LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS
partículas e tecnologia

[MACHINE LEARNING *at Colliders*]

Rute Pedro | 24th March

Café com Física | Universidade de Coimbra

POCI/01-0145-FEDER-029147
PTDC/FIS-PAR/29147/2017

Fundação para a Ciência e a Tecnologia · Lisb@2020 · COMPETE 2020 · PORTUGAL 2020 · UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional · ATLAS EXPERIMENT · Big Data HEP

Outline

Machine Learning: key concepts

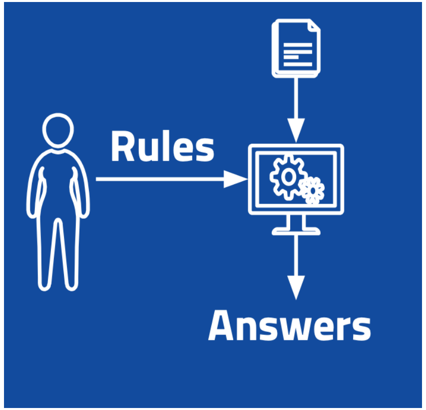
ML applications to Particle Physics

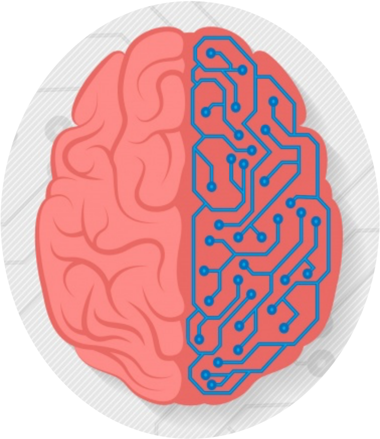
ML for Anomaly Detection: a tool for New Physics searches

What is Machine Learning?

Traditional Computation

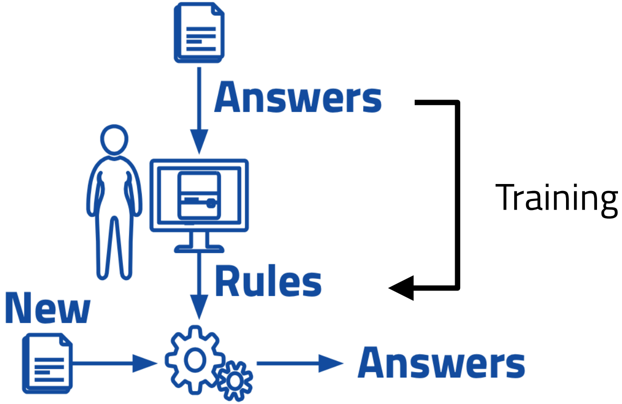
The task is programmed by the user as a pre-defined set of rules/algorithms to apply to data





Machine Learning (ML)

The program learns from data what are the necessary rules to execute a task/objective defined by the user: Training

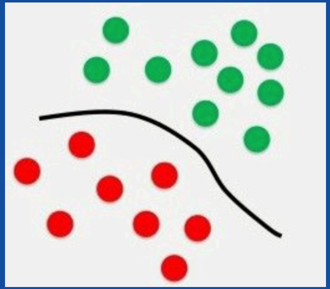


From M. C. Romão [LIP seminar]

ML tasks

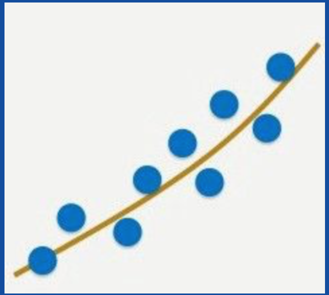
Classification

Discrete prediction



Regression

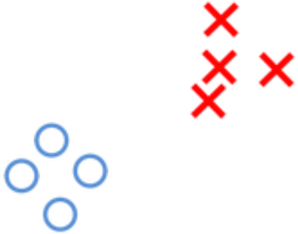
Real-value prediction



Learning types

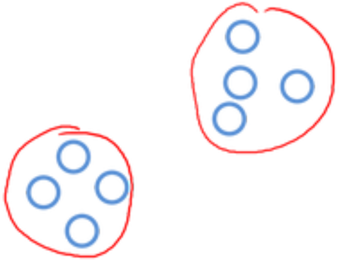
Supervised

(E.g. Simulation in Particle Physics)

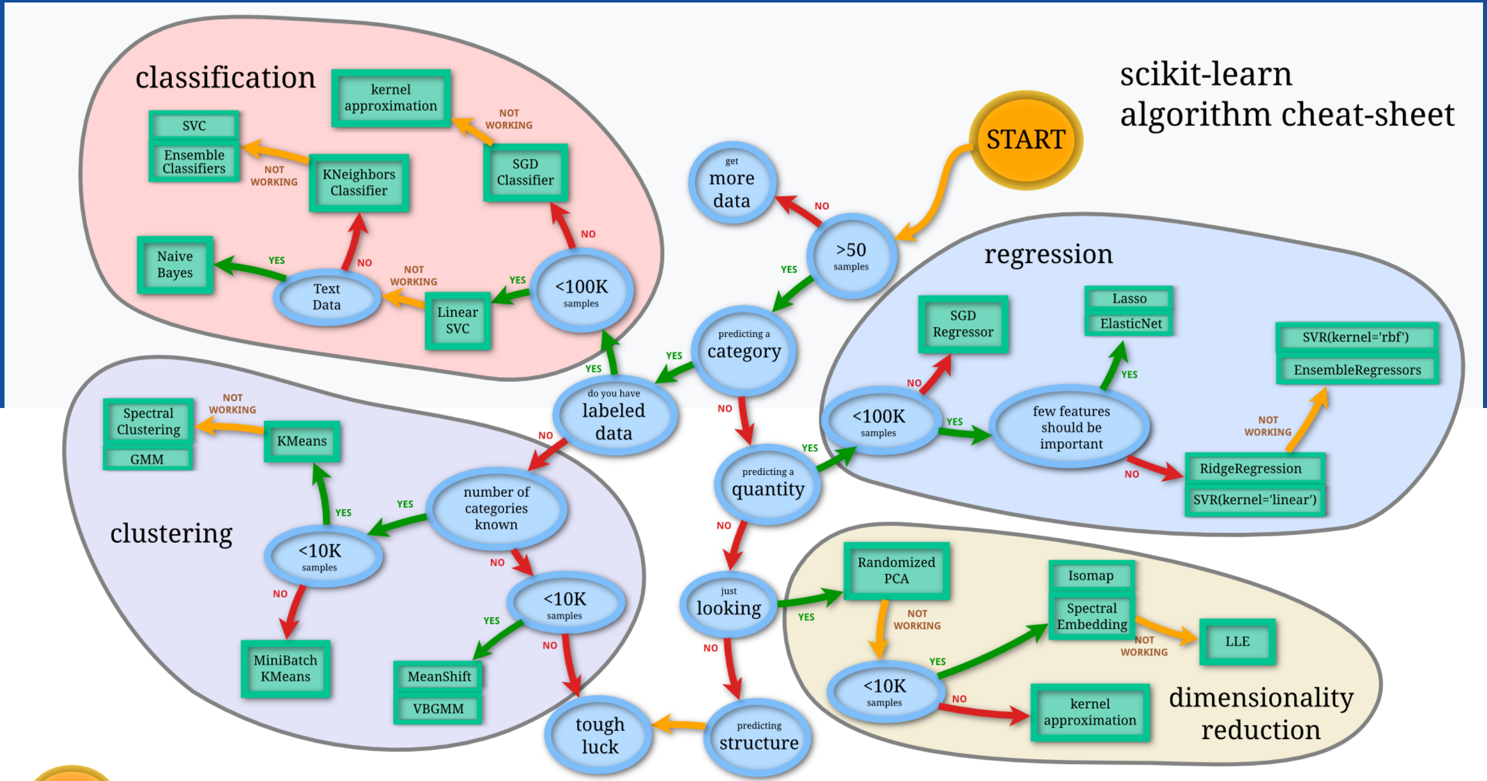


Unsupervised

(E.g. clustering)



... an entire ecosystem



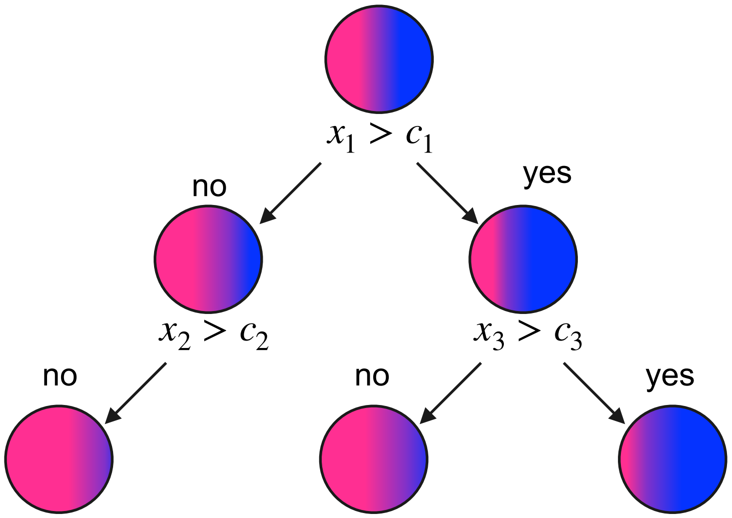
Back

Scikit-Learn: excellent ML library to start with, Python-based

Besides algorithms, it also contains data

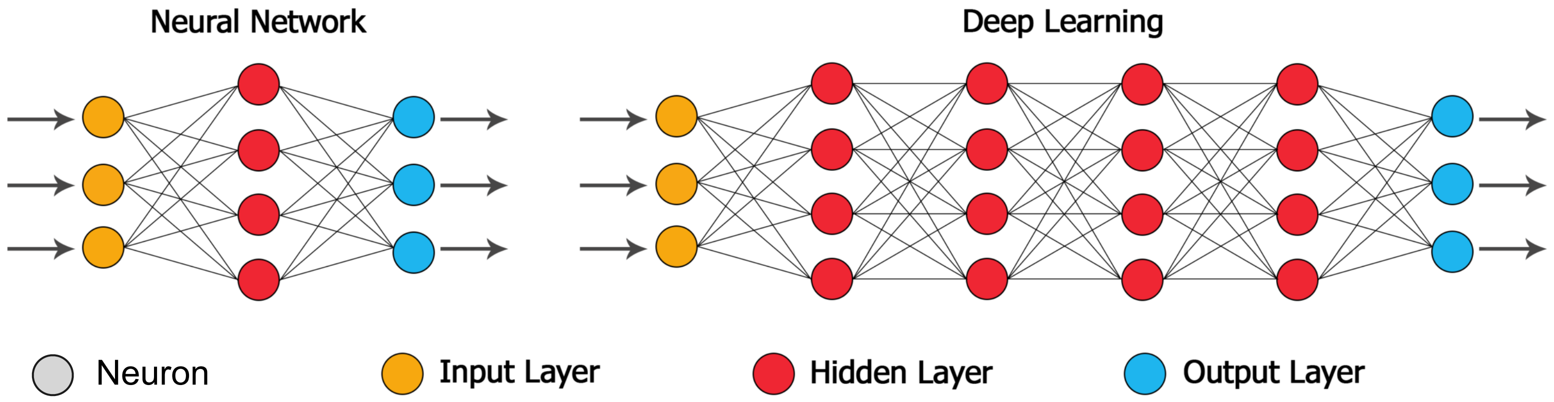
Shallow Learning

Decision Tree



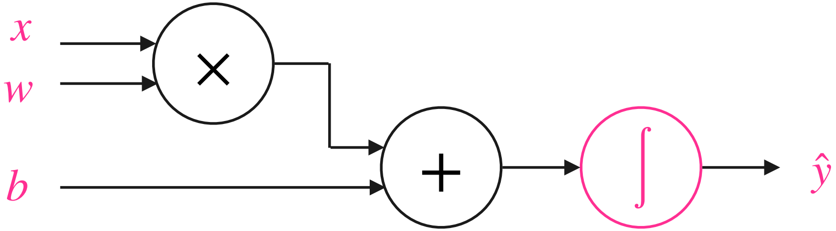
- $\vec{x}$ input features
- Labeled samples of data: **blue**/pink
- Partitions the data to increase sample purity
- Finds optimal criteria $x_i > c_i$ to separate data categories
- Category prediction based on the label of the majority samples of the end leaf
- Core of the most popular algorithms used in LHC event classification (Boosted Decision Trees)

