Reactor Neutrino Energy Reconstruction with Machine Learning Techniques for the JUNO Experiment

Arsenii Gavrikov[1,2] (1st year PhD at UNIPD)

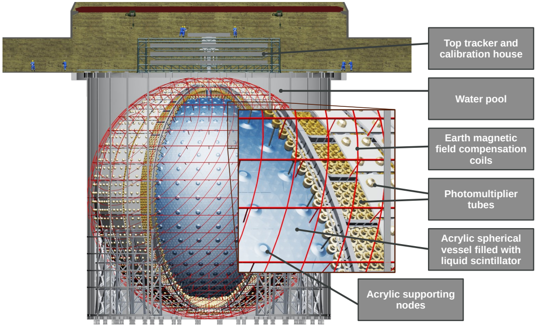
[1]Dipartimento di Fisica e Astronomia dell'Università di Padova, Padova, Italy

[2]INFN Sezione di Padova, Padova, Italy

MAYORANA School 2023, Modica, Italy

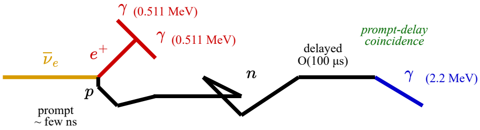
Introduction to the JUNO experiment

1. Jiangmen Underground Neutrino Observatory:
   - **multipurpose** experiment
   - under construction
   - 53 km away from **8 reactor cores** in China
   - data taking expected in ∼2024
   - JUNO Collaboration:
     - 76 institutions
     - 716 collaborators



2. The main goals of JUNO:
   - neutrino mass ordering **(3σ in 6 years)**
   - precise measure of oscillation parameters $\sin^2\theta_{12}, \Delta m^2_{21}, \Delta m^2_{31}$

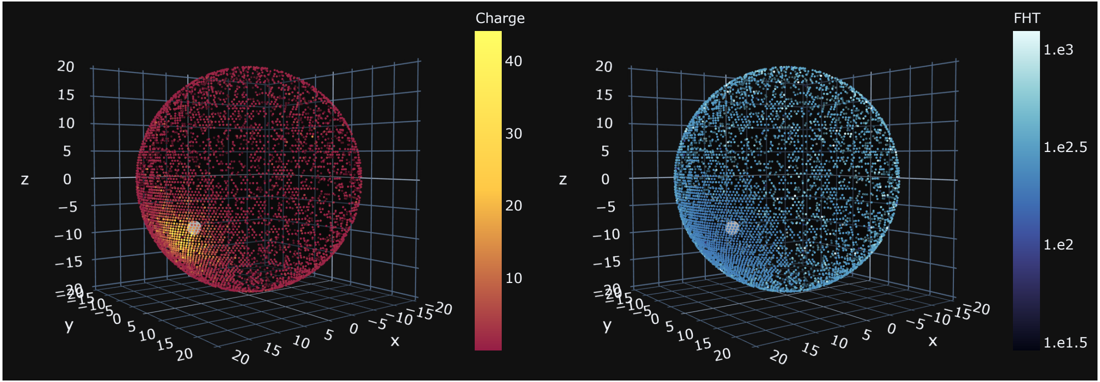
3. The Central Detector:
   - detection channel: $\overline{\nu}_e + p \to e^+ + n$;
   - deposited energy converts to optical light
   - **the largest** liquid scintillator detector: 20 kt
   - 77.9% photo-coverage: 18k 20”, 26k 3” photo-multiplier tubes (PMTs)



Machine Learning (ML) in particle physics

- ML methods are used at all levels of data processing in many particle physics experiments:
  - signal/background discrimination
  - event selection in the trigger
  - event simulation
  - anomaly detection
  - identification, etc.
- Why is ML useful for particle physics?
  - **Faster**. More precisely, with proper training
  - **Adequate** for many purposes simultaneously: event simulation, analysis, reconstruction, identification, etc.
  - **GPU friendly** by construction, which is important for big data processing
- Machine-learning algorithms use statistics to find patterns in massive amounts of data
- Our task is a supervised learning problem (regression)

Problem statement



An example of a positron event with deposited energy $\sim$**6 MeV**. The grey sphere — the primary vertex.

