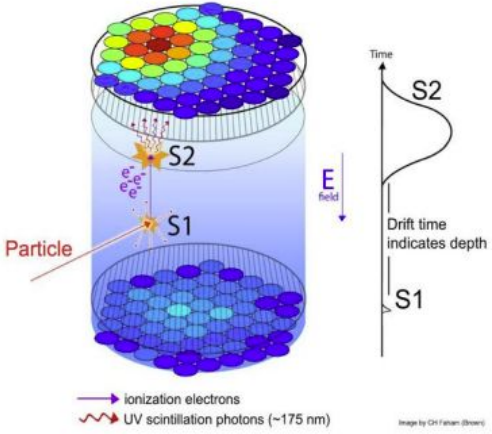


Example of simulated event (1SE)

The circles indicate the positions of the PMTs in the top array. Numbers in circles correspond to the numbers of photons from S2 detected by each PMT



Two-phase emission detector technique



Sensitive to the single ionization electron (SE) signal. CEvNS response is expected to be of several electrons.

more information — D.Rudik "The RED-100 experiment"

Background conditions

The RED-100 is working at shallow depth, unlike other similar detectors (LUX, Xenon1T).

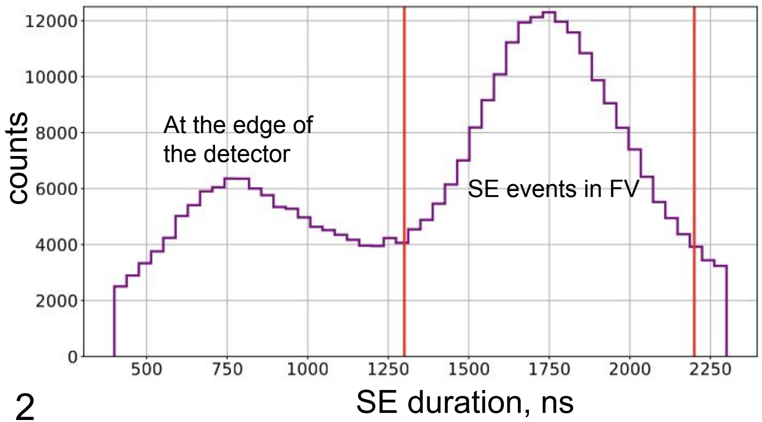
—high radioactivity level

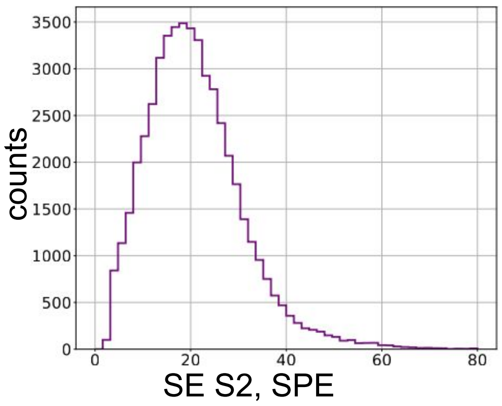
—significant background from spontaneous emission of SE

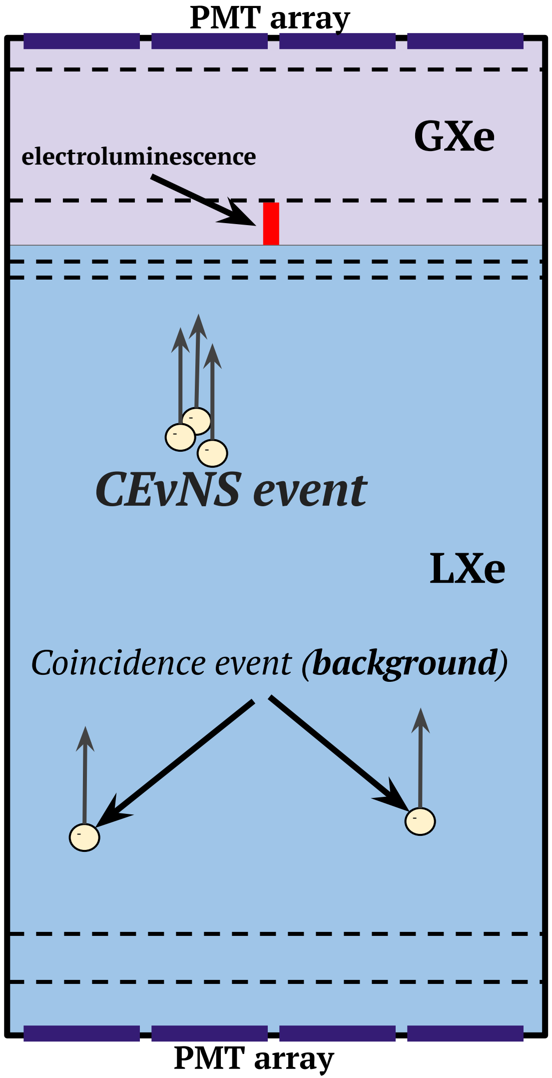
—effective cut is a need

Background event — coincidence of two or more spontaneous SE events (sometimes 2SE or 3SE).

CEvNS event — several electrons, coming from one point.



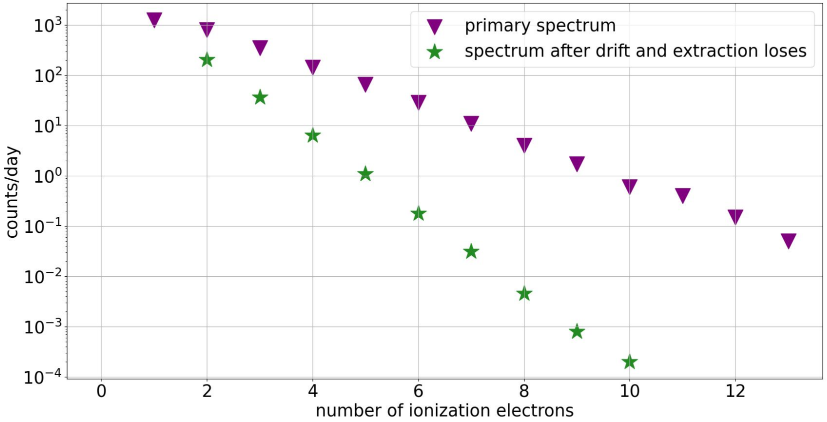




Simulation

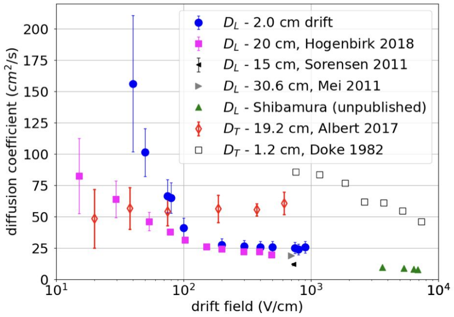
— ML solution requires training and validation data
— detailed modelling of events was performed

1. Recoil nuclei spectrum (GEANT4)
2. Ionization in LXe (GEANT4+NEST)
3. Electron drift in LXe (NEST+lifetime measured experimentally)
4. Diffusion
5. Extraction (NEST+experimental ionization yield)
6. Electroluminescence (NEST+experimental light yield)
7. Optical distribution (experimental light response functions (LRFs), see next slides)



Diffusion description:

$$n(\vec{x},t) = \frac{N}{4\pi D_T t\sqrt{4\pi D_L t}} \exp\left[\frac{-\left(x^2+y^2\right)}{4D_T t}\right] \times \exp\left[\frac{-\left(z-v_d t\right)^2}{4D_L t}\right]$$

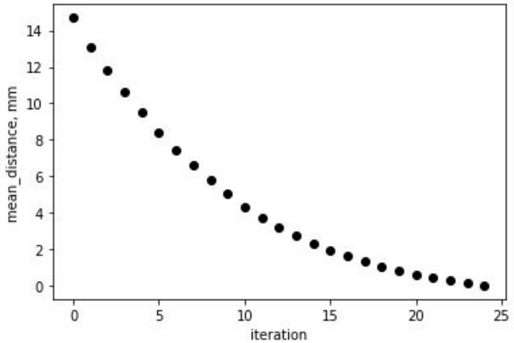


Measurements of electron transport in liquid and gas Xenon using a laser-driven photocathode, O. Njoya et. al https://arxiv.org/abs/1911.11580

