# Application of Machine Learning for the analysis of the Compton scattering data

**Vahe Sokhoyan**

**A2 Collaboration Meeting**
**04.05.2023**

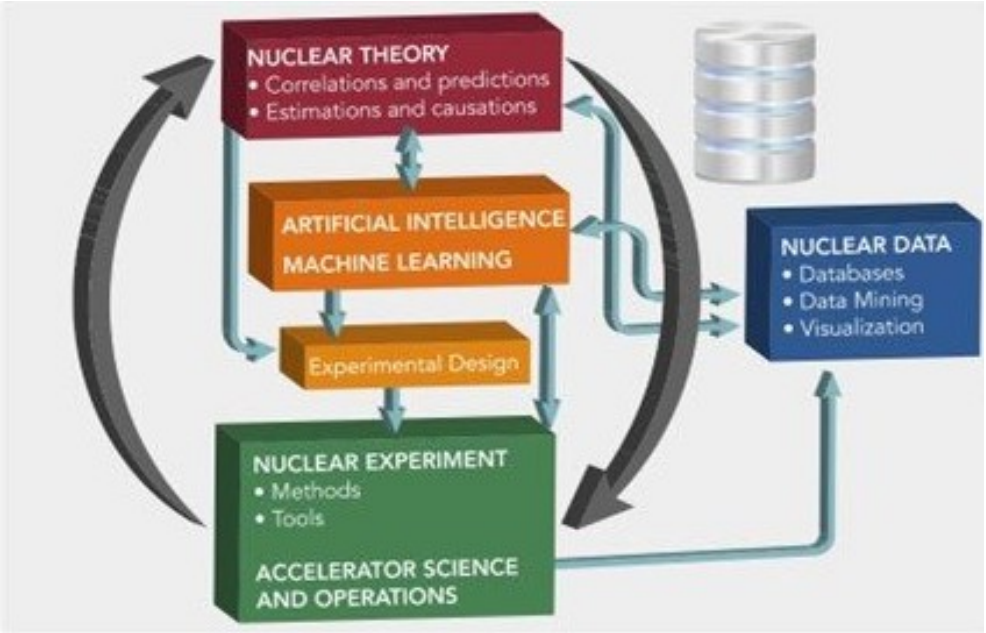


# Motivation

## High precision experiments (including Compton scattering):

- Limited by the presence backgrounds
- Introduction by uncertainties in the process of background identification
- Background subtraction leads to increased errors
- In many cases the information about correlations of kinematic variables is lost
- Preserving the correlations is important (event-by event approach), for phenomenological analysis of experimental results

- Multivariate analysis with event-by-event handling of the data!
- Classification and selection of signal instead of background subtraction
- Alternative to "conventional" reaction analysis is Machine Learning: Building models performing tasks without explicit instructions



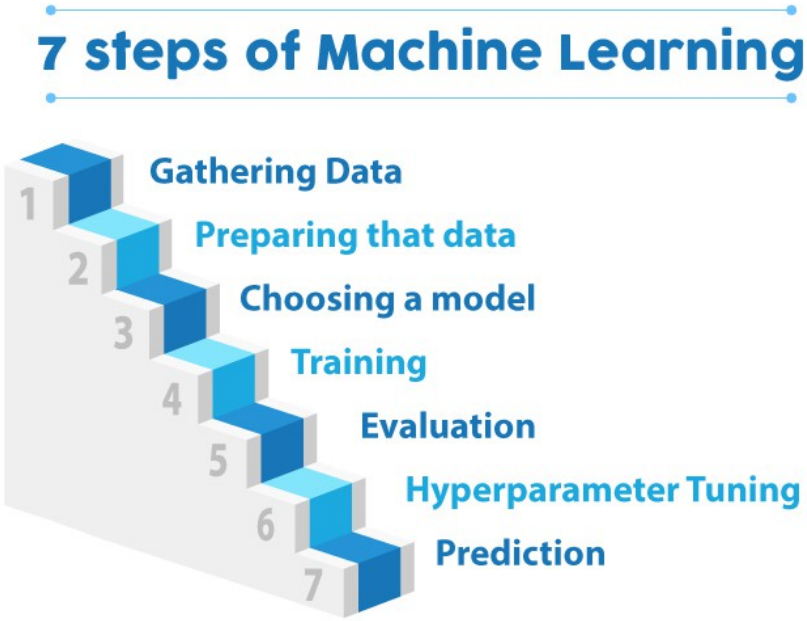
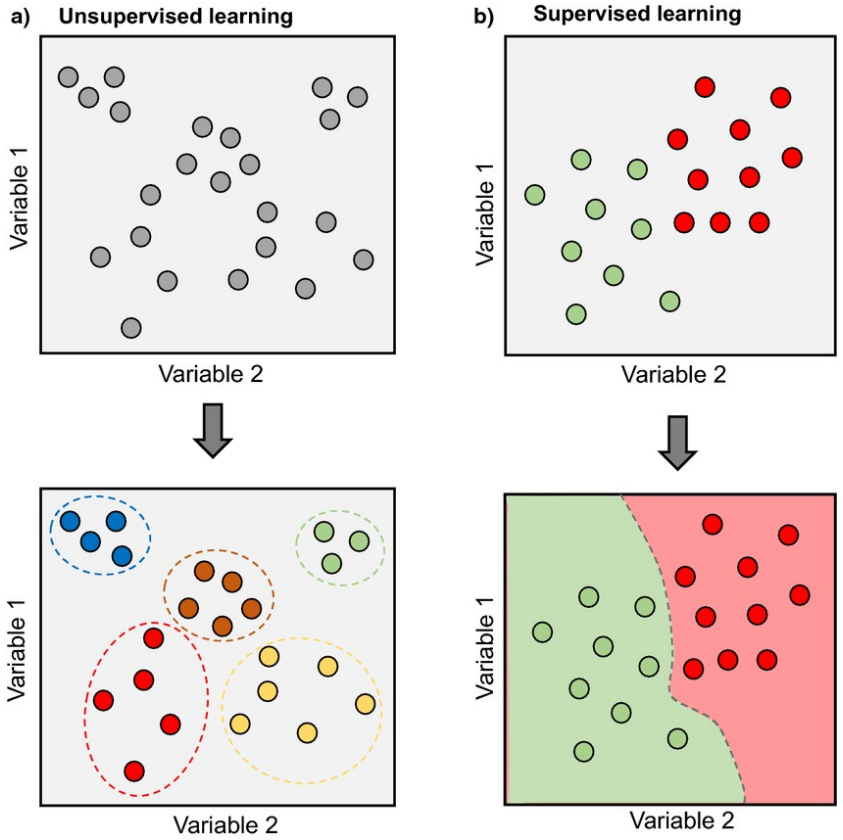




Image courtesy of DOE's Thomas Jefferson National Accelerator Facility (JLab)

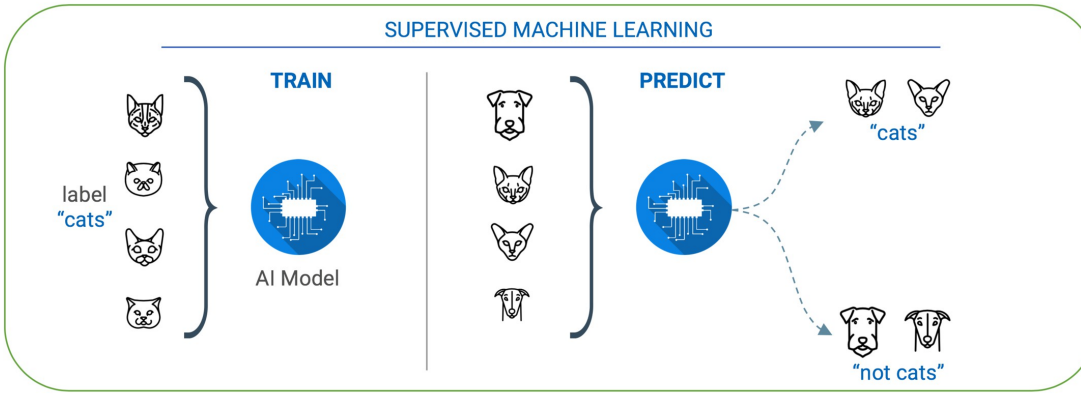
# Supervised and unsupervised Machine Learning

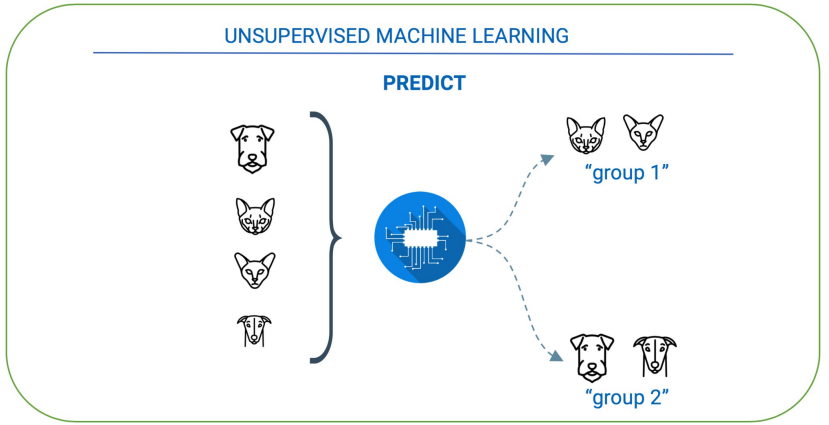



Morimoto, Juliano & Ponton, Fleur. (2021)

→ Supervised learning with initially "known" patterns and labeled data
→ Unsupervised learning via clustering for the data without labels
→ Semi-supervised learning, e.g., with labels created via clustering, or via additional information (and in our case MC simulation)

**In this work:**
→ Multidimensional clustering of events with further classification (only statistical)
→ Training models on MC or mixture of MC and data samples



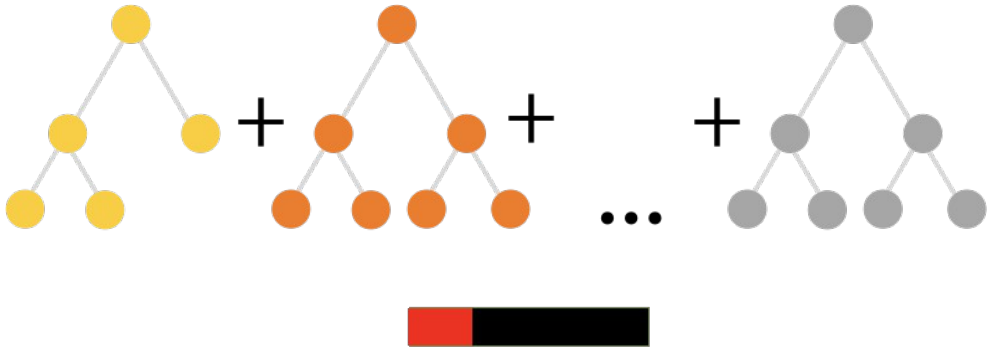

https://abeyon.com/how-do-machines-learn/

# Machine Learning algorithms

- Multiple algorithms based of different principles are available: Deep Learning with neural networks, decision trees, random forest, boosted decision trees, support vectors machines (SVM), ....

