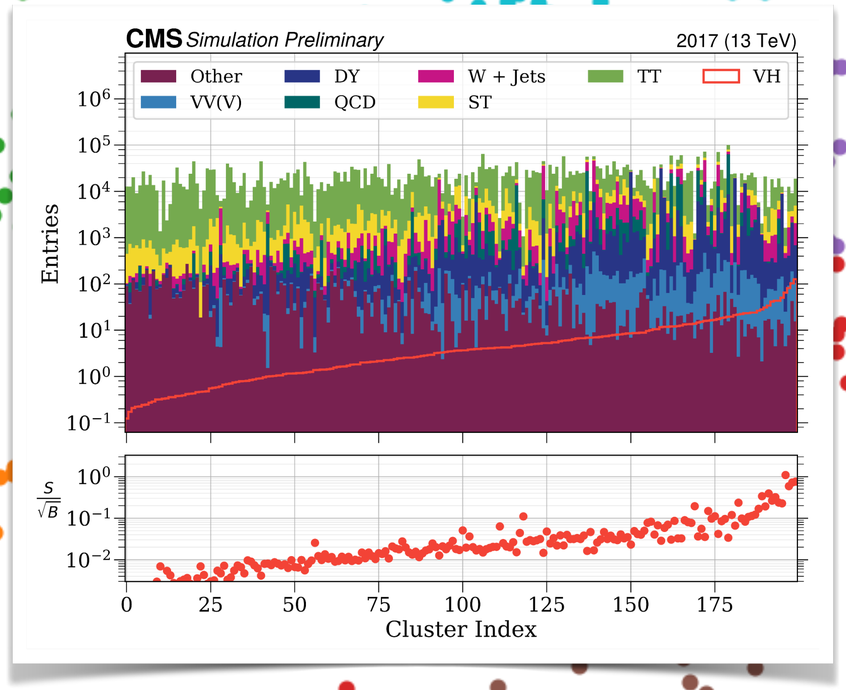
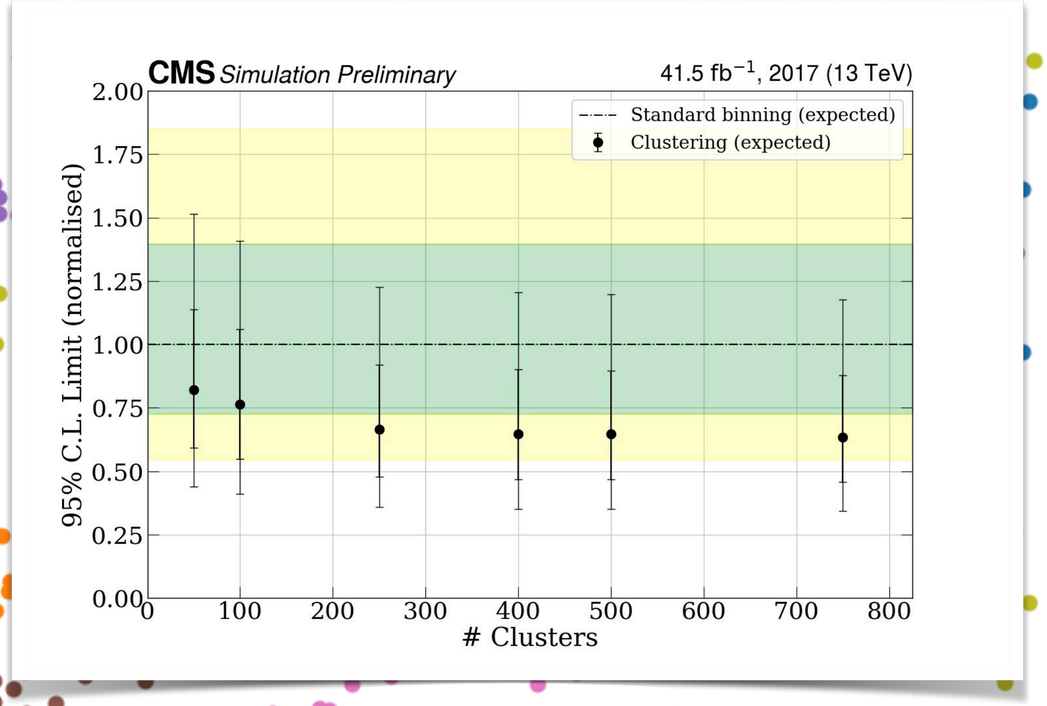




Binning high-dimensional classifier output for HEP analyses through a clustering algorithm

Svenja Diekmann, Niclas Eich, Martin Erdmann

CHEP 2023

9th May 2023

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

RWTH AACHEN UNIVERSITY

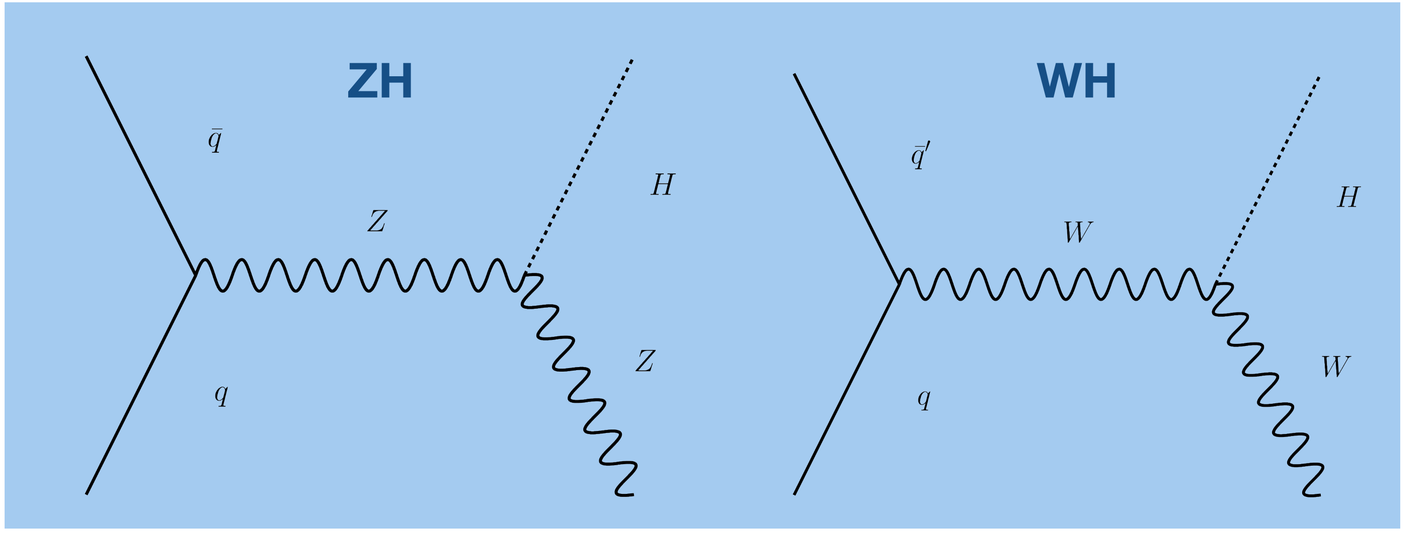
Introduction

Analysis Context:

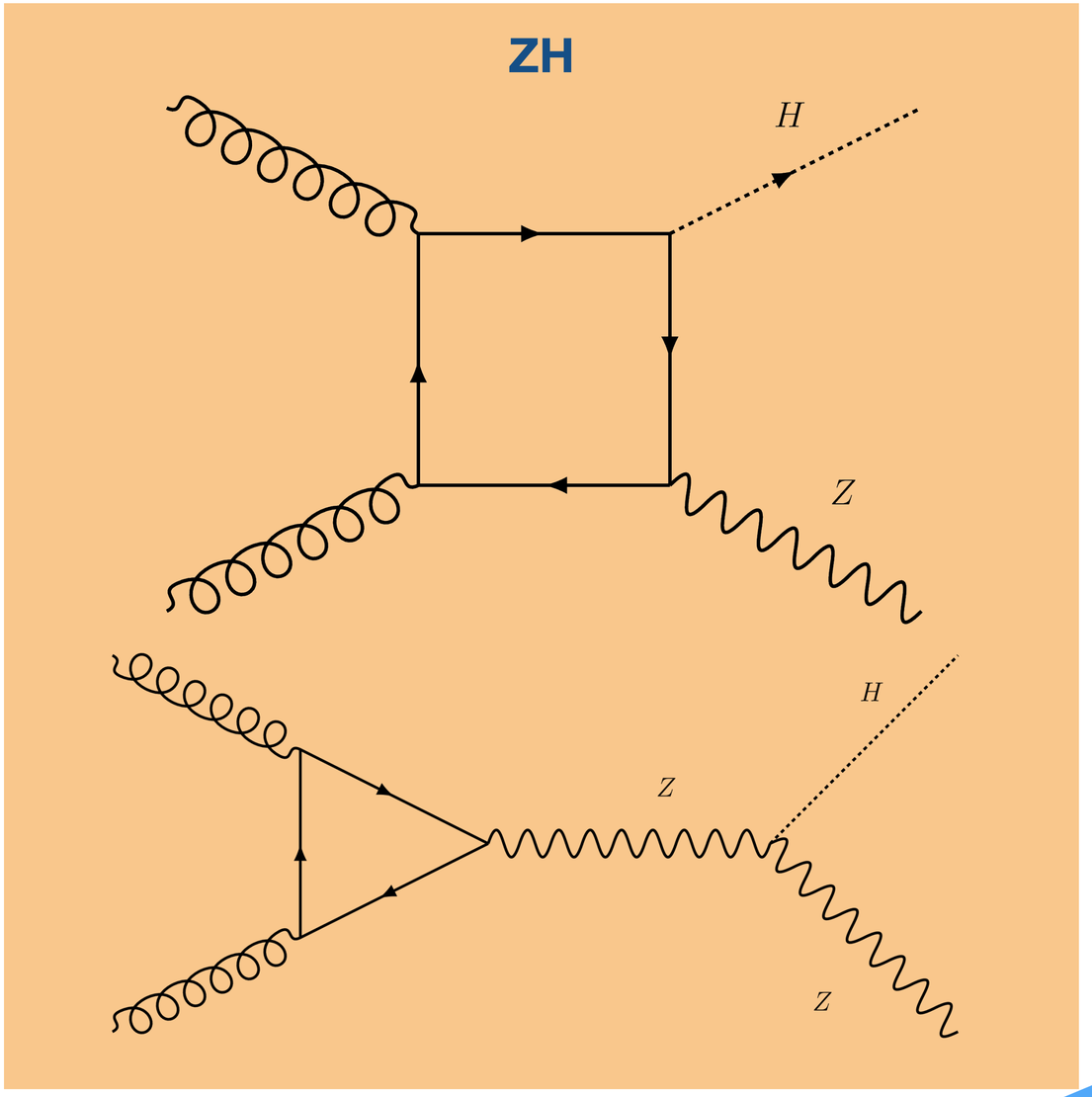
- search for VH at
- specialised analysis for gg →ZH production
- yet unmeasured production
- small cross section:
  - need for sensitivity enhancement