Charge at PMT

First Hit Time (FHT) at PMT

PMT position

<u>**We want to reconstruct**</u>:

Deposited energy E_{dep} with resolution $< 3\%$ @ 1 MeV

Datasets

- Two datasets: for training and for testing
- generated by the Monte Carlo method
- detector + electronics simulation
- using the official JUNO software

Data description:

1. positron events
2. uniformly spread in the volume of the central detector
3. $E_{\text{kin}} \in [0, 10]$ MeV. $E_{\text{dep}} = E_{\text{kin}} + 1.022$ MeV

- **Training dataset:**
  4. **2.25 million** events
  5. uniformly distributed in kinetic energy E_{kin}
- **Testing dataset:**
  4. subsets with discrete kinetic energies:
  5. 0, 0.1, 0.3, 0.6, 1, 2, ..., 10 [MeV]
  6. $\sum =$ **0.7 million** events: each subset contains 50k

Aggregated features

We use aggregated information from the whole array of PMTs as features for models:

1. `AccumCharge` — the accumulated charge on fired PMTs
2. `nPMTs` — the total number of fired PMTs
3. Coordinates of the center of charge:

$$(x_{\text{cc}},\ y_{\text{cc}},\ z_{\text{cc}}) = \vec{r}_{\text{cc}} = \frac{\sum_{i=1}^{N_{\text{PMTs}}} \vec{r}_{\text{PMT}_i} \cdot n_{\text{p.e.},i}}{\sum_{i=1}^{N_{\text{PMTs}}} n_{\text{p.e.},i}}$$

   and its radial component: $R_{\text{cc}} = |\vec{r}_{\text{cc}}|$

4. Coordinates of the center of FHT:

$$(x_{\text{cht}},\ y_{\text{cht}},\ z_{\text{cht}}) = \vec{r}_{\text{cht}} = \frac{1}{\sum_{i=1}^{N_{\text{PMTs}}} \frac{1}{t_{\text{ht},i}+c}} \sum_{i=1}^{N_{\text{PMTs}}} \frac{\vec{r}_{\text{PMT}_i}}{t_{\text{ht},i}+c},$$

   and its radial component: $R_{\text{cht}} = |\vec{r}_{\text{cht}}|$

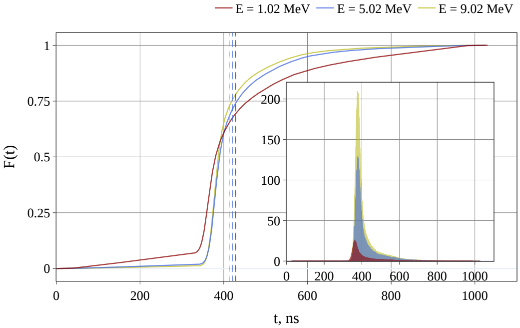
5. $\gamma_z^{\text{cc}} = \frac{z_{\text{cc}}}{\sqrt{x_{\text{cc}}^2+y_{\text{cc}}^2}}$
6. $\gamma_y^{\text{cc}} = \frac{y_{\text{cc}}}{\sqrt{x_{\text{cc}}^2+z_{\text{cc}}^2}}$
7. $\gamma_x^{\text{cc}} = \frac{x_{\text{cc}}}{\sqrt{z_{\text{cc}}^2+y_{\text{cc}}^2}}$
8. $\theta_{\text{cc}} = \arctan \frac{\sqrt{x_{\text{cc}}^2+y_{\text{cc}}^2}}{z_{\text{cc}}}$
9. $\phi_{\text{cc}} = \arctan \frac{y_{\text{cc}}}{x_{\text{cc}}}$
10. $J_{\text{cc}} = R_{\text{cc}}^2 \cdot \sin\theta_{\text{cc}}$
11. $\rho_{\text{cc}} = \sqrt{x_{\text{cc}}^2 + y_{\text{cc}}^2}$
12. with 7 similar features for the components of the center of FHT

