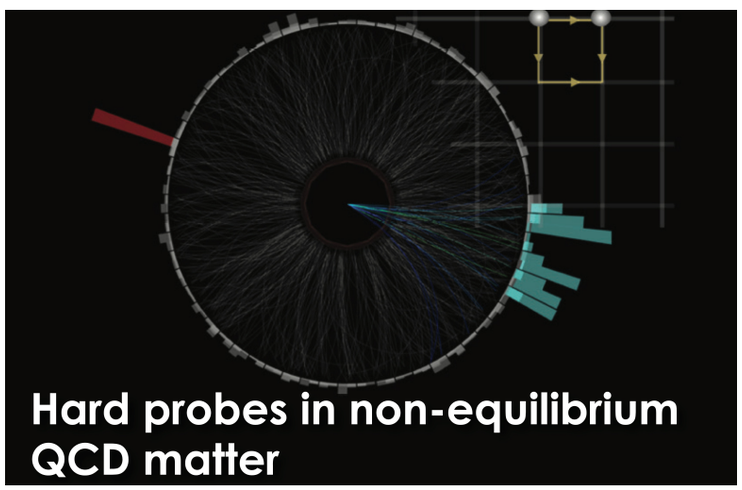
Machine Learning Applications in CMS

Arun Nayak

(Institute of Physics, Bhubaneswar)



ICTS | INTERNATIONAL CENTRE for THEORETICAL SCIENCES

TATA INSTITUTE OF FUNDAMENTAL RESEARCH

Event Selection

In most of the HEP data analysis, we try to separate events of our interest (signal) from that of the non-interesting events (backgrounds) to perform further analysis or to make some statistical statement.

For example

- Particle Identification (Separating particles in the detector)
  - Electrons from pions
  - Taus from jets
- Event classification
  - Separating Signal from Backgrounds
    - e.g. ttH, H → bb from SM ttbb

Event Selection

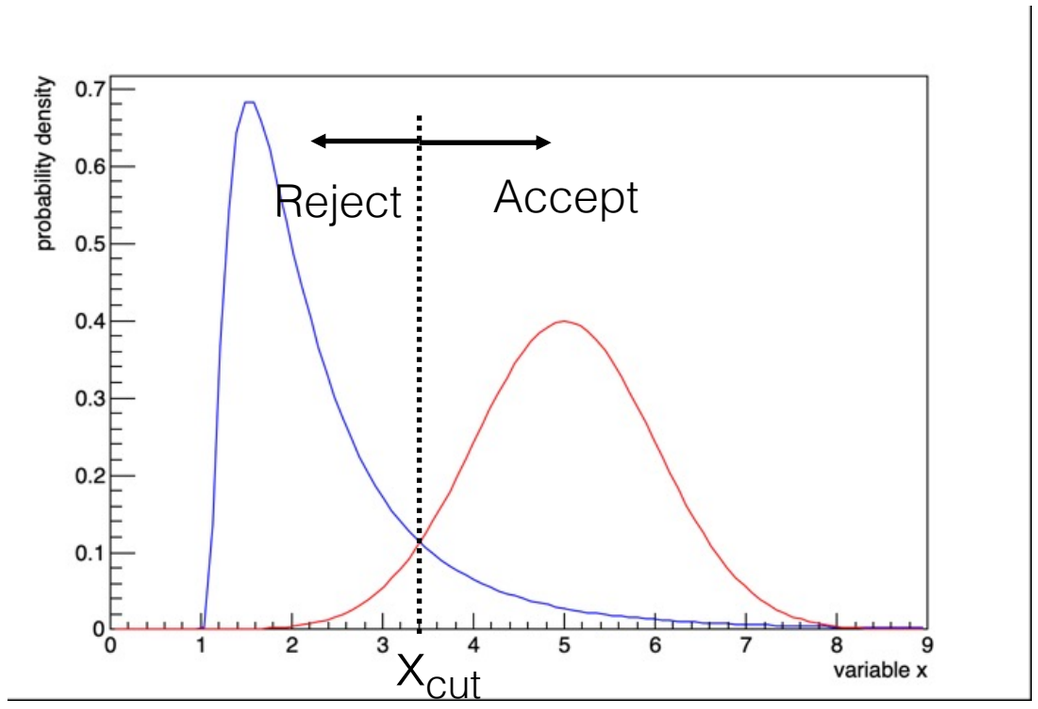
In most of the HEP data analysis, we try to separate events of our interest (signal) from that of non-interesting events (backgrounds) to perform further analysis or to make some statistical statement.

For example

- Particle Identification (Separating particles in the detector)
- Event classification

How to perform the event classification?

- Look at the distributions of some of the measured observables (variables)
  - Can also be functions of variables
- Find the observable(s) having some good discrimination between the classes
- Apply some threshold(s) (cuts) to accept / reject events
  - Cut-based selection



How to choose the cut?

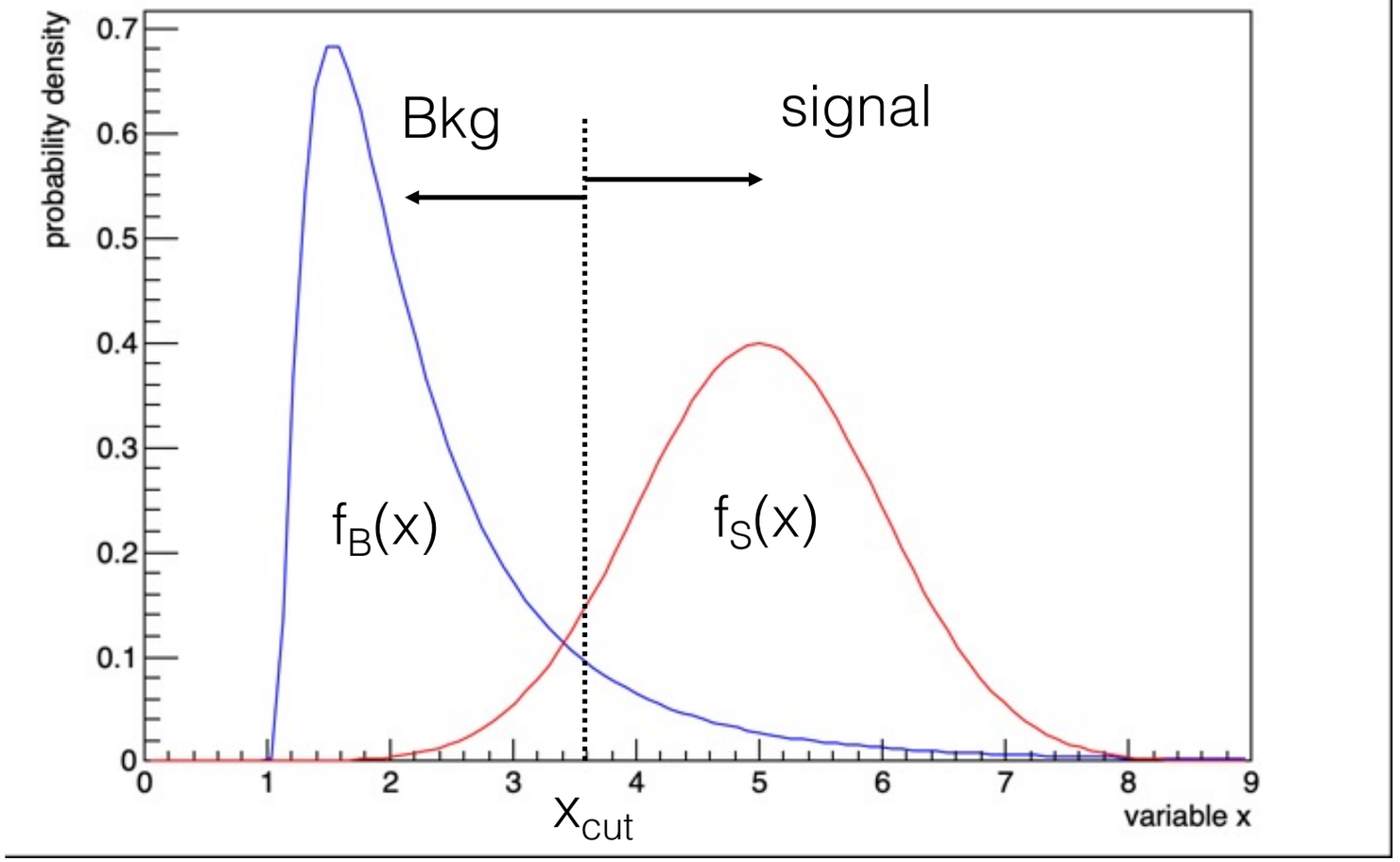
Choose x_{cut} depending on the desired "signal efficiency" ,

Signal Eff. = $\int_{x_{cut}}^{\infty} f_S(x)dx$

and background mis-identification probability

mis-ID. Prob. = $\int_{x_{cut}}^{\infty} f_B(x)dx$

$f_S(x)$ and $f_B(x)$ are signal and background probability distributions for the observable "x"



$x < x_{cut}$ → background

$x > x_{cut}$ → signal

How to choose the cut?

Choose x_{cut} depending on the desired "signal efficiency" ,

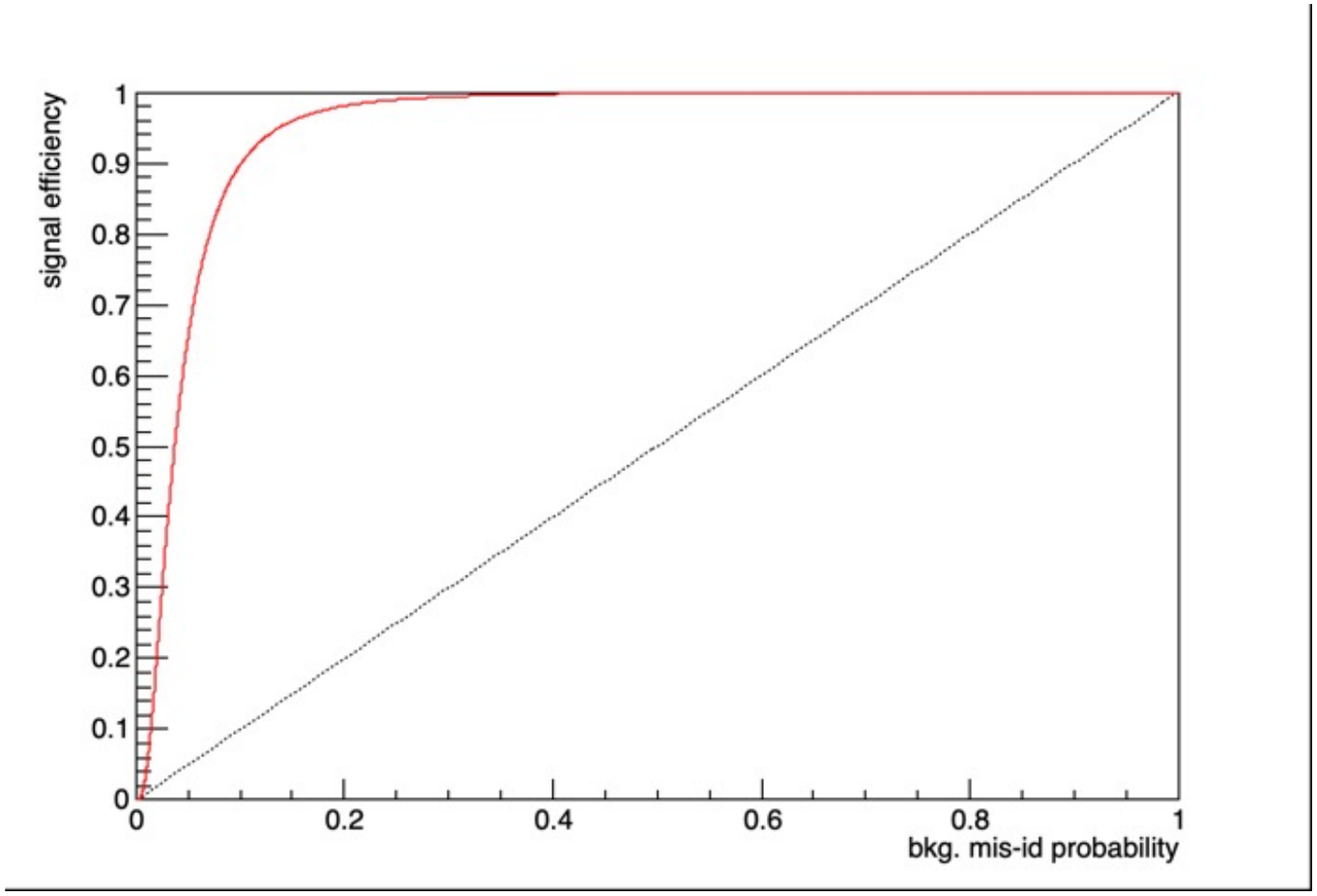
Signal Eff. = $\int_{x_{cut}}^{\infty} f_S(x)dx$

and background mis-identification probability

mis-ID. Prob. = $\int_{x_{cut}}^{\infty} f_B(x)dx$

$f_S(x)$ and $f_B(x)$ are signal and background probability distributions for the observable "x"

Scan the values of x_{cut} to get a curve of signal efficiency vs background mis-identification prob.



Receiver Operating Characteristics (ROC)

Area Under the Curve (AUC) provides the discrimination power of the observable

Suppose the relative fraction of signal and backgrounds are not known, the problem becomes that of a parameter estimation:

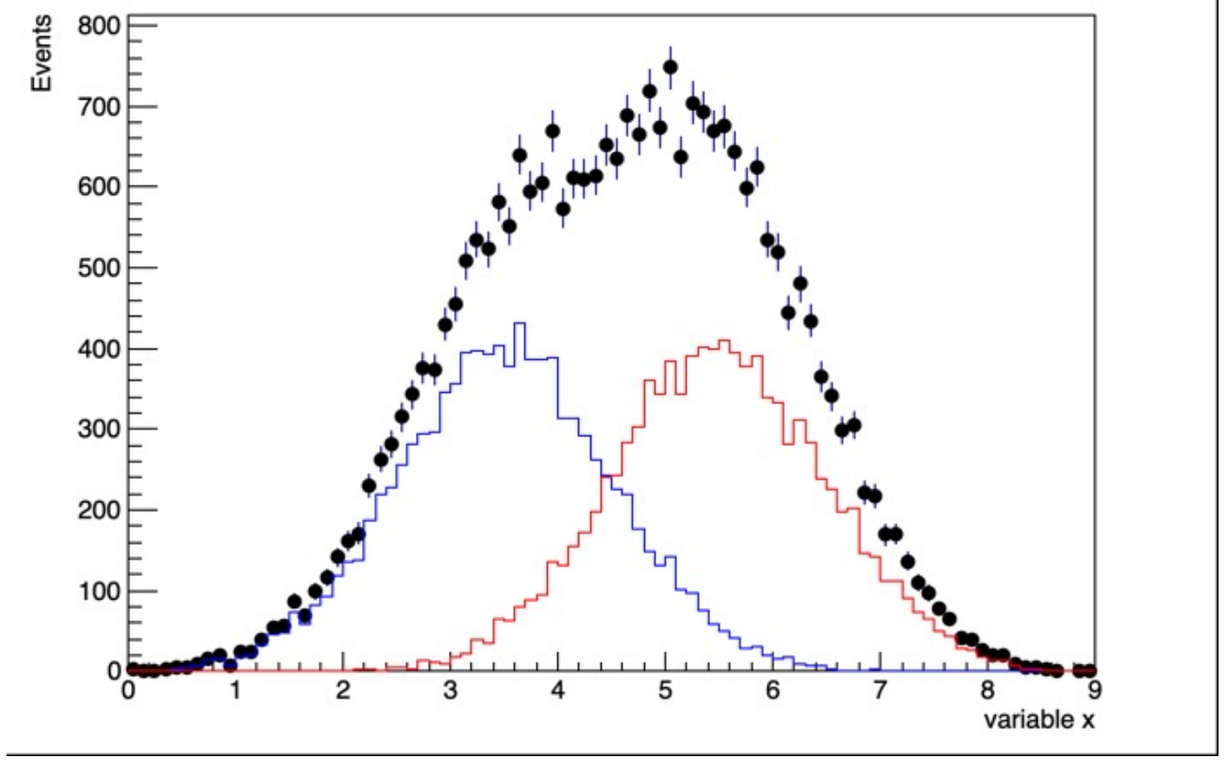
The observed distribution in data can be written as

$$f(x, \alpha_s) = \alpha_s f_S(x) + (1 - \alpha_s) f_B(x)$$

α_s and $\alpha_B = 1 - \alpha_s$ are fractions of signal and backgrounds, respectively.

Fit to data and estimate α_s.

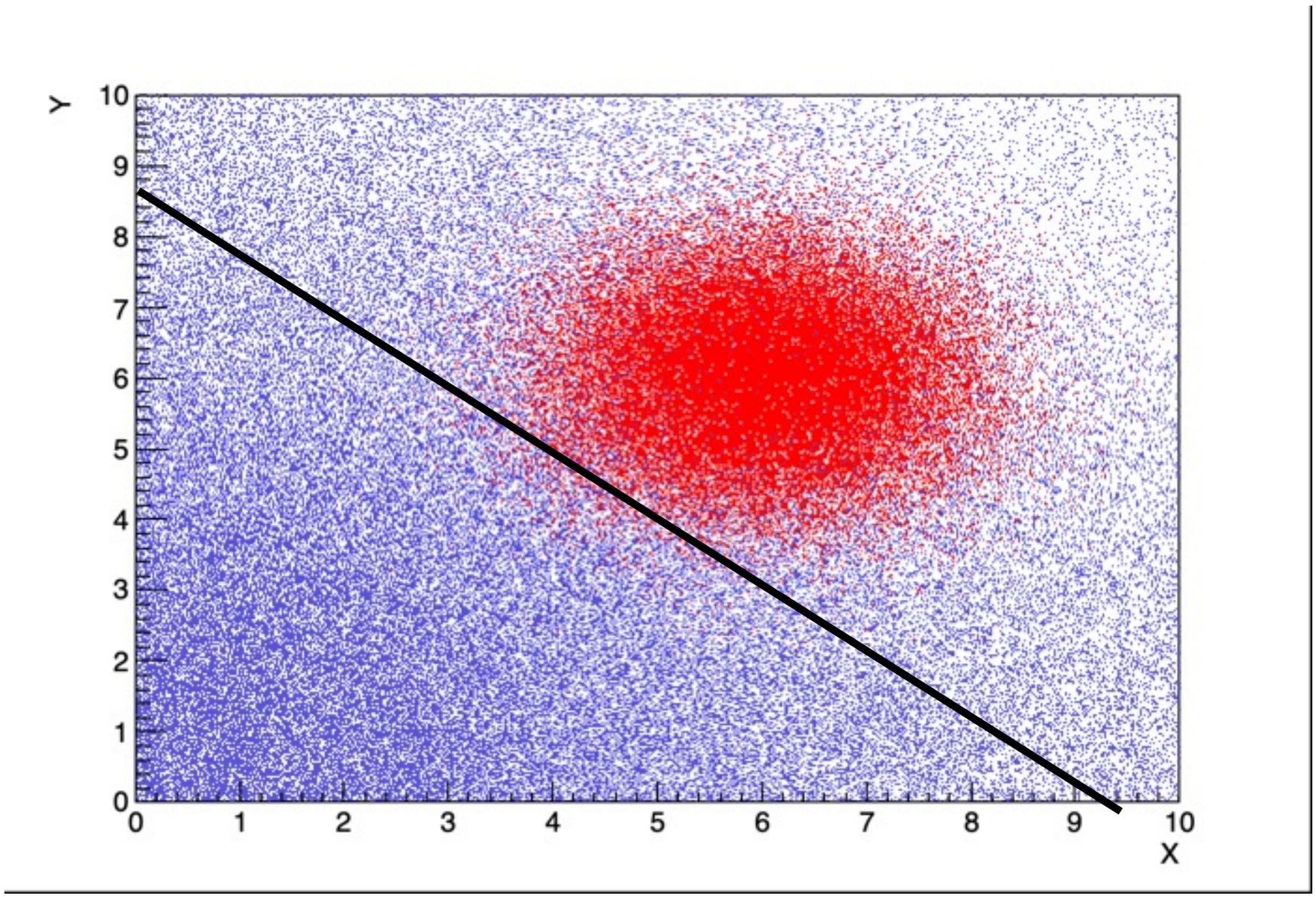
No. of signal Ns in total number of events N_{tot}, $N_s = \alpha_s N_{tot}$



Better Separation $\Longrightarrow$ Better estimation of α_s

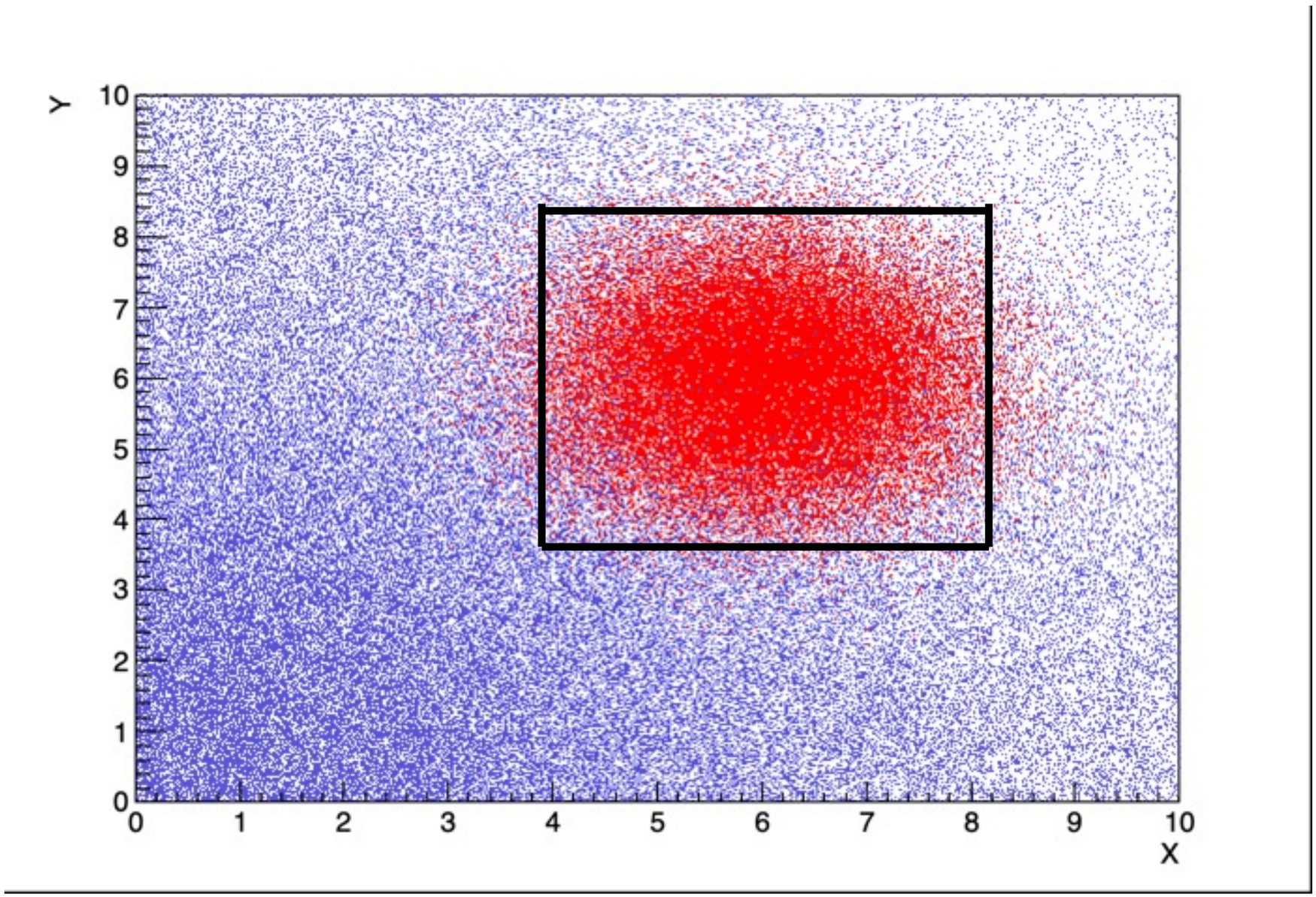
Is the simple threshold cut optimal?

Simple cut-based selections may not be optimal for more than 1 dimensions



Is the simple threshold cut optimal?

Simple cut-based selections may not be optimal for more than 1 dimensions



What is the optimal selection?

The selection performance can be considered optimal if it achieves the selection that corresponds to the largest possible signal efficiency for a fixed mis-identification probability

The optimal classifier is a cut on the **likelihood ratio (LR)**

$$LR(\vec{x}) = \frac{f_S(\vec{x})}{f_B(\vec{x})}$$ (Neyman-Pearson Lemma)

For >1d:

If the variables are independent,

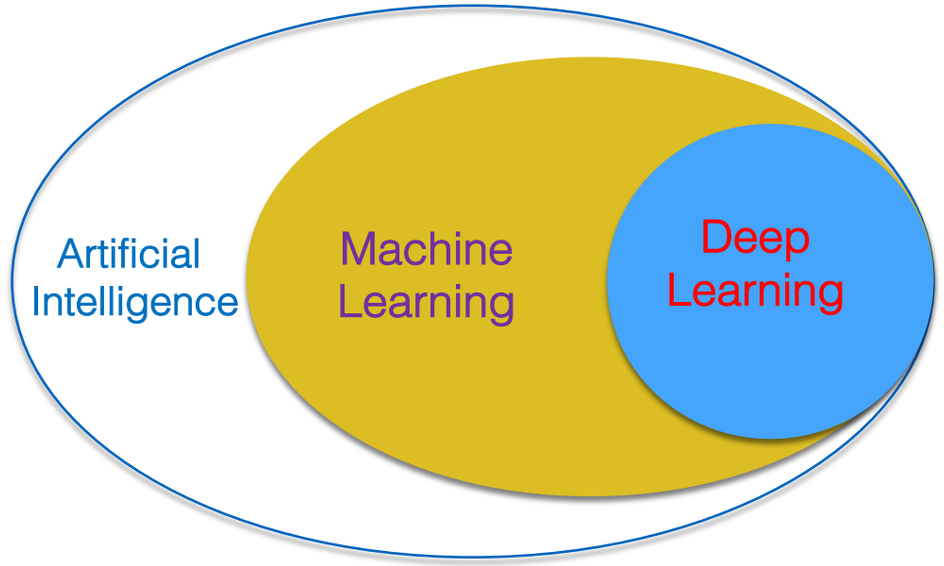
$$f(\vec{x}) = f(x_1, \ldots, x_n) = \prod f_i(x_i)$$

If variables are correlated, it may not be easy to obtain the likelihood distribution

Machine learning for classification helps in estimation of the LR using a limited training examples

Machine Learning

A type of artificial intelligence that allows computers to learn from data without being explicitly programmed



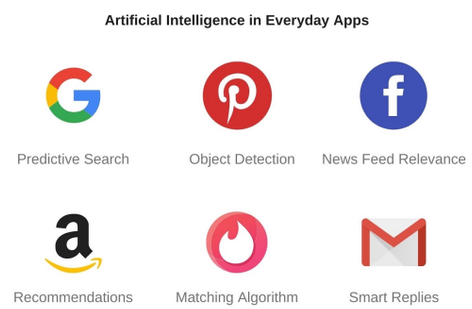
How does it work?

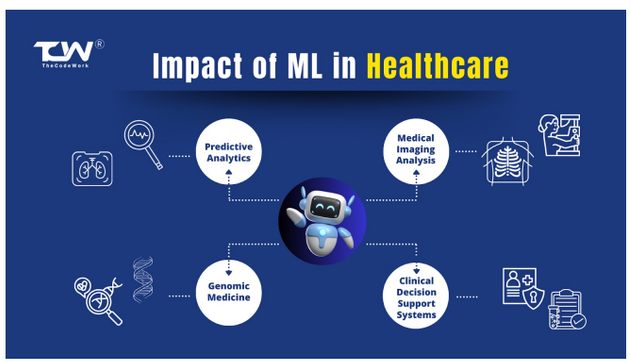
- Design a model (to represent something in the world)
    - Learn the parameters of the model through data
    - Apply to make predictions

Accuracy depends on the network architecture and the amount of data used for training

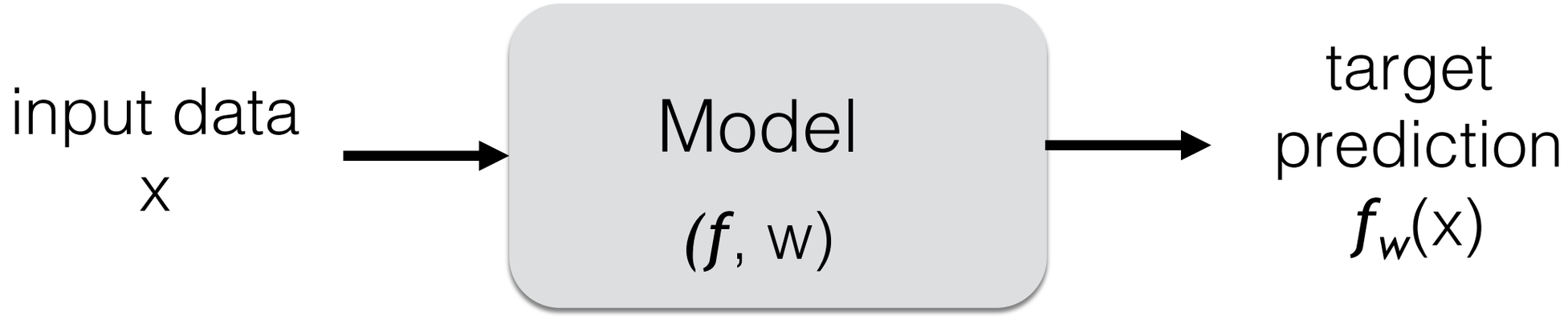
Wide ranging applications in everyday life:

From simple google search to advancement of healthcare





Supervised Learning

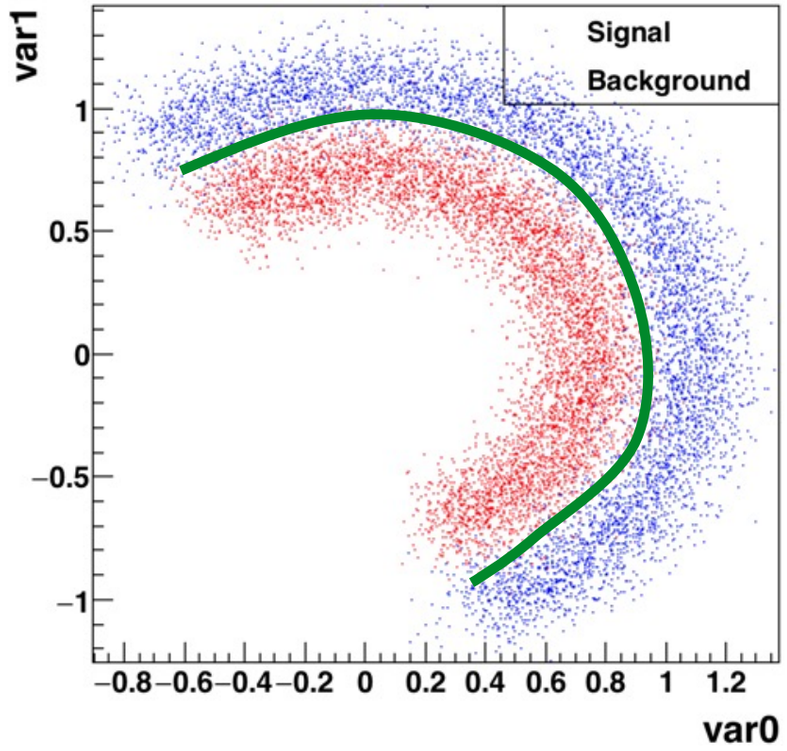


Given an independent, identically distributed (i.i.d) dataset $\{(x_1,y_1), \ldots, (x_n,y_n)\}$, find w that minimizes Loss: $L\,(w) = \frac{1}{n}\sum_i L(f_w(x_i), y_i)$

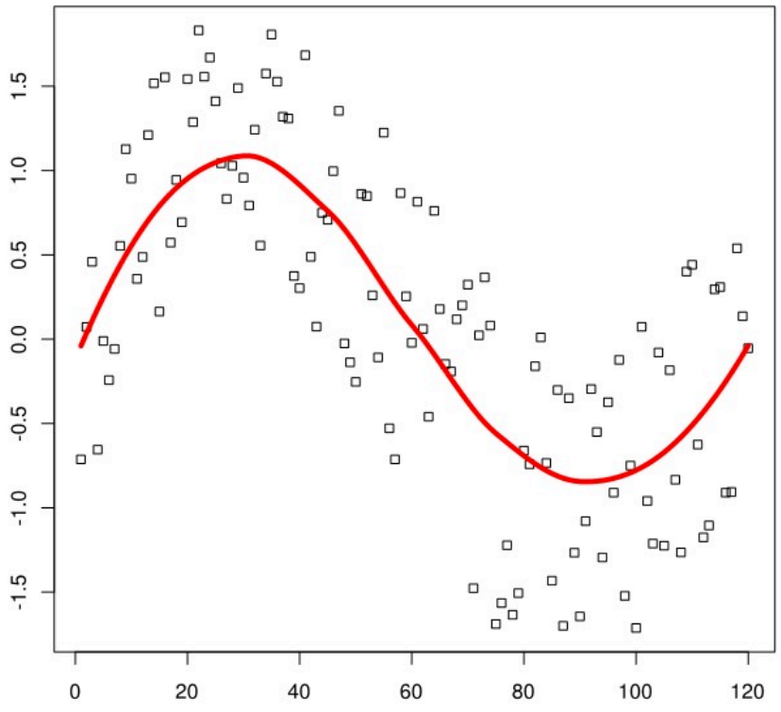
Where $f_w(x_i)$ are the predictions and y_i are the truths.