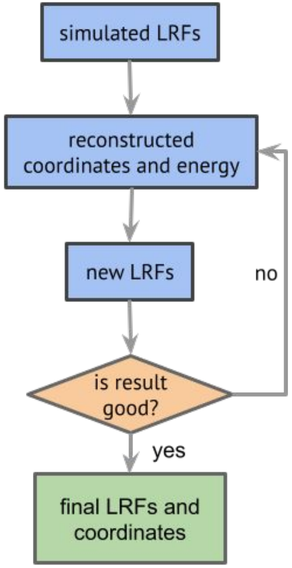
LRF calculation

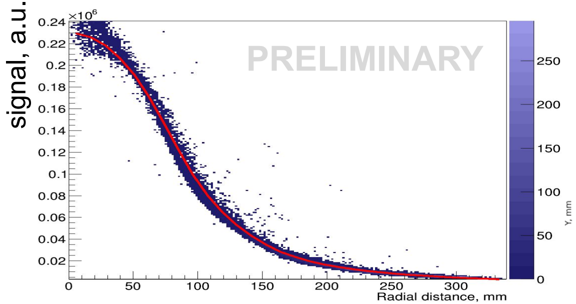
Reconstruction

— ANTS2 package for modelling and reconstruction
— we use light response functions (LRFs), that are the maps of signal vs light emission point for each PMT
— reconstruction algorithm is based on minimization of error between the observed signal distribution among PMTs and that expected from calculation using LRFs
— both s2 energy and coordinates are reconstructed

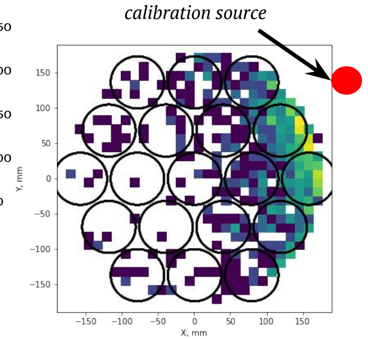


Mean distance between coordinates reconstructed on the i-th iteration vs coordinates reconstructed on the last iteration (LRFs with axial symmetry)

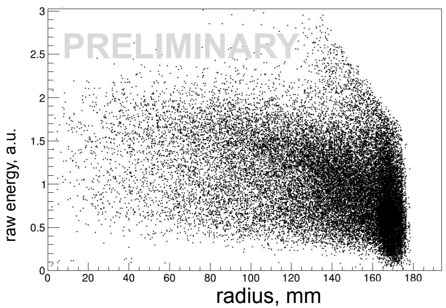




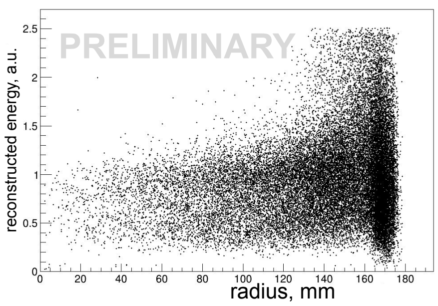
Example of LRF (red line) for PMT from second ring scaled on reconstructed energy



example of reconstructed XY distribution



raw S2 energy vs. reconstructed radius



reconstructed S2 energy vs reconstructed radius

SE signal simulation

1. Each SE-event position is chosen from uniform XY distribution

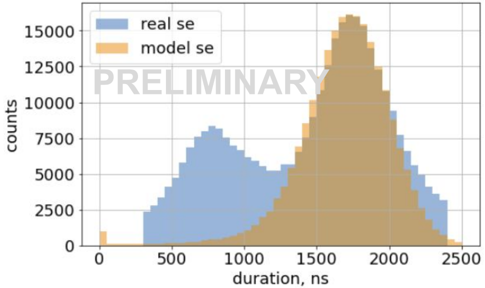
2. Number of photons per SE in the central area of the detector (27.4 photons) is scaled (using LRFs) depending on the position of the event

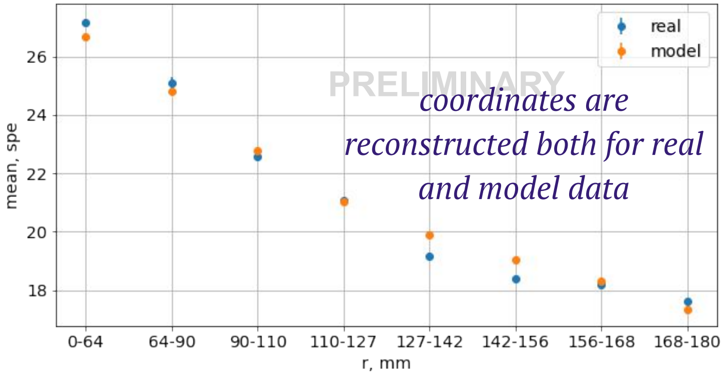
3. Final number of photons is calculated from normal distribution with
μ = *scaled number of photons*
$\sigma^2 = \mu + \sigma'^2$)
σ' is an addition sigma and it is calculated from real SPE distribution of SE events in the central area

4. Photons are distributed over the PMT with probabilities from LRFs

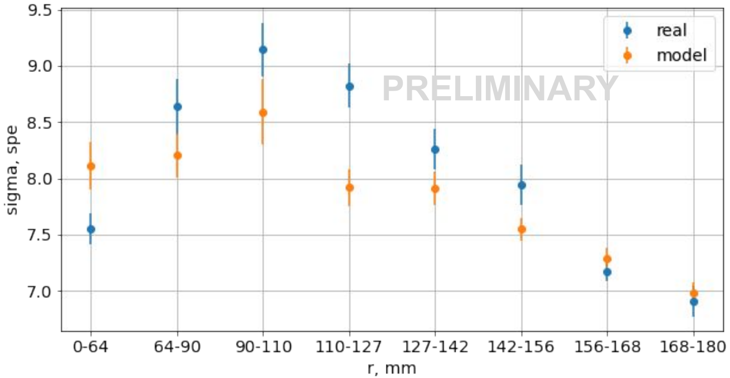
5. Duration was calculated from normal distribution with μ = *1830 ns*, σ = *230 ns*. Photons are distributed over event duration uniformly.

Every event consists of several SE events → if we can simulate SE, we can simulate everything!





coordinates are reconstructed both for real and model data



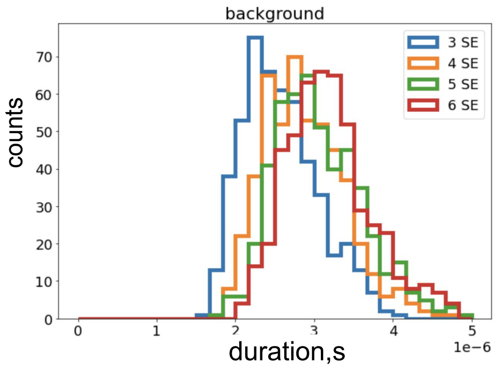
Dataset preparation

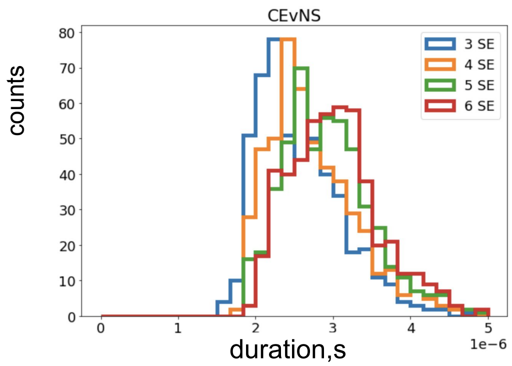
— pointlike (including CEvNS) events are constructed from several SE events time-shifted relative to each other in accordance to diffusion

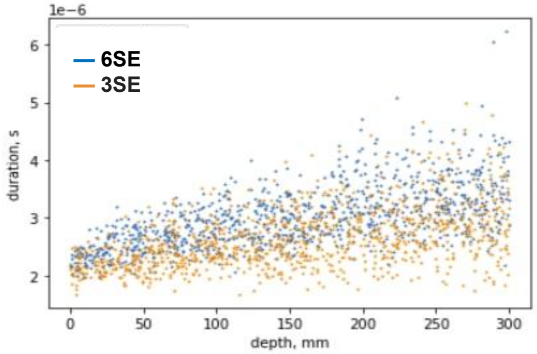
— Background events are constructed from 1SE, 2SE, 3SE pointlike events, uniformly distributed on depth and with uniform timeshifts

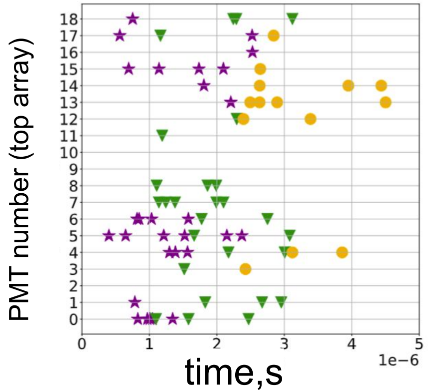
CEvNS:	**background**:
3 SE	[1+1+1] SE,[2+1] SE
4 SE	[1+1+1+1] SE,[2+1+1] SE,[3+1] SE, [2+2] SE
5 SE	[1+1+1+1+1] SE, [2+1+1+1] SE, [3+1+1] SE,...
6 SE	[1+1+1+1+1+1] SE, [2+1+1+1+1] SE,...