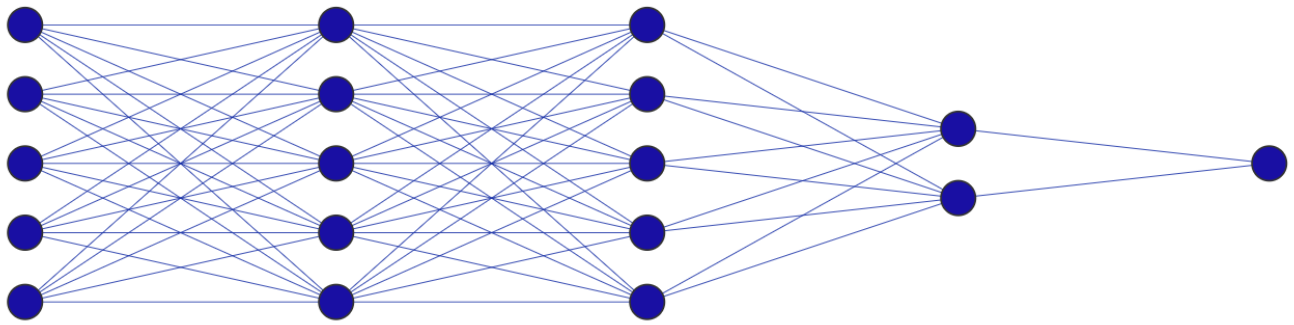
## Boosted decision trees:

→ Sequences (trees) of decisions
→ Ensemble of decision trees
→ Learning from previous training steps
→ Very well suited for classification tasks



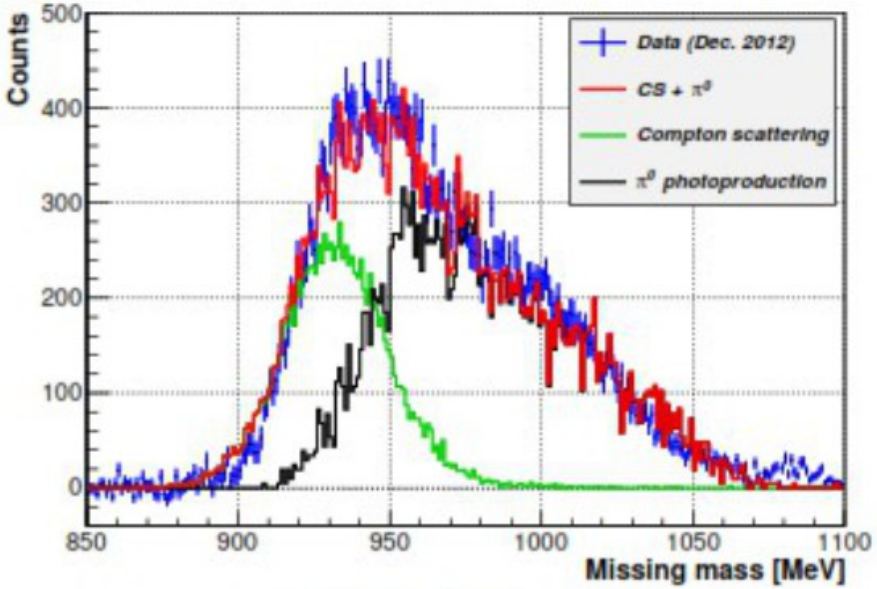
## Neural networks:

→ Structure of nodes sorted in layers with assignment/optimization of weights
→ Sensitive to patterns/low-level features
→ Requires more careful consideration when training
→ Used as an alternative approach in this work

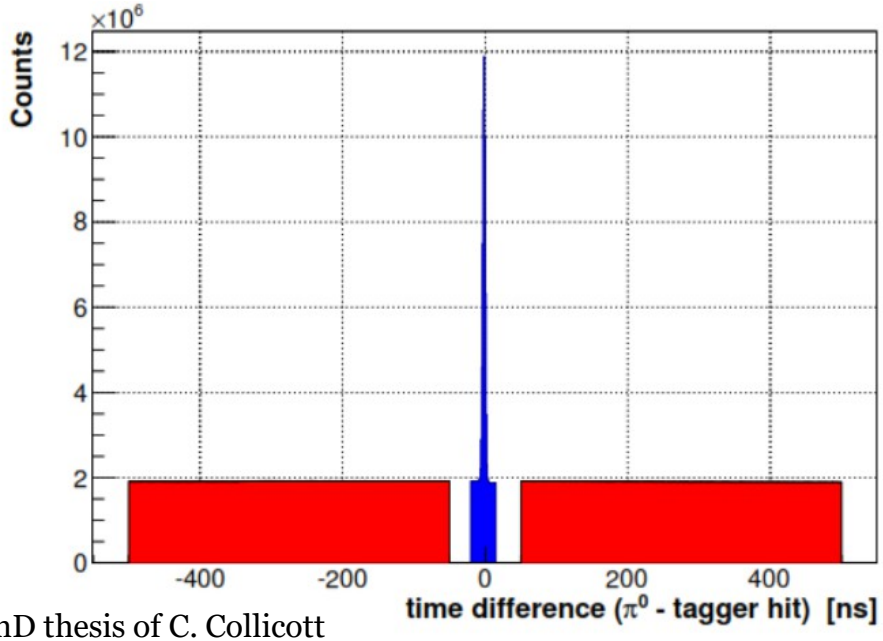


# Application of Machine Learning for Compton analysis

- Multiple (pre) analyzed data sets present for Compton scattering above pion threshold → improved background identification
- Separation of $\pi^0$ background is very challenging, in particular on an event-by event basis
- Presence of random timing background limits the accuracy of the measurements (as in most for most of the other analyses at A2 and tagged photon facilities in general)




(c) $\theta = 90°$ to $100°$

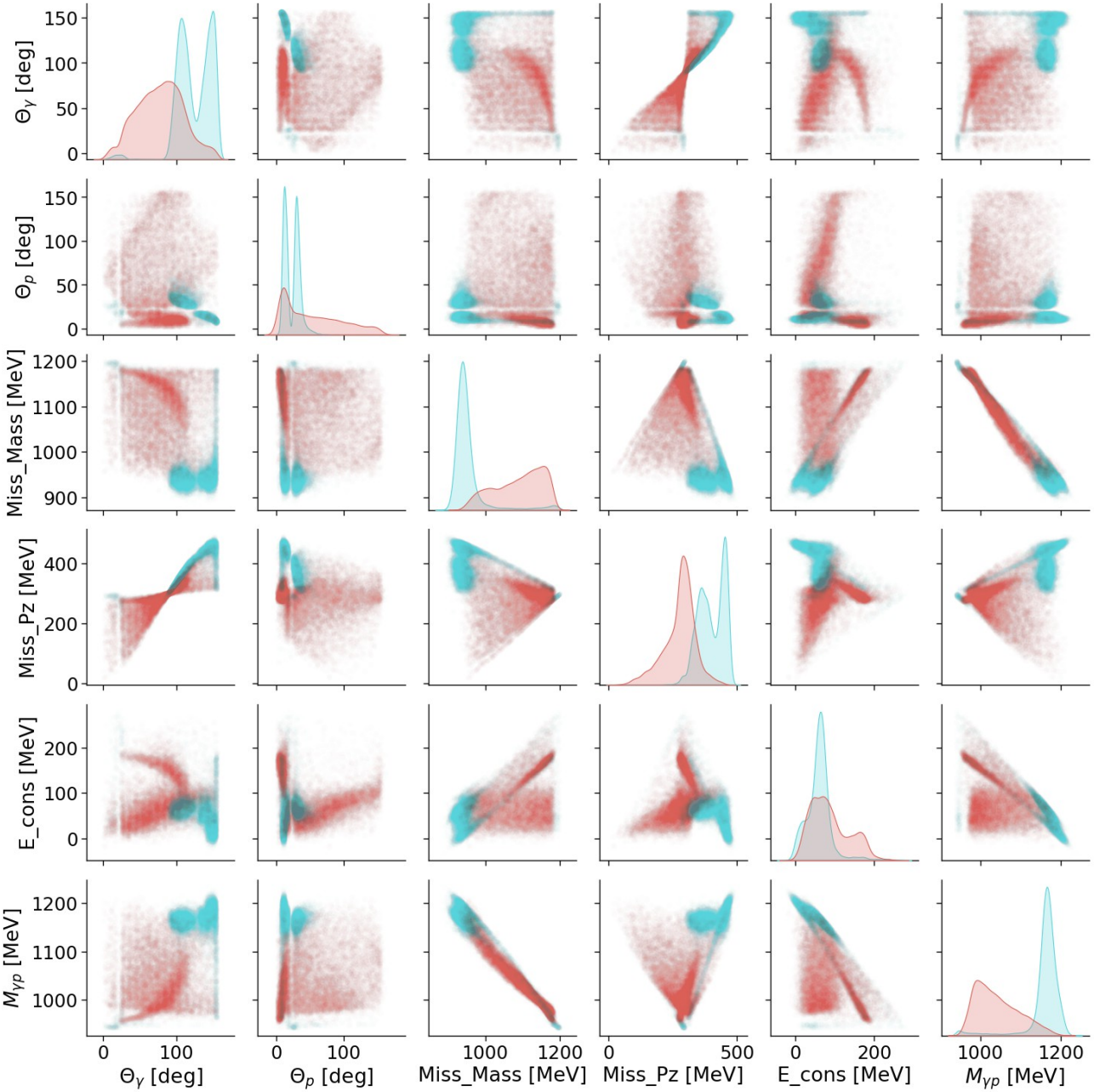



Figures from PhD thesis of C. Collicott

→ Separation of $\pi^0$ background from Compton events with Machine Learning

→ New method for time background handling without random subtraction
→ Outlook for the analysis of Compton scattering data with Machine Learning