Deep Learning



- Neural networks with many hidden layers, each with a given number of artificial neurons
- Capable of highly non-linear representations of the data
- In principle, can model any function
- Architecture -> hyper-parameters: number of layers, number of neurons/layer, …

Artificial Neuron

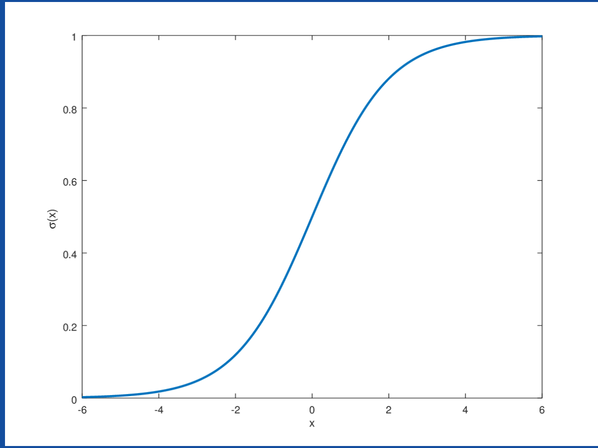


- x is the input feature
- y is the target feature (or "label")
- w, b are the model trainable parameters
- $\hat{y}$ is the output (model prediction)

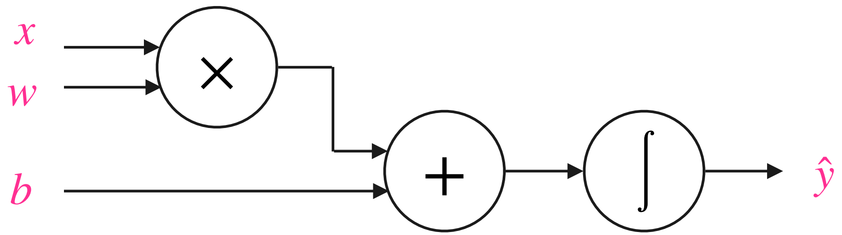
Activation function

- e.g. linear for regression
- e.g. sigmoid for classification

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



Loss function and Training Objective



Loss function L: measure of how good is $\hat{y}$ in predicting y

- e.g. Mean squared error: $L = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2$
- e.g. Binary cross-entropy: $L = \frac{1}{N}\sum_{i}^{N} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i)$

Training objective: find w, b that minimise the Loss function

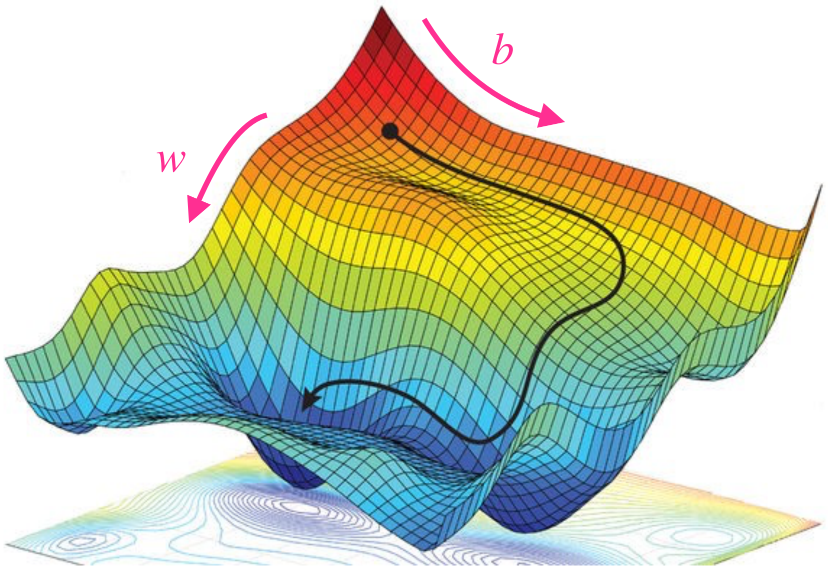
Gradient Descent and Back-propagation

Loss minimisation: **descend the Loss surface**

- $L = f(\hat{y})$
- Loss gradient $\dfrac{\partial L}{\partial \hat{y}}$

Back-propagate the Loss gradient (iteratively)

- $\dfrac{\partial L}{\partial w} = \dfrac{\partial L}{\partial \hat{y}} \dfrac{\partial \hat{y}}{\partial w}$ and update $w \leftarrow w - \alpha \dfrac{\partial L}{\partial w}$
- $\dfrac{\partial L}{\partial b} = \dfrac{\partial L}{\partial \hat{y}} \dfrac{\partial \hat{y}}{\partial b}$ and update $b \leftarrow b - \alpha \dfrac{\partial L}{\partial b}$
- α is an hyper-parameter that adjusts the learning rate



Loss surface

Practicable Deep Neural Networks

Many layers + many units

- **Vanishing gradient**: new activation functions made training possible (ReLU) (~2010)
- Advances in hardware: **GPU** increased speed of computation by 100 (~2010)
- APIs: **Keras**, **Tensorflow** (2015)

Deep learning

- Many parameters to estimate: $\{\vec{w}, \vec{b}\}$
- **Data** thirst

```
Layer (type)                 Output Shape              Param #
=================================================================
flatten_10 (Flatten)         (None, 784)               0
_________________________________________________________________
dense_22 (Dense)             (None, 128)               100480
_________________________________________________________________
activation_19 (Activation)   (None, 128)               0
_________________________________________________________________
dense_23 (Dense)             (None, 128)               16512
_________________________________________________________________
activation_20 (Activation)   (None, 128)               0
_________________________________________________________________
dense_24 (Dense)             (None, 10)                1290
_________________________________________________________________
activation_21 (Activation)   (None, 10)                0
=================================================================
Total params: 118,282
Trainable params: 118,282
Non-trainable params: 0
_________________________________________________________________
```

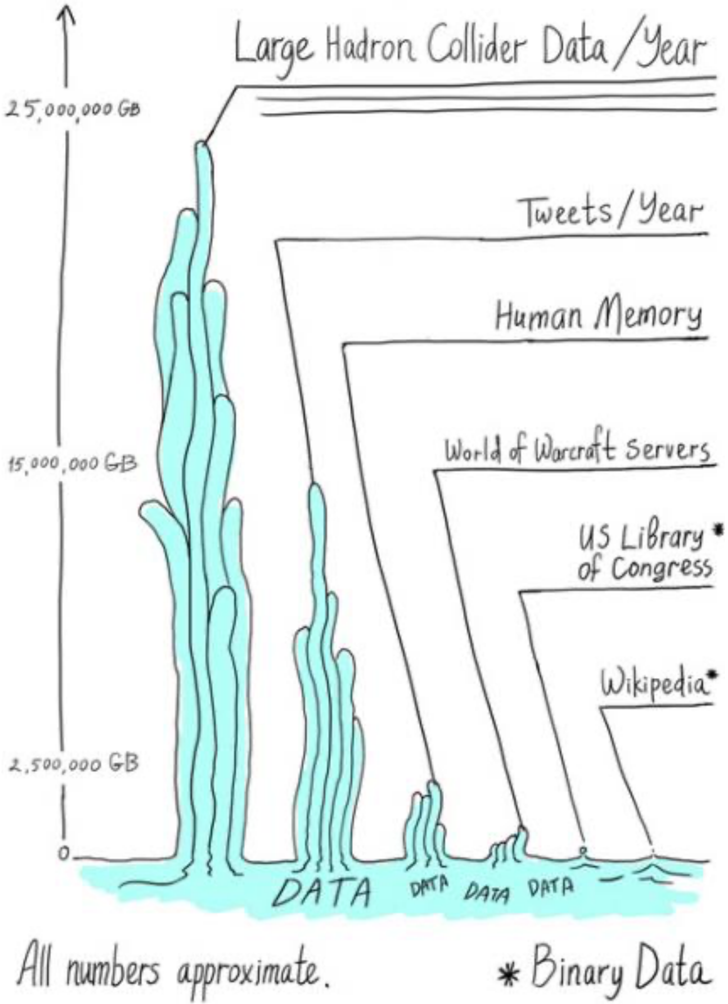
ML in Collider Physics

Rich ground for ML applications

LHC is an enormous source of data

- Number of collisions: 40 MHz, 1kHz recorded
- High data dimensionality: O(100 M) readout units