Interest in HEP

- **Classification:** The mapping function describes a decision boundary

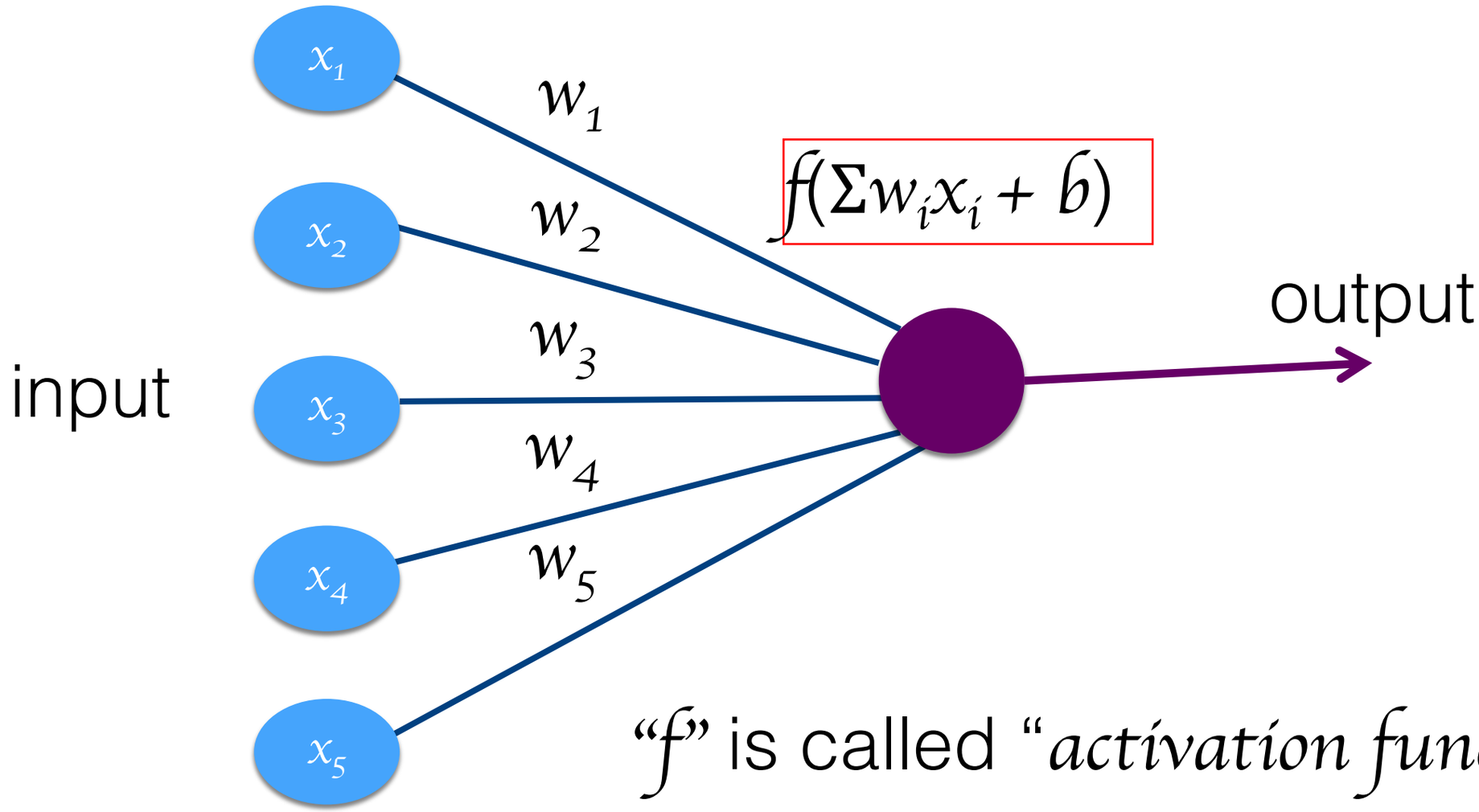


- **Regression:** The mapping function describes an approximation of the underlying functional behaviour defining the target value



Artificial Neuron

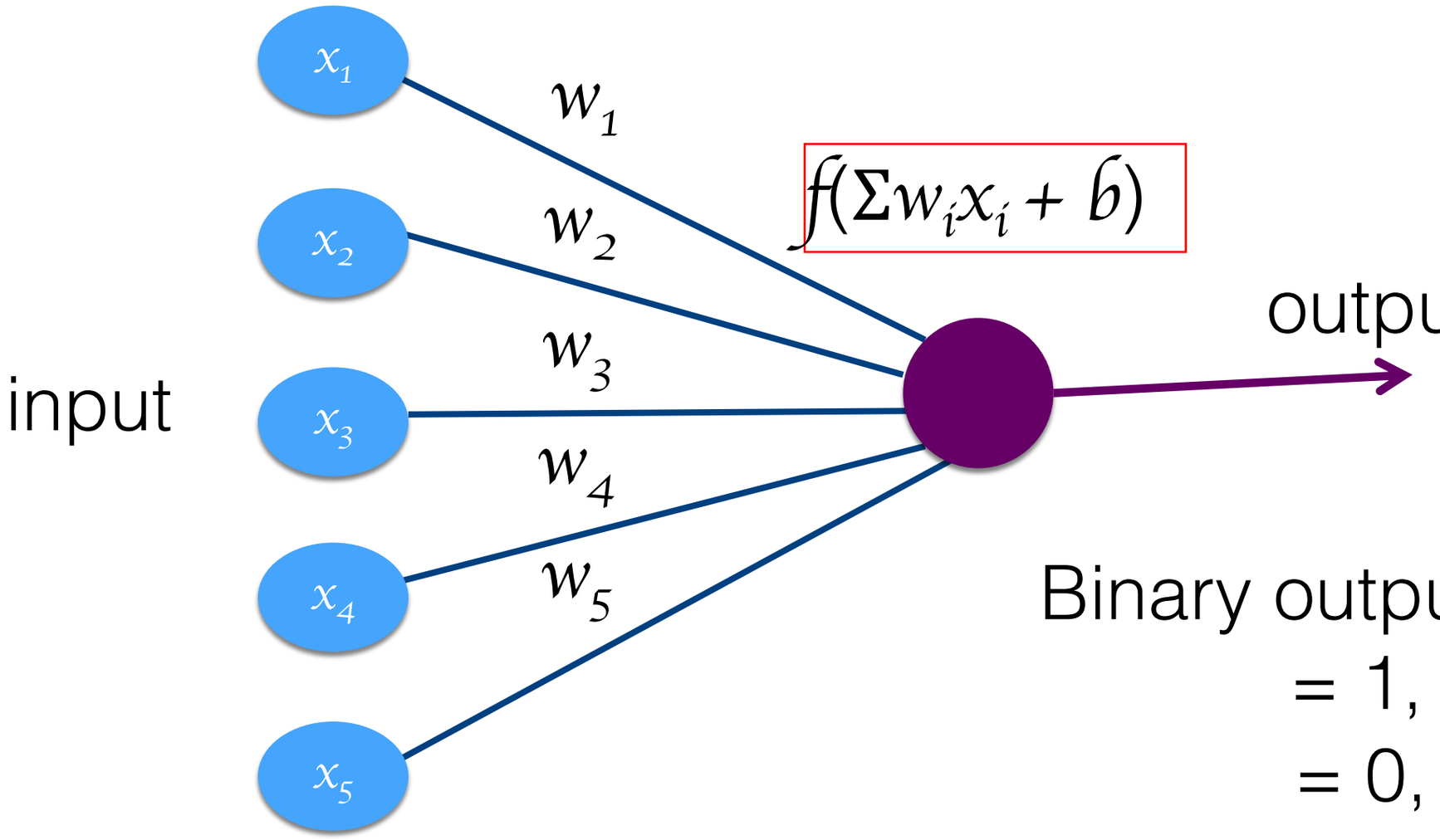
Model of an artificial neuron

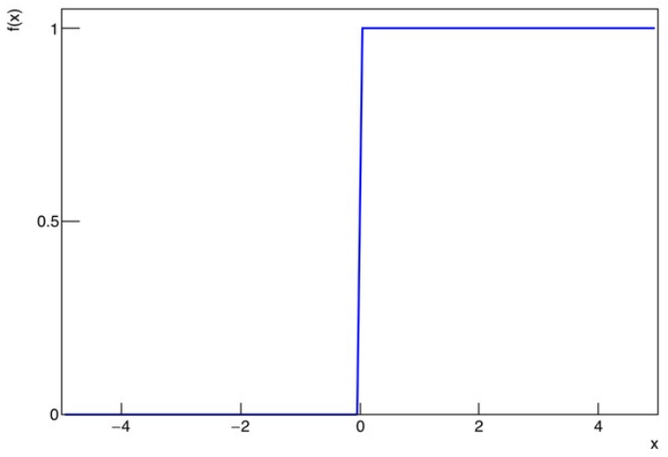


"f" is called "*activation function*"

b is a bias term → Can be represented by a node with input "1".

Perceptron



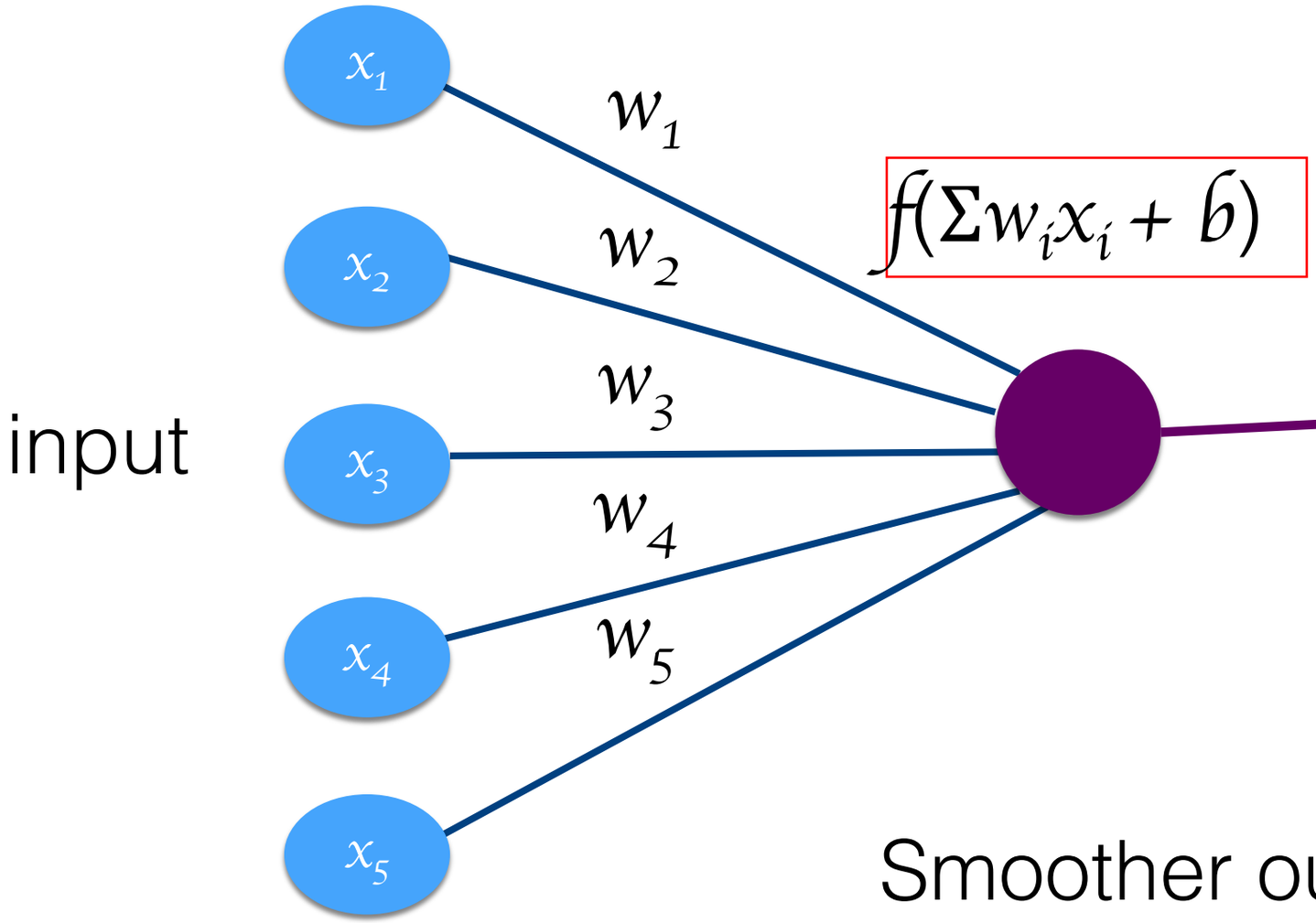


Binary output:

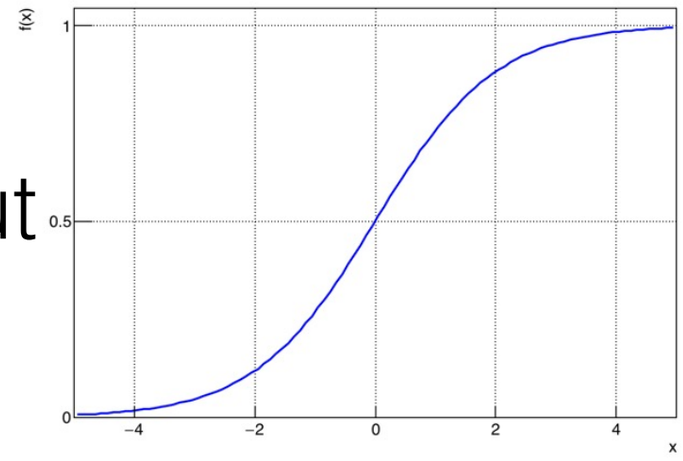
$= 1$, if $\Sigma w_i x_i + b > 0$

$= 0$, if $\Sigma w_i x_i + b < 0$

Sigmoid / logistic Neuron



$$f(x) = \frac{1}{1+e^{-x}}$$



Smoother output

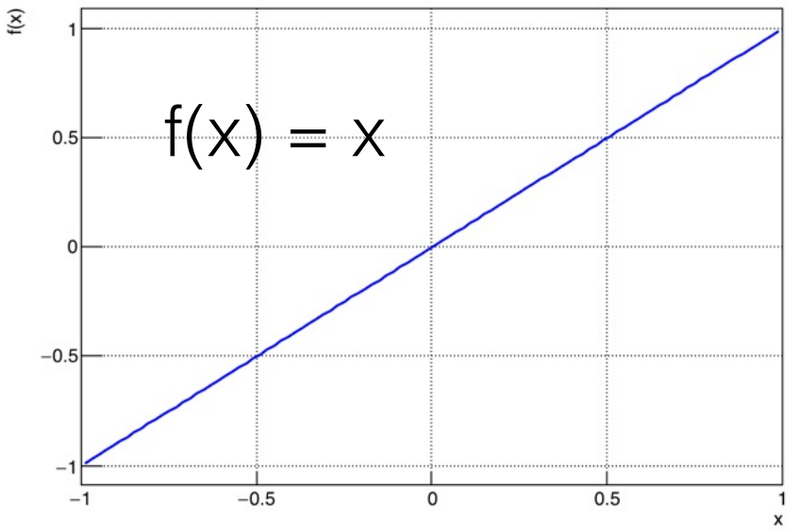
Small change in weight => small corresponding change in output

This property makes the learning possible

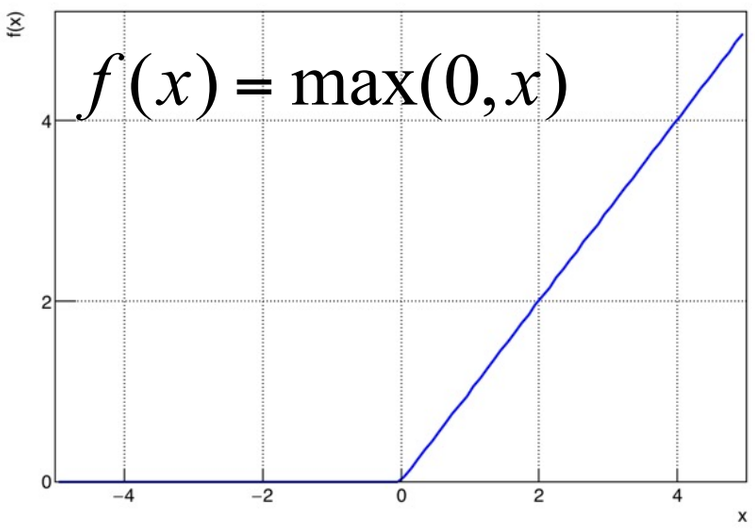
Activation Functions

It determines at what threshold the neuron will fire
OR the frequency at which a neuron fires

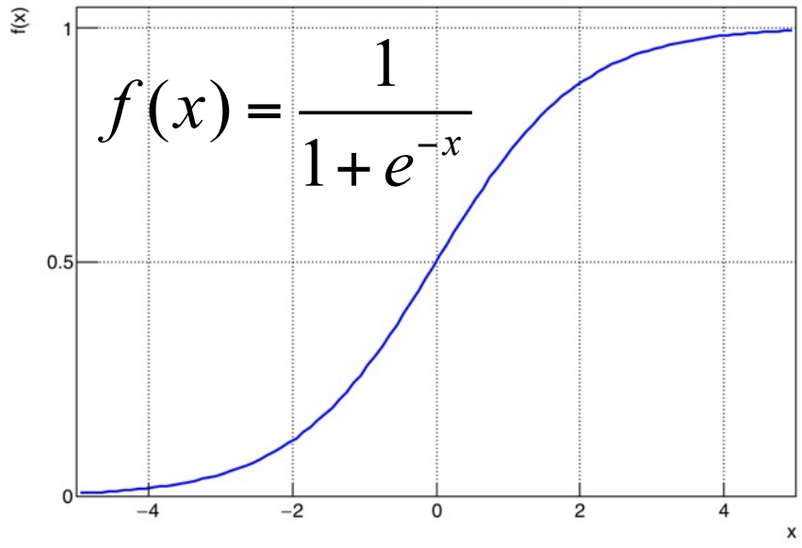
Linear Function



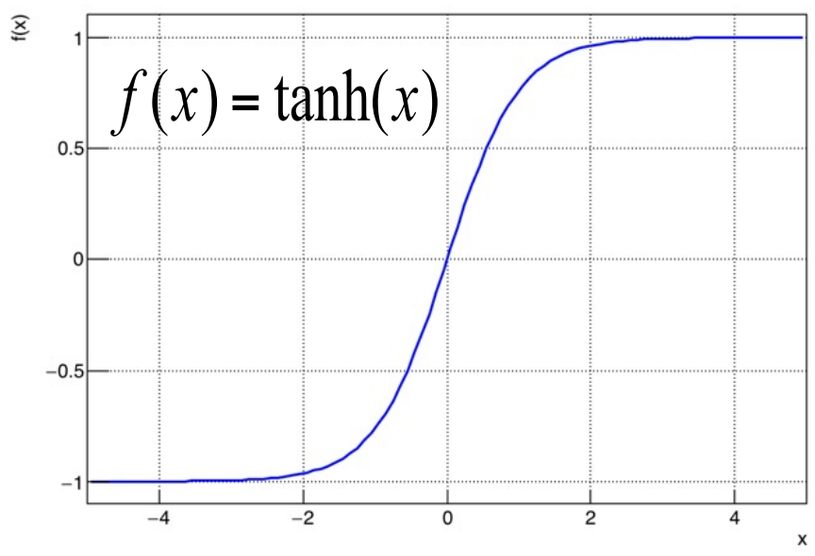
Rectified Linear Units (ReLU)



Sigmoid Function



Hyperbolic Tangent

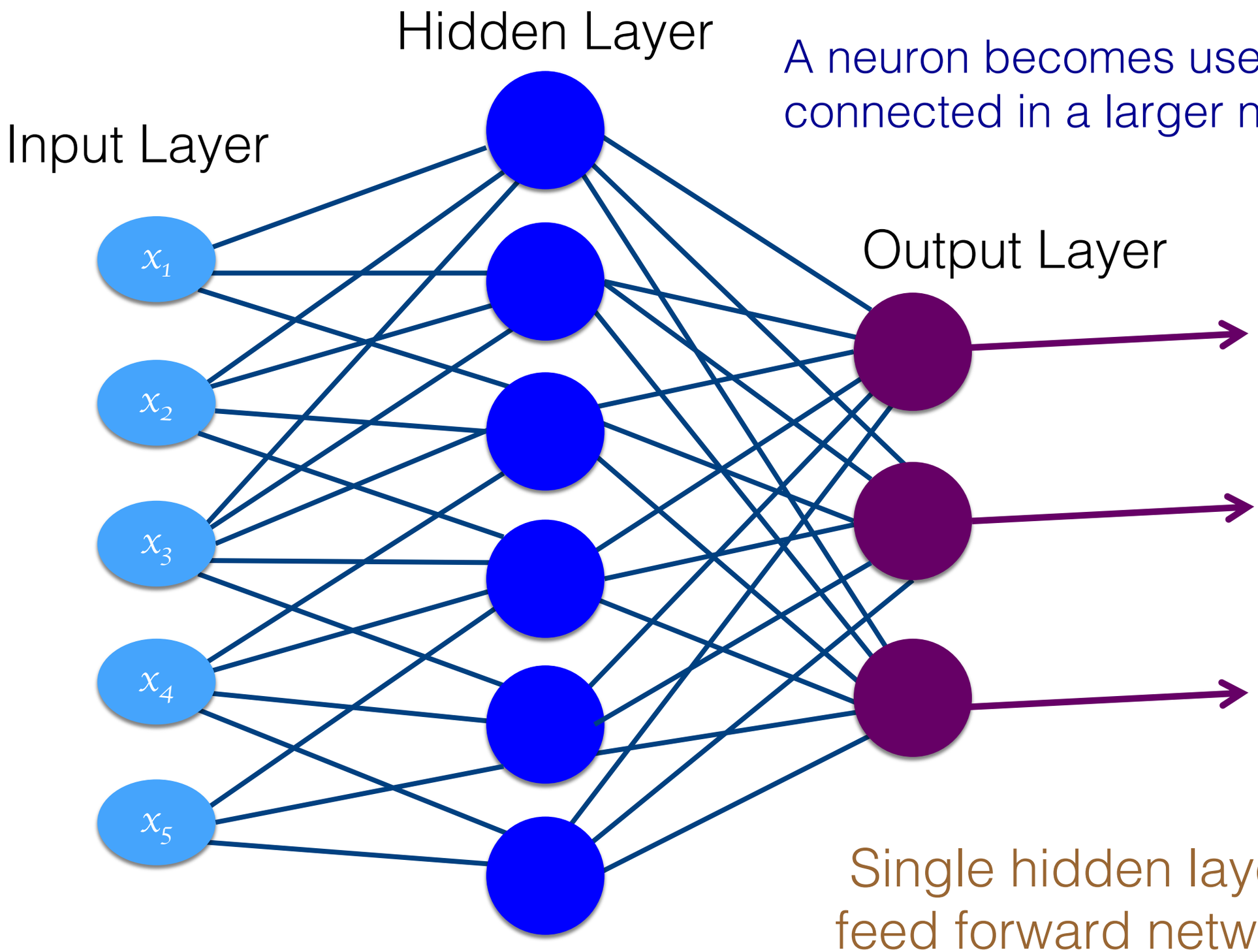


Softmax

$$f(x_i) = \frac{e^{x_i}}{\sum e^{x_i}}$$

(used in output layer of a multiclassification network)

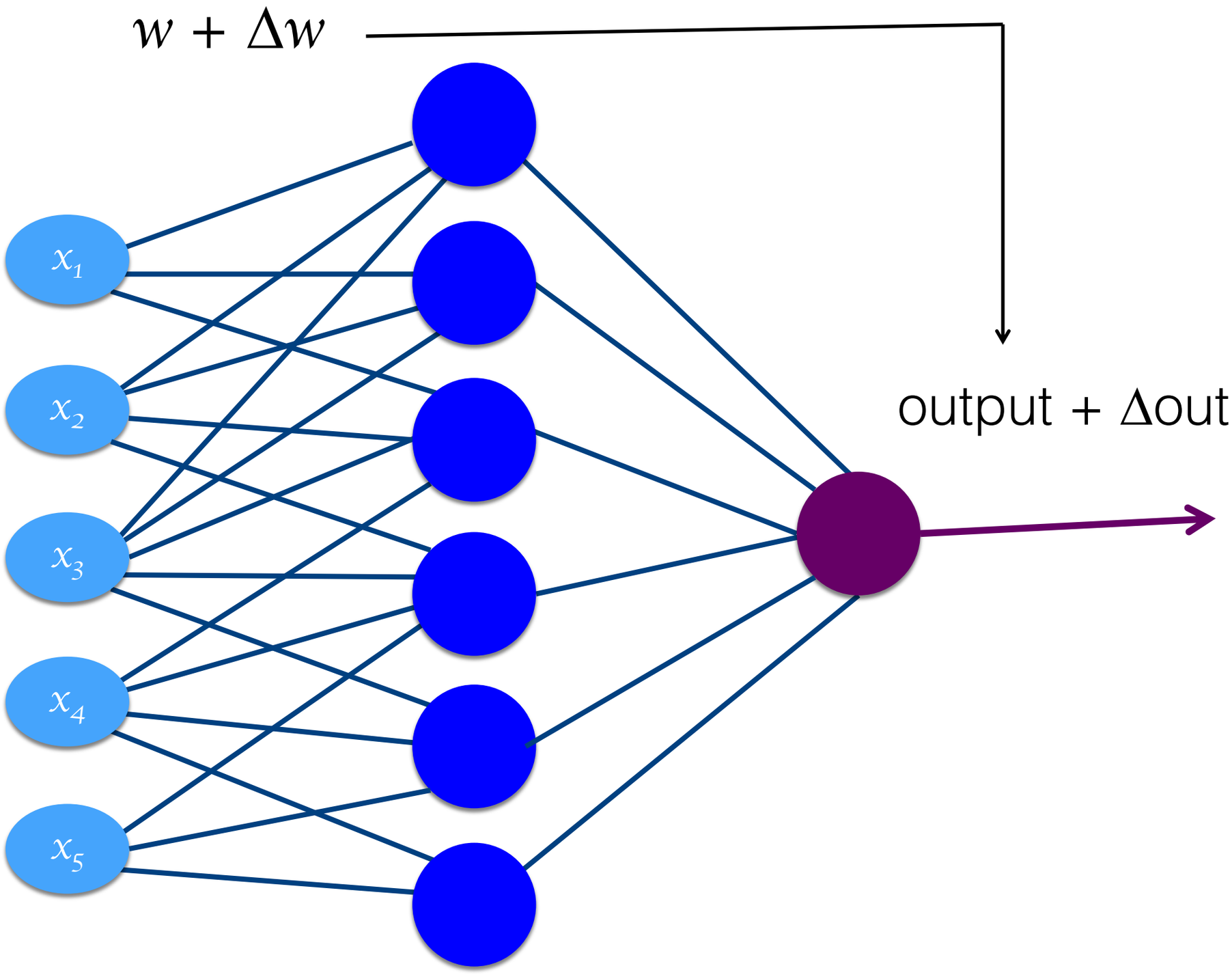
Feedforward Neural Network



A neuron becomes useful when connected in a larger network

Single hidden layer feed forward network

Training a Network



Loss Function

Loss : measure of misclassification

1. Mean Squared Error: $$L(w,b) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N}\left(y_{ik} - \hat{y}_{ik}\right)^2$$

2. Cross Entropy: $$L(w,b) = -\sum_{k=1}^{K}\sum_{i=1}^{N} y_{ik}\log\hat{y}_{ik}$$

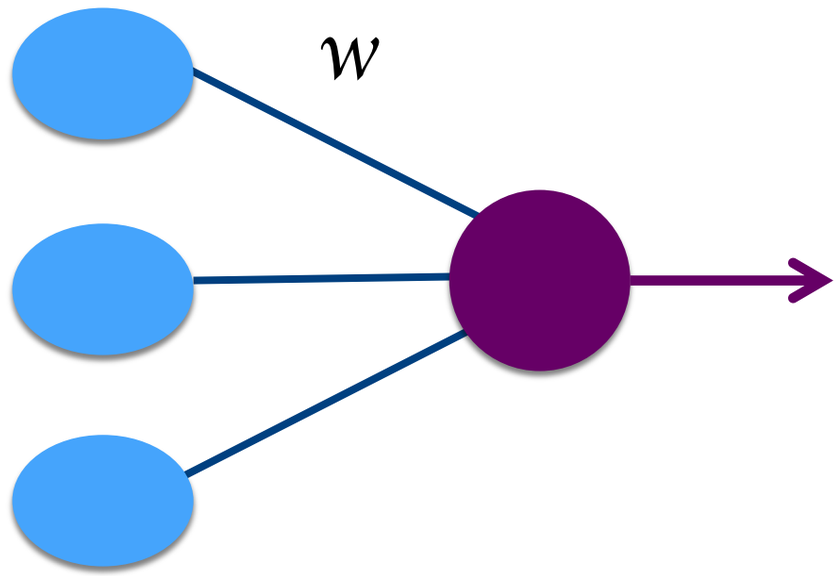
Where, index "i" is for events and "k" is for output nodes

Network Training => Minimizing Loss in an iterative way

Backpropagation

Let's consider only one output node in the network and MSE loss function

$$L(w,b) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$$



L is a function of weights and biases

→ A hypothetical surface in weight space

In every iteration, weights should be changed in a direction such that L is reduced (i.e. ΔL is –ve)

$$\Delta L \approx \sum \frac{\partial L}{\partial w_j}\Delta w_j = \nabla L \bullet \Delta w$$

Gradient descent

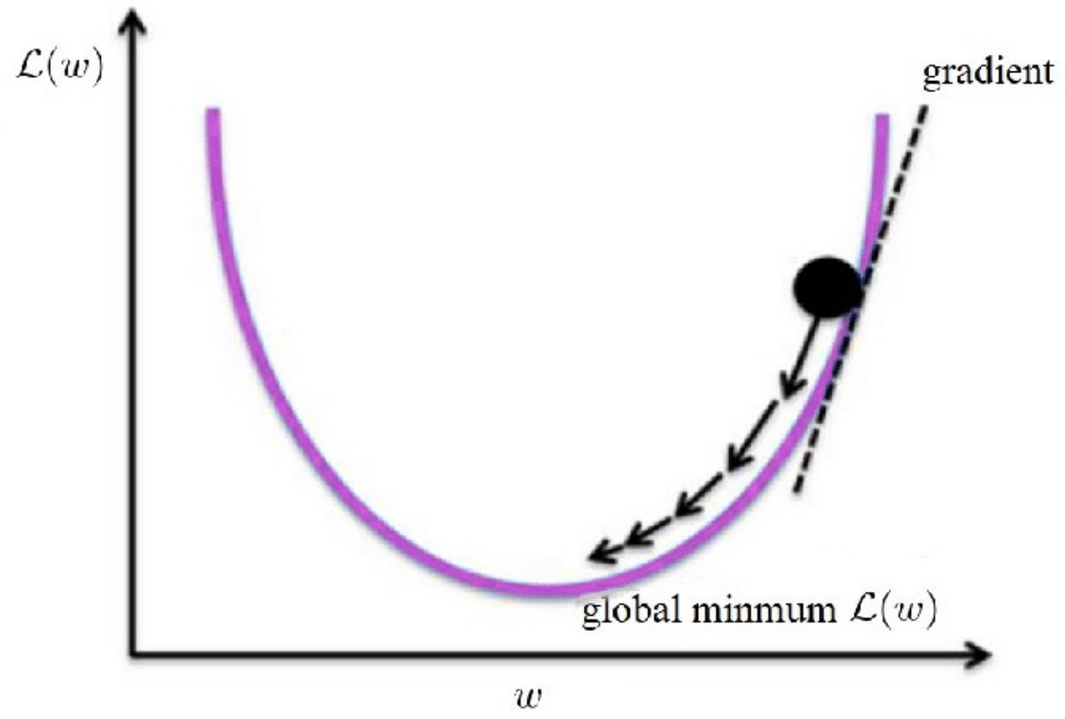
If we choose $\Delta w = -\eta \nabla L, \Rightarrow \Delta L \approx -\eta \,\|\nabla L\|^2 \Rightarrow \Delta L < 0$

Gradient descent

Starting from an arbitrary weight vector, the weights are updated after every iteration “t”, till the loss is reduced to an accepted value

$$w_j(t+1) = w_j(t) - \eta \frac{\partial L}{\partial w_j}(t)$$

η is called the learning rate

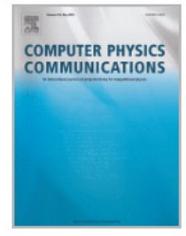


Machine Learning is not new to HEP

ELSEVIER

Computer Physics Communications

Volume 119, Issues 2–3, 2 June 1999, Pages 219-231



Neural networks in high energy physics: A ten year perspective

Bruce Denby

- Use of multivariate methods since 1980s
  - But, mostly shallow neural networks, boosted decision trees
  - Played significant role in Higgs boson discovery at LHC

Deep Learning Revolution

July 2012

CERN ABOUT NEWS

Voir en français

CERN experiments observe particle consistent with long-sought Higgs boson

4 JULY, 2012

Geneva, 4 July 2012. At a seminar held at CERN[1] today as a curtain raiser to the year's major particle physics conference, ICHEP2012 in Melbourne, the ATLAS and CMS experiments presented their latest preliminary results in the search for the long sought Higgs particle. Both experiments observe a new particle in the mass region around 125-126 GeV.