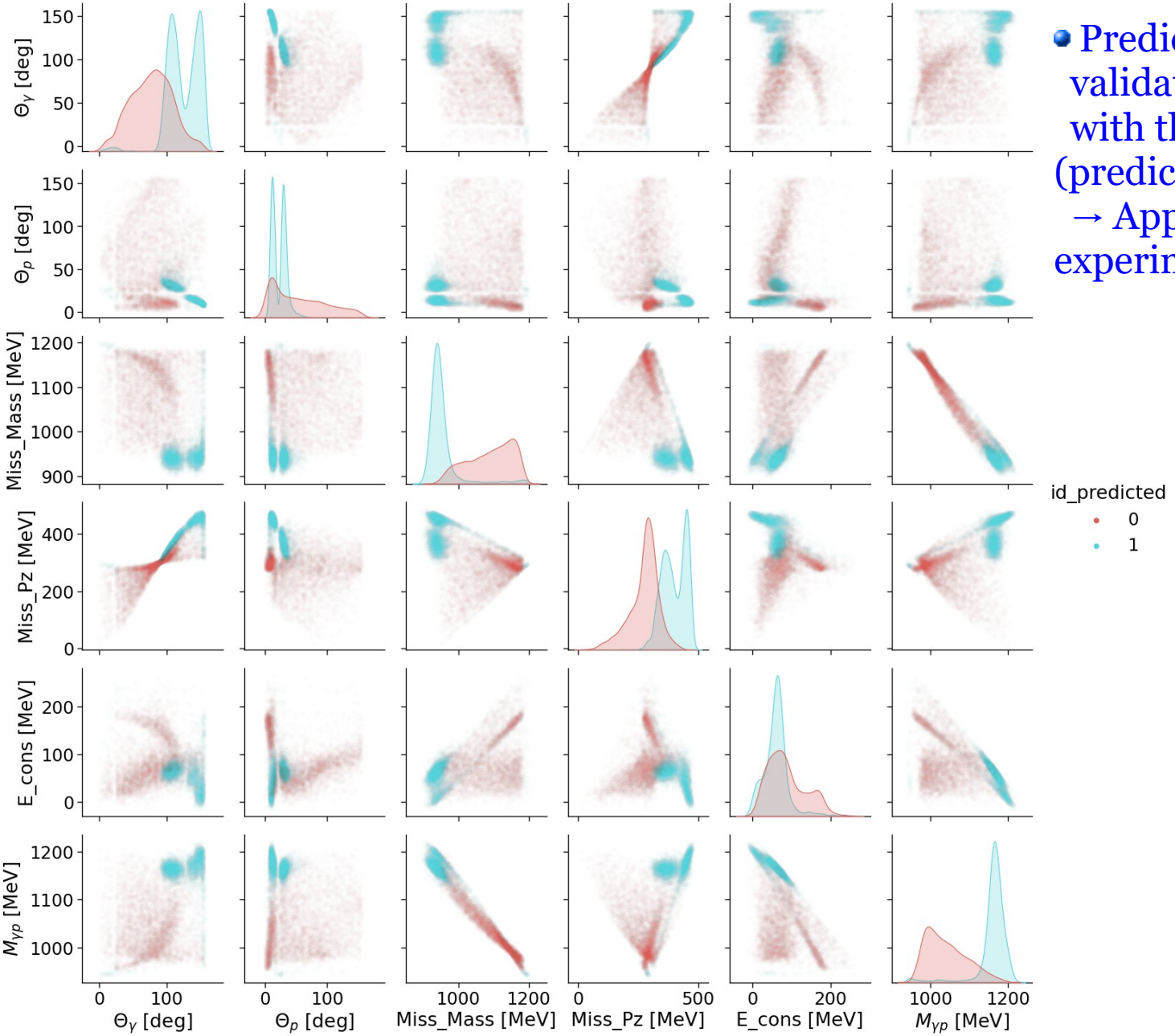
# Separation of pion and Compton events: Input




Input data for the model training:
→ 295 - 305 MeV, γ + p events
→ Notable overlap in 1D
→ Complex shapes in 2D with opportunity of separation

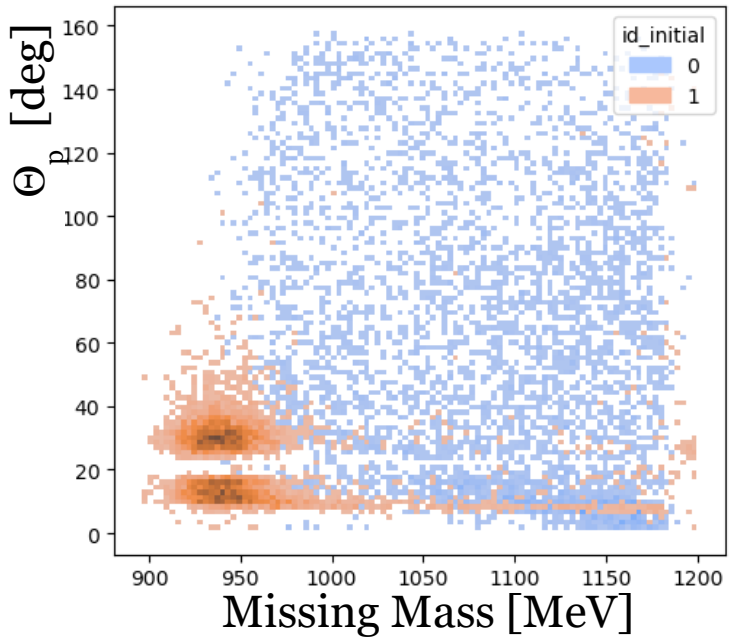
Processing the data:
→ Mix Compton and pion events
→ Reshuffle the (labeled) data
→ Split the data into training and validation data sets
→ Train and evaluate the model (boosted decision trees with XGBoost/CatBoost or neural network)

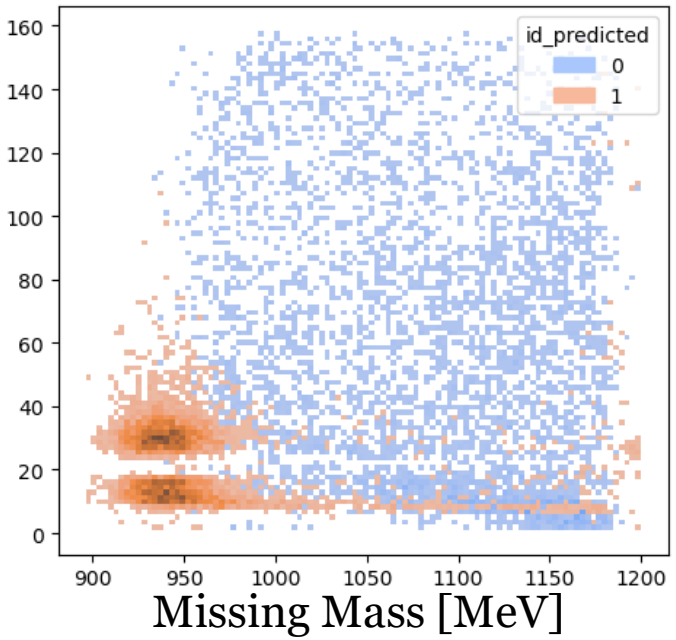
# Separation of pion and Compton events: Predictions

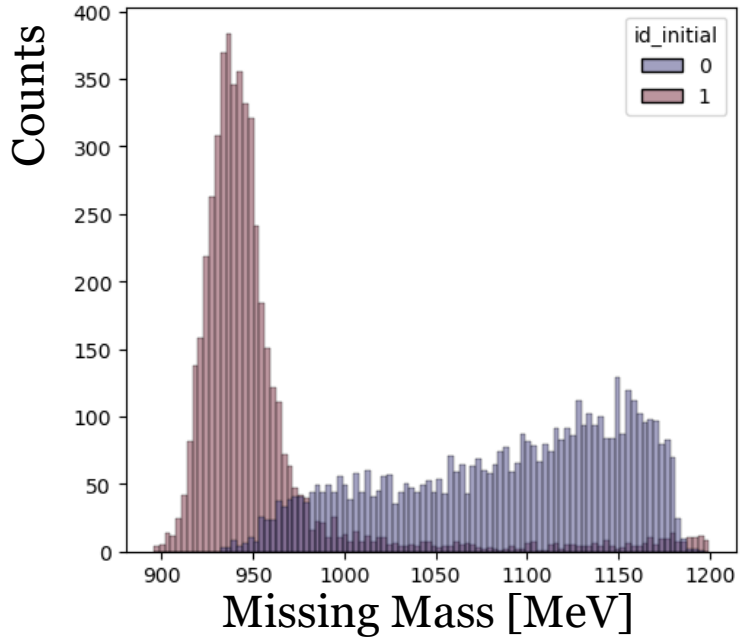


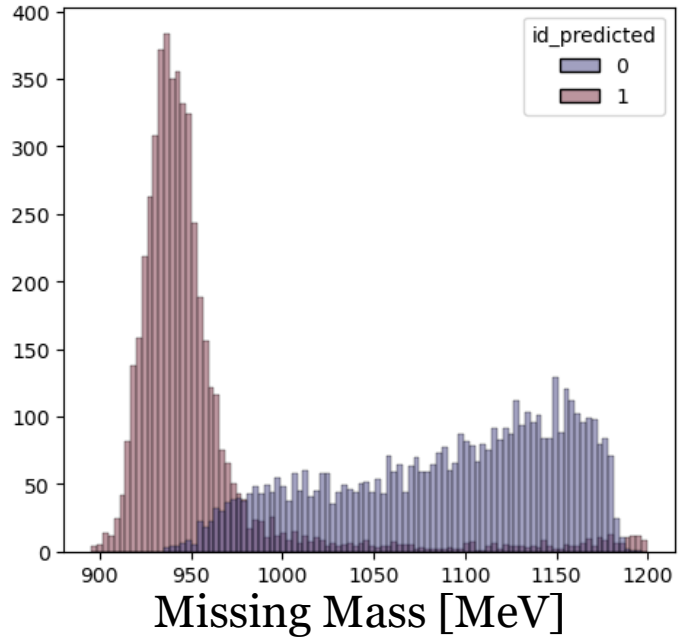

- Predicted distributions for the validation data set agree well with the initial MC data set (predicted with 99% accuracy)
  → Application on the experimental data