- Involves also large simulation datasets

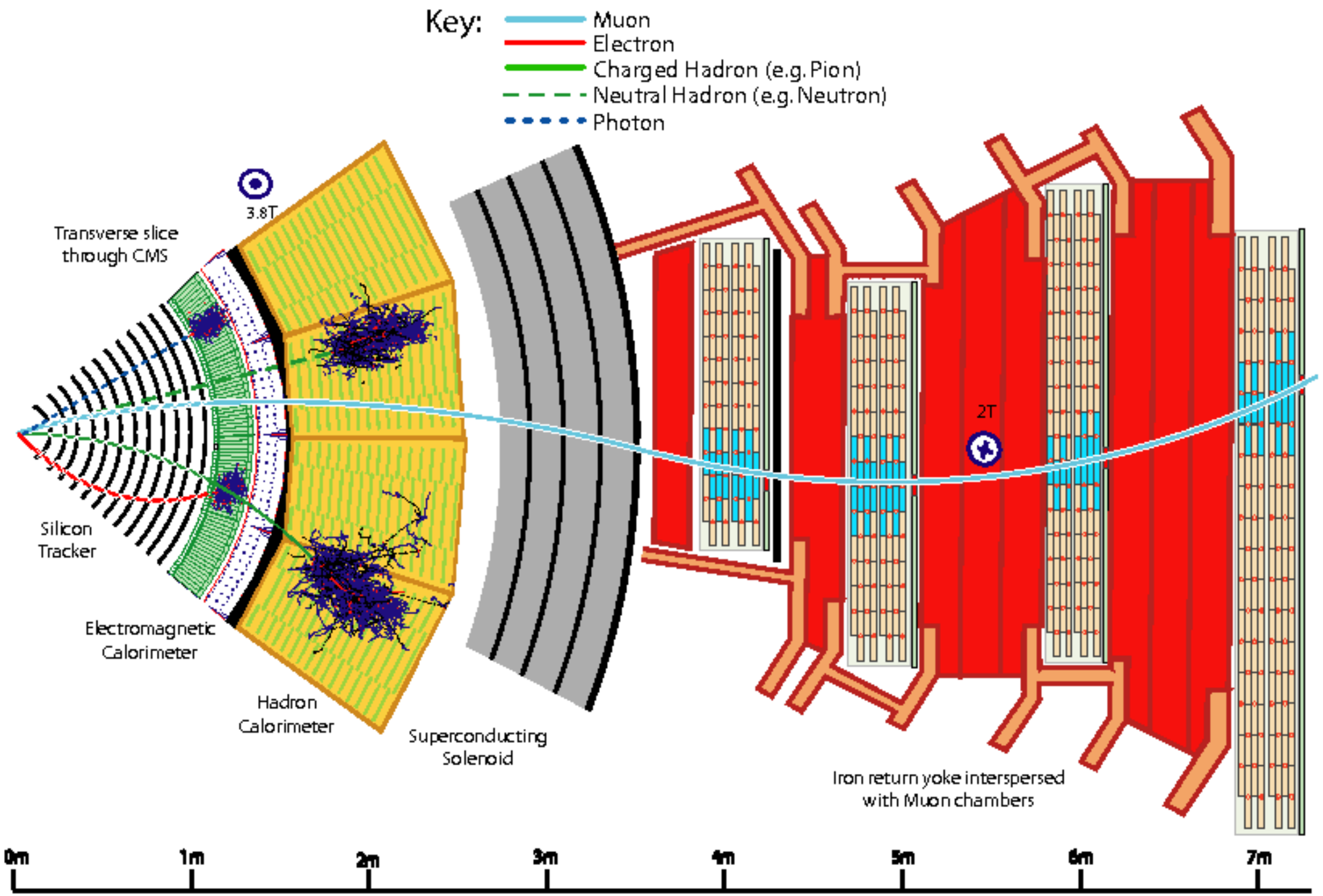


Anatomy of a collider event

CMS example

- Identify collision vertices and particles:
  - Track-finding
  - Electron/jet/muon ID/reconstruction
- Measure energy, momenta, electric charge
- Jet flavour?
- Signal topology?

ML is key in many of these tasks



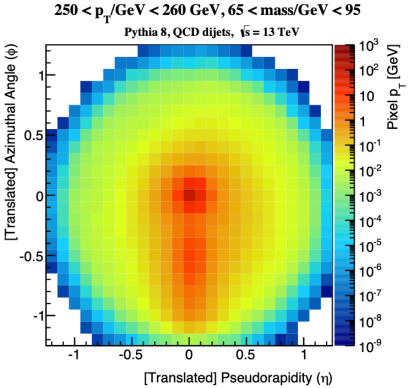
How to represent data?

… part of the definition of the ML algorithm

Tabular

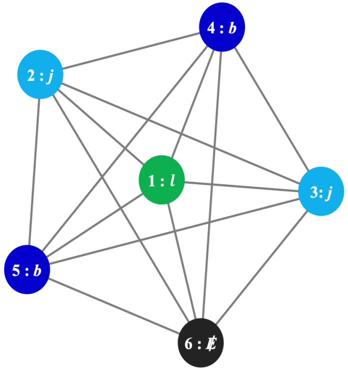
	Electron1_PT	FatJet1_PT	Jet1_PT	Muon1_PT
0	227.793961	253.598358	254.124435	0.000000
1	0.000000	225.937729	228.712021	39.127575
2	68.204712	0.000000	144.771240	0.000000
3	133.825851	229.350952	219.542404	0.000000
4	0.000000	0.000000	127.972099	0.000000
5	82.530861	259.897095	206.621994	0.000000
6	0.000000	0.000000	119.139641	0.000000
7	170.190216	0.000000	199.339508	0.000000
8	0.000000	276.407806	275.428223	219.815781
9	43.247391	240.832916	240.927399	0.000000

Image



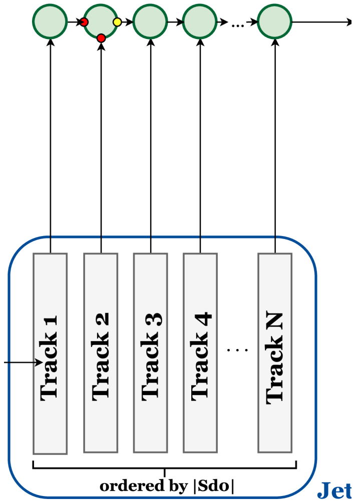
[arXiv:1511.05190]

Graph



[arXiv:1807.09088]

Sequences



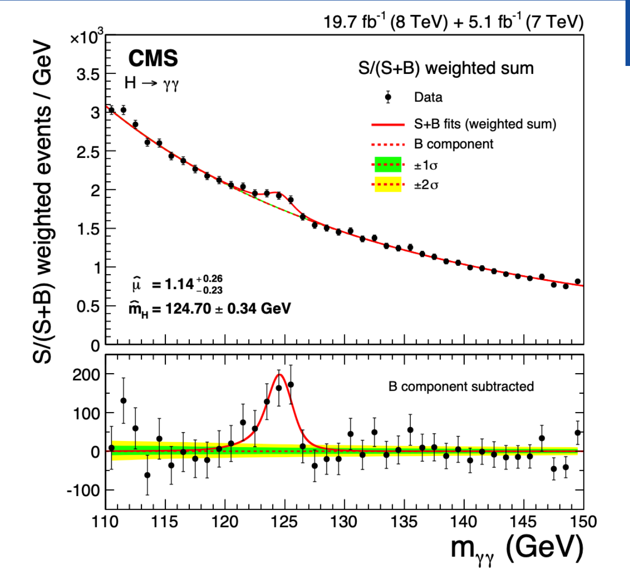
[ATL-PHYS-PUB-2017-003]

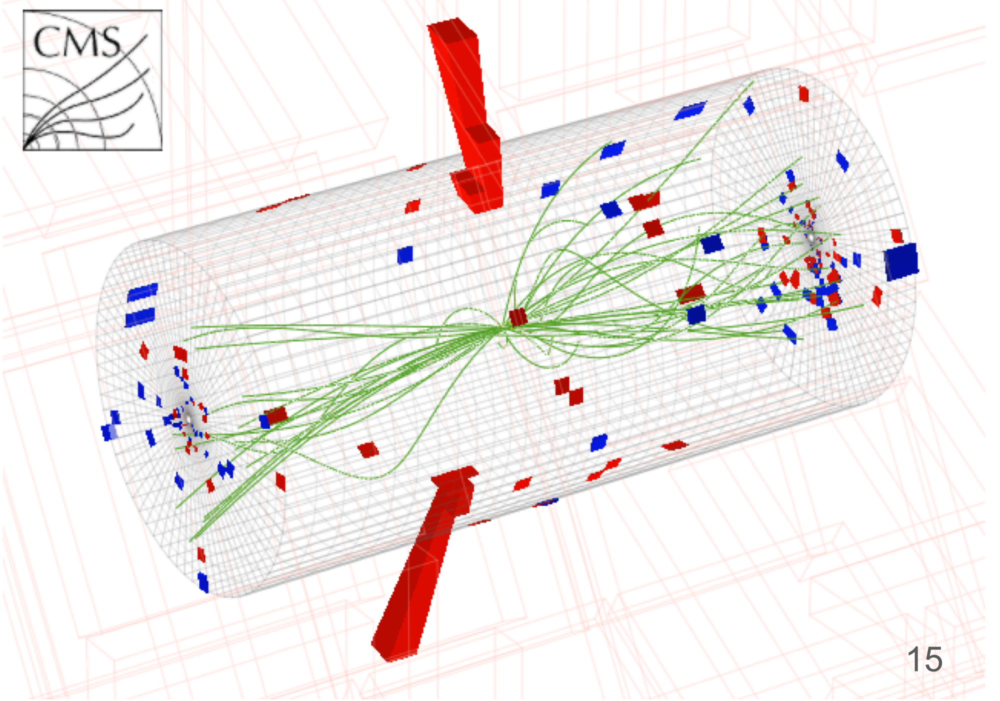
Observation of $H \to \gamma\gamma$ in CMS

1407.0558

Flagship of ML application in the LHC

- 2014: Shallow learning, before Deep learning revolution





Observation of $H \to \gamma\gamma$ in CMS

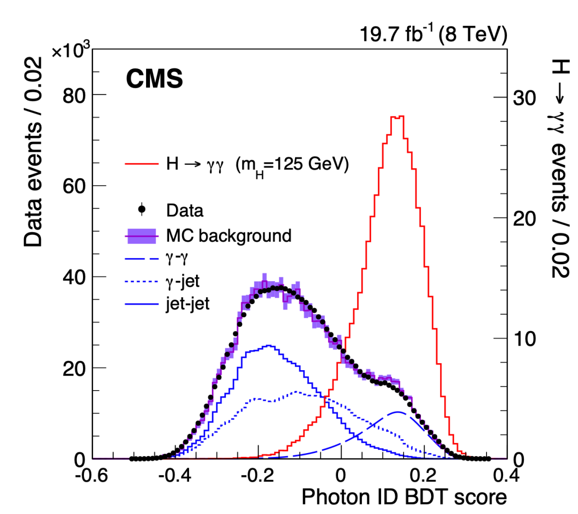
1407.0558

Boosted Decision Trees used in many aspects of the analysis

- Selection of collision vertex
- Photon identification
- Photon energy corrected with BDT regression
- Several BDT to extract signal in different categories
- …

Signal observed with 5.2σ significance

ML impact on signal sensitivity equivalent of 50% more data

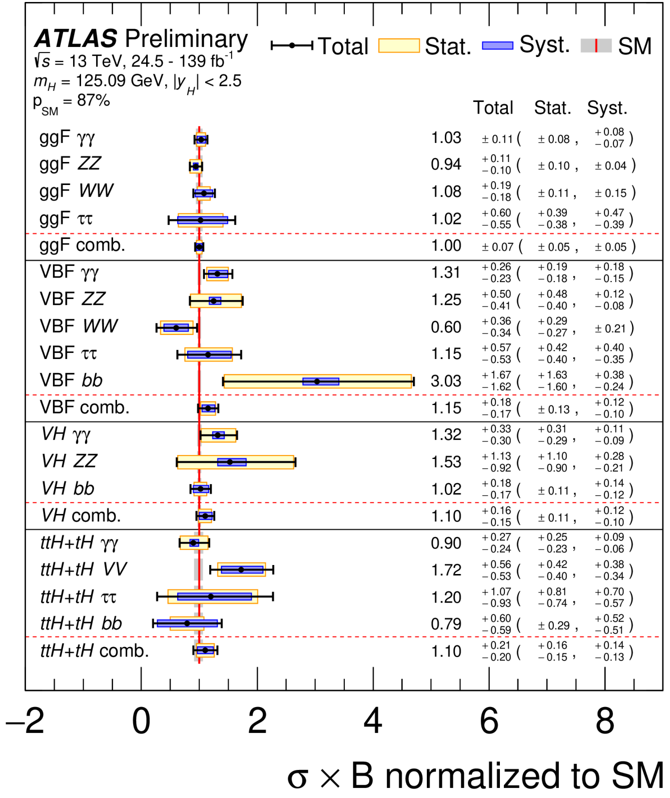


PHOTON IDENTIFICATION

- BDT discriminates photons from fakes (π^0):
  - Shower shape and isolation variables
  - Photon p_T, η

Now... ML still ubiquitous on Higgs Physics

ATLAS-CONF-2020-027



Main Higgs decay modes were observed!