Event with less than 3 ionization electrons are under the threshold









Example of simulated event

Point with different colors indicates photons from different SE

Problem: separate pointlike (CEvNS) events from not-pointlike (background)

classifier based only on the light distribution

classifier based on the light and time distribution

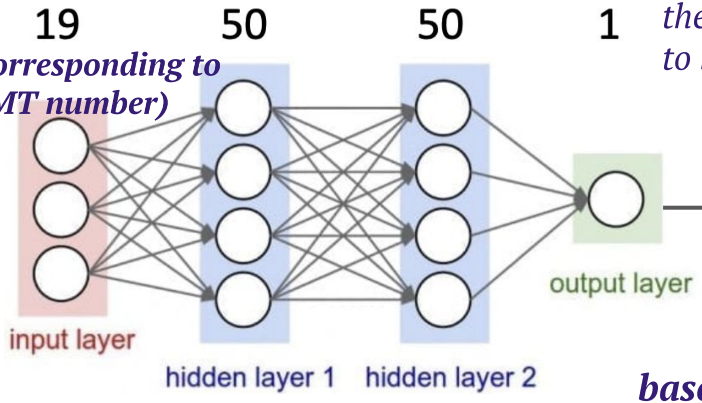
Deep learning neural network (DLNN)

Based only on the light distribution

Preprocessing
— The light response for each PMT was normalized to make a sum of 1 across PMT matrix
— NSE >3
— reconstructed radius<130 mm

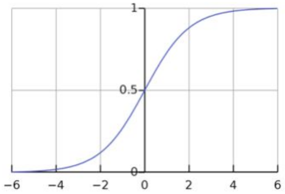
Train dataset (0.7 of all data):
~770k background events
~370k cevns events

— Bayesian optimization from *keras_tuner* was used on validation binary accuracy metric
— A common Adam optimizer was used with a BinaryCrossentropy loss function (other optimizers were also tested without any significant improvement)



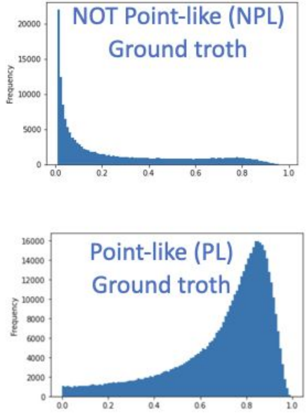
(corresponding to PMT number)

Output layer with a single neuron with sigmoid activation function to show the probability of the events to be pointlike



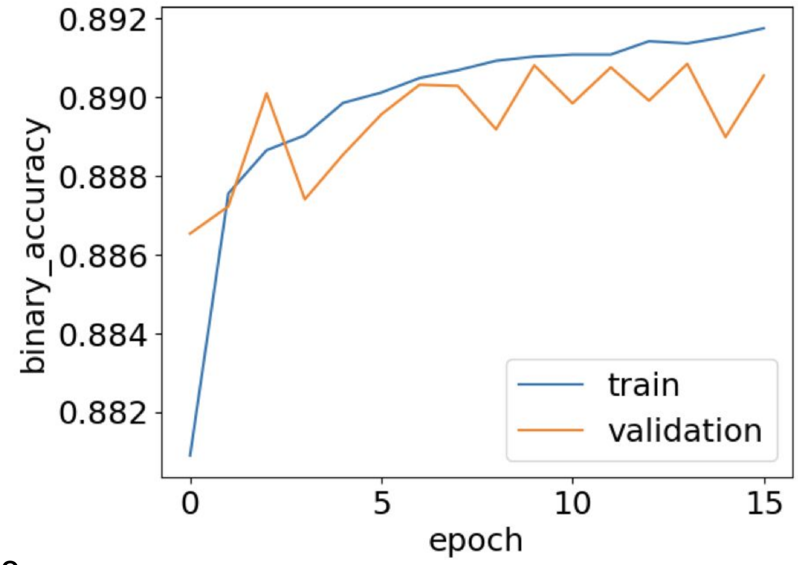
Optimized hyperparameters:
— Number of hidden layers
— Number of neurons in each layers
— Dropout/batch-normalization/no additional layers after each hidden layer
— Learning rate

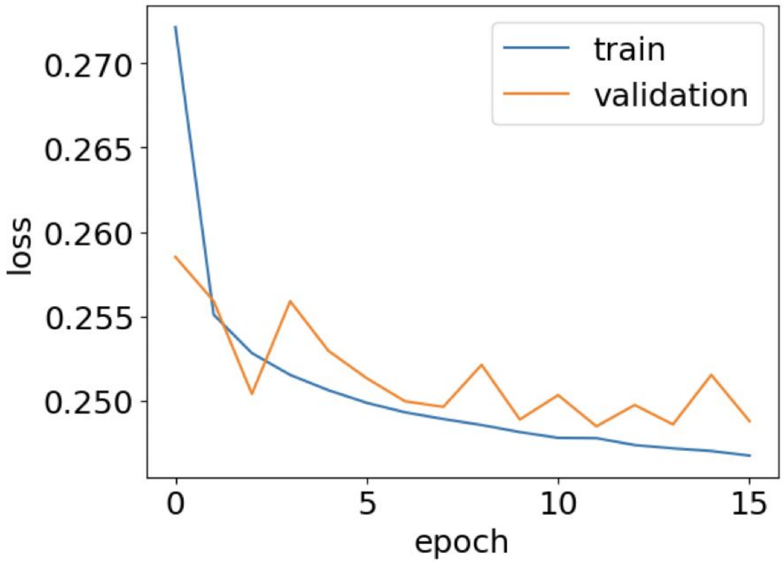
baseline configuration
(before optimization)



Deep learning neural network (DLNN)

The following DLNN structure was obtained after optimization:
— 4 hidden layers (70, 62, 72 and 44 neurons) with two batch-normalization layers after the first and third hidden layers
— Its standalone train and validation learning is presented with EarlyStopping on validation loss with patience of 4 and restoration of the best values

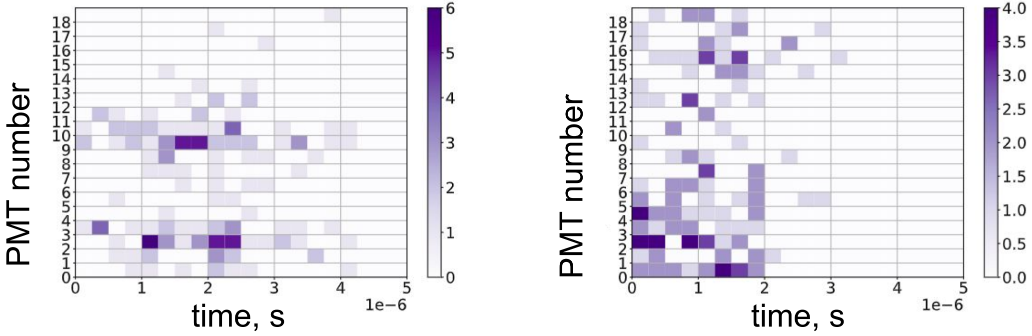




Convolutional neural network (CNN) #1

<u>Based on the light and time distribution</u>

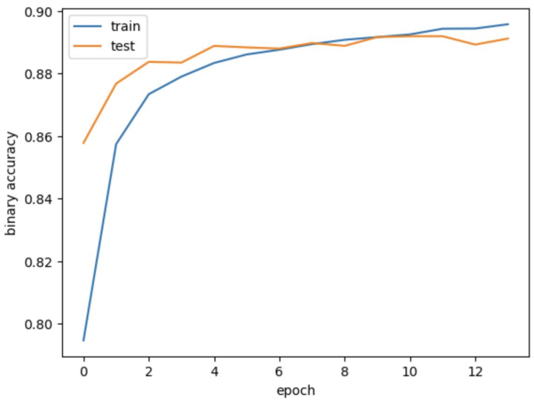
Preprocessing
— 19x19 pixels “pseudo-images” of event were constructed
— value in each pixel was divided by max signal