$$VH = (\mathbf{gg \rightarrow ZH}) + (qq \rightarrow ZH) + WH$$

Dominant Production

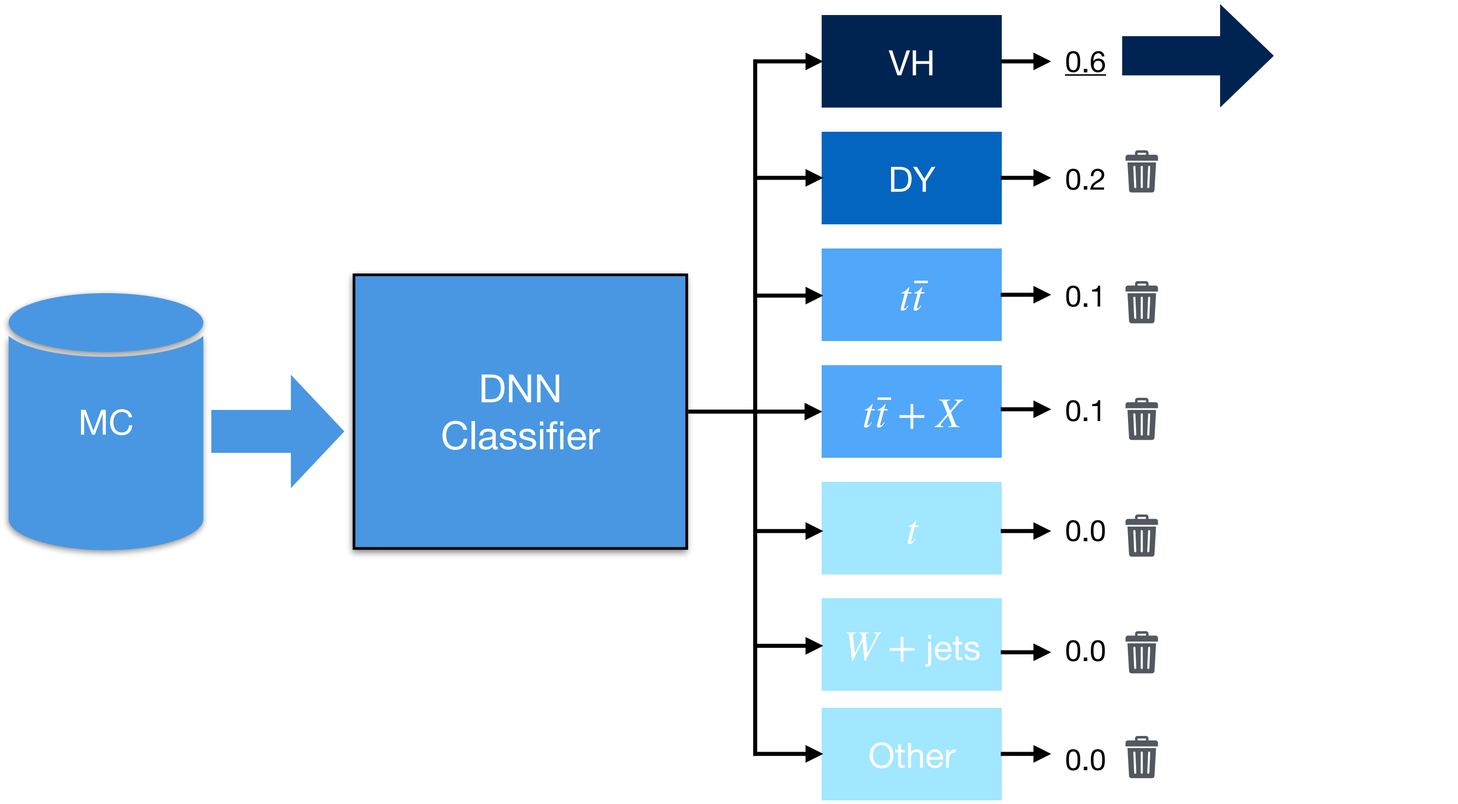


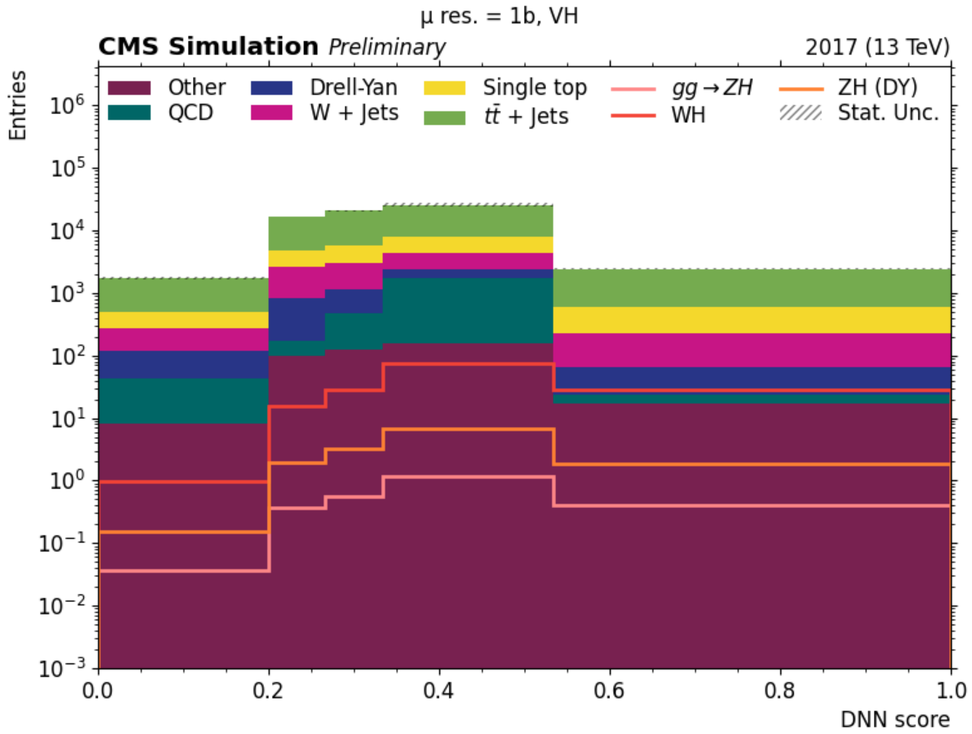
Gluon Fusion



VH-Analysis Process-classification

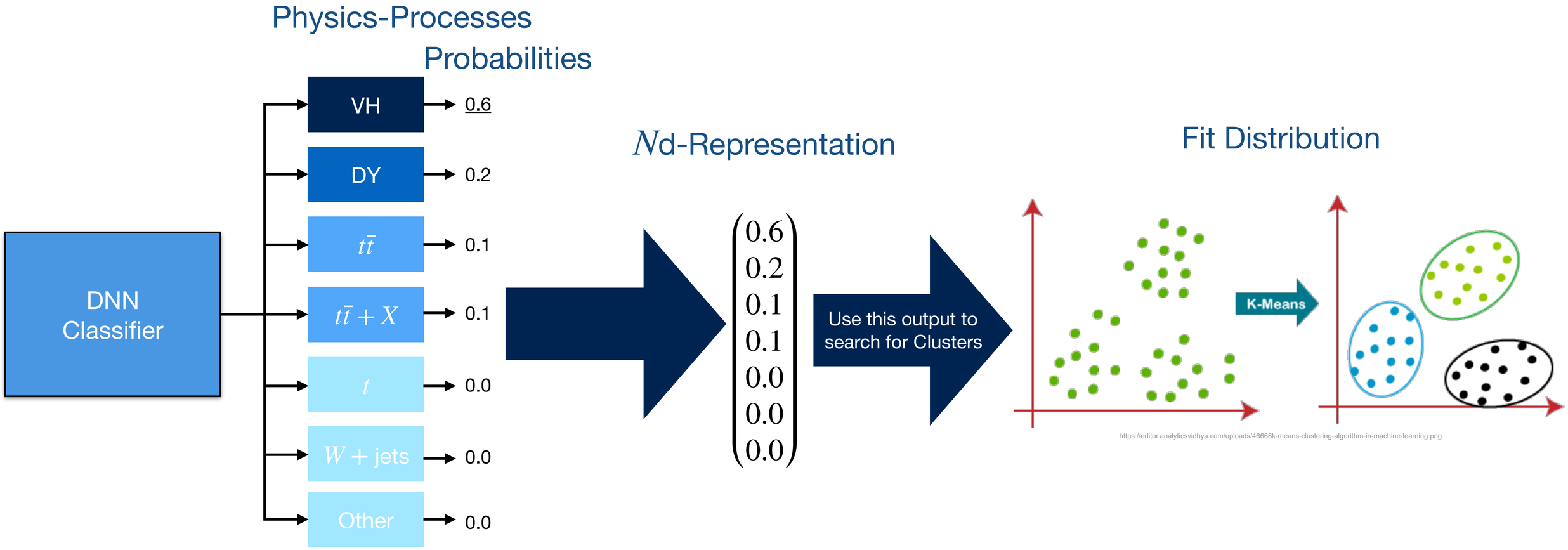
- Events are classified by process
- **Highest score** serves as fit-variable
- 1 signal region 7 control regions





Clustering with K-Means [1/2]

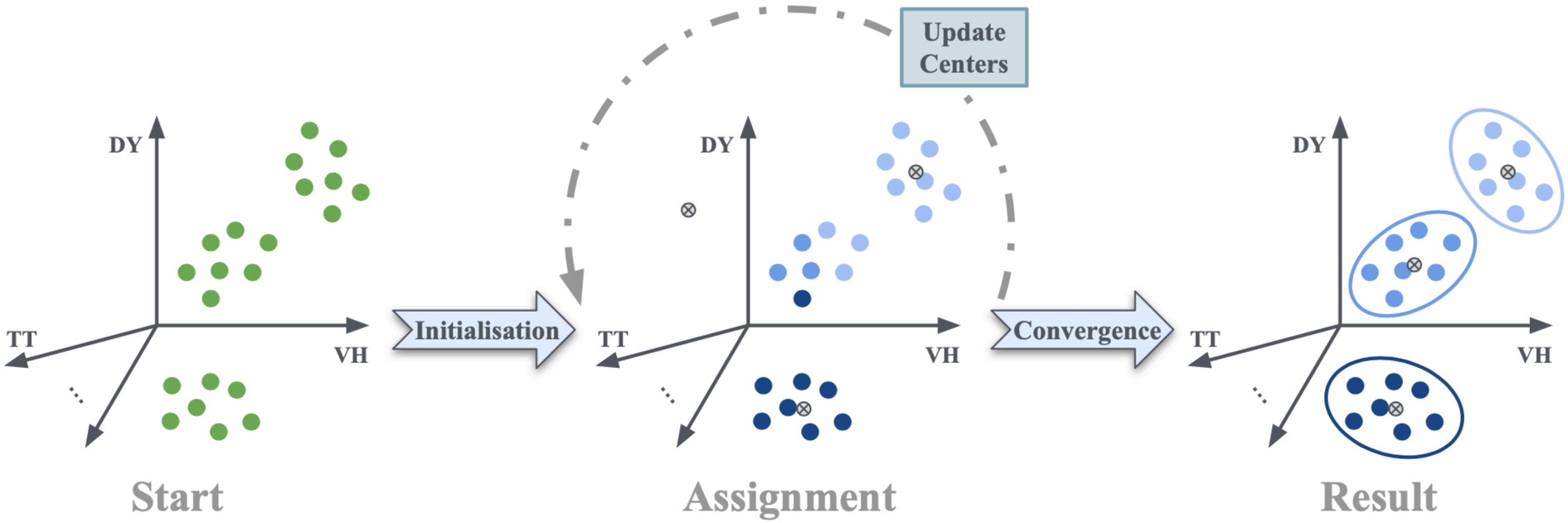
- The DNN delivers a N**d-representation** of our data
- Correlated to **physics processes**, which will be fitted
- Clustering can create bins in this Nd space without suffering under the *curse of dimensionality*