Higgs cross-section measurements:

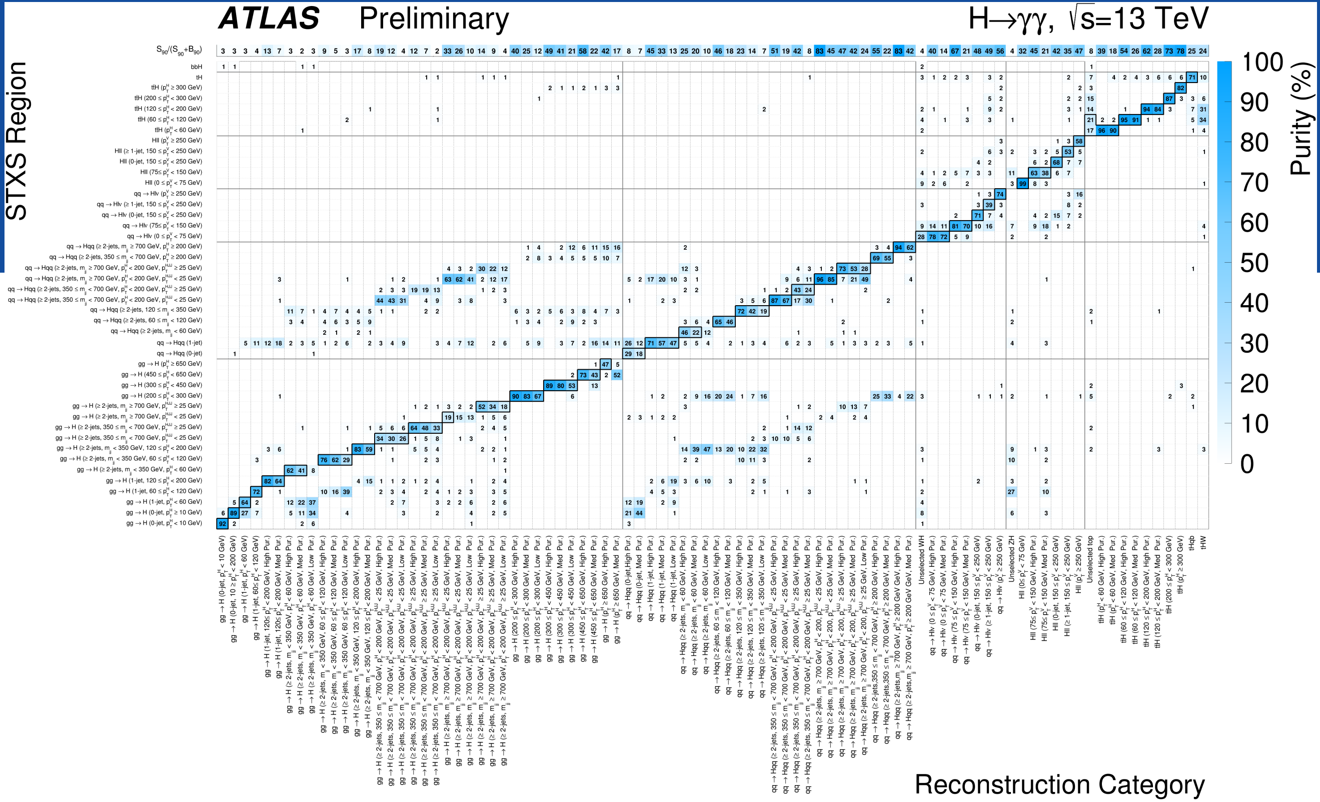
- Many production/decay channels
- Differential cross-section or in bins of the phase space

- $H \to ZZ^* \to 4\ell$: NN defining event categories (signal/bkg-like) (CMS) or as observable for fit (ATLAS)
- $H \to \gamma\gamma$: multi-class BDT to categorise 44 phase space bins (ATLAS/CMS)
- $H \to WW^*$: Deep NN signal classifier used as fit variable in the VBF production channel (ATLAS)
- $H \to \tau\tau$: Convolutional NN that reduces chance of tau mis-ID
- $H \to bb$: BDT for signal identification

See Moriond talk on the CMS/ATLAS Higgs status

Eg: $H \rightarrow \gamma\gamma$ multi-class Boosted Decision Tree

Identify the 44 signal categories



ATLAS-CONF-2020-026

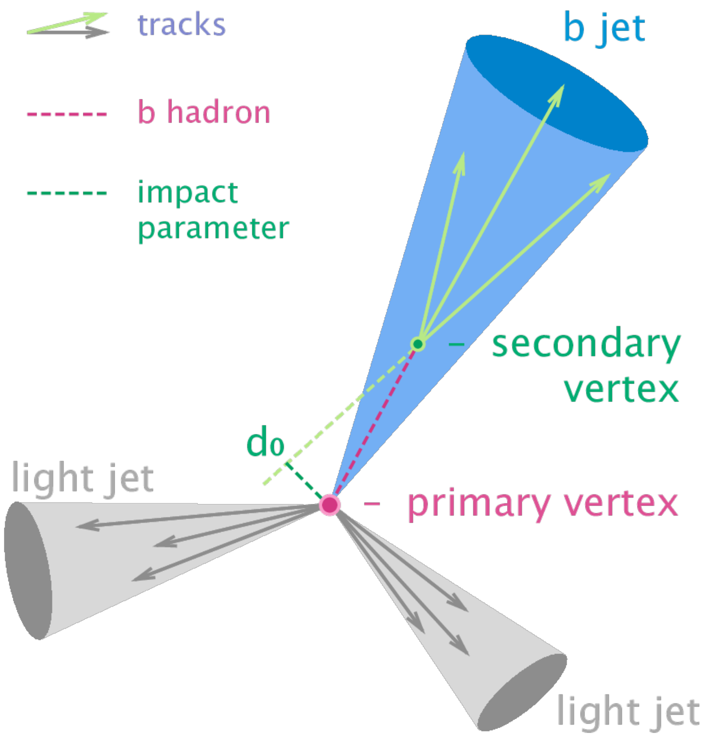
Jet Flavour identification

Essential ingredient for many physics analysis (top, Higgs...)

Per-jet probability of originating from {b, c, uds} quarks

Explore unique characteristics of heavy flavour-jets

- "Large" lifetime of b/c-hadrons (~ps)
- Displaced secondary vertex
- Soft lepton from b/c hadron decay



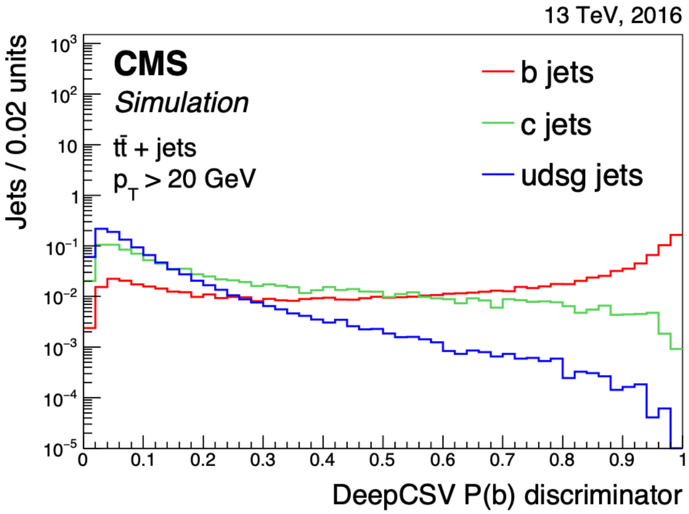
Jet Flavour identification

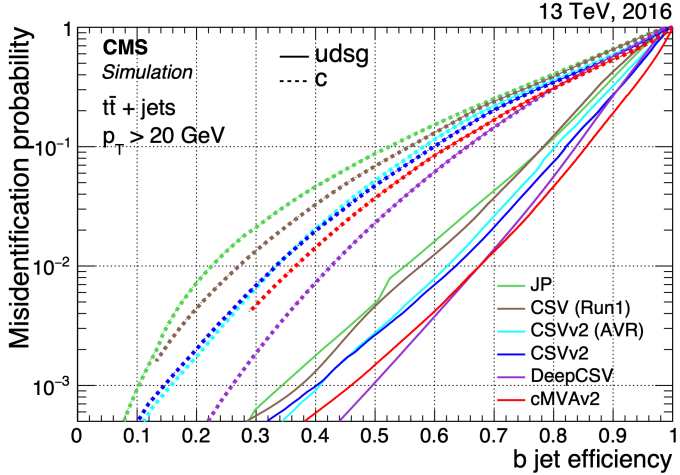
State-of-the-art Deep Learning

New **DeepCSV** (DNN) using same variables of shallow predecessor

- Number of secondary vertices (SV)
- Number of tracks from SV
- SV mass
- Radial distance $\Delta R(\text{track, jet})$
- Jet p_T, η
- ...