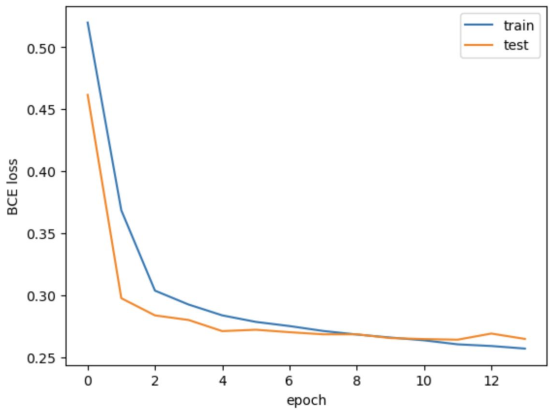


“pseudo-images” examples

Train dataset (0.75 of all data):
~300k background events
~300k cevns events

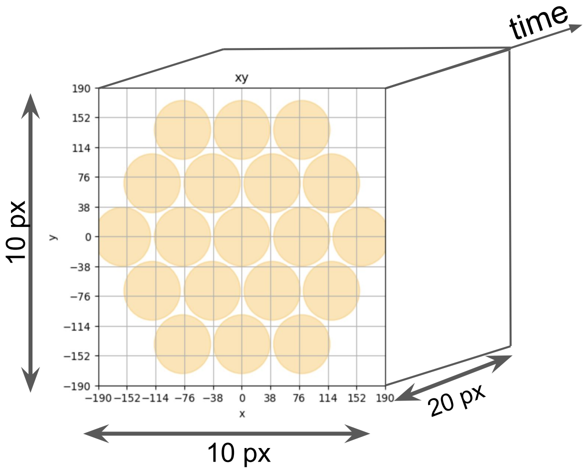
— 3 convolutional layers 3x3 with batch normalization after each other
— 4 fully connected layers
— Output layer with a single neuron with sigmoid activation function to show the probability of the events to be pointlike





Convolutional neural network (CNN) #2

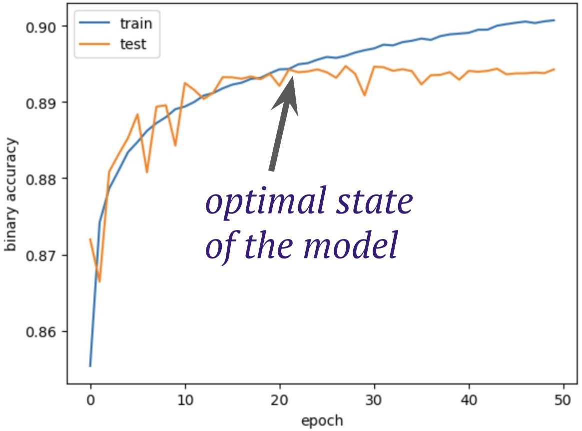
<u>Based on the light and time distribution</u>

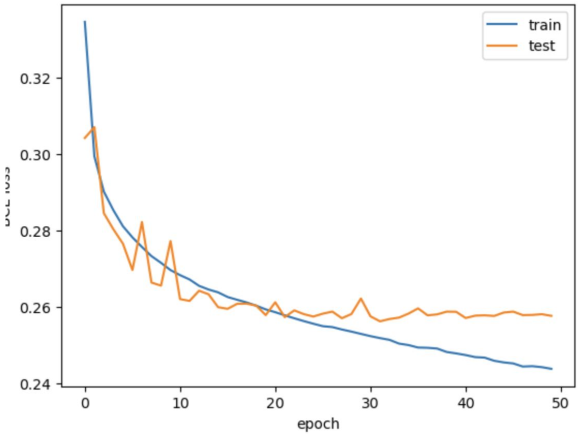


Preprocessing
— 10x10x20 pixels 3D “pseudo-images” of events were constructed
— Each pixel normalization as
(value - mean)/std, where mean and std were calculated using all dataset

Train dataset (0.75 of all data):
~400k background events
~400k cevns events

— 3 convolutional layers 3x3x5 with batch normalization after each other
— 3 fully connected layers
— Output layer with a single neuron with sigmoid activation function to show the probability of the events to be pointlike



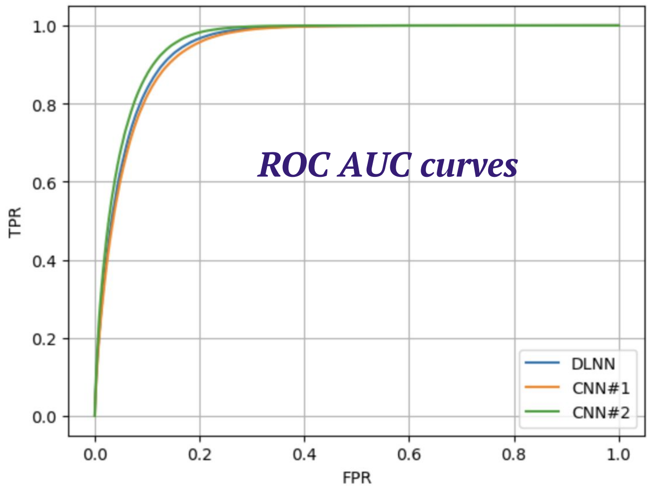


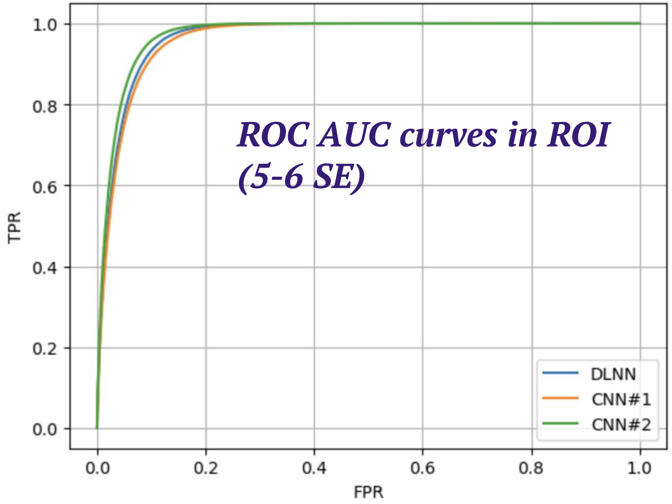
Comparison using test dataset

— general validation dataset (~600k events) was generated

DLNN : roc auc score = 0.947
CNN#1 : roc auc score = 0.943
CNN#2 : roc auc score = 0.956

in ROI (5-6 SE)
DLNN : roc auc score = 0.967
CNN#1 : roc auc score = 0.963
CNN#2 : roc auc score = 0.973

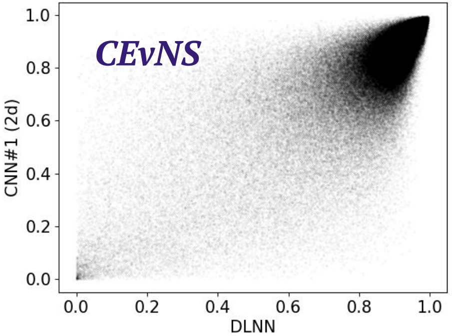


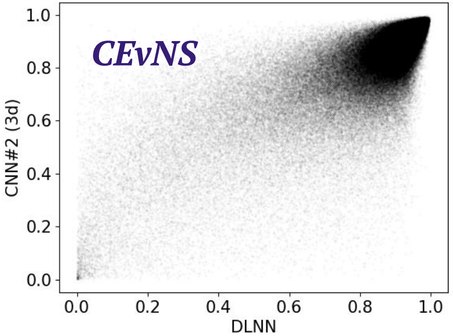


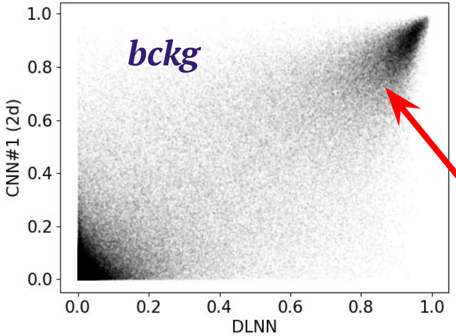
Comparison using test dataset

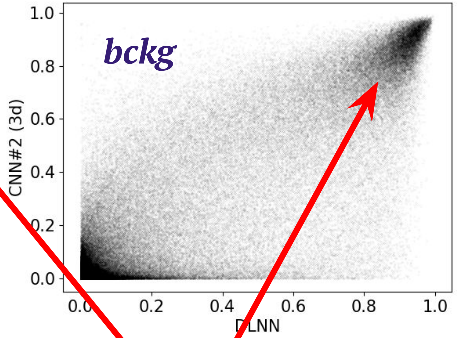
— there is a correlation between NN predictions on validation dataset

