FYS9555 Final project

«Higgs boson production mechanism classification from ATLAS Open Data»

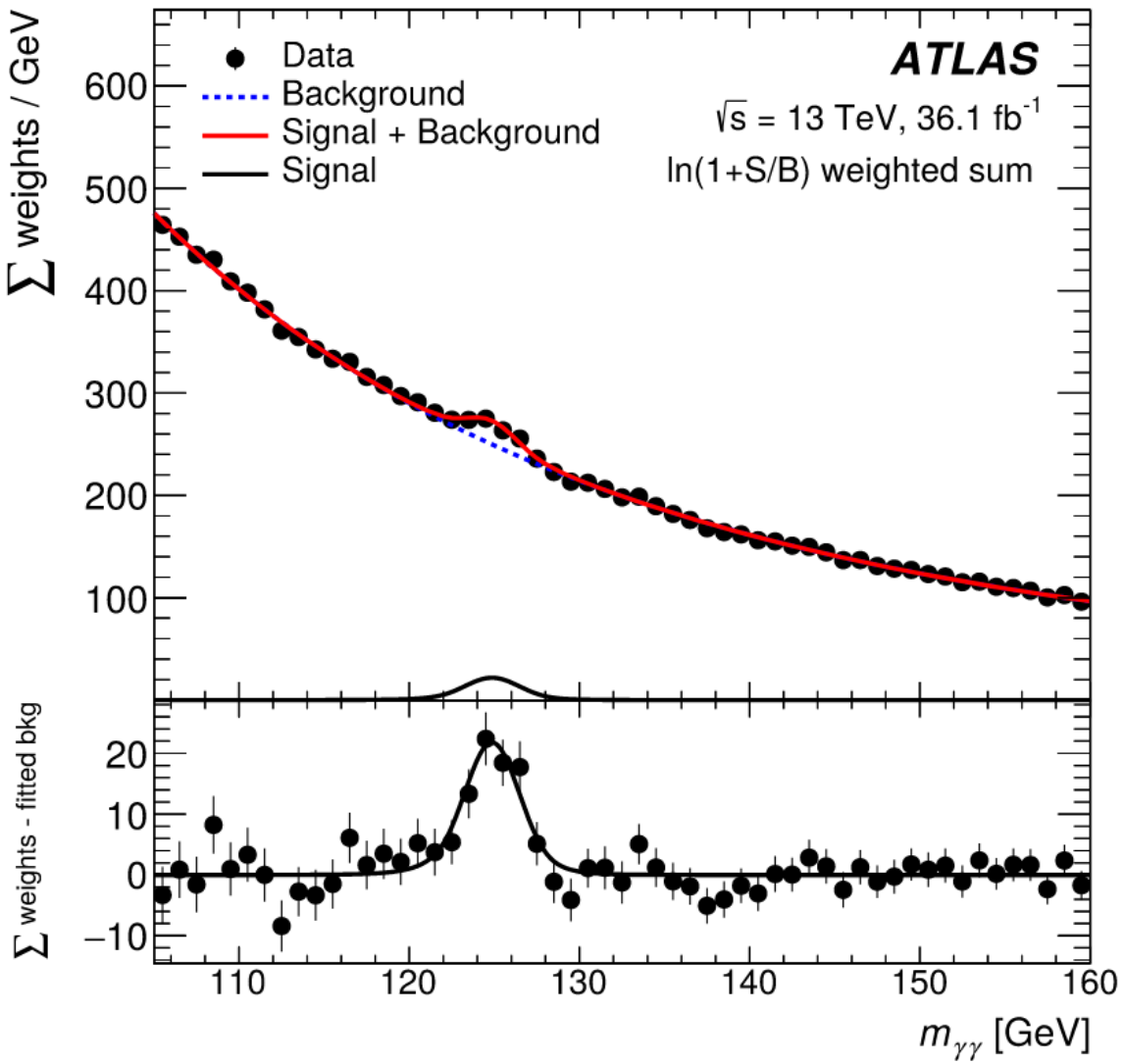
10 June 2020

Victor Ananyev

Outline

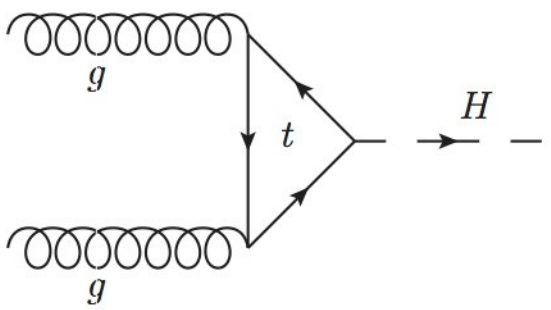
- Di-photon process
- Data acquisition
- What's inside?
- Dealing with jagged data in ML
- Fighting imbalanced data sets
- Fitting baseline ML models

Di-photon process

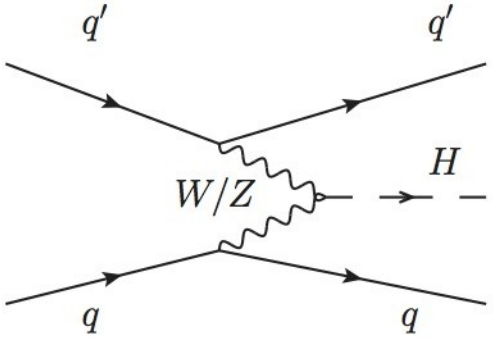


(borrowed from https://arxiv.org/abs/1806.00242)

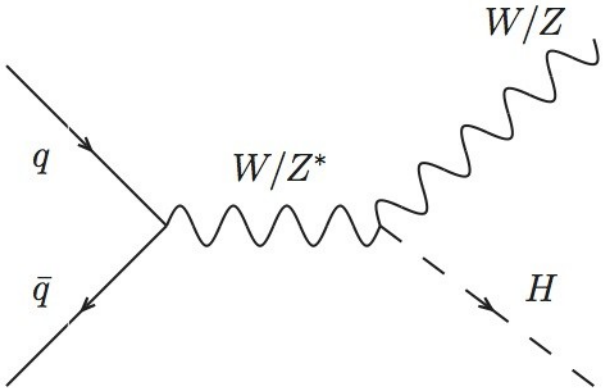
Major production mechanisms



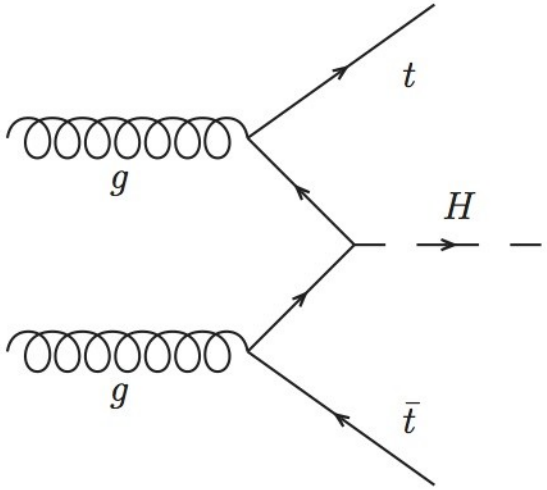
a)



b)



c)



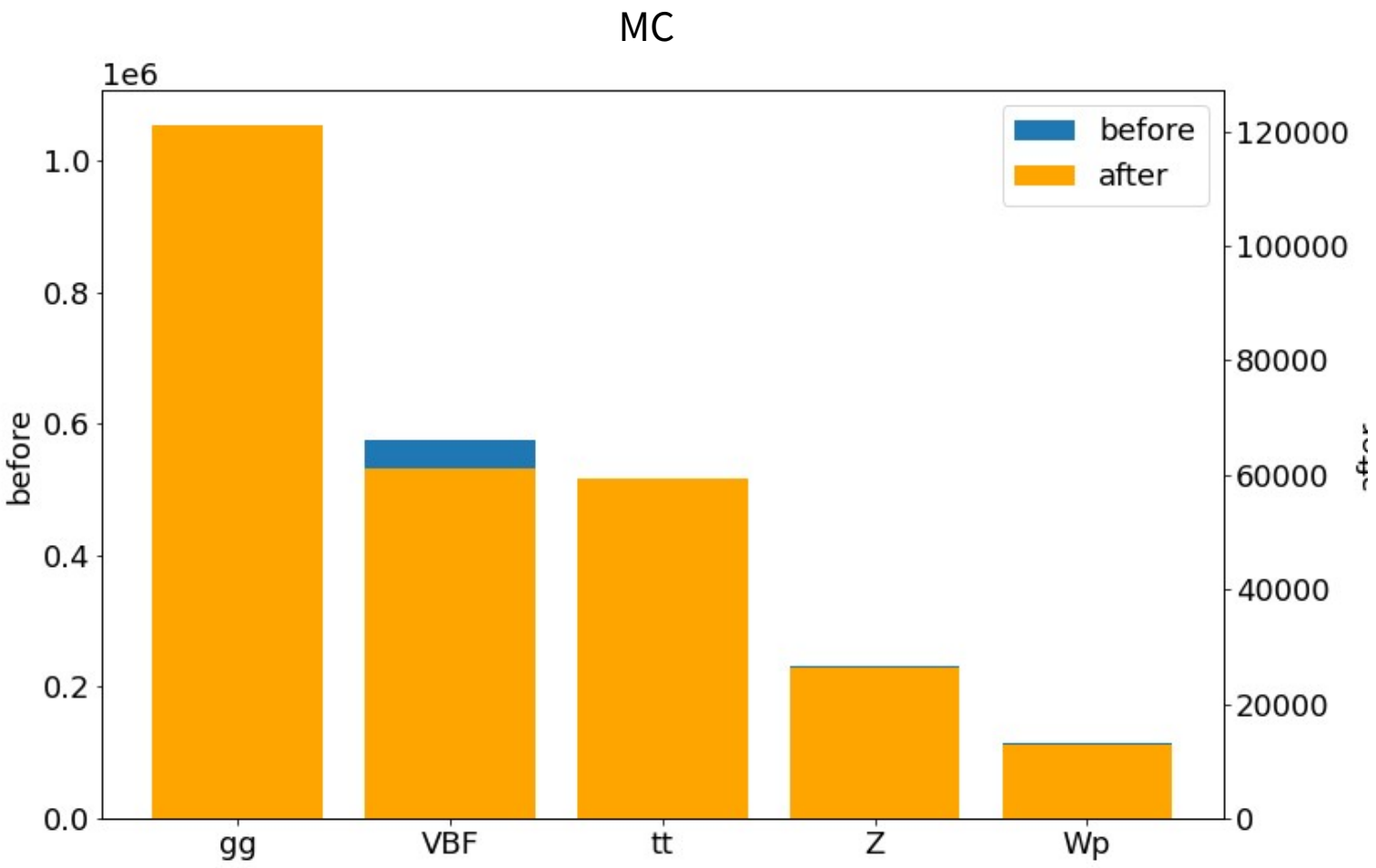
d)

(borrowed from https://grab-group.ethz.ch/research/higgs-boson.html)

Event selection for Hγγ peak

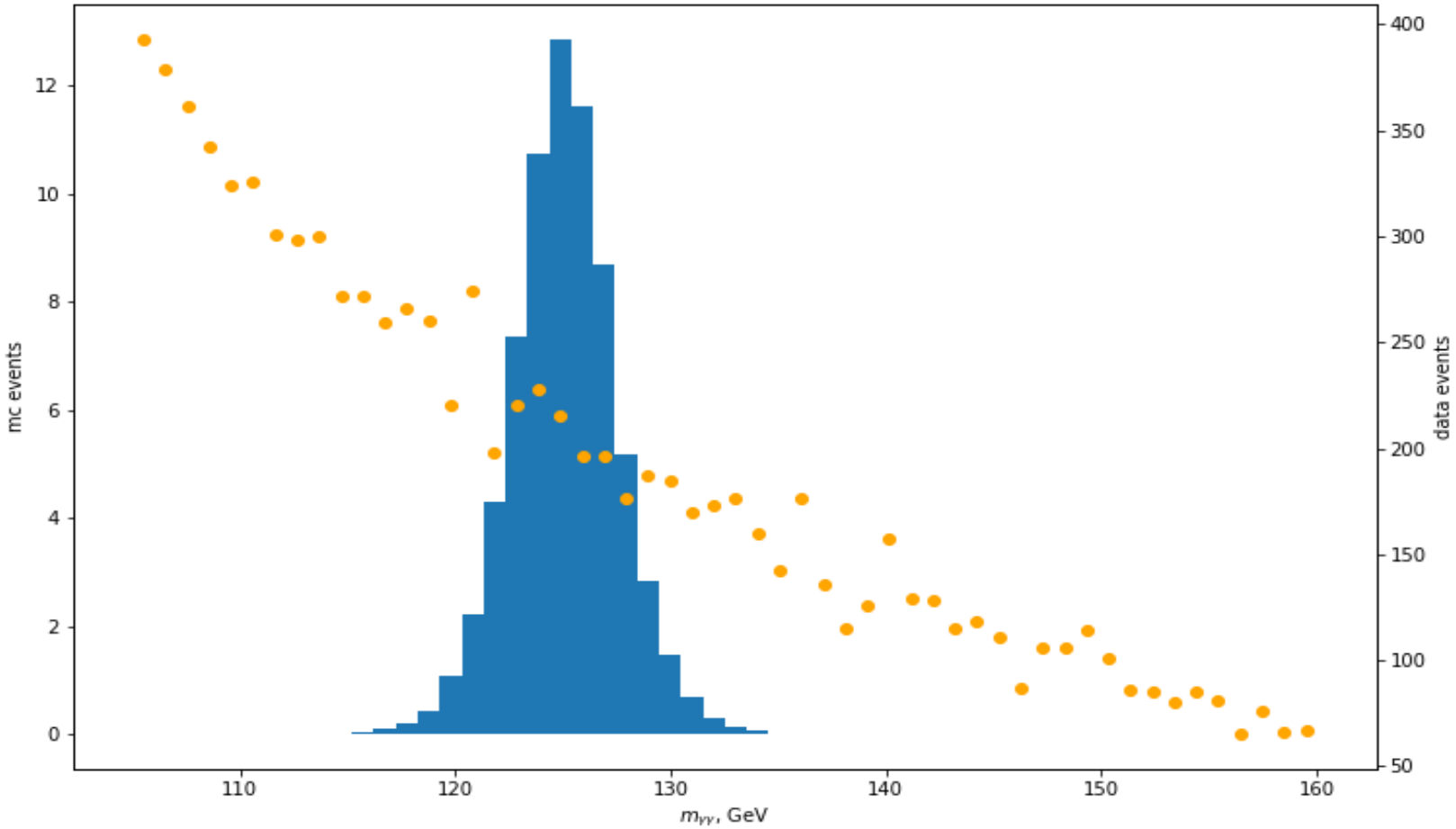
- Diphoton trigger is satisfied;
- Exactly two photons with $E_T > 35$ and 25 GeV, respectively;
- Leading and subleading photon candidates are respectively required to have $E_T/m_{\gamma\gamma} > 0.35$ and 0.25;
- Diphoton invariant mass $m_{\gamma\gamma}$ between 105 GeV and 160 GeV.

Selection efficiency ~ 10%



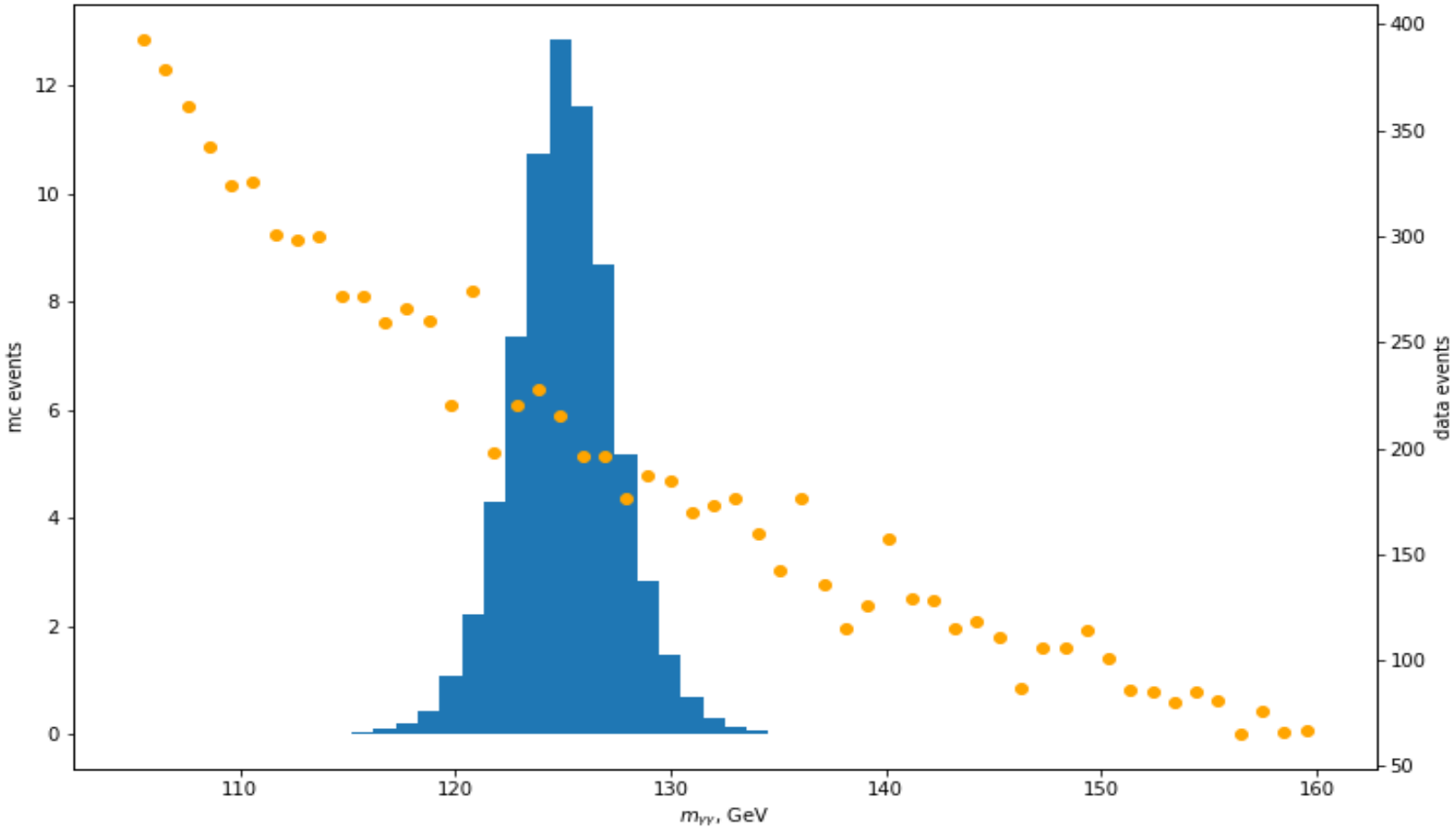
Number of events per process

Higgs mass from ATLAS Open Data



8M events in 7 minutes

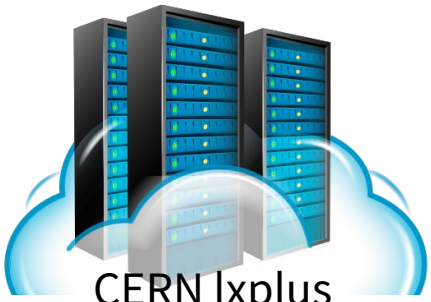
Acquire data efficiently



8M events in 7 minutes **with Python!**

Data pipeline

Processing power and data traffic



CERN lxplus

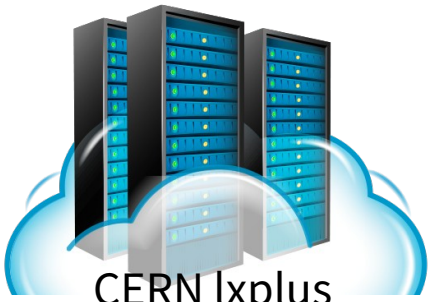
Data pipeline

CERNBox

1TB Root files storage

UpRoot (xrootd)

Processing power and data traffic



CERN lxplus

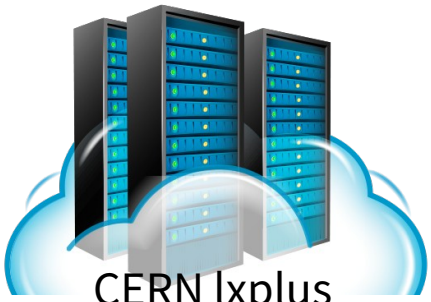
Data pipeline

CERNBox

1TB Root files storage

UpRoot (xrootd)

Processing power and data traffic



CERN lxplus

NumPy

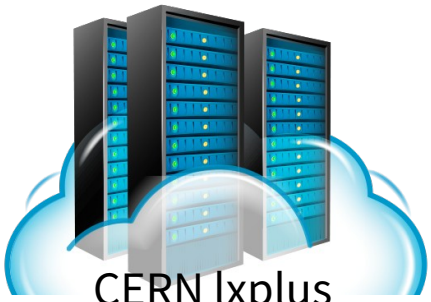
Processed files storage

Data pipeline

1TB Root files storage

UpRoot (xrootd)

Processing power and data traffic



CERN lxplus

NumPy

Processed files storage

Jupyter over SSH

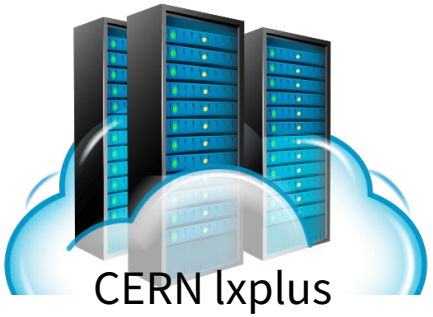
User Interface

Data pipeline

Processing power and data traffic

1TB Root files storage

UpRoot (xrootd)



NumPy

2GB – 100 GB

Batches 200MB
~ 100 000 events

Store processed
~ 2MB
~1000 events/batch

What's inside?

Simulation wide

- XSection
- SumWeights

Macro features

1 number per event

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Micro features

Jagged. List of numbers per event

- photon_n
- photon_truthMatched
- photon_trigMatched
- photon_pt
- photon_eta
- photon_phi
- photon_E
- photon_isTightID
- photon_ptcone30
- photon_etcone20
- photon_convType
- photon_pt_syst
- lep_n
- lep_truthMatched
- lep_trigMatched
- lep_pt
- lep_eta
- lep_phi
- lep_E
- lep_z0
- lep_charge
- lep_type
- lep_isTightID
- lep_ptcone30
- lep_etcone20
- lep_trackd0pvunbiased
- lep_tracksigd0pvunbiased
- lep_pt_syst
- tau_n
- tau_pt
- tau_eta
- tau_phi
- tau_E
- tau_isTightID
- tau_truthMatched
- tau_trigMatched
- tau_nTracks
- tau_BDTid
- tau_pt_syst
- tau_charge
- jet_n
- jet_pt
- jet_eta
- jet_phi
- jet_E
- jet_jvt
- jet_trueflav
- jet_truthMatched
- jet_MV2c10
- jet_pt_syst
- largeRjet_n
- largeRjet_pt
- largeRjet_eta
- largeRjet_phi
- largeRjet_E
- largeRjet_m
- largeRjet_truthMatched
- largeRjet_D2
- largeRjet_tau32
- largeRjet_pt_syst

Mask bubbling for feature extraction

- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*

- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features

[0, 0, 1, 1, 0, 1 ...]

1000 events

MASK

800

Photon Features

- photon_n
- photon_truthM…
- photon_trig…
- photon_p…
- photon_…
- photon_phi
- photon_E
- photon_isTightID
- photo…ne30
- photo…ne20
- photo…Type
- photo…yst

- lep_n
- lep_truthMatched
- lep_trigMatched
- …_pt
- …eta
- …phi
- lep_E
- lep_z0
- lep_charge
- lep_type
- lep_isTightID
- lep_ptcone30
- lep_etcone20
- lep_trackd0pvunbiased
- lep_tracksigd0pvunbiased
- lep_pt_syst

- tau_n
- tau_pt
- tau_eta
- tau_phi
- tau_E
- tau_isTightID
- tau_truthMatched
- tau_trigMatched
- tau_nTracks
- tau_BDTid
- tau_pt_syst
- tau_charge

- jet_n
- jet_pt
- jet_eta
- jet_phi
- jet_E
- jet_jvt
- jet_trueflav
- jet_truthMatched
- jet_MV2c10
- jet_pt_syst

- largeRjet_n
- largeRjet_pt
- largeRjet_eta
- largeRjet_phi
- largeRjet_E
- largeRjet_m
- largeRjet_truthMatched
- largeRjet_D2
- largeRjet_tau32
- largeRjet_pt_syst

Micro features

Mask bubbling for feature extraction

Simulation wide

- XSection
- SumWeights

Macro features

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

[0, 0, 1, 1, 0, 1 ...] > [0, 0, 0, 1, 0, 1 ...]

1000 events

MASK

MASK

800

600

Photon Features

- photon_n
- photon_truthM
- photon_trig
- photon_p
- photon_
- photon_phi
- photon_E
- photon_isTightID
- photo ne30
- photo ne20
- photo Type
- photo yst

Lepton Features

- lep_n
- lep_truthMatch
- lep_trigMa
- p_pt
- eta
- phi
- lep_E
- lep_z0
- lep_
- lep_t
- lep_i D
- lep_p 0
- le
- le vunbiased
- pvunbiased
- lep_pt_syst

- tau_n
- tau_pt
- au_eta
- phi
- E
- isTightID
- tau_truthMatched
- tau_trigMatched
- tau_nTracks
- tau_BDTid
- tau_pt_syst
- tau_charge

- jet_n
- jet_pt
- jet_eta
- jet_phi
- jet_E
- jet_jvt
- jet_trueflav
- jet_truthMatched
- jet_MV2c10
- jet_pt_syst

- largeRjet_n
- largeRjet_pt
- largeRjet_eta
- largeRjet_phi
- largeRjet_E
- largeRjet_m
- largeRjet_truthMatched
- largeRjet_D2
- largeRjet_tau32
- largeRjet_pt_syst

Micro features

Mask bubbling for feature extraction

- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features

[0, 0, 1, 1, 0, 1 ...] > [0, 0, 0, 1, 0, 1 ...] > [0, 0, 0, 1, 0, 0 ...]

1000 events

- photon_n
- photon_truthMa
- photon_trig
- photon_p
- photon_
- photon_phi
- photon_E
- photon_isTightID
- photo ne30
- photo ne20
- photo Type
- photo yst

MASK

800

Photon Features

- lep_n
- lep_truthMatch
- lep_trigMa
- p_pt
- eta
- phi
- lep_E
- lep_z0
- lep_
- lep_t
- lep_i D
- lep_p 0
- le
- le vunbiased
- pvunbiased

MASK

600

Lepton Features

- tau_n
- tau_pt
- au_eta
- phi
- E
- isTightID
- tau_truthMatched
- tau_trigMatched
- tau_nT
- tau_BD
- tau_pt_
- tau_ch

MASK

300

Tau Features

- jet_n
- jet_pt
- eta
- hi
- jet_trueflav
- jet_truthMatched
- jet_MV2c10
- jet_pt_syst

- largeRjet_n
- largeRjet_pt
- largeRjet_eta
- largeRjet_phi
- largeRjet_E
- largeRjet_m
- largeRjet_truthMatched
- largeRjet_D2
- largeRjet_tau32
- largeRjet_pt_syst

250 events

Micro features

Mask bubbling for feature extraction

- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features

[0, 0, 1, 1, 0, 1 ...] > [0, 0, 0, 1, 0, 1 ...] > [0, 0, 0, 1, 0, 0 ...]

1000 events

- photon_n
- photon_truthMa
- photon_trig
- photon_p
- photon_
- photon_phi
- photon_E
- photon_isTightID
- photo ne30
- photo ne20
- photo Type
- photo yst

MASK

800

Photon Features

- lep_n
- lep_truthMatch
- lep_trigMa
- p_pt
- eta
- phi
- lep_E
- lep_z0
- lep_
- lep_t
- lep_i D
- lep_p 0
- le
- le vunbiased
- pvunbiased

MASK

600

Lepton Features

- tau_n
- tau_pt
- au_eta
- phi
- E
- isTightID
- tau_truthMatched
- tau_trigMatched
- tau_nT
- tau_BD
- tau_pt_
- tau_ch

MASK

250

Tau Featur

- jet_n
- jet_pt
- eta
- hi
- jet_trueflav
- jet_truthMatched
- jet_MV2c10
- jet_pt_syst

MASK

- largeRjet_n
- largeRjet_pt
- largeRjet_eta
- largeRjet_phi
- largeRjet_E
- largeRjet_m
- largeRjet_truthMatched
- largeRjet_D2
- largeRjet_tau32
- largeRjet_pt_syst

250 events

Micro features

Mask bubbling for feature extraction

- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features

[0, 0, 1, 1, 0, 1 ...] > [0, 0, 0, 1, 0, 1 ...] > [0, 0, 0, 1, 0, 0 ...]

1000 events

- photon_n
- photon_truthMa
- photon_trig
- photon_p
- photon_
- photon_phi
- photon_E
- photon_isTightID
- photo ne30
- photo ne20
- photo Type
- photo yst

MASK

800

Photon Features

- lep_n
- lep_truthMatch
- lep_trigMa
- p_pt
- eta
- phi
- lep_E
- lep_z0
- lep_
- lep_t
- lep_i D
- lep_p 0
- le
- le
- le
- lep_pt_syst

MASK

250

Lepton Feature

ased

biased

- tau_n
- tau_pt
- au_eta
- phi
- E
- isTightID
- tau_truthMatched
- tau_trigMatched
- tau_nT
- tau_BD
- tau_pt_
- tau_ch

MASK

250

Tau Featur

- jet_n
- jet_pt
- eta
- phi
- jet_trueflav
- jet_truthMatched
- jet_MV2c10
- jet_pt_syst

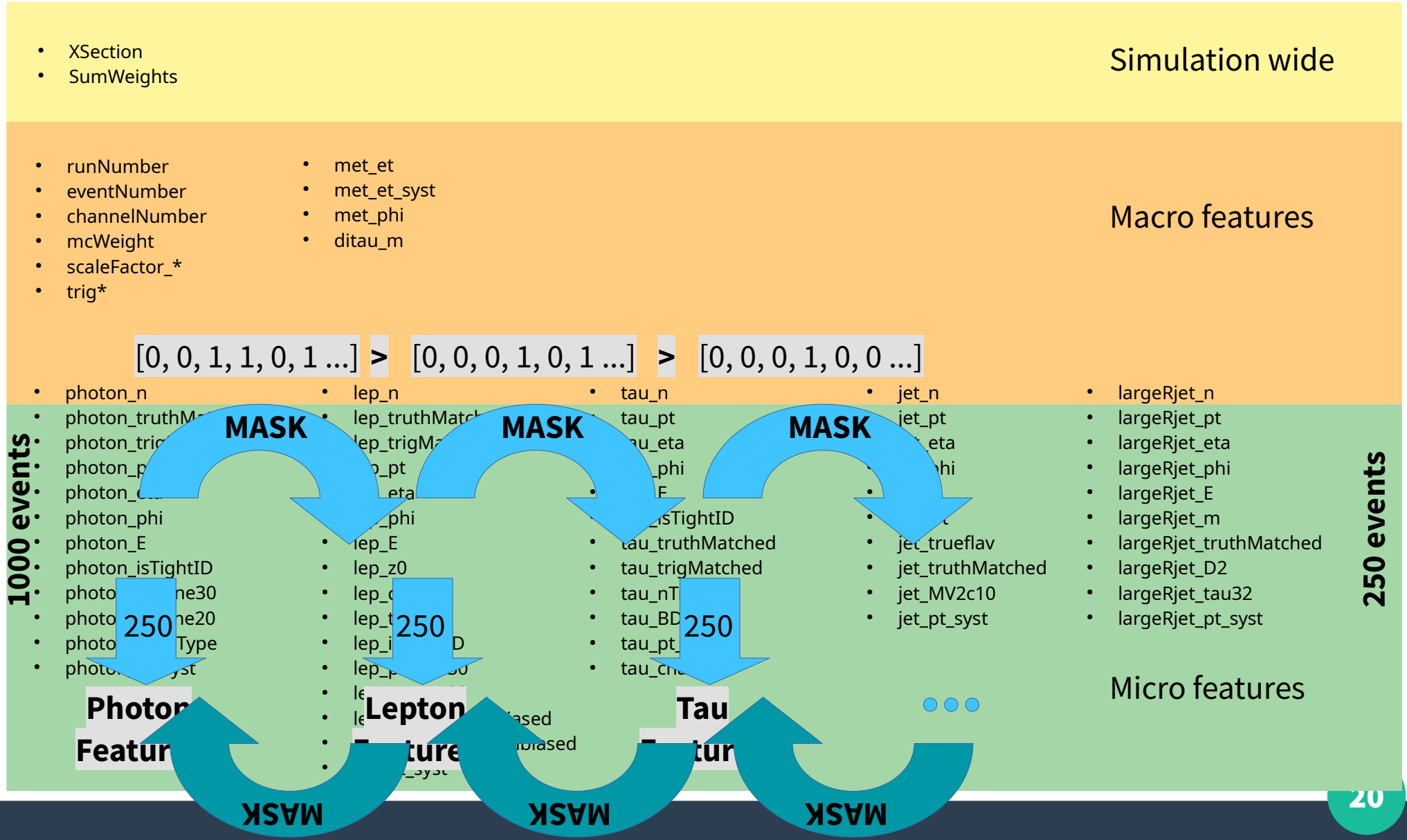
MASK

- largeRjet_n
- largeRjet_pt
- largeRjet_eta
- largeRjet_phi
- largeRjet_E
- largeRjet_m
- largeRjet_truthMatched
- largeRjet_D2
- largeRjet_tau32
- largeRjet_pt_syst

250 events

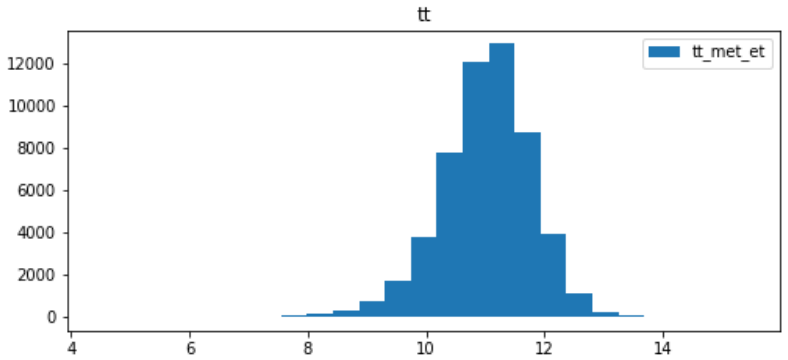
Micro features

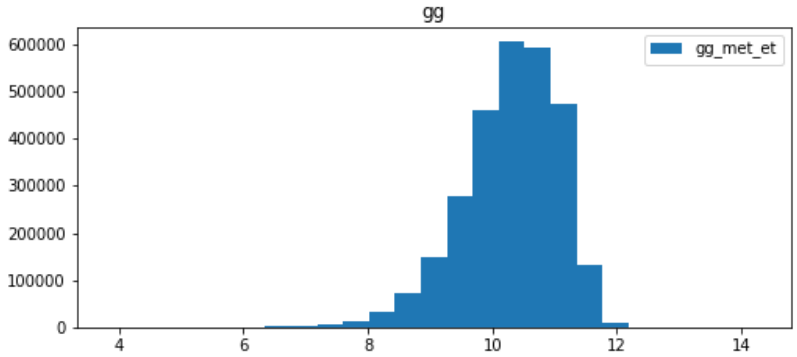
Mask bubbling for feature extraction

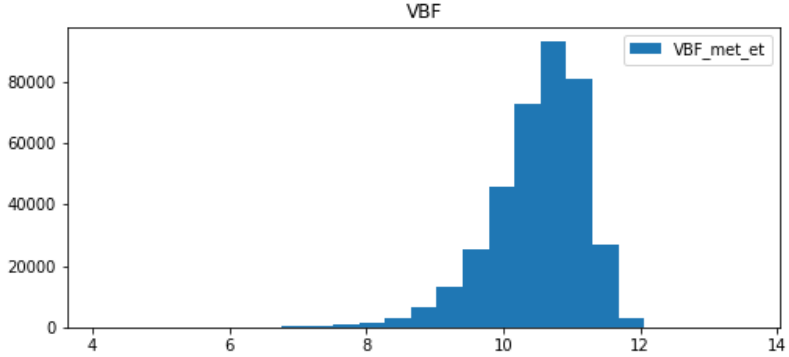


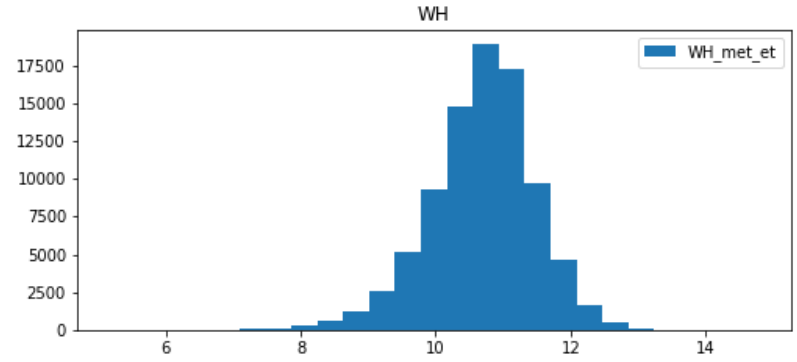
Missing energy Et (log x scale)

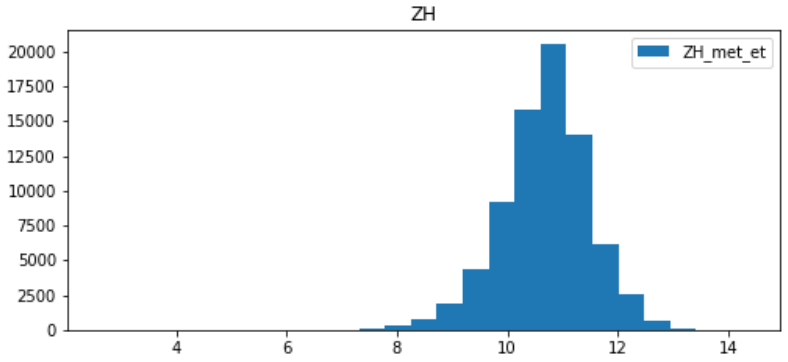
Within 100 000 events batch





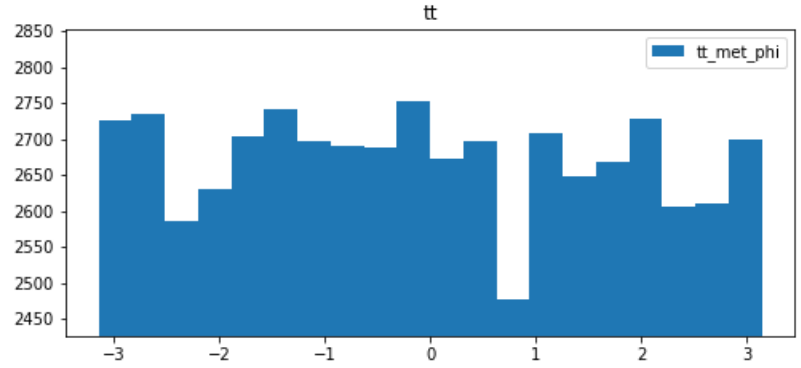


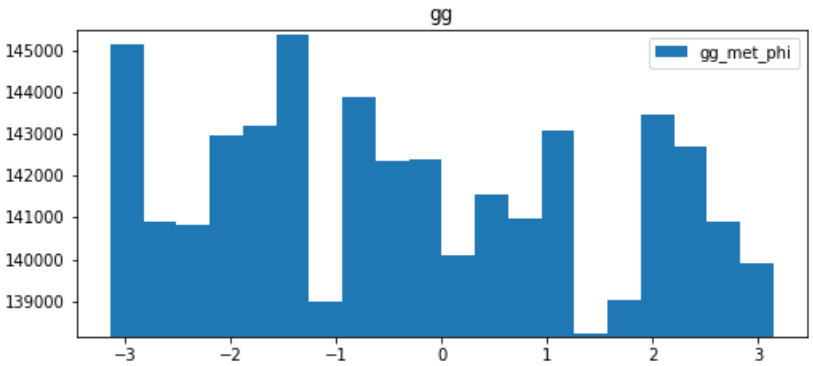


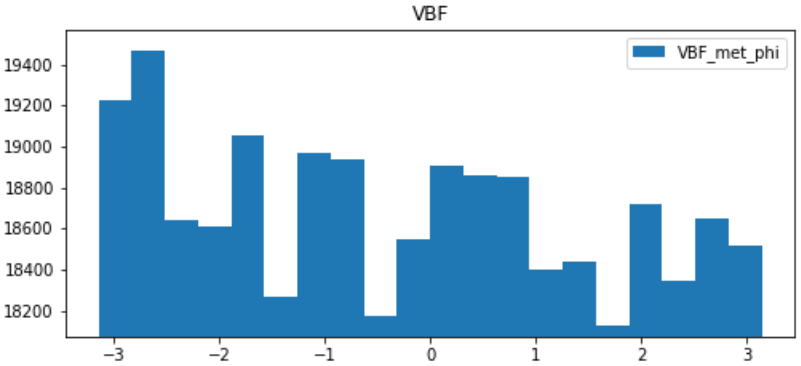


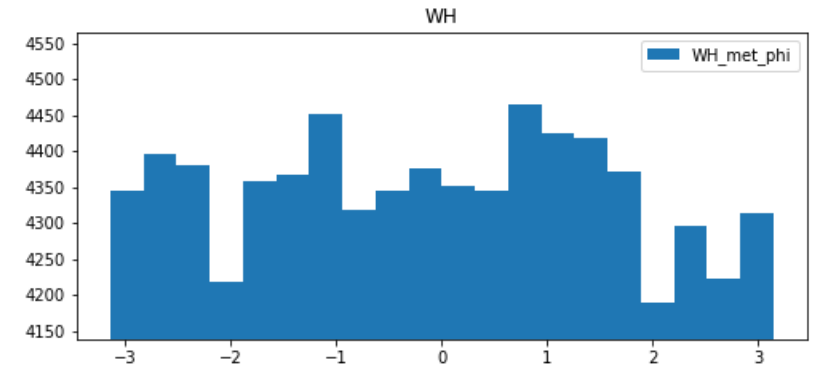
Missing energy Phi

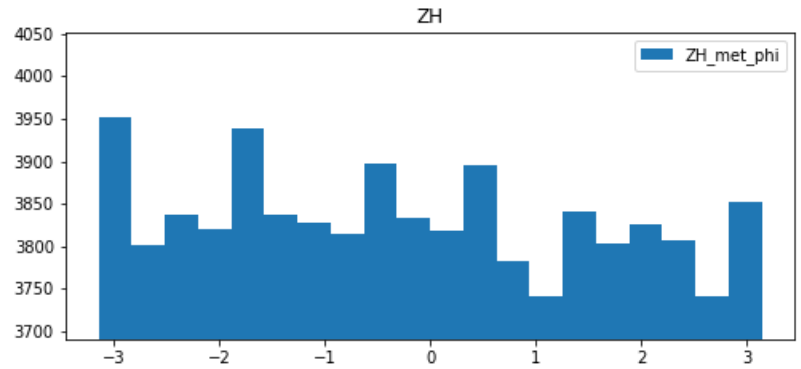
Within 100 000 events batch





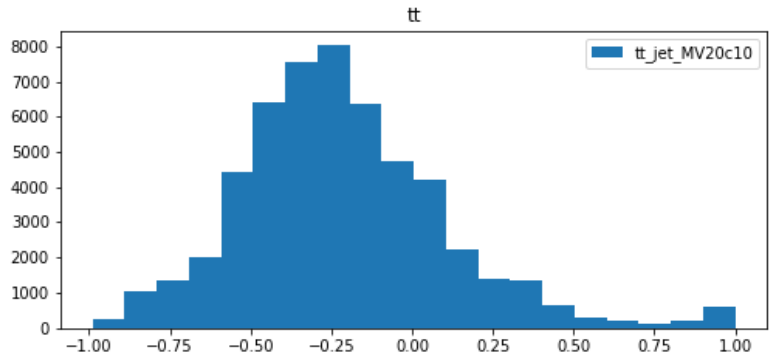


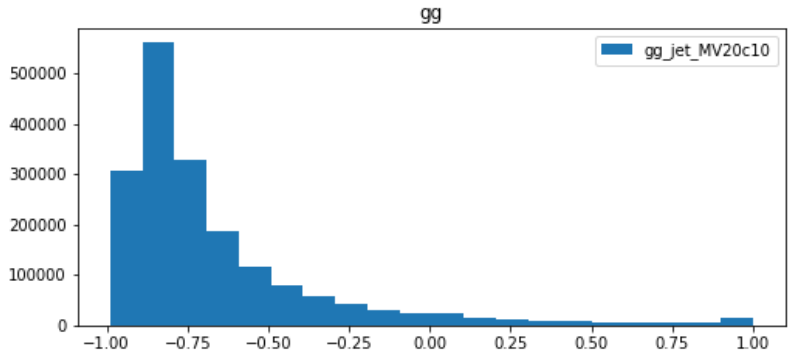


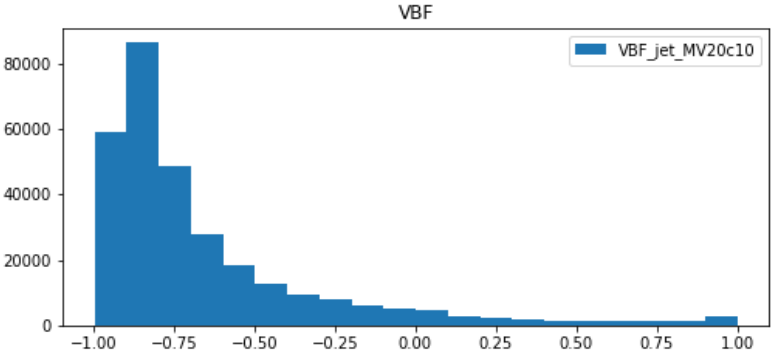


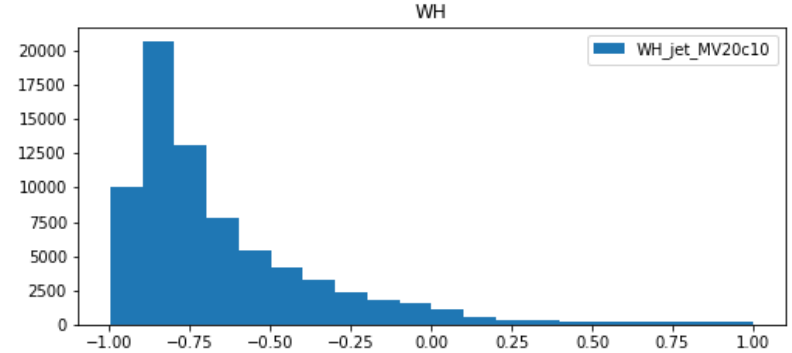
Jet b-tagging score

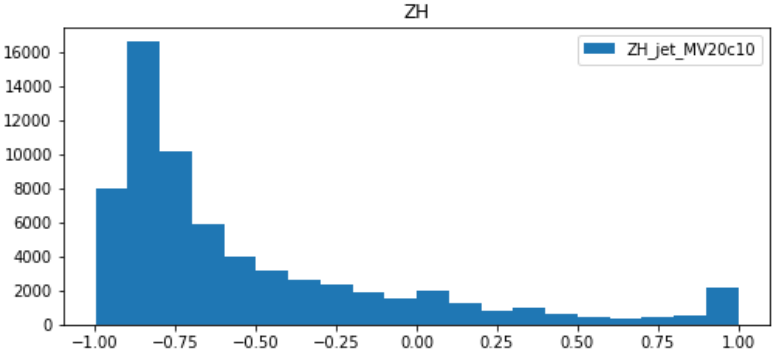
Within 100 000 events batch





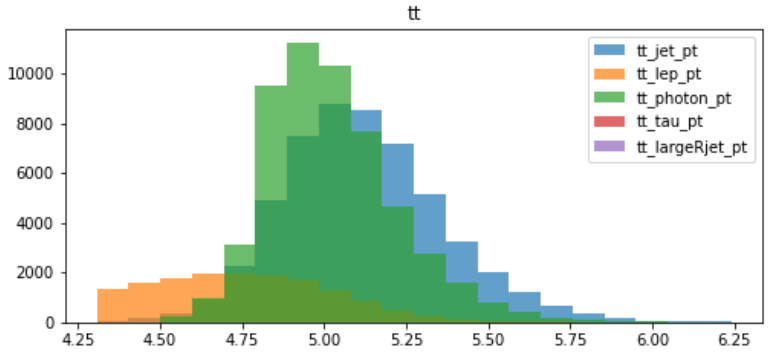


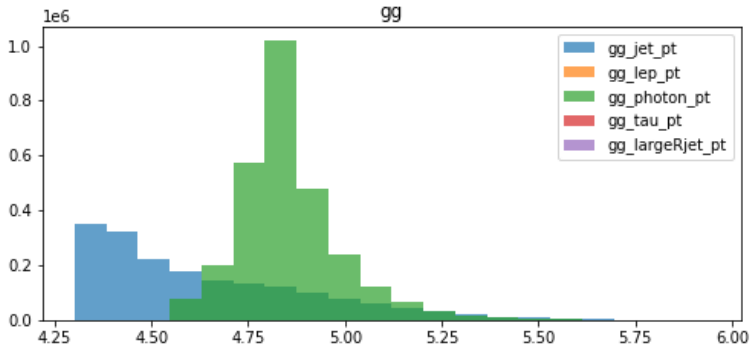


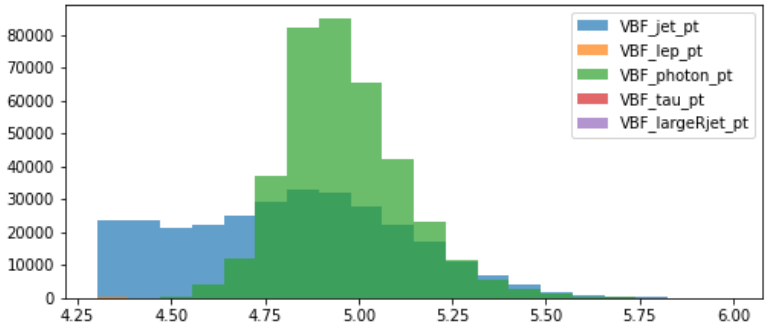


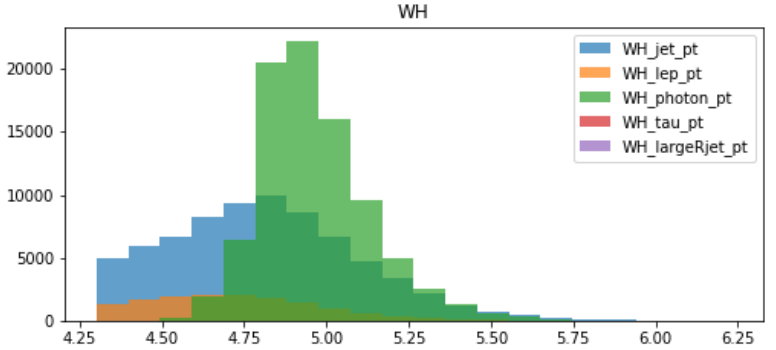
Max pT per particle type (log x scale)

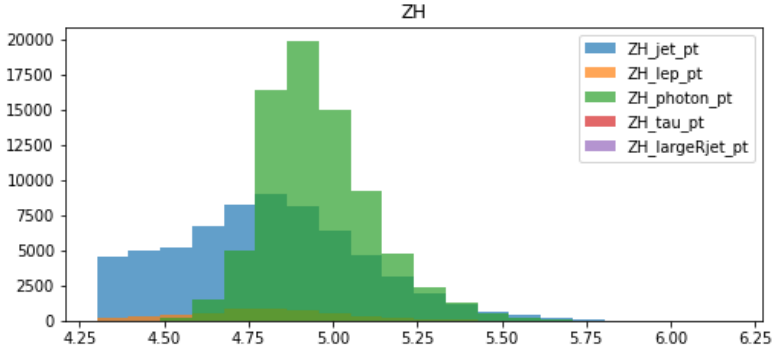
Within 100 000 events batch





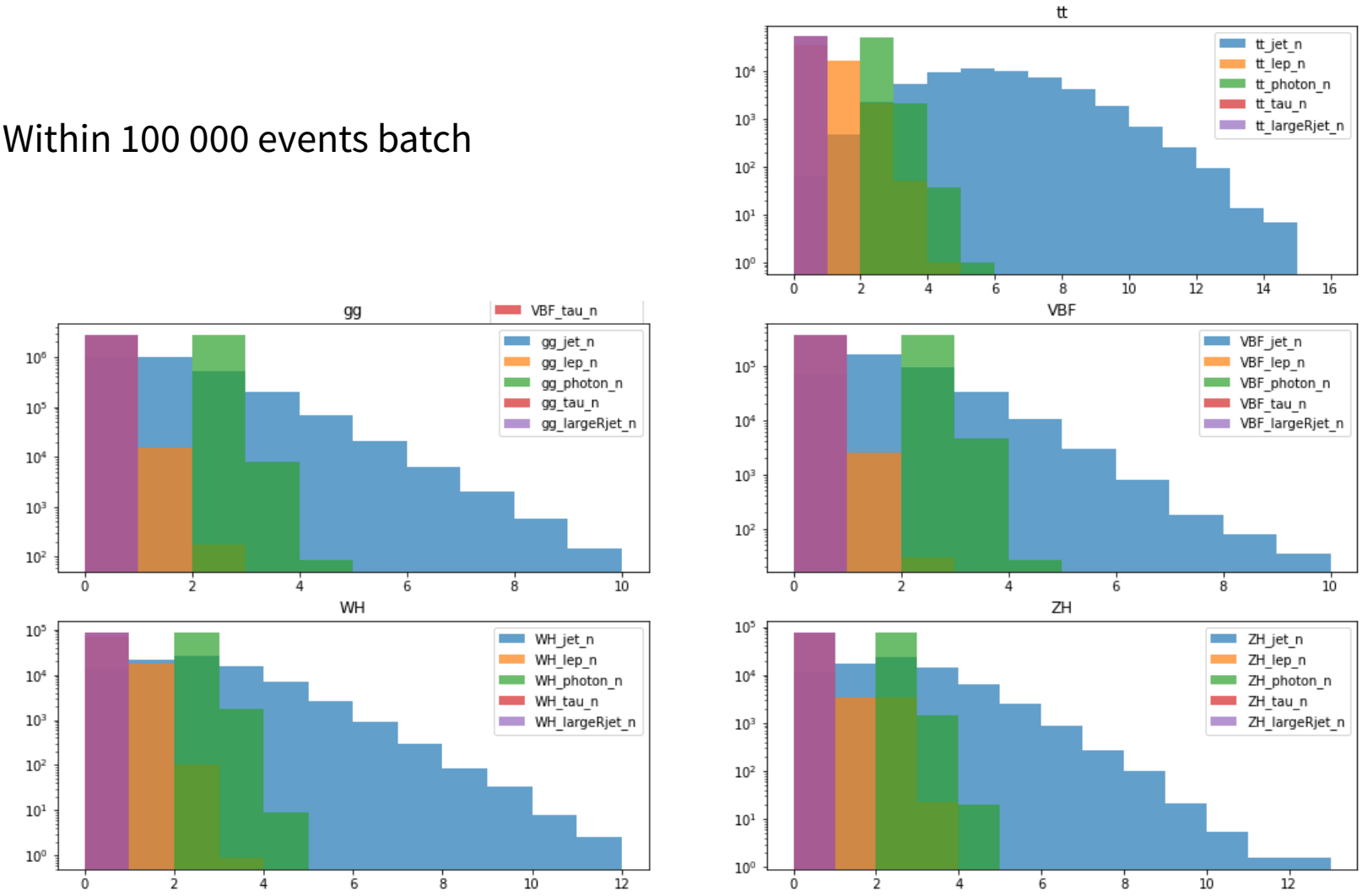




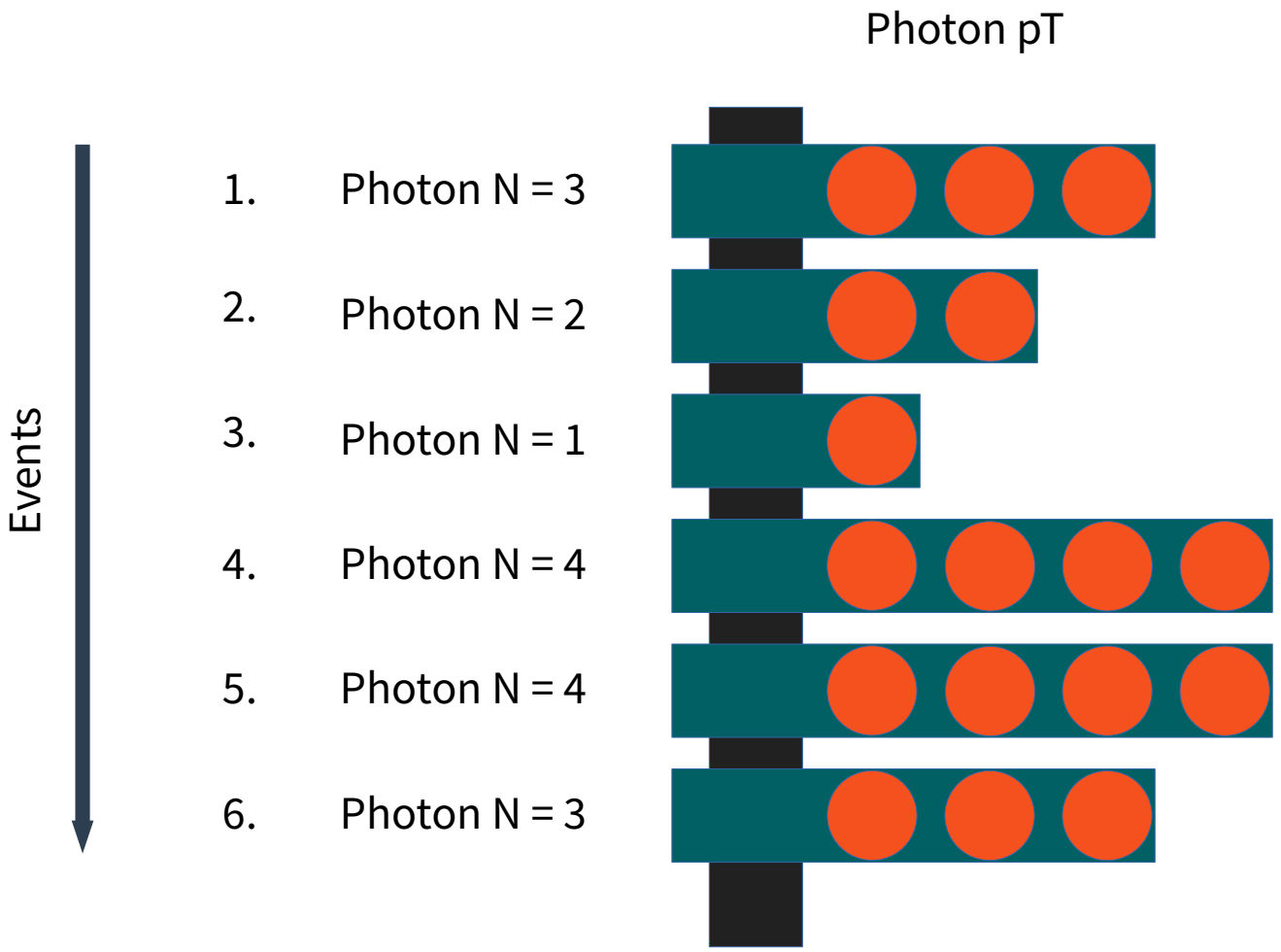


Number of particles (log y scale)

Within 100 000 events batch



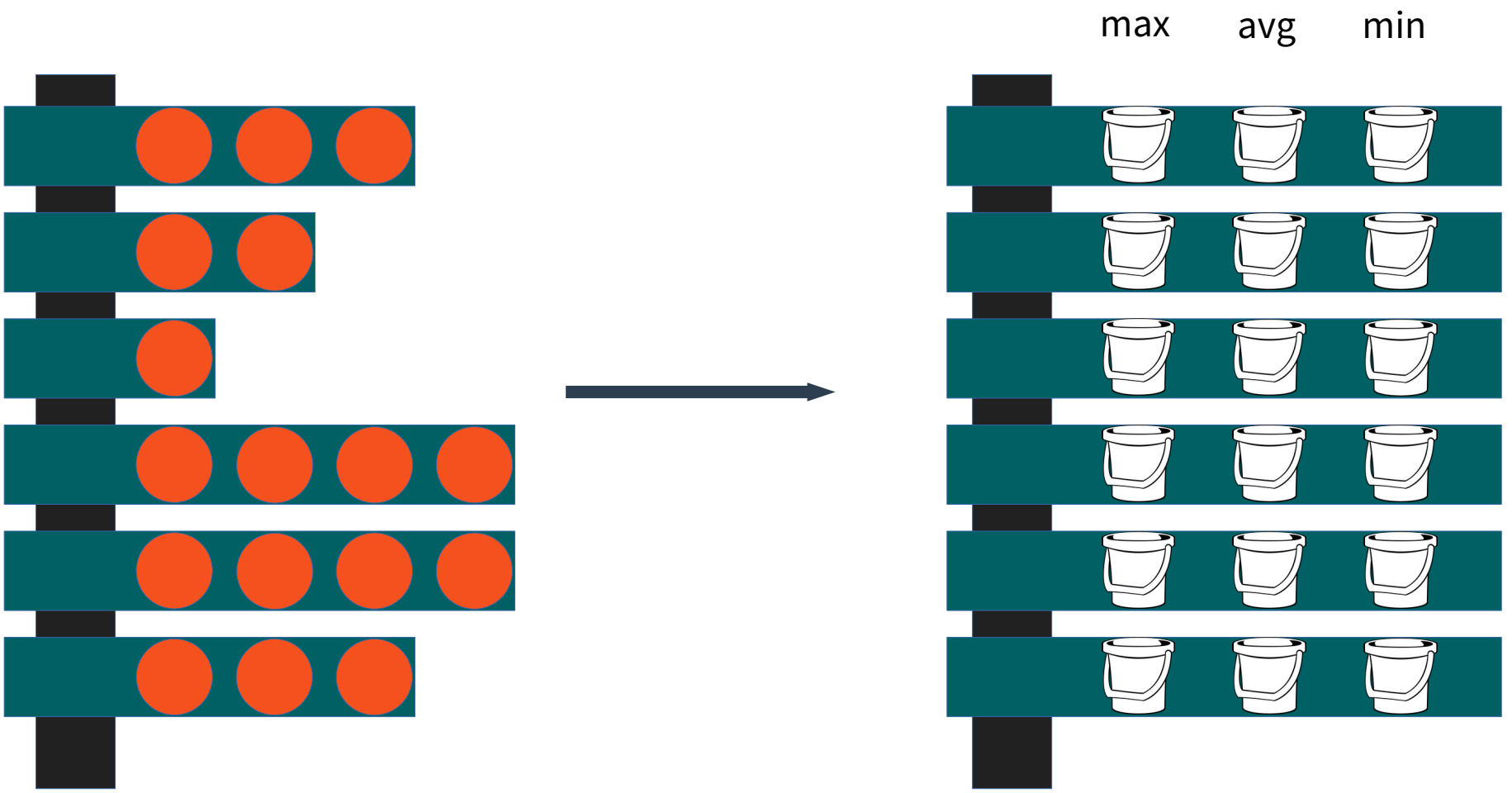
Jagged data. Structure of micro features



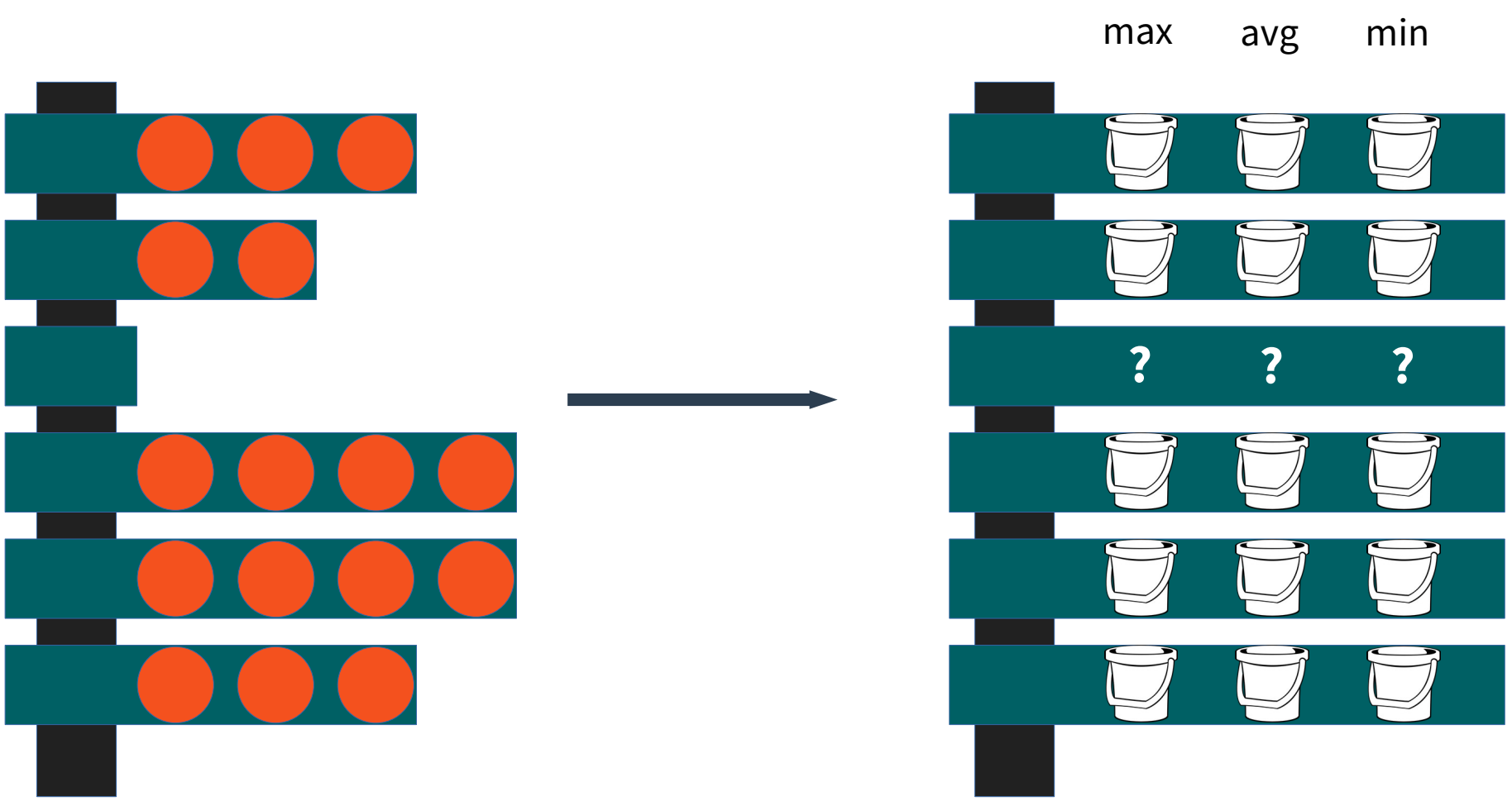
How similar are them?

?

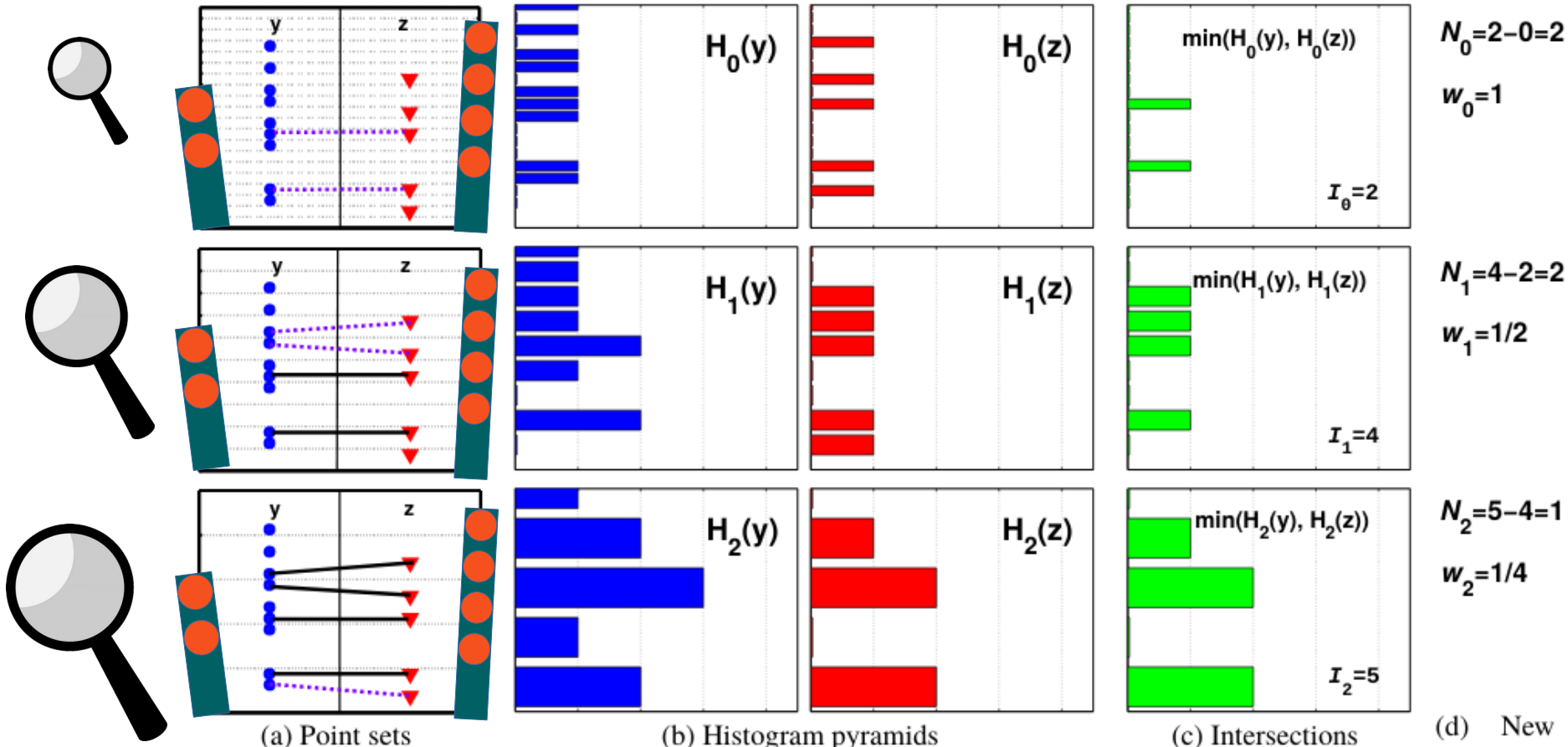
Option: Aggregation



Option: Aggregation



Option: Set Kernels (Pyramid match)



Reference: http://jmlr.csail.mit.edu/papers/volume8/grauman07a/grauman07a.pdf

Option: Set Kernels

\+ Operation in non-linearly transformed feature space at no cost (Kernel trick)

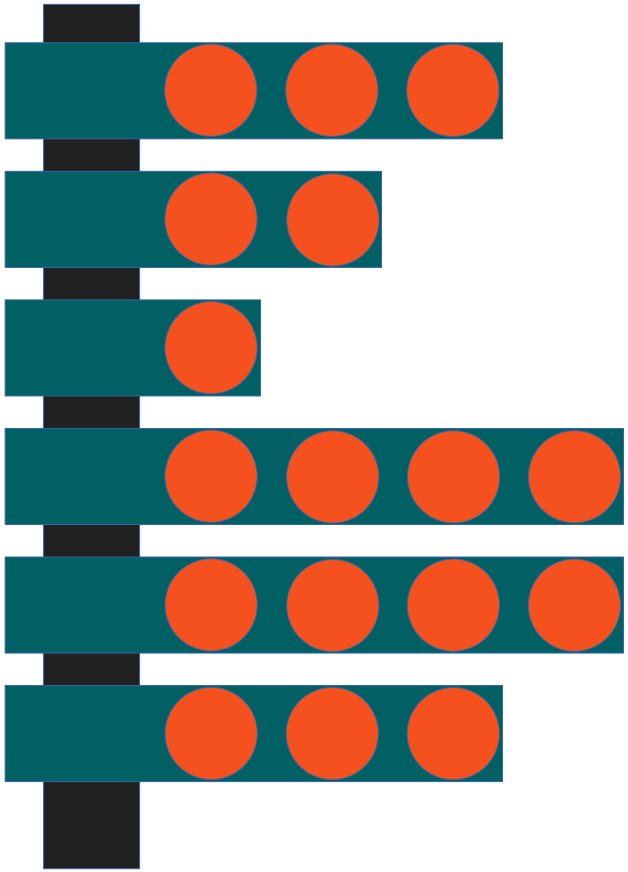
- Interpretable. Not a black box.

− Compute and store N x N matrix of distances (N – number of training instances)

- N^2 memory consumption
- Incremental learning is not easily achievable. 1 extra training point triggers N calls to kernel.

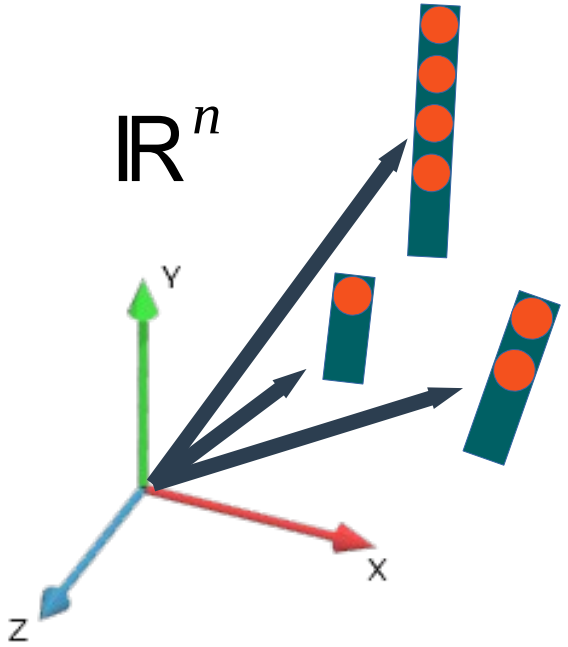
Reference: *StackExchange «What are the limitations of kernel methods?»*

Option: Embedding with NN «DeepSets»



Neural Net

«DeepSets»



Reference: https://arxiv.org/abs/1703.06114

Option: Embedding with NN «DeepSets»

\+ Order independence. Treat set of particles as a *set*

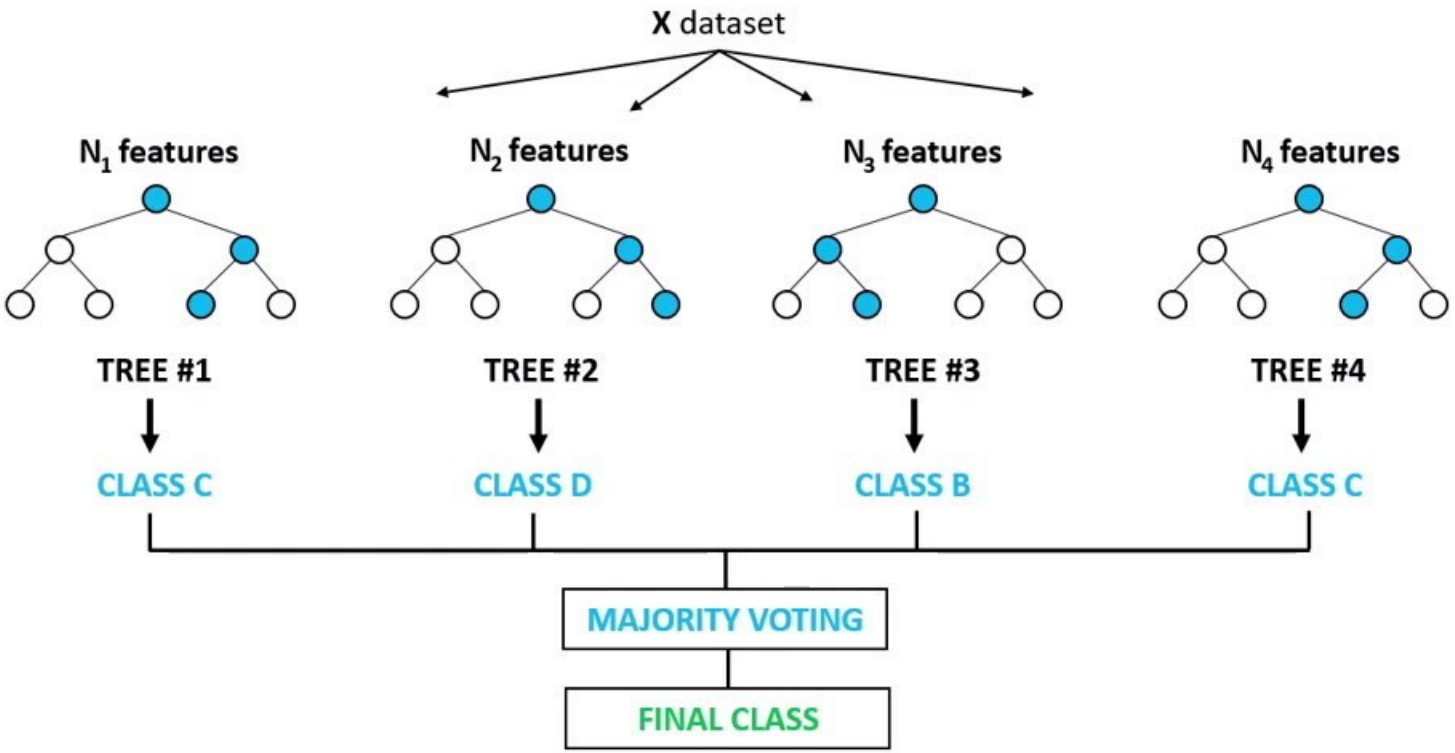
- Get feature embedding model in addition to the classifier

− Neural networks are black box

- Training requires significant amount of data

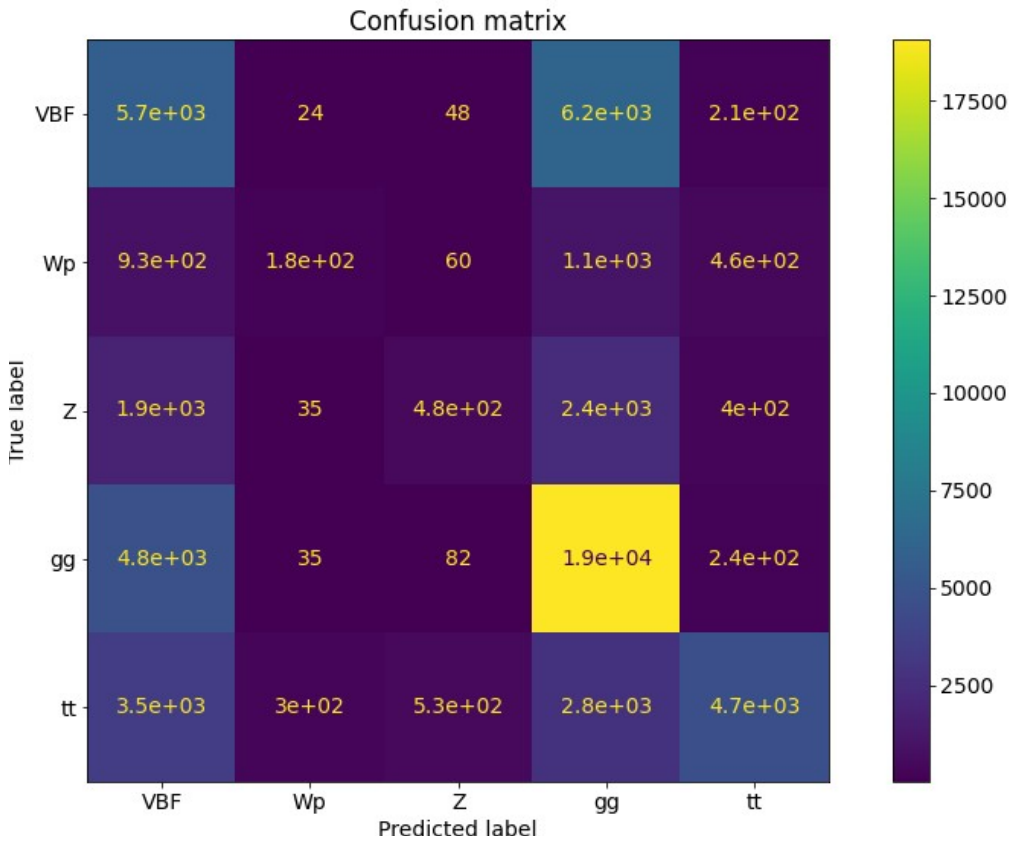
Reference: *StackExchange «What are the limitations of kernel methods?»*

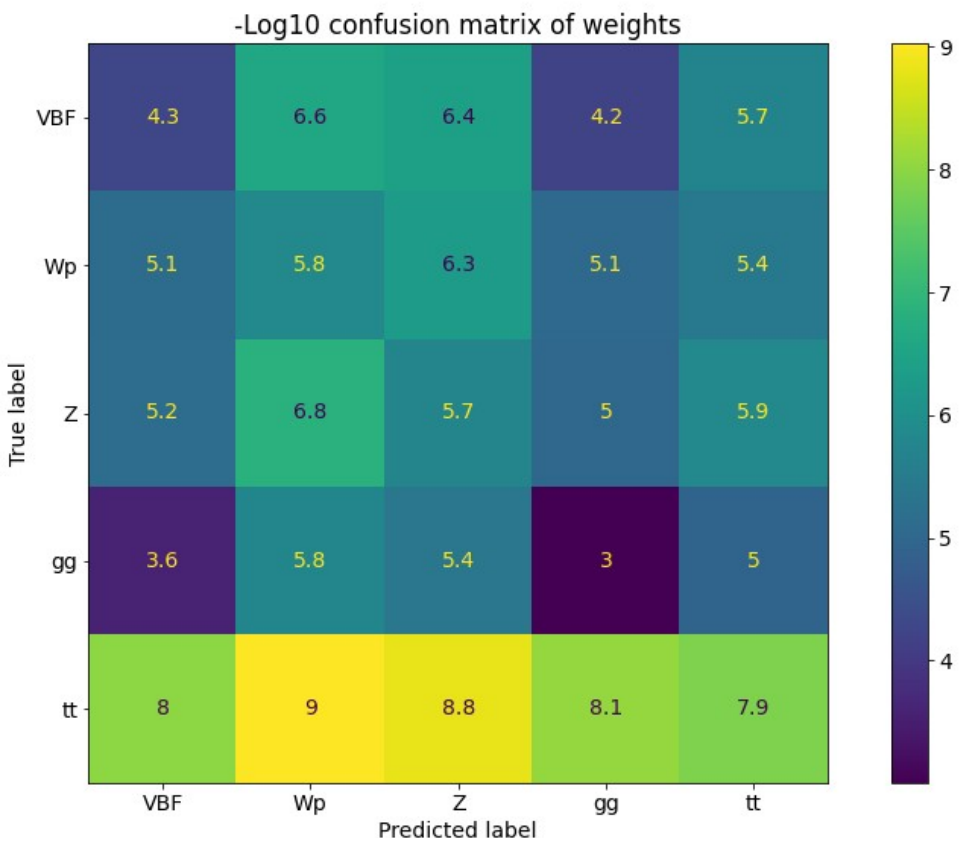
Baseline classifier: Random forest



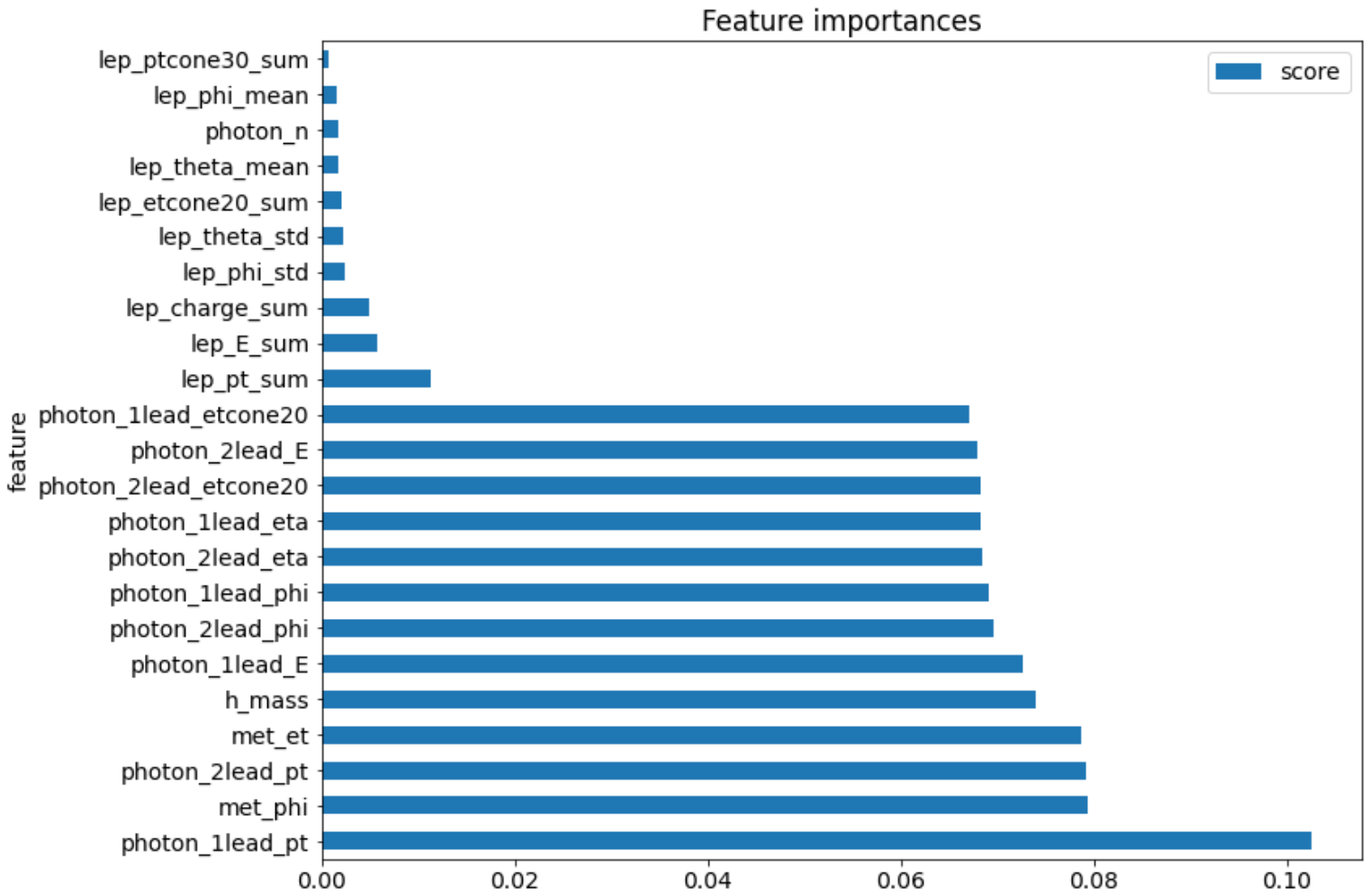
Reference: https://www.wikiwand.com/en/Random_forest#From_bagging_to_random_forests

Baseline classifier: Random forest

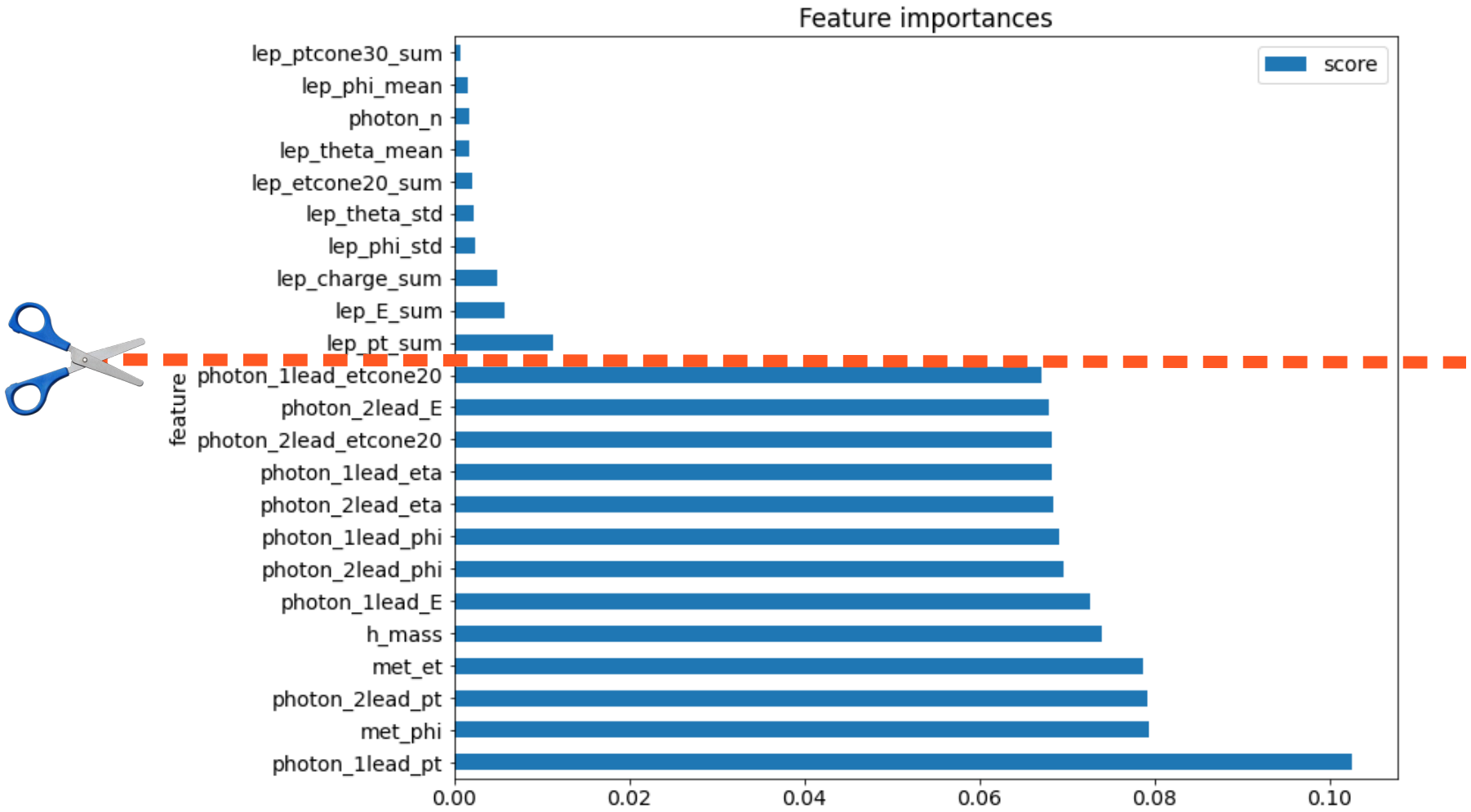




Baseline classifier: Random forest



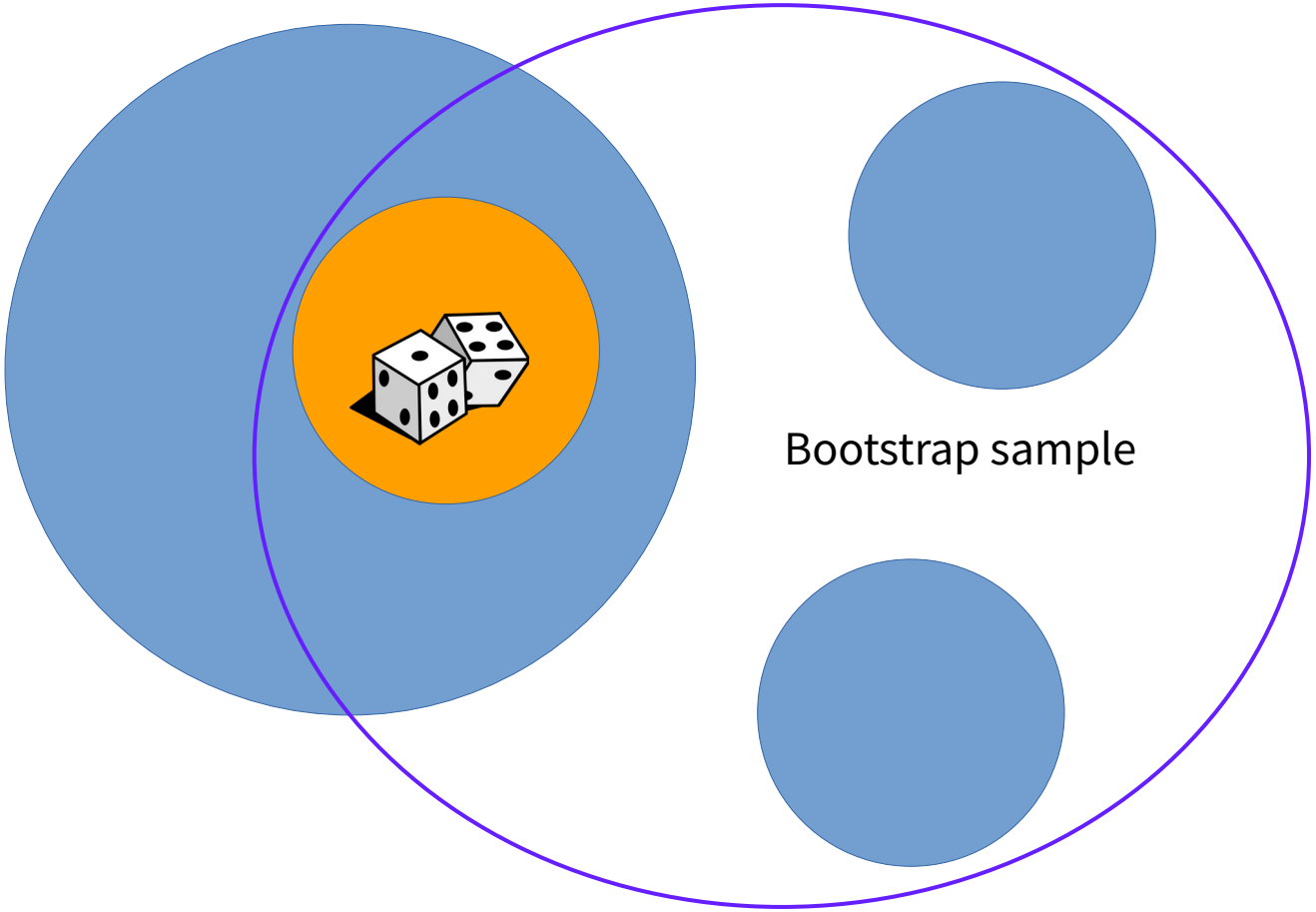
Baseline classifier: Random forest



Imbalanced data

	event_num	weight_sums	weight_frac	weight_per_event
VBF	61021	5.649031e-04	0.080207	9.257520e-09
Wp	12905	1.033396e-04	0.014673	8.007720e-09
Z	26377	9.719725e-05	0.013800	3.684924e-09
gg	121078	6.277415e-03	0.891295	5.184604e-08
tt	59236	1.730808e-07	0.000025	2.921885e-12

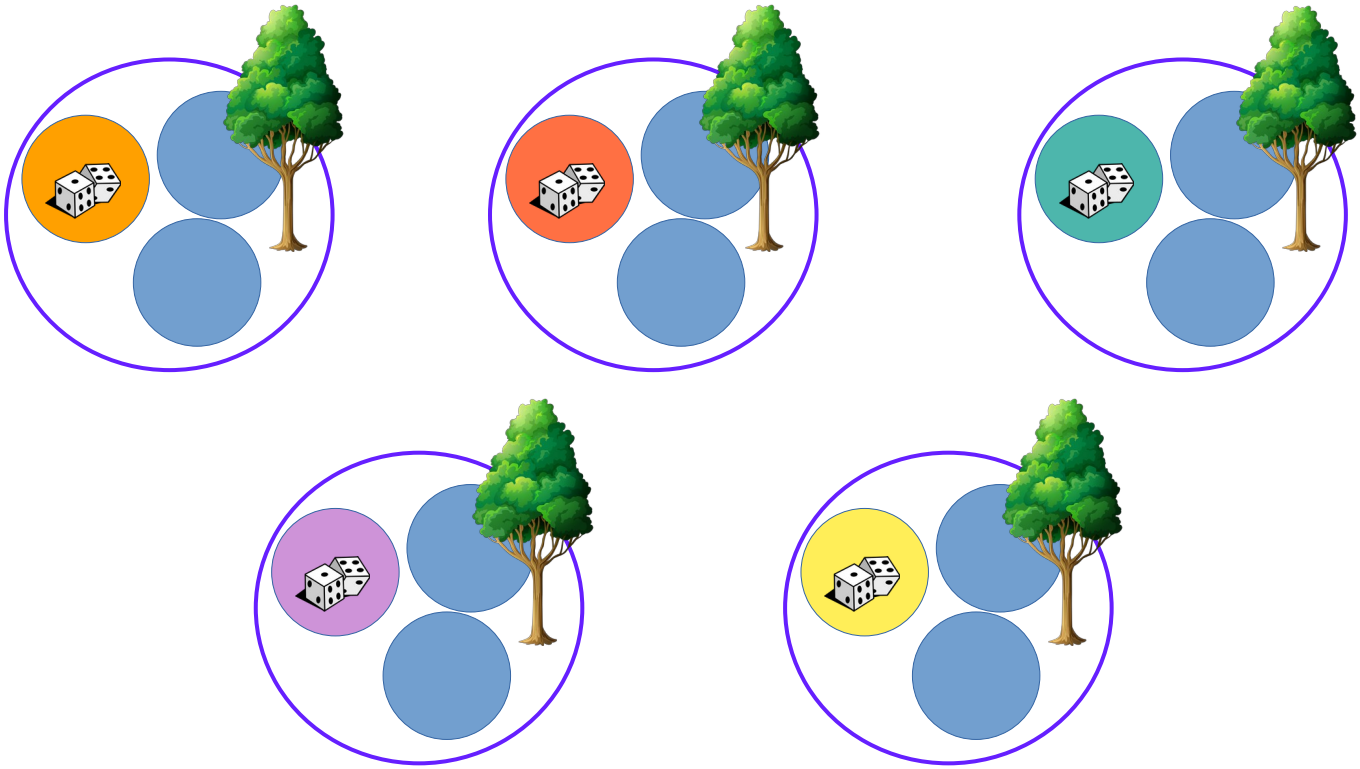
Fighting imbalanced data with bagging



Undersampling

Reference: «Bagging and Random forest for imbalanced classification»

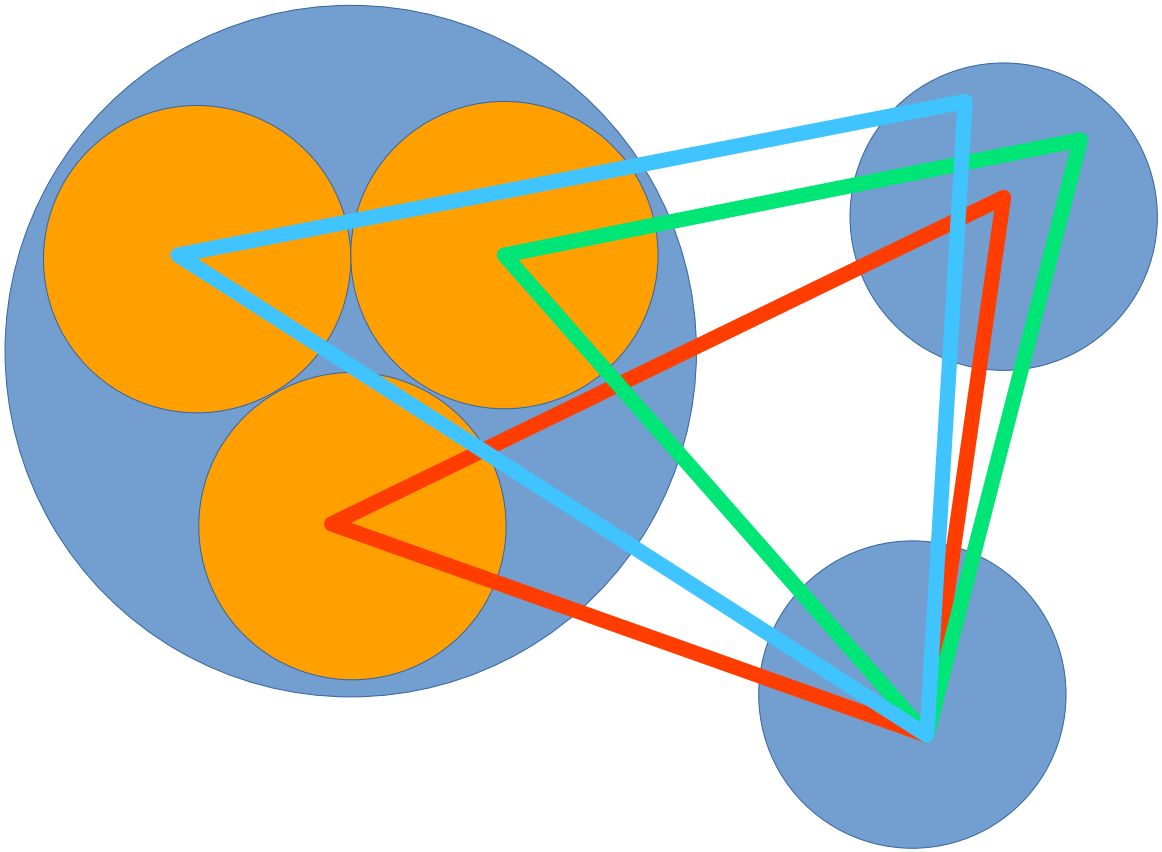
Fighting imbalanced data with bagging



Bagging

Reference: «Bagging and Random forest for imbalanced classification»

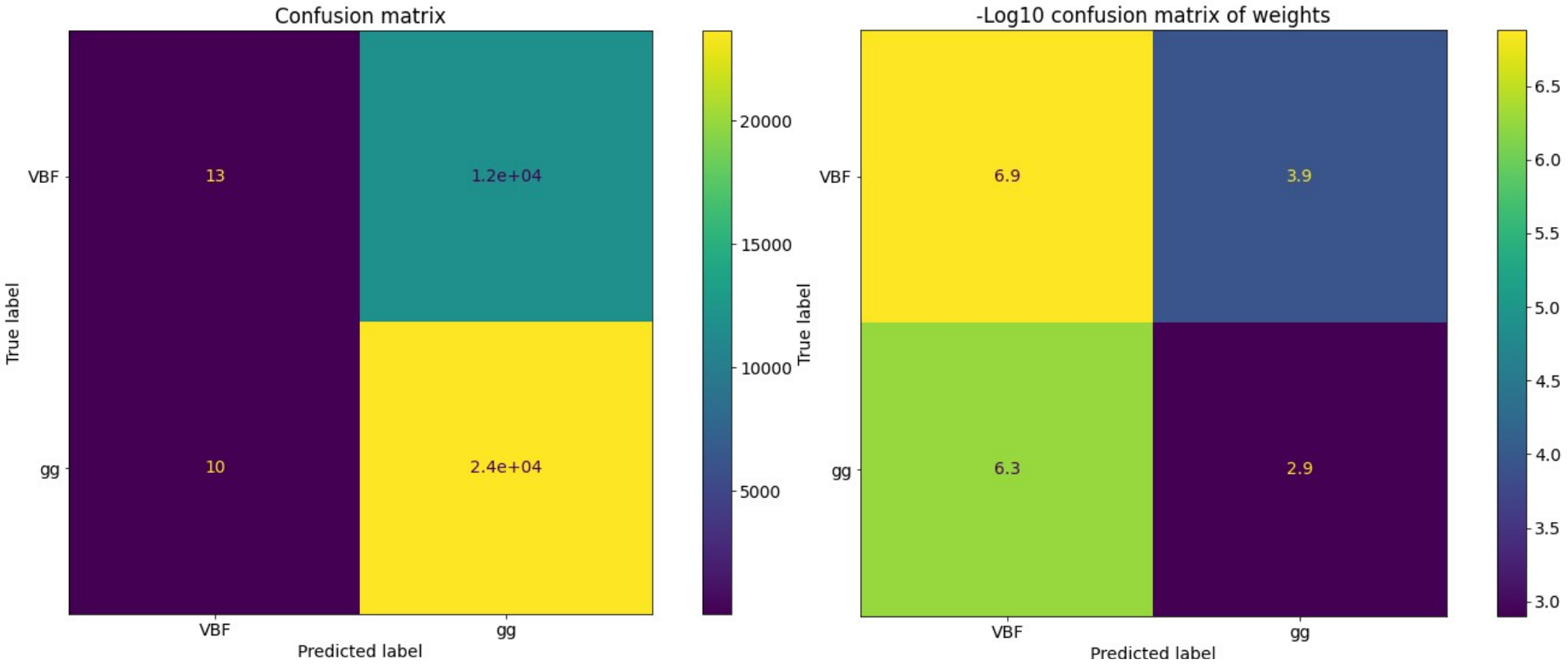
Fighting imbalanced data with bagging



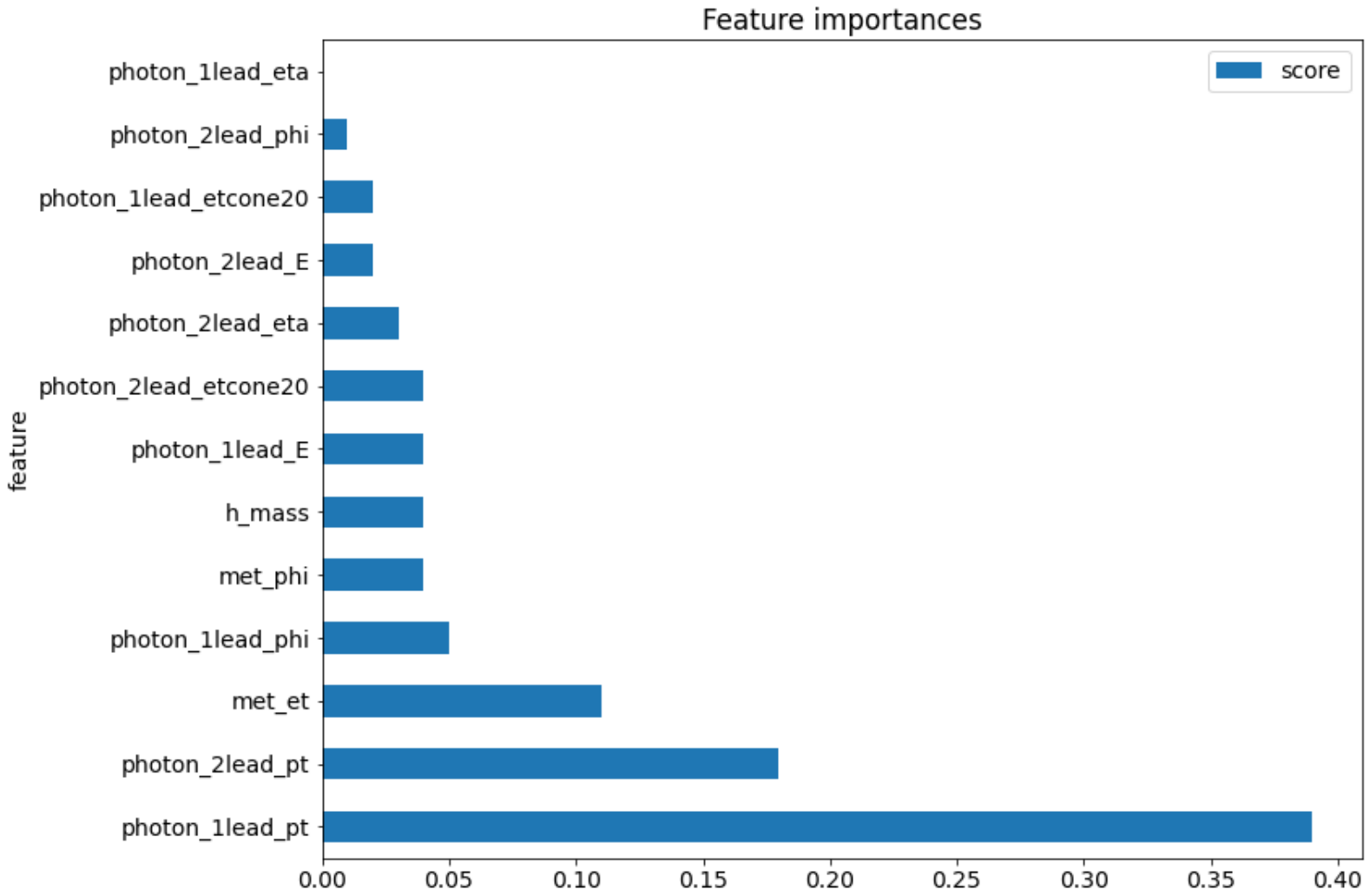
EasyEnsemble

Reference: «Bagging and Random forest for imbalanced classification»

EasyEnsemble model

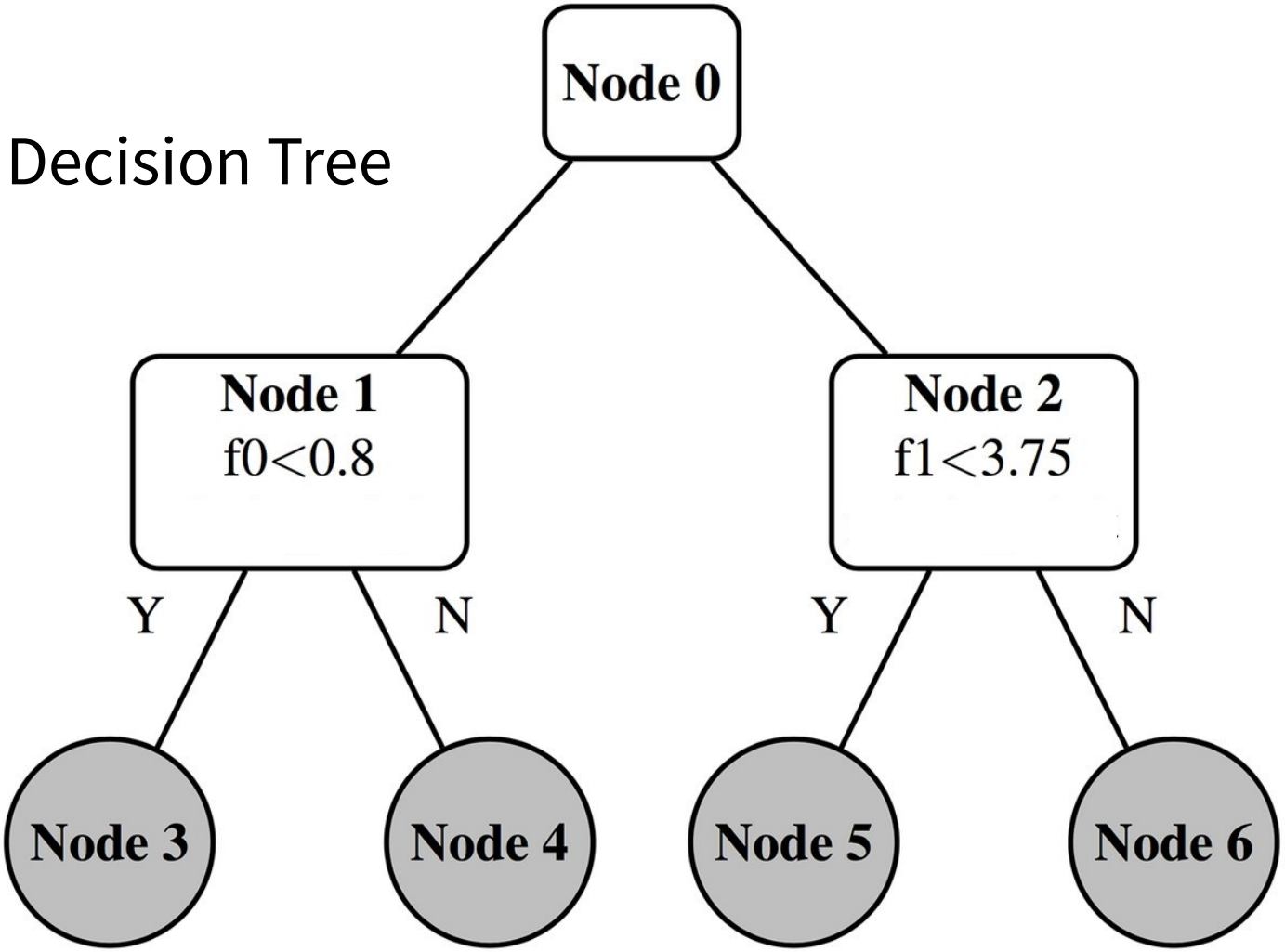


EasyEnsemble model



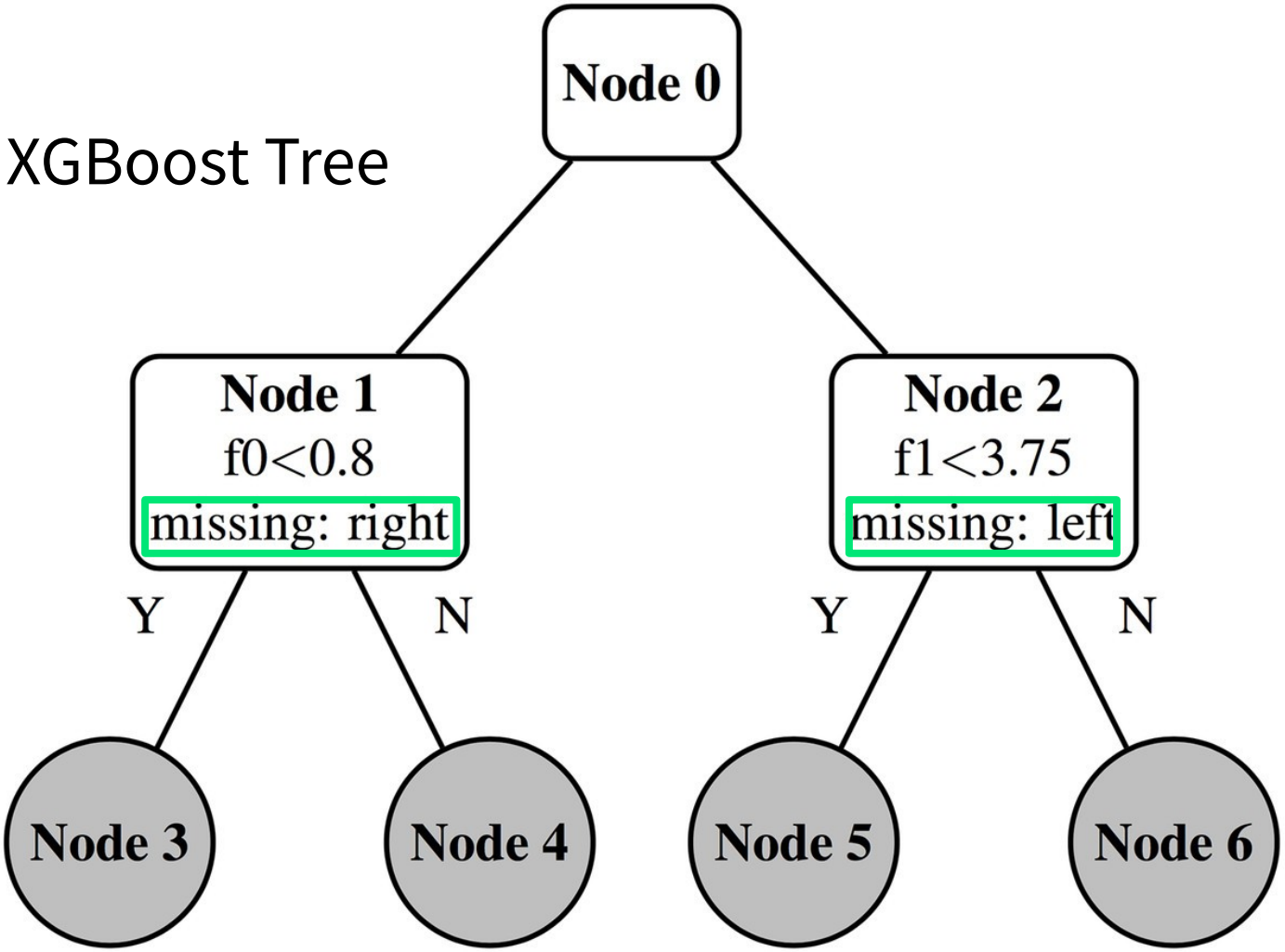
Alternative to downsampling. XGBoost

Decision Tree



Alternative to downsampling. XGBoost

XGBoost Tree

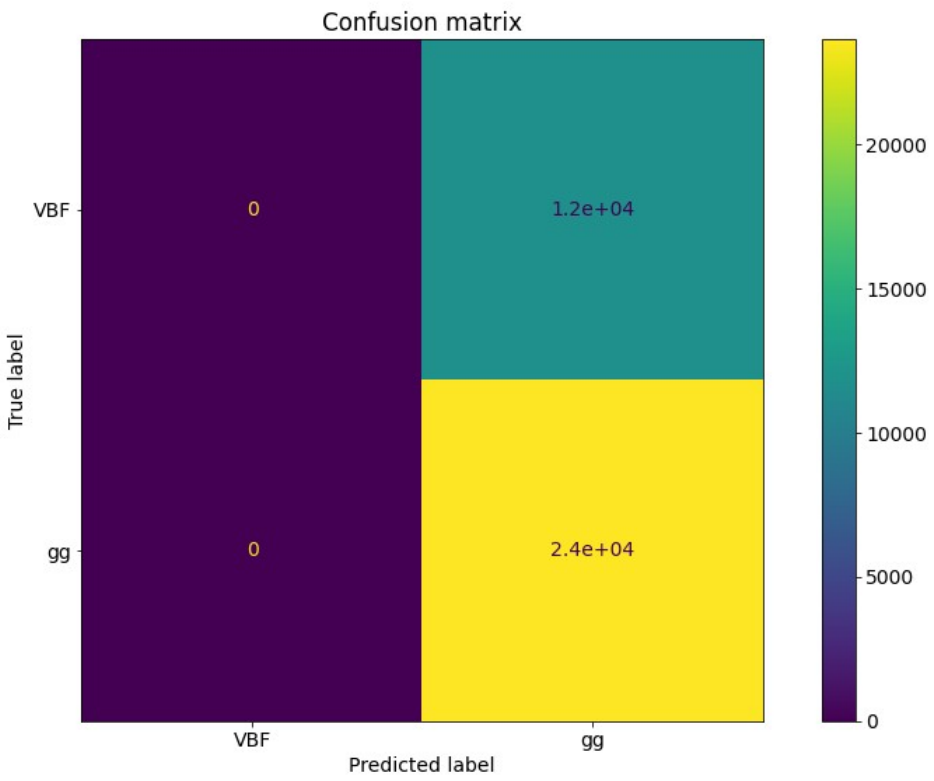


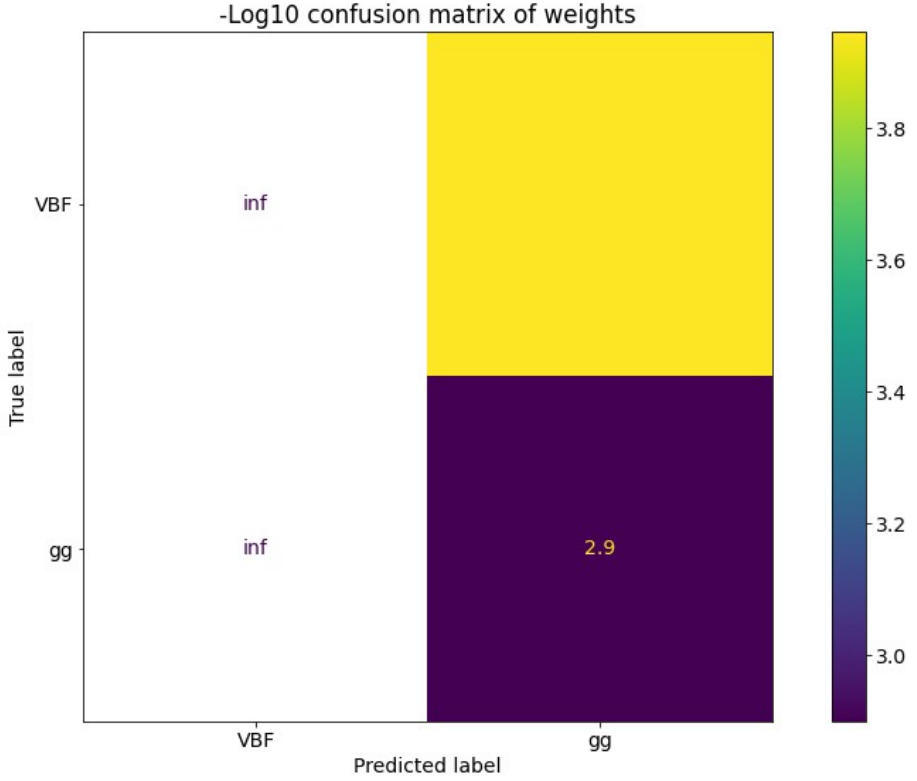
Conclusion

- It is possible to analyze HEP data with Python without loss in efficiency (in comparison to C++)
- HEP datasets are jagged. Possible solutions include:
  - feature aggregation
  - set kernels (Pyramid Match)
  - feature embedding (DeepSets or alternatives)
- HEP datasets are imbalanced. Possible solutions include:
  - downsampling
  - oversampling
  - missing data distribution fit (xgboost)

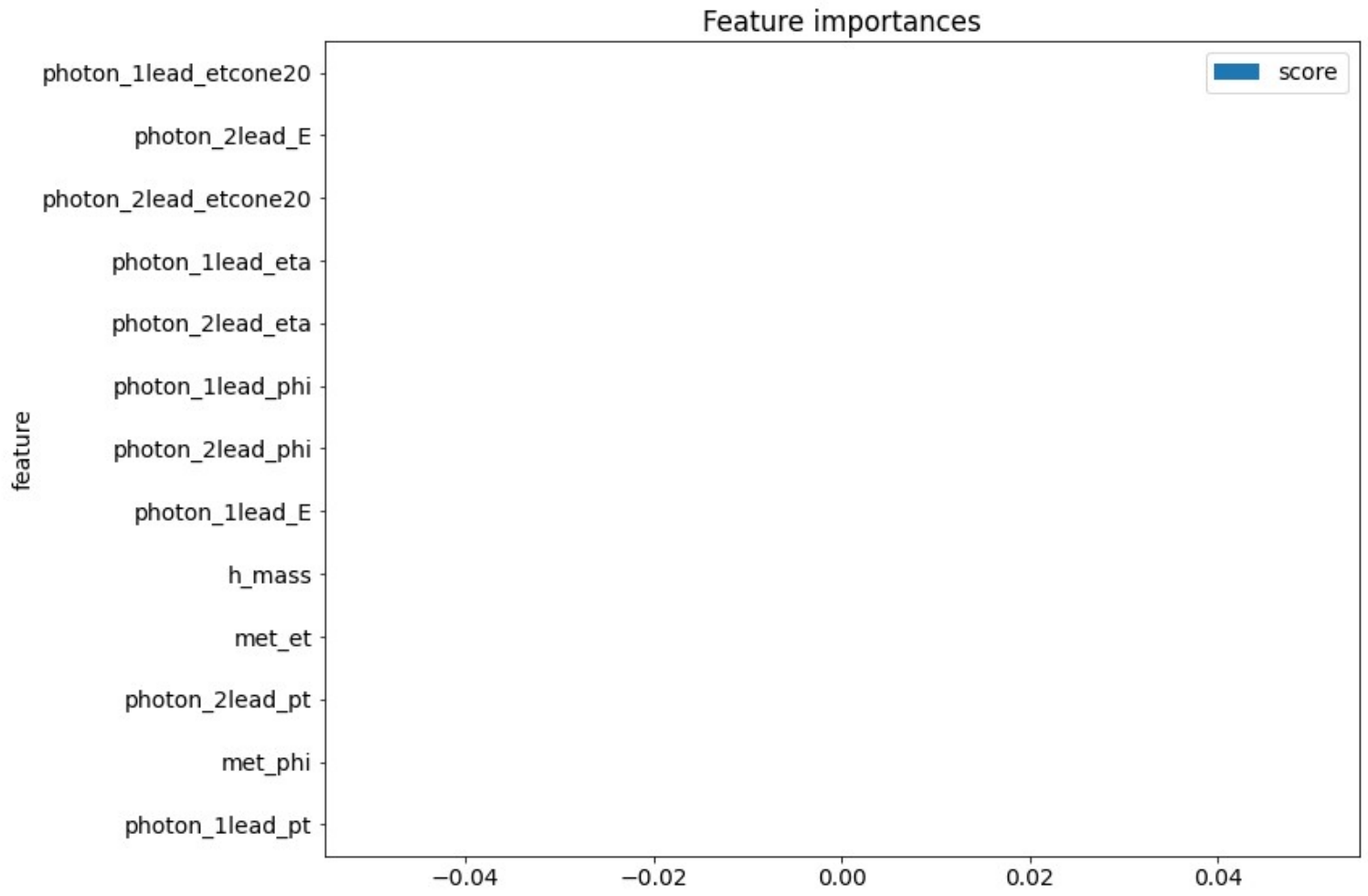
Backup

XGBoost


Confusion matrix
VBF
gg
0
1.2e+04
0
2.4e+04
True label
VBF
gg
Predicted label
20000
15000
10000
5000
0


-Log10 confusion matrix of weights
VBF
gg
inf
inf
2.9
True label
VBF
gg
Predicted label
3.8
3.6
3.4
3.2
3.0

XGBoost


Feature importances
score
photon_1lead_etcone20
photon_2lead_E
photon_2lead_etcone20
photon_1lead_eta
photon_2lead_eta
photon_1lead_phi
photon_2lead_phi
photon_1lead_E
h_mass
met_et
photon_2lead_pt
met_phi
photon_1lead_pt
feature
−0.04
−0.02
0.00
0.02
0.04