Clustering with K-Means [2/2]

K-Means Clustering Algorithm:

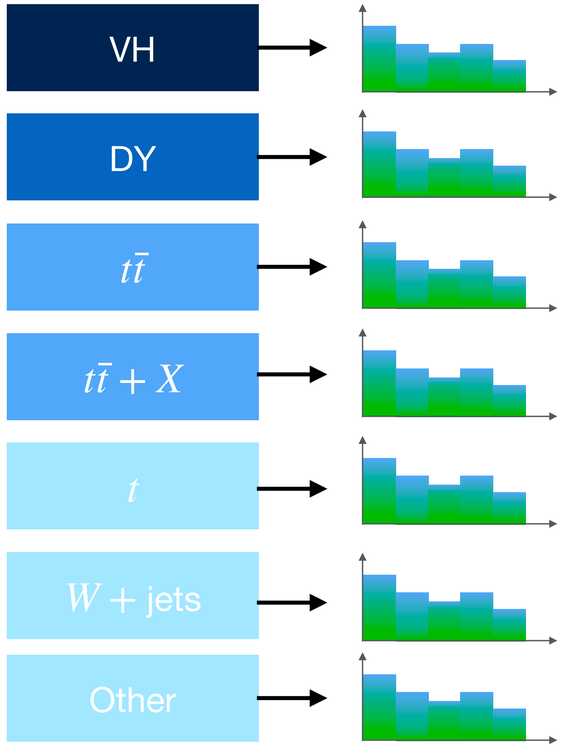
- random initialisation of K cluster centers
- eventwise assignment to closest cluster (L2 norm)
- iterative update of cluster centers by mean of assigned events
- K-Means already used in this analysis [1]