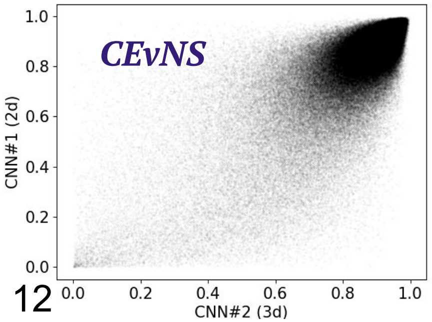
***2d distributions with NNs predictions** (probability of pointlikeness according to NNs)*



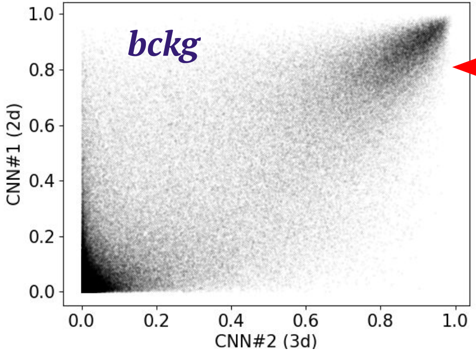








Pointlike events concentrate in one place

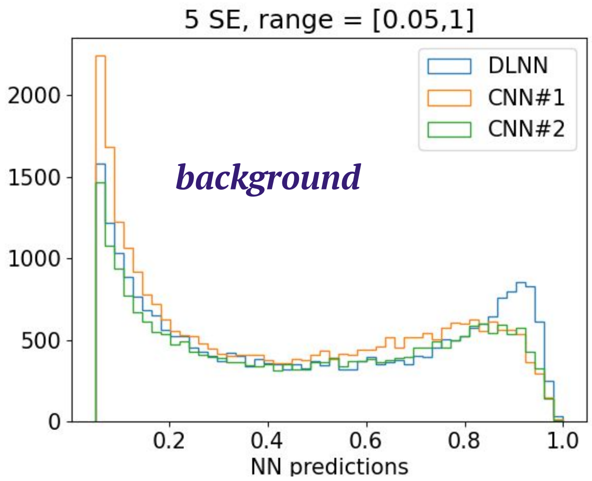


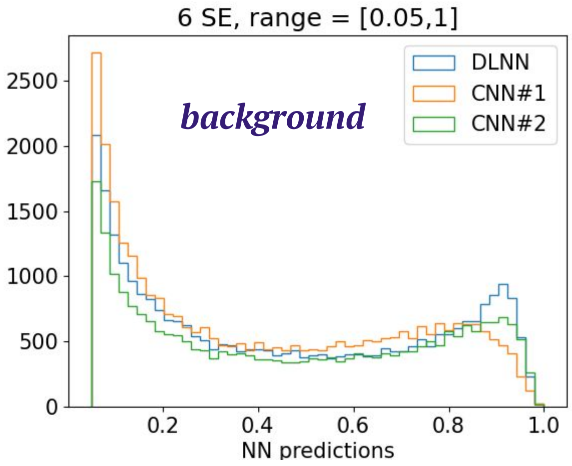
a lot of background events with high probability to be pointlike

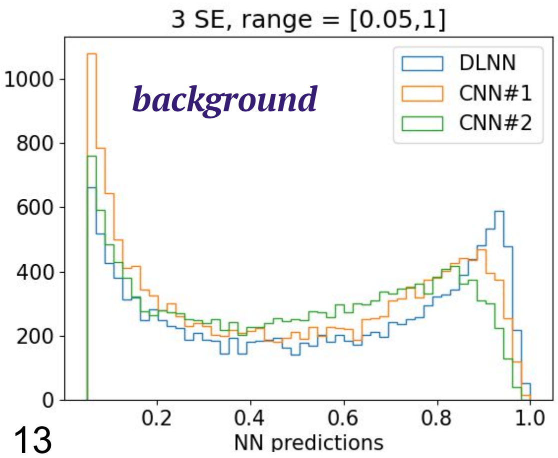
Comparison using test dataset

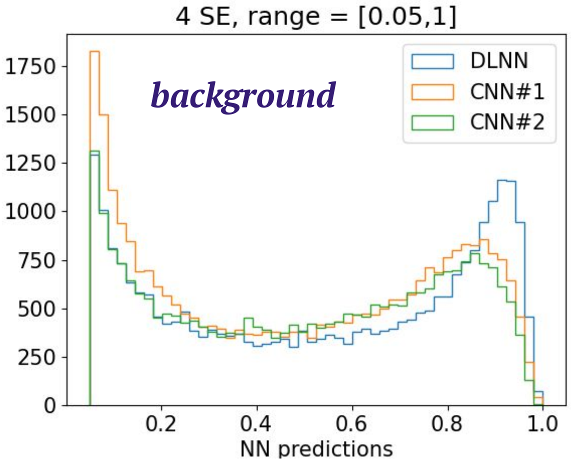
— CNNs are a bit better in background events recognizing
— but still have “pointlike peak”
— DLNN is better in pointlike events recognizing, especially in 3-4 SE region (next slide)

