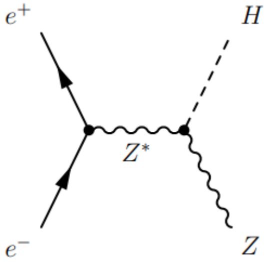
Application of machine learning for the analysis of four-jet final states in the CEPC experiment

A. Staritsyna, M. Chadeeva

Motivation

Measuring the parameters of the Higgs boson and studying rare processes in experiments at the e^+e^- collider

- CEPC - the project of an e^+e^- collider for experimenting on the Higgs boson
- In e^+e^- annihilation, the production of Higgs boson in association with the Z boson has the largest cross-section at $\sqrt{s} = 240$ GeV
- Channels $Z \to q\bar{q}(b\bar{b}, c\bar{c}, etc.)$, $H \to b\bar{b}$ - low reconstruction accuracy, but high probability
- Many background sources, including those due to wrong jet clustering
- **Goal: effective reconstruction of the Higgs boson mass from 4-jet events**



Signal and backgrounds

Delphes package (fast detector simulation and reconstruction)
FastJet package (Jet clustering algorithm - exclusive ee kt)
Particle Flow approach (reconstructed Particle Flow Objects within jets – PFO)

Signal:

- $\boldsymbol{e^+e^- \to Z(q\bar{q})H(b\bar{b})}$ $\boldsymbol{(\sigma \approx 80\ \text{fb})}$

Two fermions backgrounds:

- $e^+e^- \to l^+l^-$
- $e^+e^- \to \nu\bar{\nu}$
- $\boldsymbol{e^+e^- \to q\bar{q}}$ $\boldsymbol{(\sigma \approx 54000\ \text{fb})}$

Four fermions backgrounds:

- Leptonic $(l^+l^-\nu_l\bar{\nu}_l)$
- Semi-leptonic $(l^+l^-q\bar{q}$ or $o\nu_l, \bar{\nu}_l q\bar{q}$ $)$
- **Hadronic $(4q)$ $(\sigma \approx 7500\ \text{fb})$**

Higgs backgrounds:

- $e^+e^- \to Z(q\bar{q})H(WW^*)$
- $e^+e^- \to Z(q\bar{q})H(gg)$
- $e^+e^- \to Z(q\bar{q})H(\tau^+\tau^-)$
- $e^+e^- \to Z(q\bar{q})H(c\bar{c})$

Pre-selection

- $N_{jets} = 4$
- Total $N_{PFO} \geq 40$
- Charged $N_{PFO} \geq 25$
- Min$(N_{PFOinJet}) \geq 6$
- 2-nd min$(N_{PFOinJet}) \geq 10$
- Max$(D_{xy}, D_z$ significance$)\geq 5$
- Second max$(D_{xy}, D_z$ significance$)\geq 5$
- Aplanarity ≥ 0.06
- Total energy ≥ 170 GeV
- Max(photon energy) ≤ 40 GeV
- Mim(jet energy) ≥ 15 GeV
- Total invariant mass ≥ 150 GeV
- $m_Z \in [55; 110]$ GeV (Z candidate)
- $m_H \in [90; 150]$ GeV (H candidate)
- $(m_H - m_Z) \in [-20; 65]$ GeV
- $cos(\theta_{sum}) \leq 0.9$

Expected number of events after pre-selection

- Leptonic and semi-leptonic backgrounds: $\leq 0.1\%$, $\leq 8\ 500$
- Signal: 57%, **250 000**
- Higgs backgrounds : $0\% - 10\%$, **21 000**
- Hadronic (4 jets): $15\% - 35\%$, **8 700 000**
- $q\bar{q}$: 1.03%, **3 121 000**

Model Architecture: Transformer for classification

Input features

Jets (4×20):

- Energy, $p_x, p_y, p_z, \phi, \cos\theta$
- Number of neutral PFO, kaons, pions, protons
- Max(D_{xy} significance), D_{xy}
- 2-nd max(D_{xy} significance)
- Number of PFO with D_{xy} significance $> 3, > 5$
- Max(D_z significance)
- 2-nd max(D_z significance), D_z
- Number of PFO with D_z significance $> 3, > 5$

Event shape characteristics:

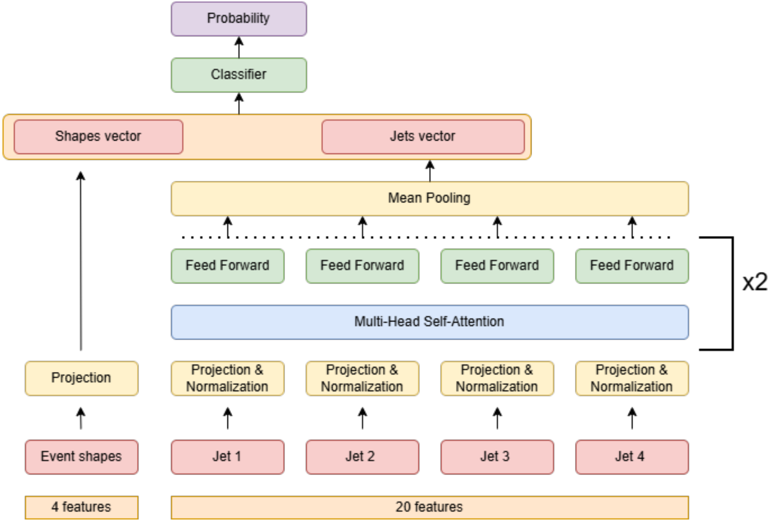
- Thrust
- Sphericity
- Planarity
- Aplanarity

Input Samples

- Signal (1:15)
- Higgs backgrounds (4:15)
- Hadronic (4 jets) (9:15)
- $q\bar{q}$ (1:15)

- Each channel (1:15) - 1 000 000 events
- Supervised learning

Model Architecture: Transformer for classification



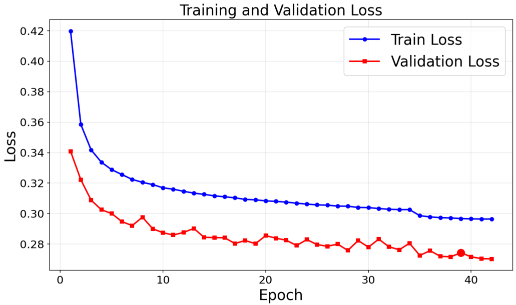
Transformer parameters

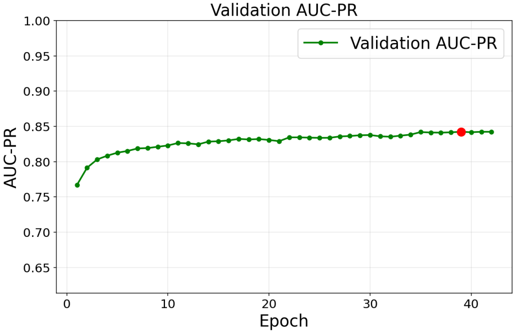
- $d_{\text{model}} = 128$, $n_{\text{head}} = 4$
- $n_{\text{layers}} = 2$, $d_{\text{FF}} = 256$
- Dropout $= 0.3$, GELU

MLP Classifier

- Linear$(256 \to 64)$ + GELU + Dropout(0.1)
- Linear$(64 \to 32)$ + GELU
- Linear$(32 \to 1)$ + Sigmoid

Model training (no Pre-selection)





Learning parameters

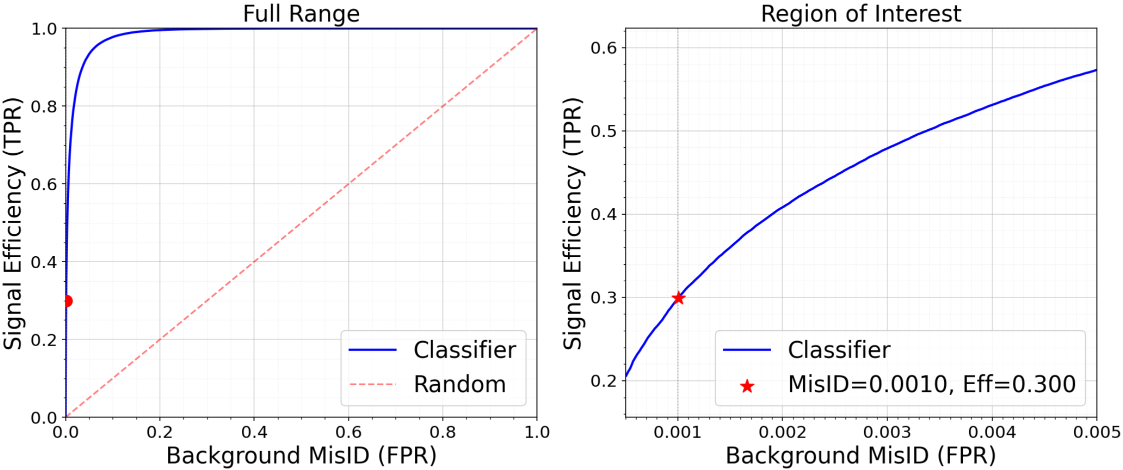
- batch_size = 512, epochs = 42, lr = 0.0001, weight_decay = 10^{-4}