Improved efficiency





[1712.07158](https://arxiv.org/abs/1712.07158)

Jet Flavour identification

Deep Sets

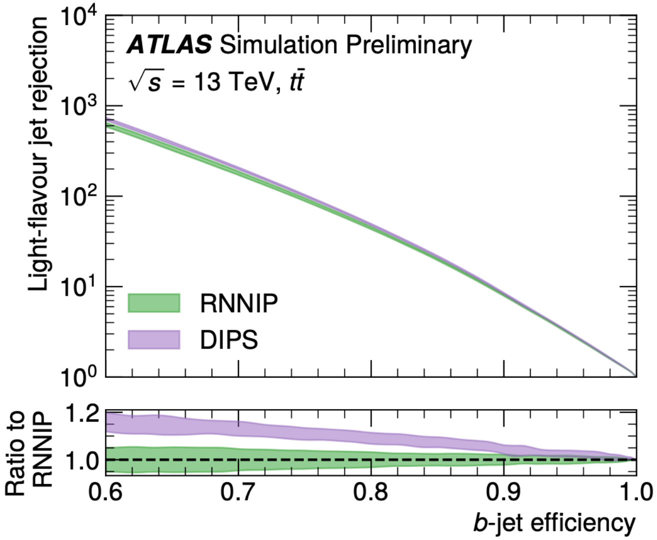
Tagging generally involve a variable number of inputs (tracks)

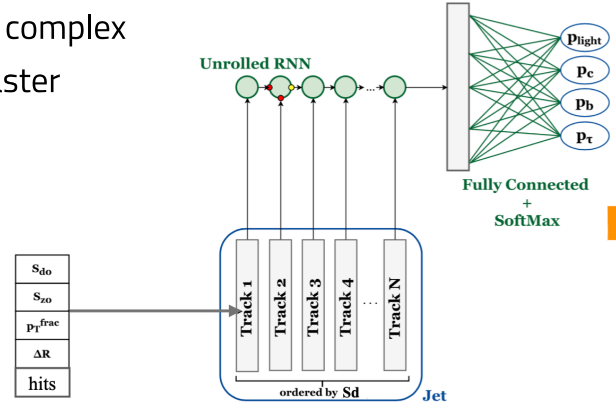
Usually addressed by **Recursive NN**

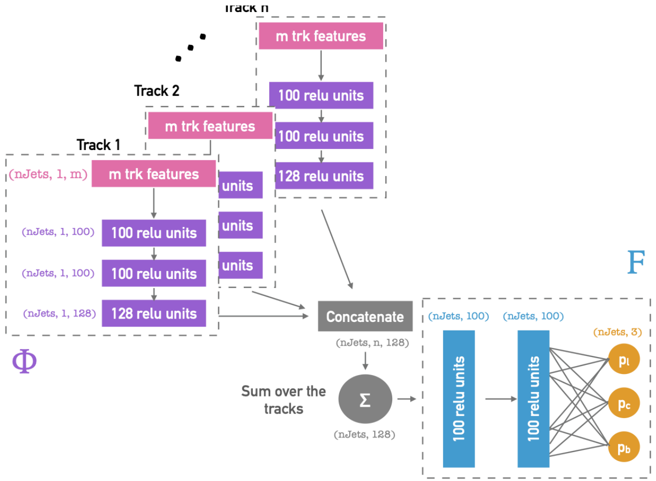
- Natural language processing, order matters (words in sentence)

When order does not matter

- Replace RNN by **DNN + sum**
- Less complex
- 4x faster



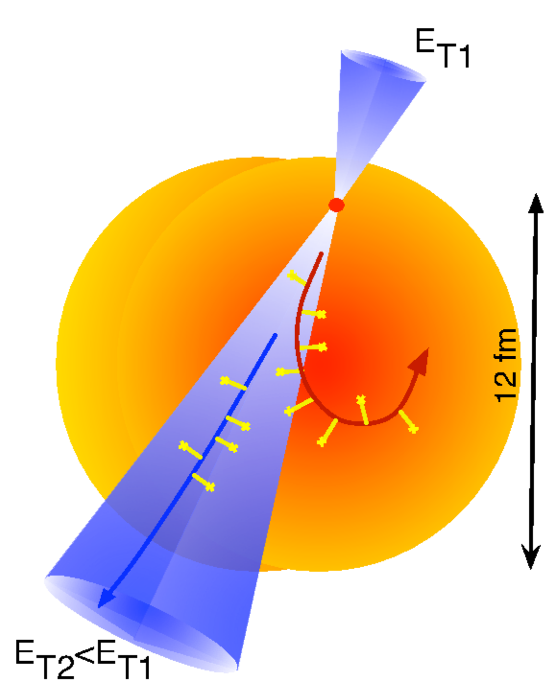




ATL-PHYS-PUB-2020-014

Classification of Quenched Jets

Jet quenching is one of the most important signatures of the quark-gluon plasma (QGP) formed at collisions of relativistic heavy ion collisions at the LHC



- Quenched jets are useful probes to study this particular form of matter
- Classification of quenched jets allow to obtain pure samples of jets which have interacted with the medium
- Useful, f.i., to study the mechanism of jet suppression and the QGP properties

Convolutional NNs to classify Quenched Jets

very soon on arXiv

Classification of **jet images** trained on jets simulated in vacuum versus jets with QGP medium

- The convolutional kernel have the parameters to learn: $\{\vec{w}, \vec{b}\}$
- Scan the image looking for successively detailed discriminant patterns

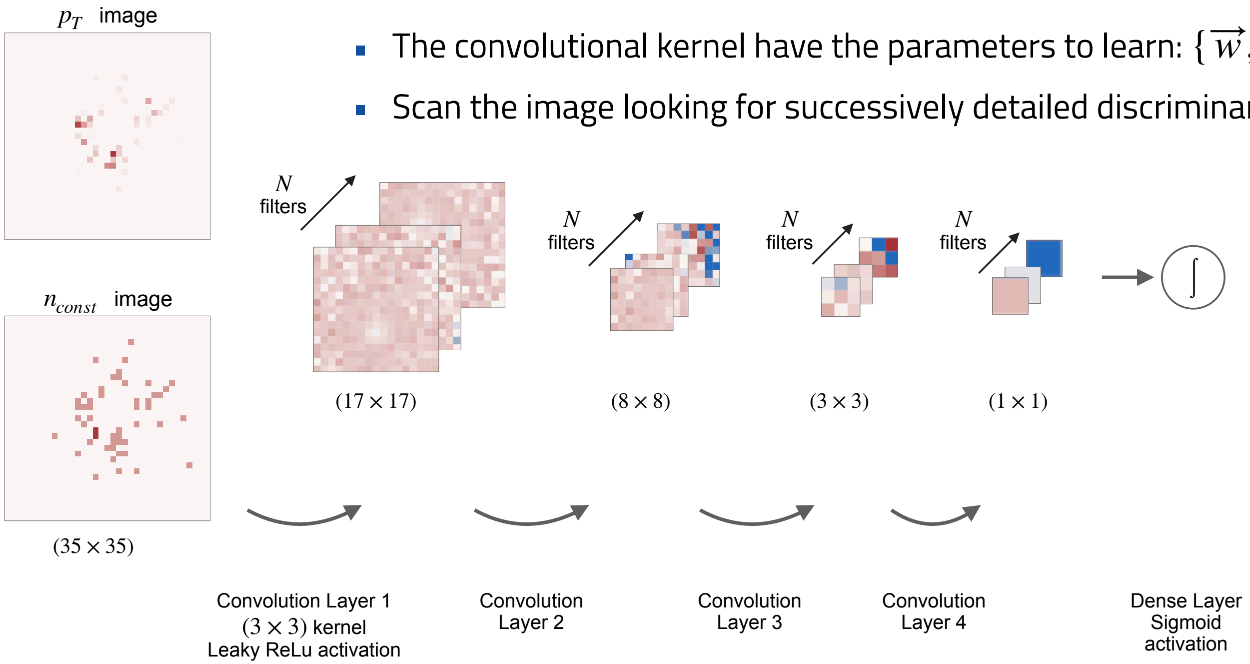


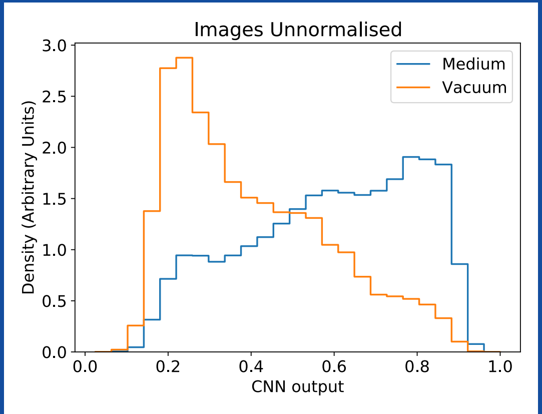
Image pixels (η, ϕ):

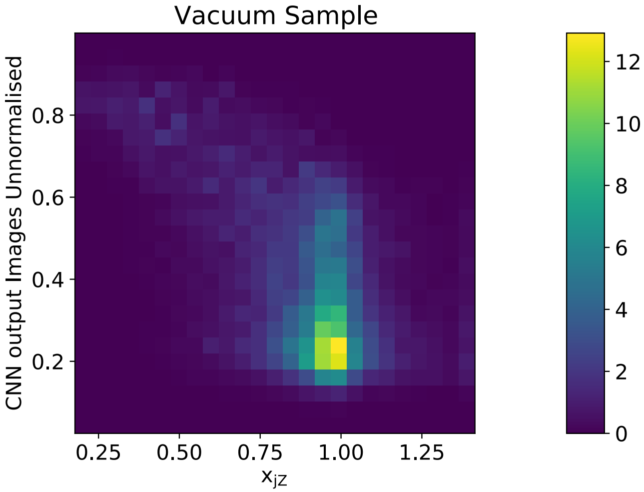
- Jet p_T
- Number of jet constituents

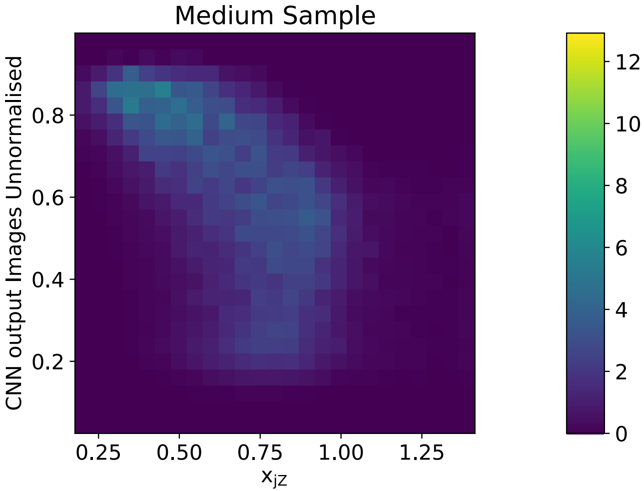
L. Apolinário et al.