# Classification of pion and Compton events



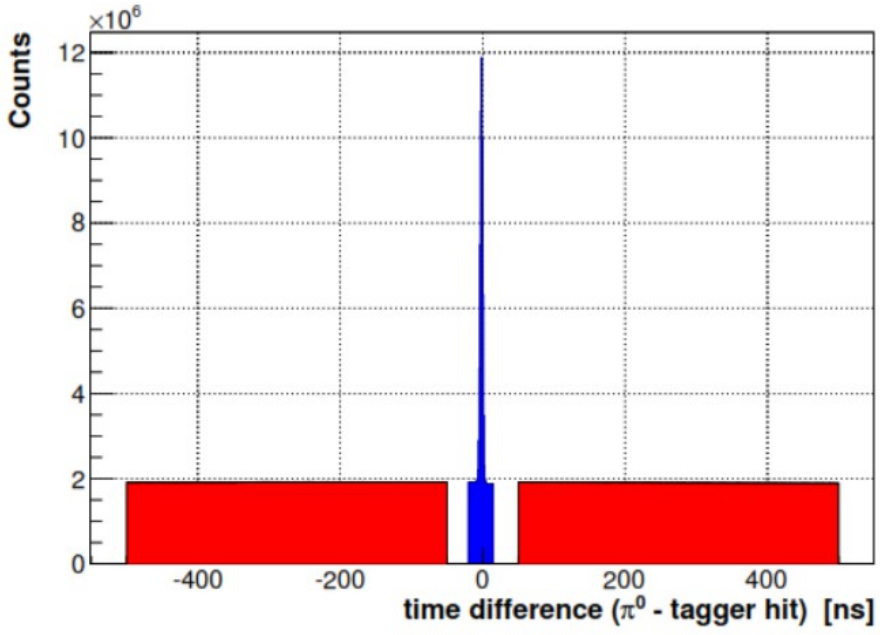



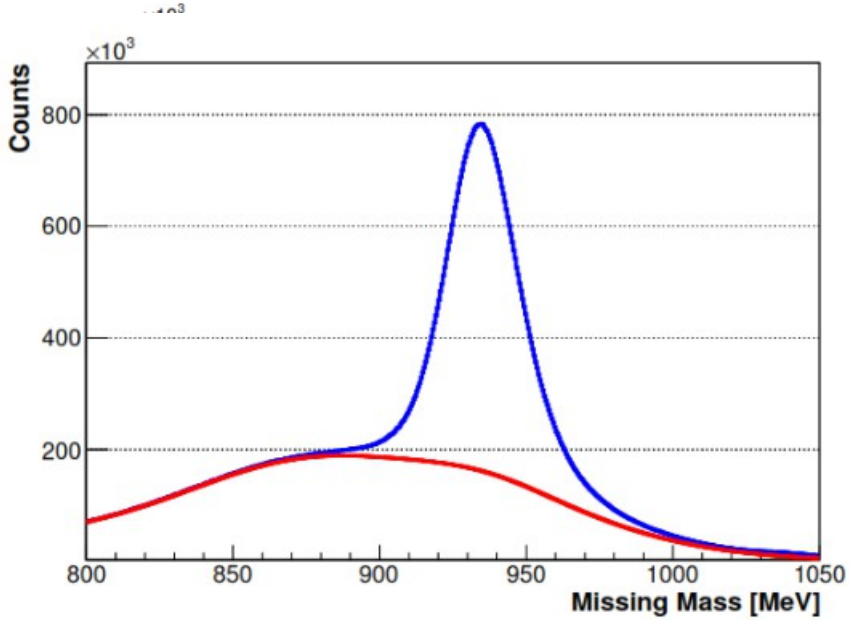




Initial and predicted data agree (also in the overlap region)

# Machine Learning approach for handling random background




Figures from PhD thesis of C. Collicott

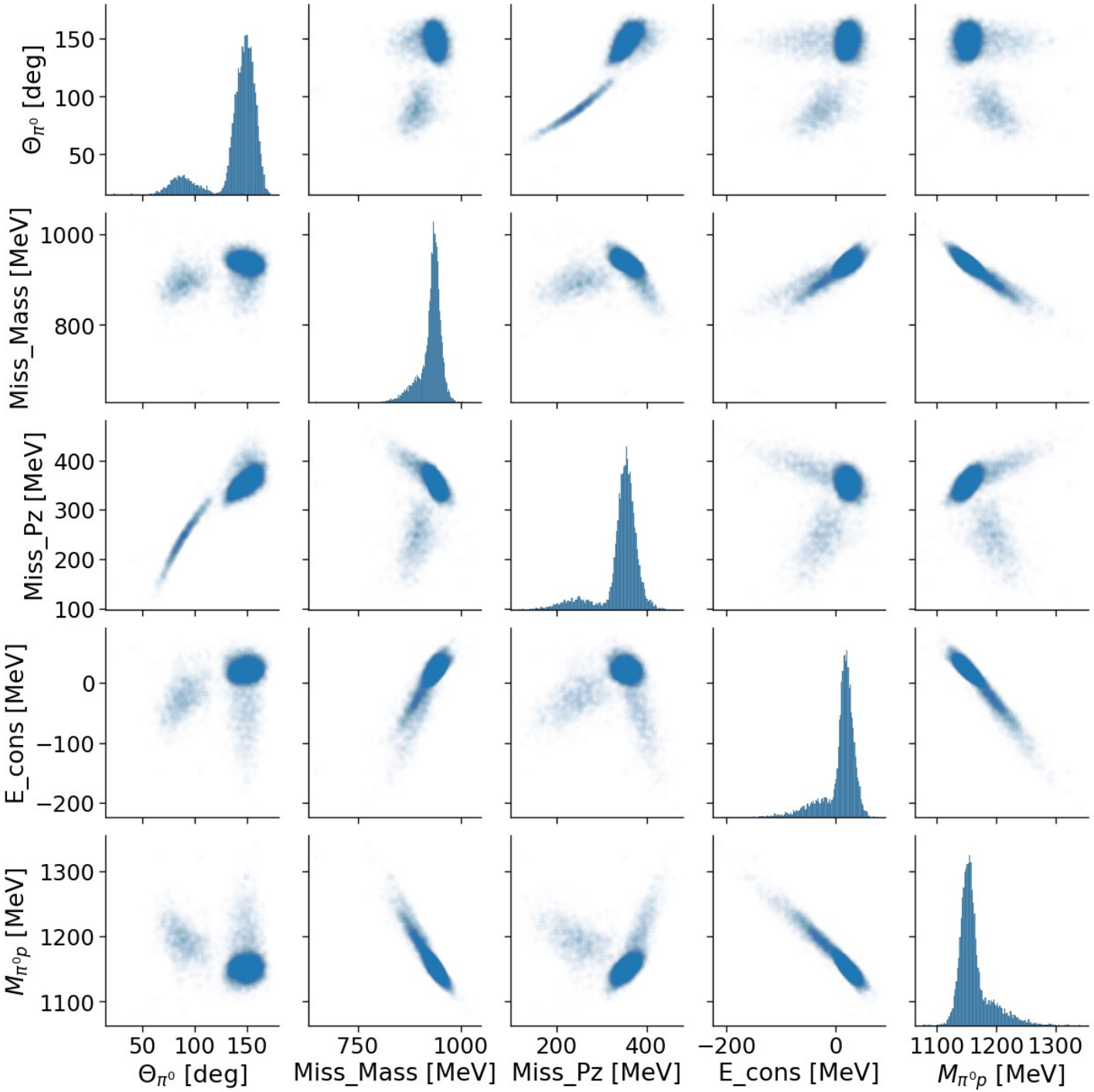
- Handle timing background needed for ML-based data analysis
- Limits precision of many experiments due to subtraction of the background in the classical method

**Separation of the prompt (signal) events with Machine Learning:**

→ Multidimensional clustering without labels (purely statistical approach)

→ MC-based approach using the simulation of the known reaction and measured background for training ML models (requires agreement between data and MC)

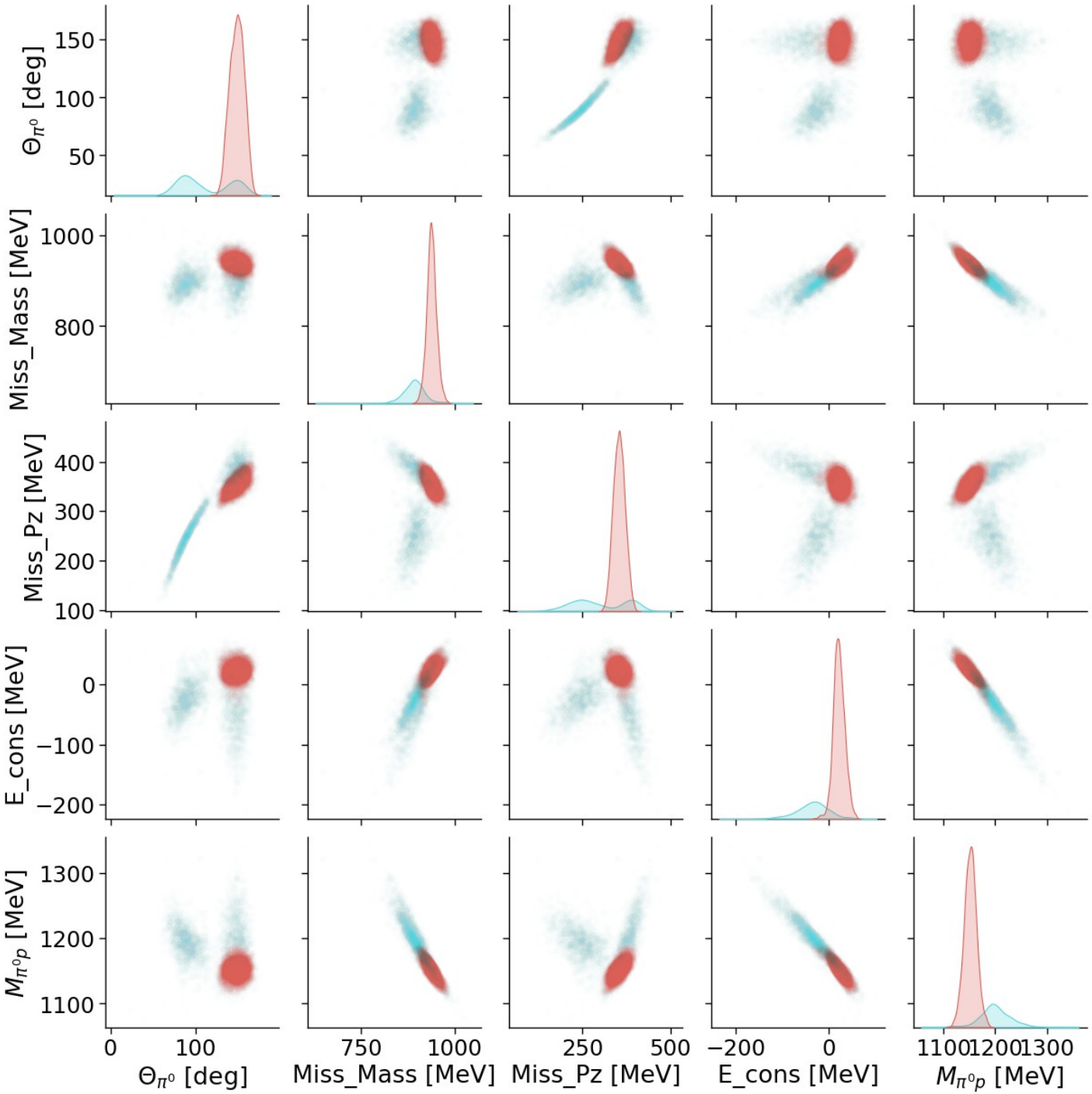
# Separation of random background with DBScan




→ $\pi^0$ events in the prompt peak (-2; 2) ns
→ 240 -260 MeV
→ $m\gamma\gamma$, $\Delta\phi$, $\Delta\Theta$ cuts