region of interest

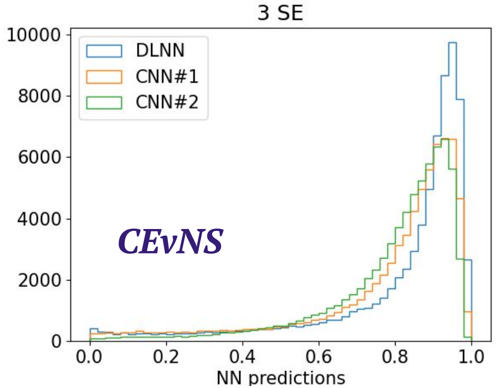


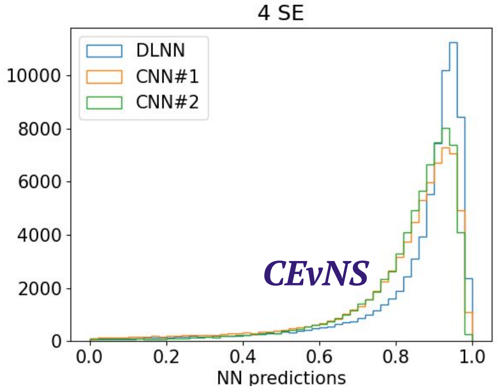


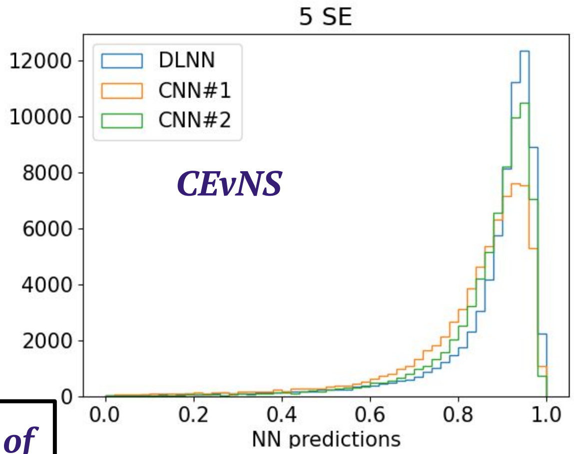




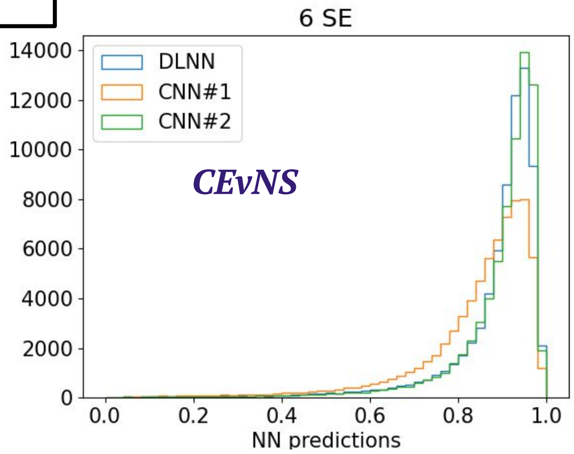
Comparison using test dataset







region of interest

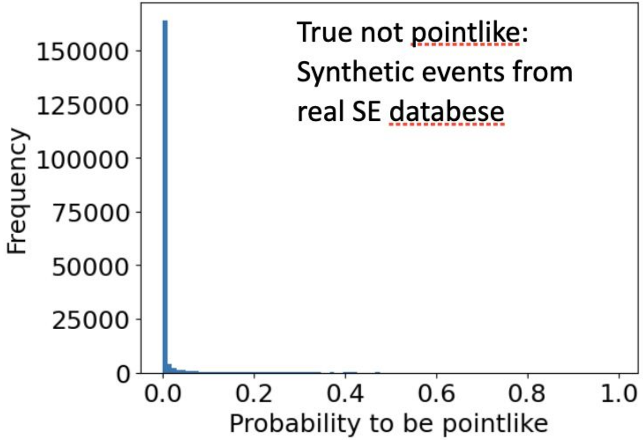


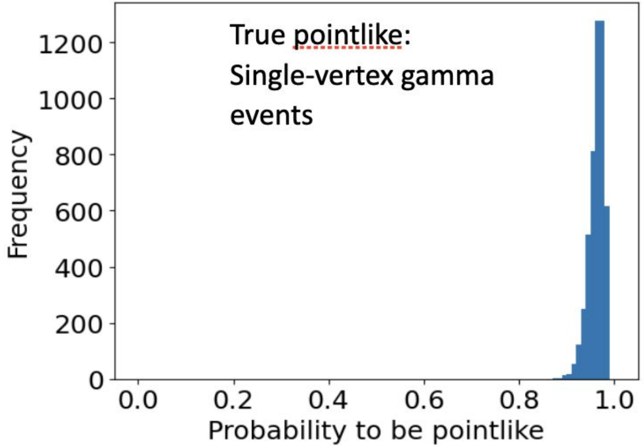
Results:

	3 SE	4 SE	5 SE	6 SE
signal reduction	DLNN: 19% CNN#1: 22% CNN#2: 21%	DLNN: 13% CNN#1: 17% CNN#2: 14%	DLNN: 9% CNN#1: 14% CNN#2: 9%	DLNN: 7% CNN#1: 12% CNN#2: 6%
bckg reduction	DLNN: 83% CNN#1: 83% CNN#2: 86%	DLNN: 83% CNN#1: 84% CNN#2: 86%	DLNN: 89% CNN#1: 91% CNN#2: 91%	DLNN: 92% CNN#1: 93% CNN#2: 93%

DLNN verification on real data

- Two types of real data were used to verify DLNN performance
- Randomly glued SE events from the real single SE database to form not pointlike events based on real data and distributions
- Gamma calibration dataset where it is easy to distinguish single-vertex events (as point-like dataset)
- Results:
  - More than 99% rejection of not poinlike events
  - 100% of pointlike gamma events survived





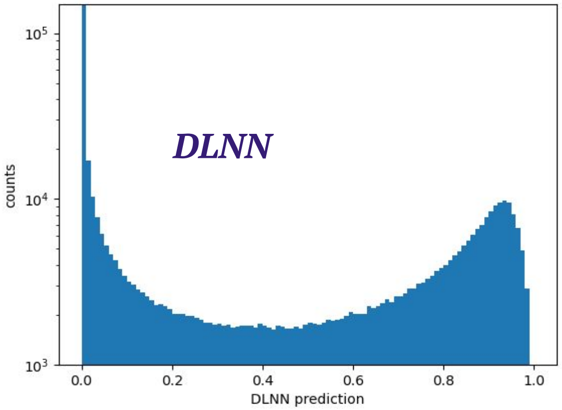
Testing on reactor OFF data

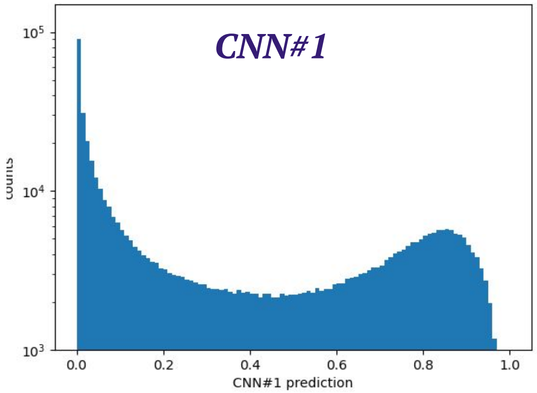
— significant part of real background is pointlike
— now we use optimized on sensitivity 2d cut based on DLNN and CNN#1:

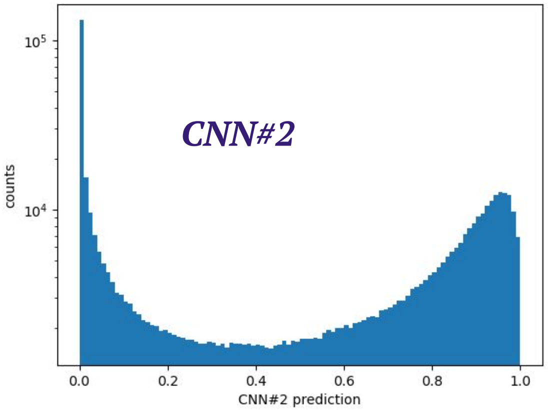
DLNN threshold: 0.6
CNN#1 threshold: 0.2

Background and signal reduction in ROI (r<130mm, duration <5000ns)

	~5SE	~6SE
signal (MC) reduction	11%	6%
bckg reduction	64%	54%





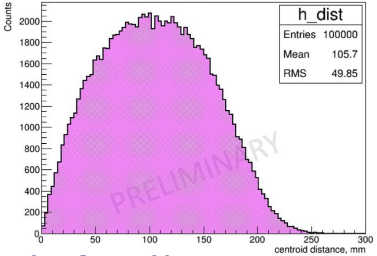


NNs predictions on real data. *A lot of background events with high probability to be pointlike.*

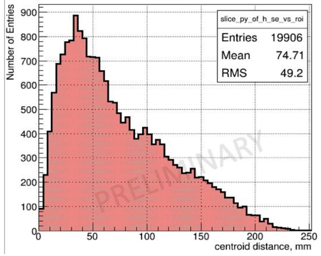
Real data problems

— “poinlike peak” is larger than in MC data (predicted by all models)
— duration of the events in the ROI is growing to the higher values
— It is possible if several SEs merged with each other

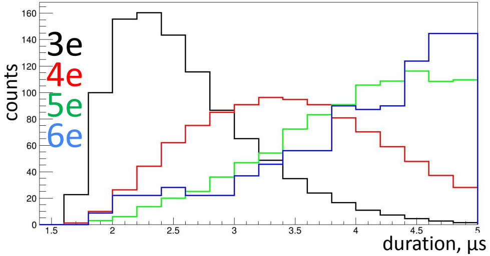
Distance between 2 random SEs from spatial distribution



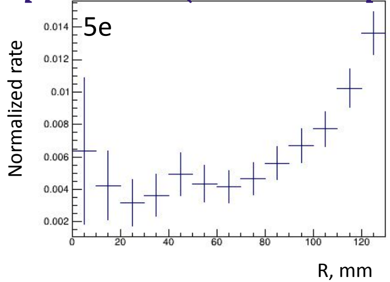
Distance between 2 consequent SEs from real data