CNNs to classify Quenched Jets

- Good separation between vacuum and medium jets
- CNN output correlated with energy loss
- Interesting result since medium sample is not pure in quenched jets





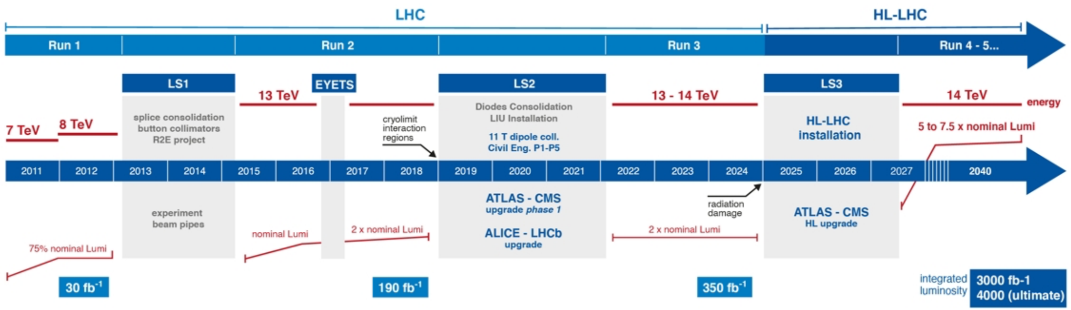


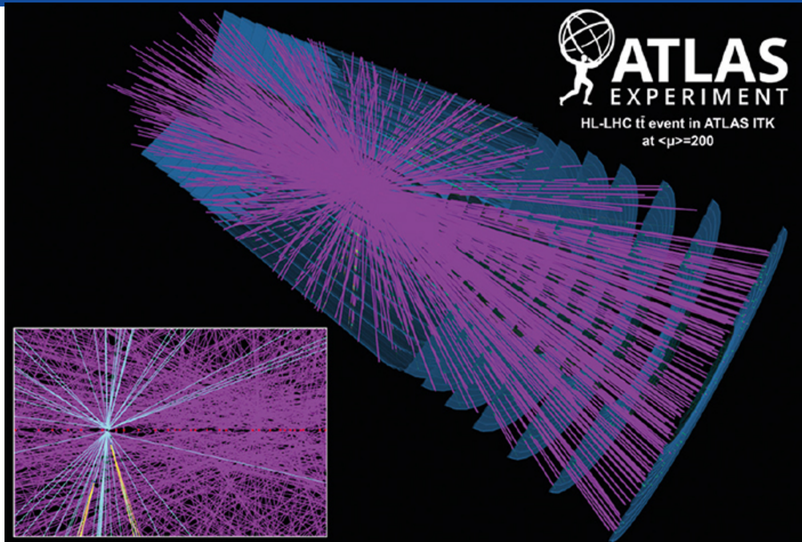
Big Data HEP

ML in the future of collider physics
HL-LHC upgrade

Many challenges and opportunities where ML can be a handle

- High pile-up: collisions per bunch crossing $33 \to 140$
- Noisy environment: ambiguous track hits reconstruction, collision vertex finding, pile-up energy subtraction,…
- Big data phase: 3000 fb^{-1}, increased need for simulation



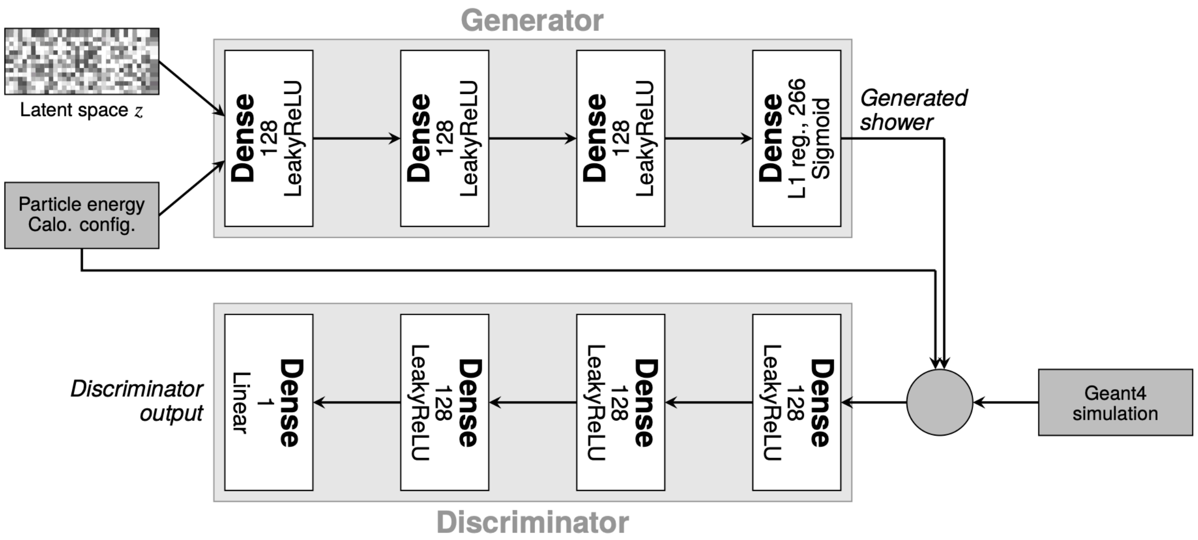


Calorimeter simulation

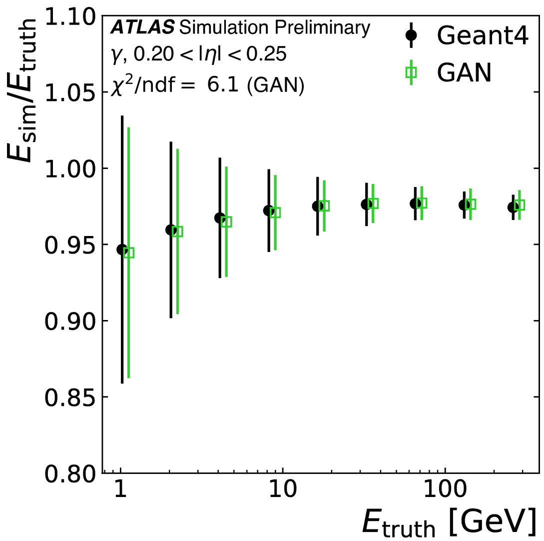
Generative algorithms with Adversarial training

Measurements rely on comparisons between data and simulation (~1000 M for a typical analysis)

- Calorimeter showering is the heaviest load (particle multiplicity and overlap)
- Generate synthetic showers given a particle and the calorimeter geometry
- Train the generator by comparing synthetic to Geant4 showers



π^0 simulation



ML role in the search for New Physics

Towards generic signal detection

A primary LHC goal remains to conquer: no sign of New Physics so far!...

ML used in direct searches, classifiers trained to recognise specific signals

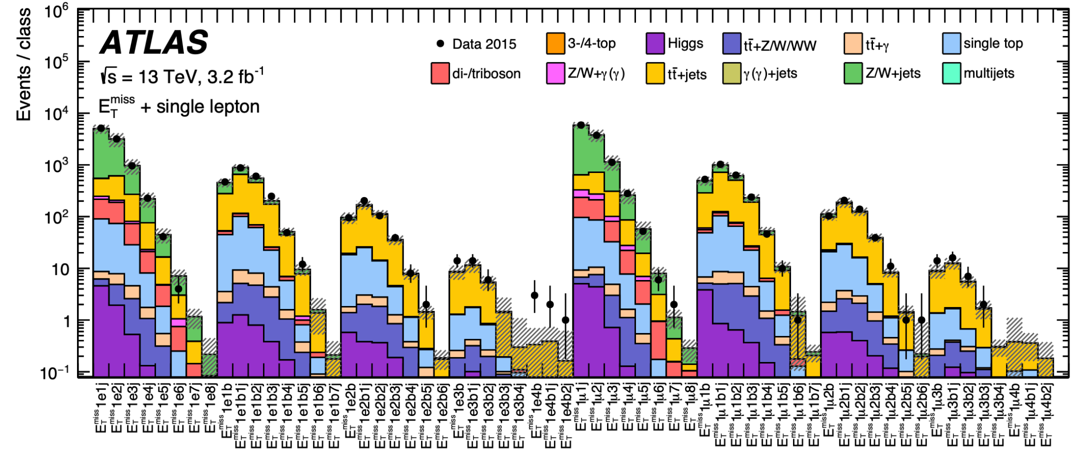
Can ML contribute to increase the generality of NP searches, extending their reach?

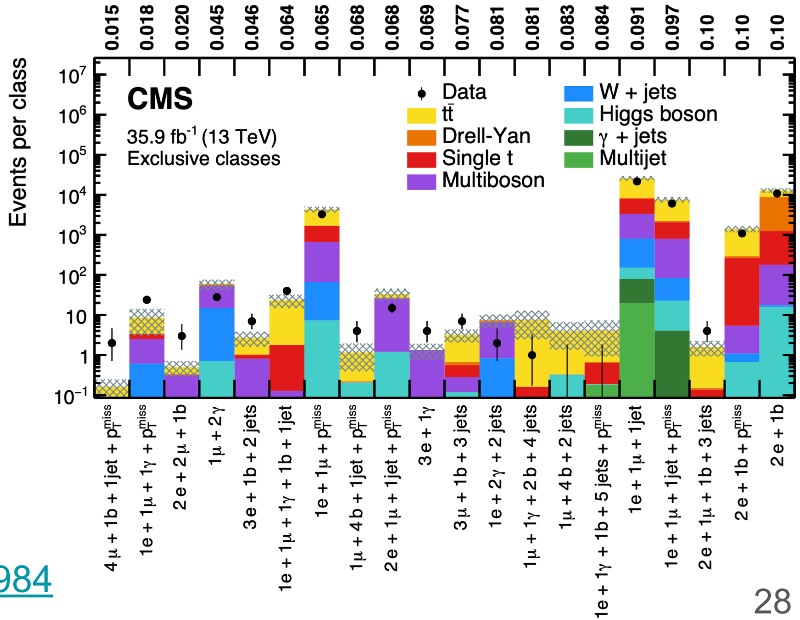
Generic searches for New Physics

Non-ML

Categorise events by particle type/multiplicity and search for disagreement with SM

- Low sensitivity to small deviations of the Standard Model (anomalous couplings)
- Can't help us at trigger level…





2005.02983

2010.02984

Anomaly Detection as a New Physics search

- Anomaly detection: many techniques available…
- What is more suited to HEP collider searches?