Approach Comparison

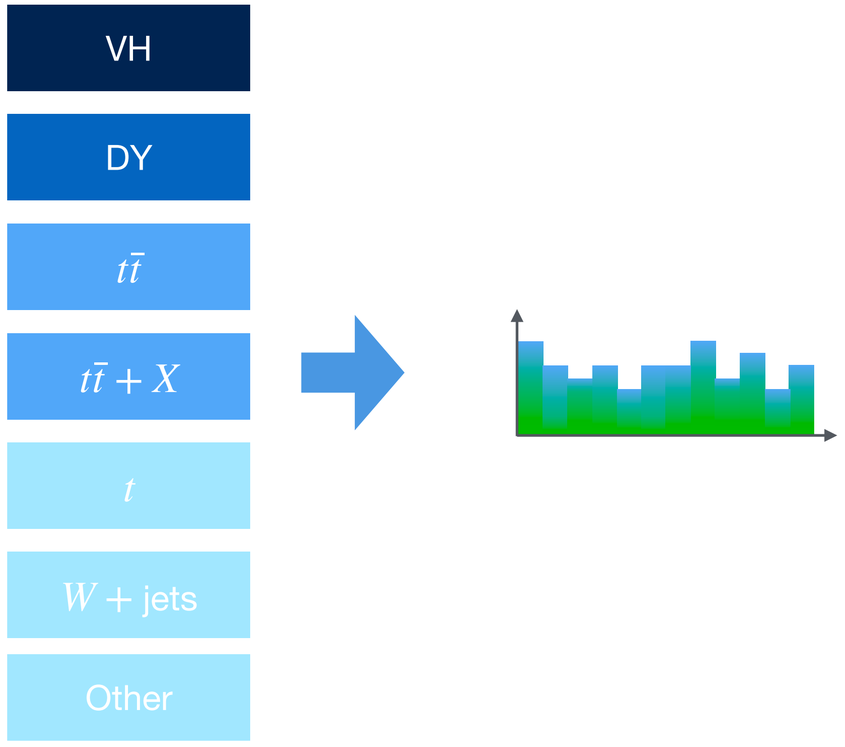
Standard Binning

- N Histograms
- Each has m bins
- Need to be tuned independently



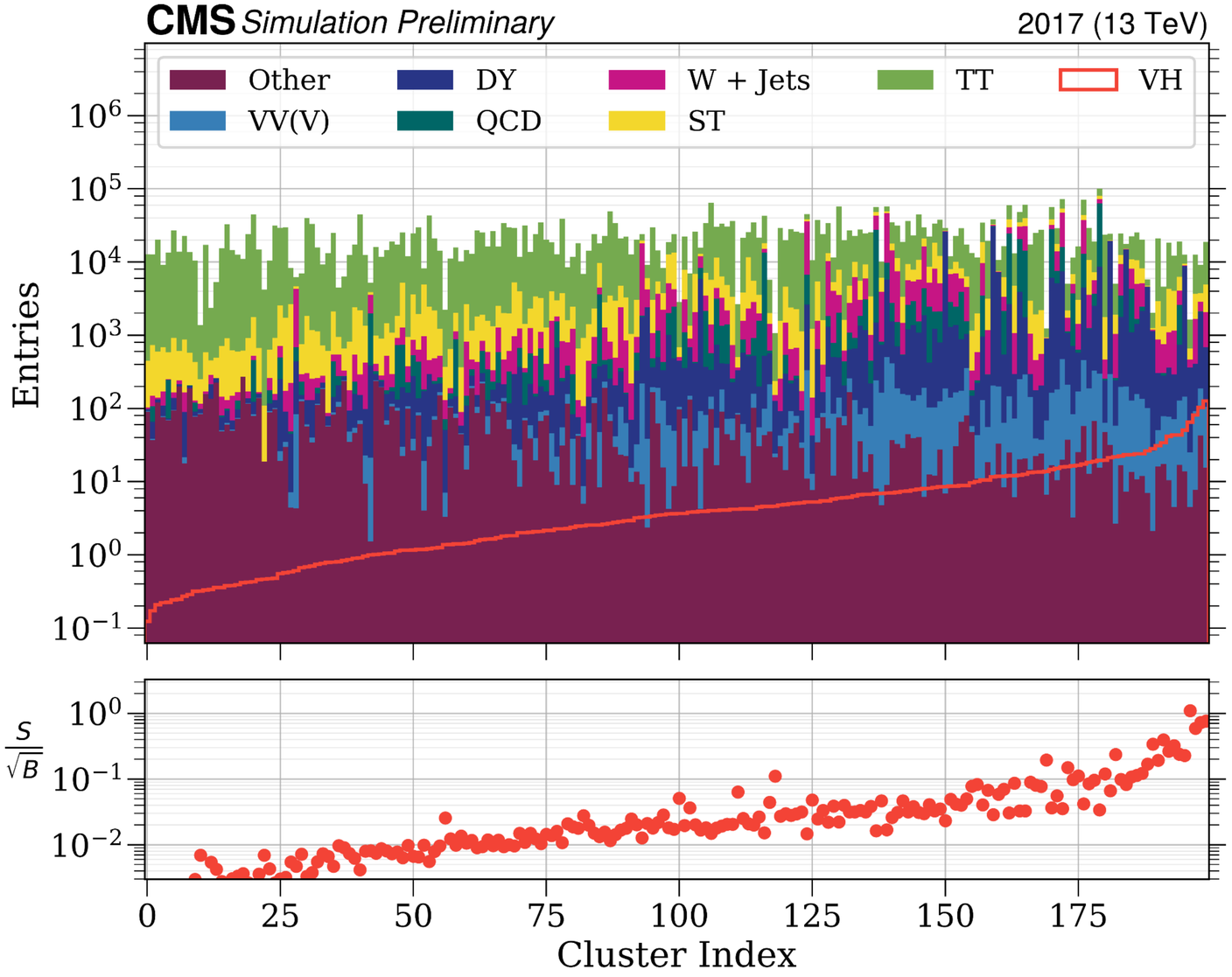
Clustering Binning

- 1 Histogram
- K bins
- Number of tuning parameters very limited



Clustered distribution

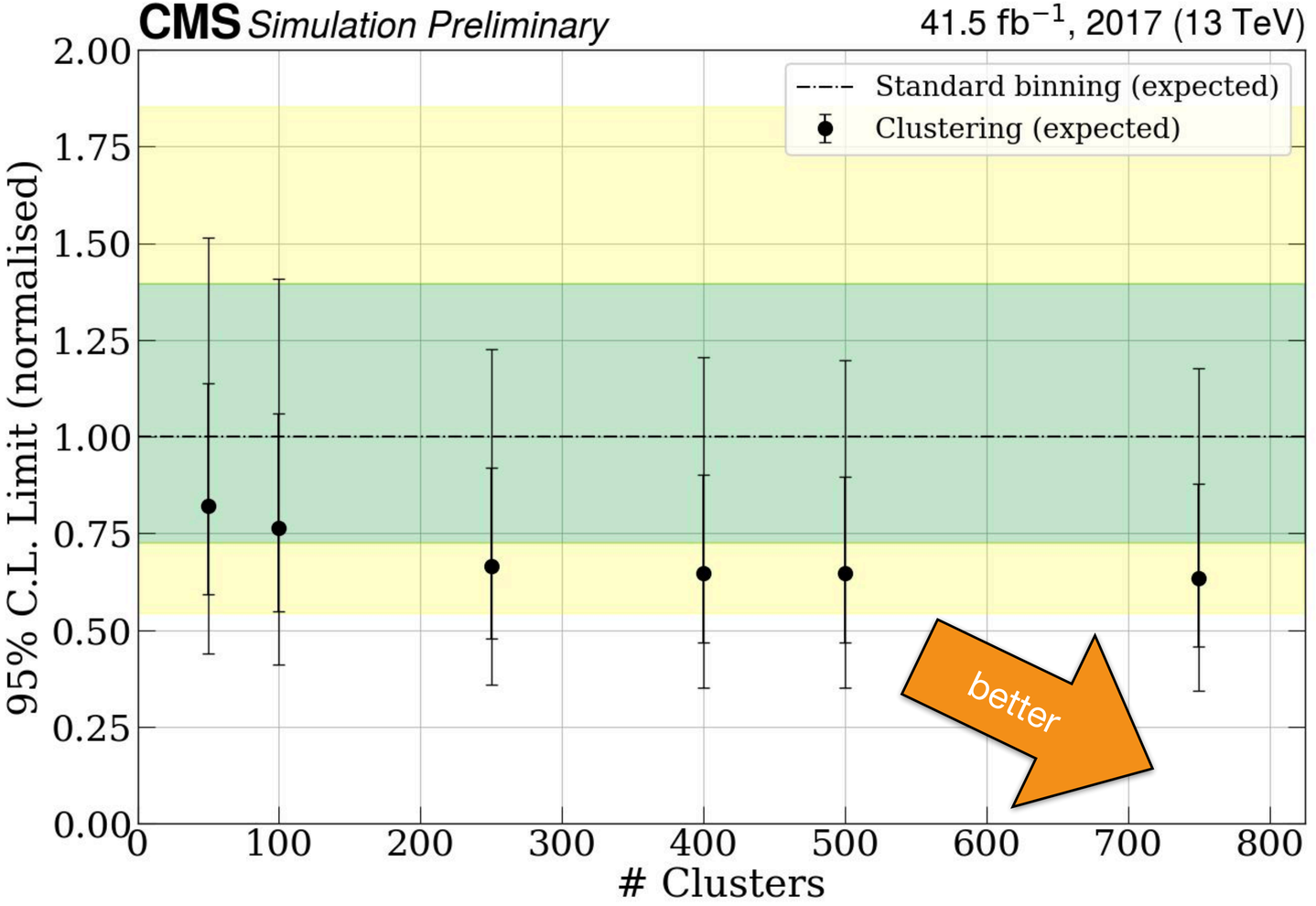
- Clusters sorted by absolute signal yield
- All bins contain a reasonable amount of events
- Some clusters are enriched in specific backgrounds
- High purity bins towards right side



Sensitivity Improvement

Benchmark results:

- evaluation of 95% C.L. limit
  - sensitivity increase for high
  - number of clusters
  - improvement compared to
    - the standard method (~ 400 bins):
    - statistical uncertainties
    - background normalisation



Bias test - MC dependency

- Clusters are determined on MC events

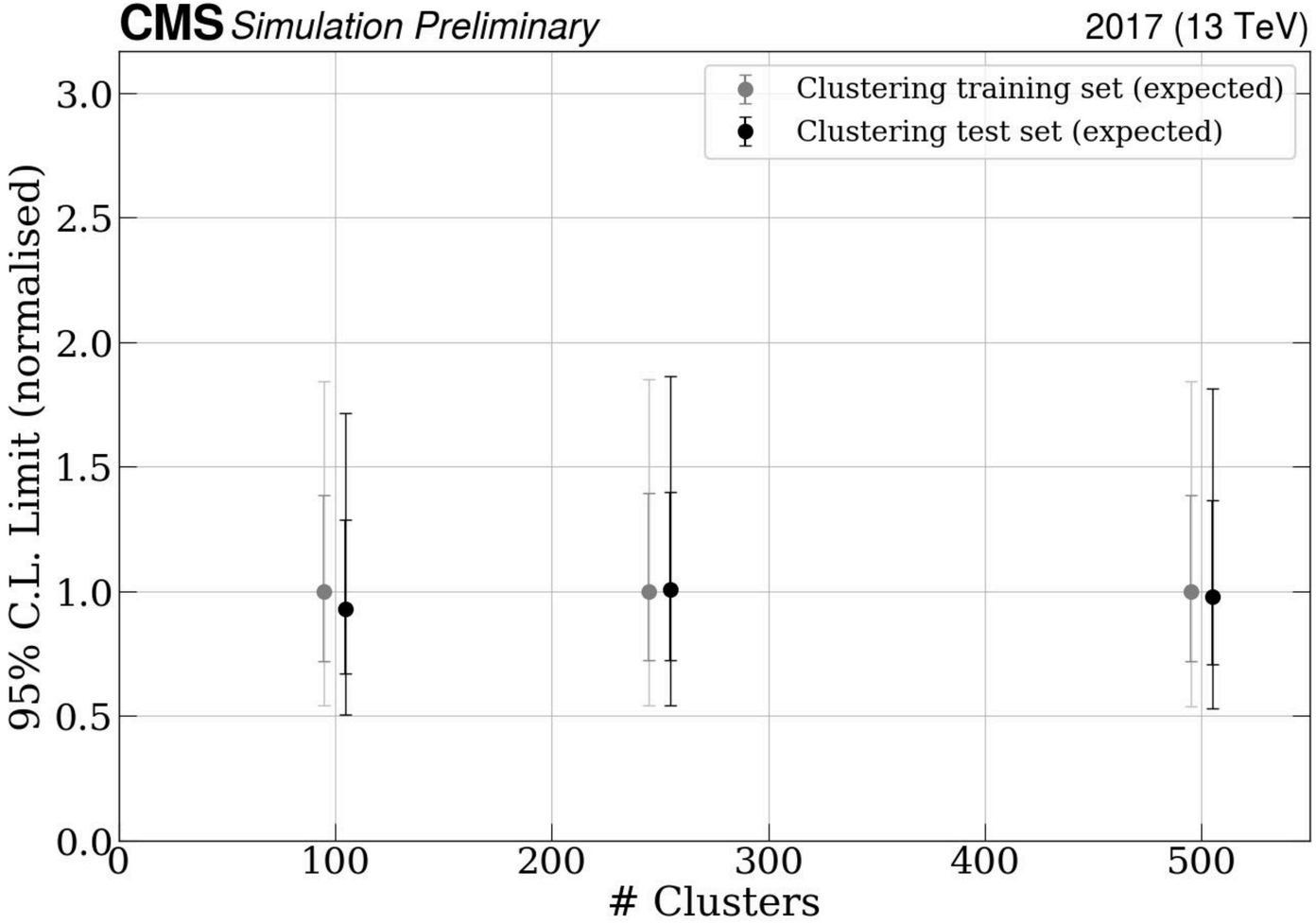
  ➜ Might introduce bias!

