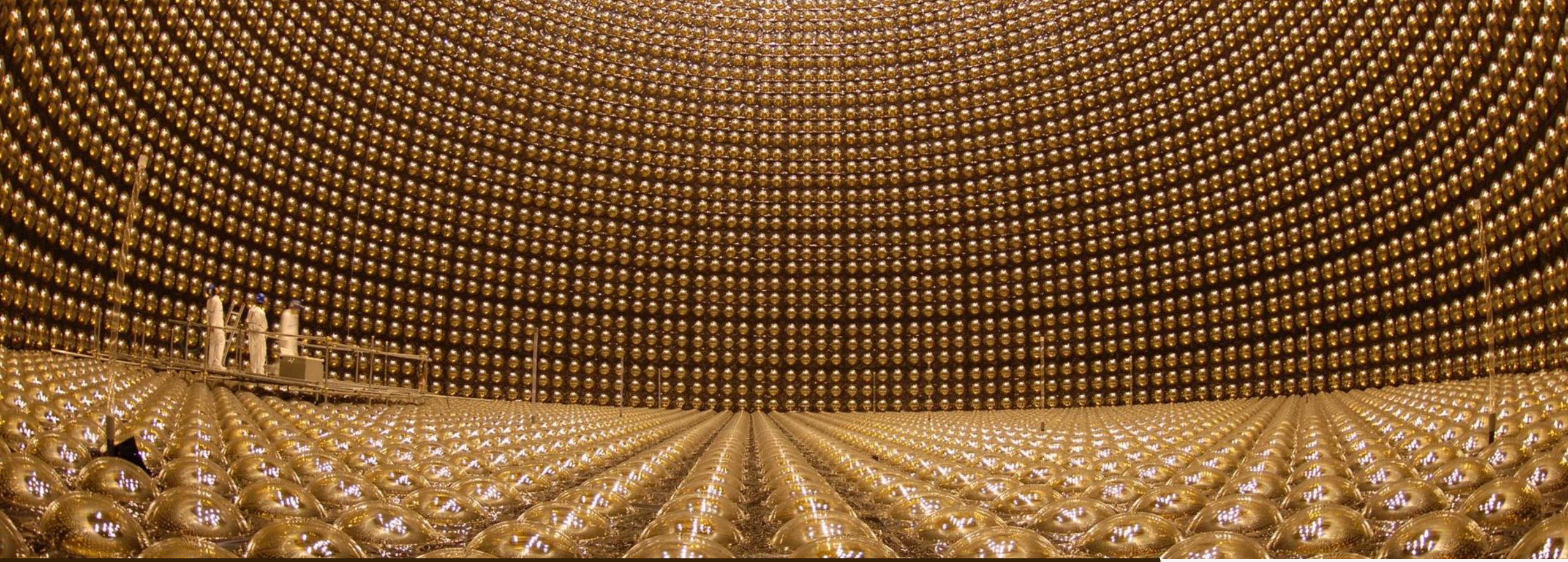
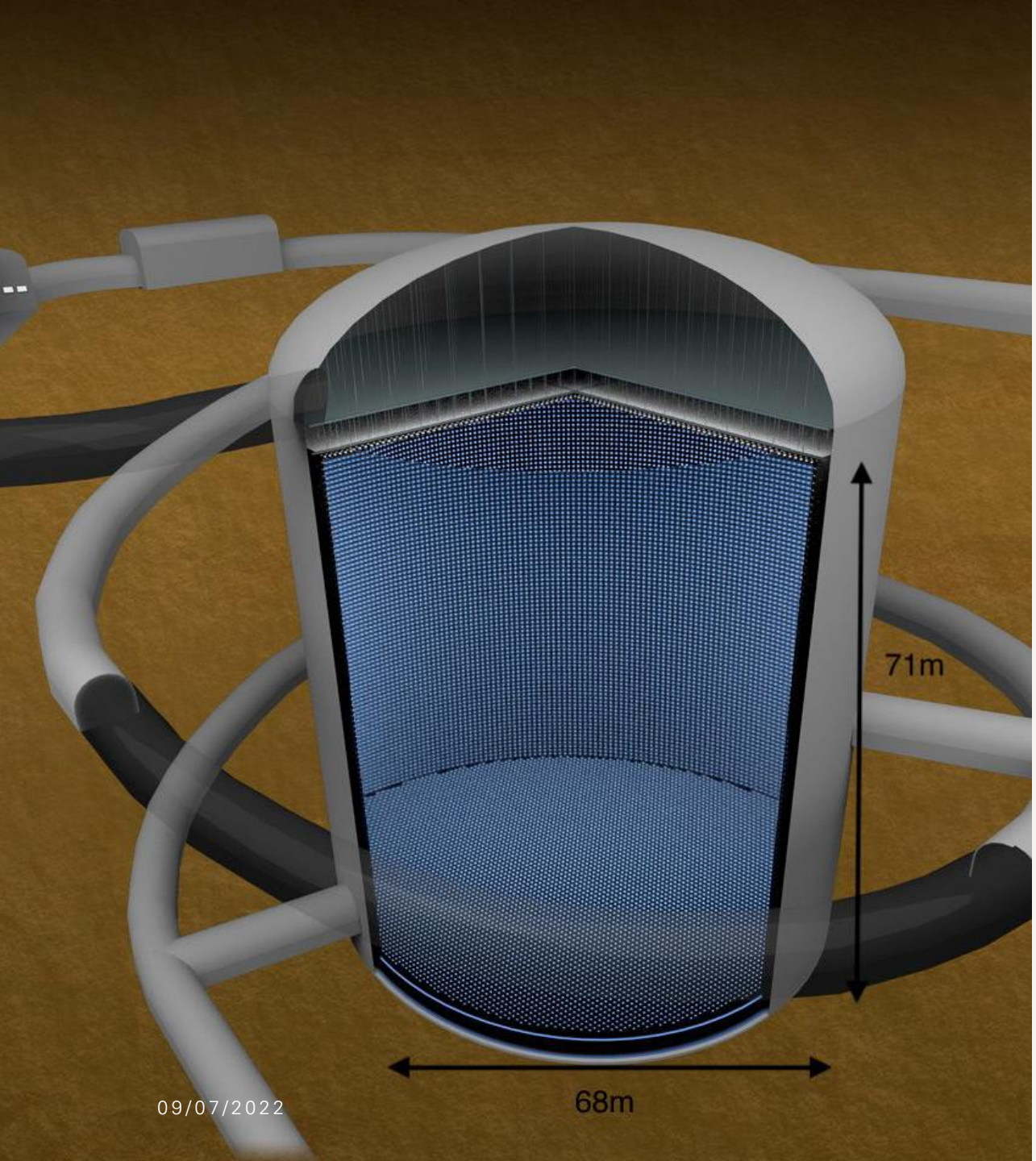