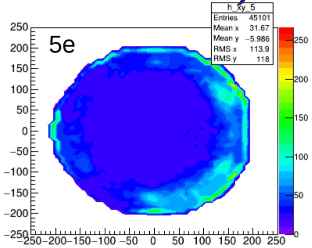
Background events durations

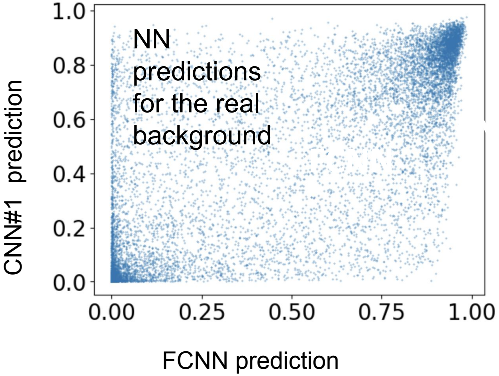


Bckg rate per radius (normalized per rings area)

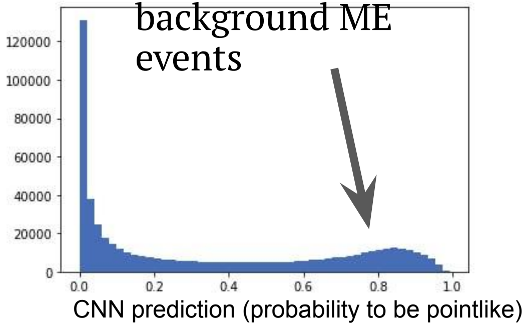


XY distribution for 5SE

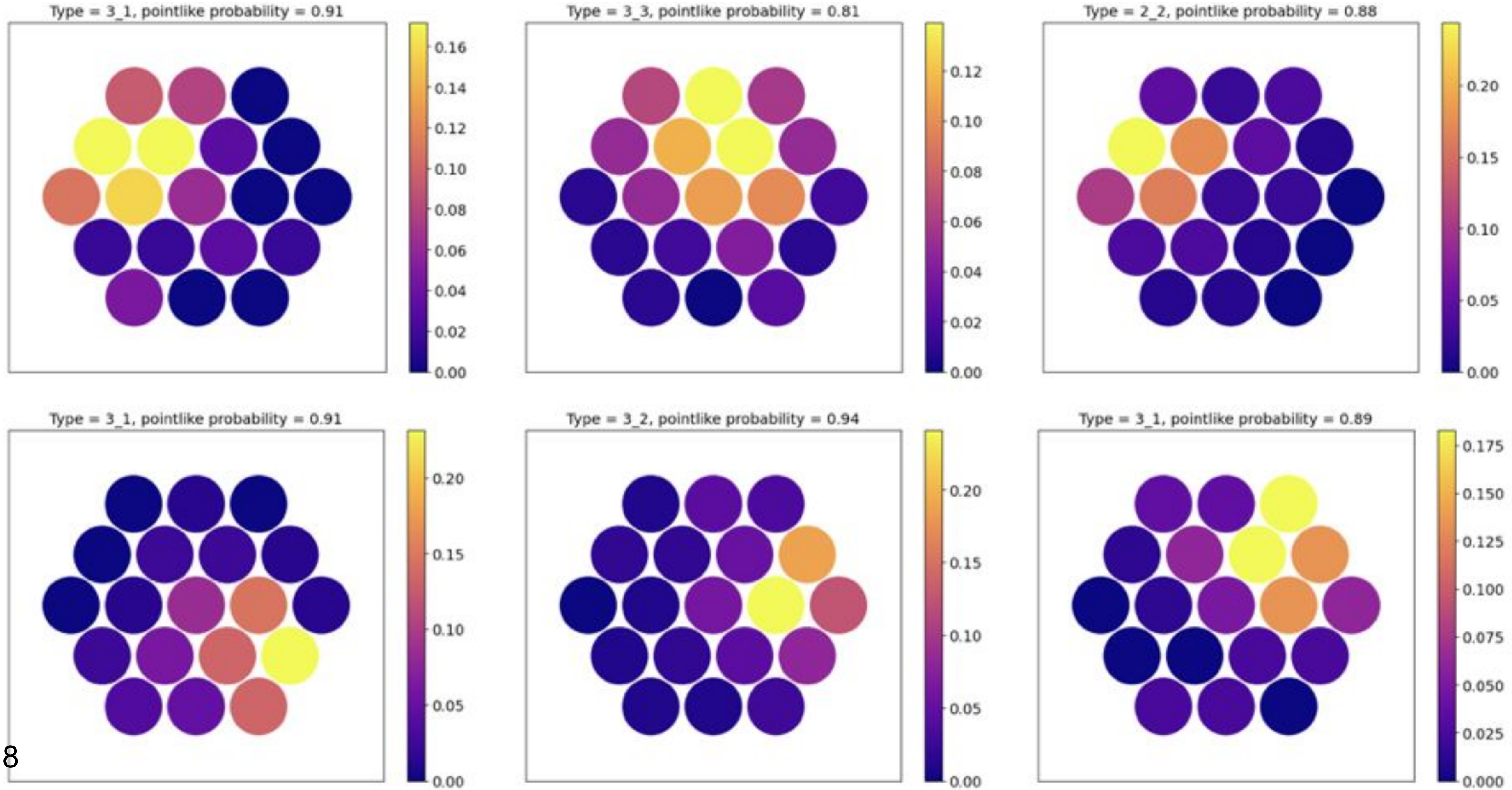




Highly correlated background ME events



Examples of MC background events with P>0.8



Summary

1. Light response functions were reconstructed using the iterative procedure with gamma-calibration data
2. Detailed simulation of 3-6 SE events in RED-100 was performed

3. Two NN approaches to pointlike event selection were tested
4. NNs show good results at MC events, but reality is more comlicated

DNN:

- + fast learn and optimization
- + less size of input data

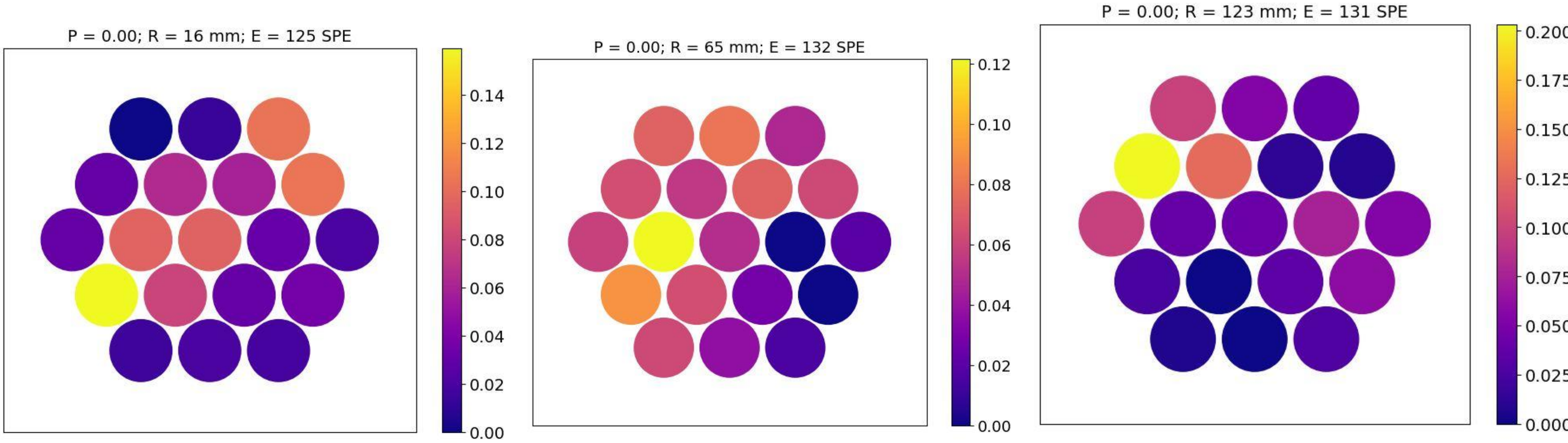
CNN:

- + use all available information about the event
- + maybe there are possibilities to improve

4. 2D optimized cut will be used in the further analysis

Thank you for your attention!