Testing-Strategy:

- Dataset split into two equal sets
  - Training
  - Test
- Clustering performed on training set
- Evaluated on sets separately

➜ Almost no difference

➜ Clustering is **robust** towards MC dataset introduced bias

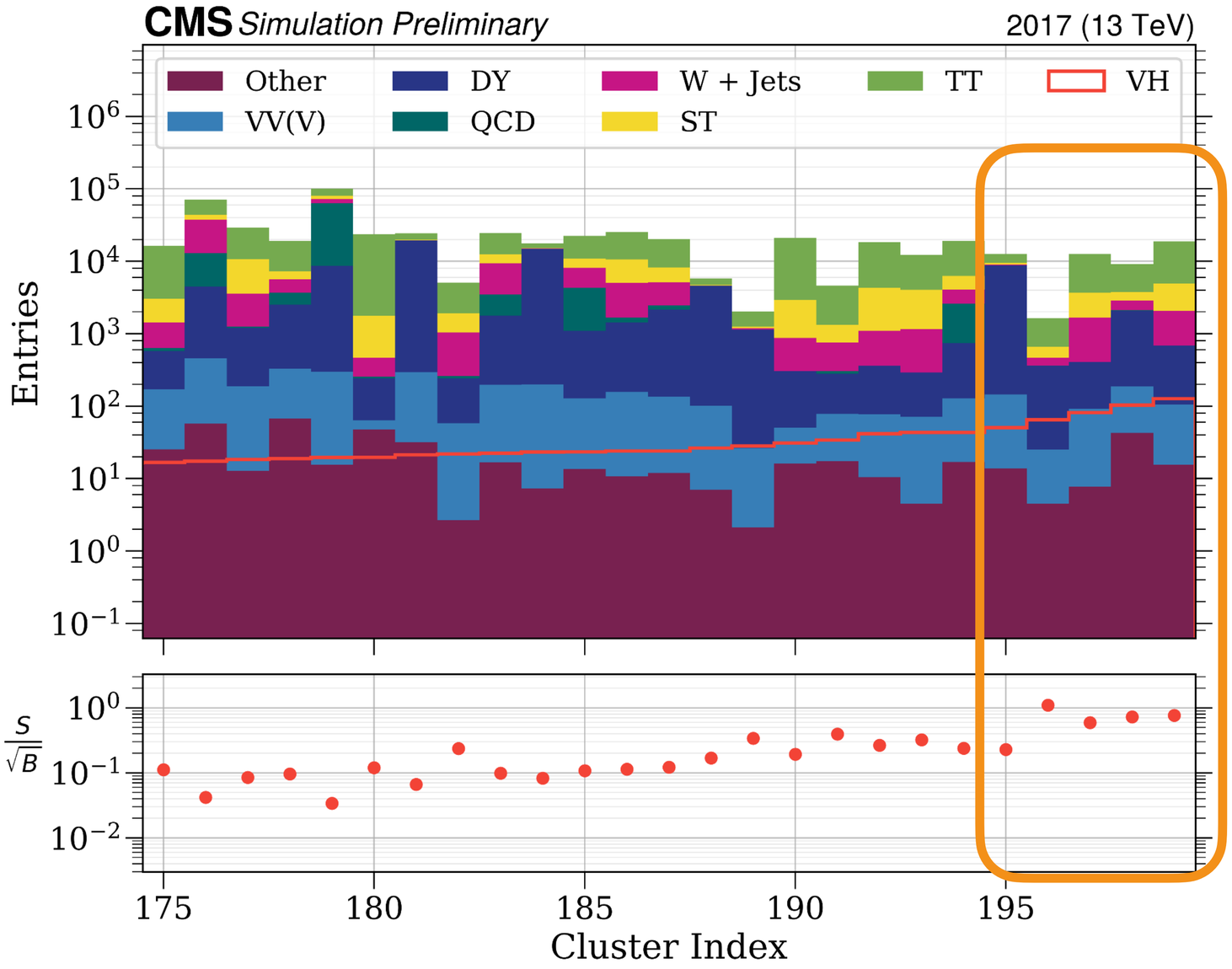


Top 25 clusters

- Clusters are determined in 8d-space

➜ Cluster center coordinate is a 8d vector

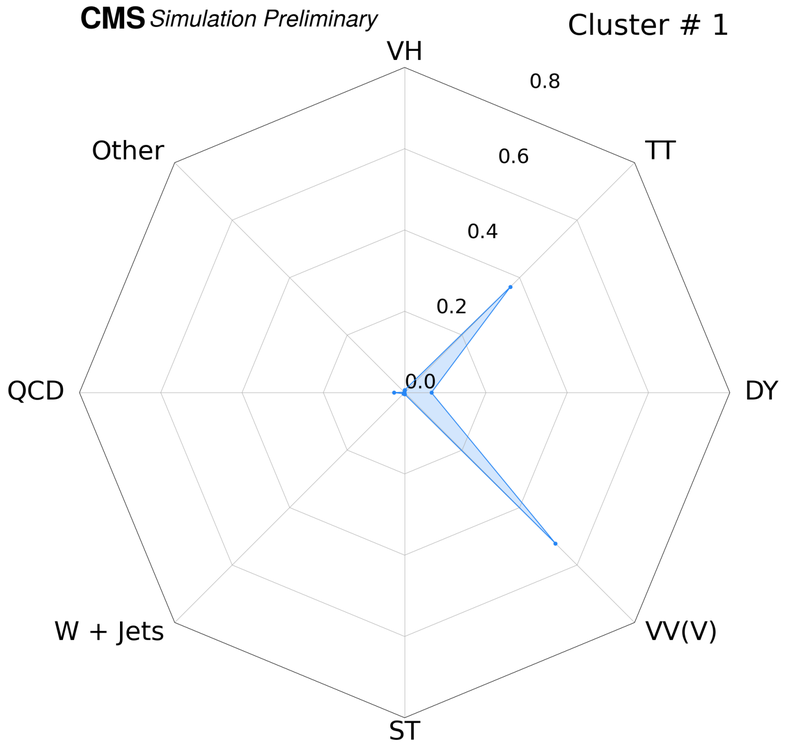
Try to get some insight how the clusters are constructed!