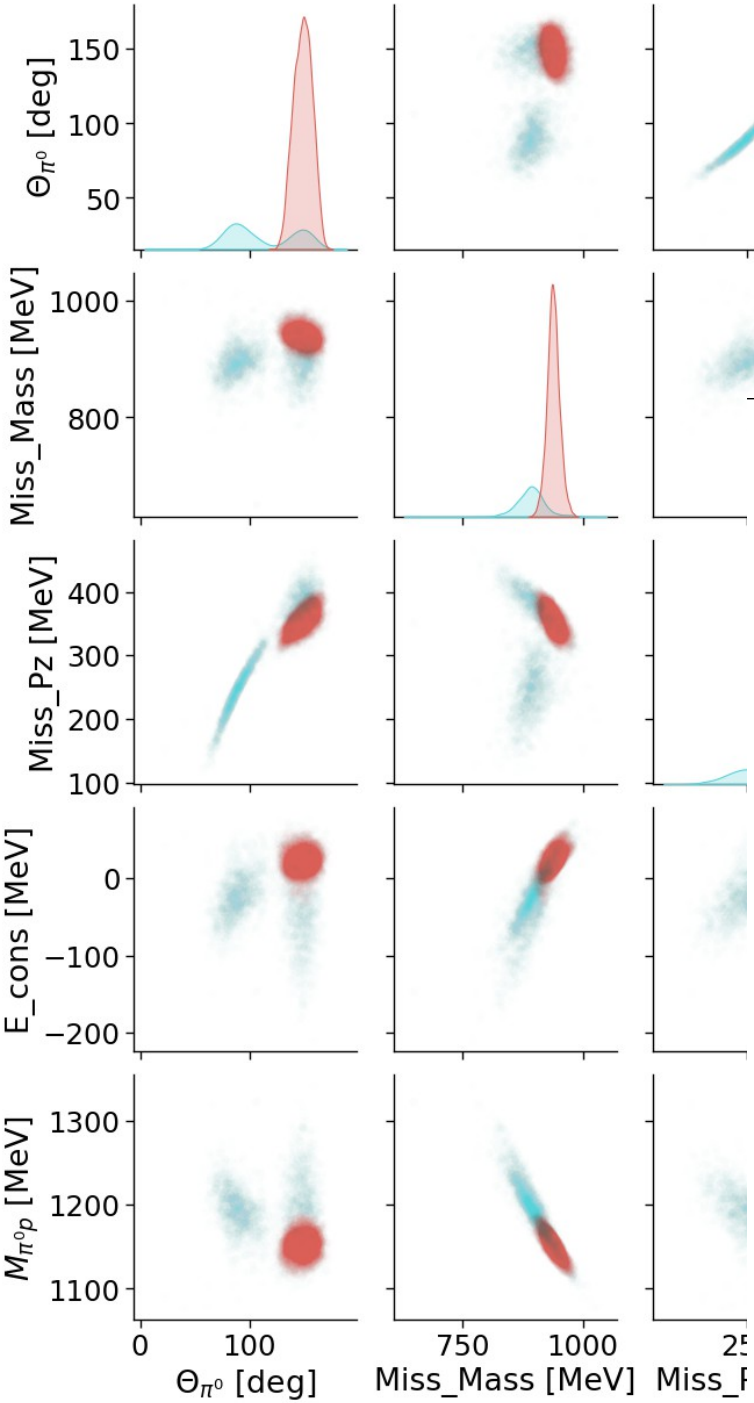
→ Unsupervised approach
→ Separation with density-based clustering algorithm (DBScan)

# Separation of random background with DBScan

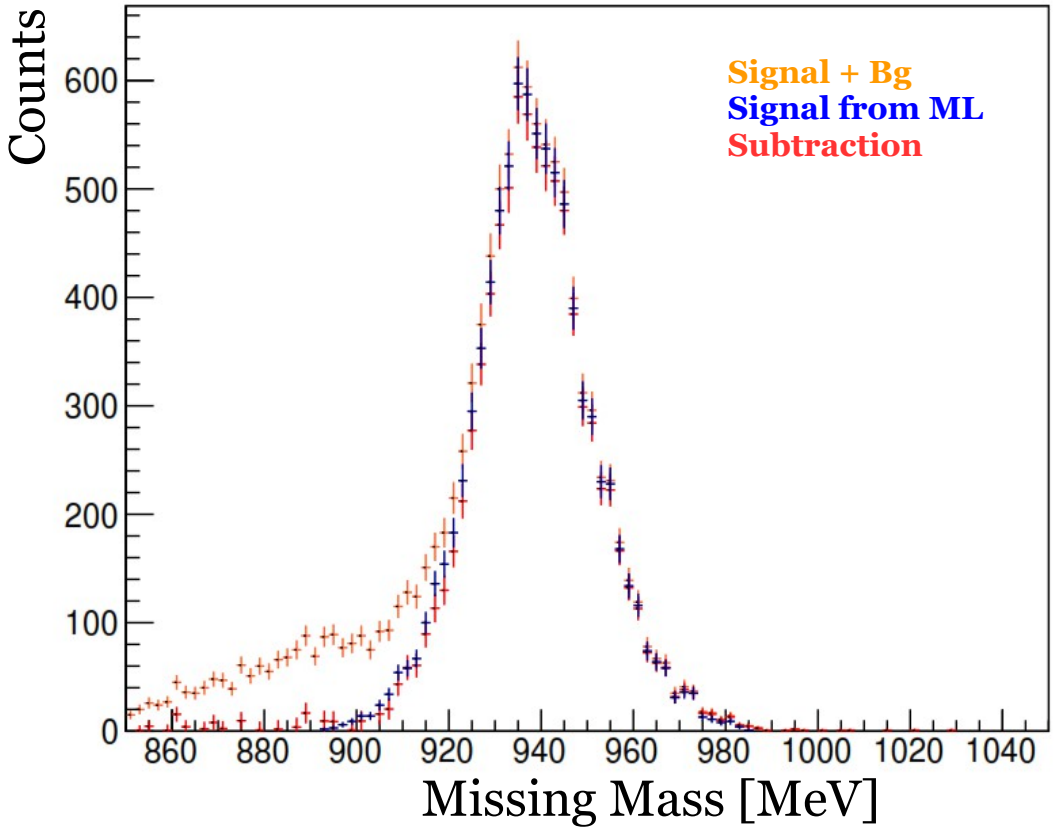



- Good agreement between subtraction and clustering!
  → Still, there is room for improvement

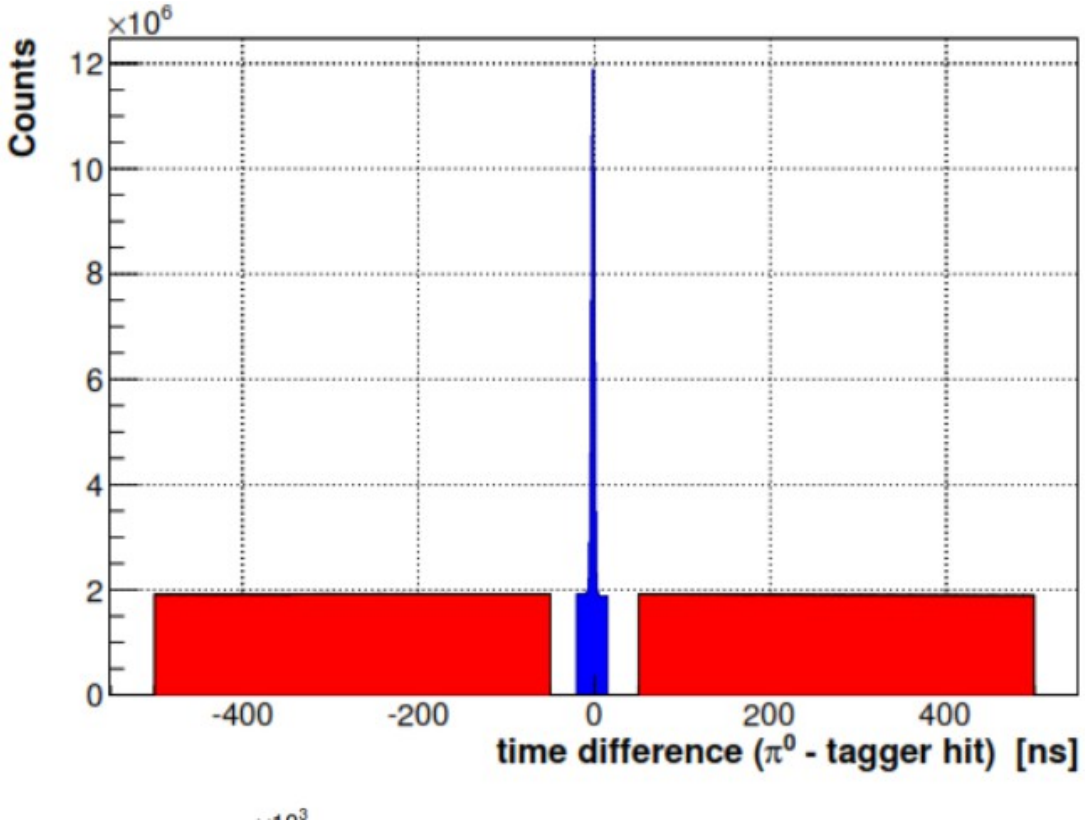
# Separation of random background with DBScan




- Good agreement between subtraction and clustering!
  → Still, there is room for improvement