Early-stopping criteria

- $G = Loss_{valid} - Loss_{train}$
- 7 epoch: $(G \geq 0)\&(\Delta G > 0.01||\varepsilon_G > 0.1||\varepsilon_{Loss_{valid}} > 0.1)$
- 7 epoch: AUC-PR improvement less 0.001

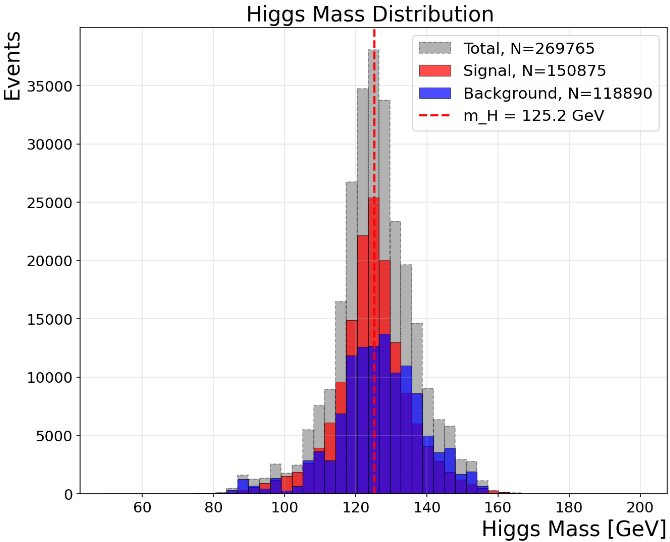
Model training (no Pre-selected data)

Curve for working point selection - signal efficiency vs background misidentification.



Red star corresponds to 0.1% background misidentification (working point).

Model test (no Pre-selection), non pre-selected data

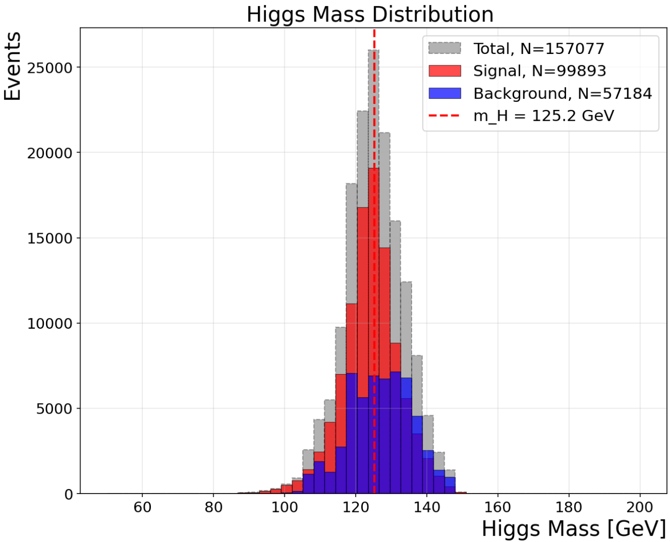


- Reconstruct Higgs bozon (m_{jj}) and Z bozon from jets
- Calculate m_{recoil} from Z bozon $(E_Z, \bar{p}_Z)$:
  $E_{recoil} = E_{tot} - E_Z$,
  $\bar{p}_{recoil} = -\bar{p}_Z$
- $m_H = (m_{jj} + m_{recoil})/2$

Weighted statistics

- **Total efficiency 34%**
- **Total purity 56%**

Model test (no Pre-selection), Pre-selected data

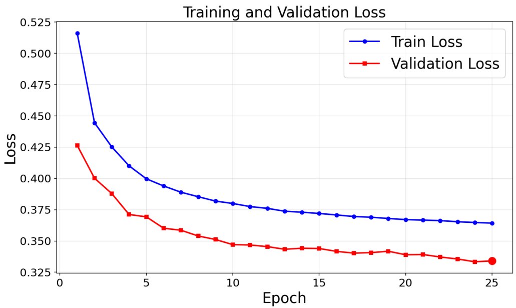


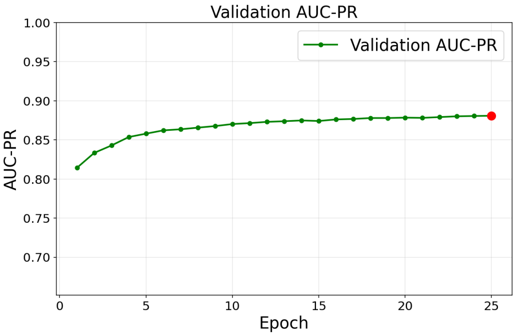
- Reconstruct Higgs bozon (m_{jj}) and Z bozon from jets
- Calculate m_{recoil} from Z bozon $(E_Z, \bar{p}_Z)$:
  $E_{recoil} = E_{tot} - E_Z,$
  $\bar{p}_{recoil} = -\bar{p}_Z$
- $m_H = (m_{jj} + m_{recoil})/2$

Weighted statistics

- **Total efficiency 23%**
- **Total purity 64%**

Model training (with Pre-selection)





Learning parameters

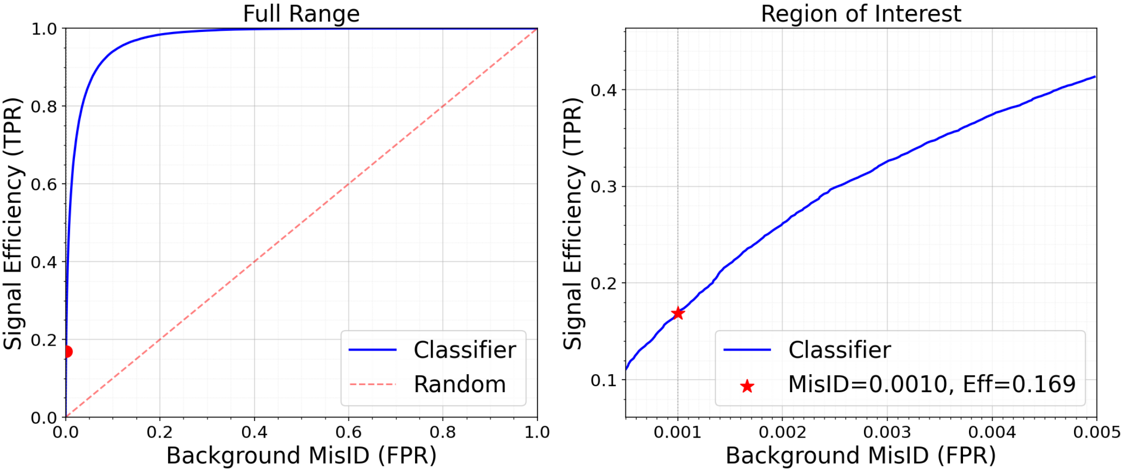
- batch_size = 512, epochs = 25, lr = 0.0001, weight_decay = 10^{-4}

Early-stopping criteria

- $G = Loss_{valid} - Loss_{train}$
- 7 epoch: $(G \geq 0)\&(\Delta G > 0.01||\varepsilon_G > 0.1||\varepsilon_{Loss_{valid}} > 0.1)$
- 7 epoch: AUC-PR improvement less 0.001

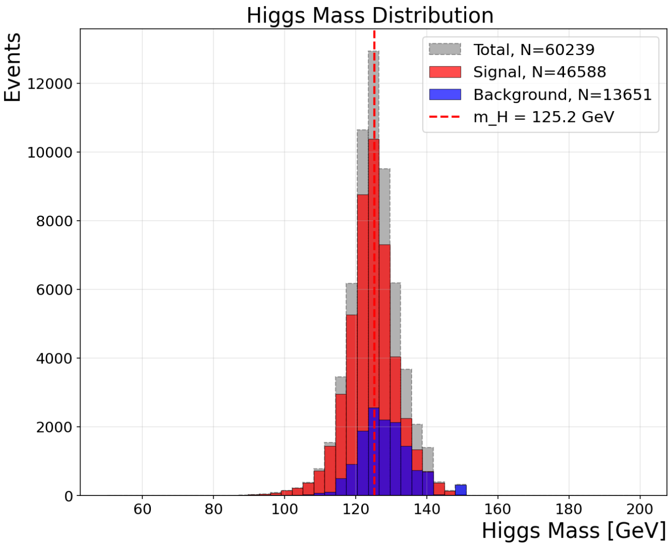
Model training (with Pre-selection)

Curve for working point selection - signal efficiency vs background misidentification.



Red star corresponds to 0.1% background misidentification (working point).

Model test (Pre-selection), Pre-selected data



- Reconstruct Higgs bozon (m_{jj}) and Z bozon from jets
- Calculate m_{recoil} from Z bozon $(E_Z, \bar{p}_Z)$:
  $E_{recoil} = E_{tot} - E_Z,$
  $\bar{p}_{recoil} = -\bar{p}_Z$
- $m_H = (m_{jj} + m_{recoil})/2$

Weighted statistics

- **Total efficiency 11%**
- **Total purity 77%**

Results

Model training results

- Application of machine learning for Higgs analysis in the planned CEPC experiment
- Transformer model in the PyTorch framework is used for classification task for background rejection
- The best approach - training at pre-selected data
- Signal efficiency 11%
- Purity 77%

Further plans

- Optimize the selection criteria
- Add $H(b\bar{b})Z(ff)$ background
- Optimize the input features
- Test on full-simulation data, statistical cross-check on different test subsamples

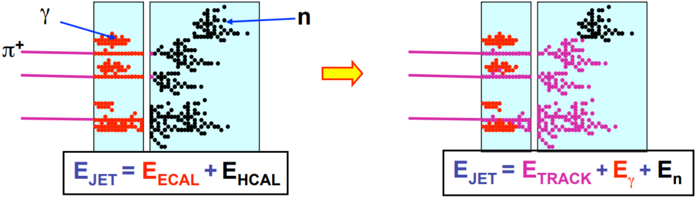
Additional slides

Particle Flow Approach

When reconstructing the jets, the **Particle Flow** method is used.
Jet's structure:
$\sim$60% - charged particles, $\sim$25-30% - photons, and the rest - neutral hadrons.
Information from detectors is taken into account:

- Tracker
- Electromagnetic calorimeter
- Hadron calorimeter - the main contribution to the error

Improving the resolution of the hadron calorimeter is important for reconstructing jets during clustering.



https://indico.cern.ch/event/124299/contributions/93837/subcontributions/8961/attachments/74346/106
PFLOW.pdf

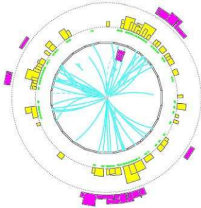
Definition and requirements

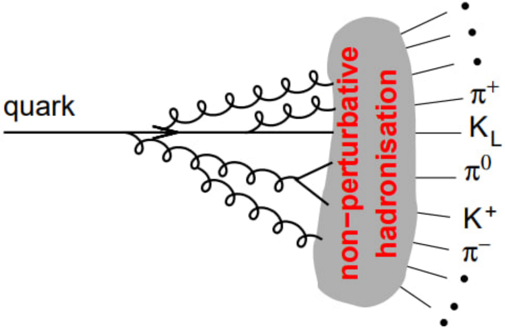
Jet's definition depends on:

- Clusterization method (which particles get put together into a common jet?)
- Recombination scheme

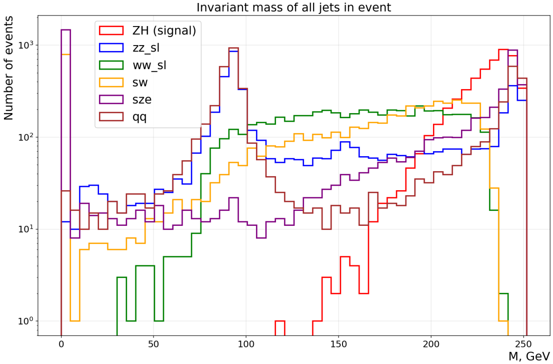
Jets should be **invariant** to certain modifications of the event:

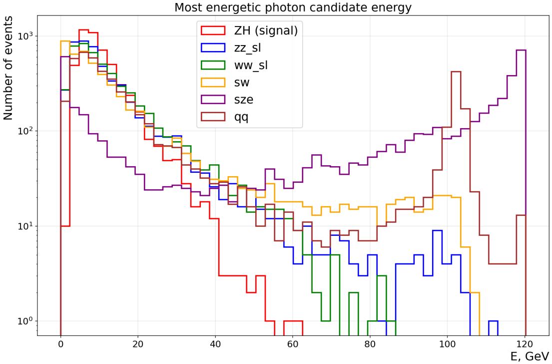
- collinear splitting
- infrared emission





Distributions of the invariant mass and maximum photon energy





Distributions of the N_{PFO} and total E