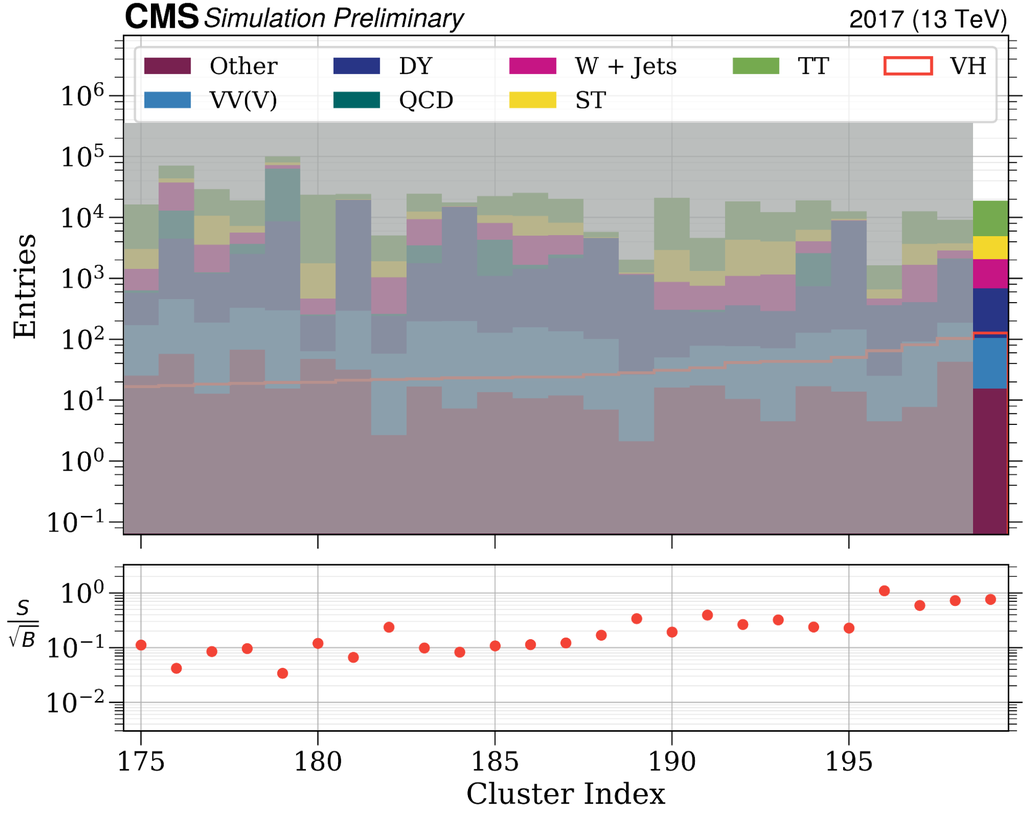


Cluster Visualisation - Top 1 Cluster

Visualisation: 8d cluster center coordinate plotted as Radar-Plot

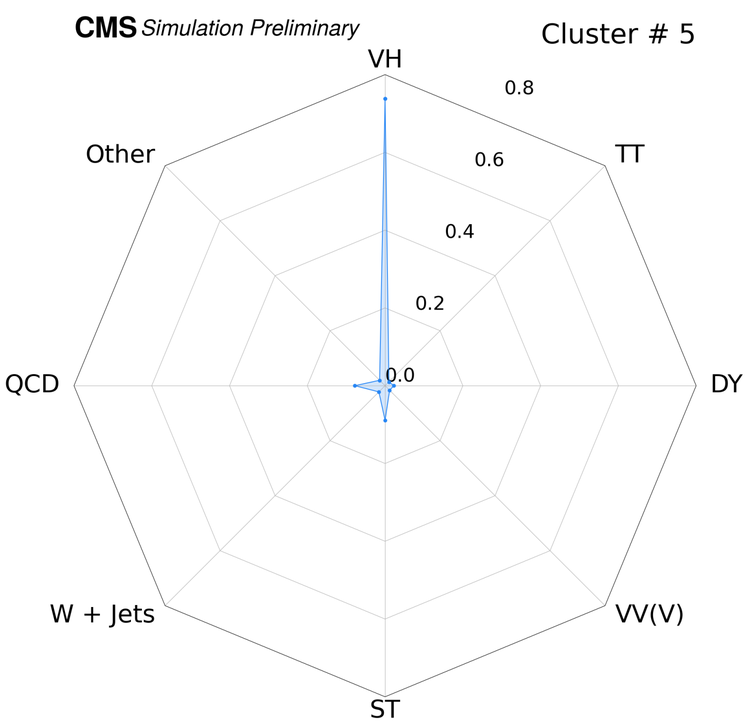
- #1 cluster has the largest signal contribution
- Main coordinate consist of backgrounds
  - $t\bar{t}$
  - VV(V)

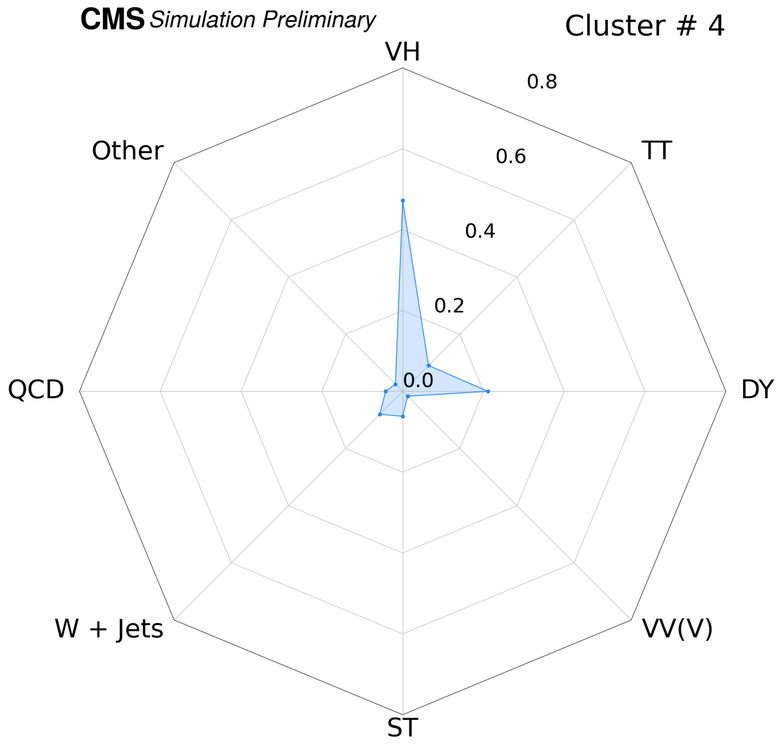


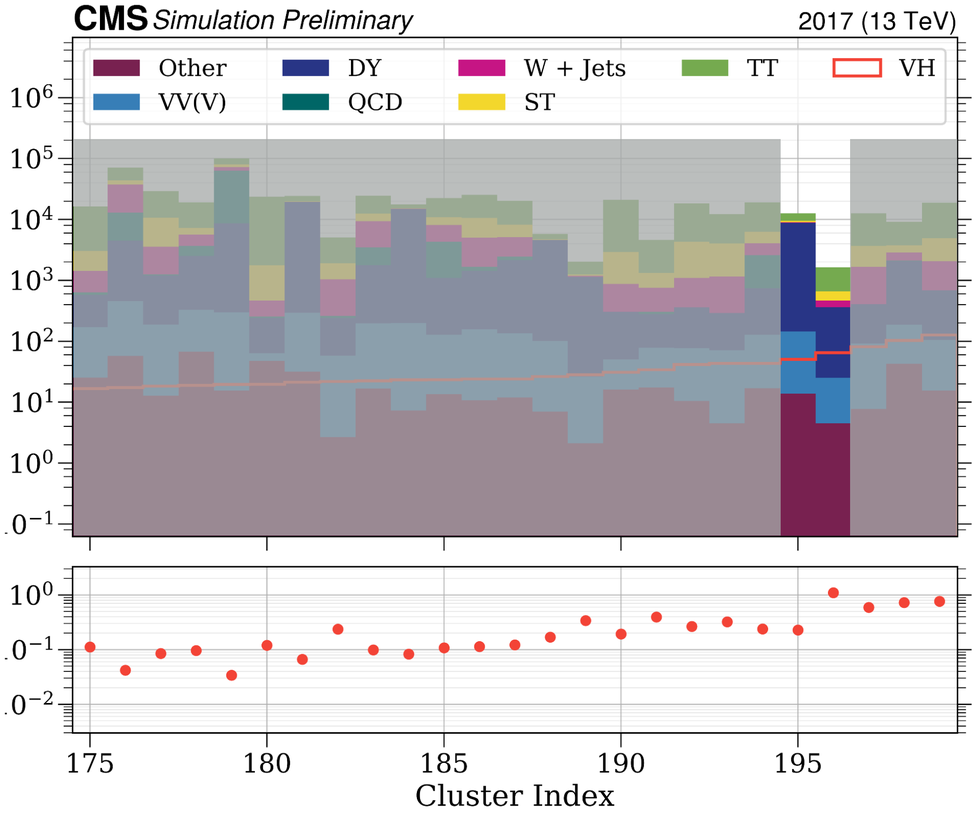


Cluster Visualisation - Top 4/5 Cluster

- High signal prediction confidence
- Better signal-background ratio



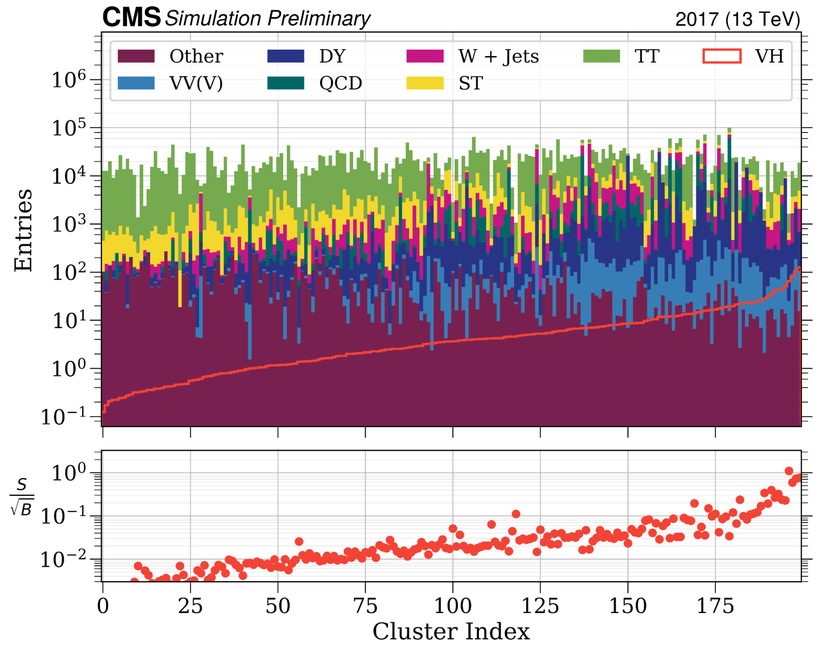


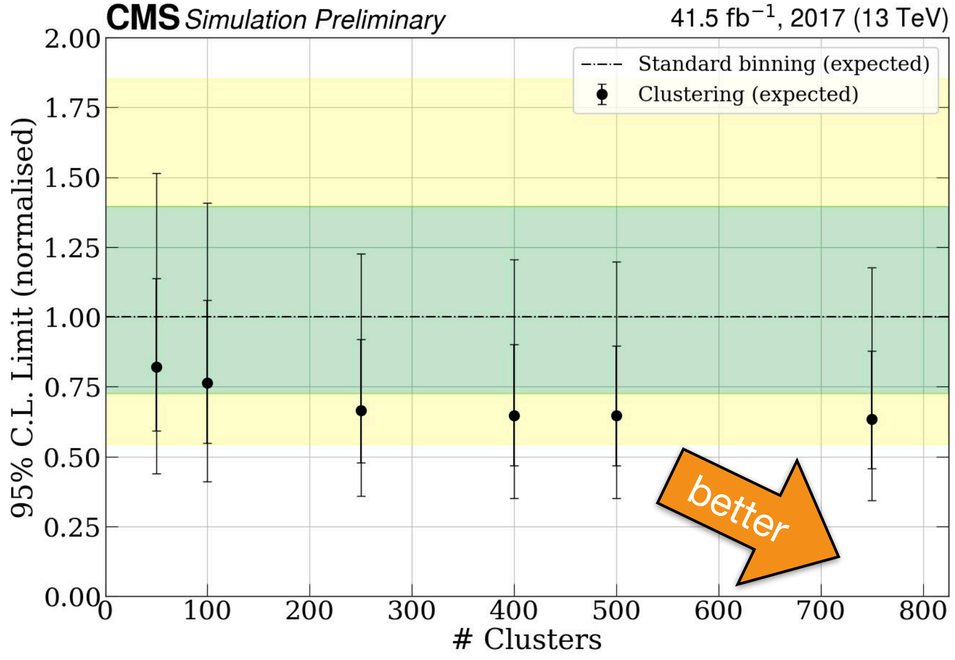


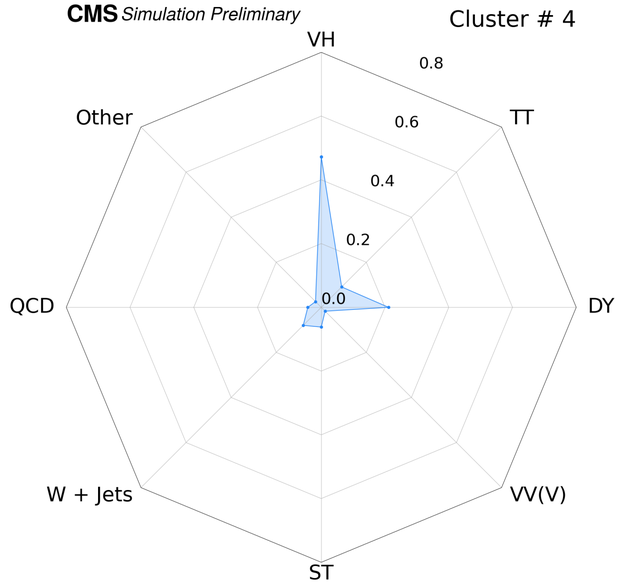
Summary

Binning High-Dimensional Classifier Output through Clustering:

- **Agnostic** binning algorithm in high-dimensional space
- No loss of information
- **Improvement of analysis sensitivity**
- Radar plots visualise cluster center coordinates
- Robust towards MC training dataset







Resources

[1] Evidence for associated production of a Higgs boson with a top quark pair in final states with electrons, muons, and hadronically decaying τ leptons at √s= 13 TeV, CMS Collaboration, JHEP 08 (2018) 066,