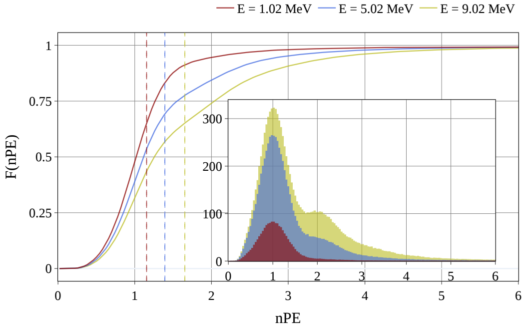
Aggregated features

13. Percentiles of FHT and charge distributions:
    - $\{ht_{2\%}, ht_{5\%}, ht_{10\%}, ht_{15\%}, \ldots, ht_{90\%}, ht_{95\%}\}$
    - $\{pe_{2\%}, pe_{5\%}, pe_{10\%}, pe_{15\%}, \ldots, pe_{90\%}, pe_{95\%}\}$

14. Differences between percentiles for FHT:
    - $\{ht_{5\%-2\%}, ht_{10\%-5\%}, \ldots, ht_{95\%-90\%}\}$

15. Moments for FHT and charge distributions:
    - $\{ht_{mean}, ht_{std}, ht_{skew}, ht_{kurtosis}\}$
    - $\{pe_{mean}, pe_{std}, pe_{skew}, pe_{kurtosis}\}$





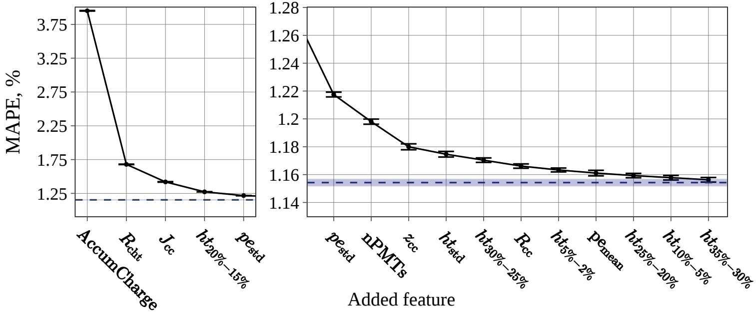
CDFs and PDFs for FHT (left) and charge (right) distributions. $R \simeq 0$ m, E_{kin} varied. Dashes lines show mean values.

Feature selection

- **Feature selection** procedure is performed with *a greedy algorithm* using Boosted Decision Trees (BDT)
- Optimized **set of features** (sorted by *importance*):

1. `AccCharge`
2. R_{cht}
3. J_{cc}
4. $\text{ht}_{20\%-15\%}$
5. pe_{std}
6. `nPMTs`
7. z_{cc}
8. ht_{std}
9. $\text{ht}_{30\%-25\%}$
10. R_{cc}
11. $\text{ht}_{5\%-2\%}$
12. pe_{mean}
13. $\text{ht}_{25\%-20\%}$
14. $\text{ht}_{10\%-5\%}$
15. $\text{ht}_{35\%-30\%}$

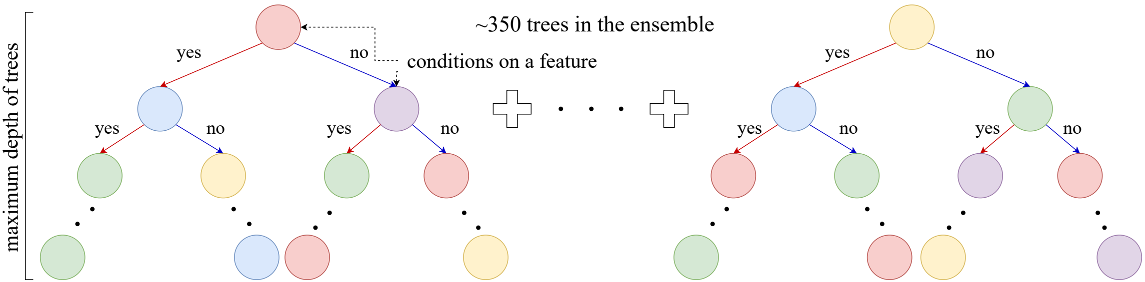
charge-related features (red) — time-related features (blue)