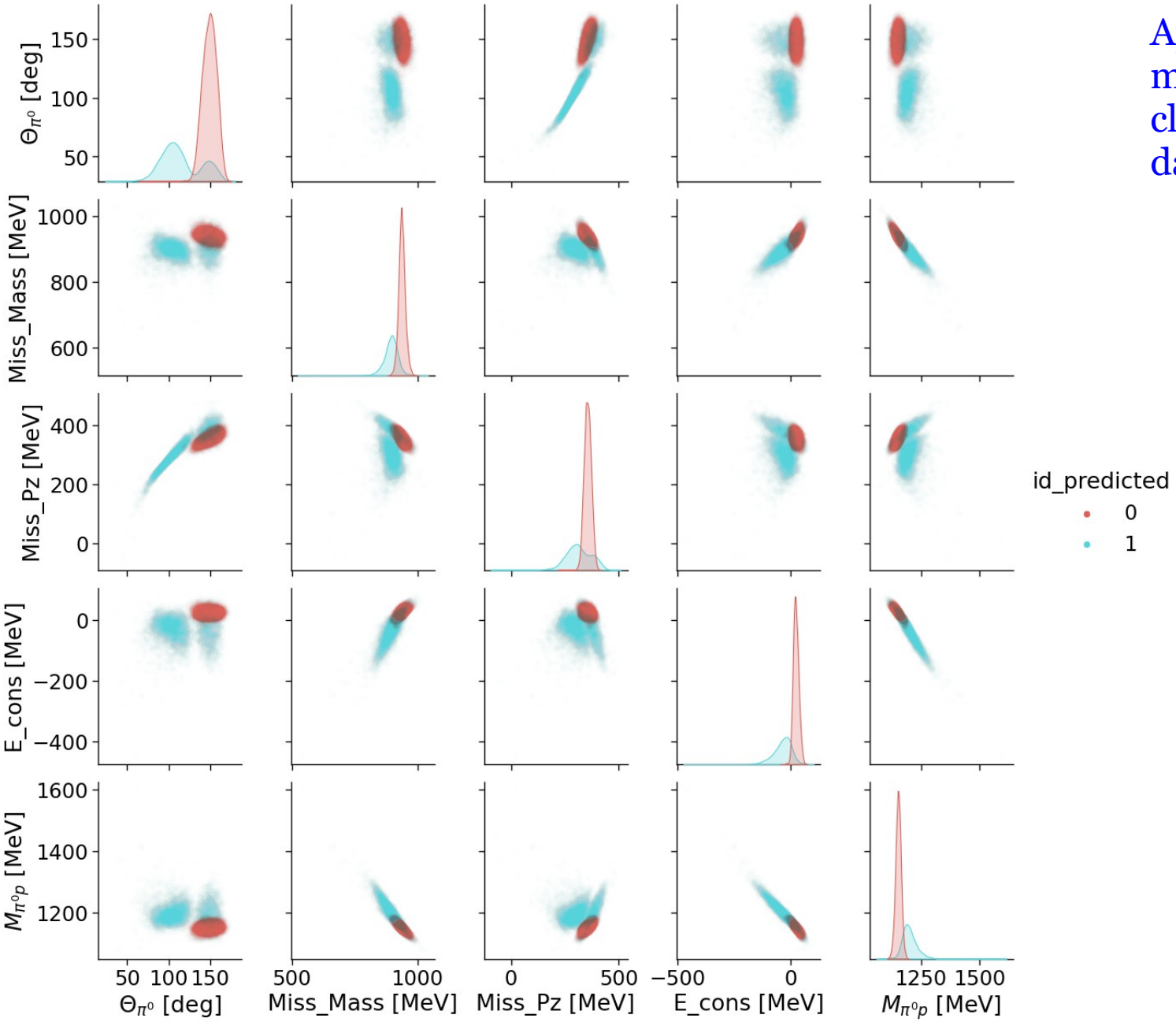


# From clustering to (semi)supervised learning



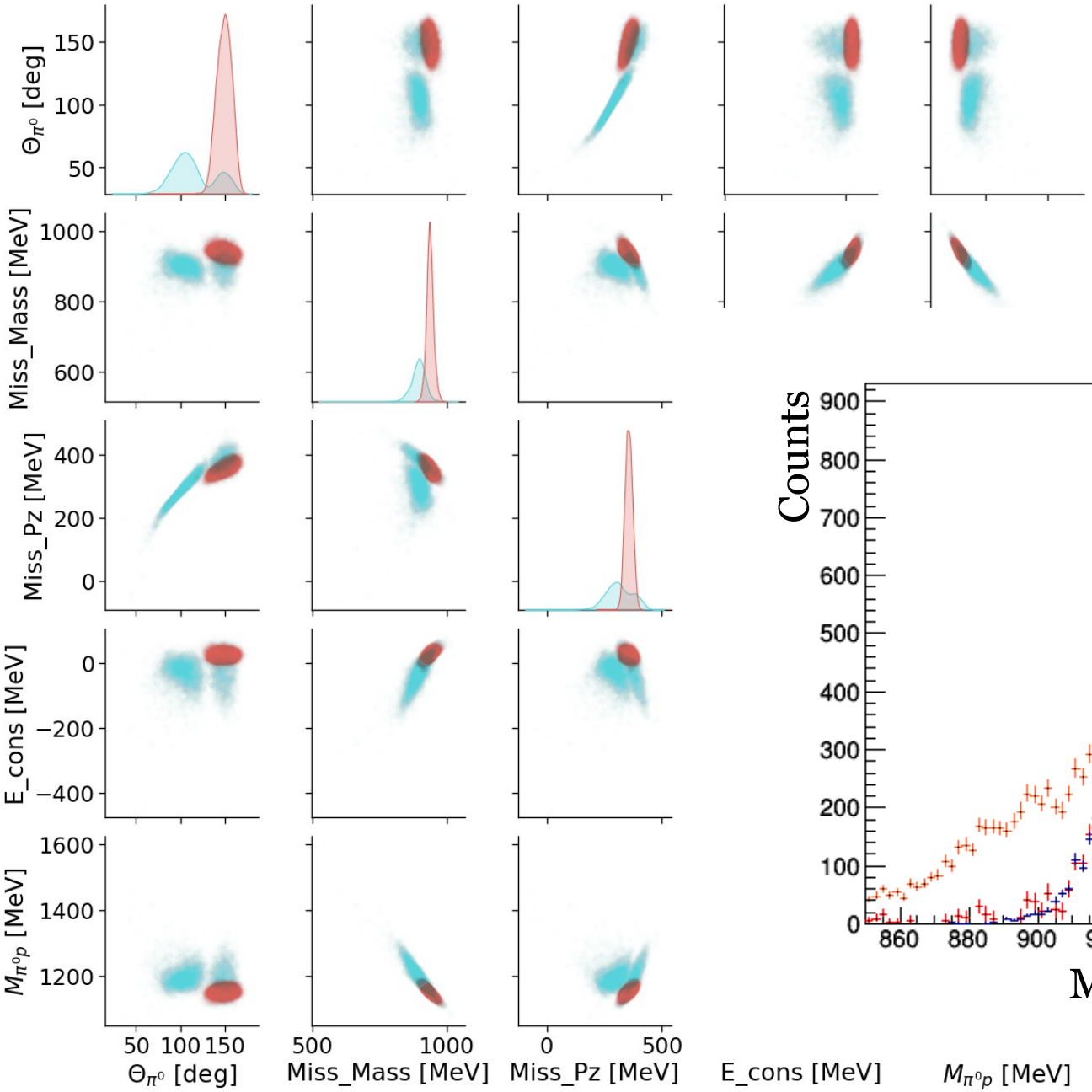

Create a model from clustering
→ Take signal from clustering
→ Combine with pure background sample outside of the prompt peak
→ Create a model (semi-supervised learning)
→ Apply the model on another data set (or part of the same data set)

# From clustering to (semi)supervised learning:



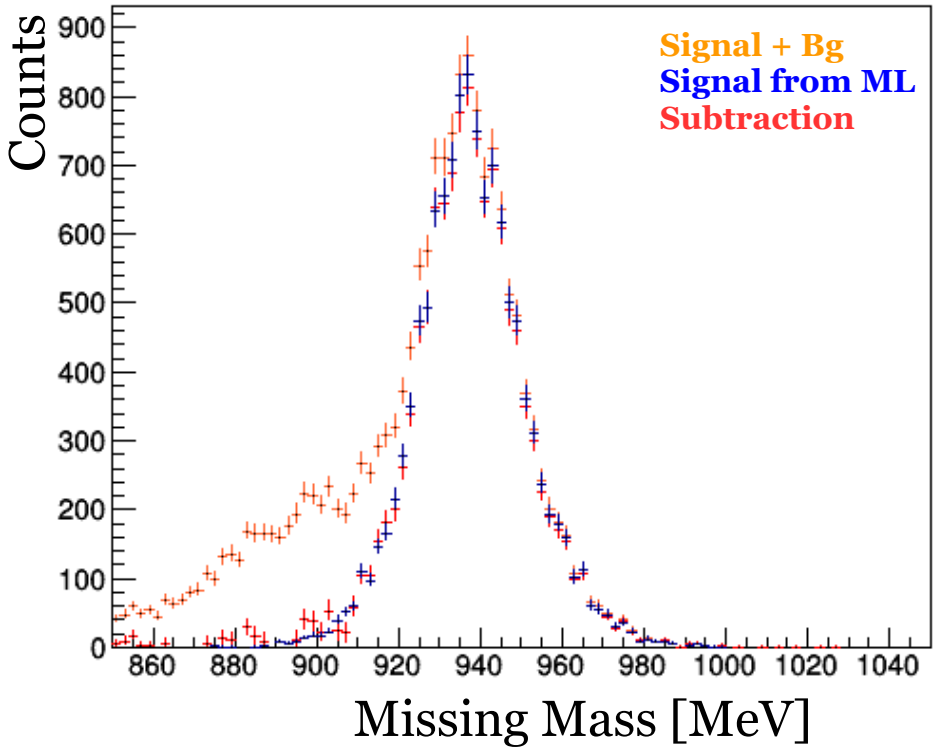

Application of the model made from clustering on another data set

# Semi-supervised learning: Application on another data set

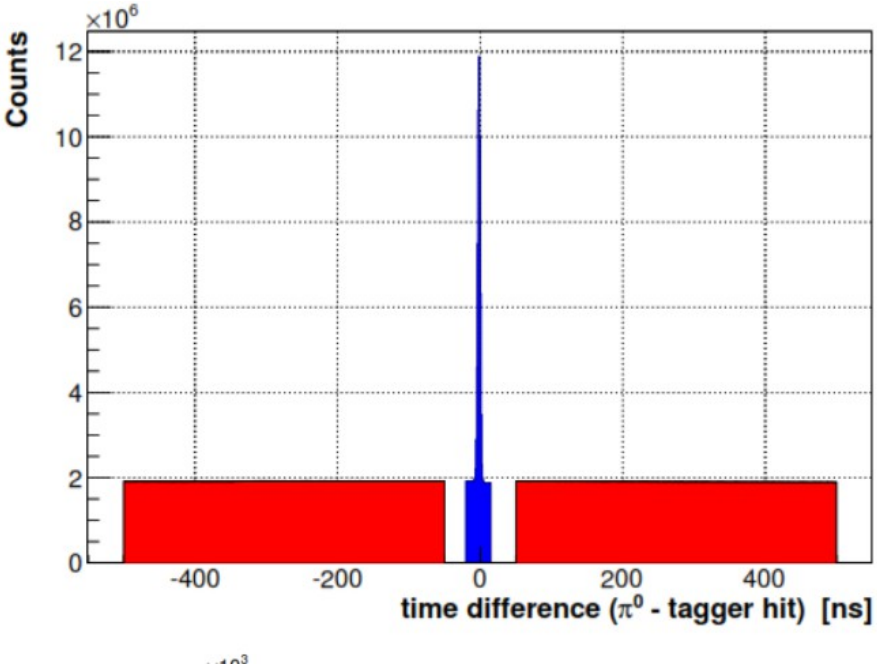


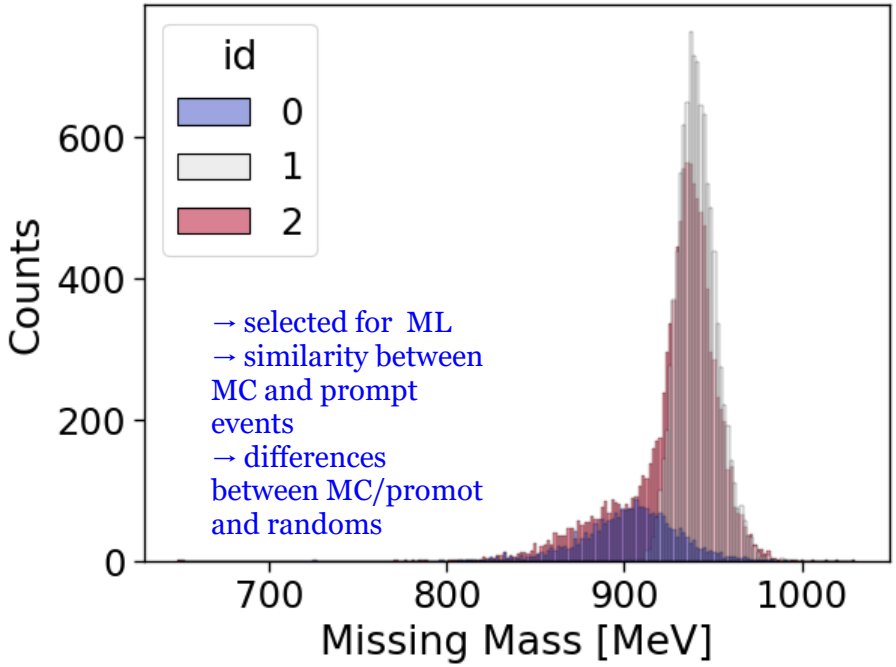

Application of the model made from clustering on another data set

→ Semi- supervised approach
→ (Even) better agreement with the subtraction method