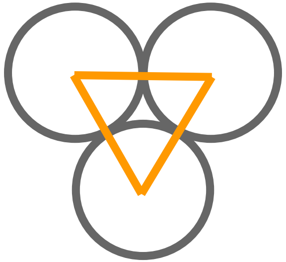
Backup


P = 0.00; R = 16 mm; E = 125 SPE
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
P = 0.00; R = 65 mm; E = 132 SPE
0.00
0.02
0.04
0.06
0.08
0.10
0.12
P = 0.00; R = 123 mm; E = 131 SPE
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200

Point-like event discrimination

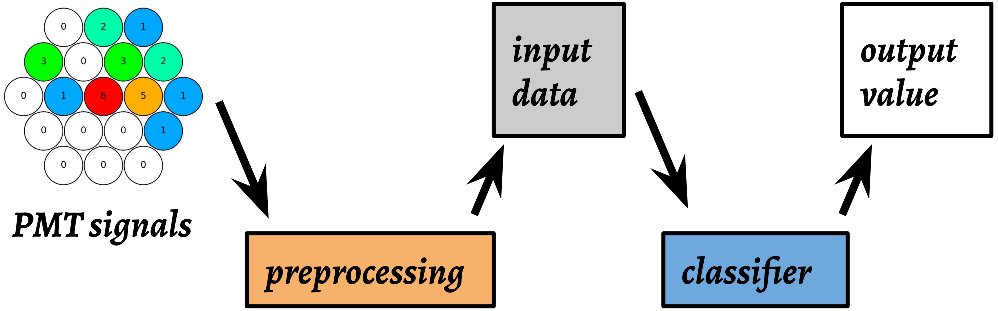
The MC and the data are not relevant now!

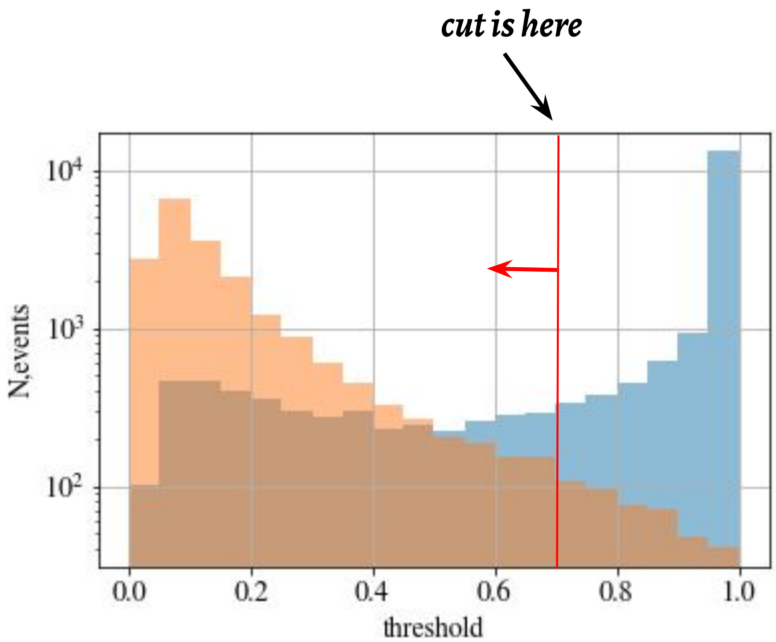
Using event classification based on total signals in PMTs. Tried several ML approaches (linear models, decision trees etc.), selected AdaBoost



Triangle cluster of three PMTs

Input signals are distribution of fraction of a signal in PMTs and three-PMT clusters, both sorted by signal size.



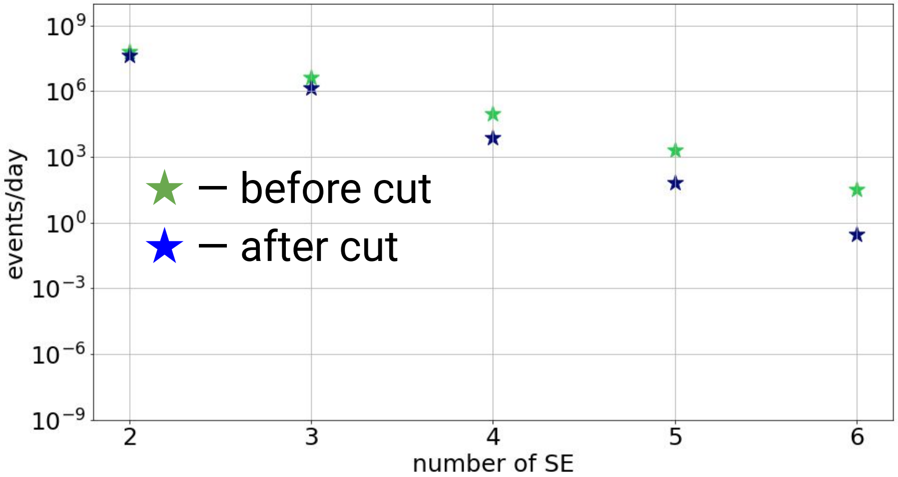


The output value of the classifier.
Orange spectrum corresponds to CEvNS events, while blue is background.

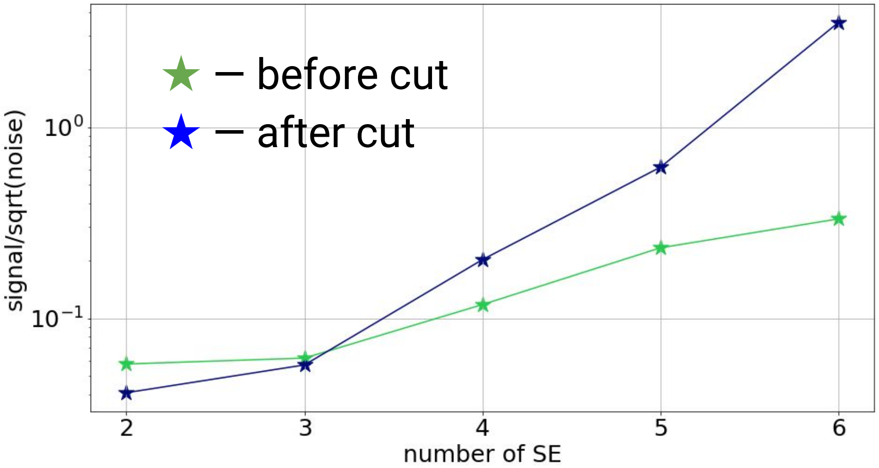
Discriminator results

on the simulated data

The MC and the data are not relevant now!



Background reduction (1 day)



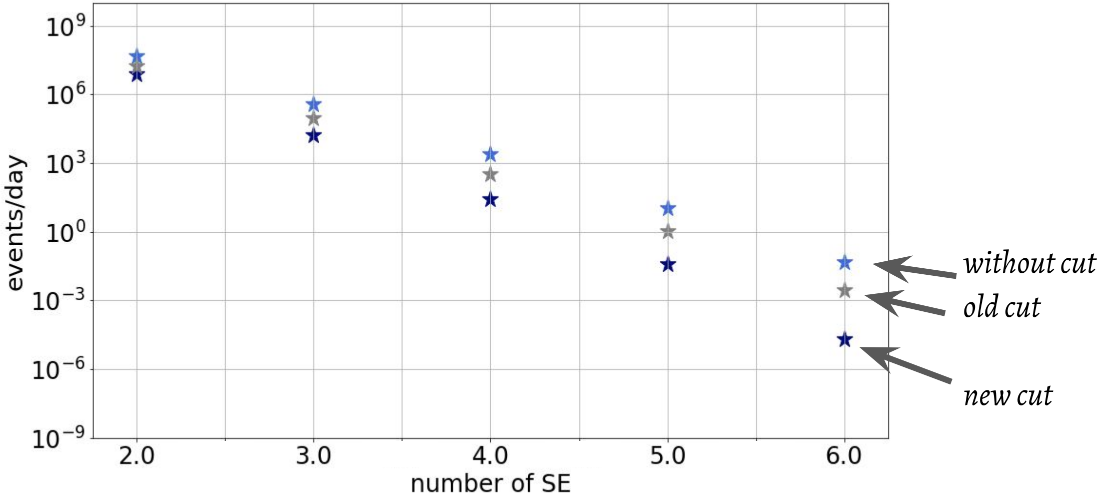
Signal/$\sqrt{\text{Background}}$, (1 day)

Discriminator results

The MC and the data are not relevant now!

comparison with old cut

(Dmitry Rudik, Status of the RED-100 experiment, ICPPA 2020)



Results of background reduction (only 1, 1+1, 1+1+1…), events from test run.
Classifier trained on simulation, tested on real data.

Number of SE	No cut	Old cut	ML cut
2	465	283	290.3
3	129	78	79.4
4	35.5	21.7	22
5	10.6	6.4	6.9
6	1.9	1.2	1.6

Signal reduction

First ground-level laboratory test of the two-phase xenon emission detector RED-100, Akimov D. et.al., JINST 2020