January 2012

Imagenet classification with deep convolutional neural networks

Authors	Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton
Publication date	2012
Journal	Advances in neural information processing systems
Volume	25

IMAGENET

Nobel Prize in Physics 2024

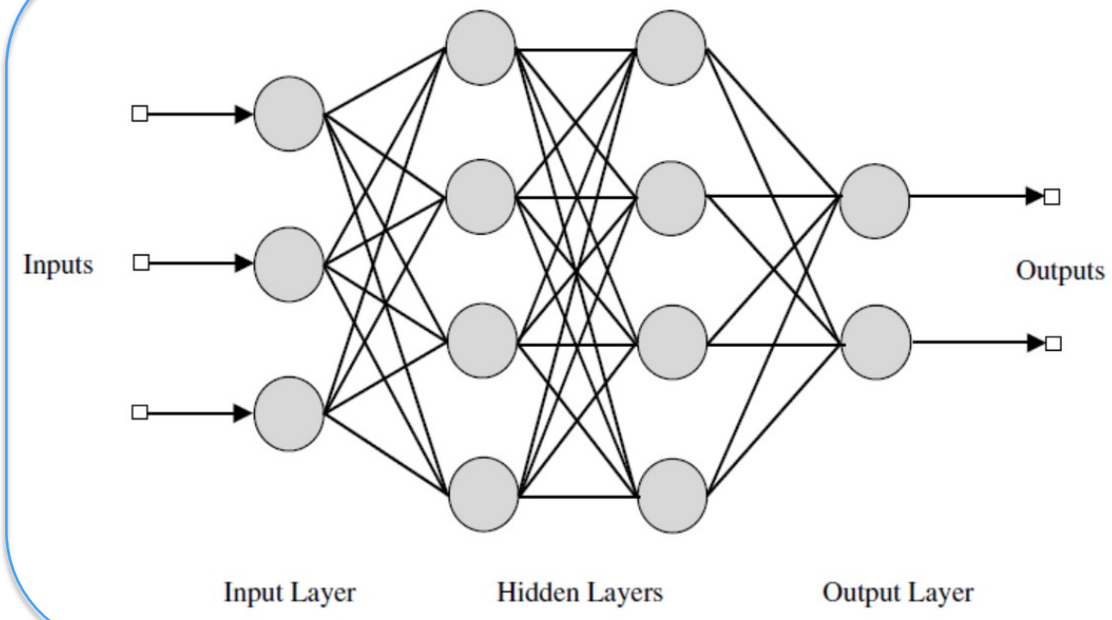
The emergence of AlexNet (a Deep CNN) revolutionized the deep learning applications

A large varieties of deep learning techniques developed over last 10 – 12 years:

- CNN, RNN, GAN, GNN etc…., with wide range of applications.
  - Driven by advanced ML architectures, increased computing resources, GPUs, availability of large amount of data

ML Applications @ LHC Experiments

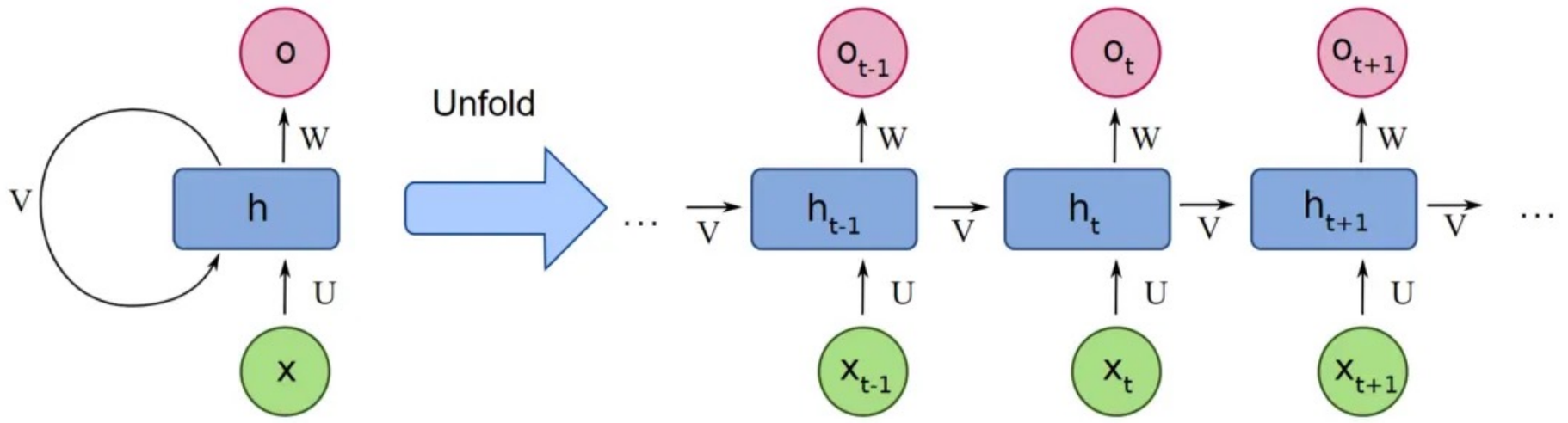
	Run-1	Run-2	Run-3
Architectures	BDTs, DNNs	CNNs, GraphNN	Transformers, Diffusion models, Normalizing Flows, Unsupervised models
Applications	Classification, Regression	Reconstruction, Calibration, High Level Trigger, Data Quality monitoring	Data/MC corrections, Fast Simulations, Anomaly detection, Level-1 Trigger (FPGA), ML based particle flow



Feed forward neural network

Universal function approximator

No inherent input invariance



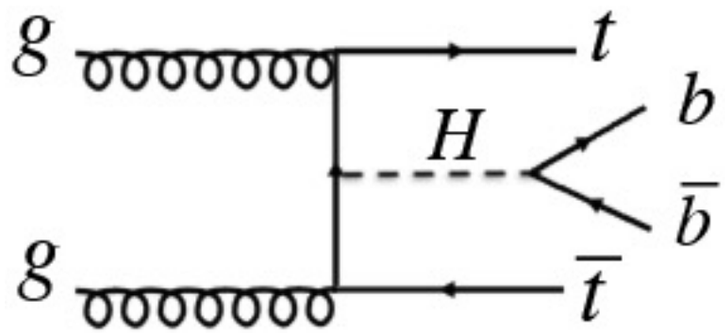
Recurrent NN: used for sequential data or time series data

NN in event classification

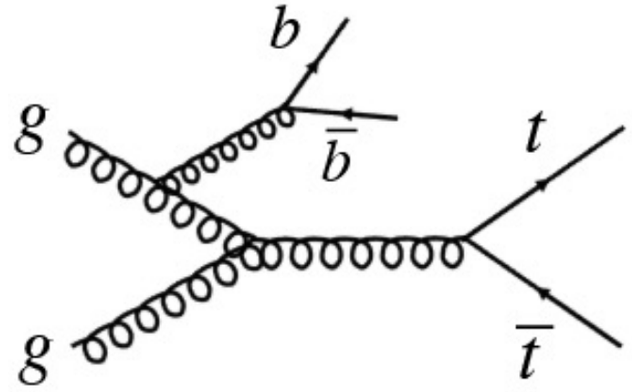
JHEP 03 (2019) 026

Separate tt+Higgs (signal) from tt+jets (background)

Signal



Background



Single lepton

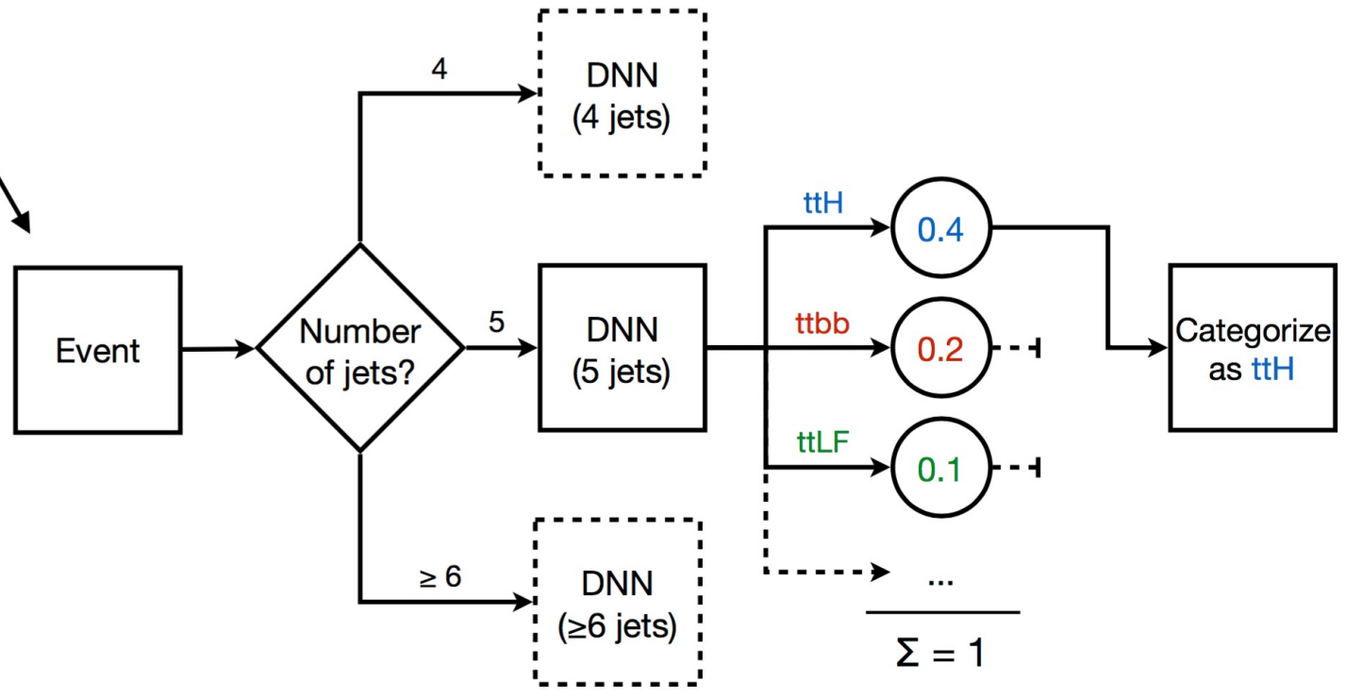
4j, ≥3b | 5j, ≥3b | ≥6j, ≥3b

DNN (MEM is input)

Categorise by most probable process
tt̄H, tt̄+bb̄/b/2b/cc̄/lf

4j, ≥3b, *tt̄H node* | 5j, ≥3b, *tt̄H node* | ≥6j, ≥3b, *tt̄H node*

3 x 6 categories



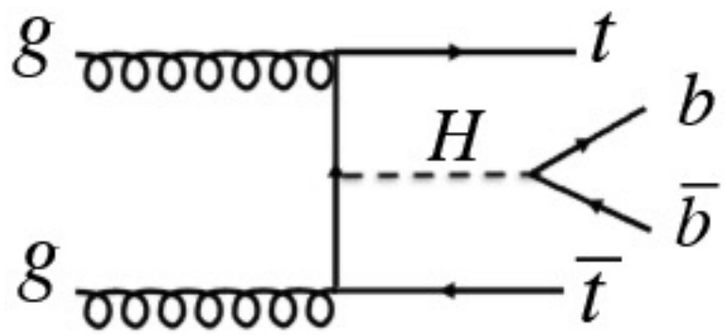
One of the most complex final state

NN in event classification

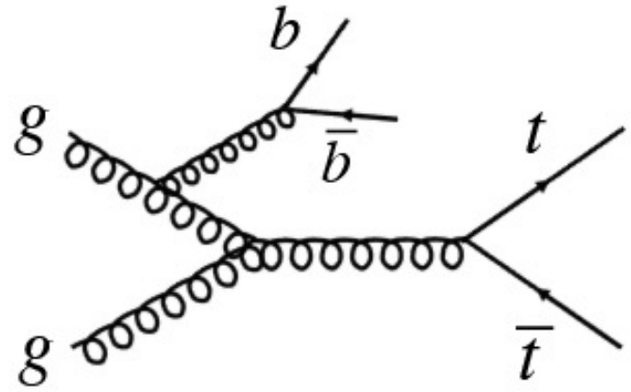
JHEP 03 (2019) 026

Separate tt+Higgs (signal) from tt+jets (background)

Signal



Background



Single lepton

4j, ≥3b | 5j, ≥3b | ≥6j, ≥3b

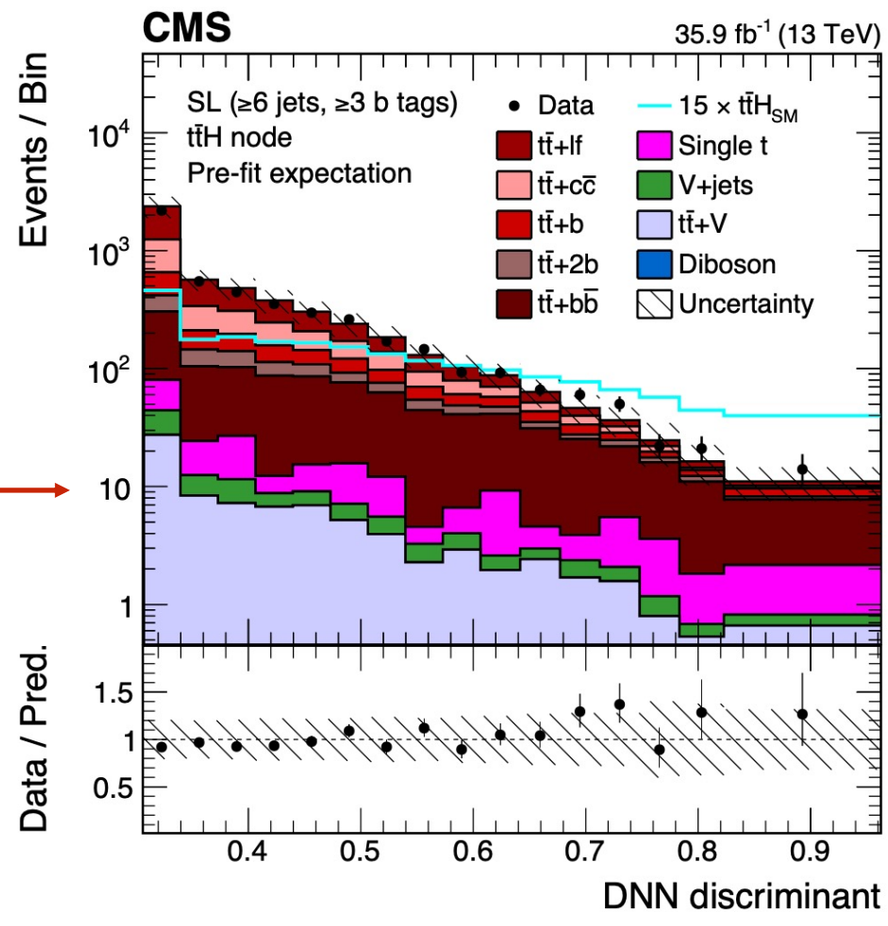
DNN (MEM is input)

Categorise by most probable process
ttH, tt+bb/b/2b/cc/lf

4j, ≥3b, *ttH node* | 5j, ≥3b, *ttH node* | ≥6j, ≥3b, *ttH node*

3 x 6 categories

One of the most complex final state

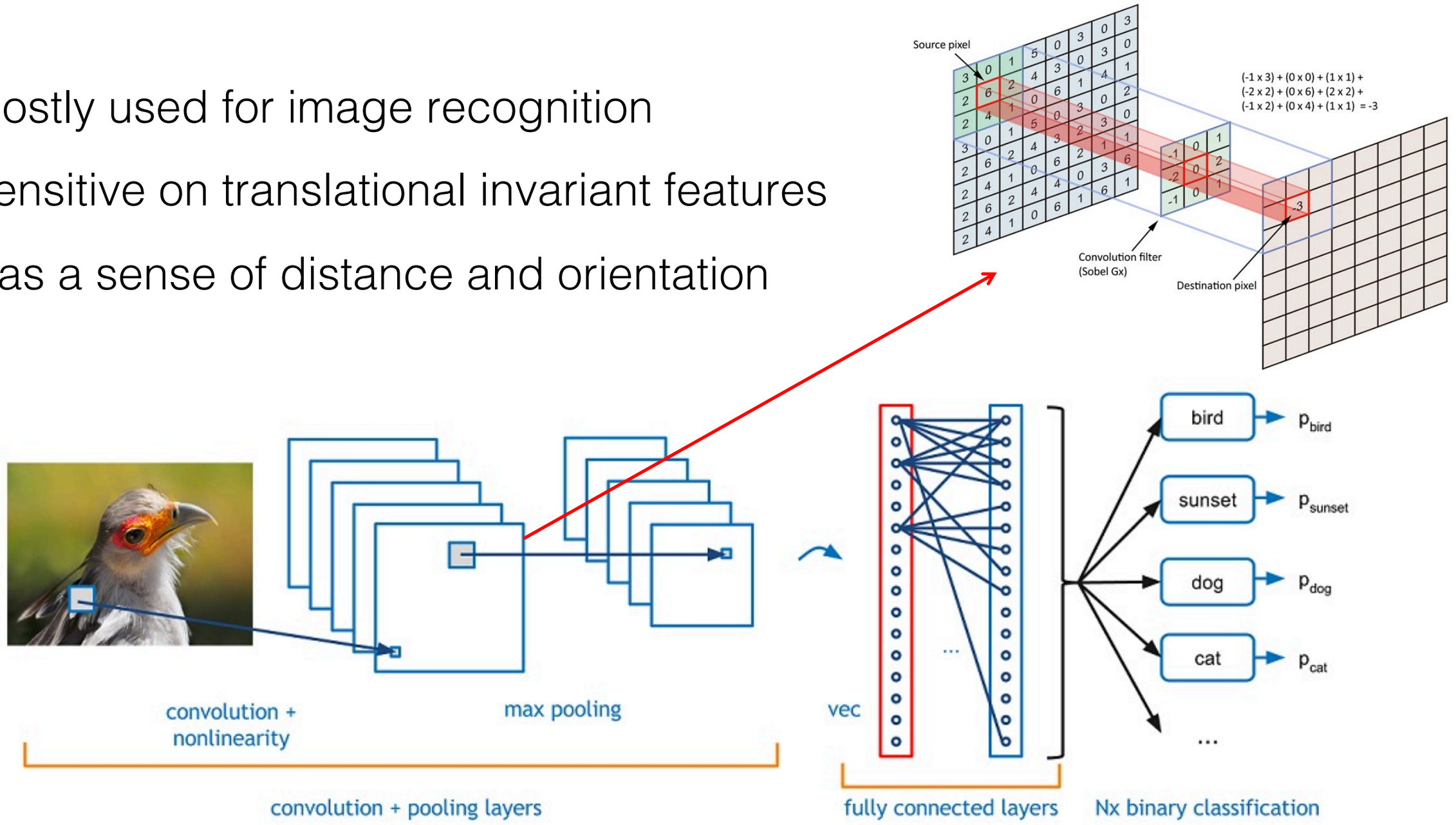


Convolutional Neural Networks (CNNs)

Mostly used for image recognition

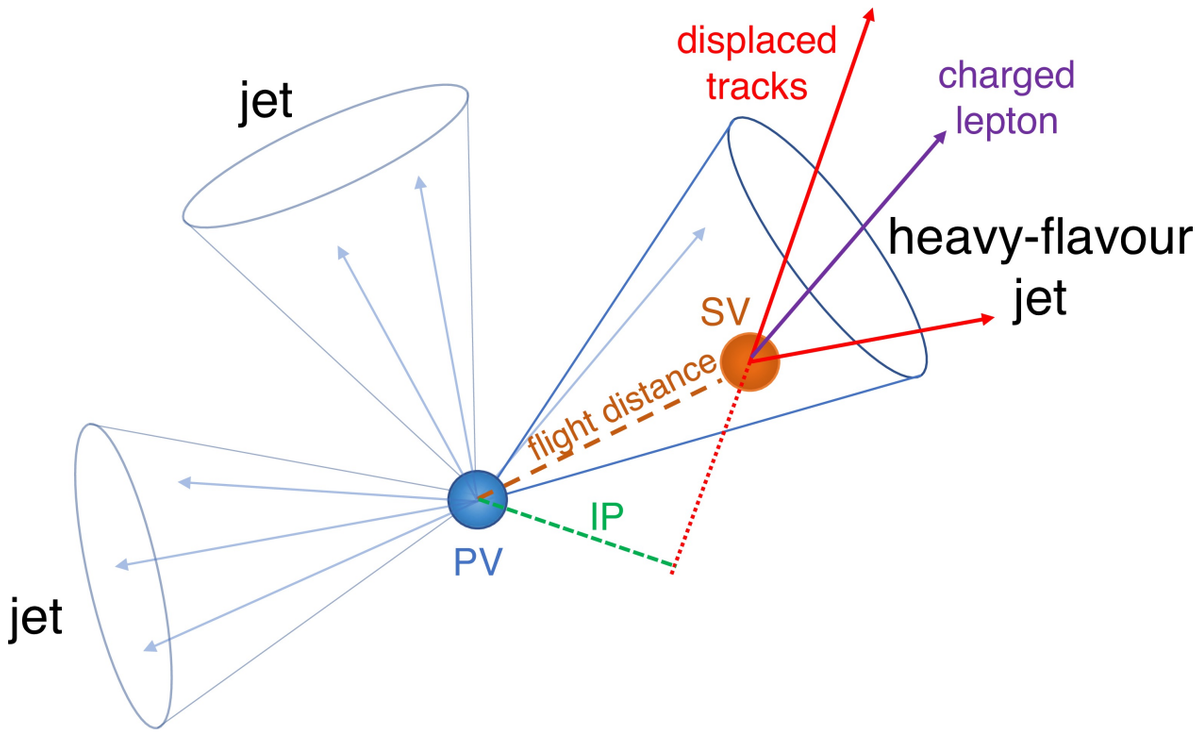
Sensitive on translational invariant features

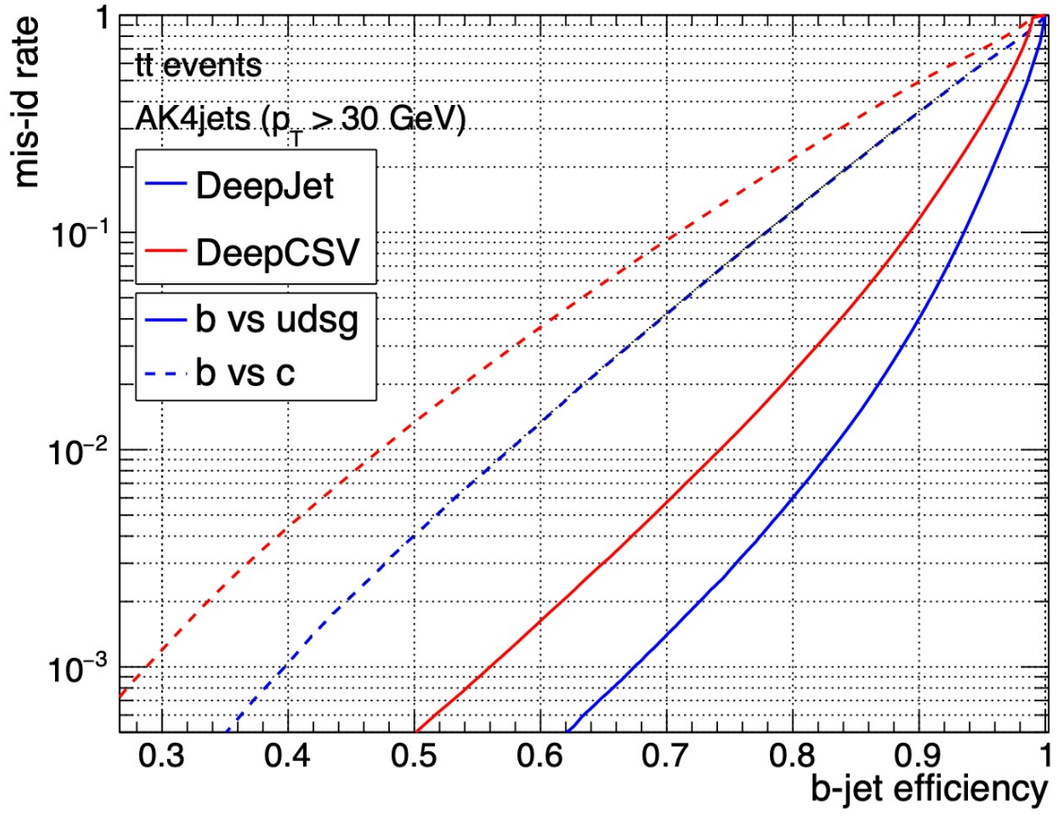
Has a sense of distance and orientation



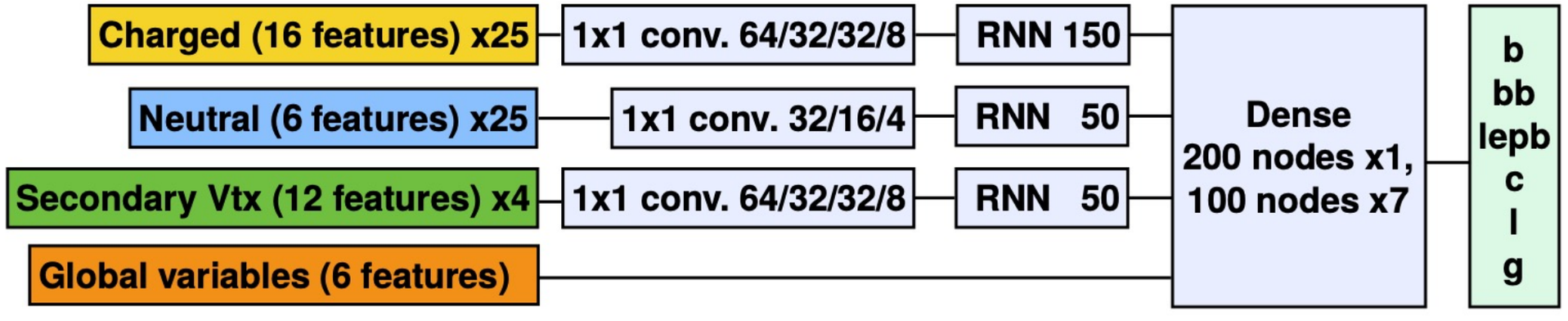
Many applications in HEP → Energy deposits in detector(s) can be considered as an image

Jet flavour classification





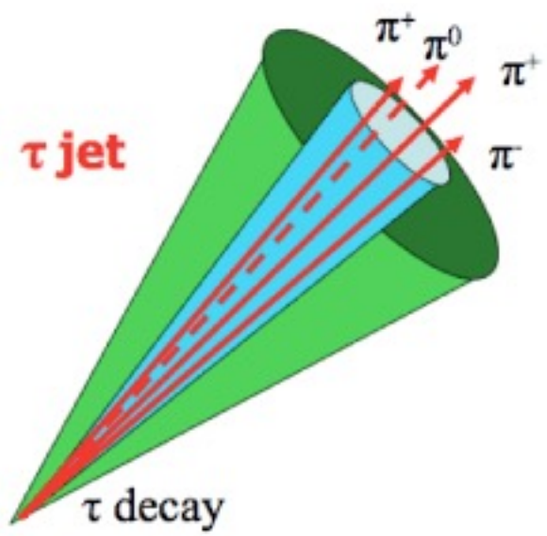
DeepJet Architecture: RNN + 1x1 CNNs for dimensionality reduction

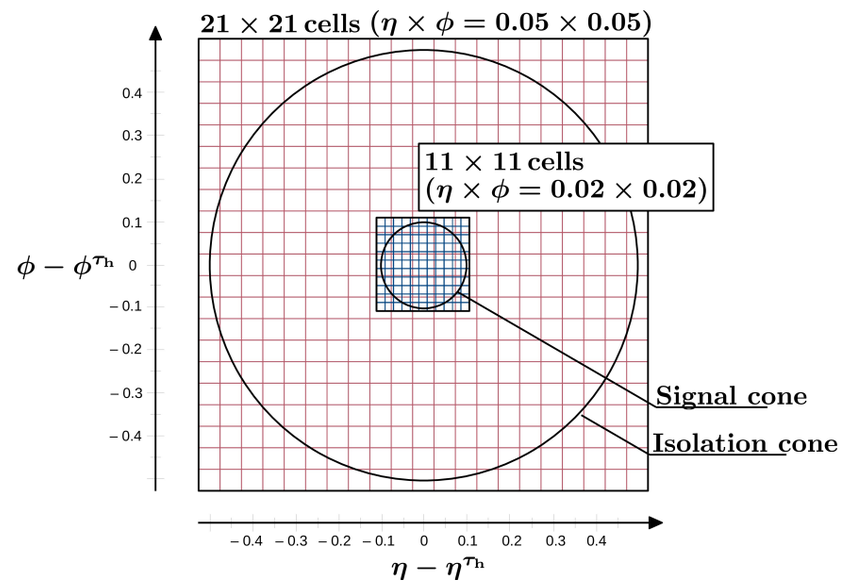


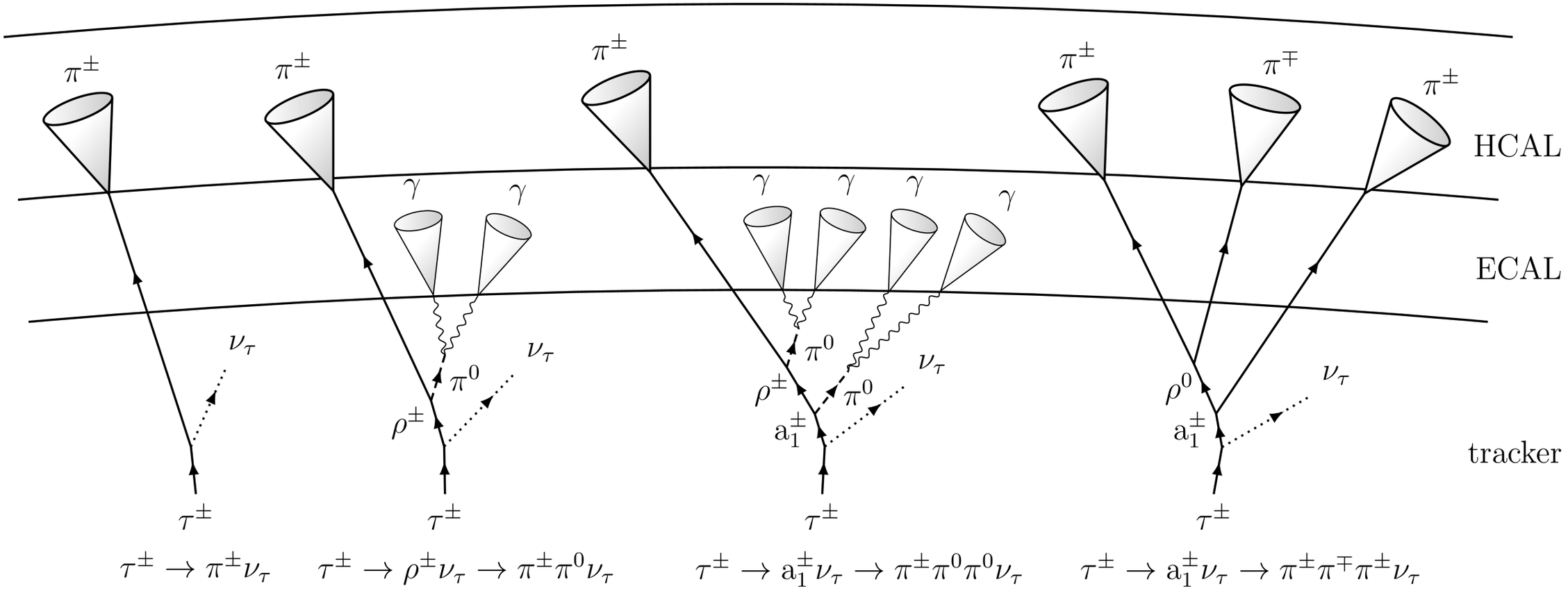
(arXiv:2008.10519)