Models description: BDT

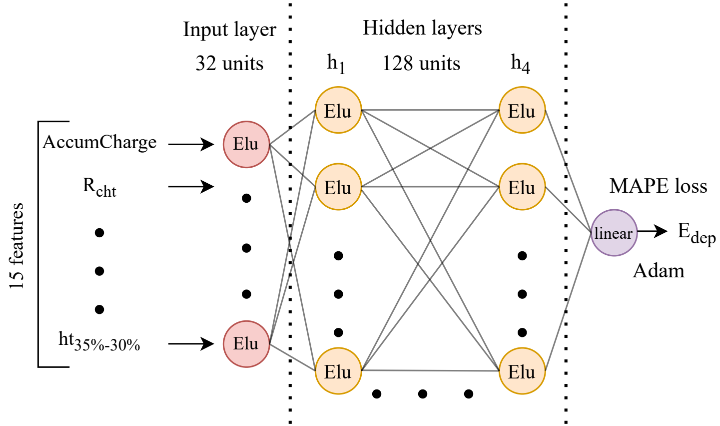
- Optimized hyperparameters (using Grid Search):
  1. The maximum depth of the tree: 11
  2. Number of trees in the ensemble: $\simeq$350
  3. Learning rate: 0.08

- The optimized set of features:
  1. 15 out of 91 features
  2. 6 charge-related
  3. + 8 time-related
  4. + number of fired PMTs



Models description: FCDNN

Fully-connected deep neural network (**FCDNN**):



- Optimization of the hyperparameters using BayesianOptimization
- Training with early stopping
- Validation dataset: 200k events
- The optimized set of features

Results

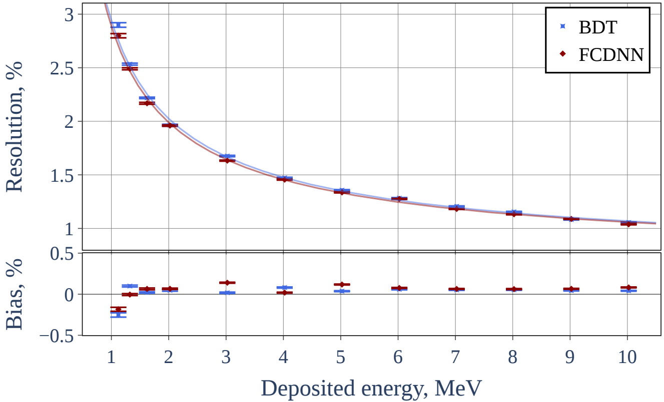
Metrics:

- Defined by a Gaussian fit of the $E_{\text{predicted}} - E_{\text{dep}}$ distributions
- *Resolution*: σ / E_{dep}, where σ — standard deviation of the fit
- *Bias* μ / E_{dep}, where μ — mean of the fit

Parameterization:

$$\frac{\sigma}{E_{\text{dep}}} = \sqrt{\left(\frac{a}{\sqrt{E_{\text{dep}}}}\right)^2 + b^2 + \left(\frac{c}{E_{\text{dep}}}\right)^2}$$

Model	$a \pm \Delta a$	$b \pm \Delta b$	$c \pm \Delta c$
BDT	2.50 ± 0.12	0.71 ± 0.05	1.38 ± 0.29
FCDNN	2.45 ± 0.09	0.71 ± 0.04	1.36 ± 0.23



JUNO energy reconstruction summary

- **Energy reconstruction** using the information collected by PMTs
- *Aggregated* features approach
- The following ML models are used: **BDT, FCDNN**
- As a result <u>*achieved*</u>:
  1. High **quality** <3% @ 1 MeV, required for physics goals of JUNO
  2. Great **computation speed**, thanks to a small set of aggregated features