Dark rate reduction with machine learning techniques for the Hyper-Kamiokande experiment

Aurora Langella, Bernardino Spisso, Lucas N. Machado
on behalf of the Hyper-Kamiokande Collaboration

ICHEP 2022

09/07/2022

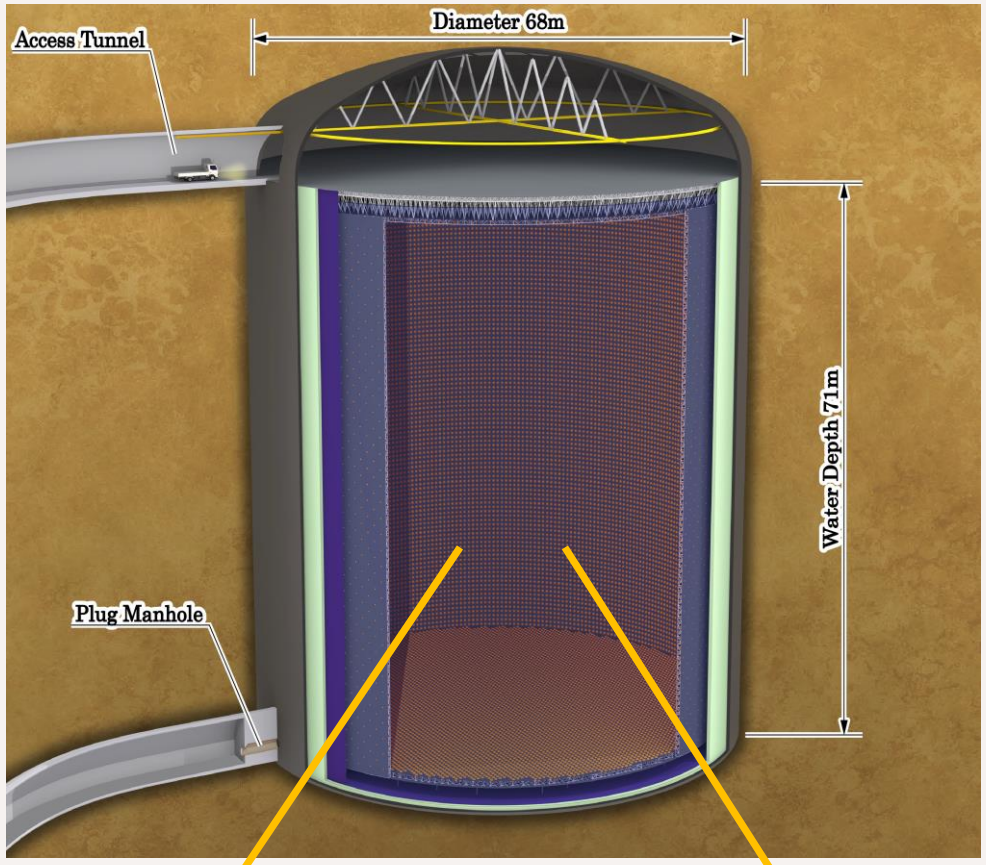


Hyper-Kamiokande

Hyper-Kamiokande (Hyper-K) is to be the next generation of large-scale water Cherenkov detectors that will be built in Japan.
It aims to obtain exciting results in many fields such as:

- CP violation in the lepton sector;
- Neutrino oscillations (solar, atmospheric and accelerator neutrinos);
- Astrophysical neutrinos;
- Proton decay;

Hyper-K design



Hyper-K will adopt the successful strategies used to study neutrino oscillations in Super-Kamiokande, K2K and T2K. Main improvements will be:

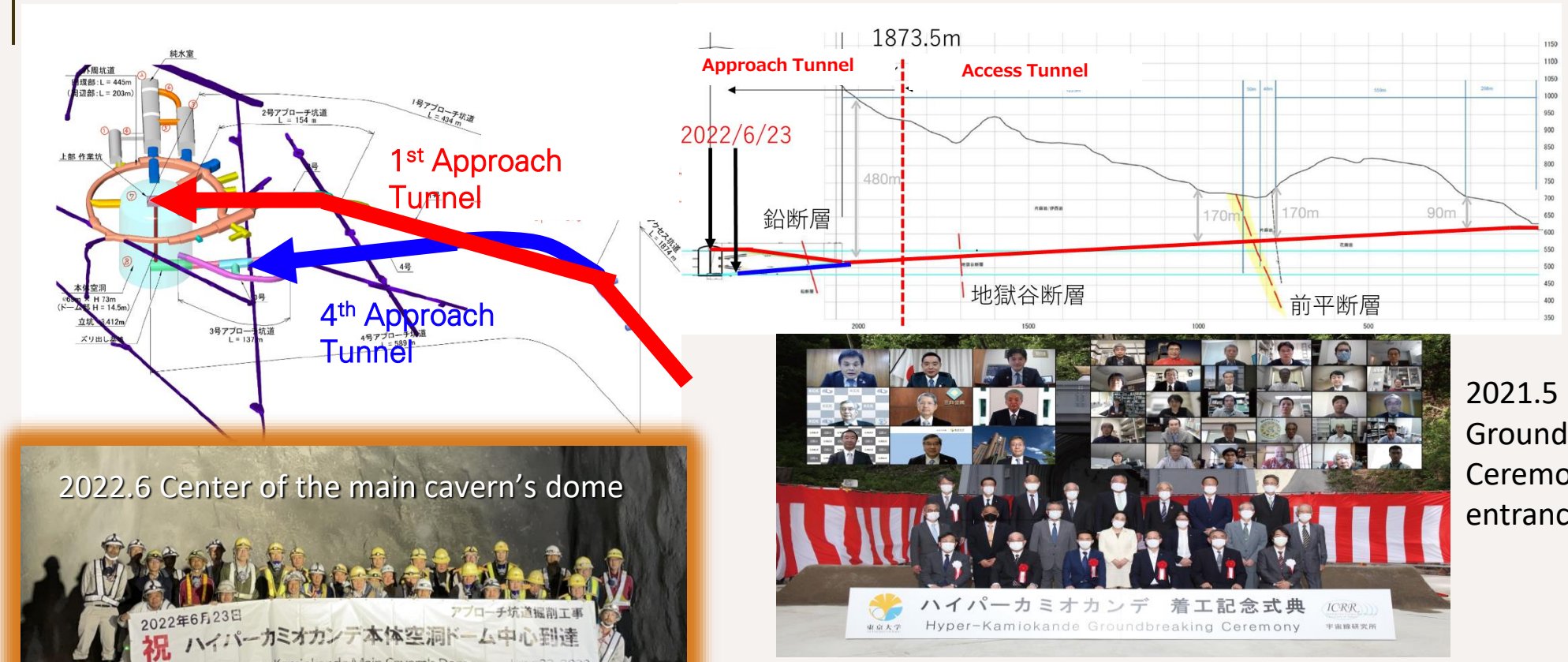
- Larger detector for increased statistics;
- Improved photo-sensors for better efficiency;
- Higher intensity beam and updated/new near detectors for accelerator neutrino part.

Hyper-K Far Detector

The Hyper-K Far Detector (HK-FD) will be charachterized by:

Cylindrical tank: Φ = 68 m and H = 71 m
Fiducial volume: ~0.19 Mtons; × 8 SK → HK-FD
Baseline design: ~20% photo-coverage with 20'000 20'' B&L PMTs combined with few thousands of multi-PMT modules.

Construction of the detector



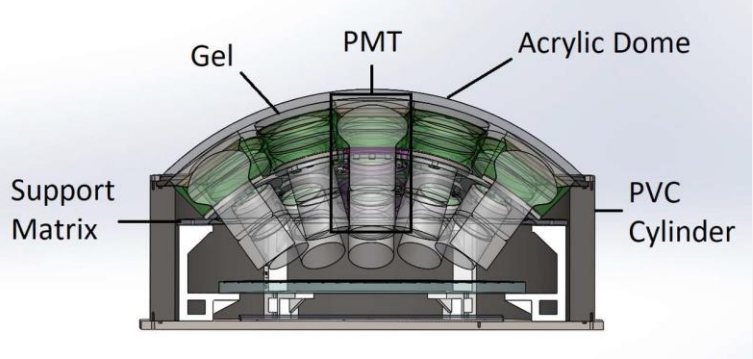
2021.5 Groundbreaking Ceremony @ entrance

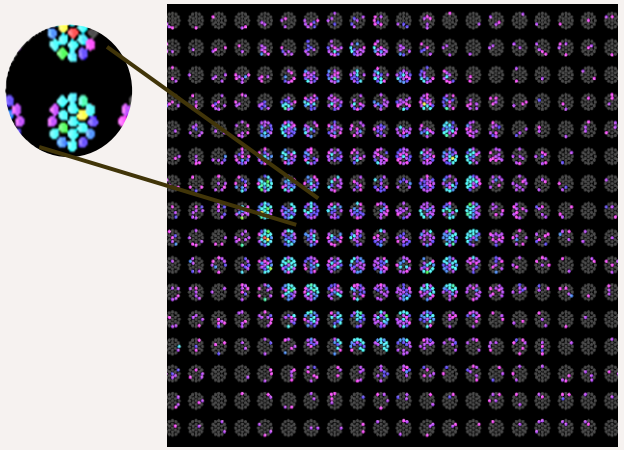
- Cavern excavation is about to begin on schedule
- Hyper-K will start its operation in 2027

The multi-PMT option

The multi-PMT (mPMT) was first designed for the KM3NeT experiment.
A new design, optimized for Hyper-K requirements, has been realized. It consists of 19 3" PMTs inside a vessel, each one with a different orientation.
The mPMT has several advantages as a photosensor module:

- Increased granularity;
- Superior photon counting;
- Improved angular acceptance;
- Extension of dynamic range;
- Intrinsic directional sensitivity;
- Local coincidences.





Dark rate reduction: motivation

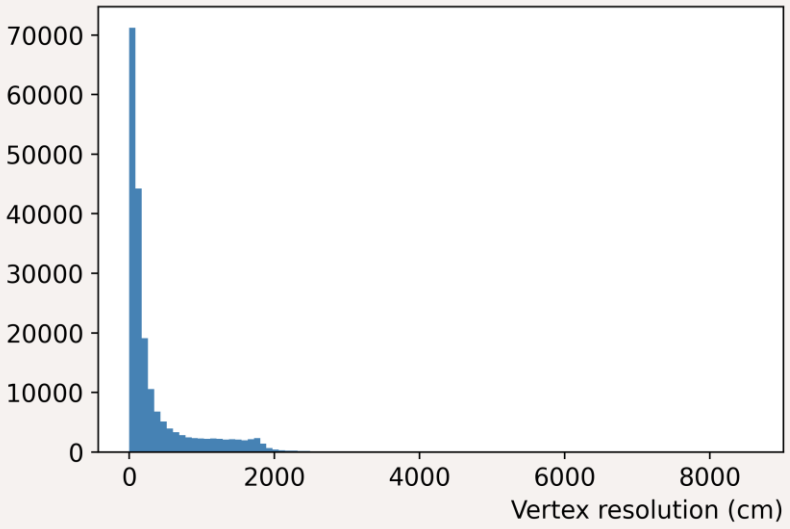
One of the strengths of the mPMT module lies in the possibility of improving the vertex resolution. This can lead to an enlargement of the fiducial volume as well as a better energy resolution and a possible lower threshold, thus improving the low energy neutrino detection. Simulation studies on the performance of the mPMT showed that these improvements is mainly due to the lower dark rate of the 3" PMTs.

→ Efforts are being made to reduce dark rates for the hybrid configuration of the Hyper-Kamiokande Far Detector using multivariate statistical methods.
Our goal is to train a machine learning method to reject heavily mis-reconstructed events that are probably due to dark rate.
We choose to adopt the Boosted Decision Tree (BDT) method trained with **scikit-learn** machine learning libraries.

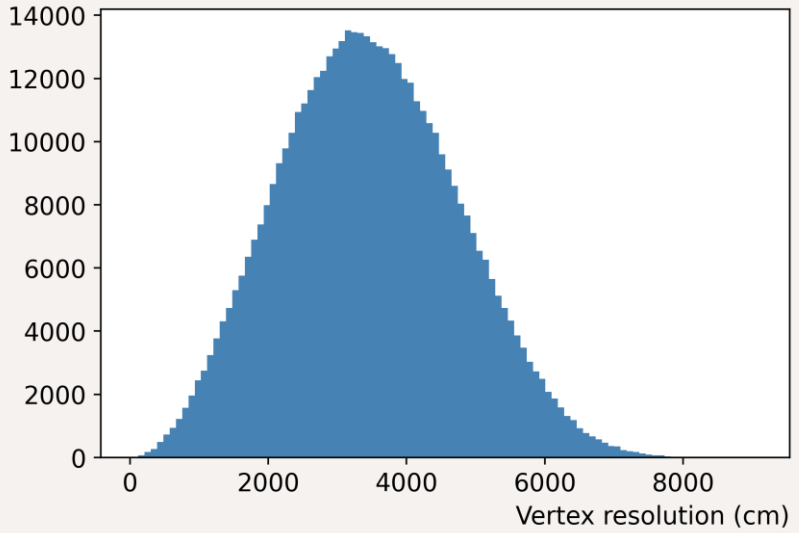
Effects of dark rate on vertex resolution

The consequence of dark rate is clearly visible when studying the vertex resolution distribution for a given particle energy.
Considering a sample of electrons with energy = 3.5 MeV, 3'' PMT dark rate = 600 Hz and 20'' PMT dark rate = 4.2 kHz we get:

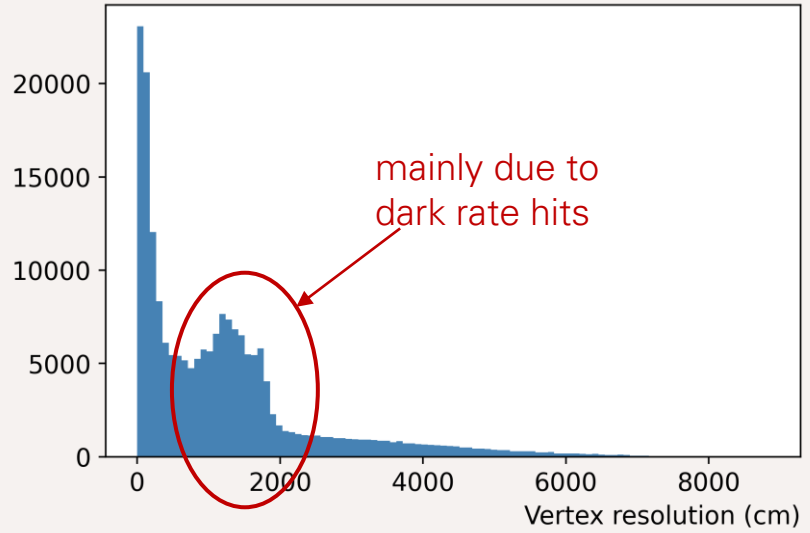
No dark rate



Only dark rate



Signal + dark rate



Simulation Configuration

Detector Setup

Hyper-K hybrid configuration with:

- 20k B&L PMTs with dark rate of 4.2 kHz;
- 2k mPMTs with 3'' PMT dark rate of 600 Hz;

Training dataset

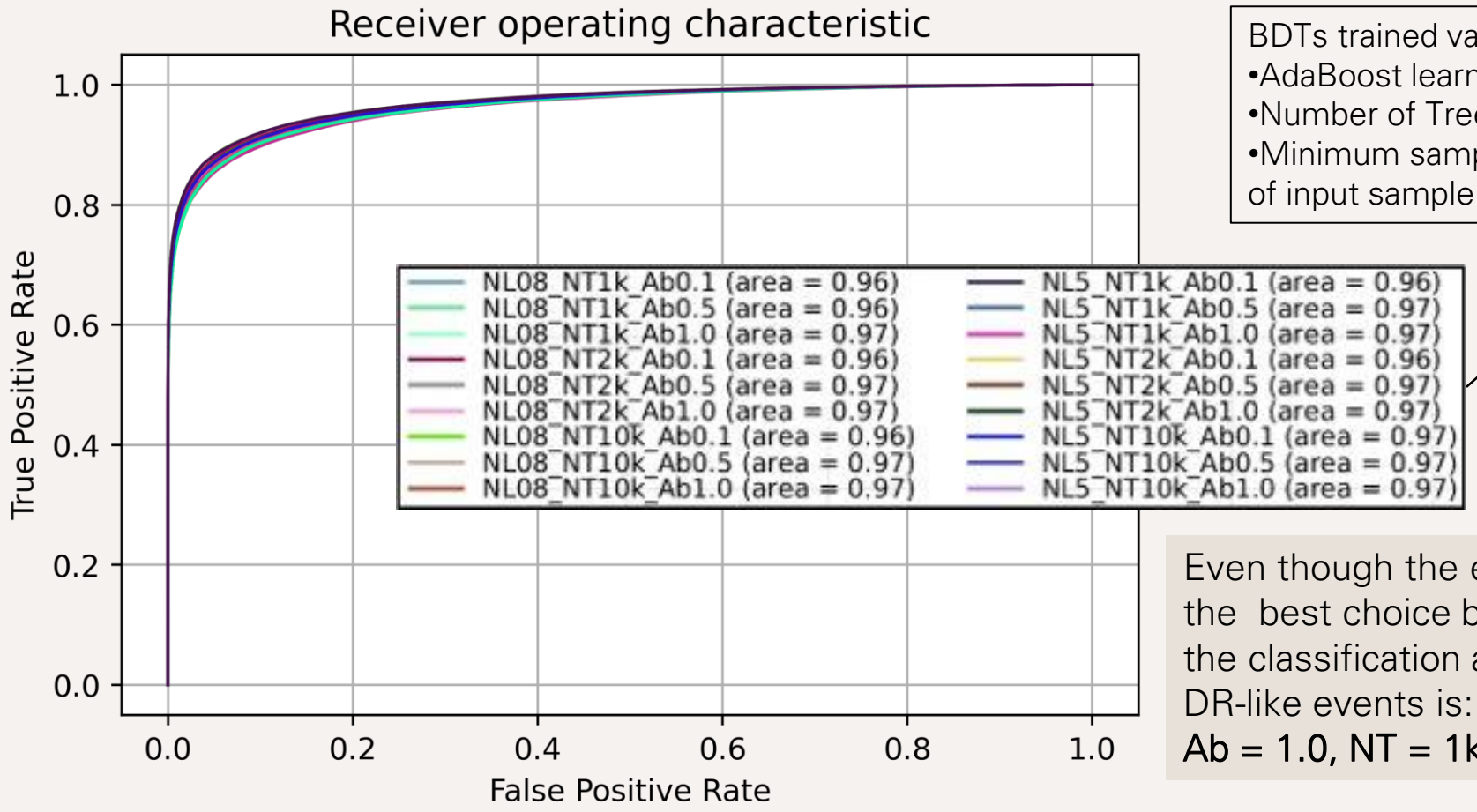
→ Signal: 3 MeV electrons with gaussian profile (0.5 MeV sigma) + dark rate events;

→ Background: only dark rate events;

Validation dataset

→ Electrons with energies in the [1-18] MeV range + dark rate events.

BDT performance evaluation



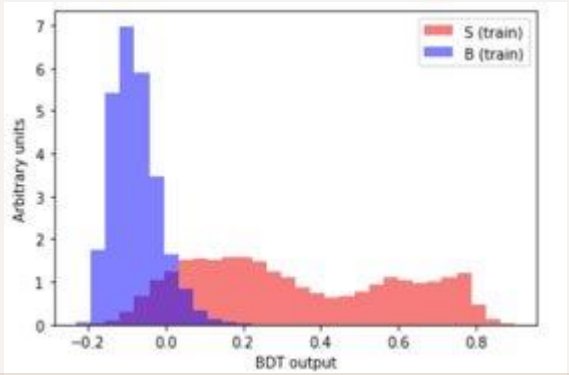
BDTs trained varying the hyperparameters:
- AdaBoost learning rate (Ab) = 0.1, 0.5, 1.0
- Number of Trees (NT) = 1k, 2k , 10k.
- Minimum samples in leaf nodes (NL) = 0.08% or 5% of input sample.

Even though the efficiencies are very similar, the best choice based on computation time for the classification and reduction efficiency for the DR-like events is:
Ab = 1.0, NT = 1k and NL = 5%.

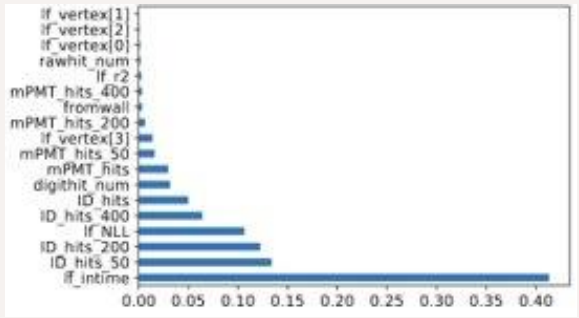
BDT performance evaluation

For the chosen BDT (Ab = 1.0, NT = 1k and NL = 5%):

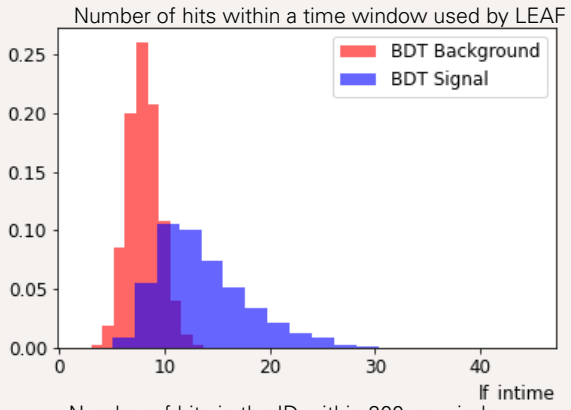
Signal-Background separation

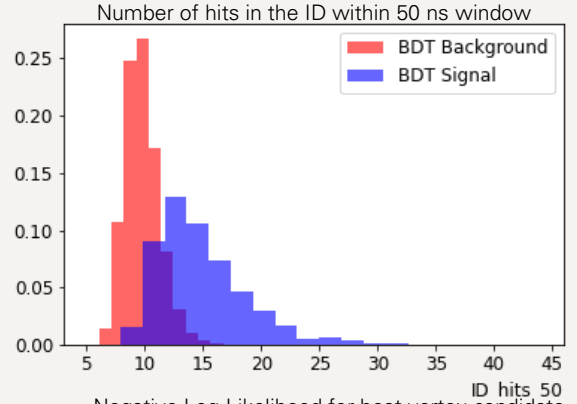


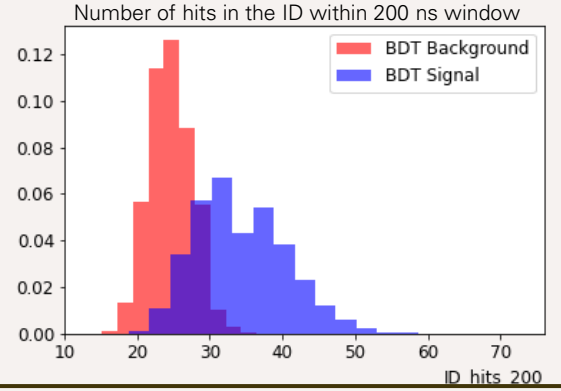
Variable importance for best BDT

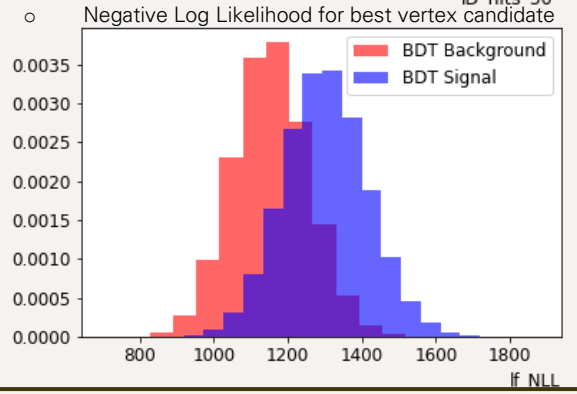


Most important variables



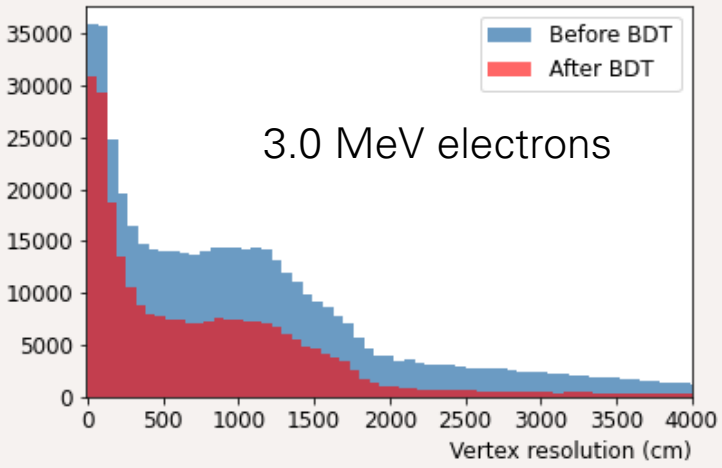


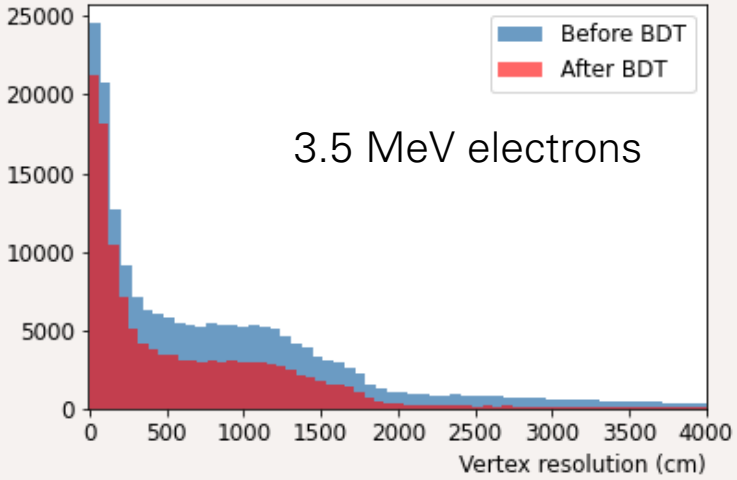


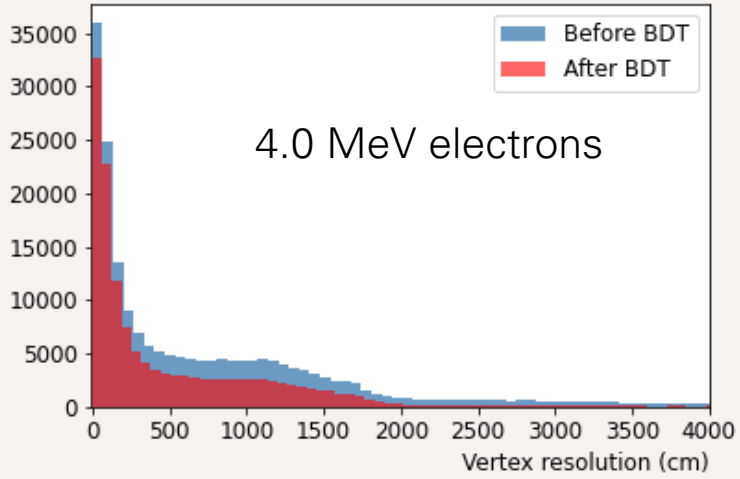


Improvements in the vertex resolution

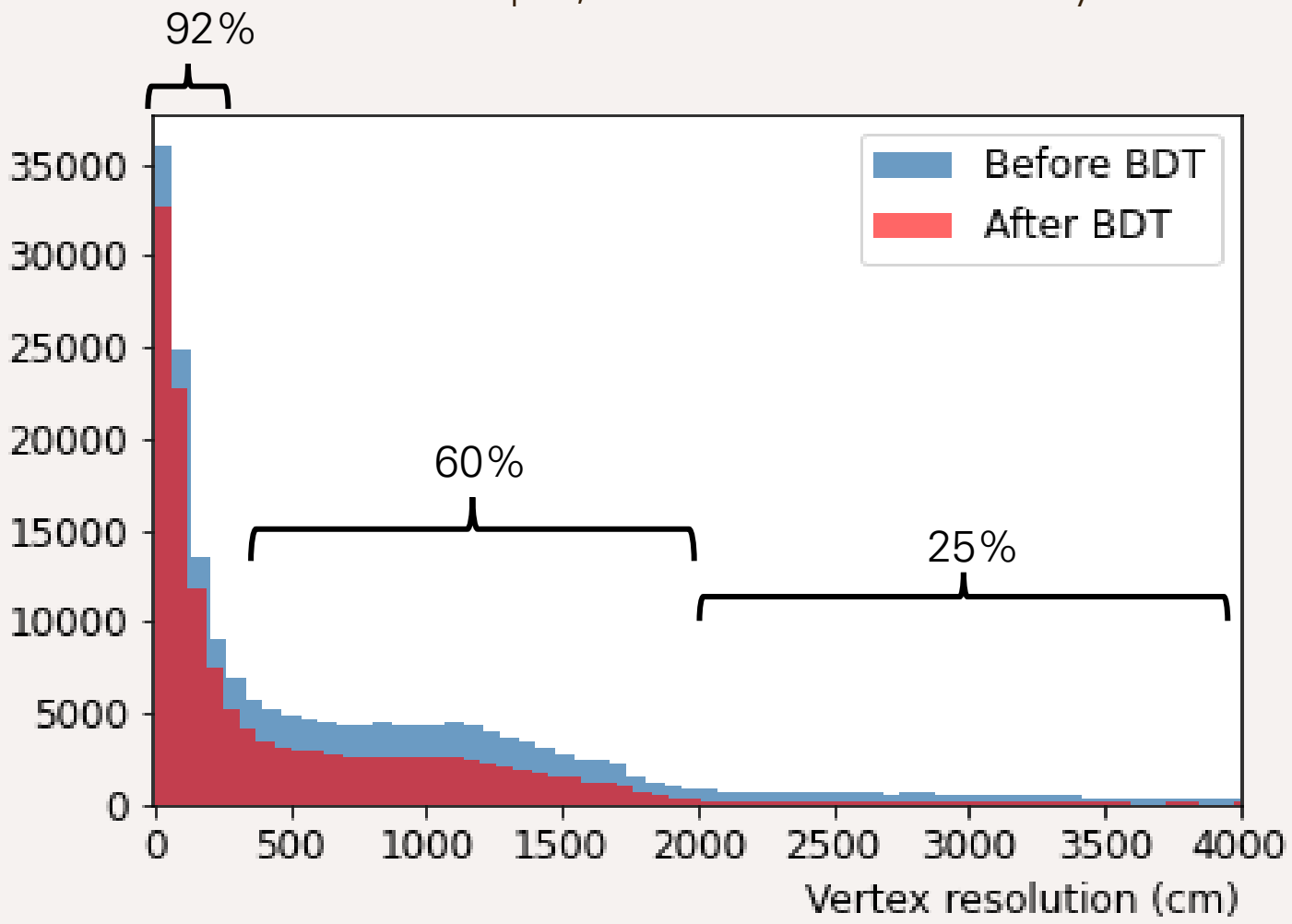
To evaluate the efficency of the BDT in reduce the dark rate, we compared the vertex resolution distribution with and without applying a BDT cut for electrons of different energies.







For the 4 MeV sample, the BDT cut efficiency:



We identified three blocks to evaluate the BDT efficiency:

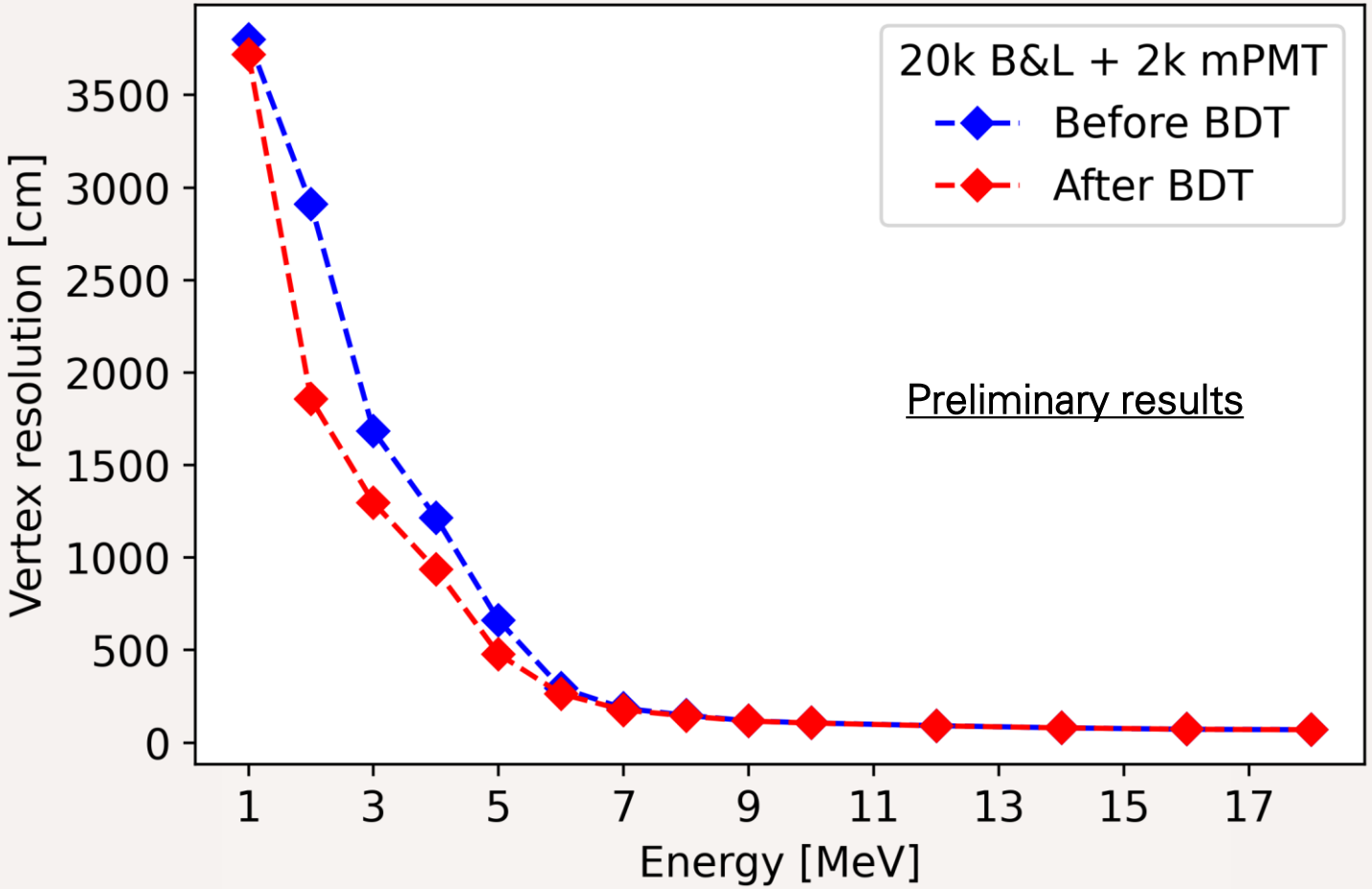
1. Vertex resolution < 250 cm
This region contains almost only signal events.

2. 250 < Vertex resolution < 2000 cm
This region is a mixture of signal events and dark rate hits (mostly dark rate).

3. Vertex resolution > 2000 cm
This region contains only dark rate events.

Improvements in the vertex resolution

By defining the vertex resolution as the point including 68% on distribution of distance between MC and reconstructed vertex, we evaluated it for electron with different energies.



A clear improvement in the low energy region is visibile for a BDT trained with 3 MeV electrons

Summary & new prospects

- Hyper-Kamiokande will start acquire data in 2027;
- The Hyper-K Far Detector will consist of an hybrid configuration with 20'' B&L PMTs and mPMTs;
- Efforts are being made to reduce the overall dark rate of Hyper-K FD through machine learning techniques;
- BDTs have proven to be efficient in reject mis-reconstructed events due to dark rate hits thus leading to a better vertex resolution in the MeV energy range;
- Other machine learing methods will be tested;
- Mixed energy dataset will be used for training to avoid a possibile bias;
- To improve the sensitivity to low energy neutrinos, new studies will be done to test the efficiency of BDT in reducing background events coming from radon decay.

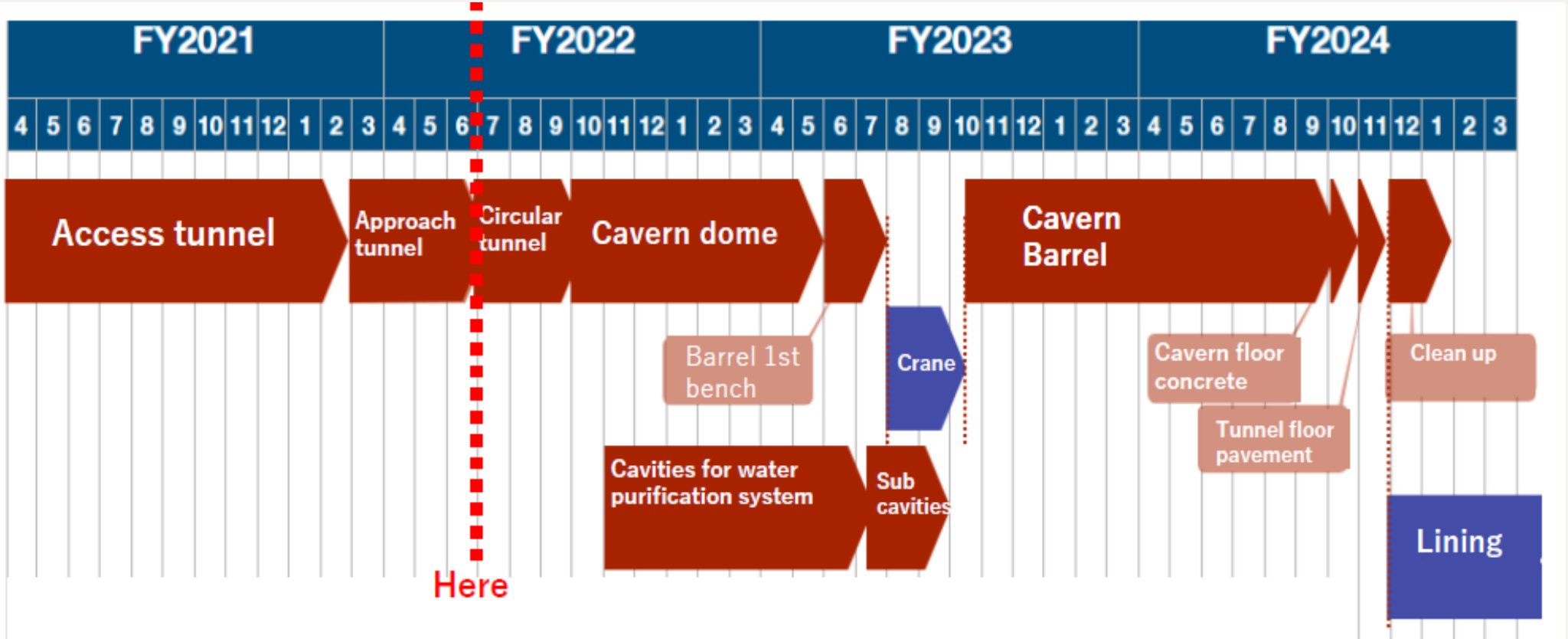
Thank you

8/05/20XX

SPARES

8/05/20XX

Status and schedule of construction



Hyper-K software

Simulations are done with the following tools:

- **WCSim:** a very flexible Geant4 based program, adopted by the T2K and Hyper-K Collaborations for developing and simulating large water Cherenkov detectors that relies on ROOT and Geant4.

- **LEAF (Low Energy Analysis Frame):** a reconstruction tool developed by the Hyper-Kamiokande Collaboration that reconstructs vertex position, direction and hit time of low energy events (few MeV - few tens MeV).

LEAF variables

- **lf vertex[0]:** Reconstructed vertex, X.
- **lf vertex[1]:** Reconstructed vertex, Y.
- **lf vertex[2]:** Reconstructed vertex, Z.
- **lf vertex[3]:** Reconstructed vertex, T.
- **mPMT hits:** Number of hits in the mPMTs.
- **mPMT hits 50:** Number of hits in the mPMTs within 50 ns window.
- **mPMT hits 200:** Number of hits in the mPMTs within 200 ns window.
- **mPMT hits 400:** Number of hits in the mPMTs within 400 ns window.
- **ID hits:** Number of hits in the ID.
- **ID hits 50:** Number of hits in the ID within 50 ns window.
- **ID hits 200:** Number of hits in the ID within 200 ns window.
- **ID hits 400:** Number of hits in the ID within 400 ns window.
- **lf NLL:** Negative Log Likelihood for best vertex candidate.
- **lf ctime:** CPU time for event reconstruction.
- **lf intime:** Number of hits within a time window used by LEAF.
- **digithit num:** Number of digitized hits.
- **rawhit num:** Collection of hits, including dark noise.

Vertex resolution

For each event the vertex resolution is defined as:

$$Vertex\ resolution = \sqrt{(trueX - recX)^2 + (trueY - recY)^2 + (trueZ - recZ)^2}$$

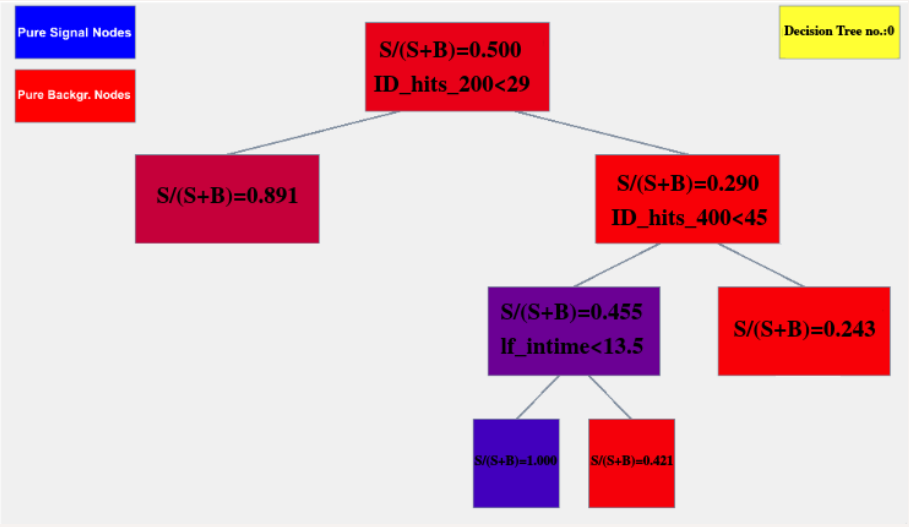
Boosted Decision Tree

Boosted Decision Trees (BDTs) are machine learning techniques consisting of tree structured classifiers.

A Decision Tree (DT) is a pattern recognition technique that can bu used for binary classification. The dataset goes through a series of sequential nodes and each event is classified by each node as signal or noise. The split ends according to a certain criterion for the stop. The dataset will then result divided into a number of leaves each belonging to one of the two classes.
At each node weights are chosen in such a way that the difference between the Gini index of the original leaf and the split leaves (weighted by the proportion of events) is maximised.

$$\text{Gini index} = p(1-p)$$

where p is p the signal purity of a certain node.



The BDT is a boost of decision trees, each one trained on the same samples ensemble and the output of the BDT is the average output of all the trees.
The boosting algorithm used is AdaBoost.

AdaBoost

AdaBoost, which stands for "Adaptive Boosting" and was proposed by Freund and Schapire in 1996, was was the first highly successful boosting algorithm developed for classification binary.
For every new decision tree, data that were misclassified in the previous one are given an adjusted weight for the following tree. The weight of each event in the new decision tree is multiplied by the boost weight α boost, given by:

$$\alpha_{boost} = (\frac{1-\varepsilon}{\varepsilon})^{\beta_{boost}}$$

Where ε is the classification error of the tree and β is the learning rate of the classifier. Thus, events that are not classified well will have greater weight in the next tree. Once all DTs have been trained, the class of an event is defined as:

$$y_{boost}(x) = \frac{1}{N}\sum_{k=0}^{N} \ln(\alpha_k)\, h_k(x)$$

Where α_k is the weight and h_k the class assigned by the k-th tree to event x.