Tau lepton identification

Decay mode	Meson resonance	$\mathcal{B}$ [%]
$\tau^- \to e^- \bar{\nu}_e \nu_\tau$		17.8
$\tau^- \to \mu^- \bar{\nu}_\mu \nu_\tau$		17.4
$\tau^- \to h^- \nu_\tau$		11.5
$\tau^- \to h^- \pi^0 \nu_\tau$	$\rho(770)$	26.0
$\tau^- \to h^- \pi^0 \pi^0 \nu_\tau$	$a_1(1260)$	9.5
$\tau^- \to h^- h^+ h^- \nu_\tau$	$a_1(1260)$	9.8
$\tau^- \to h^- h^+ h^- \pi^0 \nu_\tau$		4.8
Other modes with hadrons		3.2
All modes containing hadrons		64.8

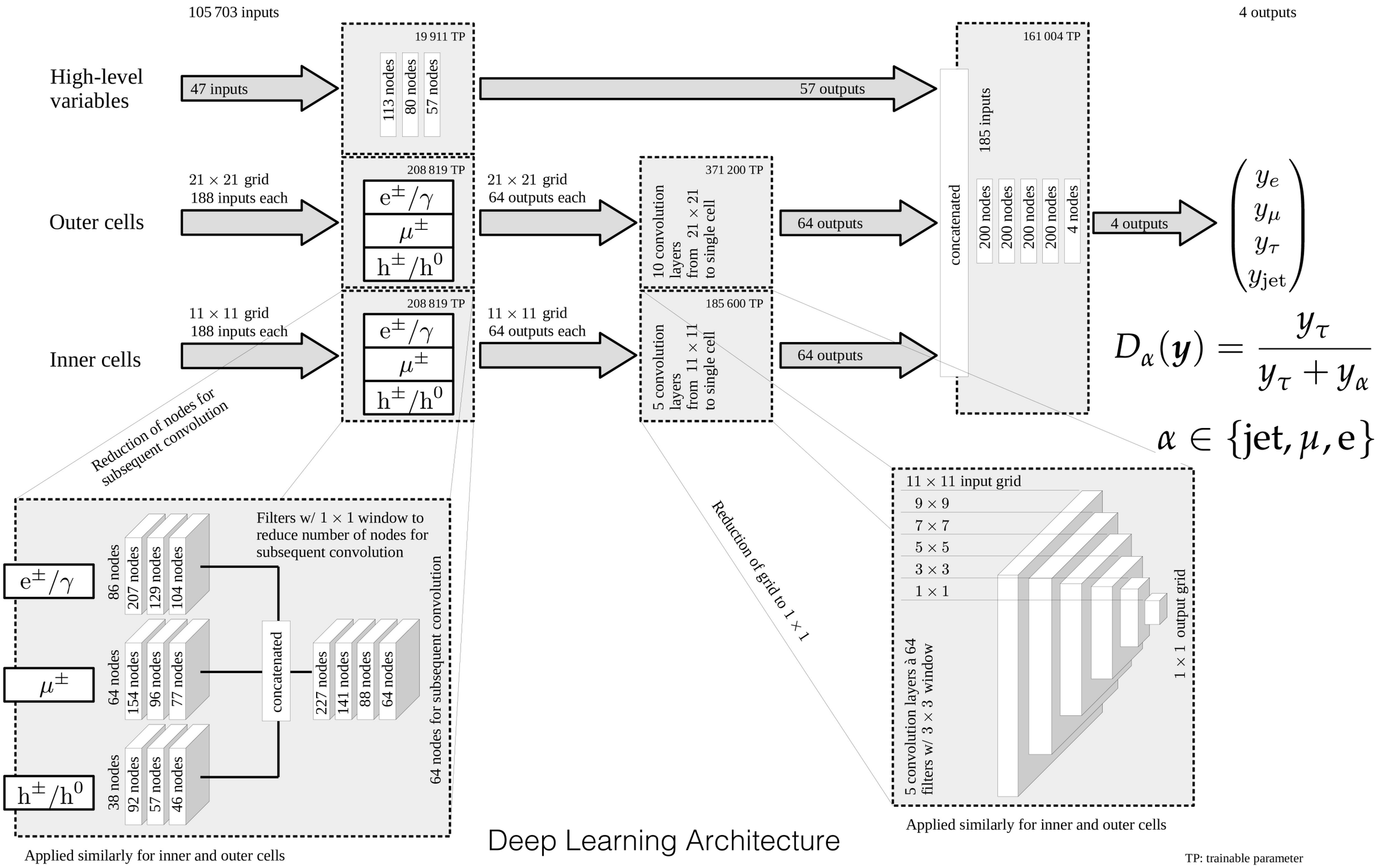






Tau lepton identification

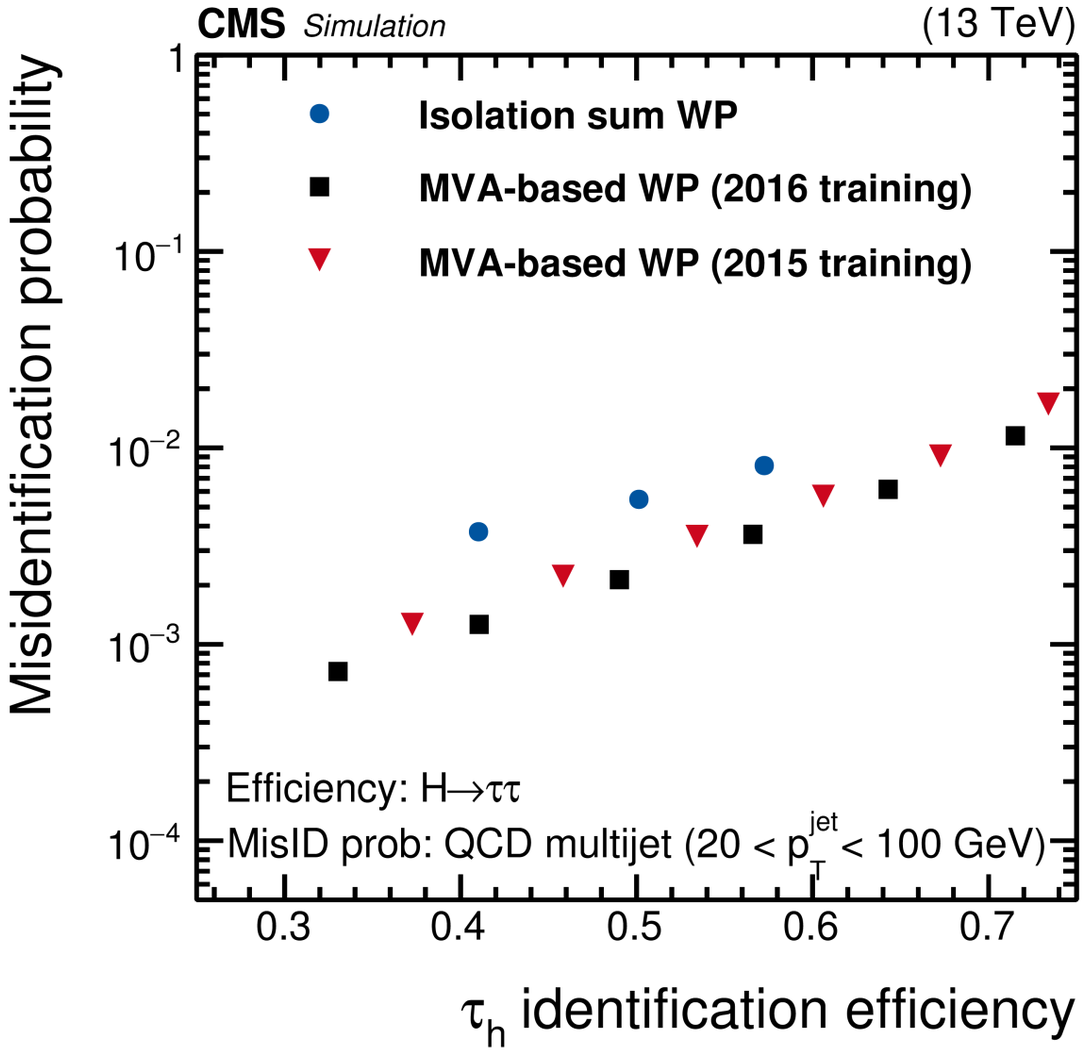
arXiv:2201.08458

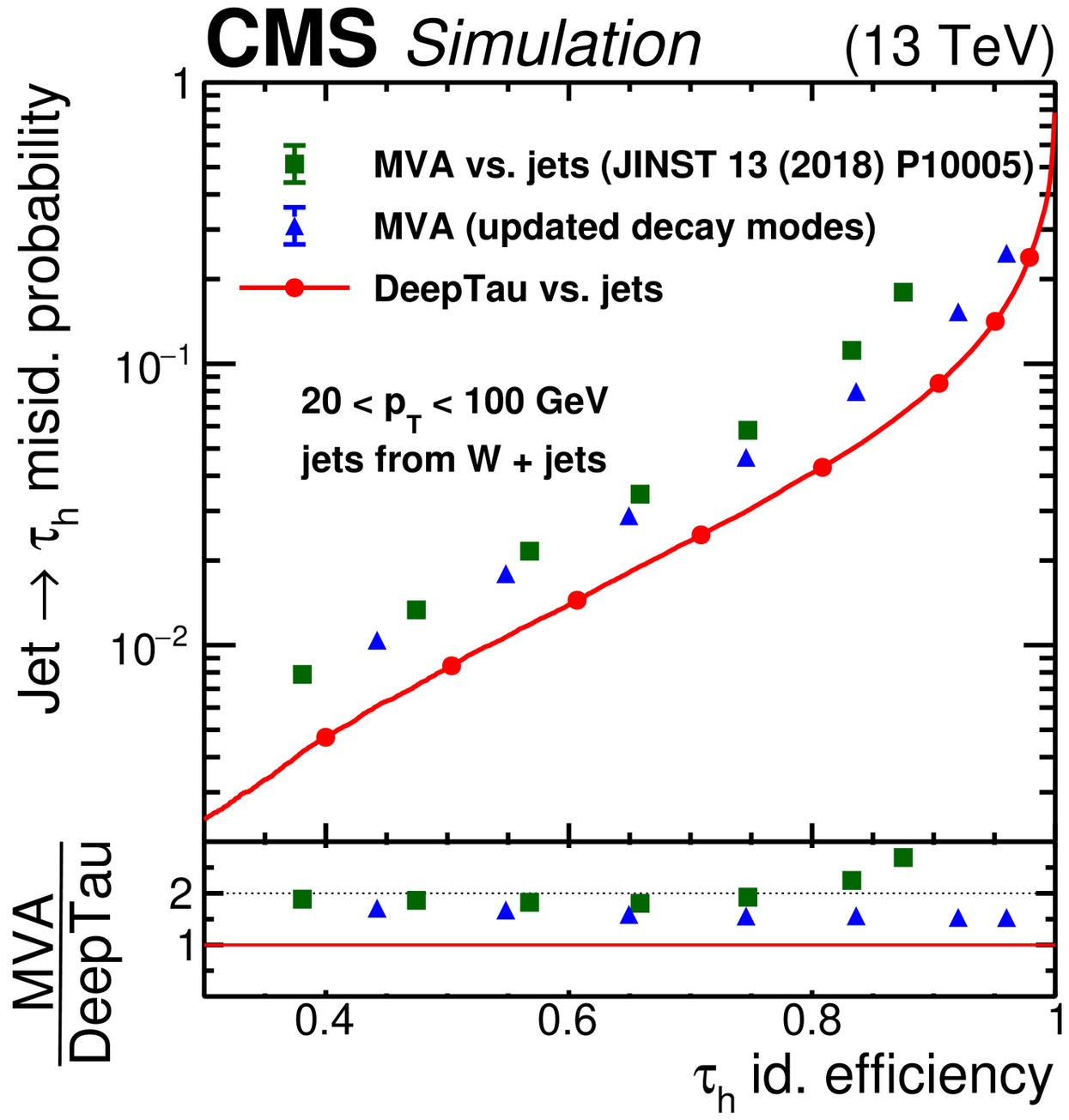


Deep Learning Architecture

TP: trainable parameter

Tau lepton identification

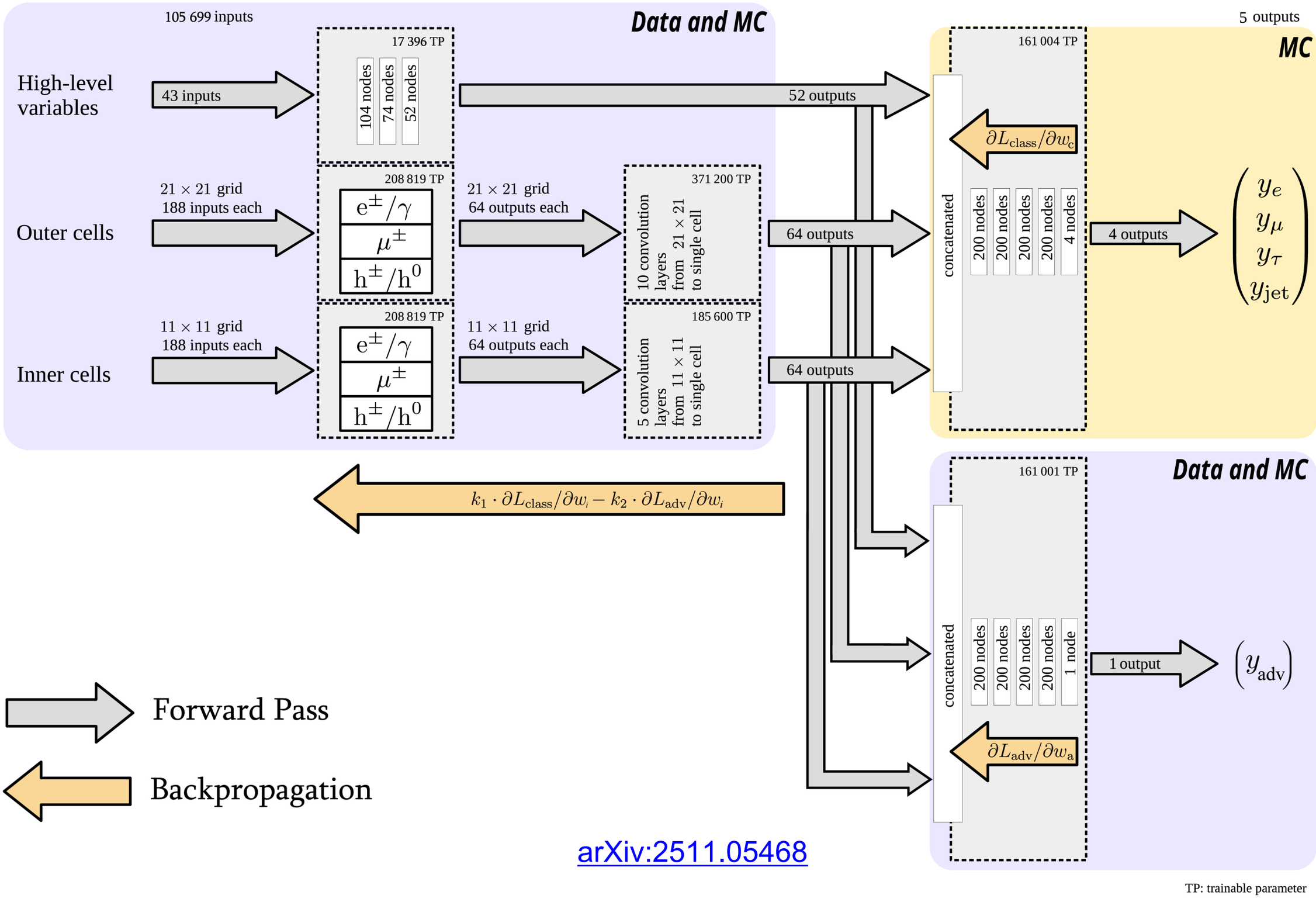




Cut based → BDT based → Deep Learning

Factor of 2 improvement using low-level inputs through CNN compared using only high level features

Improving Data/MC agreement using domain adaptation

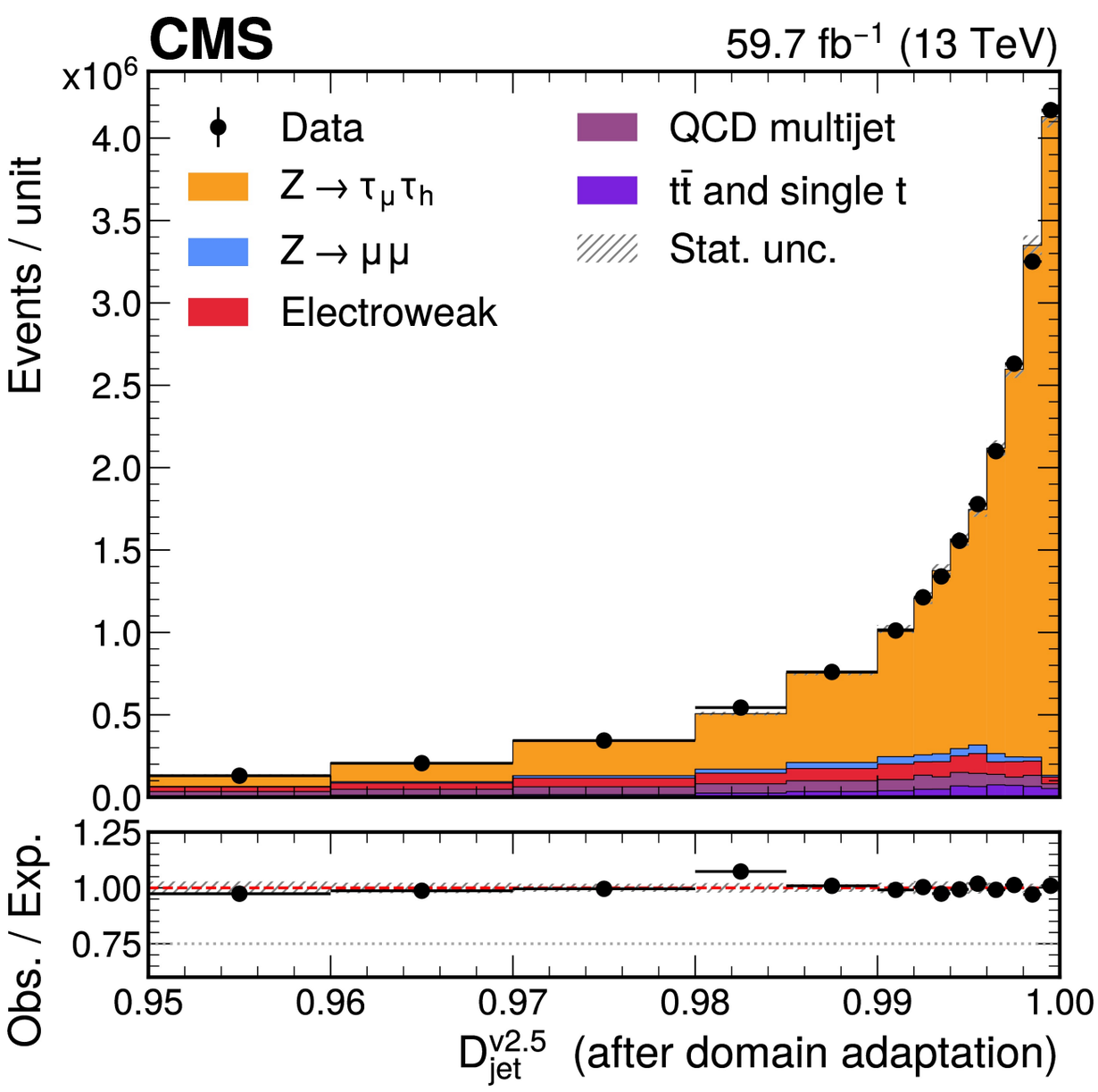


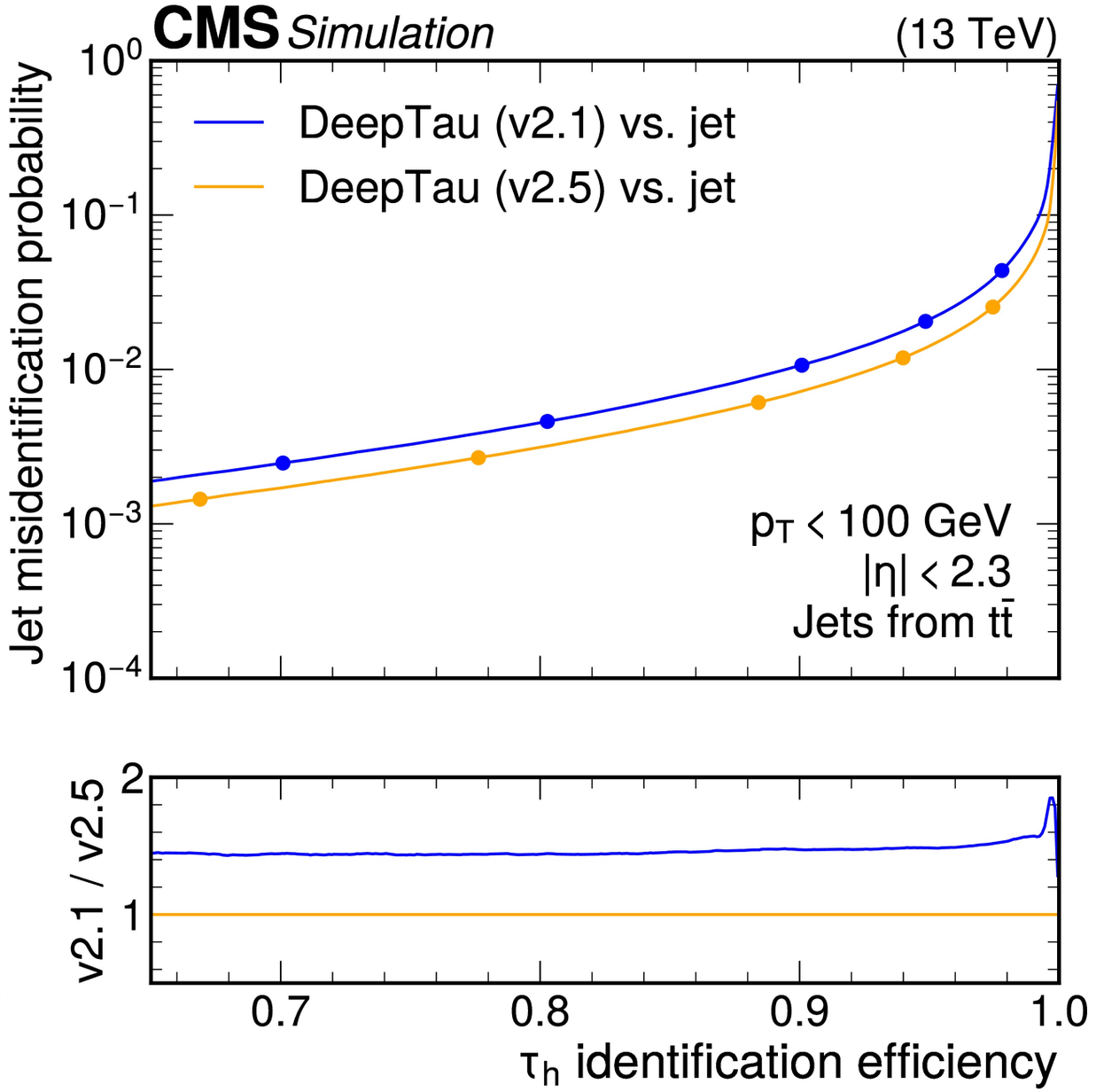
arXiv:2511.05468

TP: trainable parameter

Improves Data/MC agreement for discriminator

Also, reduces fake rate for same efficiency

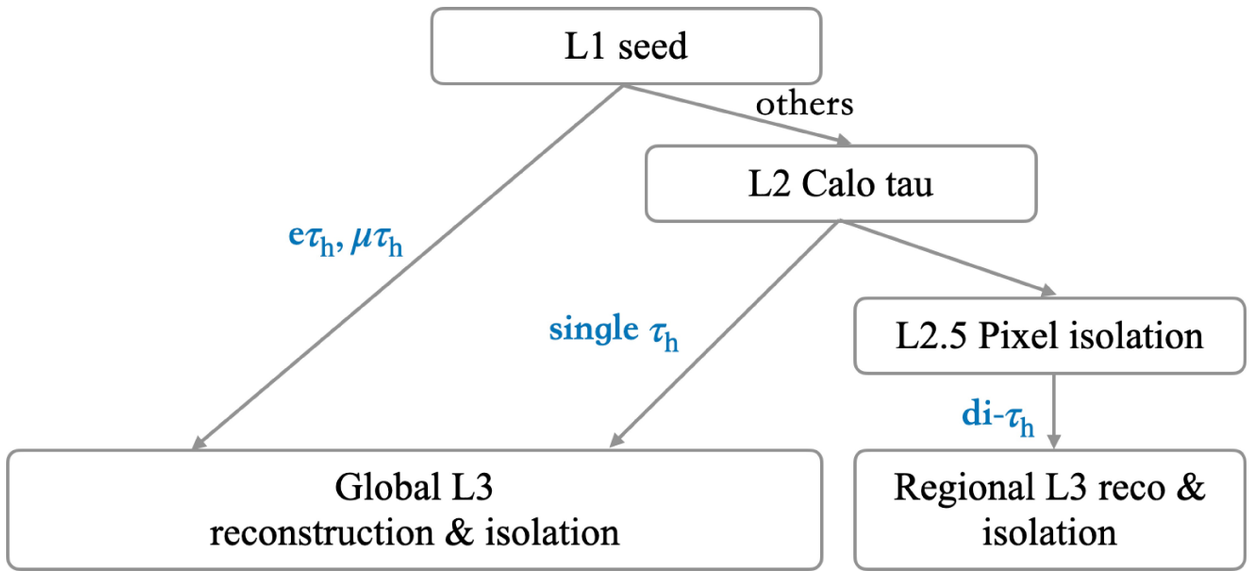


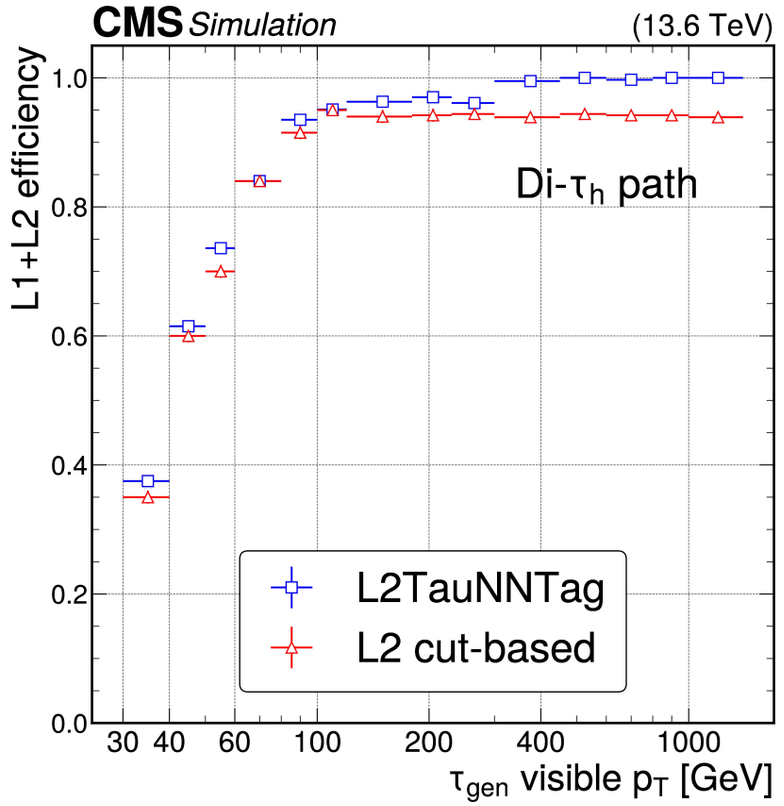


Improved Tau identification at High Level Trigger

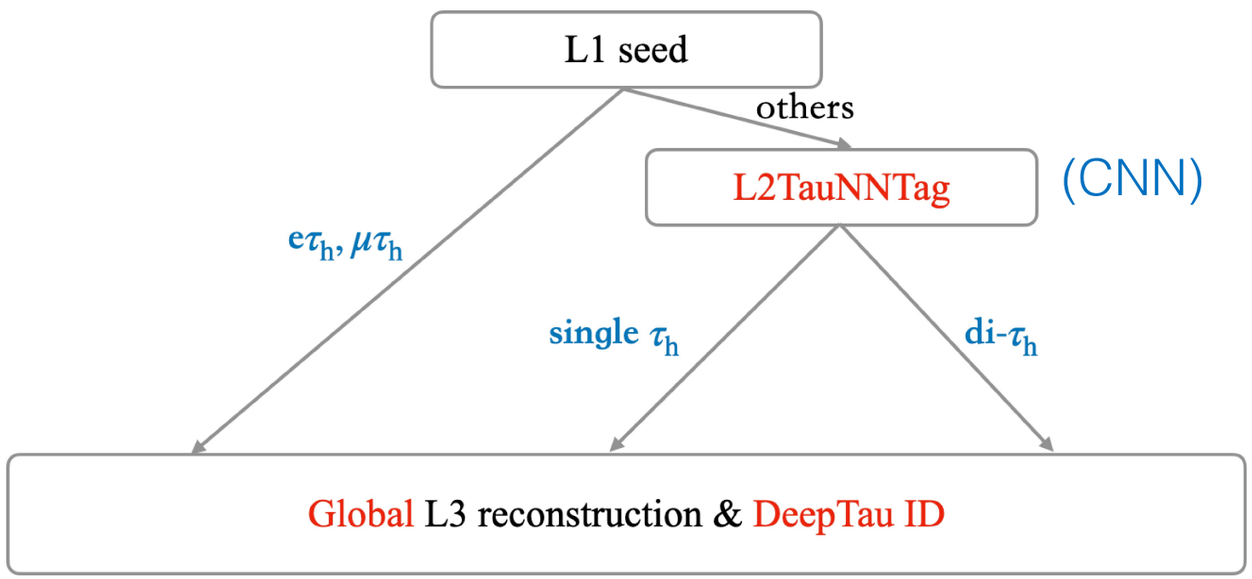
arXiv:2602.11359

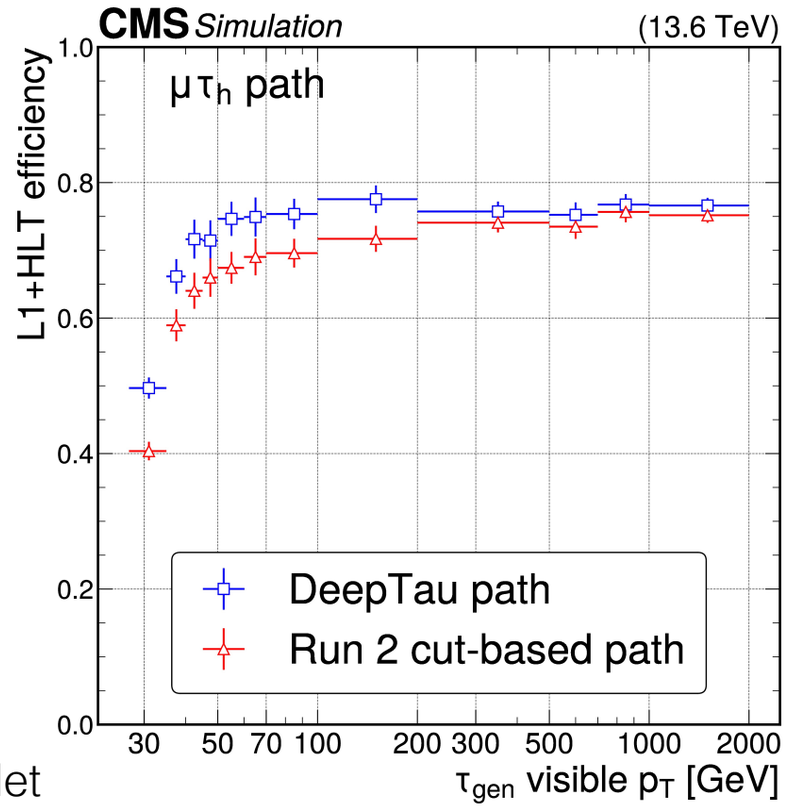
τ_h candidate reconstruction at the HLT in Run 2





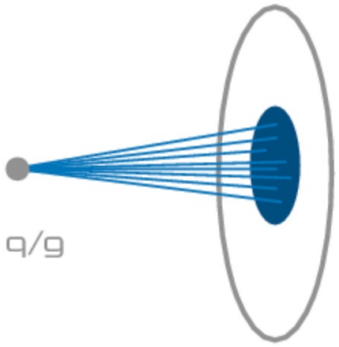
τ_h candidate reconstruction at the HLT in 2022-23 (Run-3)

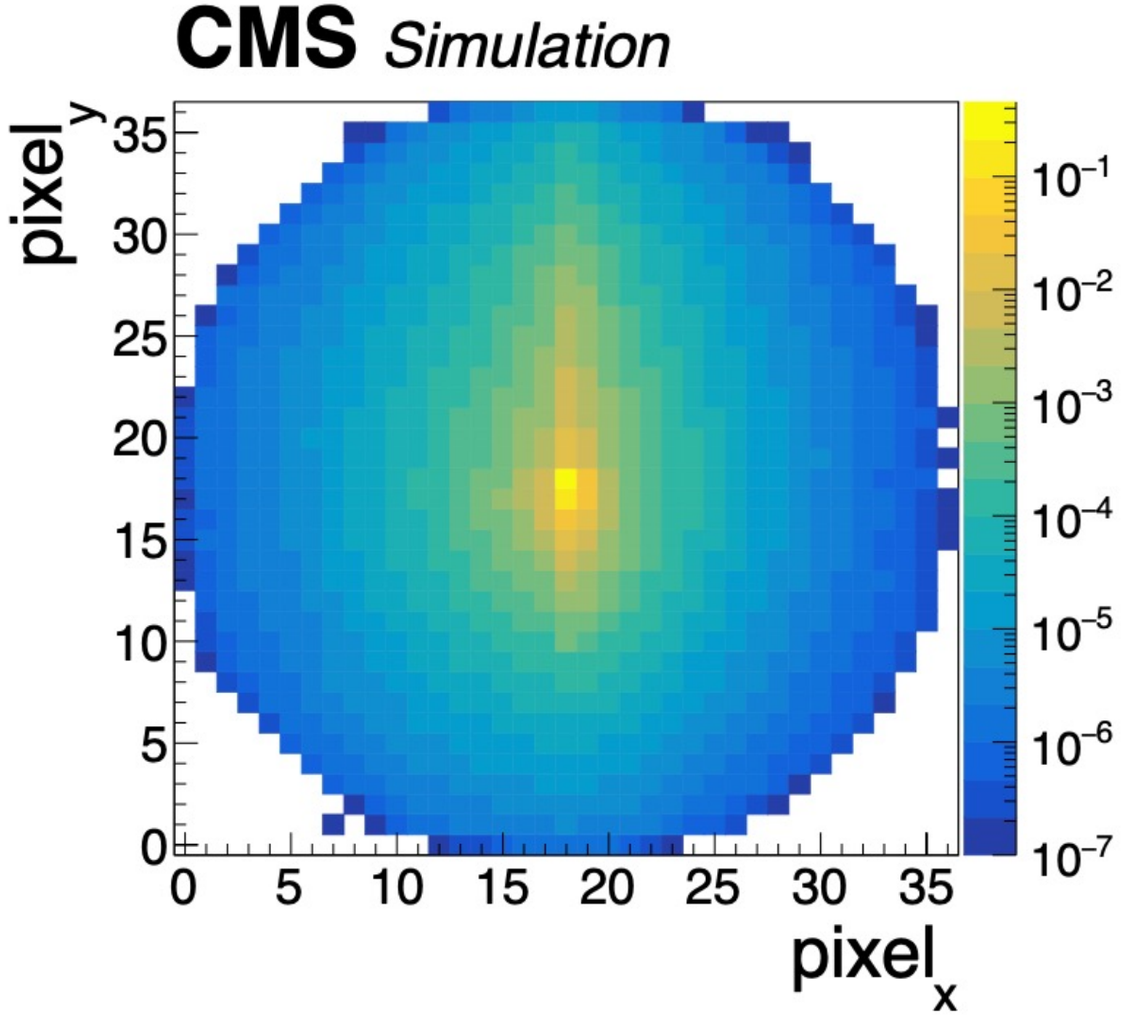




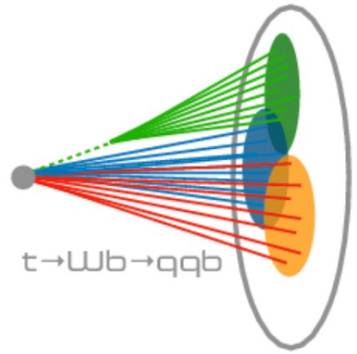
Identifying Boosted and Merged Objects

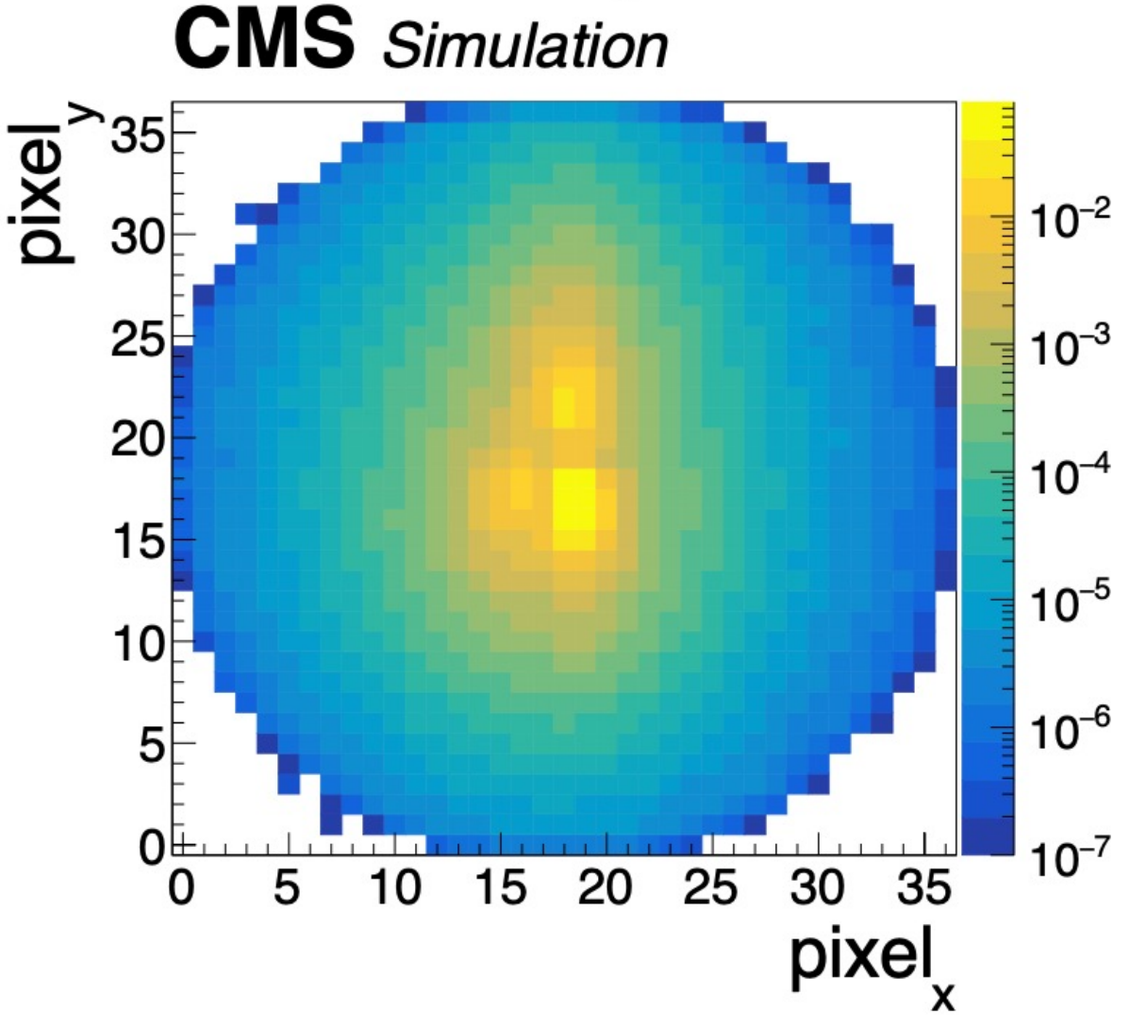
Quark / Gluon jets





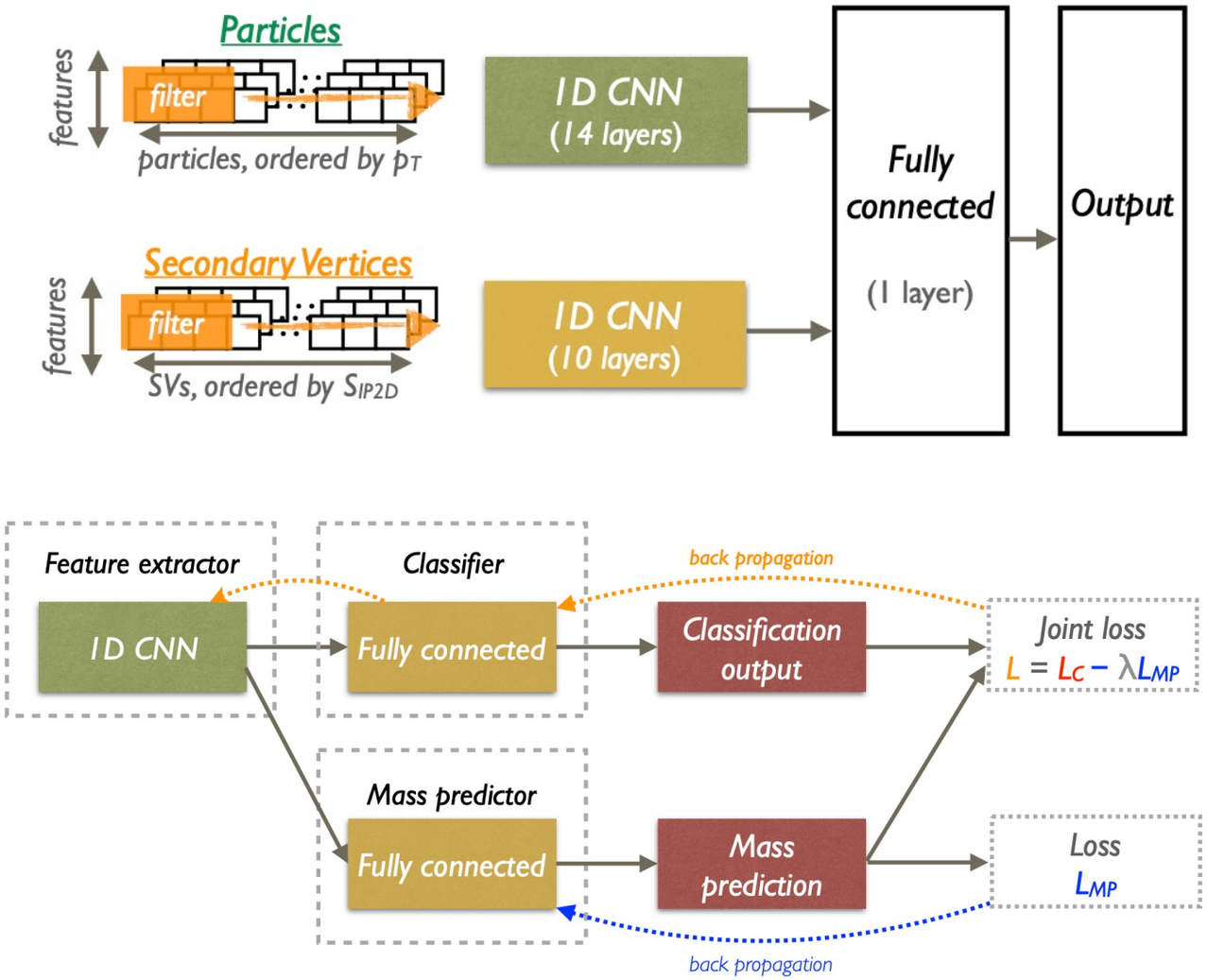
Boosted top quark jets



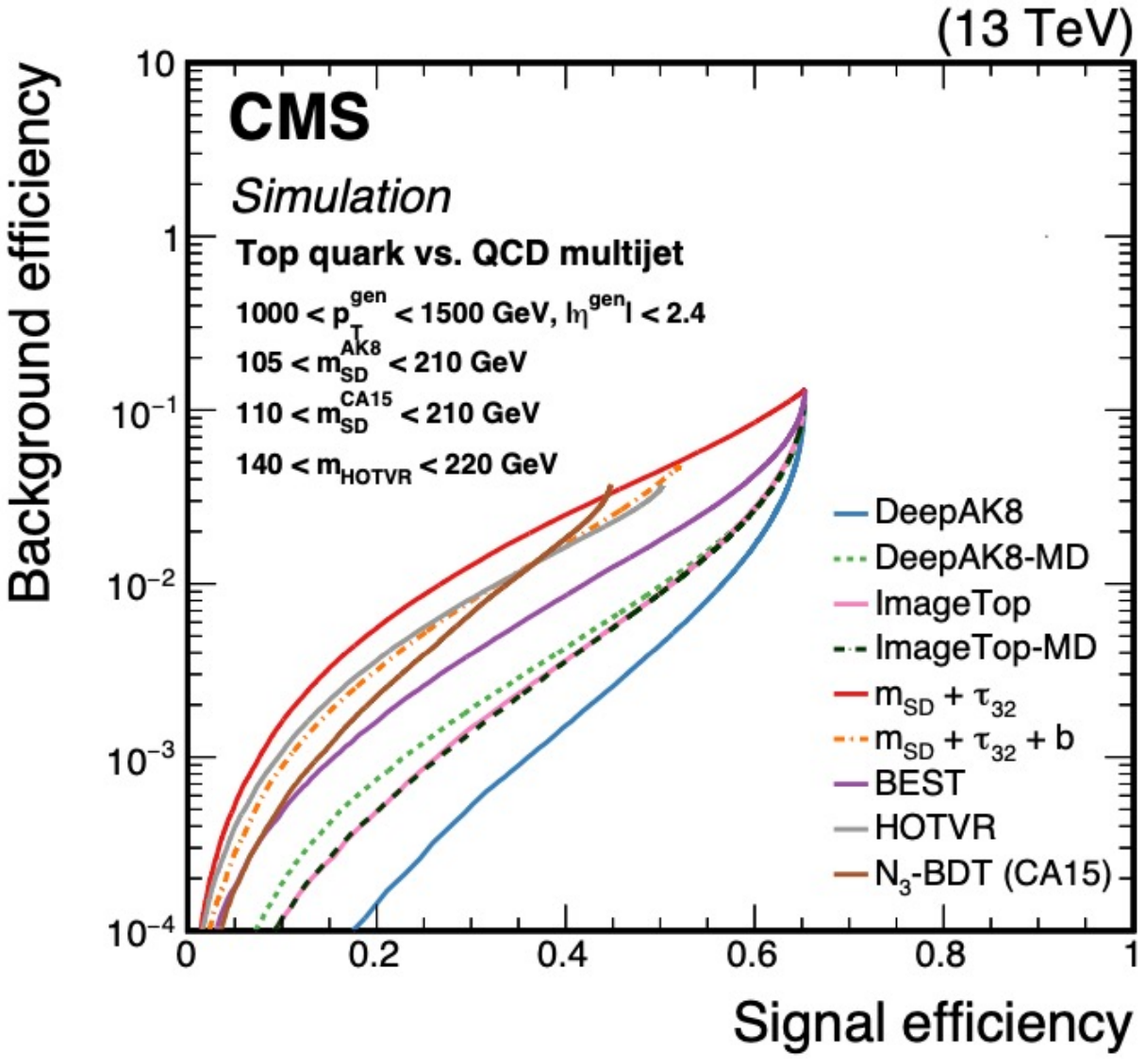


Identification of boosted top quark

DeepAK8
(arXiv:2004.08262)



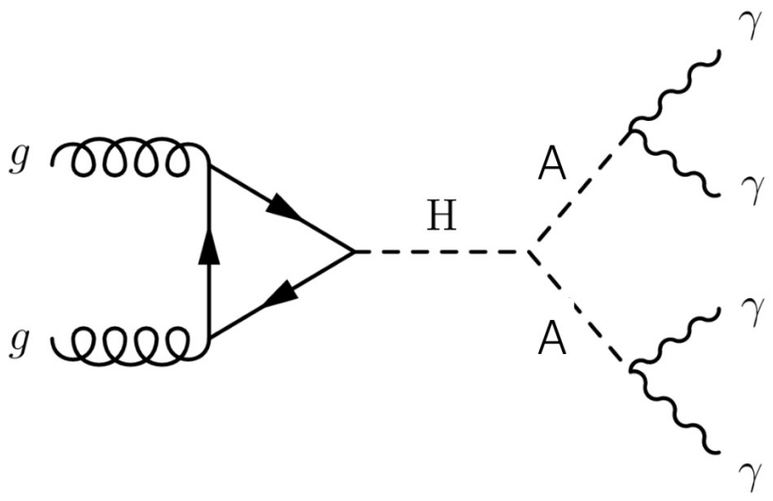
DeepAK8-MD



Large improvements compared to traditional substructure algorithms

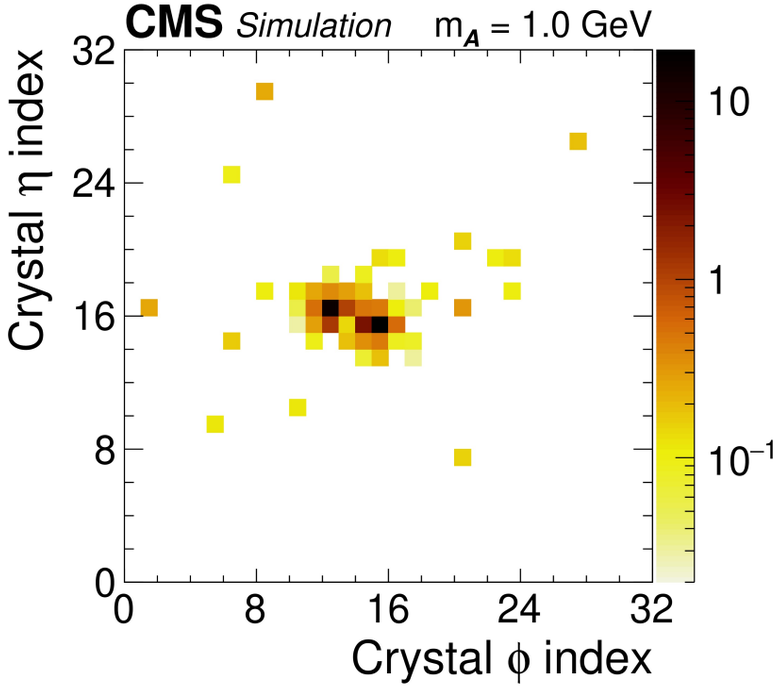
Reconstructing Merged diphotons

[Phys. Rev. D 108 (2023) 052002](#)
[PRL 131 (2023) 101801](#)

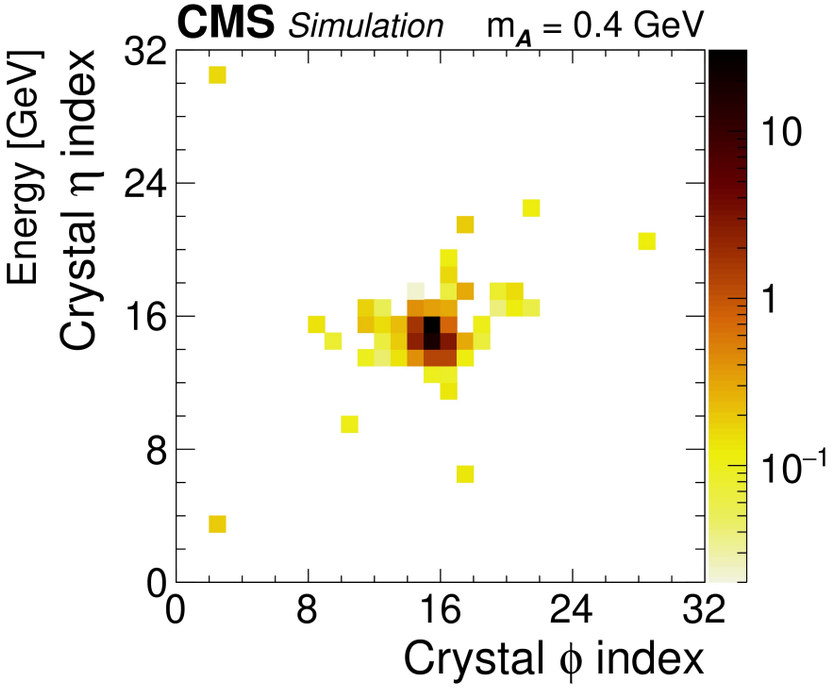


- Impossible to reconstruct using standard reconstruction methods
- Employs a novel end-to-end ML-based particle reconstruction framework, based on RESNET CNN architecture
  - Uses minimally processed detector data as input
  - Outputs regressed m_A

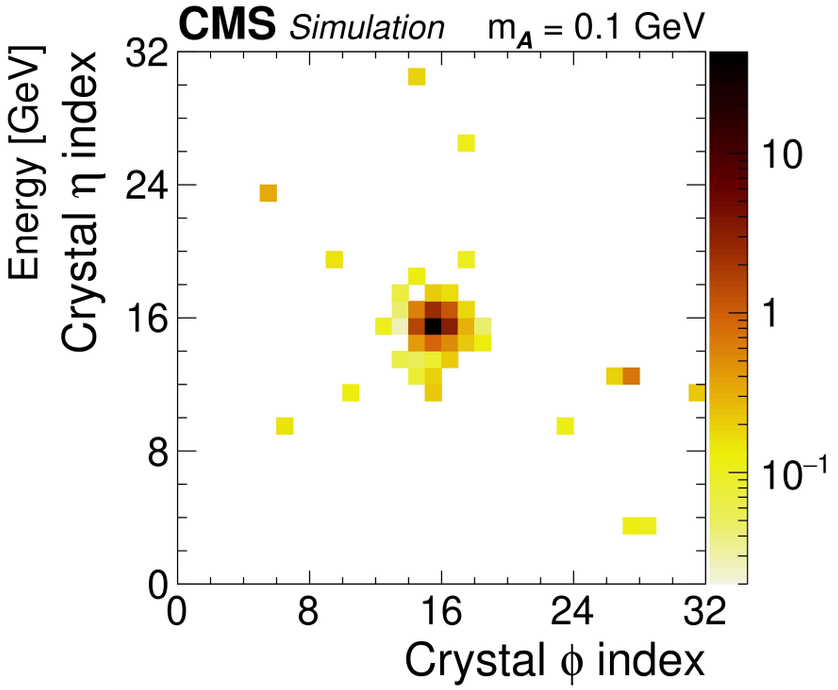
ECAL crystal granularity: $\Delta\eta \times \Delta\phi = 0.0175 \times 0.0175$



m_A= 1.0 GeV, boost = 50



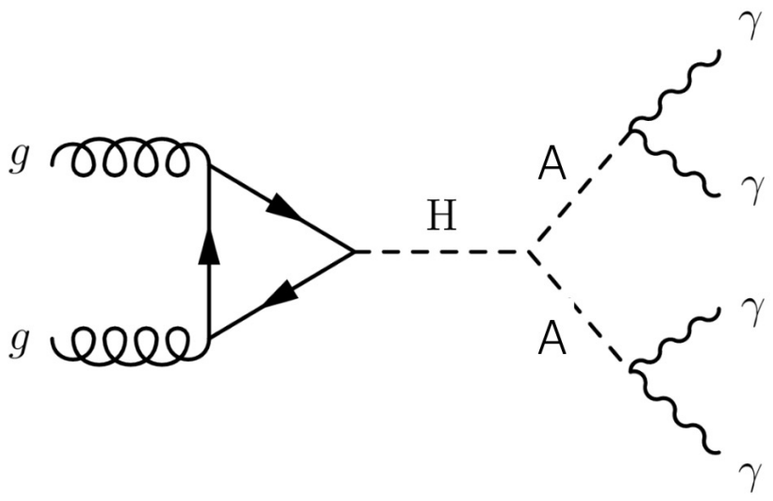
m_A= 0.4 GeV, boost = 150



m_A= 0.1 GeV, boost = 625

Reconstructing Merged diphotons

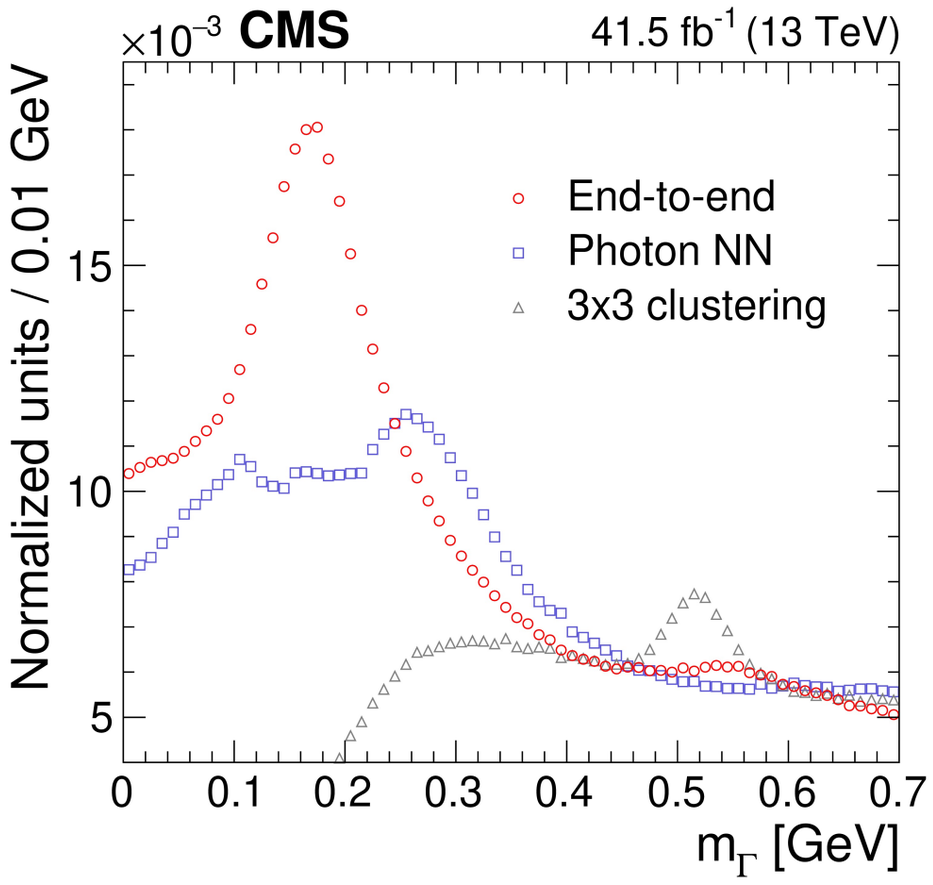
[Phys. Rev. D 108 (2023) 052002](#)
[PRL 131 (2023) 101801](#)

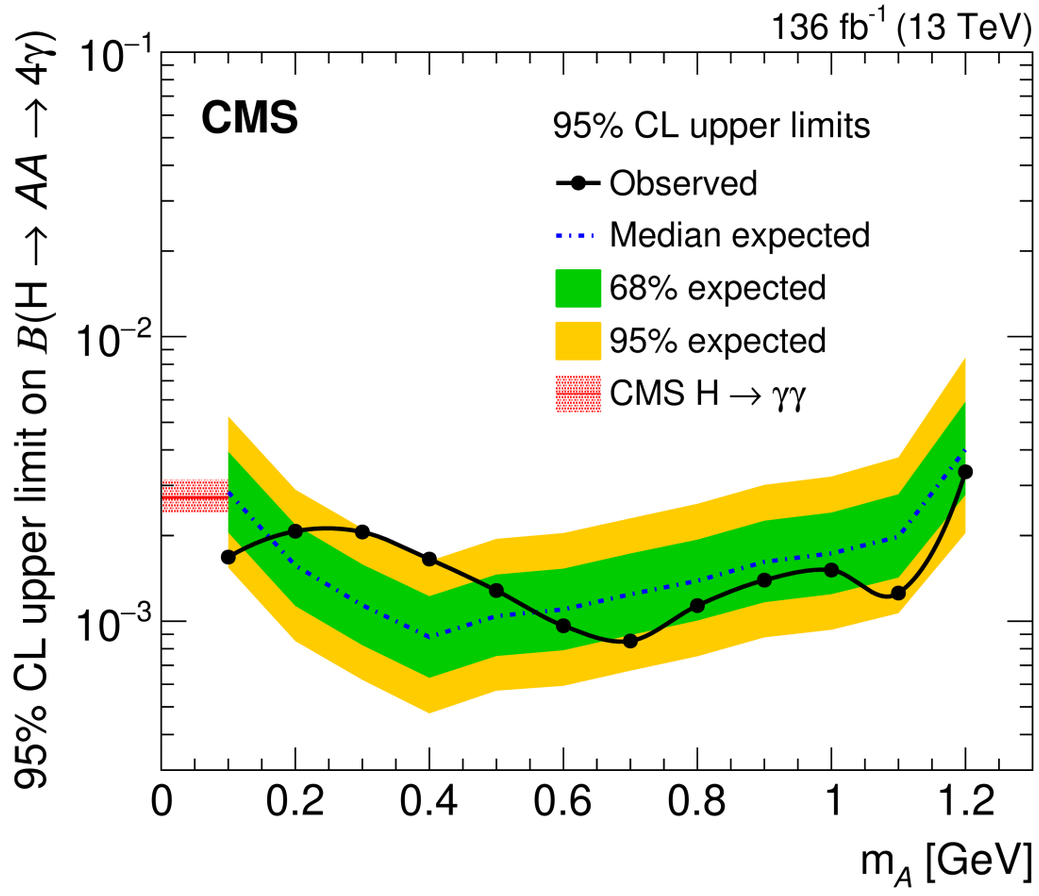


- Impossible to reconstruct using standard reconstruction methods
- Employs a novel end-to-end ML-based particle reconstruction framework, based on RESNET CNN architecture
  - Uses minimally processed detector data as input
  - Outputs regressed m_A

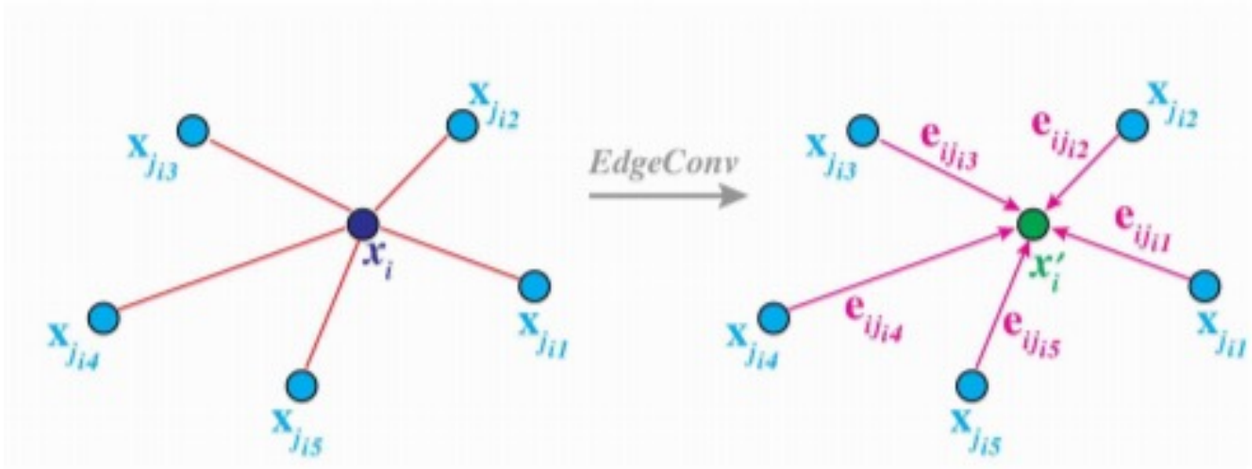
First search of its kind at CMS

Validated using $\pi^0 \rightarrow \gamma\gamma$





Graph Neural Networks (GNNs)



Designed to perform inference on data described by graphs.

Can extract information about the full graph or single nodes (objects) or edges.

arXiv:1812.08434

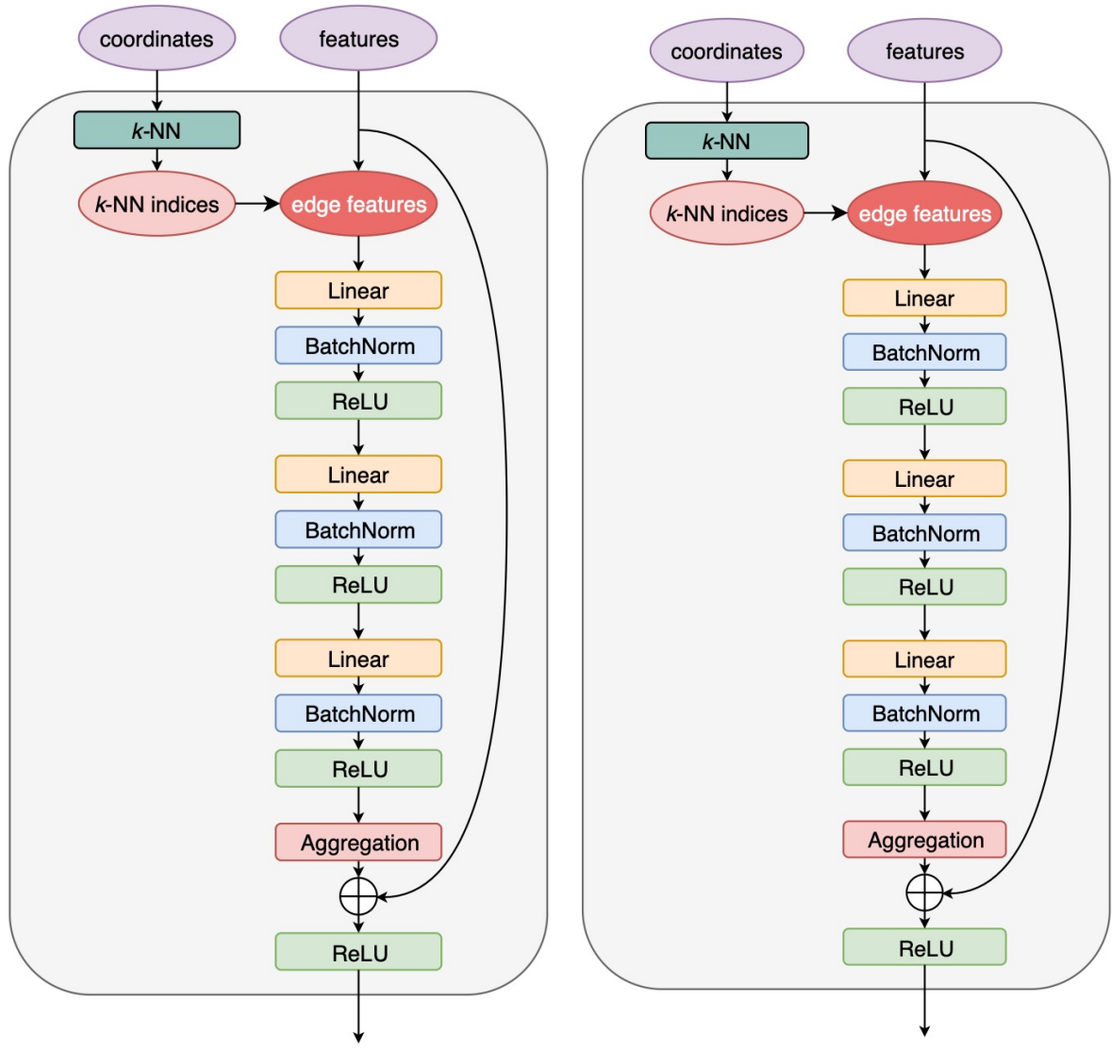
Graph neural networks: A review of methods and applications

Jie Zhou [a,1], Ganqu Cui [a,1], Shengding Hu [a], Zhengyan Zhang [a], Cheng Yang [b], Zhiyuan Liu [a,*], Lifeng Wang [c], Changcheng Li [c], Maosong Sun [a]

ParticleNet (arXiv:1902.08570)

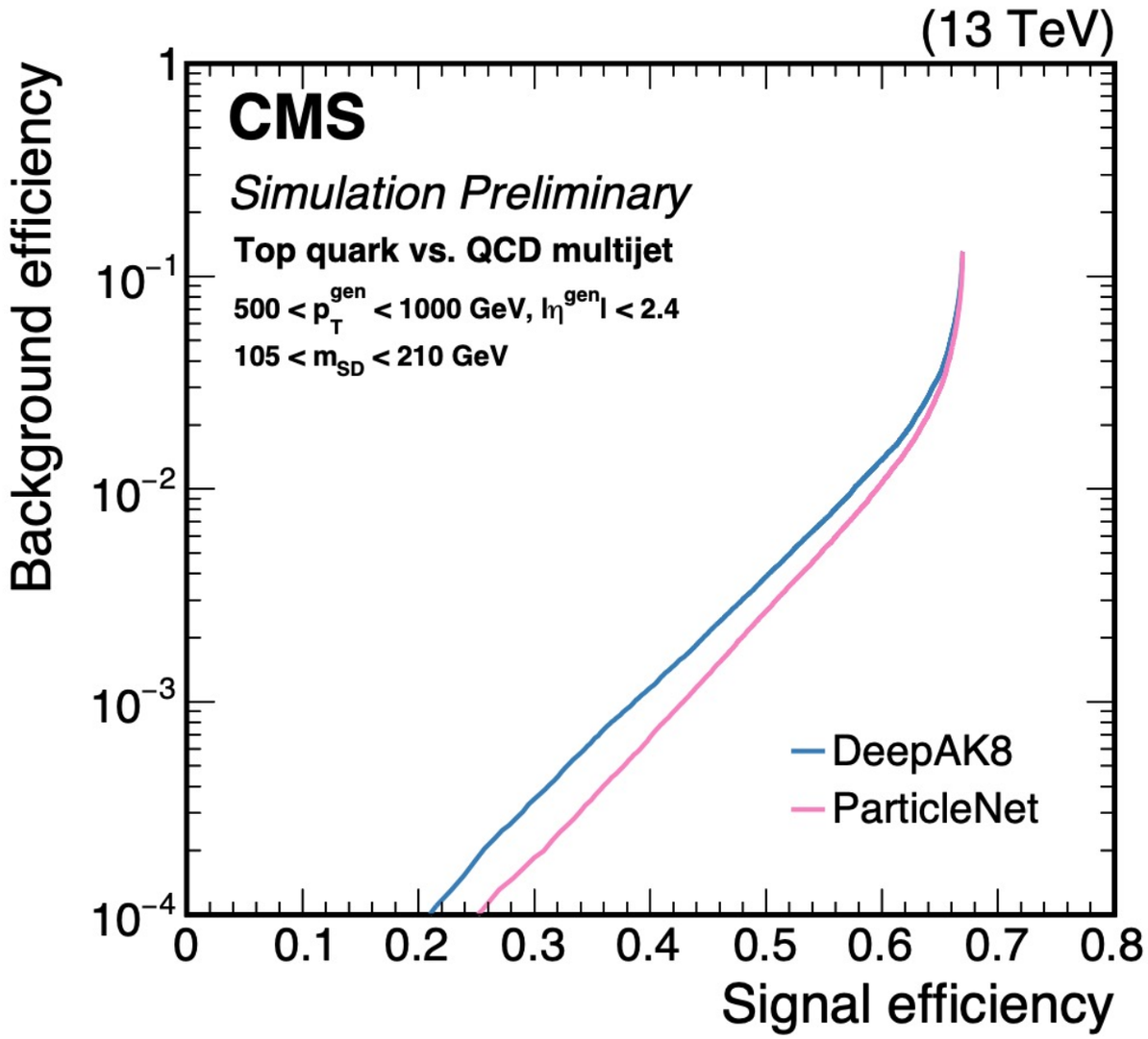
(Based on DGCNN)

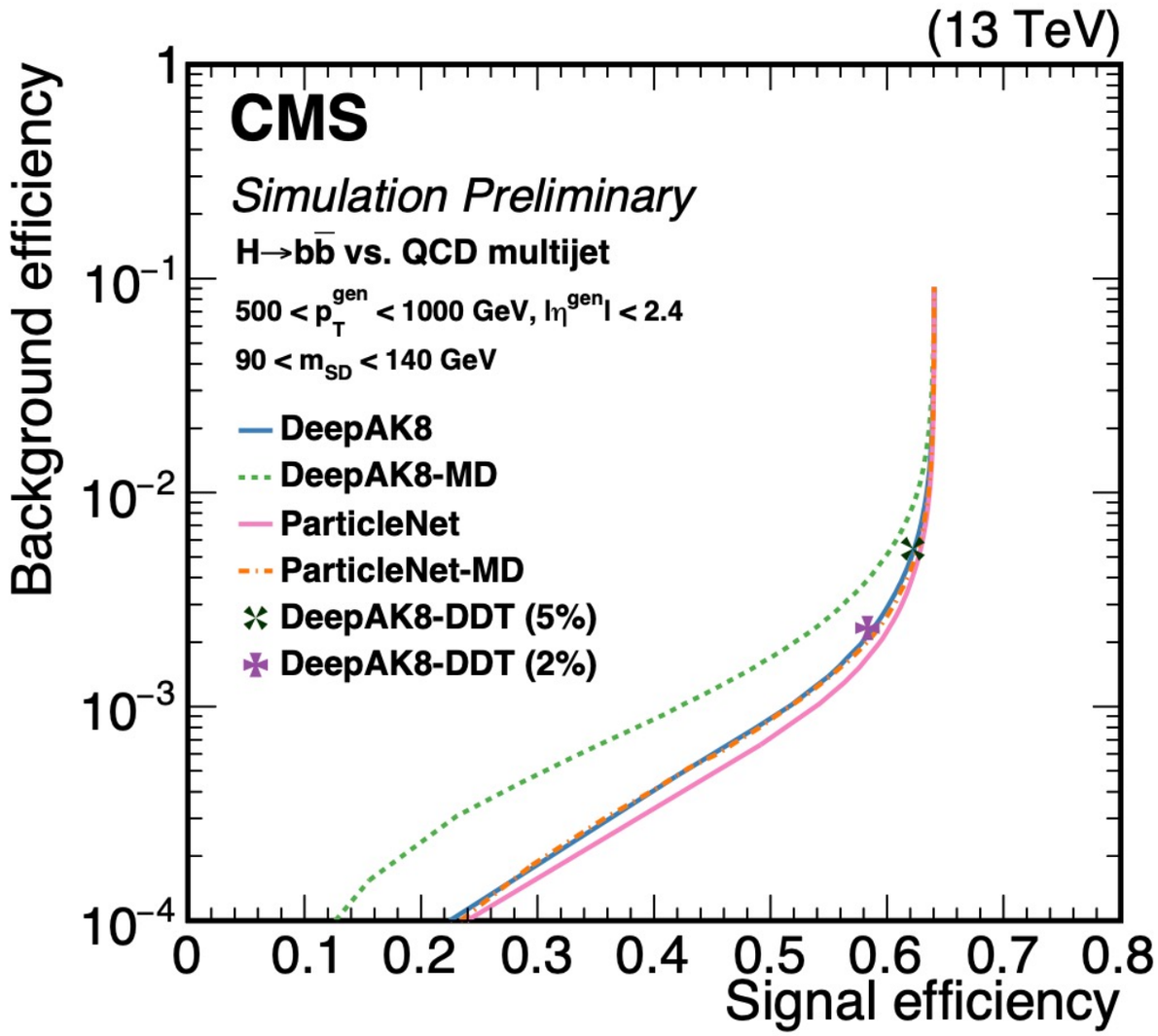
ACM Trans. Graph. 38, 146 (2019)



Used for jet Tagging → Considers jet as an unordered set of constituent particles

Further improvement in object identification using Graph CNN (ParticleNet)





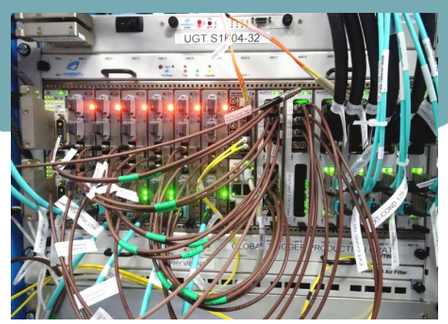
CMS DP 2020.002

CMS Trigger System

Collisions: ~ 30 MHz

Level 1 Trigger

Coarse reconstruction, limited detector systems



Fast Electronics
Latency: 4 μs

~110 kHz

High Level Trigger

Full detector information and online resolution

Computing farm

Employs an optimized version of CMS software framework to run online

Standard data stream:
~2.8 kHz, ~3.1 GB/s

Prompt offline reconstruction

Parking data stream:
~6.3 kHz, ~6.5 GB/s

Delayed offline reconstruction

Scouting data stream:
~32 kHz, ~0.8 GB/s

No offline reconstruction

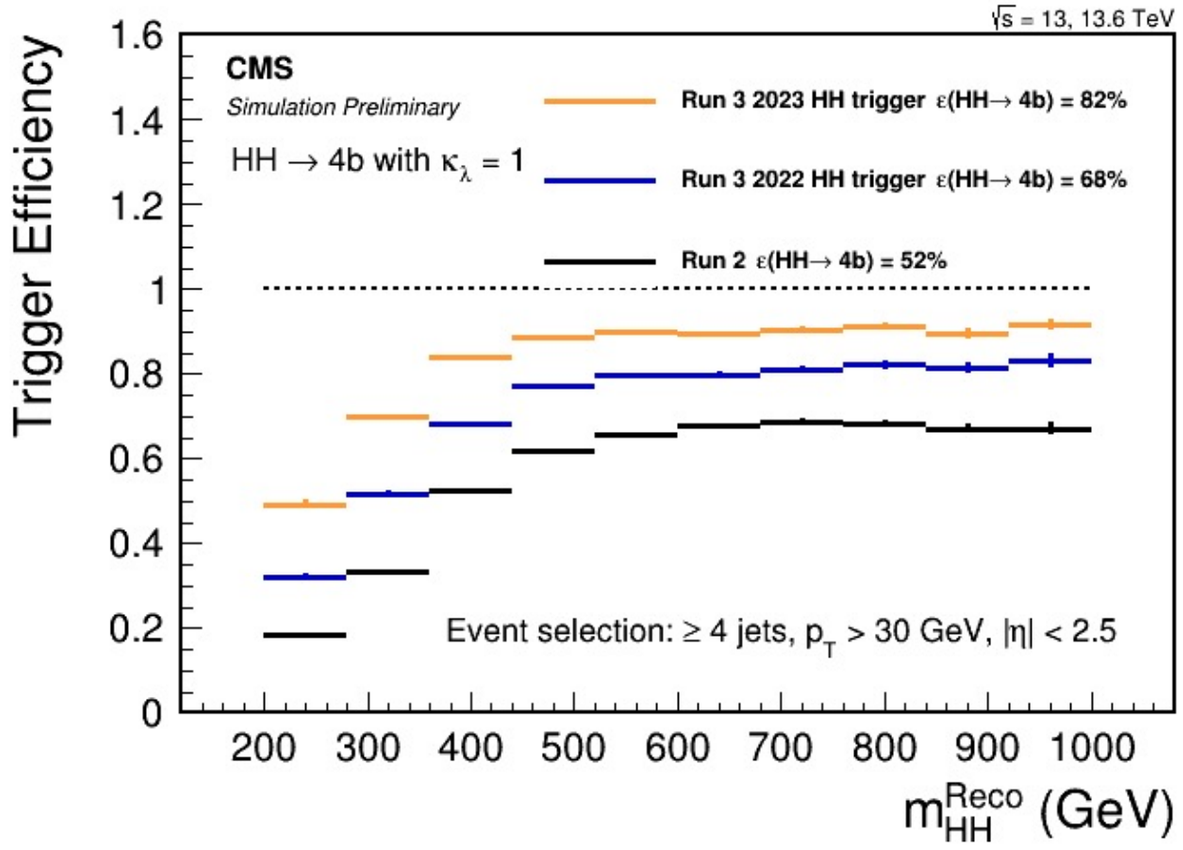
Data reconstructed and stored on disk

b-tagging @ HLT

Trigger	Requirement	Rates at HLT at 2x10^34 cm-2s-1
2023 HH trigger	HT > 280 GeV, 4 jets with pT > 30 GeV, PNet@AK4(mean 2 highest b-tag score) > 0.55	180 Hz
2022 HH trigger	4 jets pT > 70, 50, 40, 35 GeV, PNet@AK4(mean 2 highest b-tag score) > 0.65	60 Hz
2018 triple b-tag [2,3]	HT > 340 GeV, 4 jets pT > 75, 60, 45, 40 GeV, 3 b-tags with DeepCSV > 0.24	8 Hz

Improved efficiency at HLT in Run-3 using ParticleNet

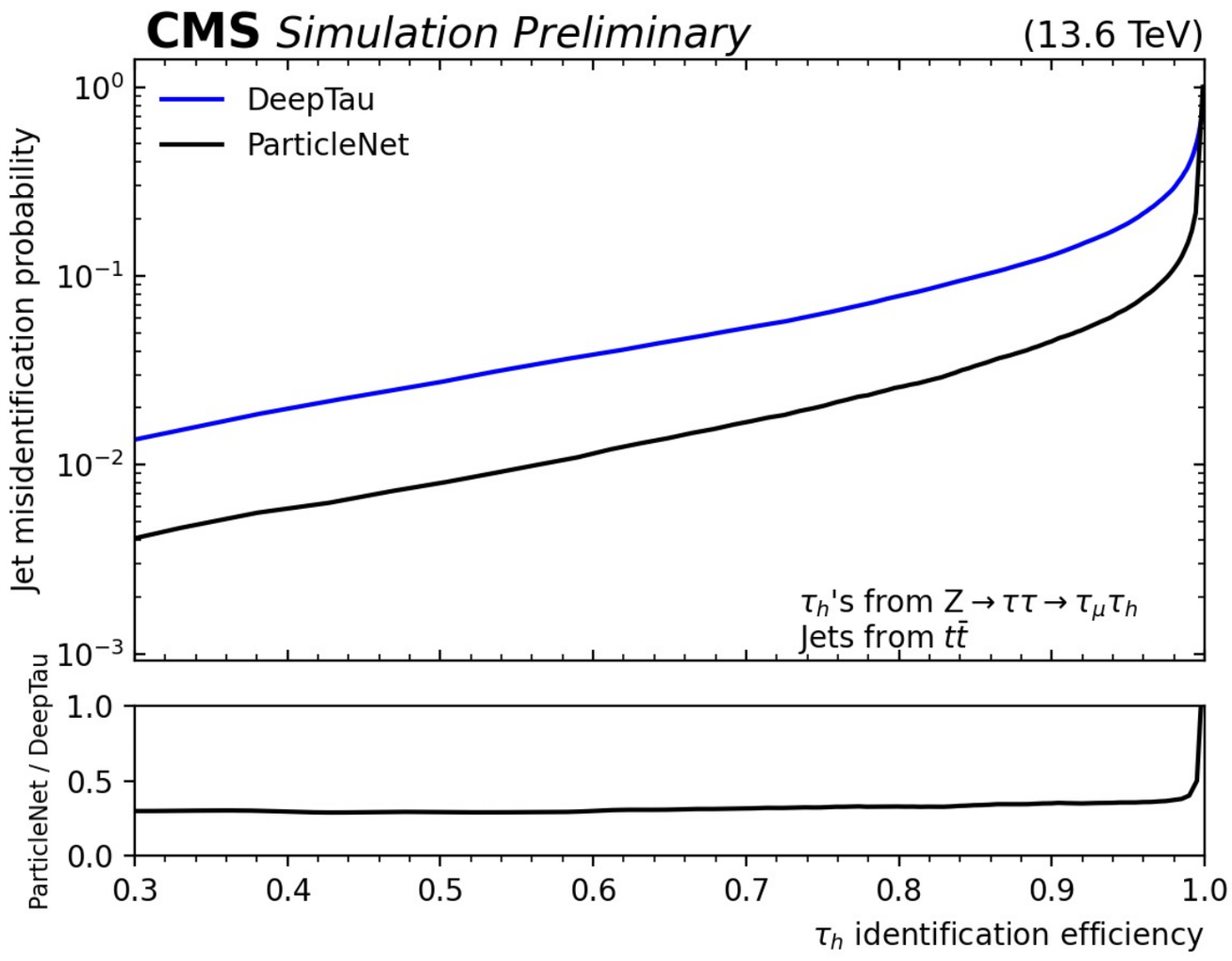
CMS-DP-2023-050

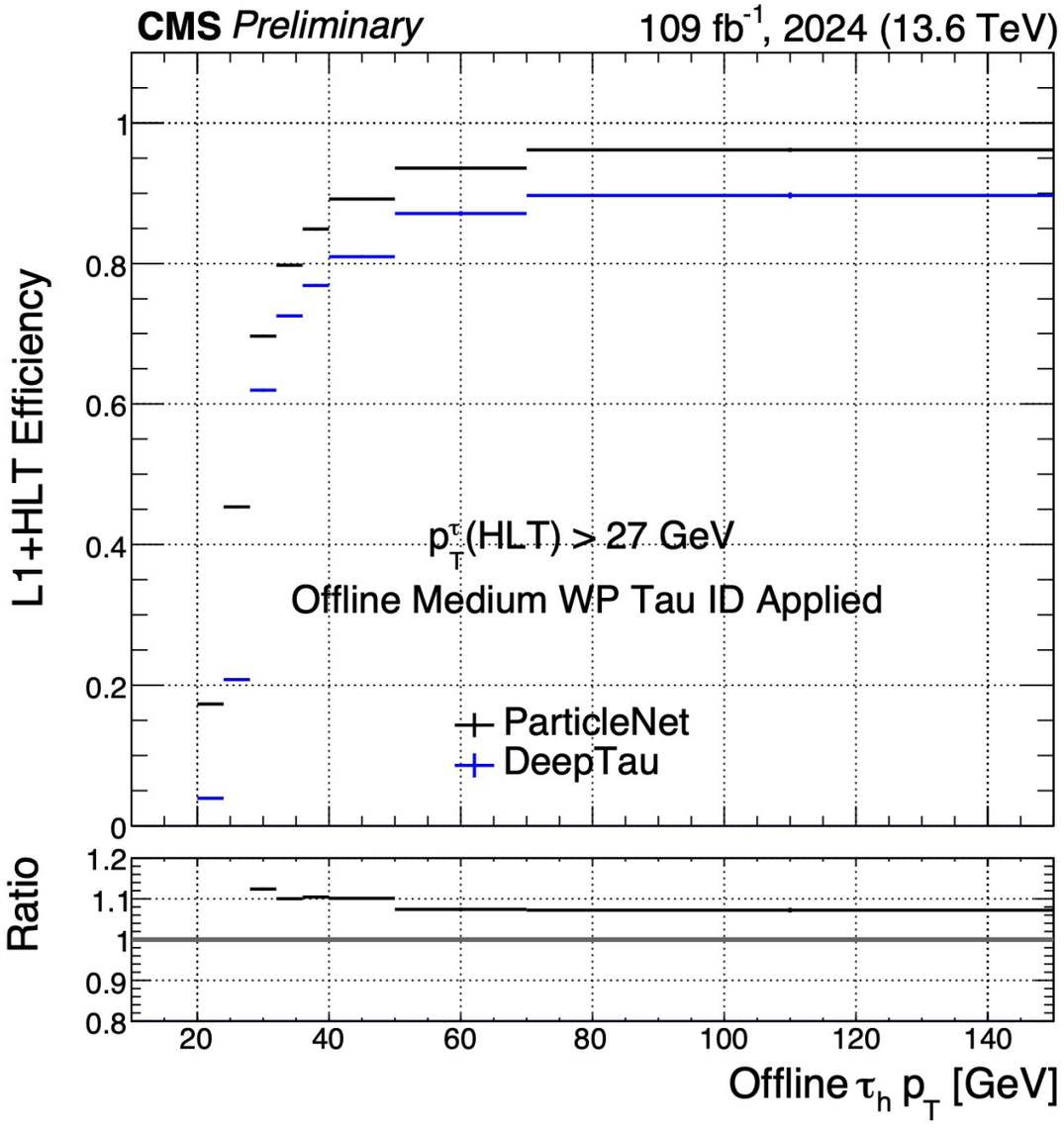


Improvement in Tau ID at HLT using ParticleNet

CMS-DP-2026/008

- Implemented since 2024
- Higher efficiency than DeepTau with comparable rate

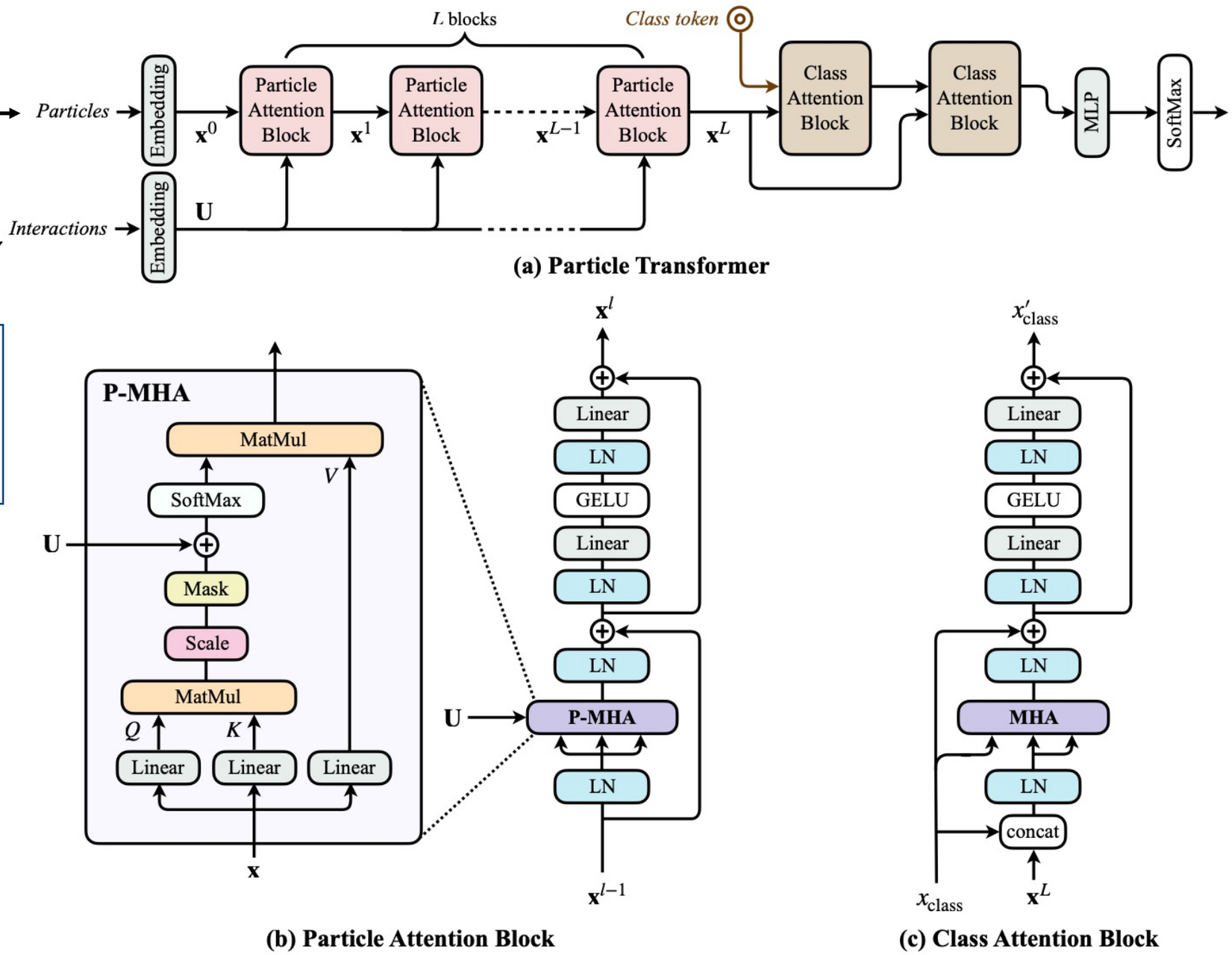




Transformers

Single particle information

Pair wise interaction features:
e.g. m, k_T, ΔR etc..



Similar to fully connected graph networks with efficient attention mechanism.

Efficiently models sequential data (arXiv:1706.03762)

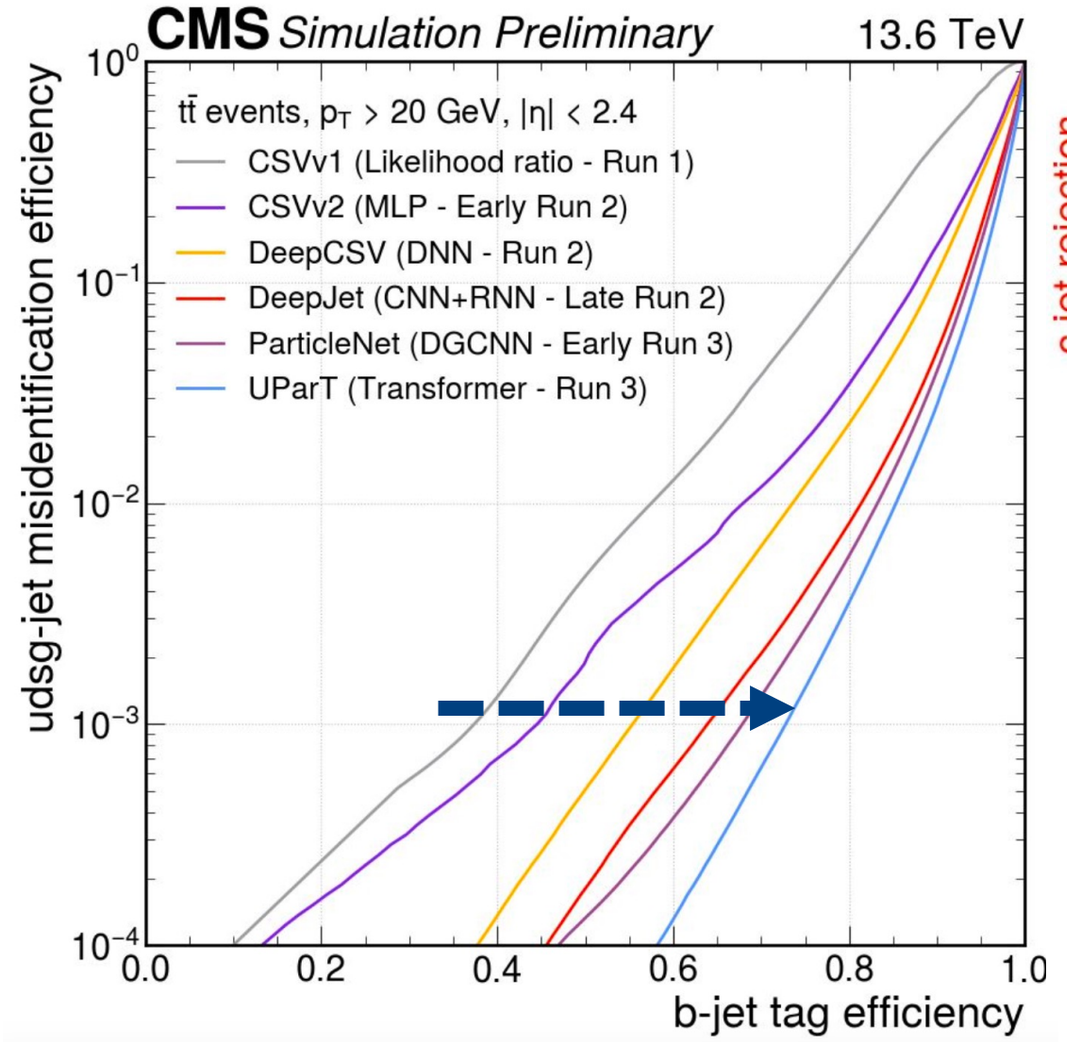
Several extensions for particle physics applications (arXiv:2202.03772)

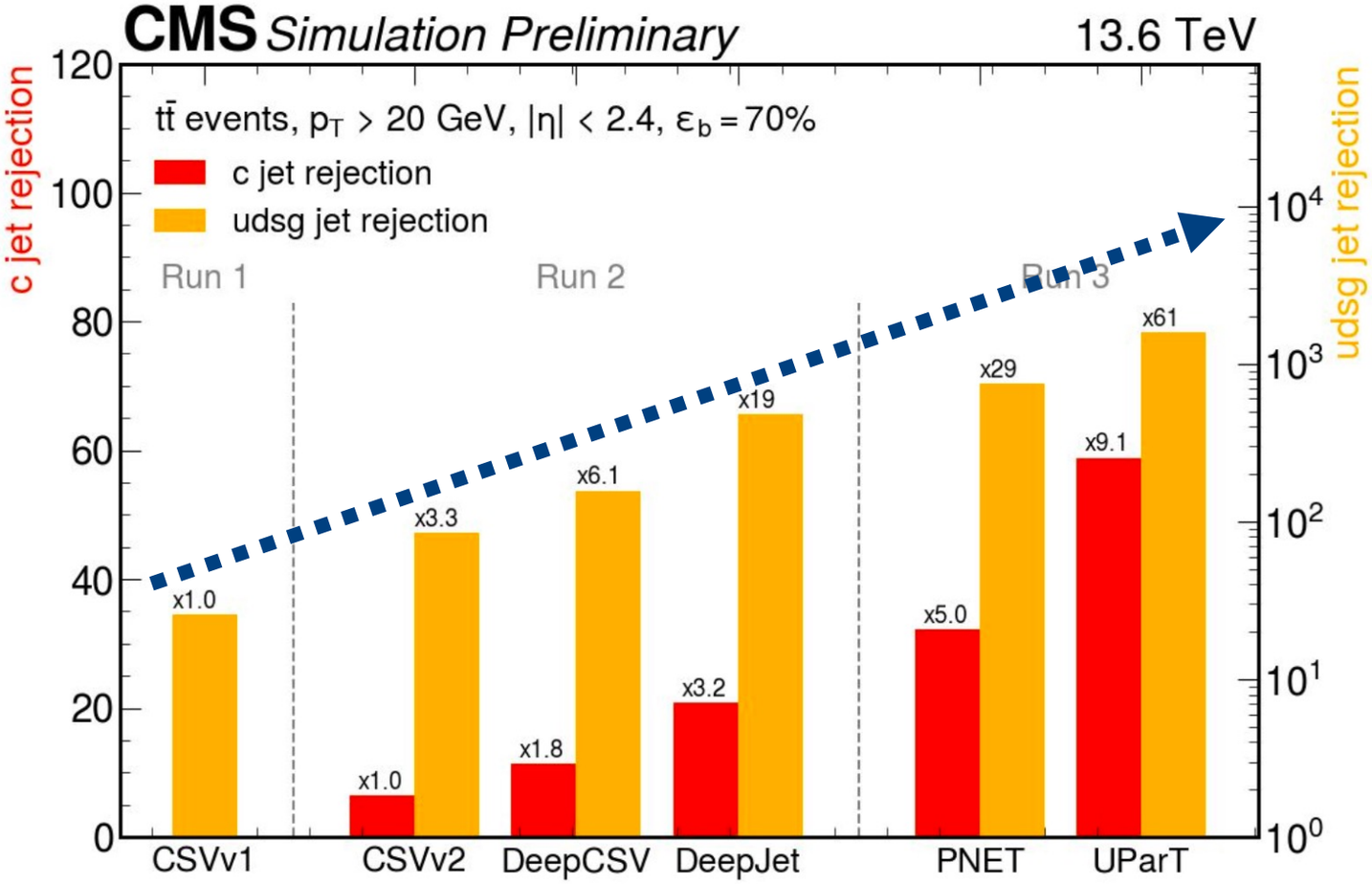
Unified approach to jet tagging [CMS-DP-2024-066]

Evolution of jet taggers → DNN to GraphNN → ParticleTransformer

Includes also jet energy regression and resolution

Adversarial training to improve Data/MC robust ness

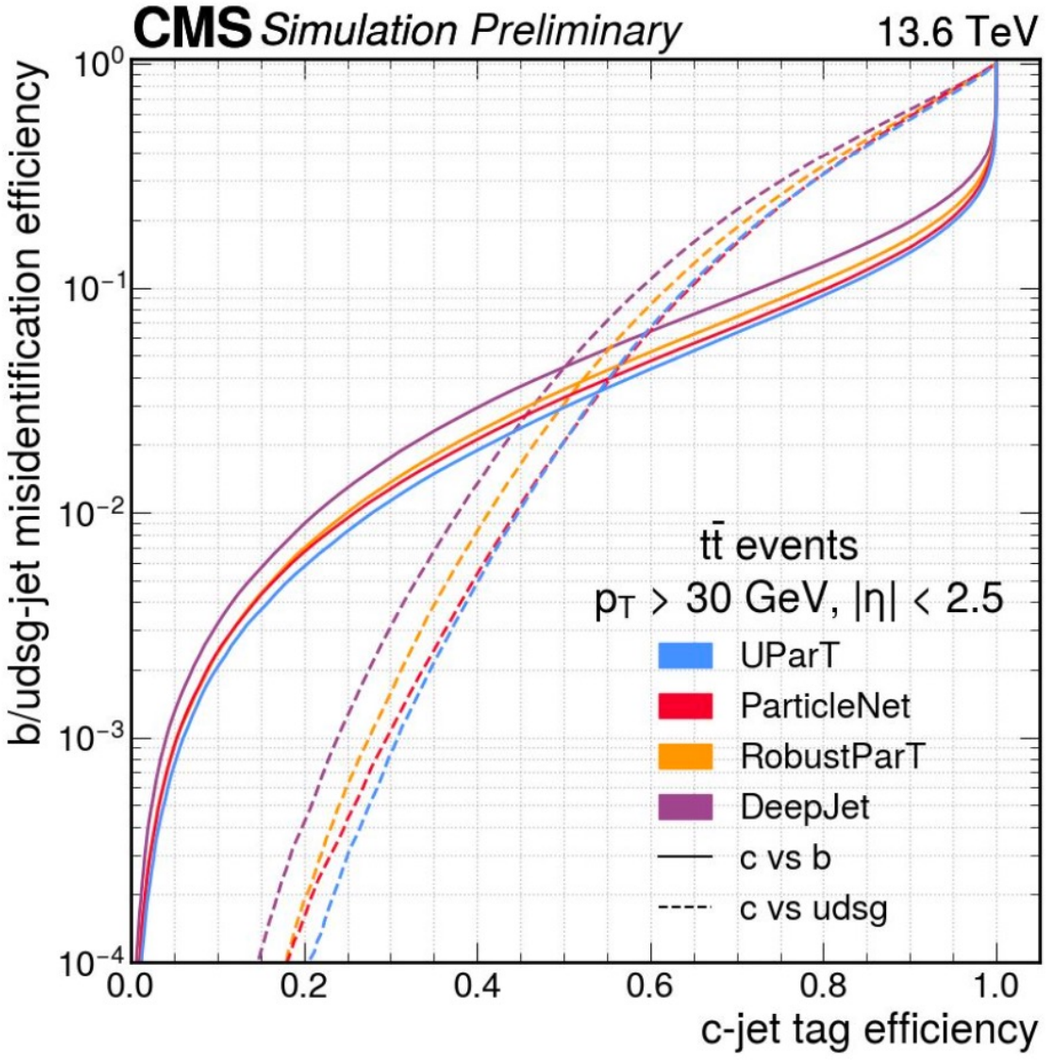


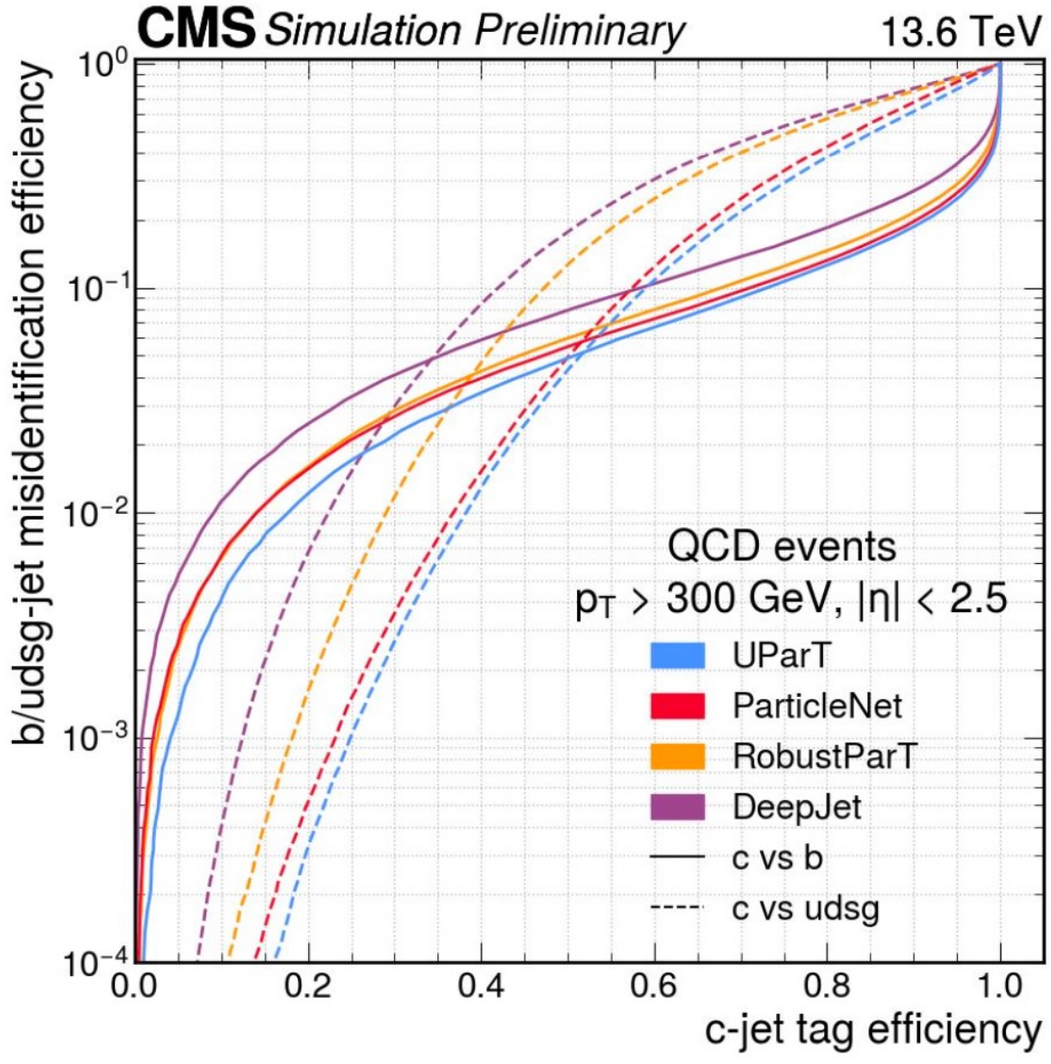


Improvement in performance with time

Improved c-jet tagging

CMS-DP-2024-066





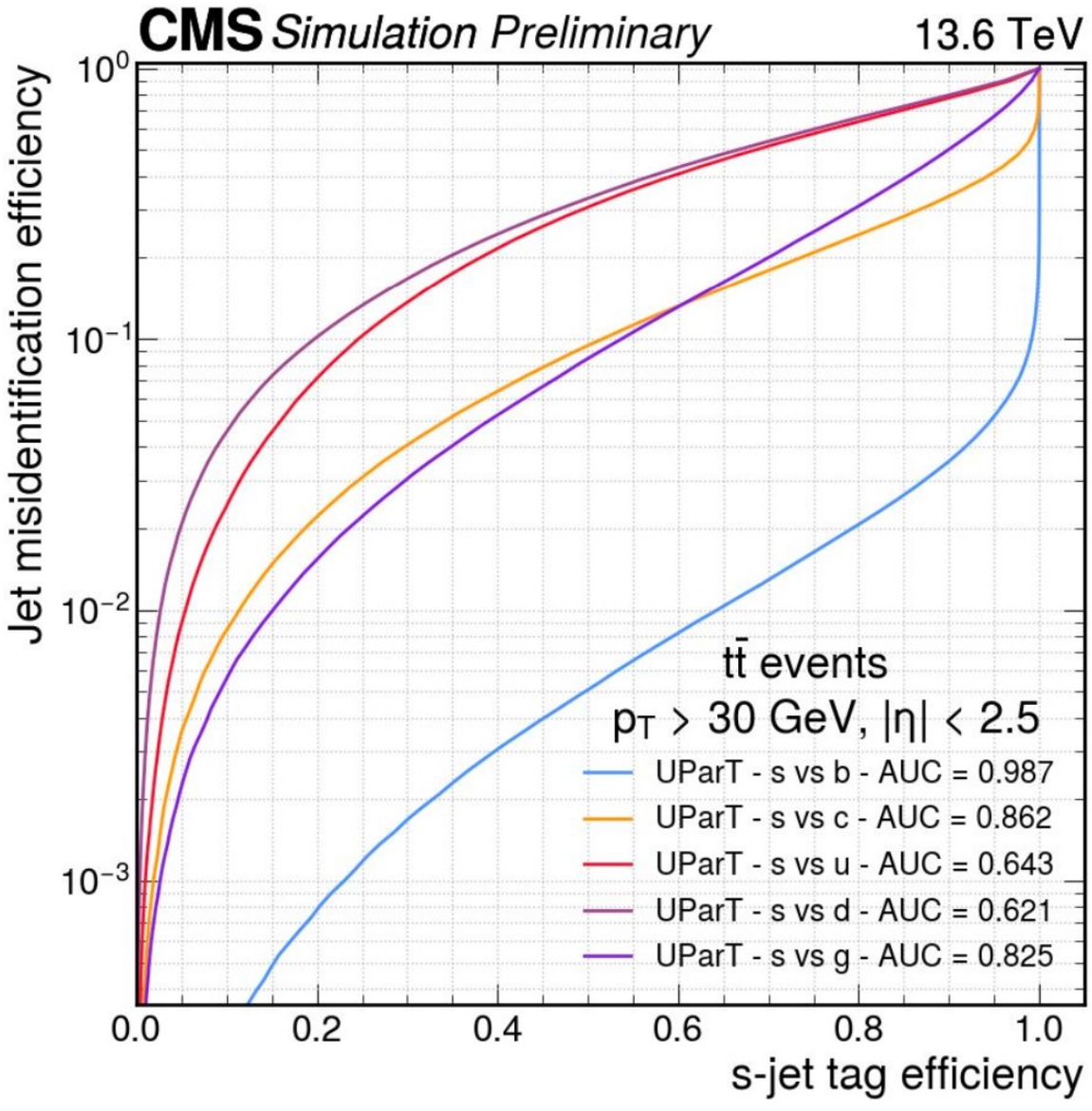
State-of-the-art performance for both b and light jets rejections

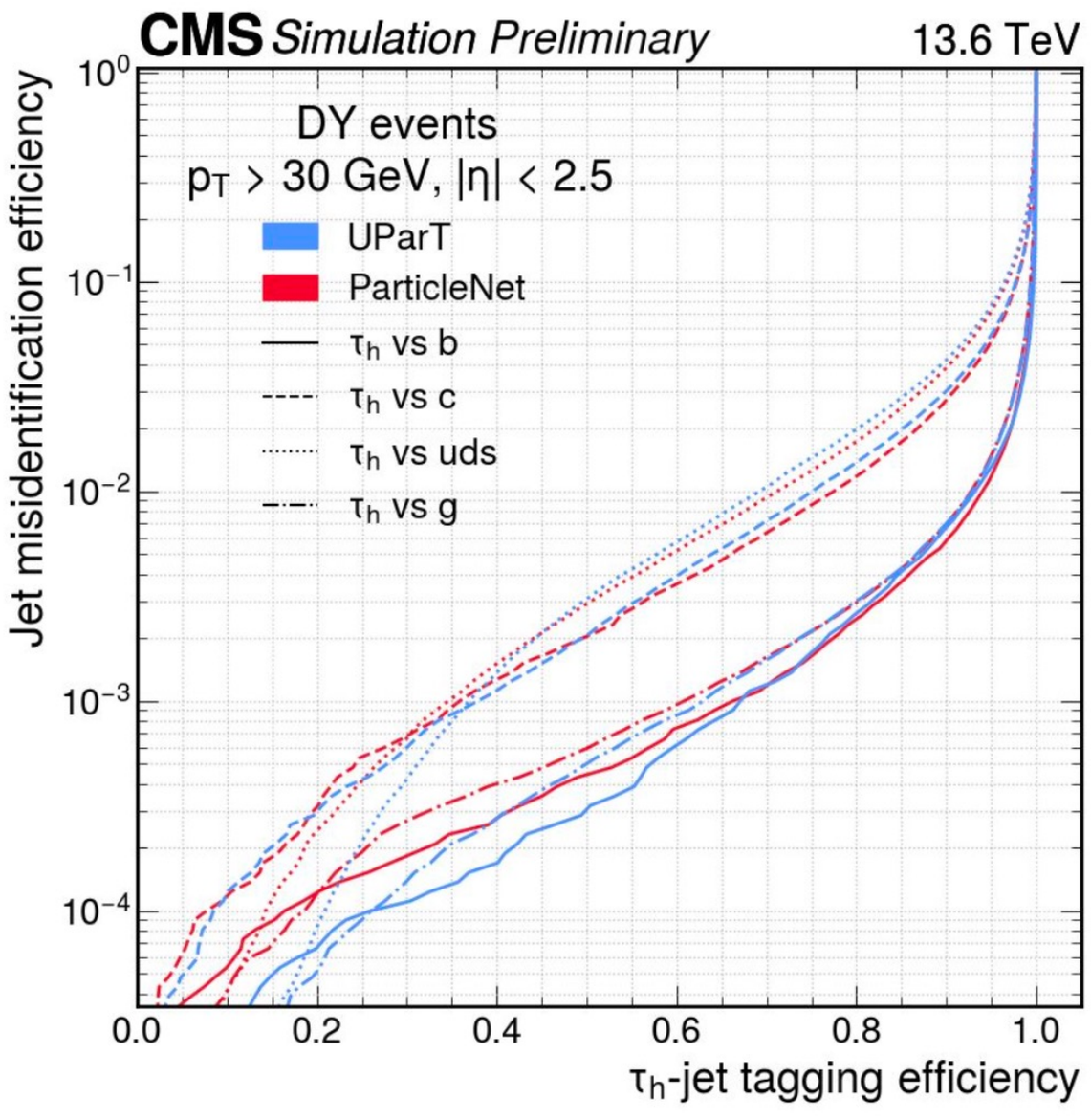
Extended class: tau and strange-jet tagging

CMS-DP-2024-066

ParticleNet / UParT can improve further tau-tagging performance

First time tagging of s – jets (with low efficiency)





ML Based Particle Flow Reconstruction

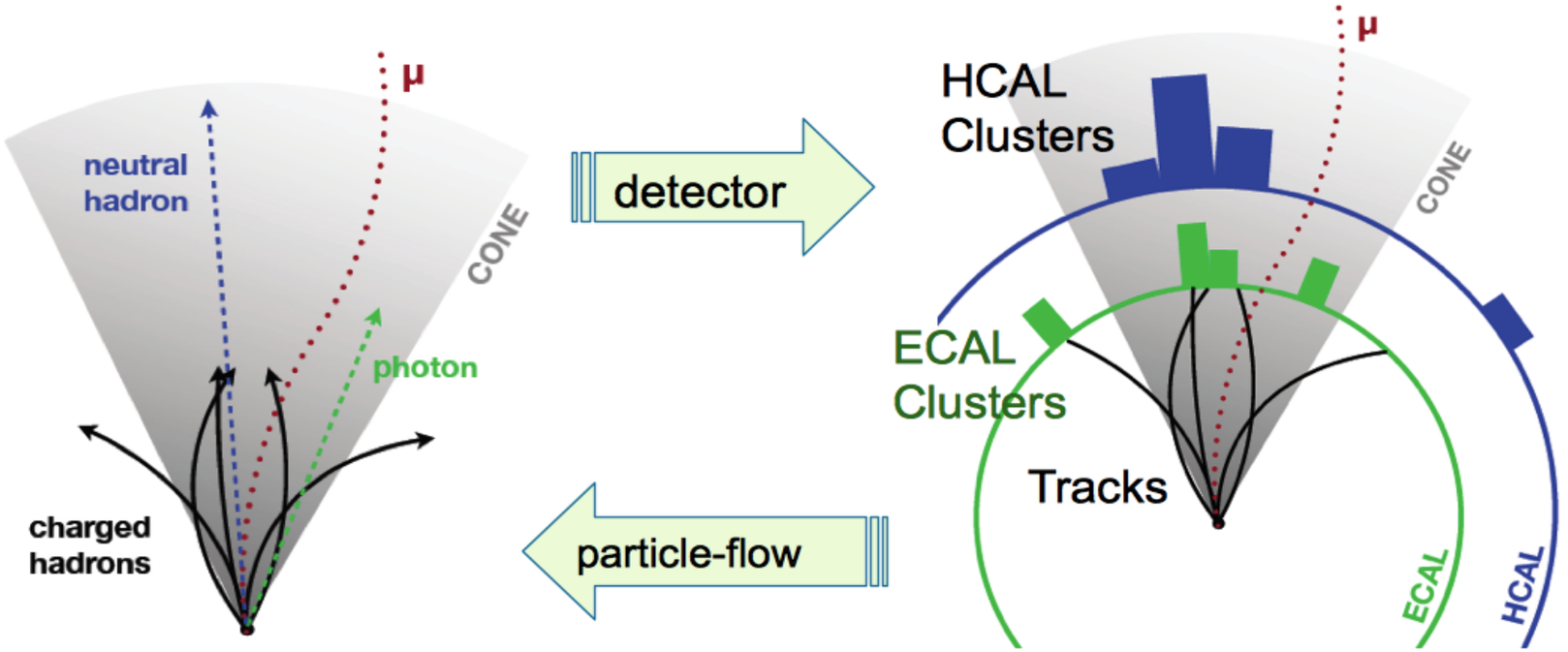
Global event reconstruction ➔ reconstruct and identify each individual particle in an event

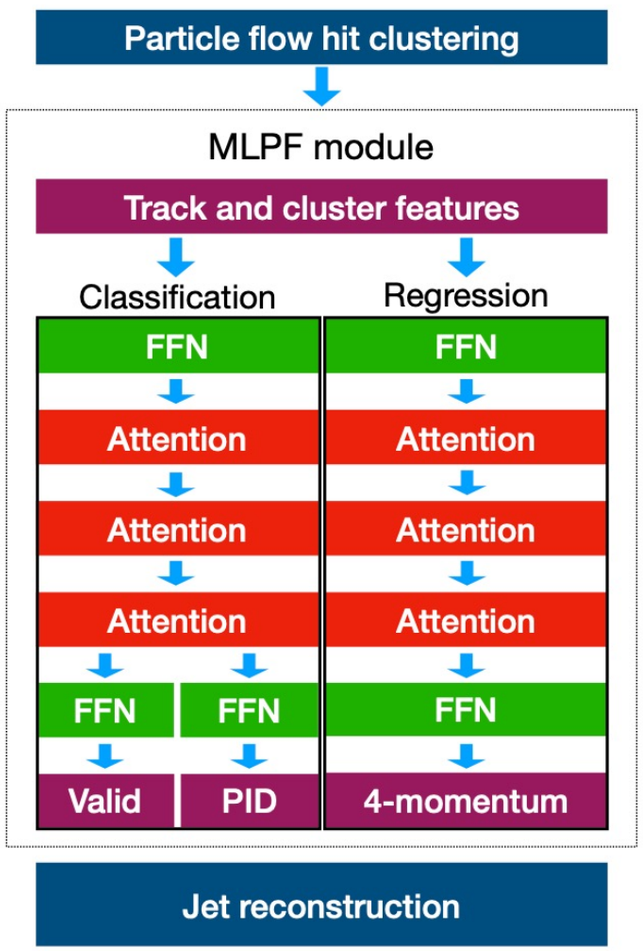
Current PF algorithm:

- Iterative, rule-based methods, such as proximity-based linking
  - i.e. tracks and calorimeter clusters are associated based on geometric closeness in detector coordinates
- Tries to optimally combine all detector information

ML for Particle-Flow:

- Uses a transformer based architecture
- Predicts all particles simultaneously at one inference step
- Expected to benefit more in HL-LHC, with very high Pileup and more granular detector
- Validated using Run-3 data





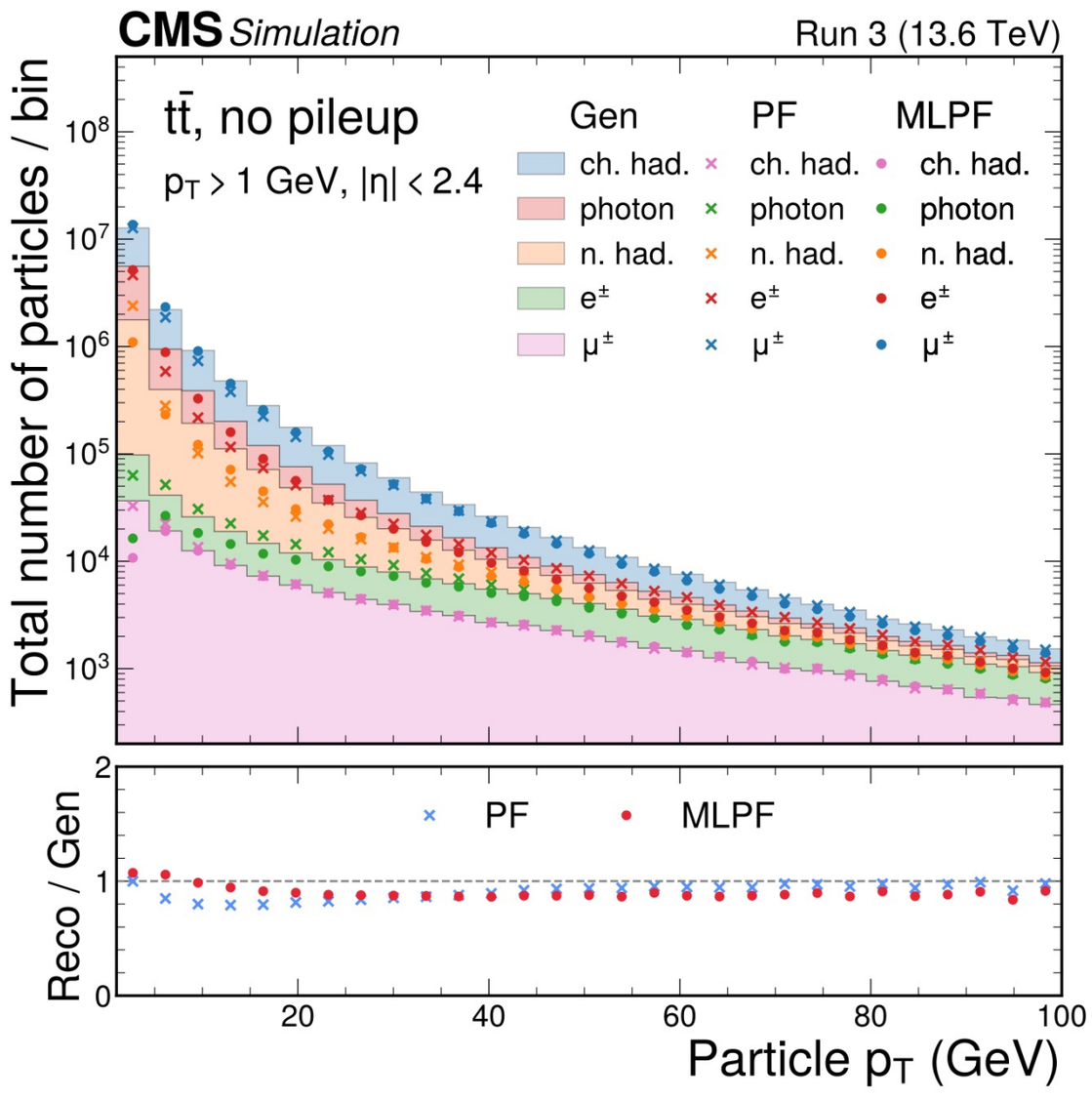
JINST 12 (2017) P10003

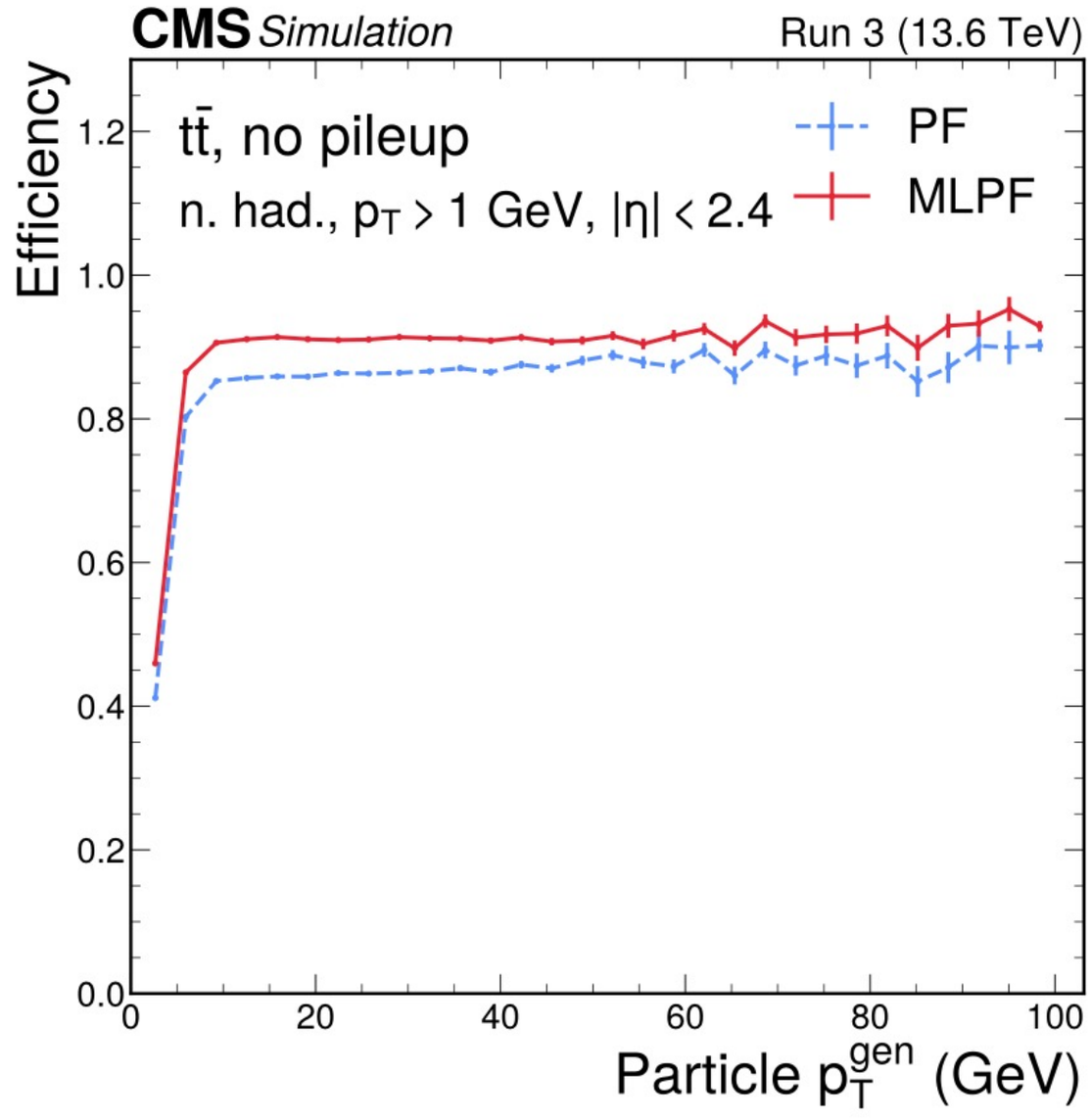
ML Based Particle Flow Reconstruction

CMS-PFT-25-001

Performance validated using tt and QCD multijets events

- Similar performance at particle level reconstruction & identification
- Improved efficiency for neutral hadron for similar fake rate

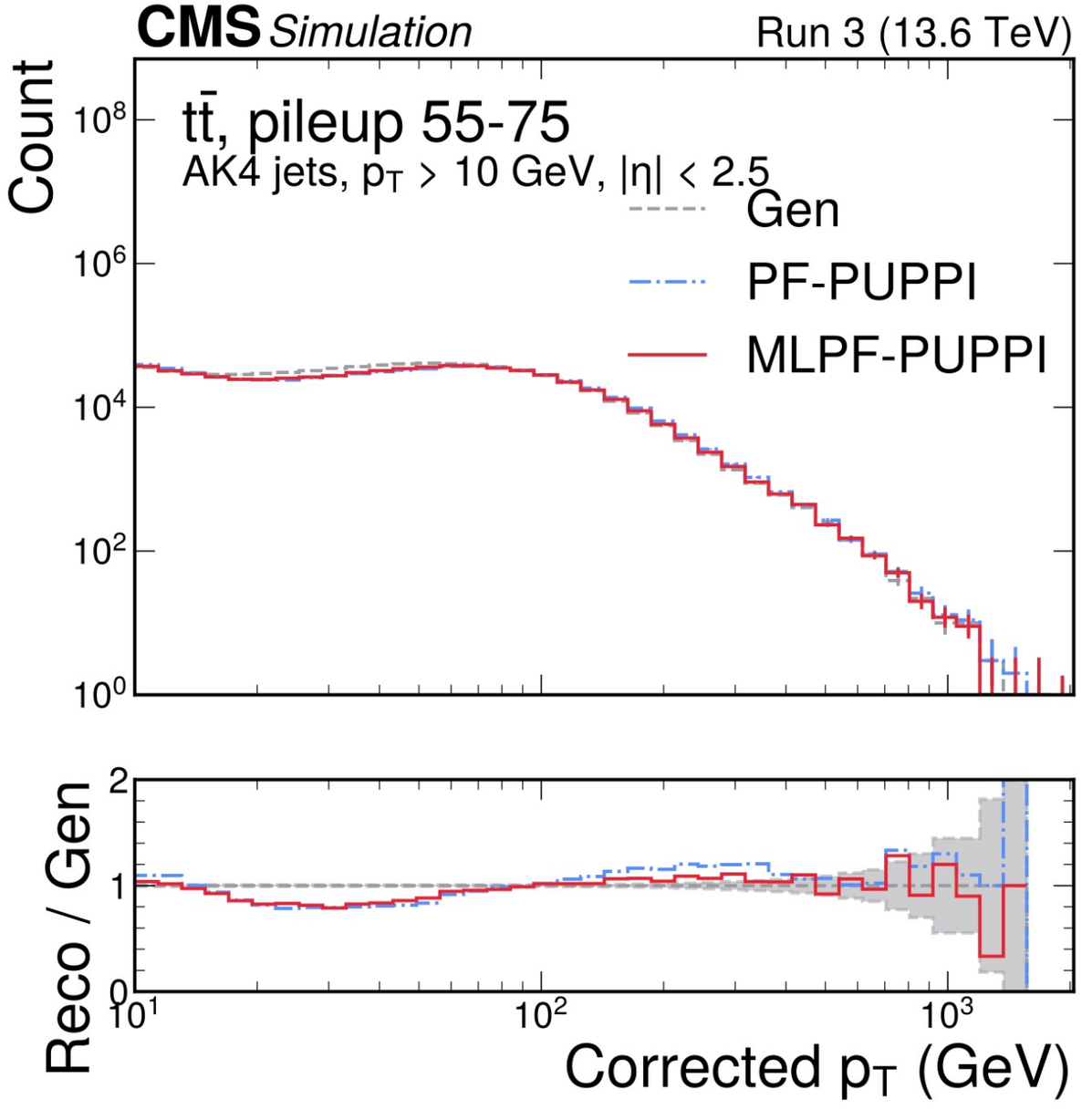


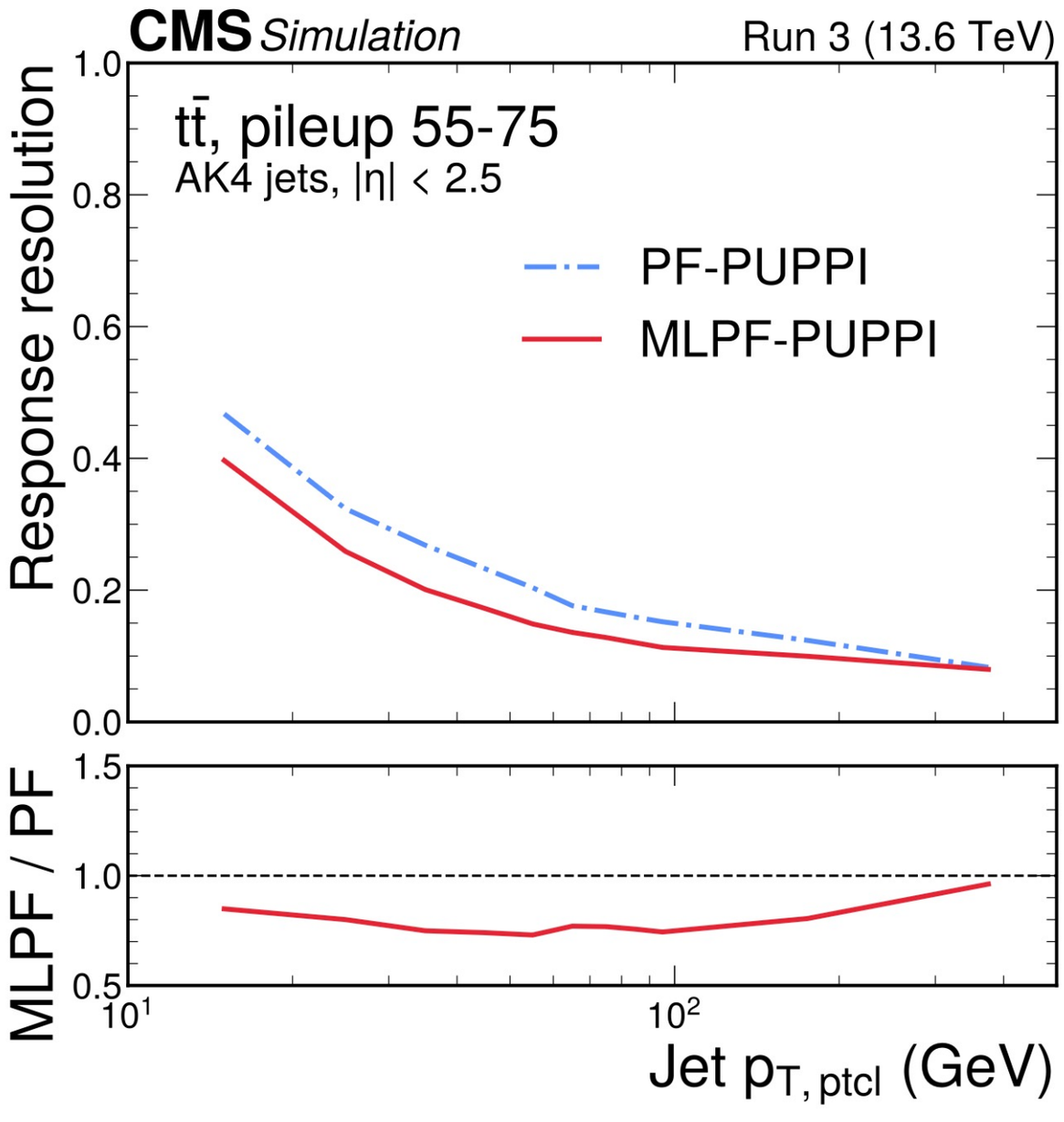


ML Based Particle Flow Reconstruction

CMS-PFT-25-001

Similar jet response compared to PF, and improved jet p_T resolution

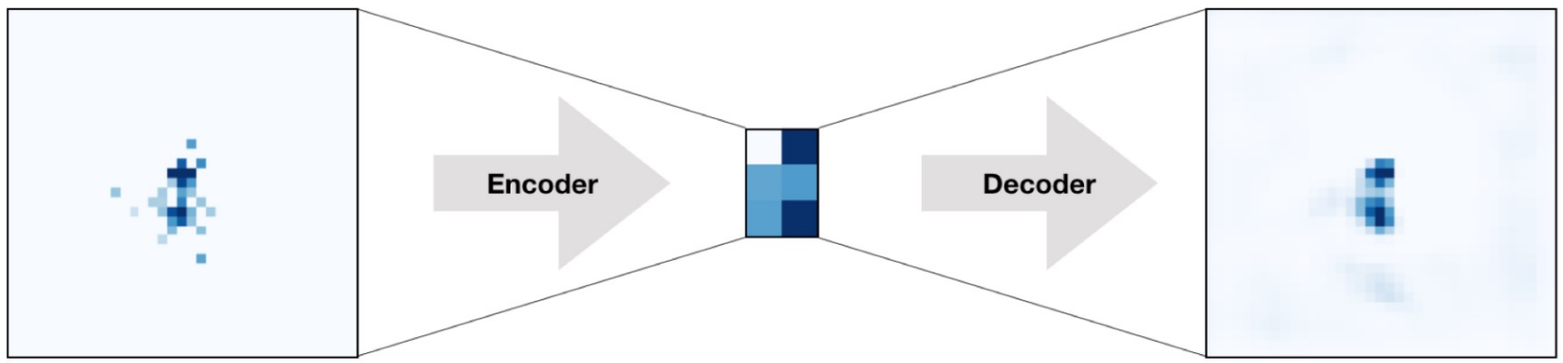




Anomaly Detection

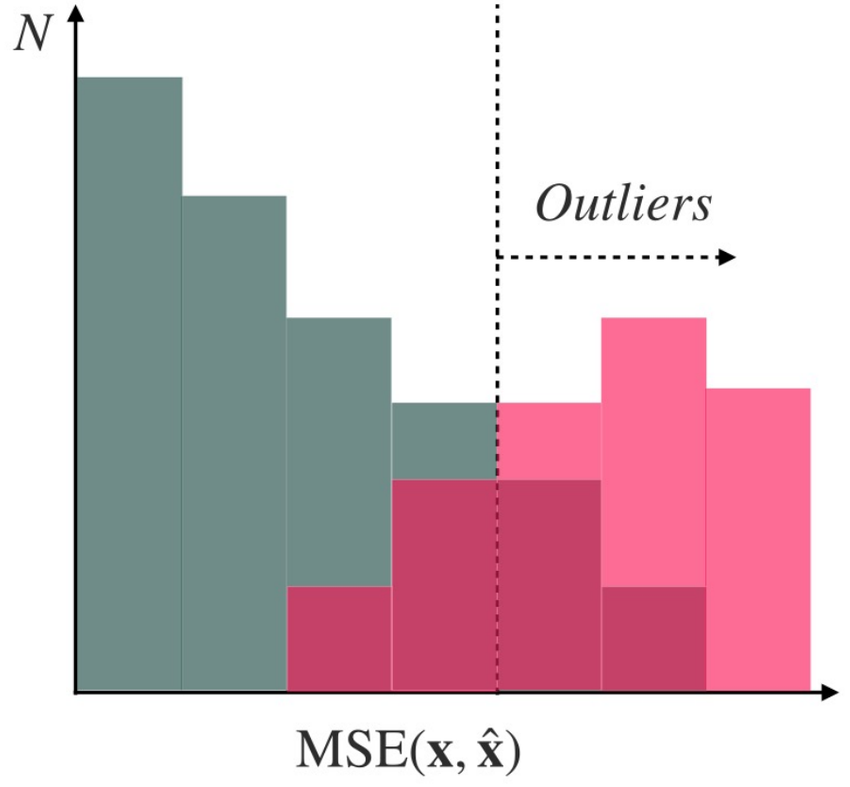
Unsupervised = no labels

e.g. Autoencoders



arxiv:1808.08992
arXiv:1808.08979

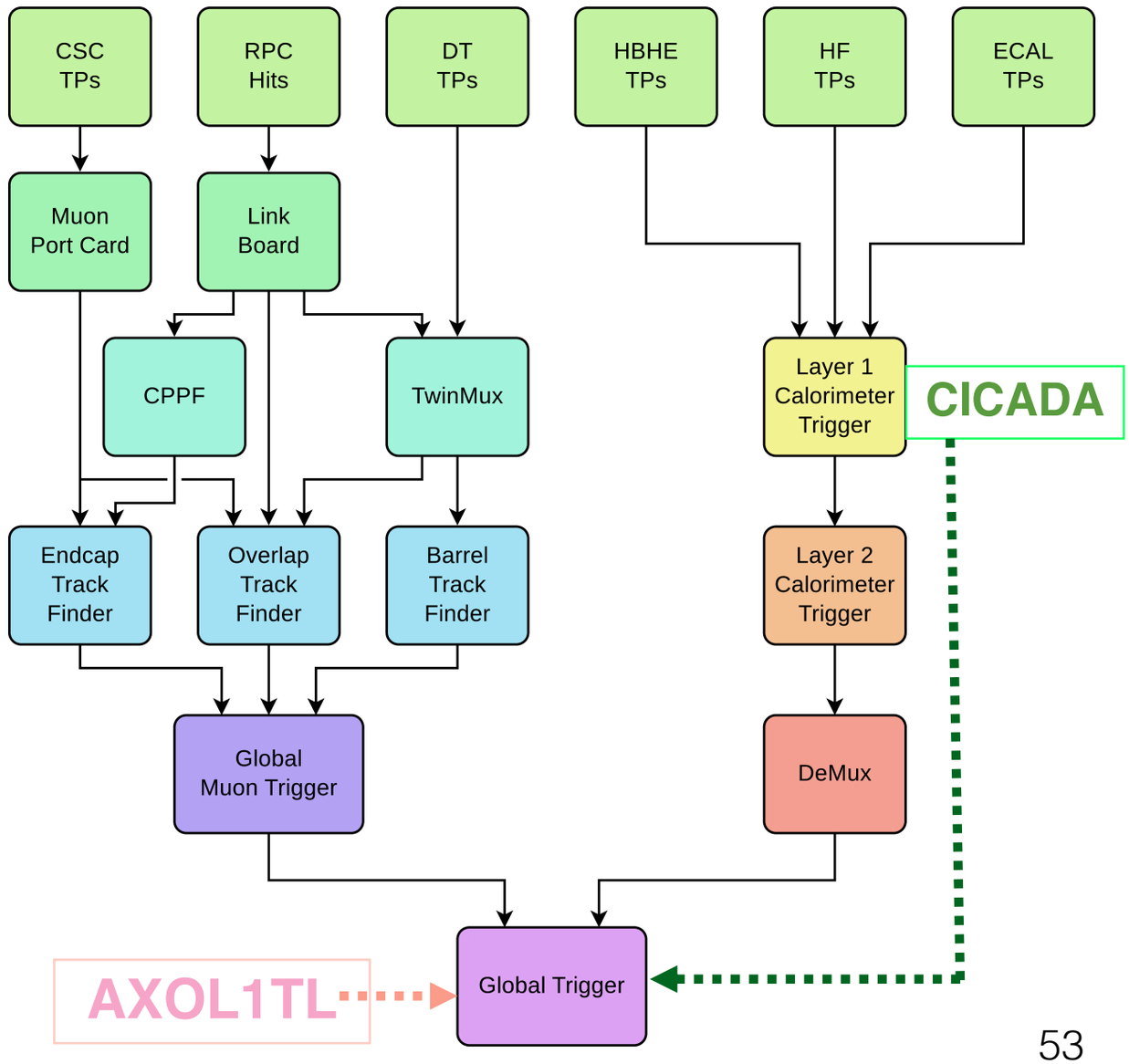
- Train network Compress events and then decompress them to get the original input
  - Anomaly events should have higher loss
- Can **train directly on data**, no labels needed
  - New physics searches in a model independent way
  - **Useful for triggering** events with unknown signatures



ML based anomaly detection trigger

- Trigger is essential for recording data at LHC
  - But most triggers are designed aiming for specific signatures → Model dependent
    - Maybe new physics events NOT simply selected at the trigger level

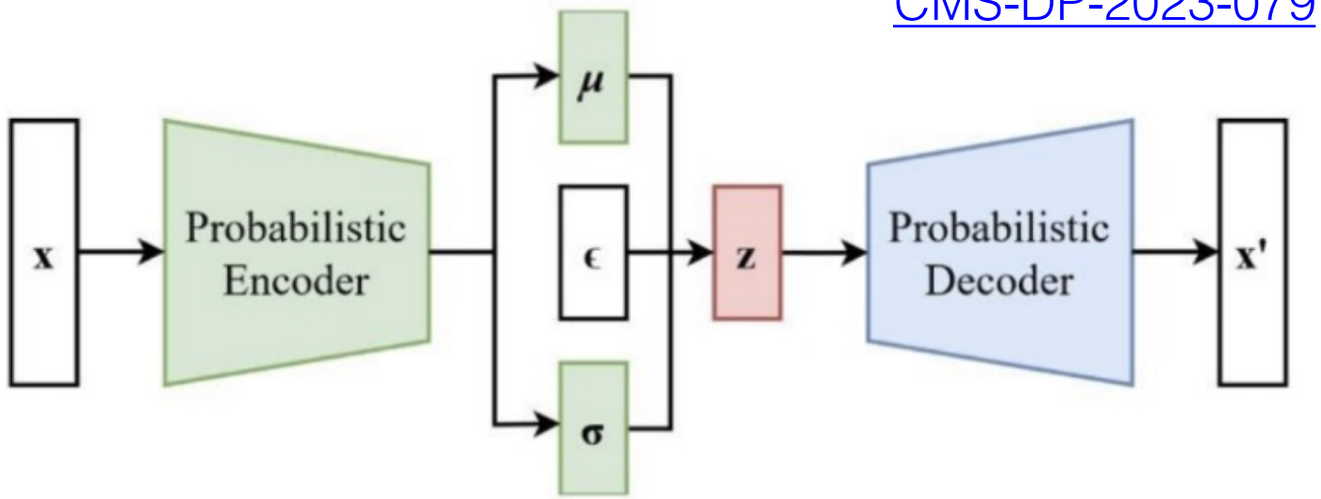
- **ML-based anomaly triggers first time in CMS** @ LHC Run-3
  - **A**nomaly e**X**traction **O**nline **L1 T**rigger **L**ightweight in L1Global Trigger
  - **C**alorimeter **I**mage **C**onvolutional **A**nomaly **D**etection **A**lgorithm in L1 Calo Layer 1
- Implementation in FPGA using [HLS4ML]



AXOL1TL

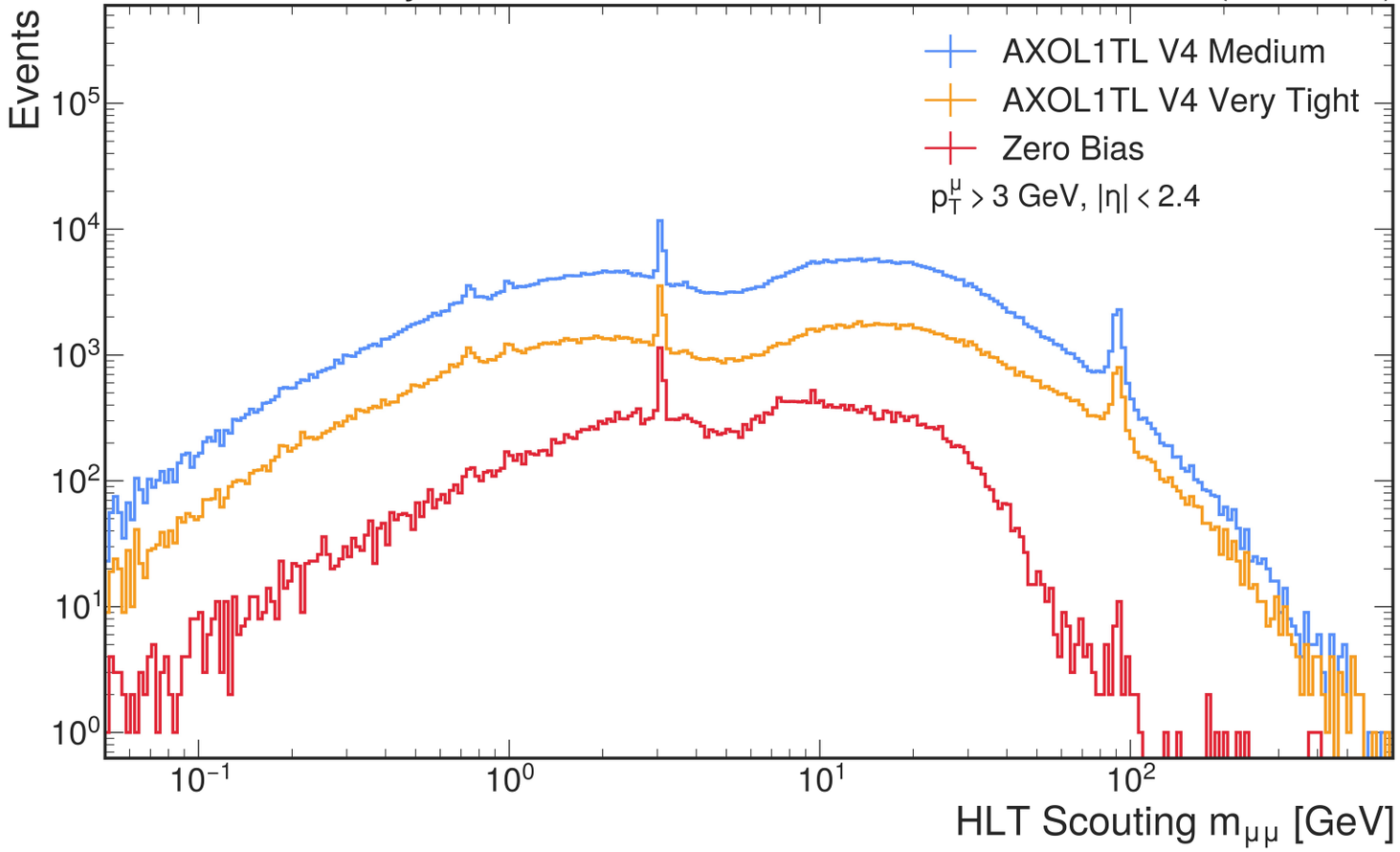
CMS-DP-2023-079

Variational autoencoder (VAE) trained on ZeroBias data to detect data outliers



$$\text{Loss} = (1-\beta)\|x-\hat{x}\|^2 + \beta\frac{1}{2}(\mu^2+\sigma^2-1-\log\sigma^2)$$

Reconstruction term — Full regularization term



Invariant mass distribution of muon pairs collected from HLT Scouting events by different AXOL1TL V4 trigger paths

CMS-DP-2025-061

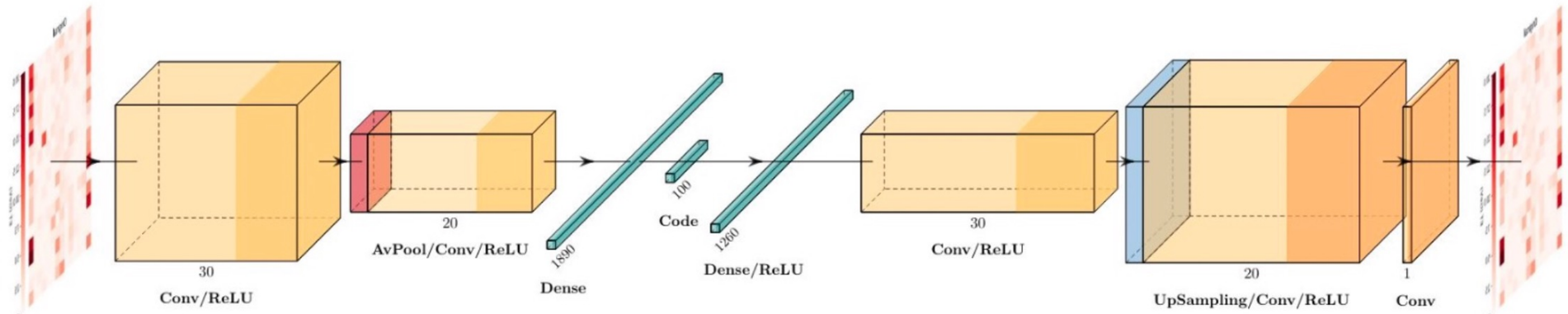
CICADA

https://cicada.web.cern.ch/

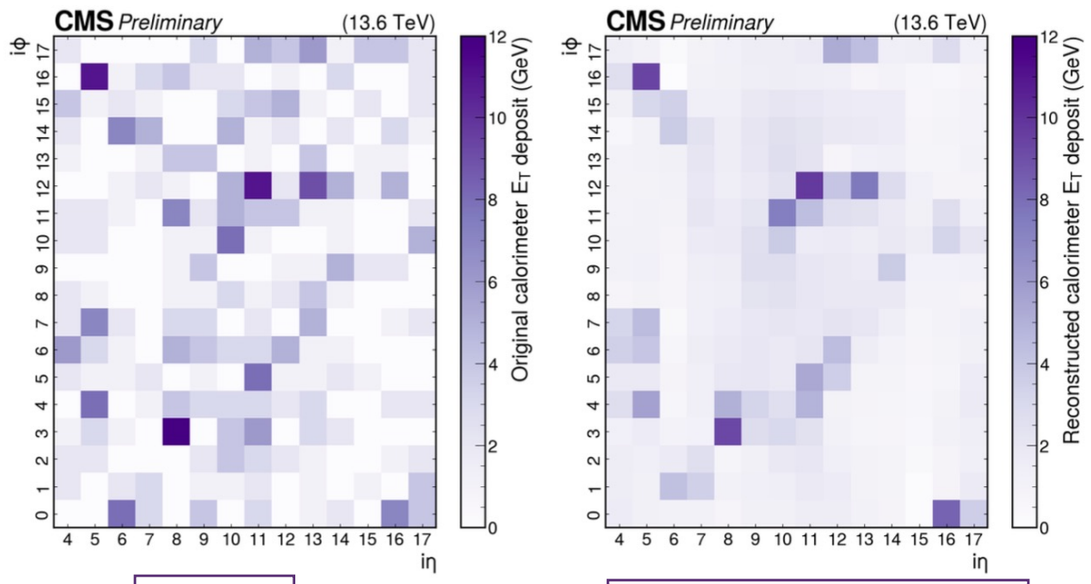
Convolutional neural network auto-encoder

CMS DP -2024/121

- Inputs from CaloLayer1
- Low level information, not dependent on jet reconstruction etc..
- MSE as anomaly score



ZeroBias

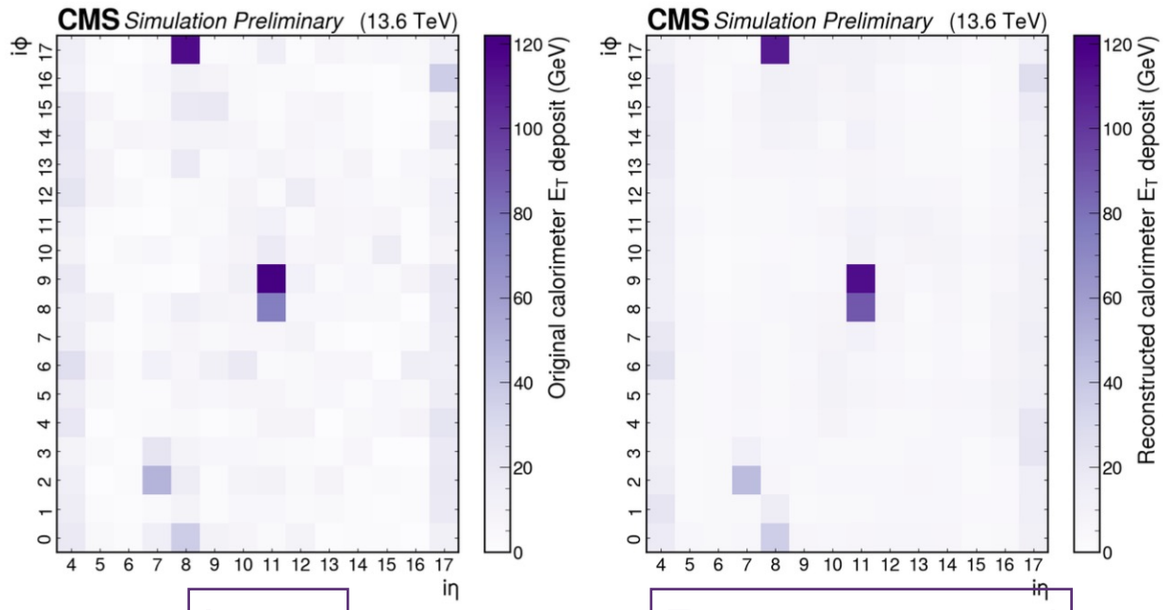


Input

Reconstructed

MSE = 2.57

SUSYGGBBH signal



Input

Reconstructed

MSE = 14.89

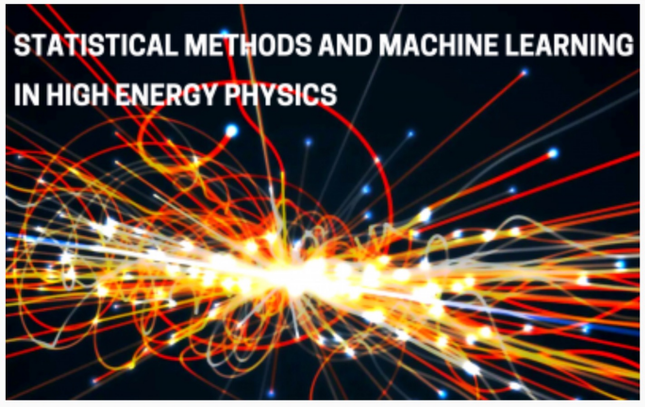
Summary

- Rapid development of Machine Learning techniques in past decade
- Deep learning revolutionized the way large data samples are analyzed in HEP experiments
    - Improvement in several areas: from online event selection to offline data analysis
        - Increased sensitivity to BSM signatures
- ML is currently used at all stages of the experiments: From Data collection to physics inference
- Discussed only a few applications from CMS experiment

Resources

A Living Review of Machine Learning for Particle Physics

https://iml-wg.github.io/HEPML-LivingReview/

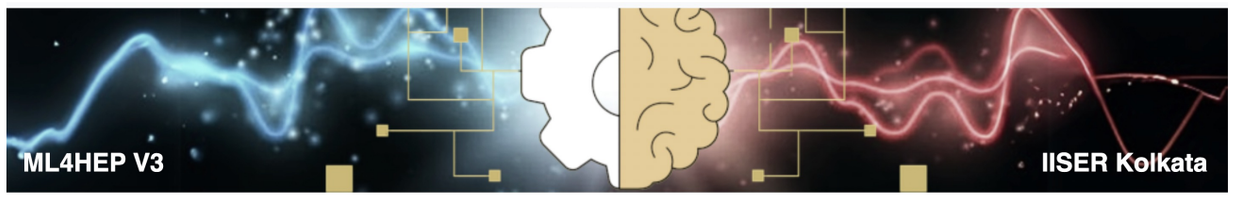


Machine learning school @ ICTS,
28 August 2023 to 08 September 2023

https://www.icts.res.in/program/ML4HEP

Machine learning school @ IOP
1st - 13th July 2024

https://iopb.res.in/ml4hep/



School and workshop on Statistical Methods and Deep Machine Learning in High Energy Physics and Astrophysics

Machine learning school @ IISERK
30th June – 12th July 2025

https://www.iiserkol.ac.in/~ml4hep/

Keep an eye for this years school @TIFR, Mumbai

Thanks