Many dreams…

- Generic searches, fully independent of BSM physics hypothesis
  - Capable of analysing full event and different event topologies at once
  - Detect resonances but also small deviations from SM physics
- Trigger-level application
  - Utmost importance: ensure that all BSM events are recorded…

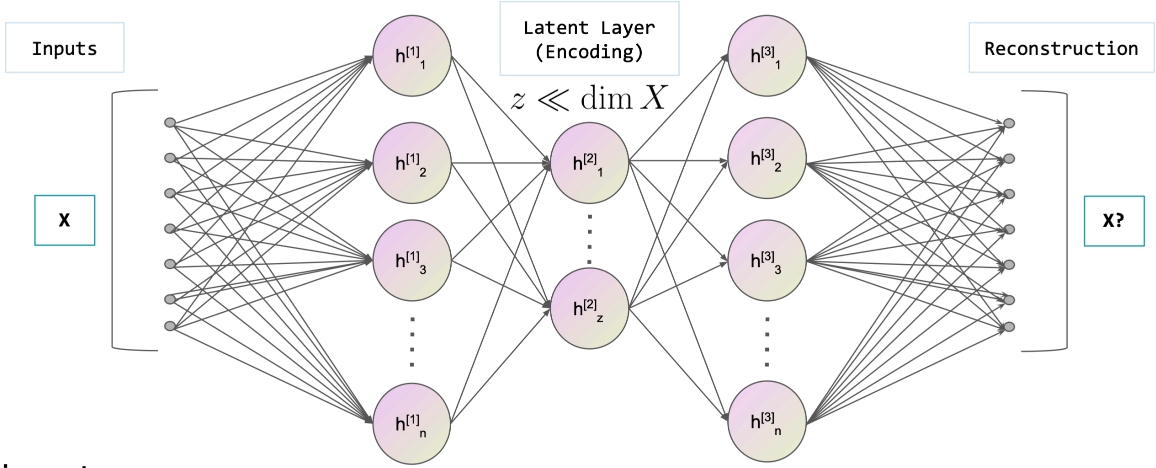
"Finding New Physics without learning about it: Anomaly Detection as a tool for Searches at Colliders"

M. C. Romão, N. F. Castro, R. Pedro

Eur.Phys.J.C 81 (2021) 1, 27

- Physics case study: tZ+X final states, dilepton channel
- How does anomaly detection (AD) perform w.r.t. fully-supervised DNNs?
- Survey of four AD techniques:
  - Auto-Encoder
  - Deep SVDD
  - Isolation Forest
  - Histogram-Based

Auto-Encoder



- Training objective is to minimize input reconstruction loss
- More common events will be better reconstructed
- Reconstruction error is a measurement of anomaly/ *outlyingness*

$\mathbf{x}_i$ the feature vector of the ith event

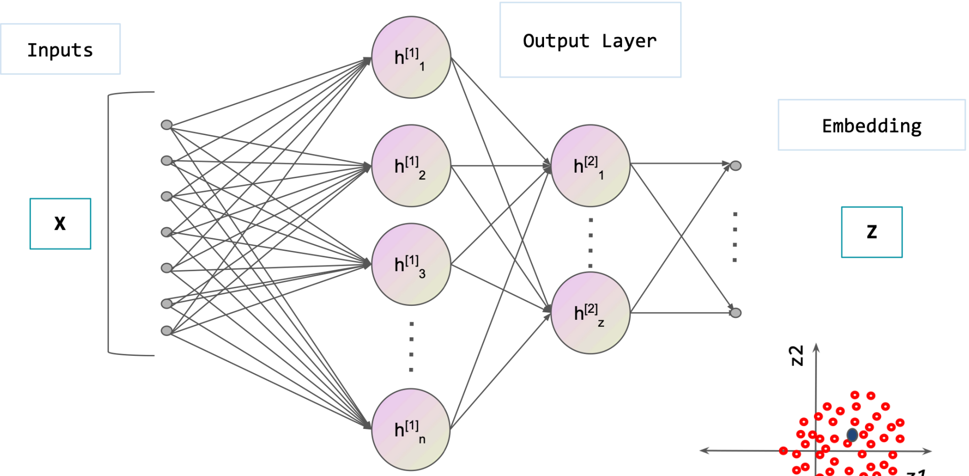
$$\min_{\mathcal{W}} \frac{1}{n} \sum_i ||\mathrm{AE}(\mathbf{x}_i, \mathcal{W}) - \mathbf{x}_i||^2$$

Deep-Support Vector Description

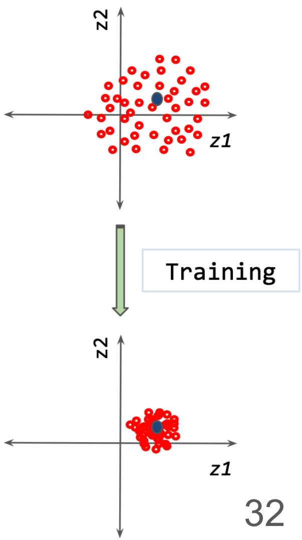
[ref]

- Map the data into an embedding space using a DNN
- Train to minimise the distance of the data points to the center of the distribution in this space

- The rarer events will be further away
- Distance to the center used as the anomaly score



$$\min_{\mathcal{W}} \frac{1}{n} \sum_i ||\mathrm{DNN}(\mathbf{x}_i, \mathcal{W}) - \mathbf{c}||^2$$

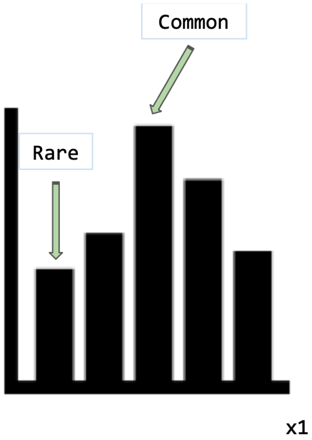


Anomaly Detection methods

Shallow techniques

Both are fast and scalable to high-dimensional data with many instances

$$\sum_j \mathrm{Log2}(\mathrm{Hist}_j)$$

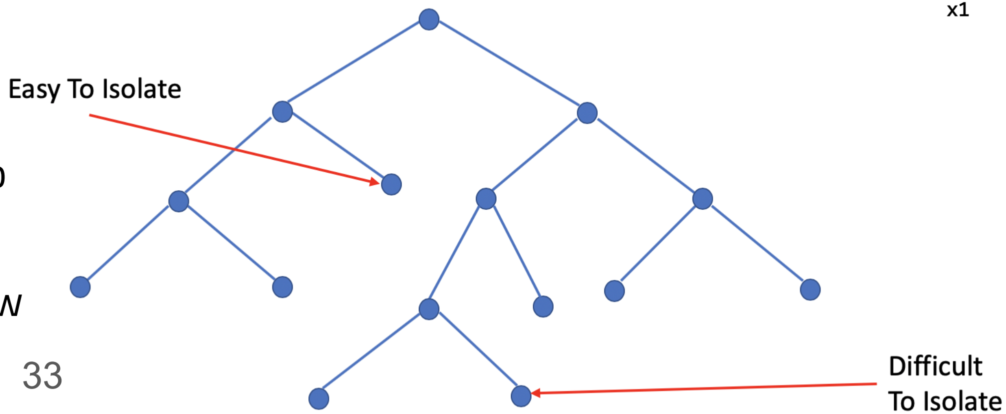


Histogram-based outlier detection (HBOS) [ref]:

- Histogram constructed per input feature *j*
- Anomaly score based on the bin height/density (Hist) where a new instance falls in

Isolation Forest (iForest) [ref]:

- Randomly pick a feature and split value to recursively partition the data
- Anomaly score given by the inverse of how many nodes it took to isolate the event



Benchmark signals and data simulation

Data: MADGRAPH5+Pythia 8+Delphes simulation

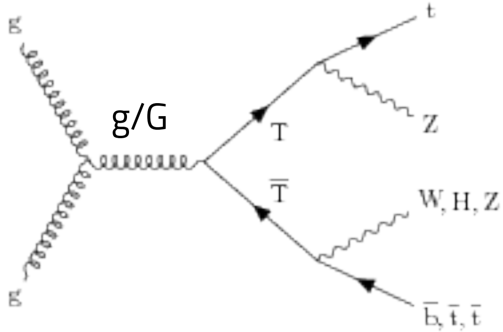
Benchmark BSM signals containing TZ+X final states:

- Vector-like T-quark pairs
  - T-quark mass = {1, 1.2, 1.4} TeV
  - Via SM gluon fusion
  - Via BSM 3 TeV heavy gluon production
- tZ production with FCNC effective vertex

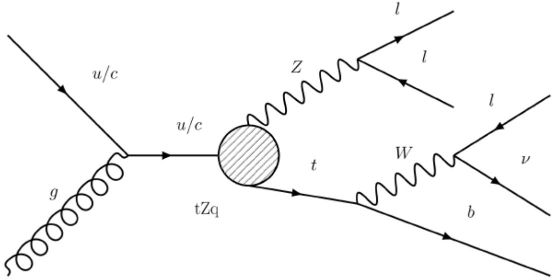
SM dominant processes: Z+jets, top pairs, di-boson

- Total ~13 M events
- Good statistical representation of all phase space
  - Samples generated in slices of pT (or scalar HT)