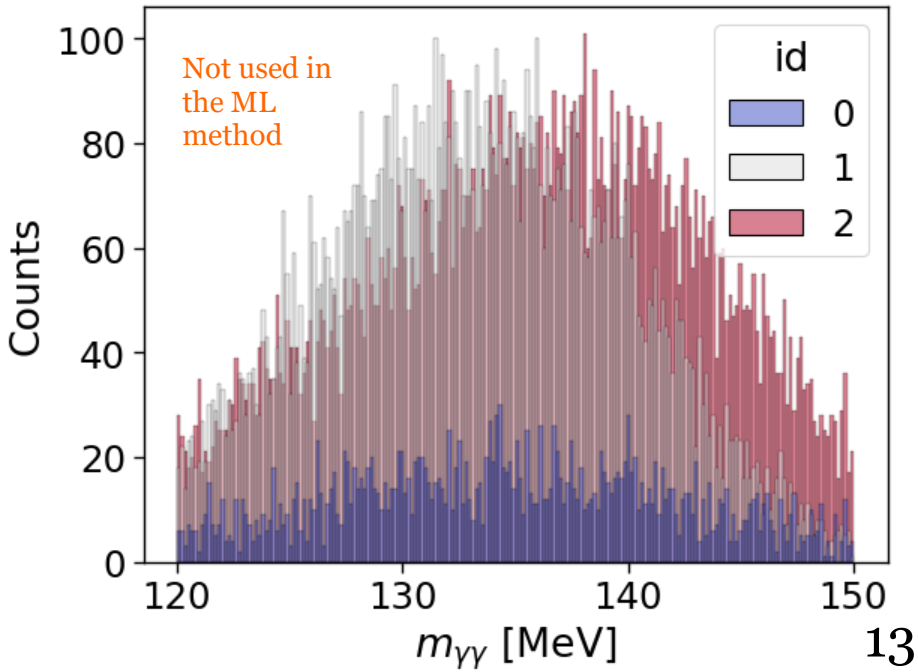


# Semi-supervised approach with simulated (MC) events

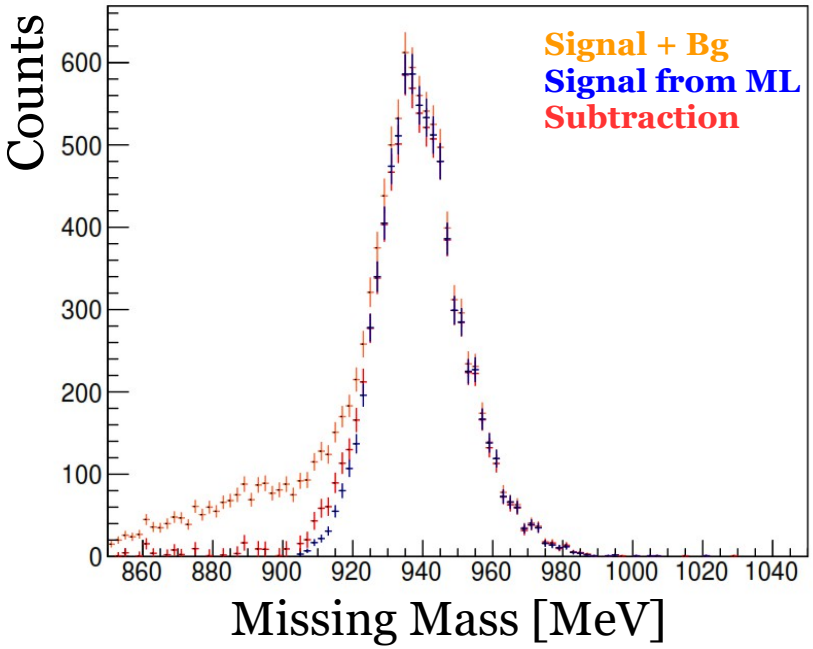


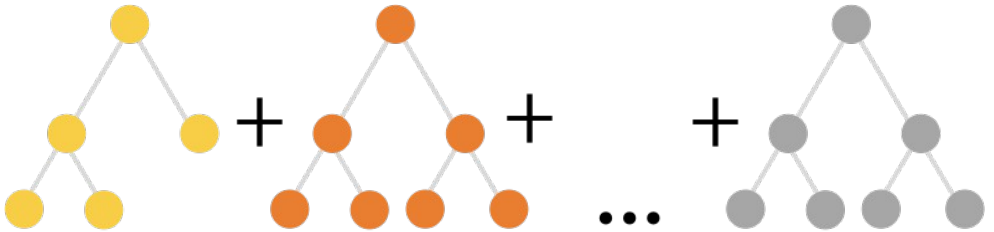




→ Create and train a model based on MC and random background
→ Select variables with agreement of MC and prompt (+bg), but different from pure background sample
→ Apply the model on events in the prompt region

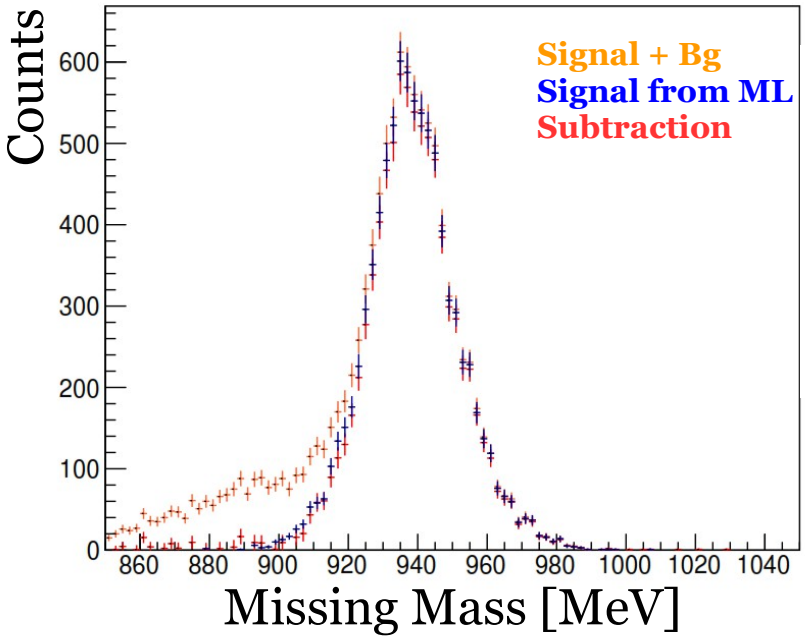


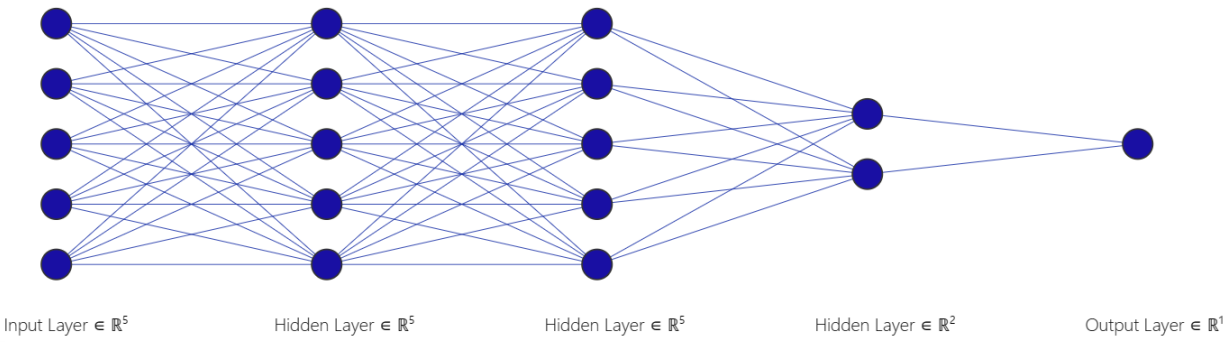
# Semi-supervised approach with simulated (MC) events



CatBoost

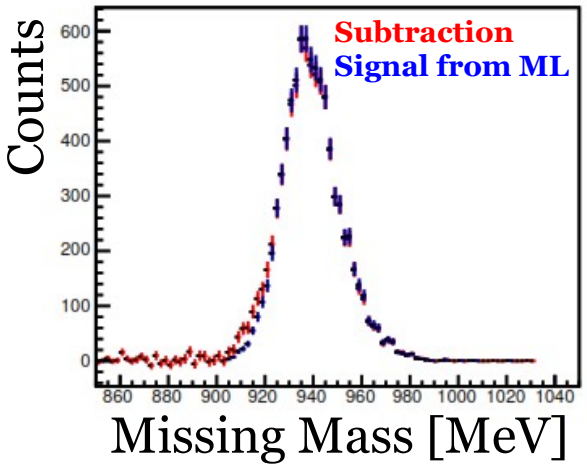


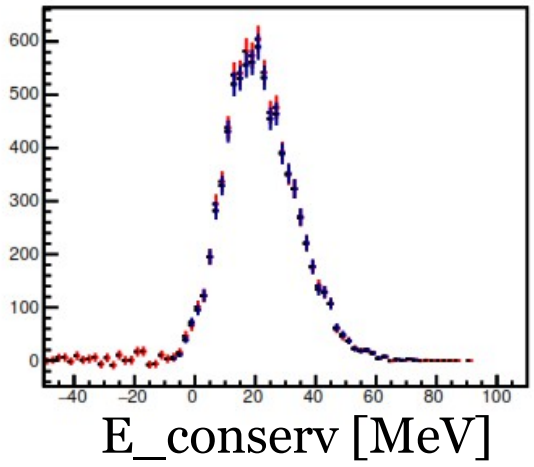



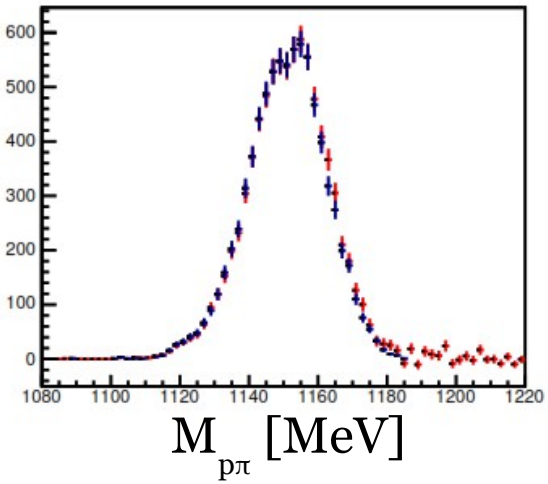

TensorFlow

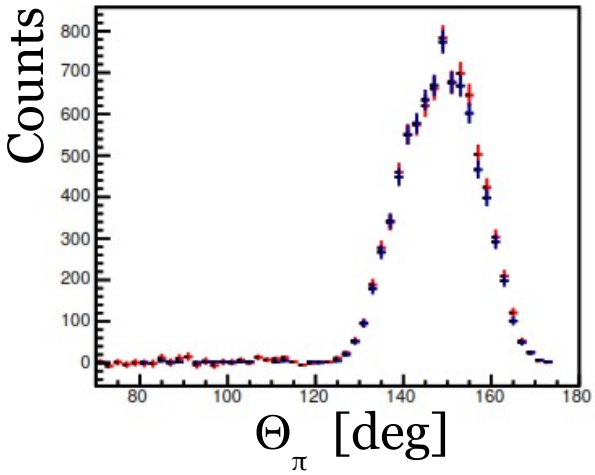
# Machine Learning + MC: Predictions for different variables

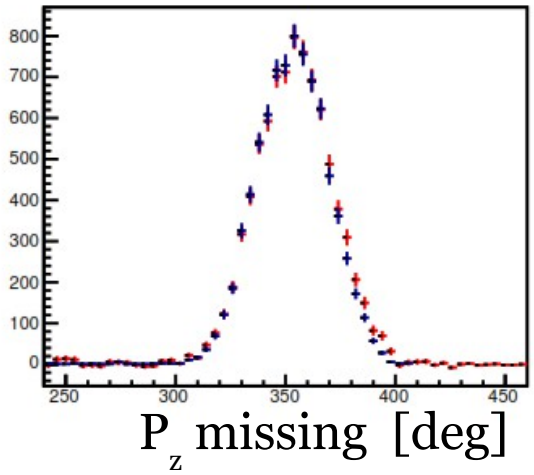
Cross check: Comparison of the MC-based semi-supervised ML approach
With the standard subtraction method!

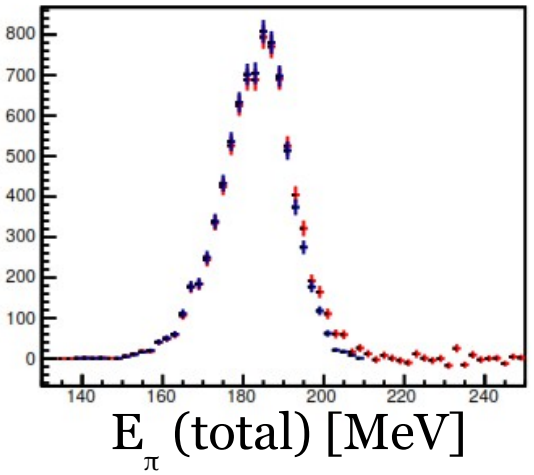


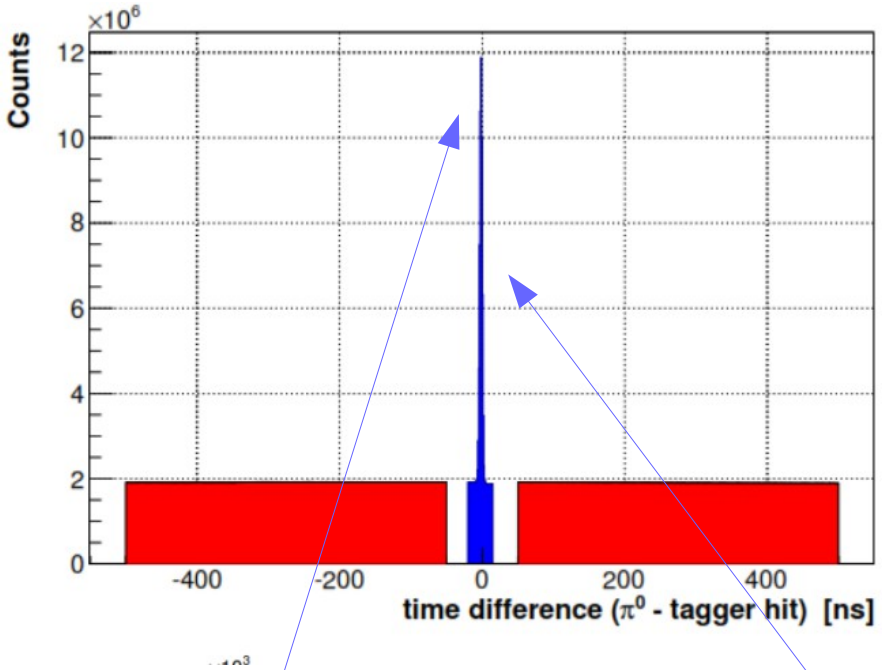



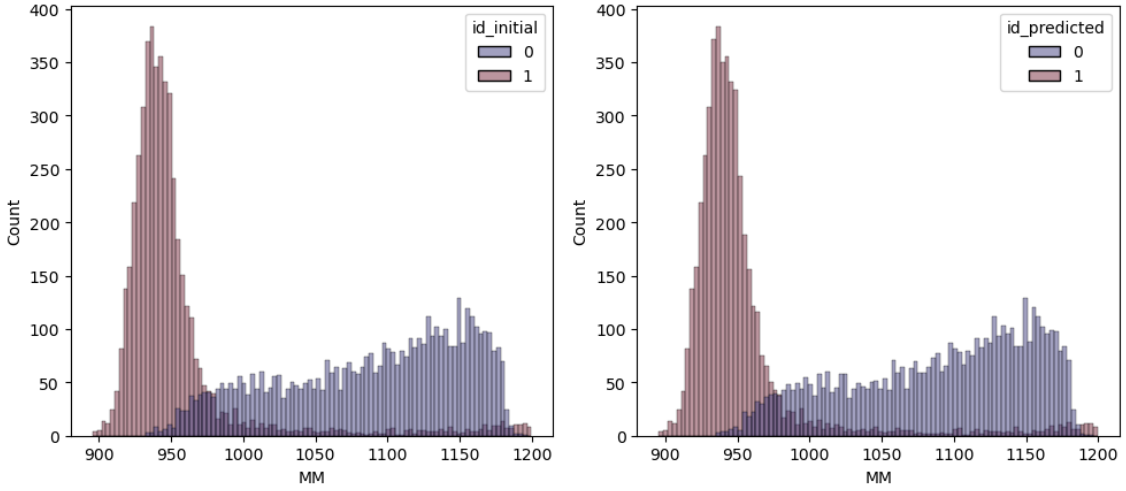

- ML predictions for the variables in good agreement with the standard method (pads 1-5)
- Small differences on the edges are due to difference with MC and can be reduced
- Prediction works with a similar quality for a variable not included as input (pad 6)

# Outlook for a full Compton analysis with Machine Learning

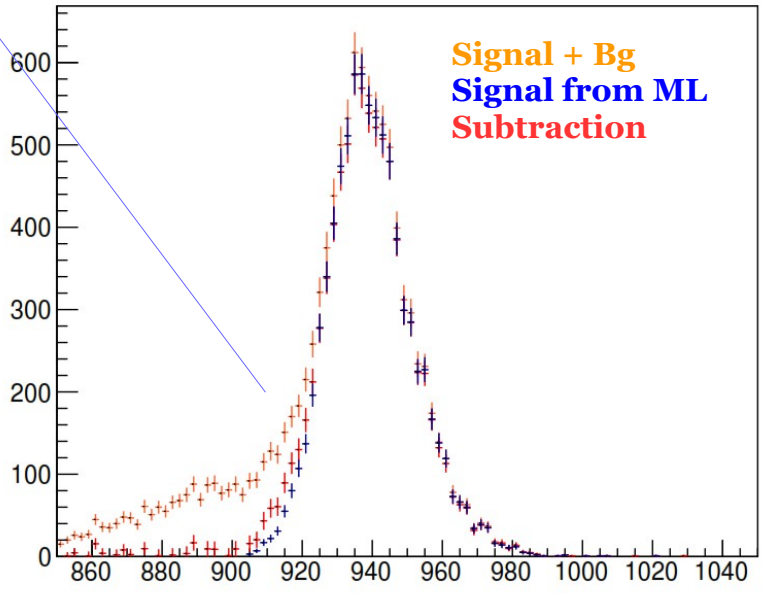



→ Analysis of 3 classes of events (Compton, pion, random)
→ Creating a Machine Learning Model with the 3 classes
→ Combinations of supervised and unsupervised approaches possible

Training MC to separate Compton/pion signals




Apply the ML-based method for random background subtraction

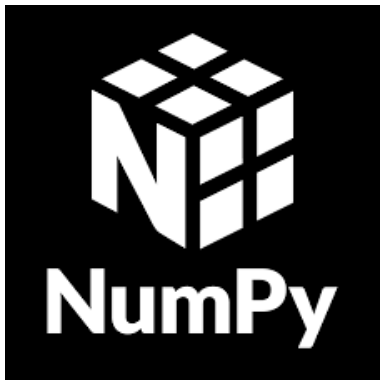


# Software used in this work (along with A2 packages)









# Summary and Outlook

**New Machine Learning-based methods proposed for A2 analyses**

- The separation of Compton scattering and $\pi^0$ events works for simulated events
- New methods for handling random background subtraction have been developed
- Monte Carlo-based method with semi-supervised Machine Learning algorithms leads to stable and accurate results
- Clustering methods is applicable for unlabeled data → requires additional handling of the data in many cases
- Verification with the classical approach is important!

**Outlook:**

- Comprehensive analysis of the Compton scattering data with combination of the developed methods
- Applicable for many of the A2 analyses and other experiments with tagged photons
- Improvement and more ambitious planning for the future experiments (Compton scattering with TPC as an active target, ...)

# Summary and Outlook

**New Machine Learning-based methods proposed for A2 analyses**

- The separation of Compton scattering and $\pi^0$ events works for simulated events
- New methods for handling random background subtraction have been developed
- Monte Carlo-based method with semi-supervised Machine Learning algorithms leads to stable and accurate results
- Clustering methods is applicable for unlabeled data → requires additional handling of the data in many cases
- Verification with the classical approach is important!

**Outlook:**

- Comprehensive analysis of the Compton scattering data with combination of the developed methods
- Applicable for many of the A2 analyses and other experiments with tagged photons
- Improvement and more ambitious planning for the future experiments (Compton scattering with TPC as an active target, ...)

## Thank you for your attention!

# Semi-supervised learning: Application on another data set