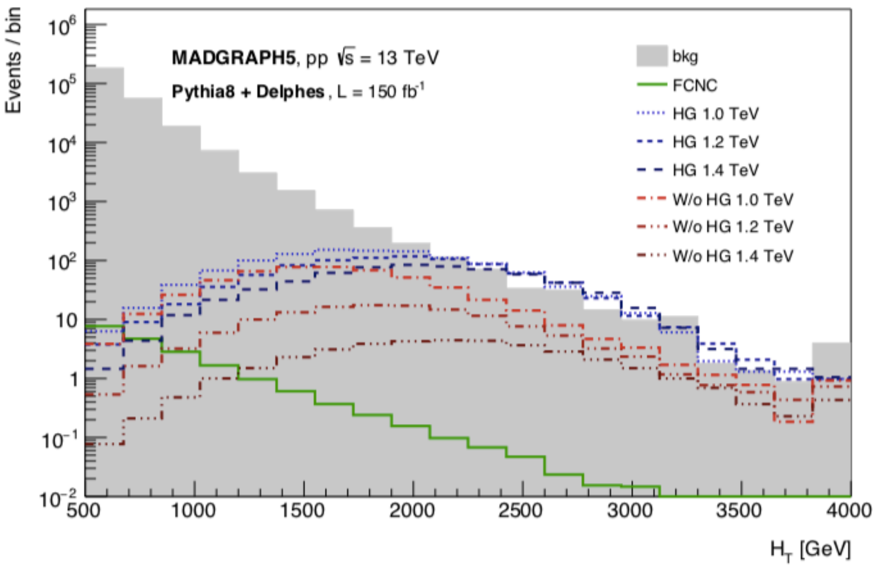
TTbar via SM gluon fusion



tZ via FCNC



Training and input features



Pre-selection

- 2 leptons
- at least 1 b-jet
- HT>500 GeV

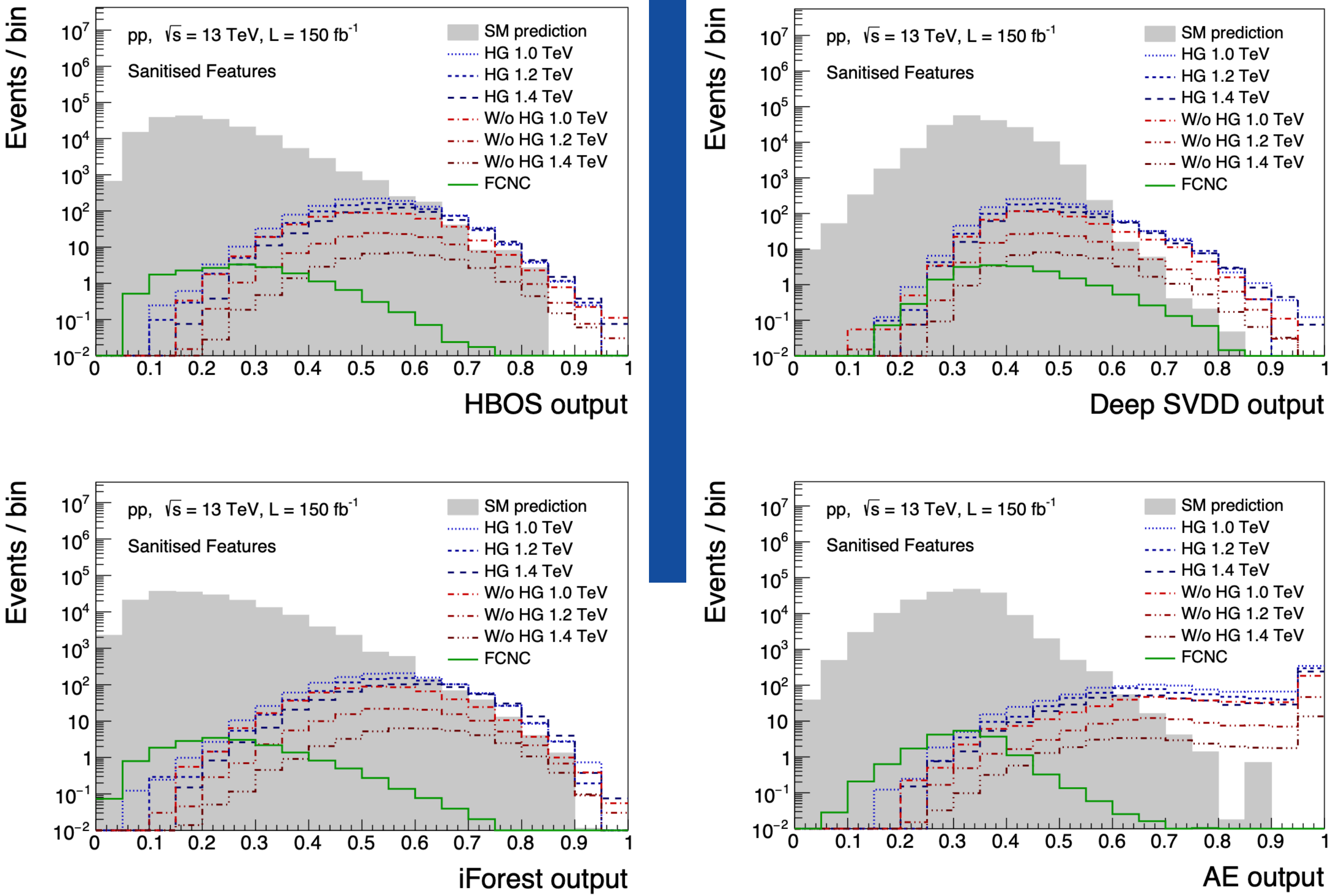
Input features

- (η, ϕ, p_T, m) of the 5 leading jets and large-radius jets;
- (η, ϕ, p_T) of the 2 leading electrons and muons;
- multiplicity of jets, large-radius jets, electrons and muons;
- (E_T, ϕ) of the missing transverse energy.

Training

- Semi-supervised learning
- Train the AD methods on the SM data

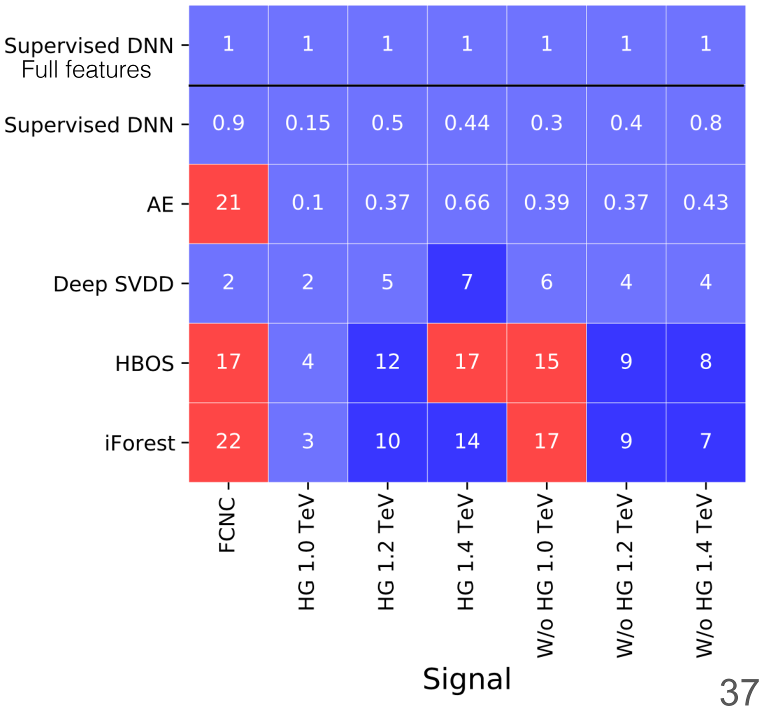
AD score for SM and benchmark signals



Comparison of the AD methods for benchmark signals

- We fit the AD output distributions to compute the upper limits on the signal strength (μ) of the benchmark signals
  - $\mu = \dfrac{\sigma_{obs}}{\sigma_{theo}}$
- Only statistical uncertainties are considered
- Maximum sensitivity degradation around O(10)
- AE is competitive for VL-tops (heavy resonance)
- Deep SVDD seems to be more suitable to small SM deviations (such as FCNC)

Upper limits on μ normalised to Supervised DNN



Summary

- **ML is a universal tool in collider experiments**, increasing the efficiency of many applications
  - Started well back-ago before Deep Learning revolution
  - Now we use increasingly lower information with deeper and more complex architectures
  - Data representation as images, sets, graphs… to take advantage of the most powerful algorithms
  - Deep Learning is also a key to address future challenges (simulation, tracking…)

- **Anomaly Detection is an imminent path for the HL-LHC big data phase**, very active R&D
  - Our conclusions so far:
    - Deep Learning AD models outperform the shallow ones
    - … but the methods have different notions of anomaly
    - Different AD algorithms are suitable to isolate different types of BSM physics
    - Use them in a complementary way?

[THANK YOU]

POCI/01-0145-FEDER-029147
PTDC/FIS-PAR/29147/2017

Anomaly Detection Training

Shallow methods:

- Principal component rotation to remove linear correlation between features

Deep SVDD:

- DNN without bias terms (prevent trivial solutions)

Deep methods:

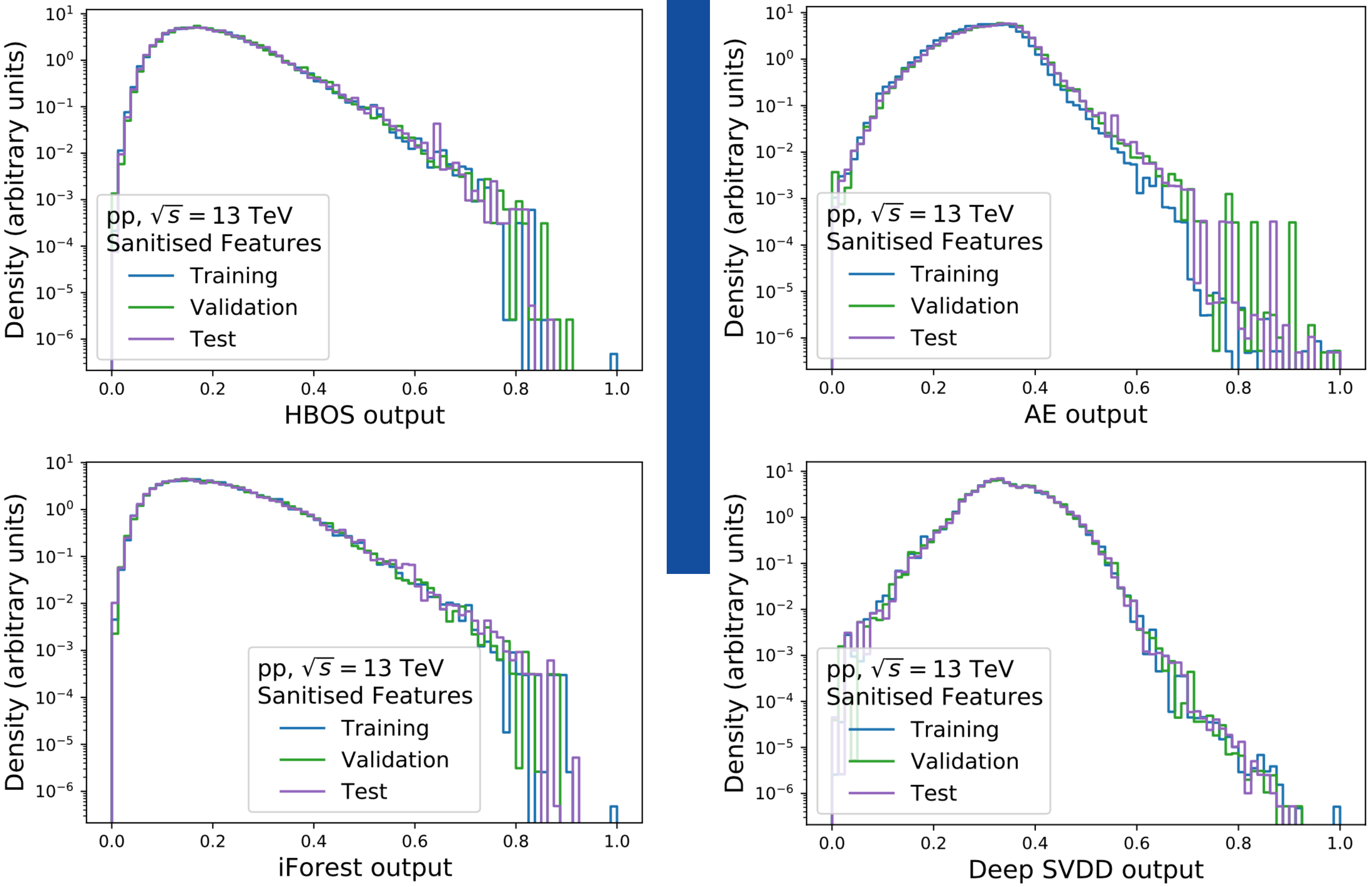
- Latent space dimension fixed to 16
- Activation function LeakyRelu
- Hyper-parameter have Bayesian optimisation based on predefined parameter range

- Semi-supervised learning
- Train the AD methods on the SM data

Hyperparameter	Possible Values
Number of Layers	$[1, 5]$
Number of Units	$[32, 256]$
Initial LR	$[10^{-8}, 10^{-3}]$
Max LR	$[10^{-3}, 10^{-1}]$
Gamma	$[0.95, 0.999]$ in steps of 0.001
Weight Decay	$\{0, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$
Clipnorm	$\{\texttt{None}, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0\}$

Hyperparameter	AE	Deep SVDD
Number of Layers	3	1
Number of Units	93	128
Initial LR	4.487459×10^{-7}	10^{-6}
Max LR	0.063960	0.02
Gamma	0.992	0.995
Weight Decay	0.0	10^{-8}
Clipnorm	100.0	None

AD score



Density (arbitrary units)
pp, $\sqrt{s} = 13$ TeV
Sanitised Features
Training
Validation
Test
HBOS output
AE output
iForest output
Deep SVDD output

Correlation between AD scores

- Shallow methods very correlated
- Most methods are not correlated
- Different notions of outlyingness
- Events in the10% outlier quantile: