Enabling online selection of rare events at LHC with Deep Neural Networks

Daniela Mascione

University of Trento and Fondazione Bruno Kessler

FONDAZIONE BRUNO KESSLER

UNIVERSITÀ DI TRENTO

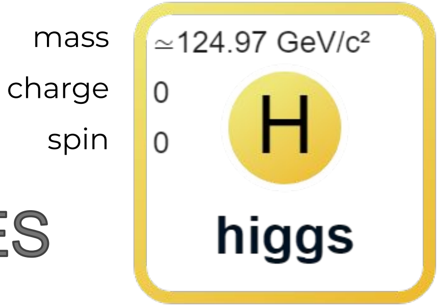
deeppp

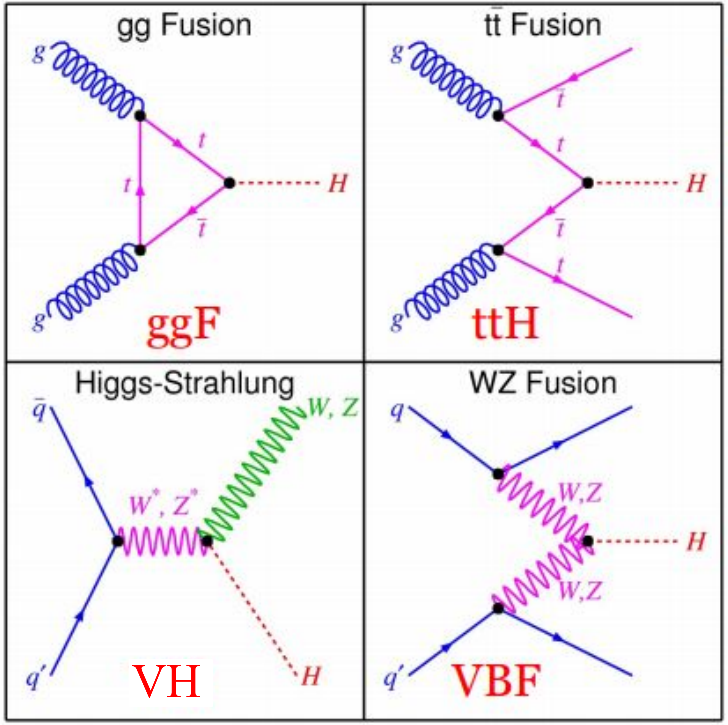
TIFPA Trento Institute for Fundamental Physics and Applications

Overview

- The Higgs boson at the LHC
  - Production/Decay Modes and $H \rightarrow b\bar{b}$ observation
- Deep Neural Networks
  - Functioning and role in observation and selection of interesting events
- Online event selection with Deep Neural Networks
  - Implementation of Deep Neural Networks at trigger level

The Higgs boson at the LHC





source

PRODUCTION MODES

1% ASSOCIATION WITH tt (ttH)

4% ASSOCIATION WITH WITH A WEAK VECTOR BOSON (VH)

7% VECTOR-BOSON FUSION (VBF)

88% GLUON FUSION (ggF)

The Higgs boson at the LHC

mass
charge
spin

DECAY MODES

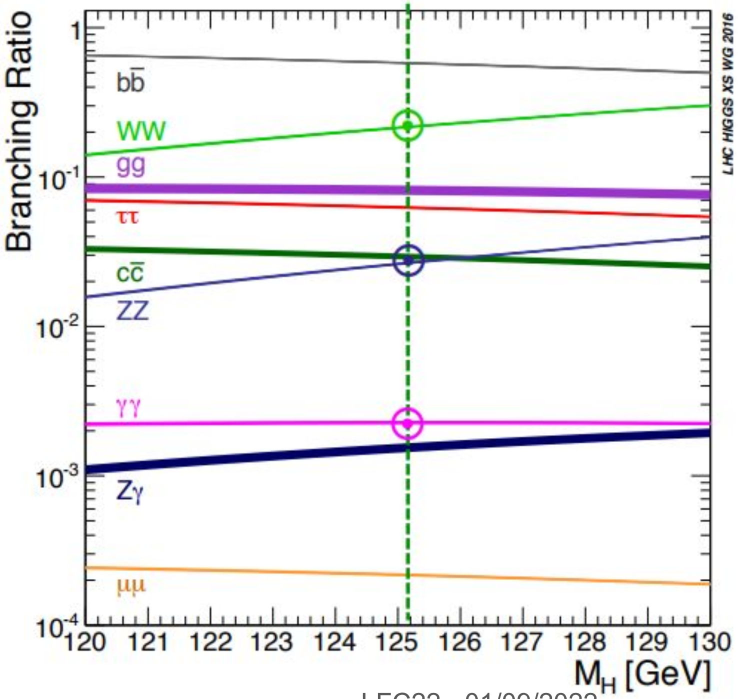
0.2% $\gamma\gamma$

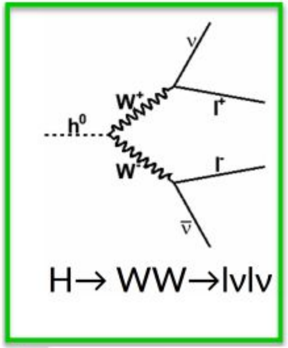
3% ZZ

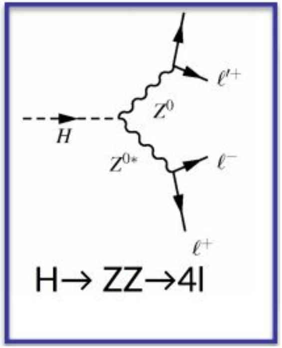
⋮

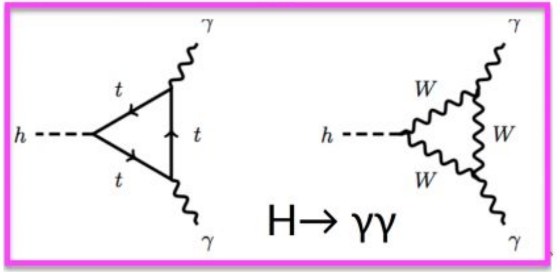
21% WW

57% $b\bar{b}$



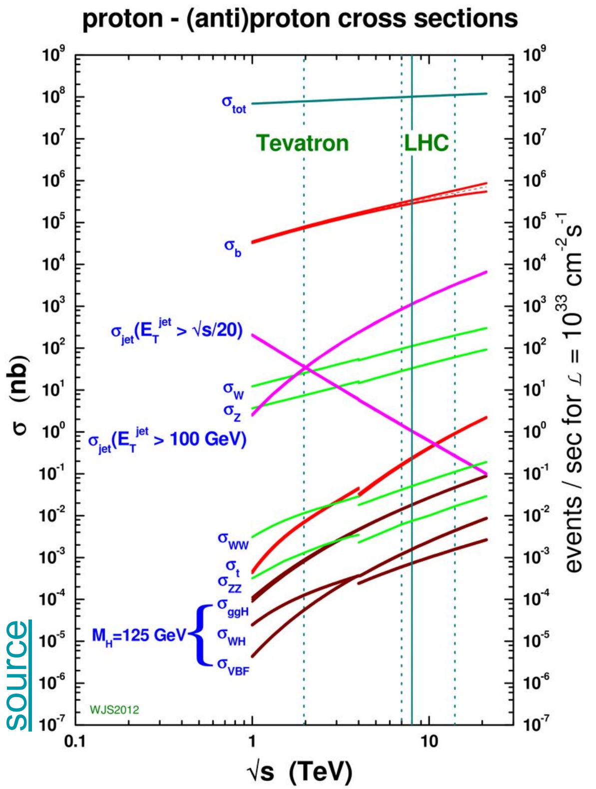






source

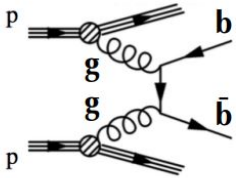
H→$b\bar{b}$ at the LHC



source

	$\gamma\gamma$	$b\bar{b}$
Branching ratio	0.2%	57%
Mass resolution	0.1%	10%

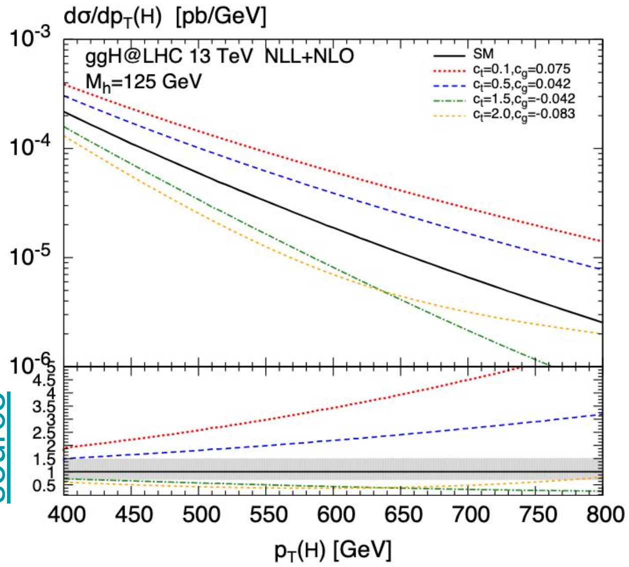
- 🙂 Favored channel to study the Higgs properties
- 🙁 Poor mass resolution
- 🙁 Overwhelming background from QCD production of *b* quarks (10^7 larger)



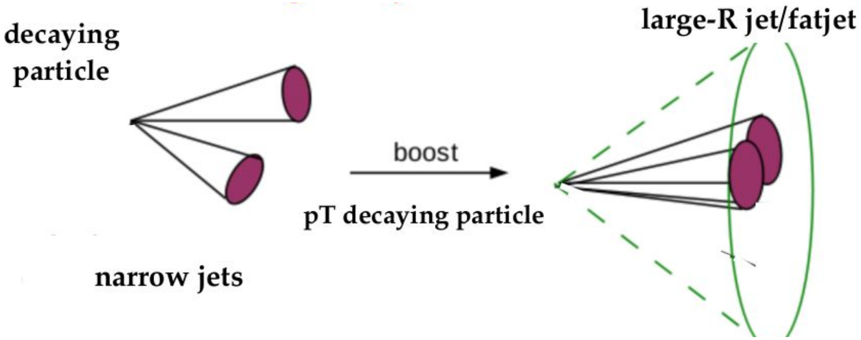
Boosted H→$b\bar{b}$

- Some events produced with a very large p_T
- Production cross-section could be enhanced at high p_T with new physics, as hypothesized by Standard Model Effective Field Theories
  Massimiliano Grazzini et al., *Modeling BSM effects on the Higgs pT spectrum in an EFT approach*, 10.1007/JHEP03(2017)115

source



When the Higgs p_T is very large the angular separation of the two *b* jets gets smaller

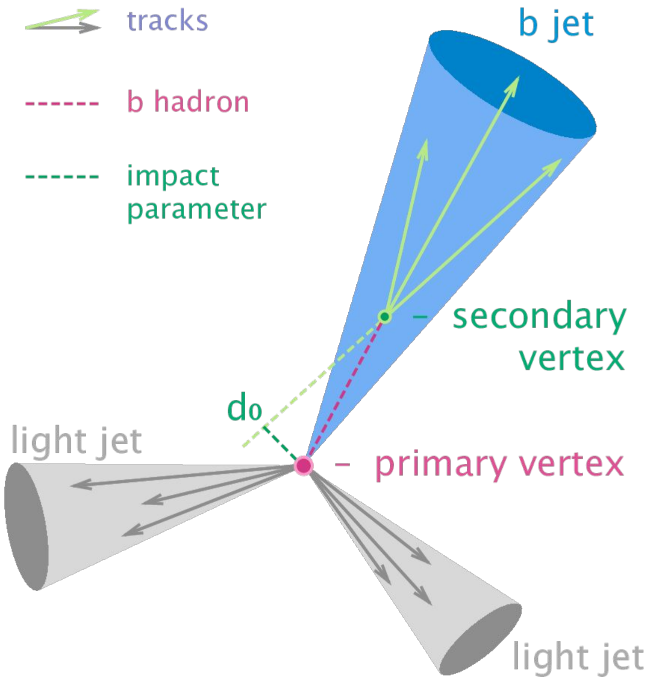


b tagging

Key ingredient to H $\rightarrow b\bar{b}$ searches:
➔ very good *b*-jet identification

Hadrons containing bottom quarks have sufficient lifetime that they travel some distance before decaying.

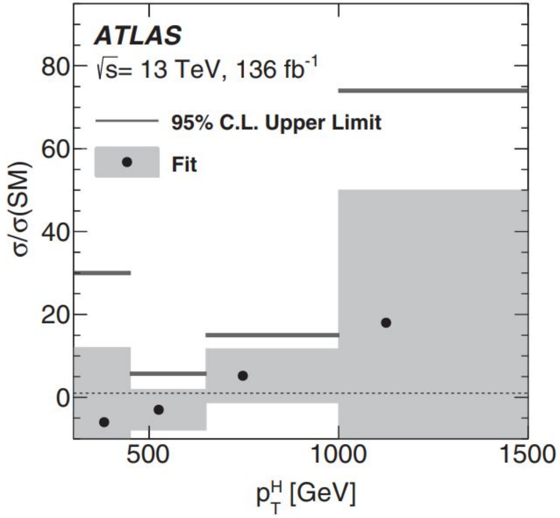
Particles that originate from a place different to where the bottom quark was formed indicate the likely presence of a *b*-jet.



Results of ATLAS and CMS

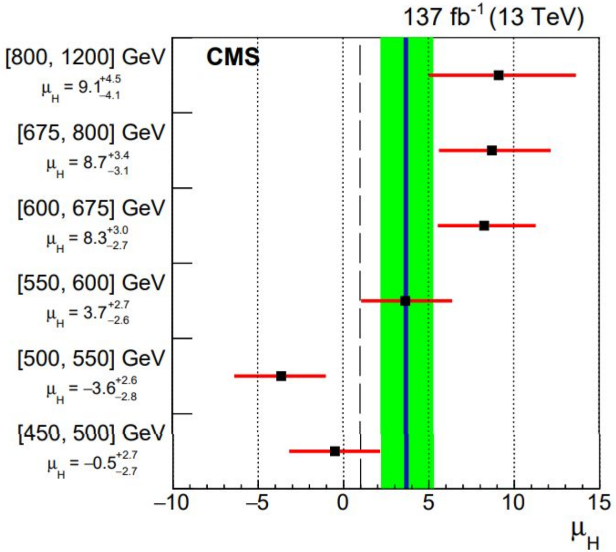
Constraints on Higgs boson production with large transverse momentum using H → b¯ b decays in the ATLAS detector

Inclusive search for highly boosted Higgs bosons decaying to bottom quark-antiquark pairs in proton-proton collisions at √s = 13 TeV



Inclusive signal strength:

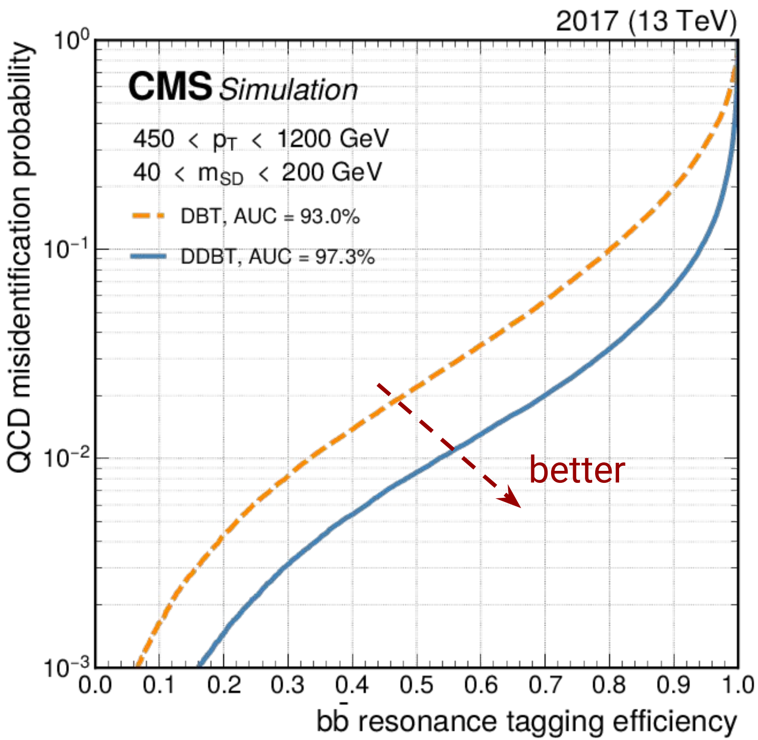
$$\mu_H = 0.8 \pm 3.2$$



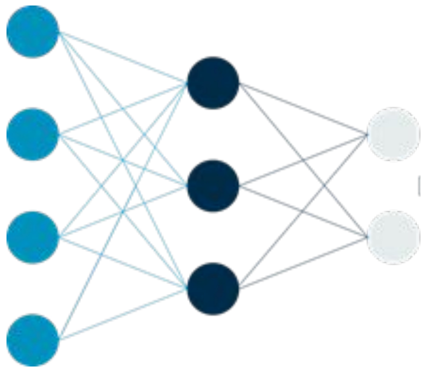
Inclusive signal strength:

$$\mu_H = 3.7^{+1.6}_{-1.5}$$

Results of ATLAS and CMS

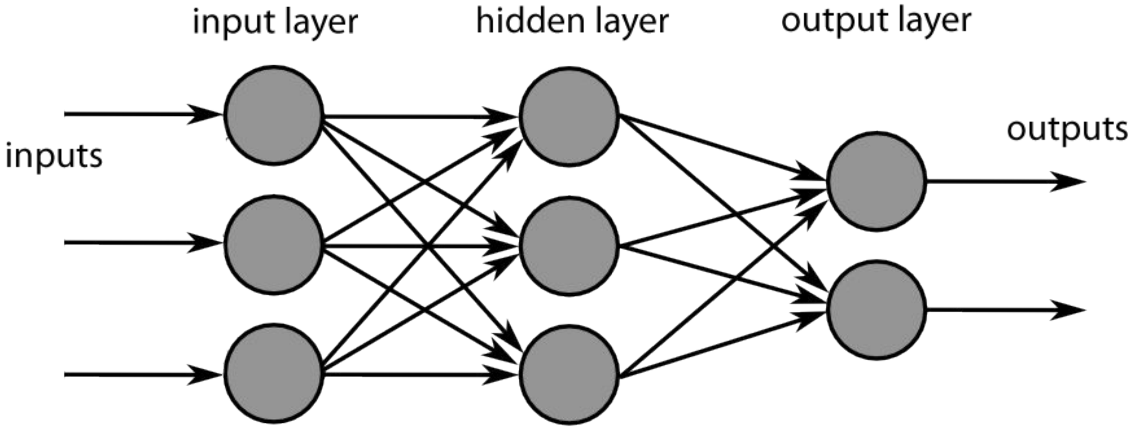


The relative precision of the μ_H measurement in CMS is improved by using a *b* tagging technique based on a Deep Neural Network

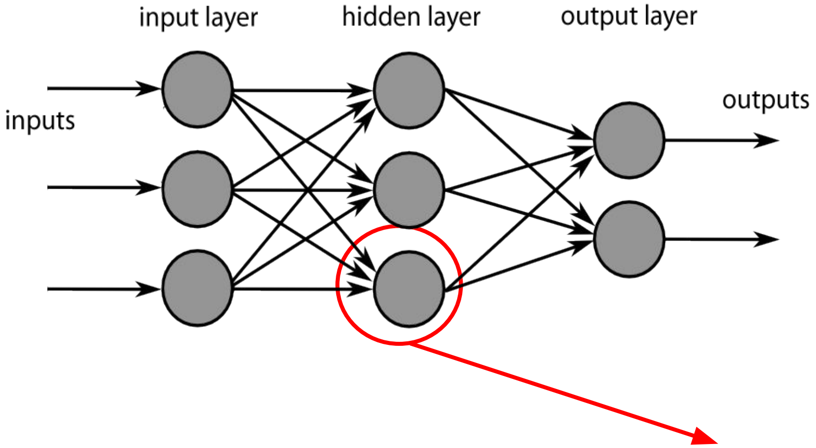


Deep Neural Networks

An Artificial Neural Network is a **computational model** that has layers of interconnected nodes.
A Deep Neural Network has more than one hidden layer.



Through training, the neural network **learns** to recognize a **pattern** in the input data.



Nodes convert weighted inputs to outputs. The **weights keep getting updated** in the process of learning.

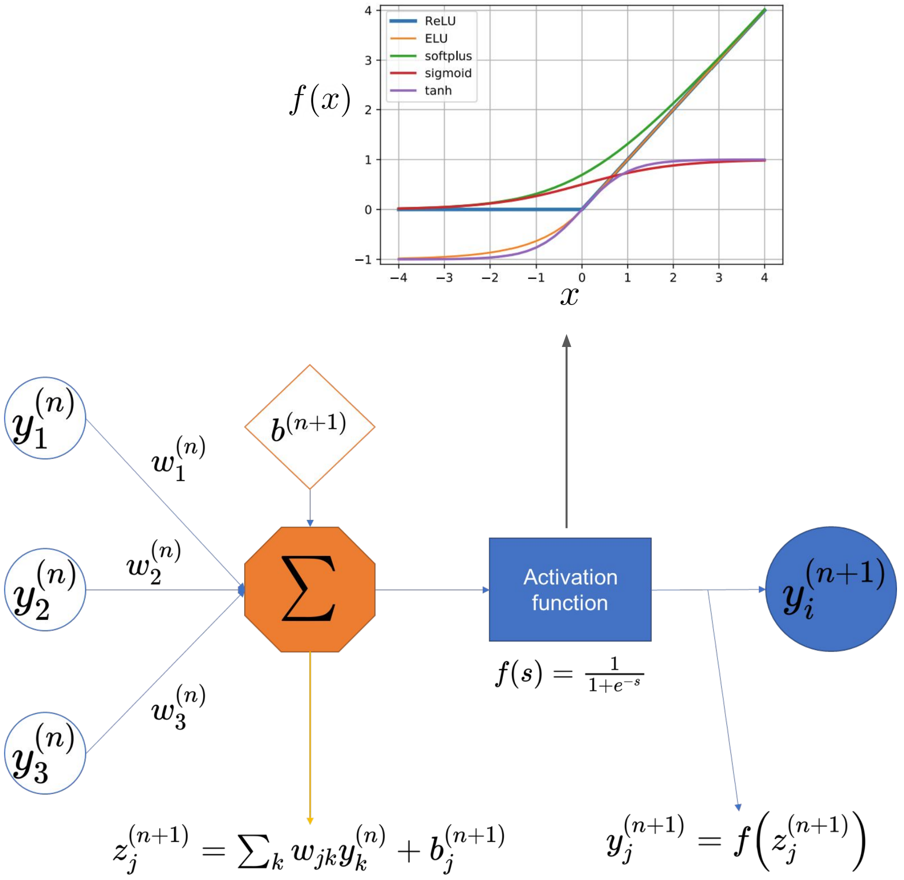
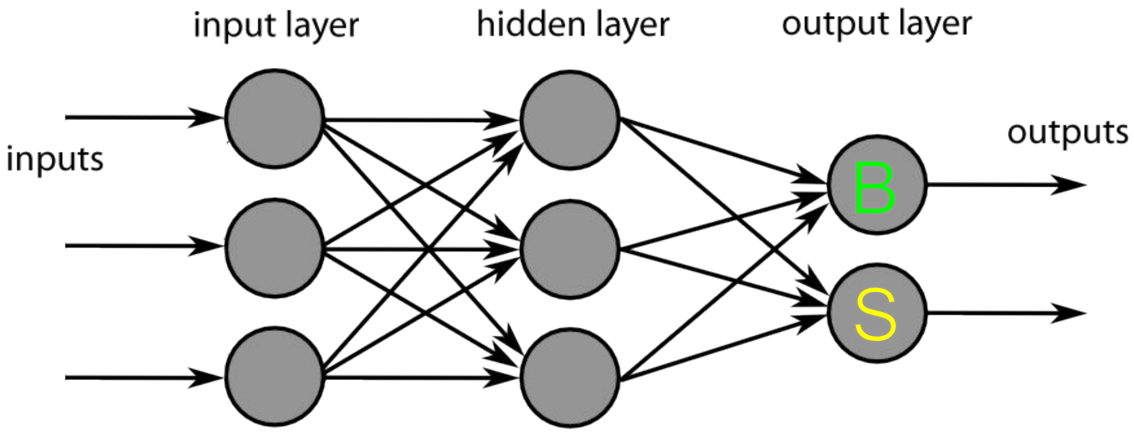


Image: F. Di Clemente

Example

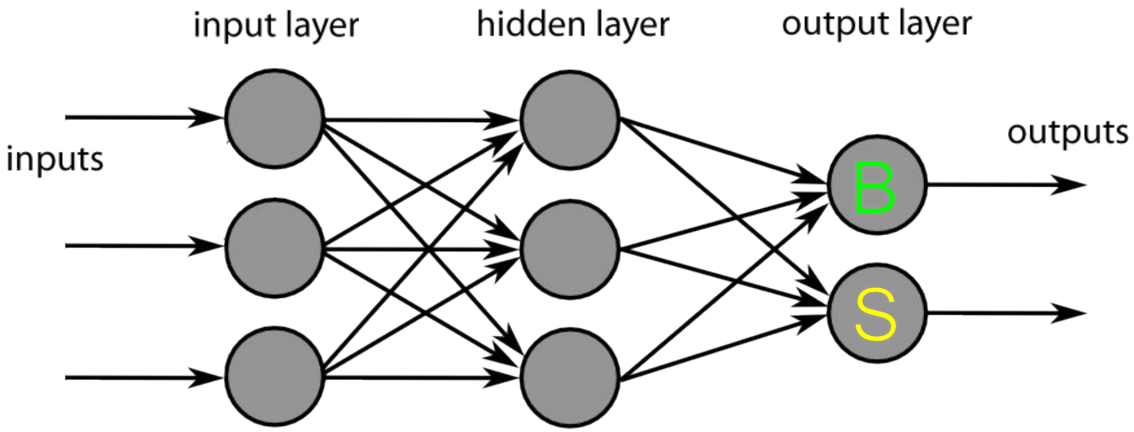
Back-
ground

Signal



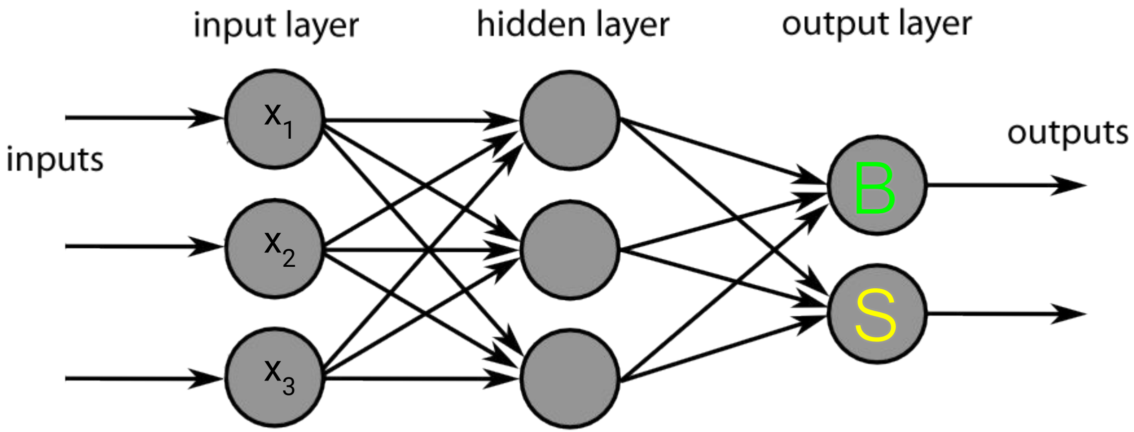
Example

Signal



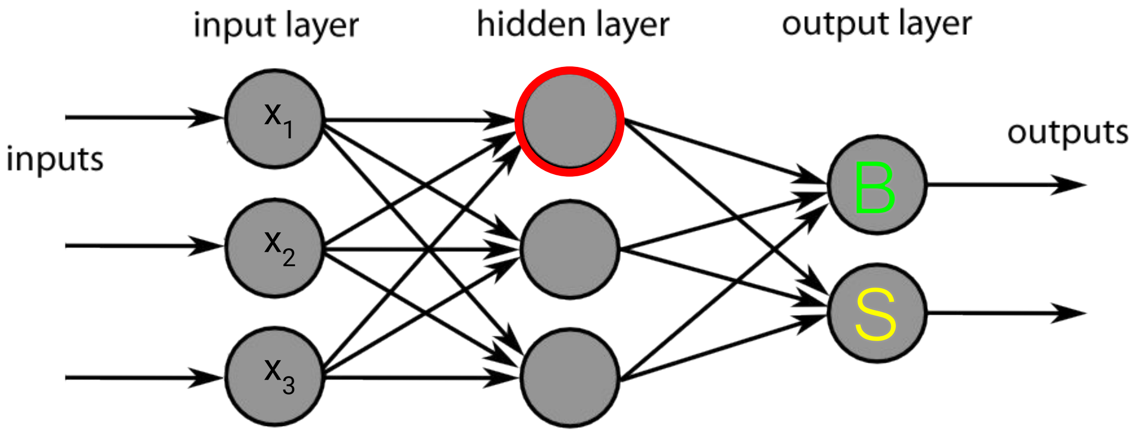
Example

Signal



Example

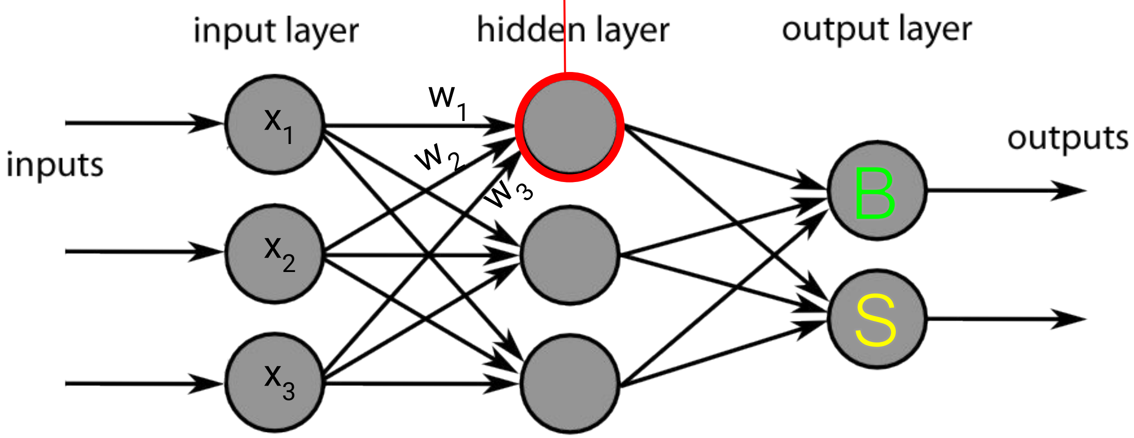
Signal



Example

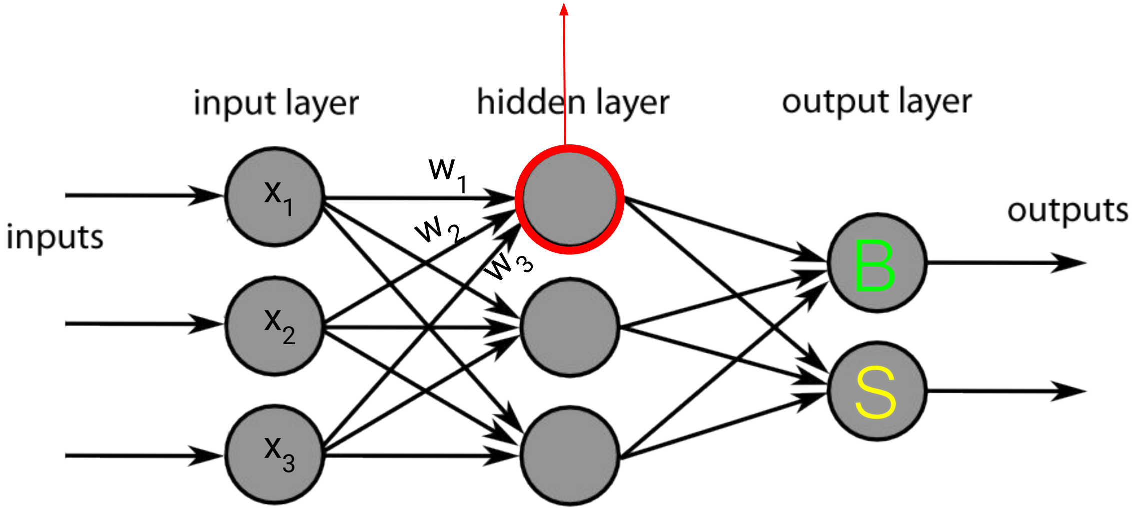
$$(x_1 * w_1 + x_2 * w_2 + x_3 * w_3) + b_1$$

Signal



Example

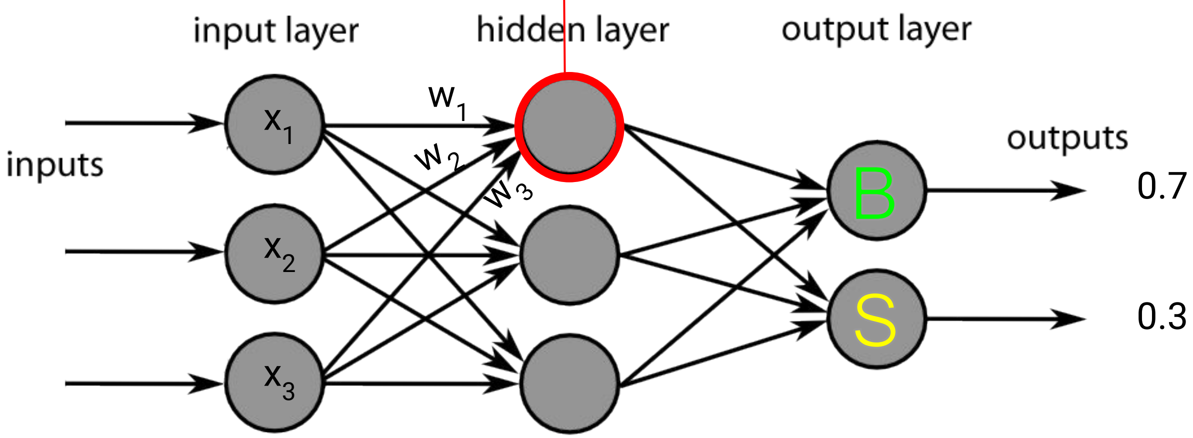
$(x_1 * w_1 + x_2 * w_2 + x_3 * w_3) + b_1$ ⟩ activation function



Signal

Example

$$(x_1 * w_1 + x_2 * w_2 + x_3 * w_3) + b_1 \rightarrow \text{activation function}$$

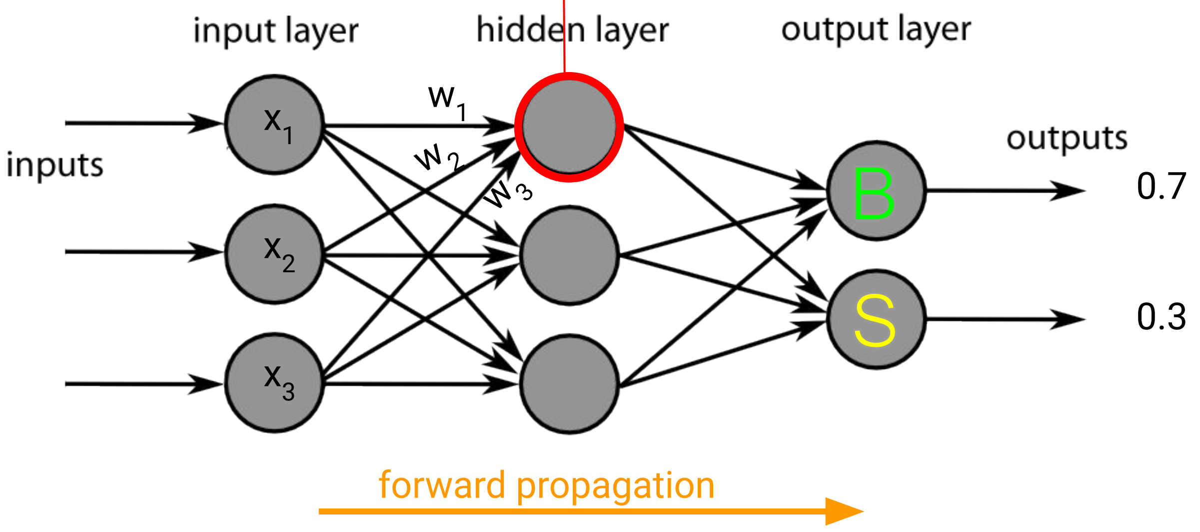


Signal

Example

$(x_1 * w_1 + x_2 * w_2 + x_3 * w_3) + b_1$ ⟹ activation function

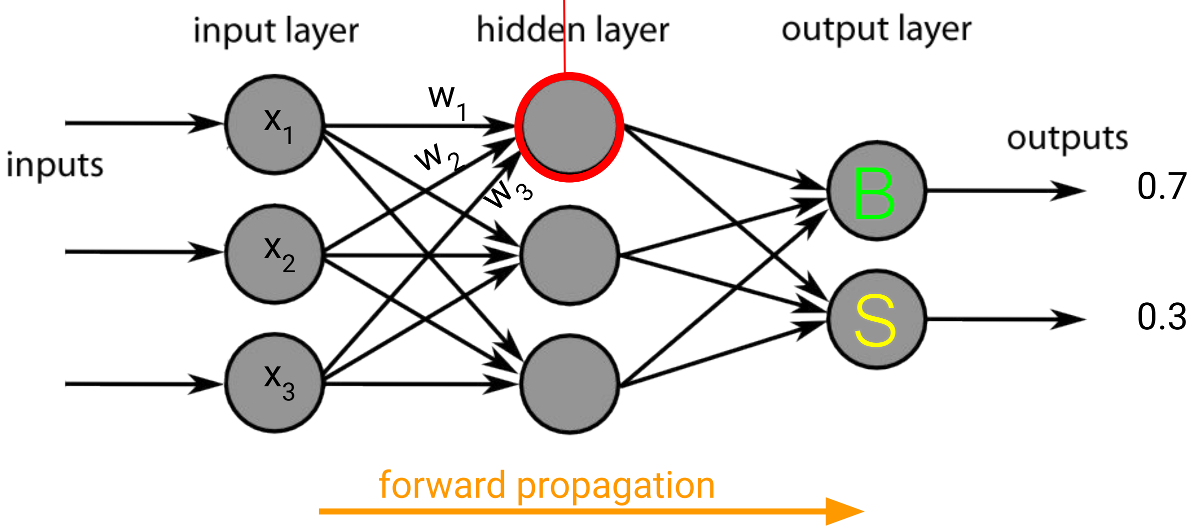
Signal



Example

$$(x_1 * w_1 + x_2 * w_2 + x_3 * w_3) + b_1 \Rightarrow$$ activation function

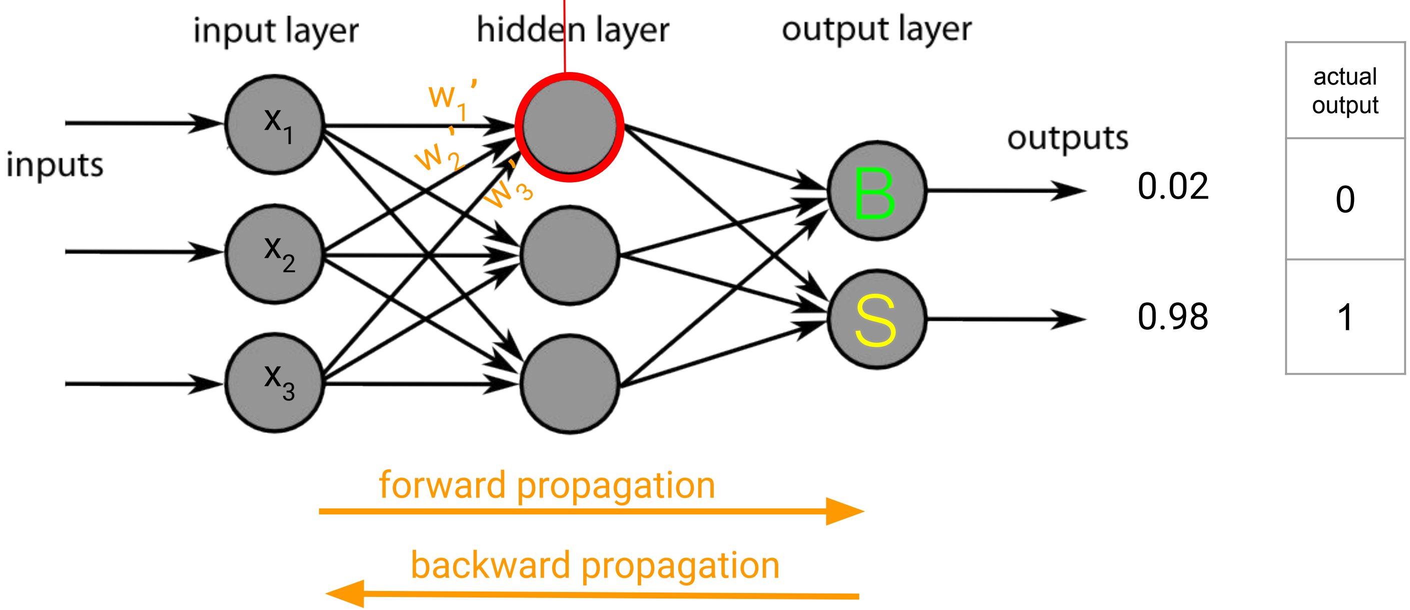
Signal



actual output
0
1

Example

$$(x_1{*}\,w_1' + x_2{*}\,w_2' + x_3{*}\,w_3') + b_1$$ ⟹ activation function

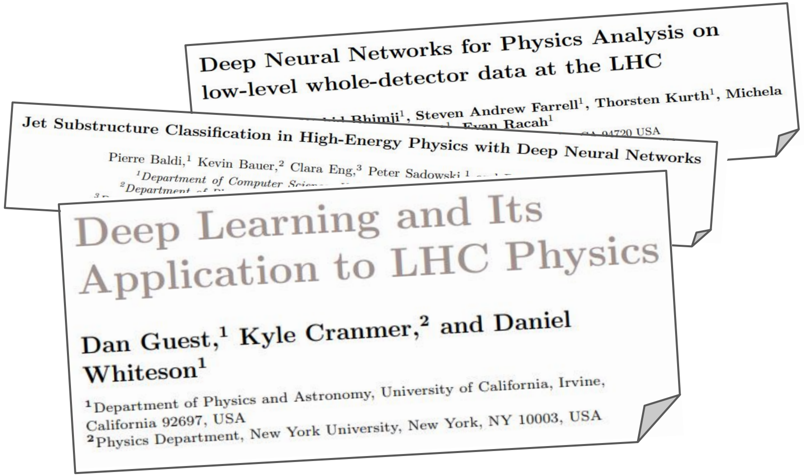


Signal

Deep Neural Networks at the LHC

Deep Neural Networks are widely used at the LHC for a variety of applications that include:

- Event selection
- Tracking
- Jet classification
- Fast simulation


Deep Neural Networks for Physics Analysis on low-level whole-detector data at the LHC

Steven Andrew Farrell[1], Thorsten Kurth[1], Michela

Evan Racah[1]

Jet Substructure Classification in High-Energy Physics with Deep Neural Networks

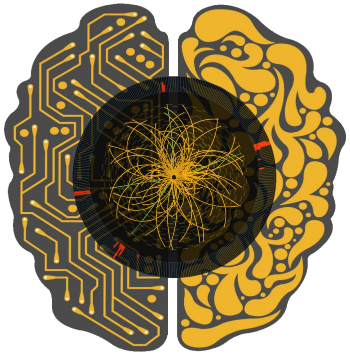
Pierre Baldi,[1] Kevin Bauer,[2] Clara Eng,[3] Peter Sadowski

Deep Learning and Its Application to LHC Physics

Dan Guest,[1] Kyle Cranmer,[2] and Daniel Whiteson[1]

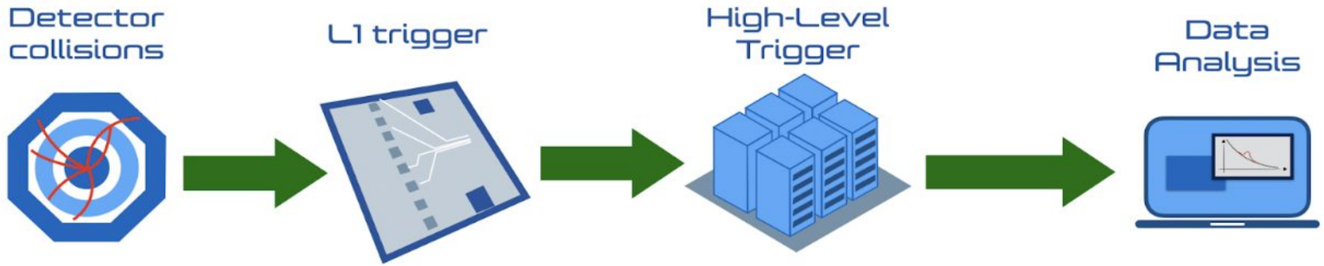
[1]Department of Physics and Astronomy, University of California, Irvine, California 92697, USA

[2]Physics Department, New York University, New York, NY 10003, USA



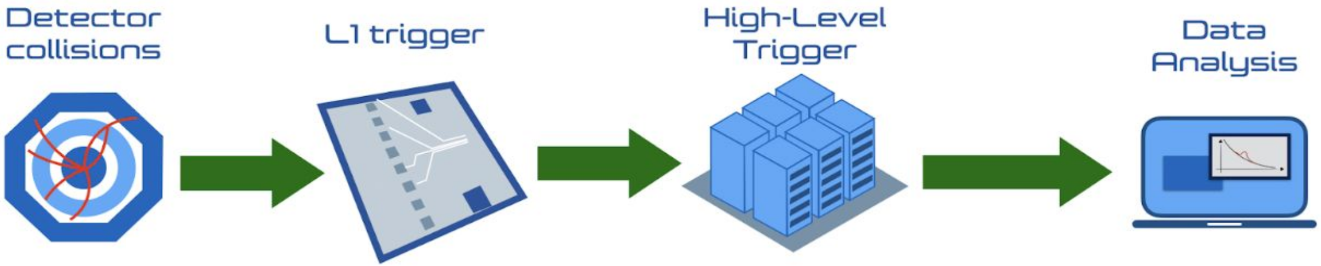
Can Deep Neural Networks be used to identify interesting events online at trigger level?

Idea



Detector collisions
L1 trigger
High-Level Trigger
Data Analysis

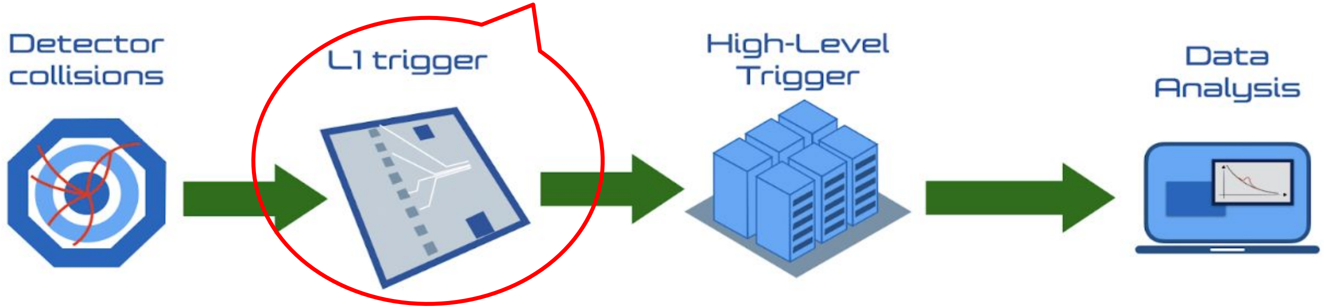
Idea



⚠ Events that are discarded by the trigger are **lost**!

Idea

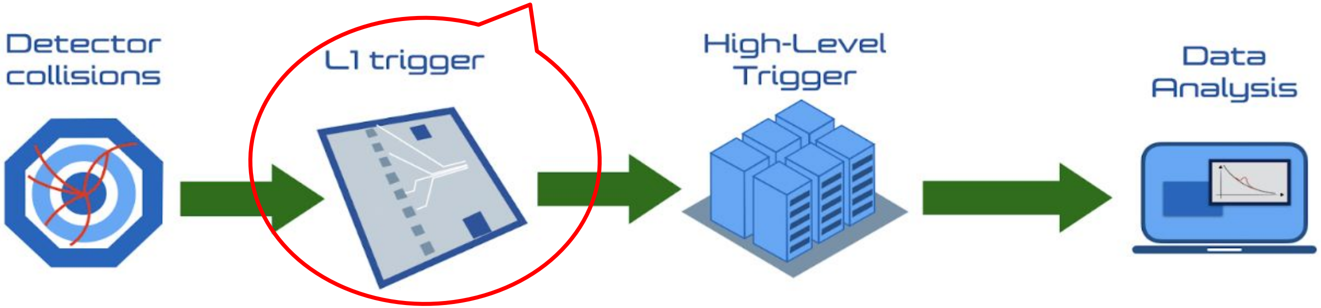
L1 of data processing typically uses custom hardware with FPGAs



Events that are discarded by the trigger are **lost**!

Idea

L1 of data processing typically uses custom hardware with FPGAs



Events that are discarded by the trigger are **lost**!

Let's run Deep Neural Networks in real-time on FPGAs to improve event selection!

Running Deep Neural Networks on FPGAs

FPGAs (Field-Programmable Gate Arrays) are programmable integrated circuits.

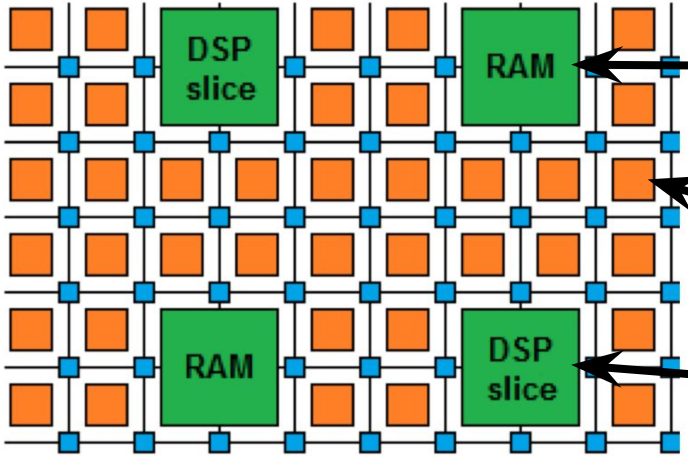


Image source

Random Access Memories to store constant values

Logic cells for simple arithmetic operations

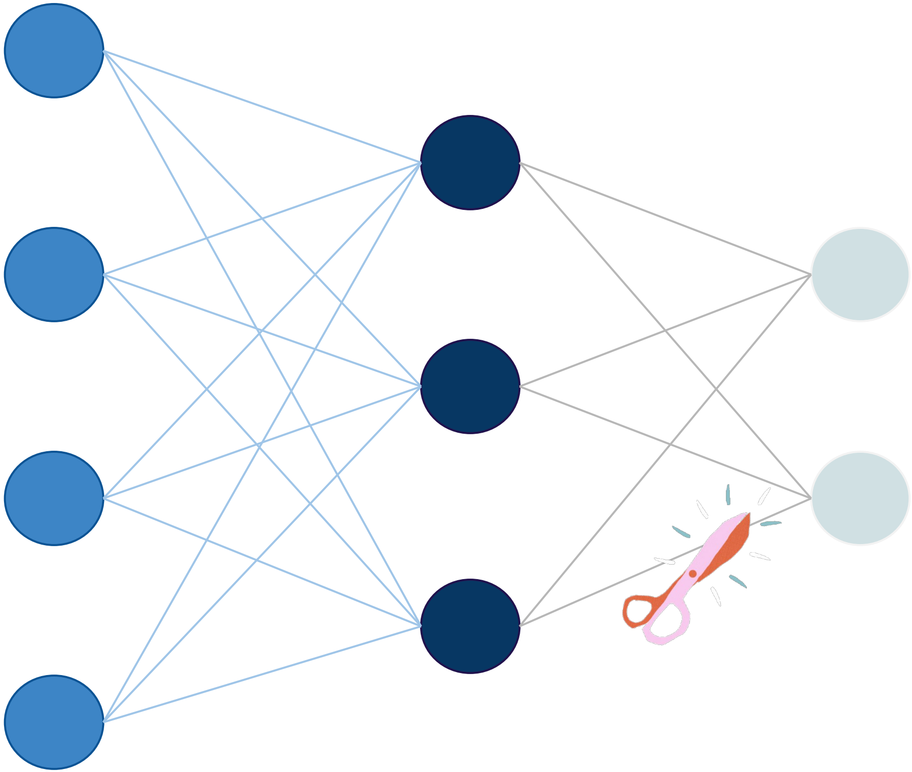
Digital Signal Processors to perform multiplications

Depending on the FPGA resources available, we should know how to **reduce the size** of a network

Pruning

One way of **reducing** the size of a neural network is **pruning**.

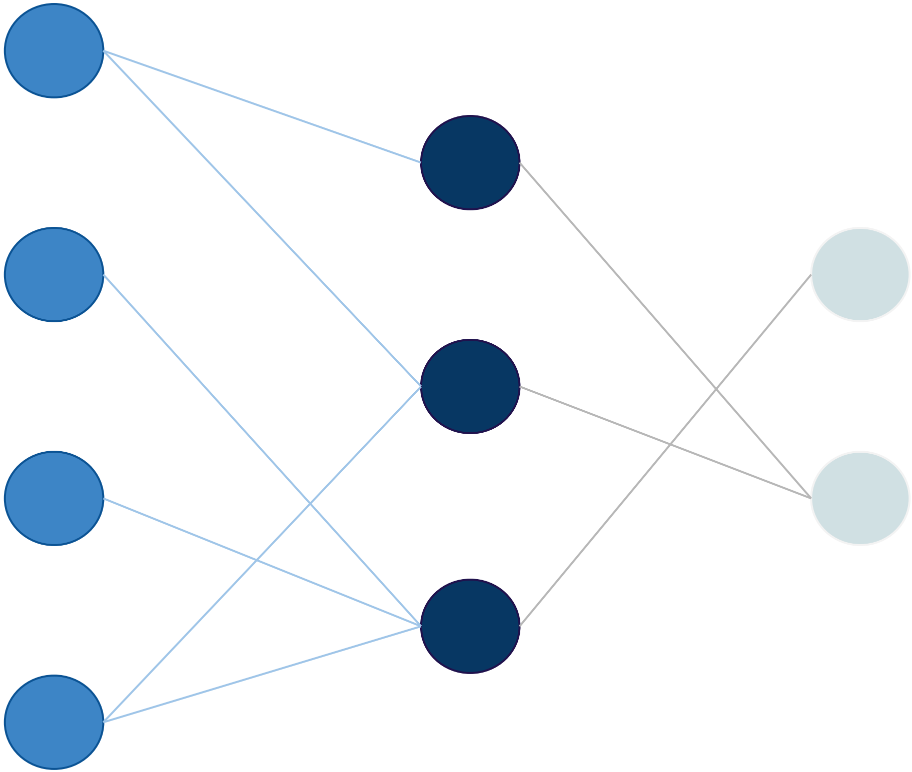
Pruning = **removing** superfluous structure



Pruning

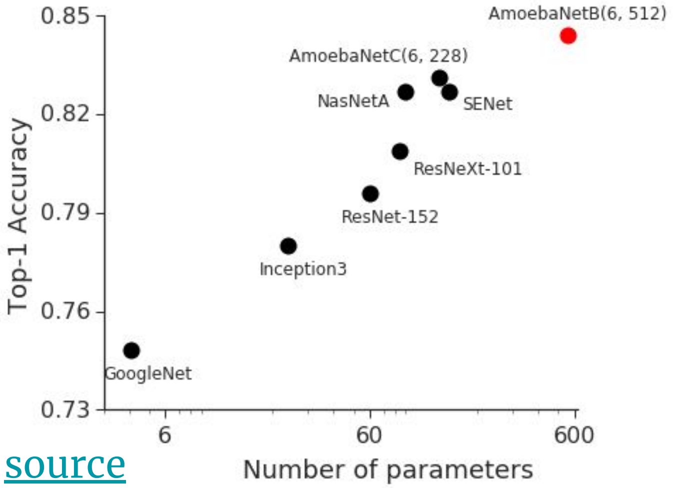
One way of **reducing** the size of a neural network is **pruning**.

Pruning = **removing** superfluous structure



Why pruning?

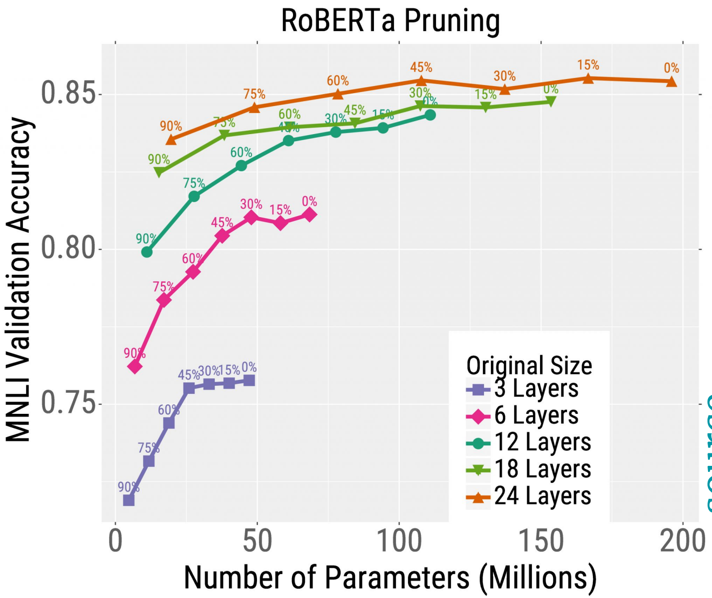
Bigger networks are usually more **accurate**



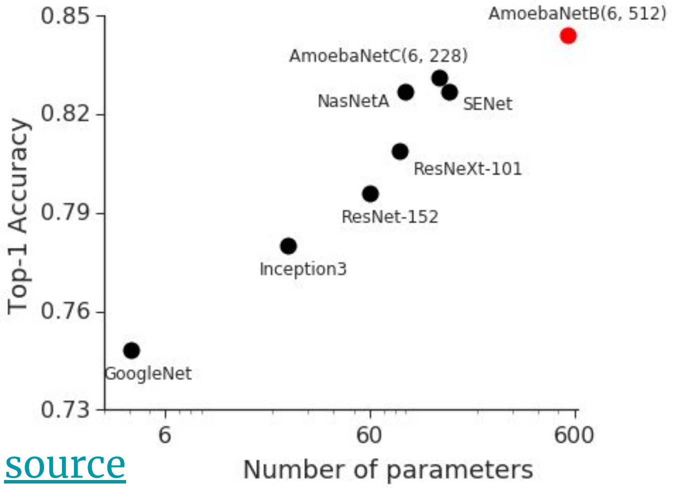
source

Why pruning?

Bigger networks are usually more **accurate**



source

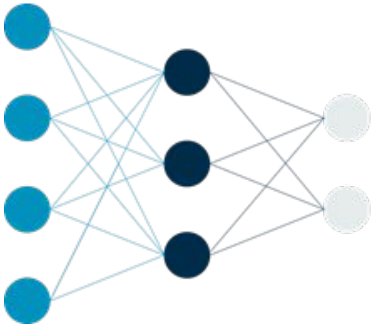


source

➔ Best to start out with very large models and prune with **minimal** performance penalty

Usual pruning scheme

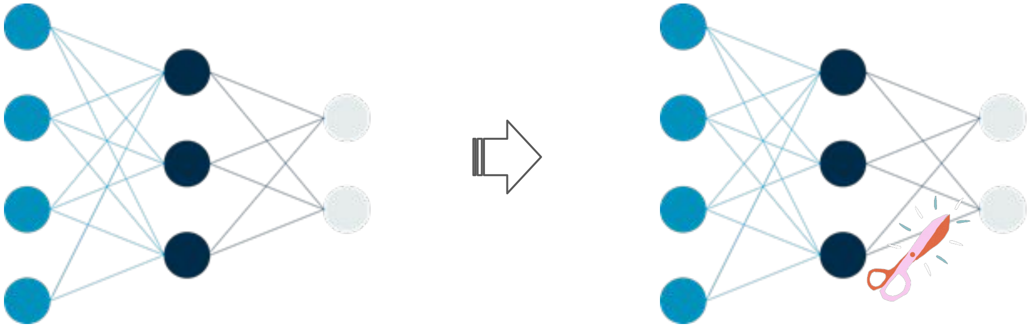
1. Train



Davis Blalock et al., *What is the state of neural network pruning?*, Proceedings of machine learning and systems 2 (2020), pp. 129–146

Usual pruning scheme

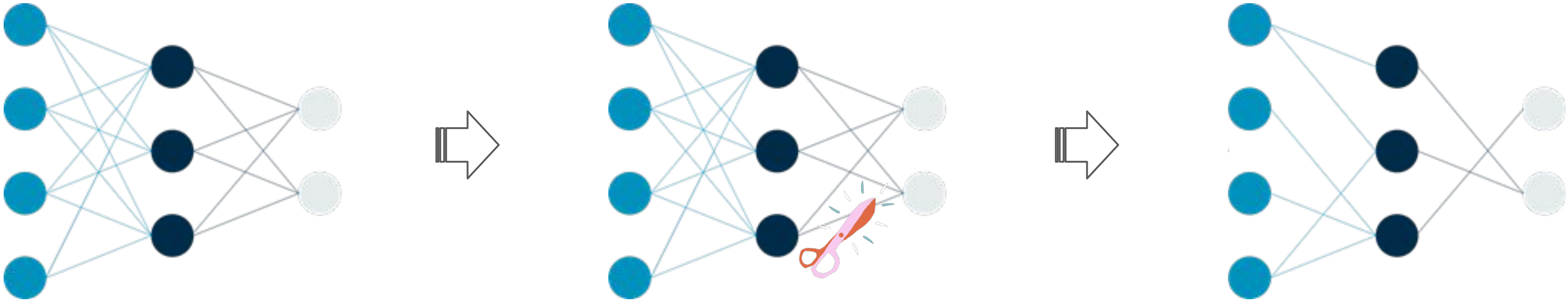
1. Train
2. Prune weights



Davis Blalock et al., *What is the state of neural network pruning?*, Proceedings of machine learning and systems 2 (2020), pp. 129–146

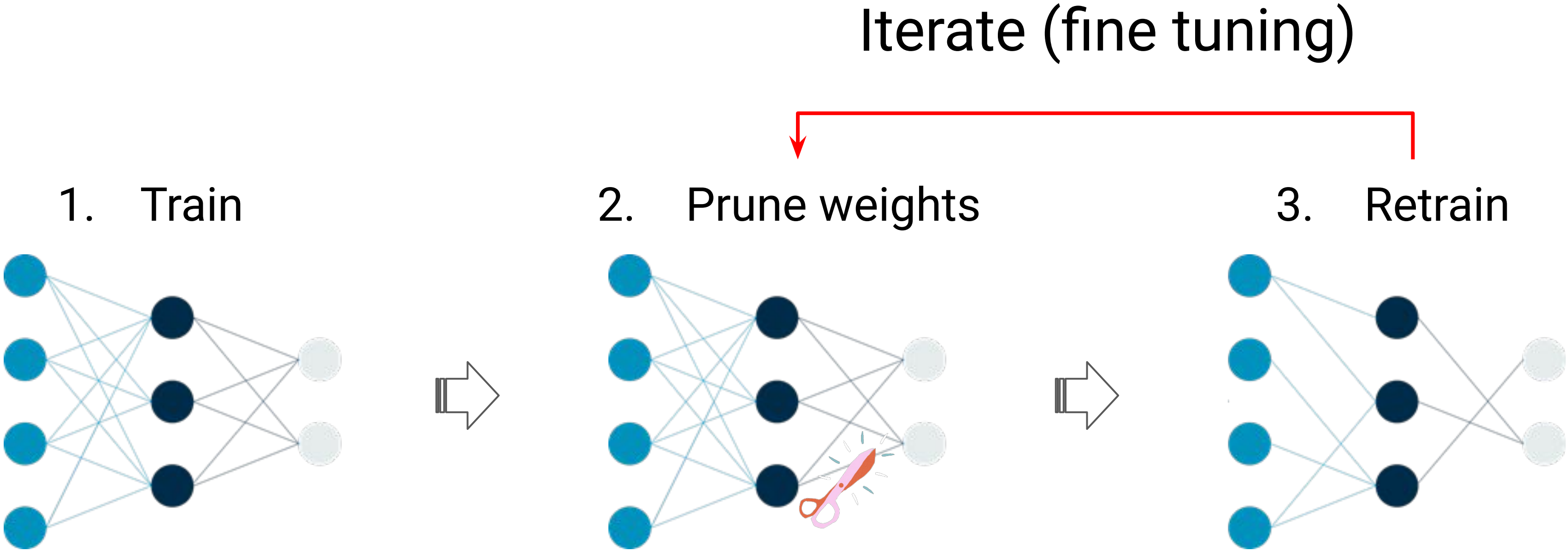
Usual pruning scheme

1. Train
2. Prune weights
3. Retrain



Davis Blalock et al., *What is the state of neural network pruning?*, Proceedings of machine learning and systems 2 (2020), pp. 129–146

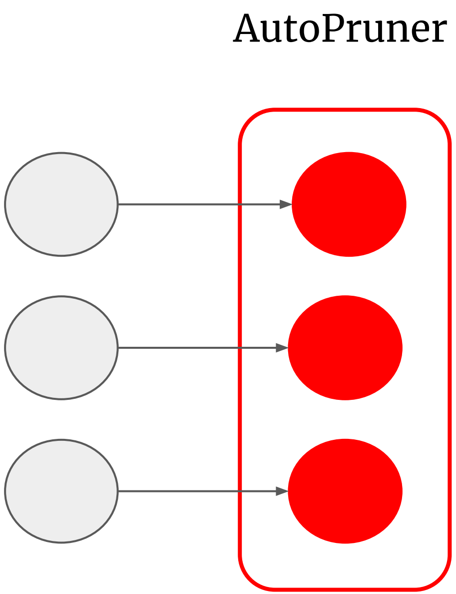
Usual pruning scheme



Davis Blalock et al., *What is the state of neural network pruning?*, Proceedings of machine learning and systems 2 (2020), pp. 129–146

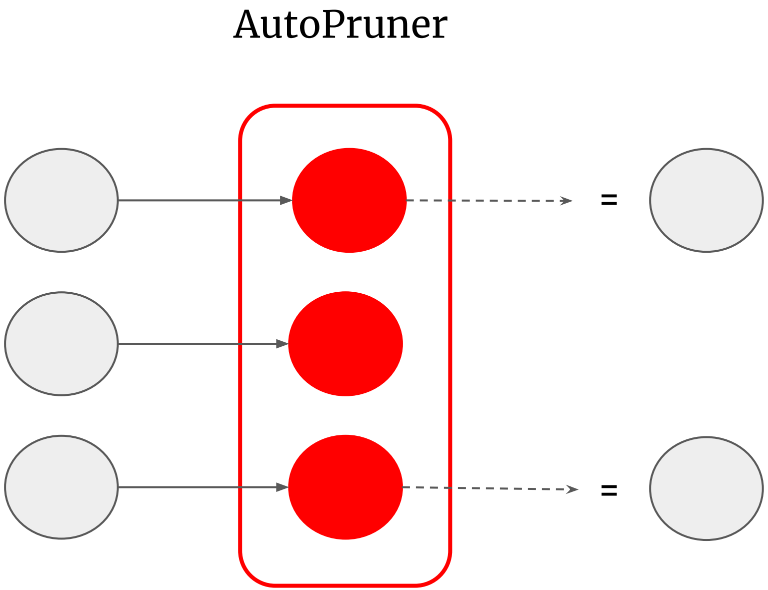
A different pruning strategy

- it can prune **nodes**
- it prunes **during training**
- the number of nodes to be pruned can be determined by the **user**
- it can determine the most suitable **network architecture**



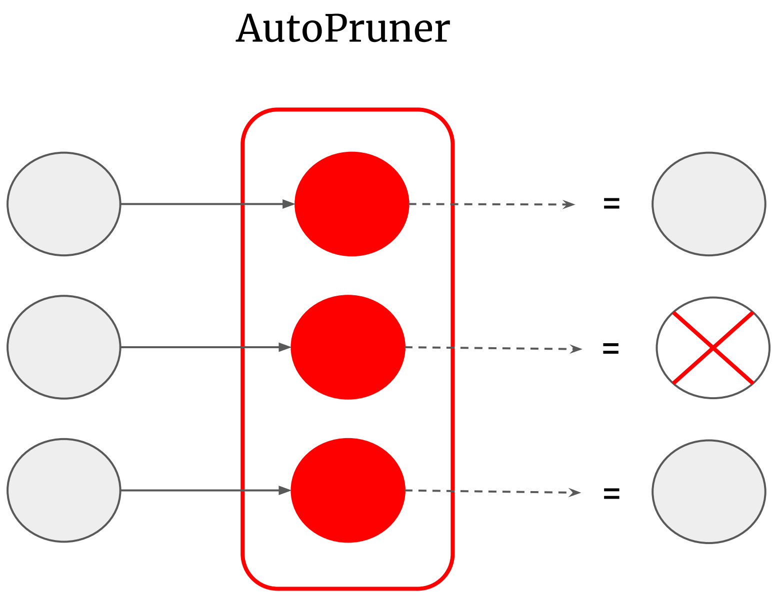
A different pruning strategy

- it can prune **nodes**
- it prunes **during training**
- the number of nodes to be pruned can be determined by the **user**
- it can determine the most suitable **network architecture**



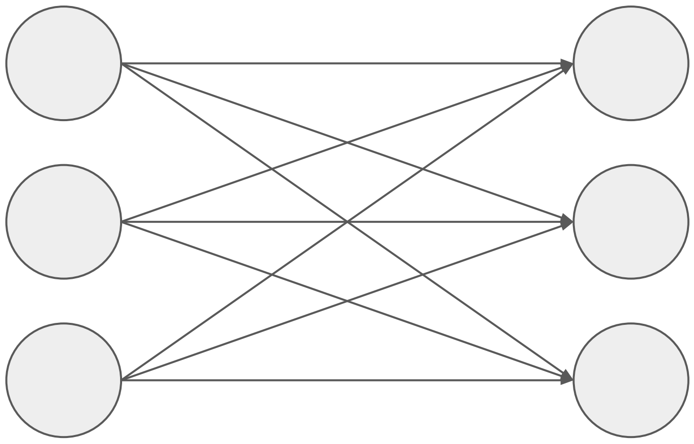
A different pruning strategy

- it can prune **nodes**
- it prunes **during training**
- the number of nodes to be pruned can be determined by the **user**
- it can determine the most suitable **network architecture**



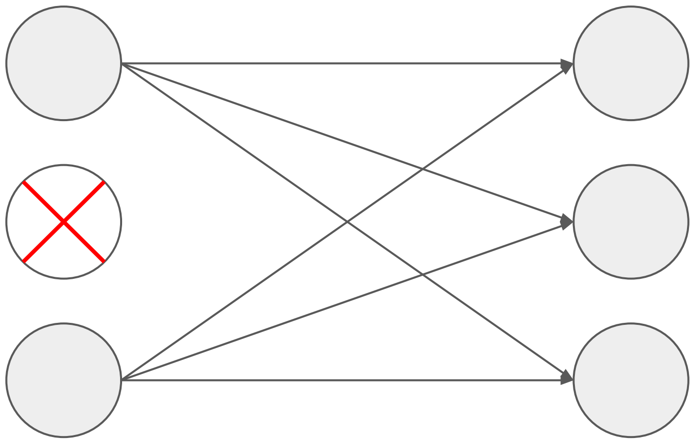
A different pruning strategy

- it can prune **nodes**
- it prunes **during training**
- the number of nodes to be pruned can be determined by the **user**
- it can determine the most suitable **network architecture**



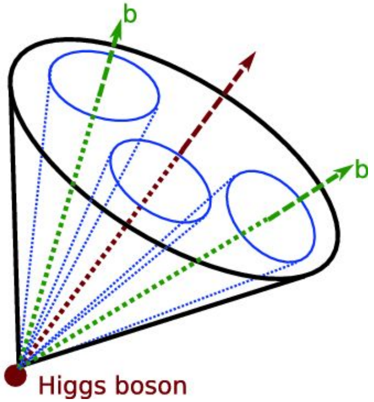
A different pruning strategy

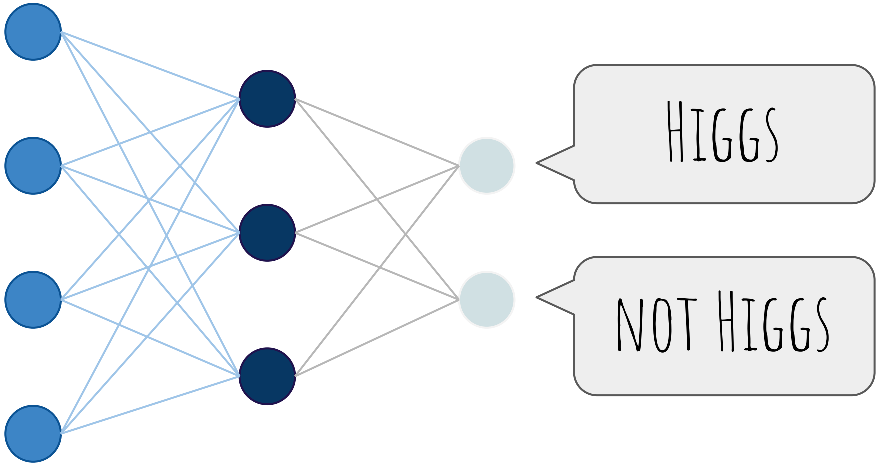
- it can prune **nodes**
- it prunes **during training**
- the number of nodes to be pruned can be determined by the **user**
- it can determine the most suitable **network architecture**



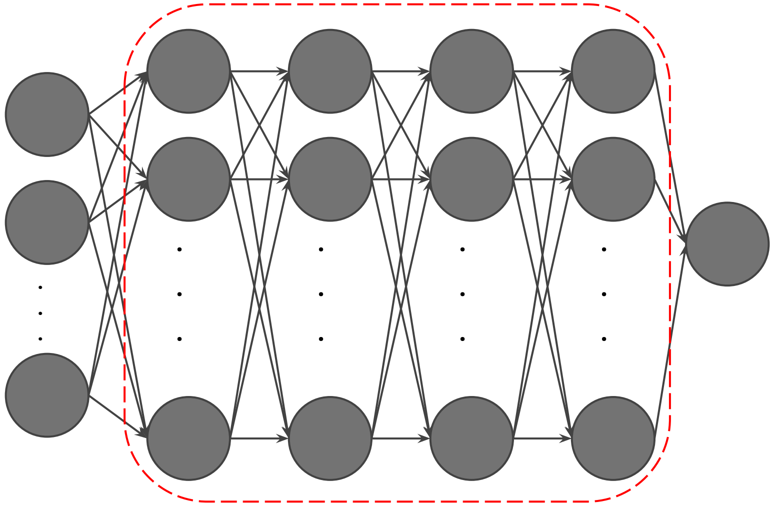
Use case

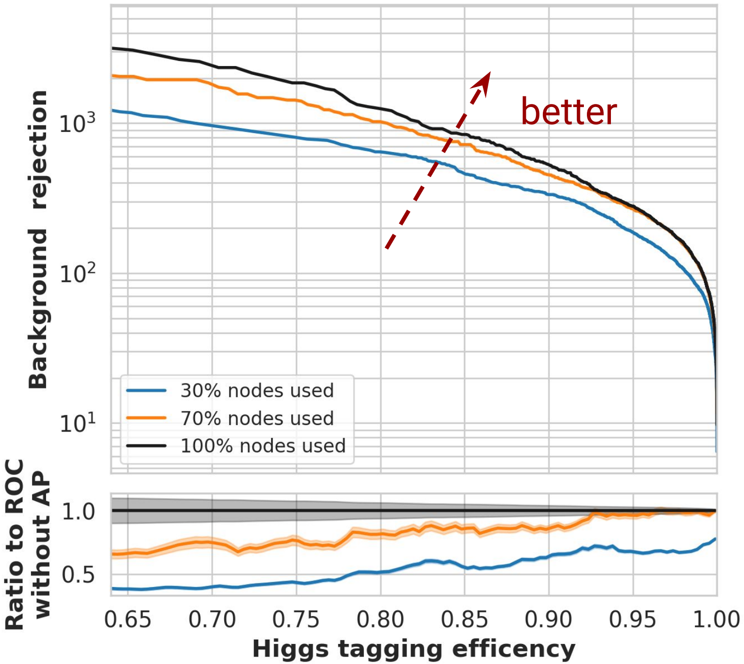
Identify jets that contain both the *b* quarks from boosted Higgs decay in *pp* collision experiments using Deep Neural Networks





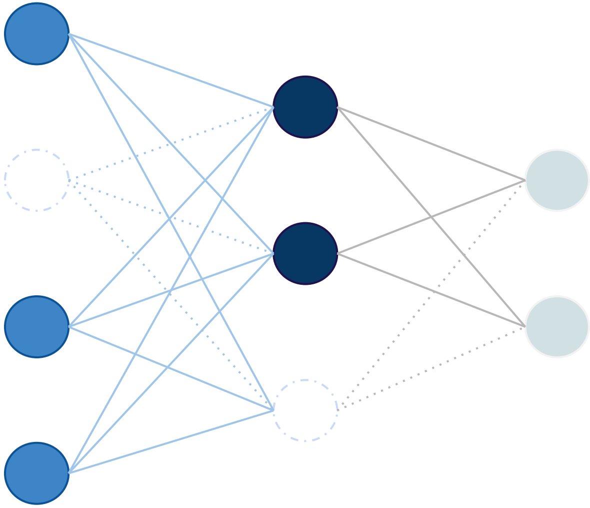
Results



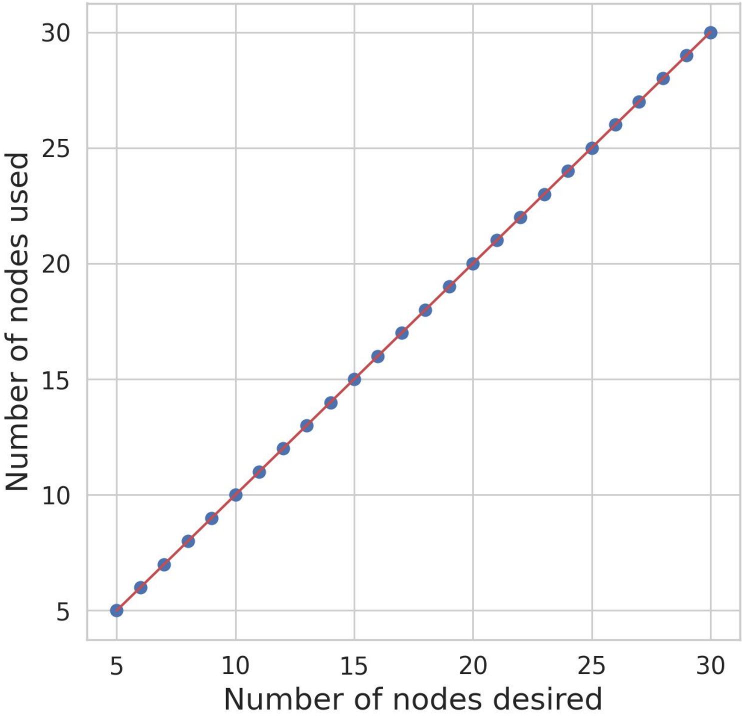


The performance increases with the percentage of nodes used, as expected: AutoPruner is really **switching off** nodes

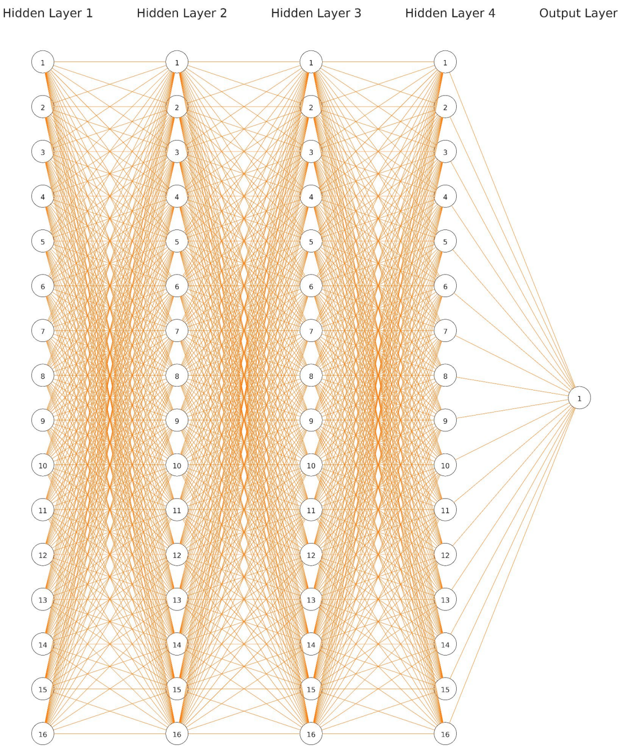
Results



The total number of nodes used is **always** equal to the required number

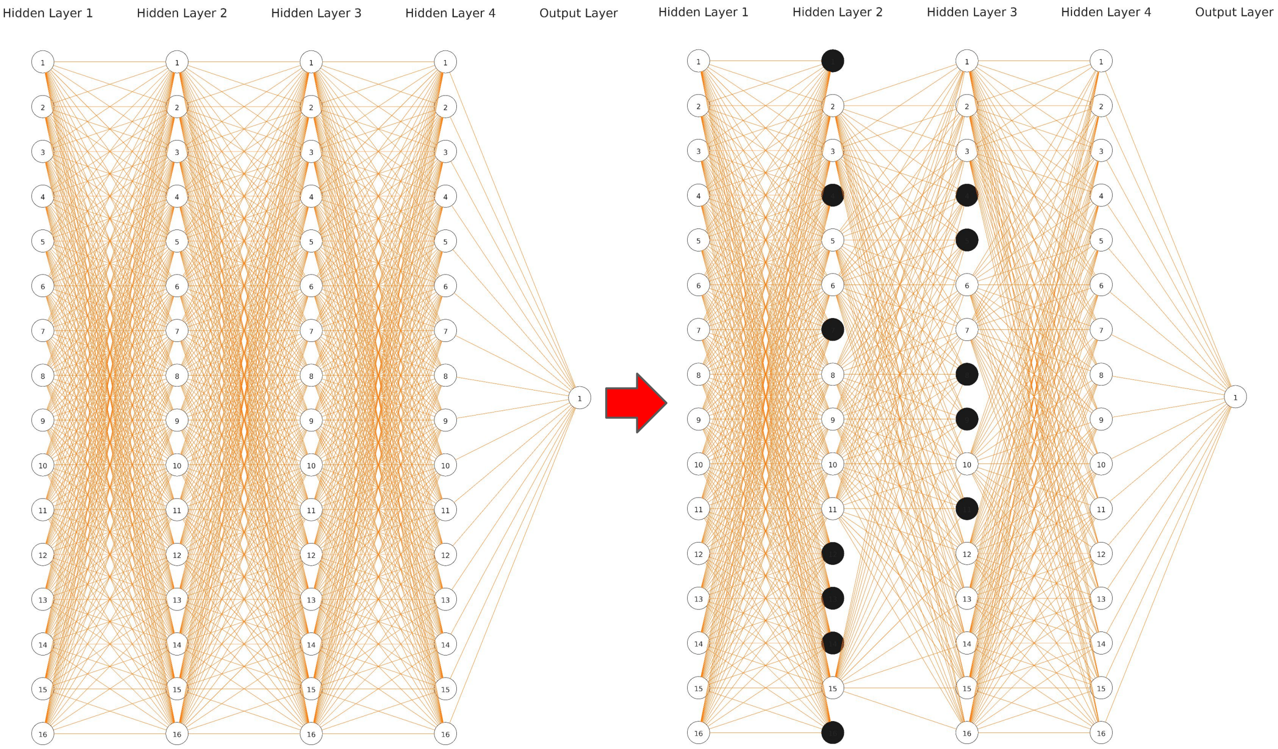


Results

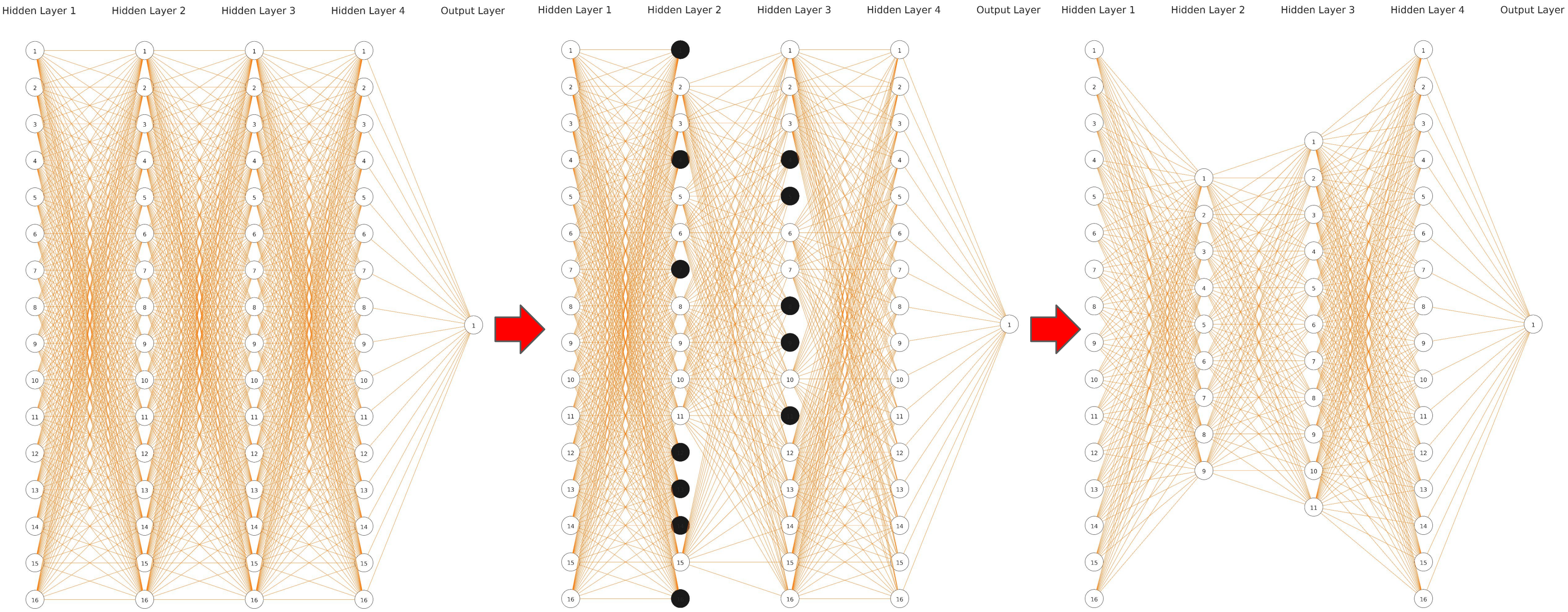


Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Hidden Layer 4
Output Layer

Results

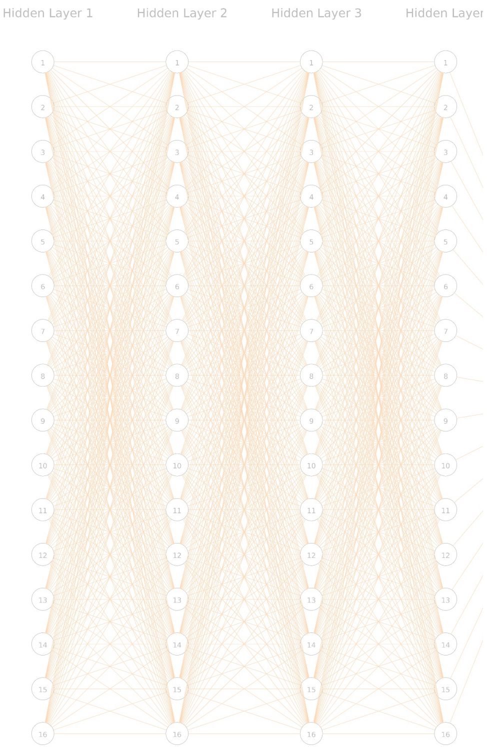

Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Hidden Layer 4
Output Layer
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Hidden Layer 4
Output Layer

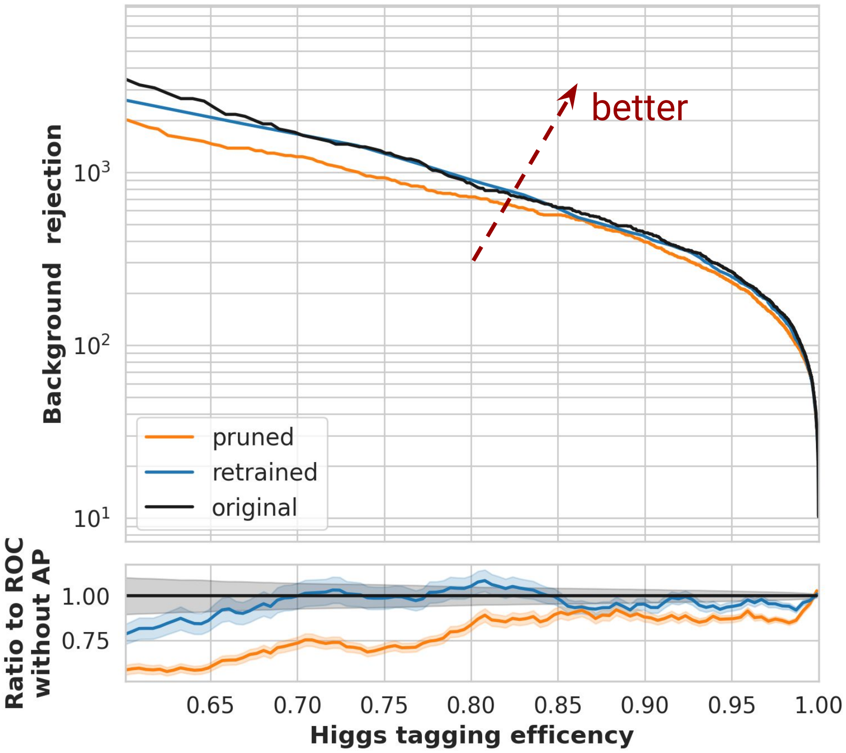
Results

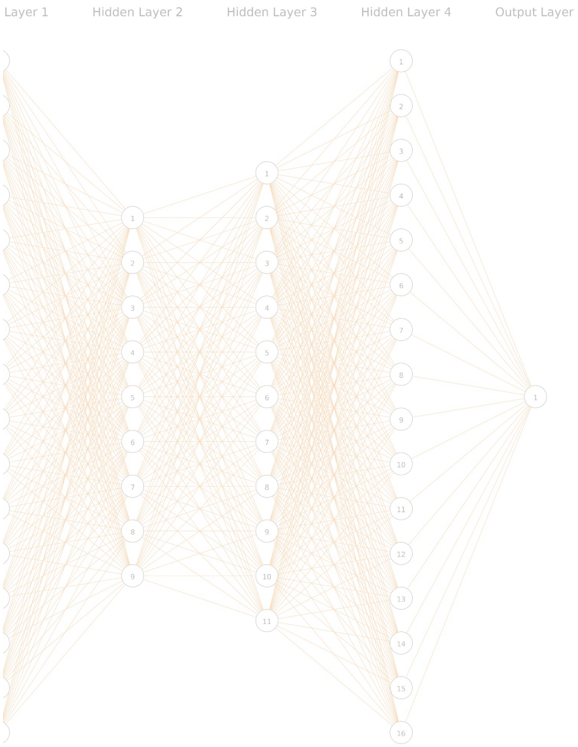


Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Hidden Layer 4
Output Layer
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Hidden Layer 4
Output Layer
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Hidden Layer 4
Output Layer

Results

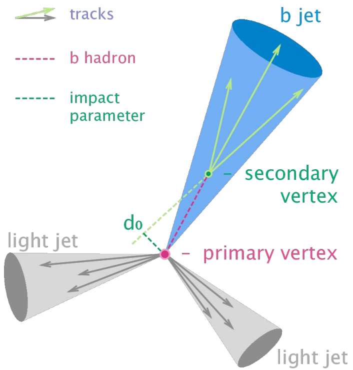

Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Hidden Layer


better
Background rejection
10^3
10^2
10^1
pruned
retrained
original
Ratio to ROC without AP
1.00
0.75
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Higgs tagging efficency


Layer 1
Hidden Layer 2
Hidden Layer 3
Hidden Layer 4
Output Layer

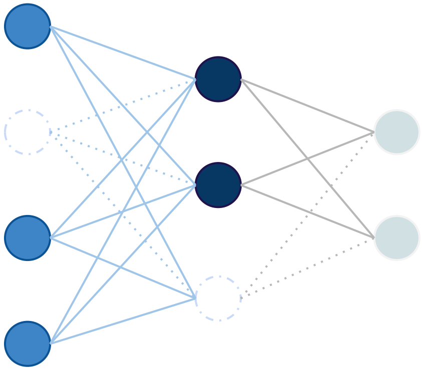
Future perspectives

Apply AutoPruner to Deep Neural Networks currently used in the *ATLAS Flavour Tagging Working Group* to **improve** tagging algorithms

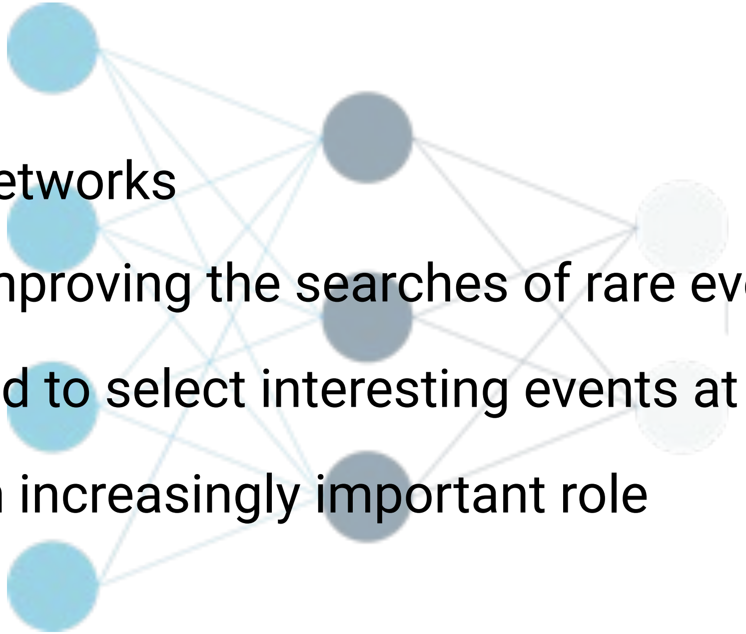


ATLAS Flavour Tagging Working Group

Investigate how our pruning strategy can improve the significance level of predictions by **reducing** the propagation of **uncertainties**



Summary



Deep Neural Networks

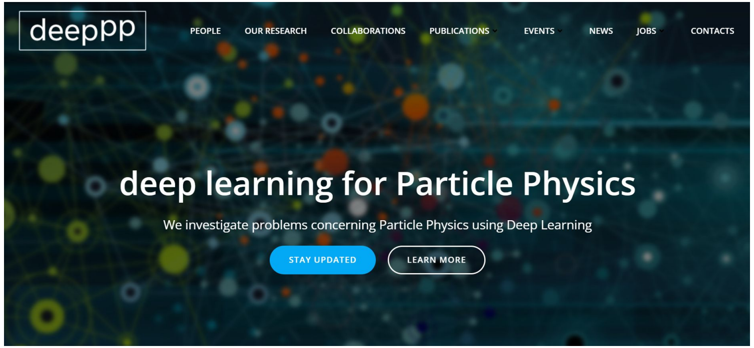
- can help improving the searches of rare events
- can be used to select interesting events at trigger level
- will play an increasingly important role

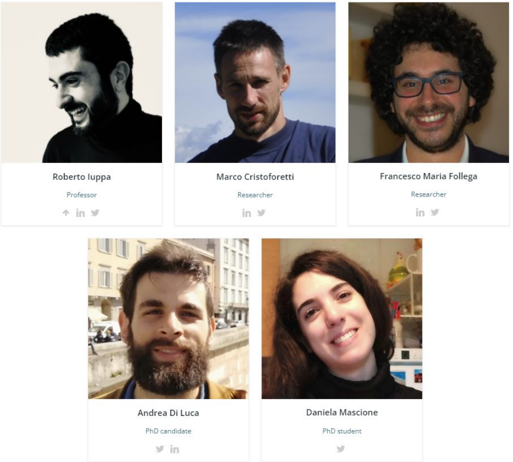
Acknowledgements

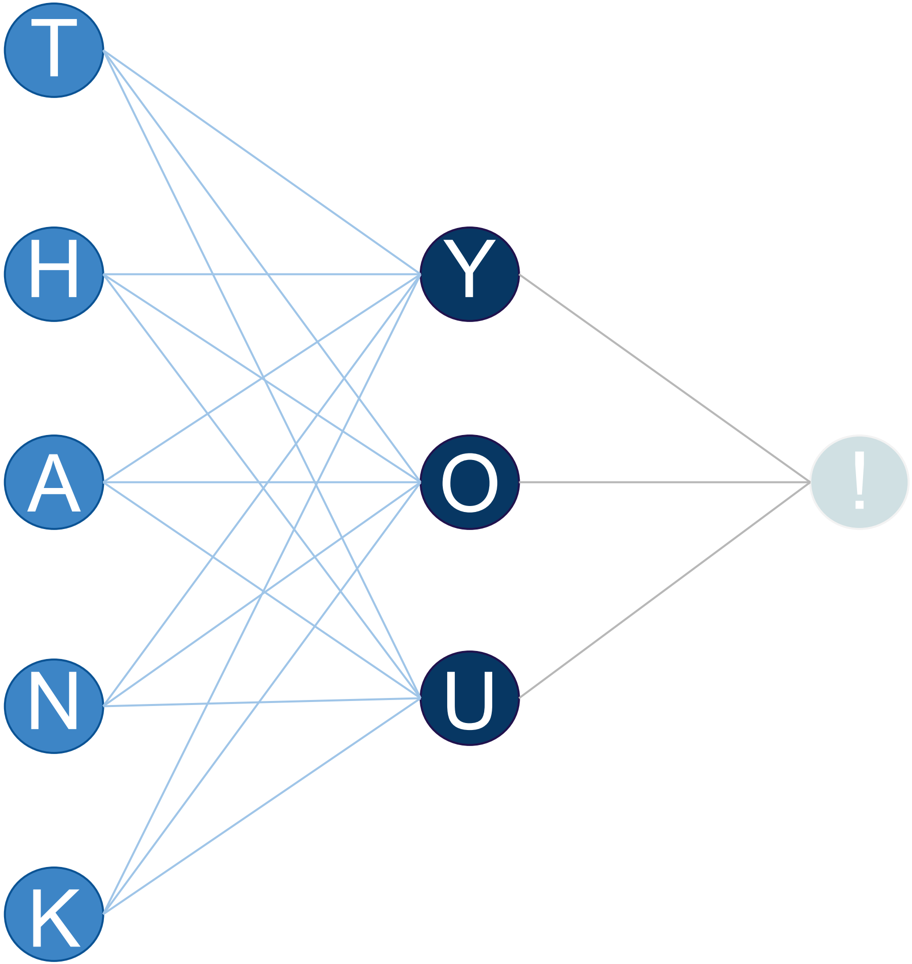
This work is a joint effort of the deepPP group of the University of Trento and FBK

You can find more about about Deep Learning applications in Particle Physics and our work here: https://www.deeppp.eu/

deeppp

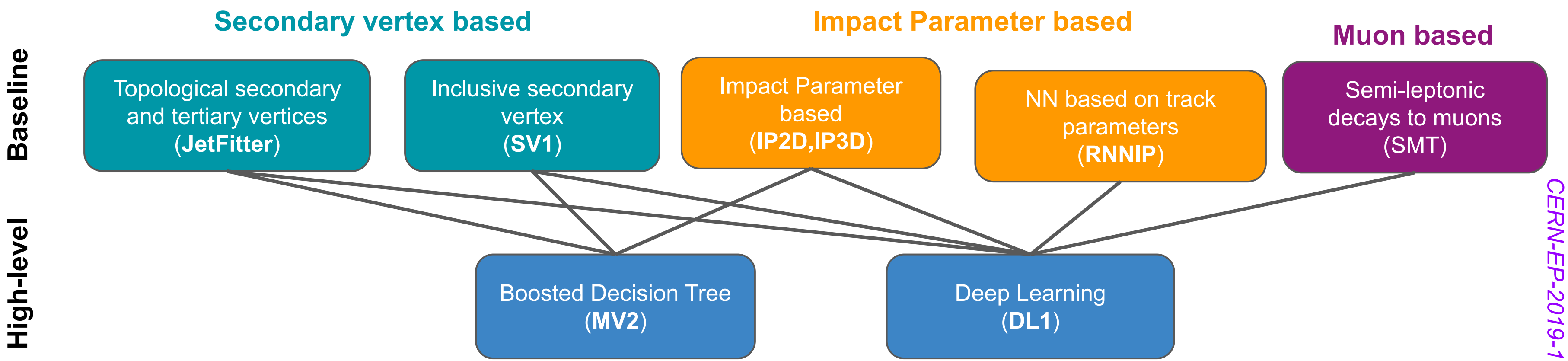





T
H
A
N
K
Y
O
U
!

ADDITIONAL MATERIAL

Flavour Tagging Strategies in ATLAS



CERN-EP-2019-132

3 types of algorithms were designed employing topologies of b-hadrons

- **Impact parameter (IP) based**
- **Secondary/tertiary vertex (SV) based**
- **Soft Muon based**

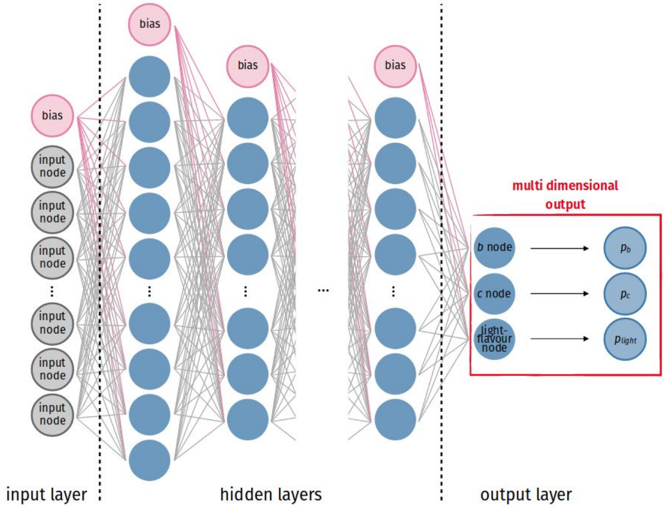
High-level taggers (MV2 e DL1) **combine all this information**

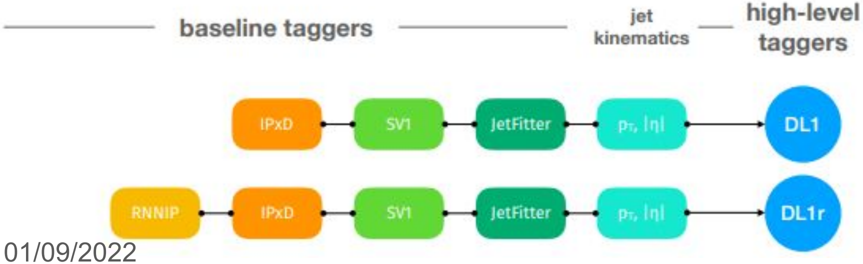
Deep-Learning Flavour Tagger (DL1) - Architecture

- Neural Network with fully connected layers with 8 hidden layers with Relu activation function.
- Multi-class output (also allows c-tagging without dedicated training):

$$\mathrm{DL1}_{b-score} = \ln\left(\frac{p_b}{f_c \cdot p_c + (1-f_c)\cdot p_{light-flavour}}\right)$$

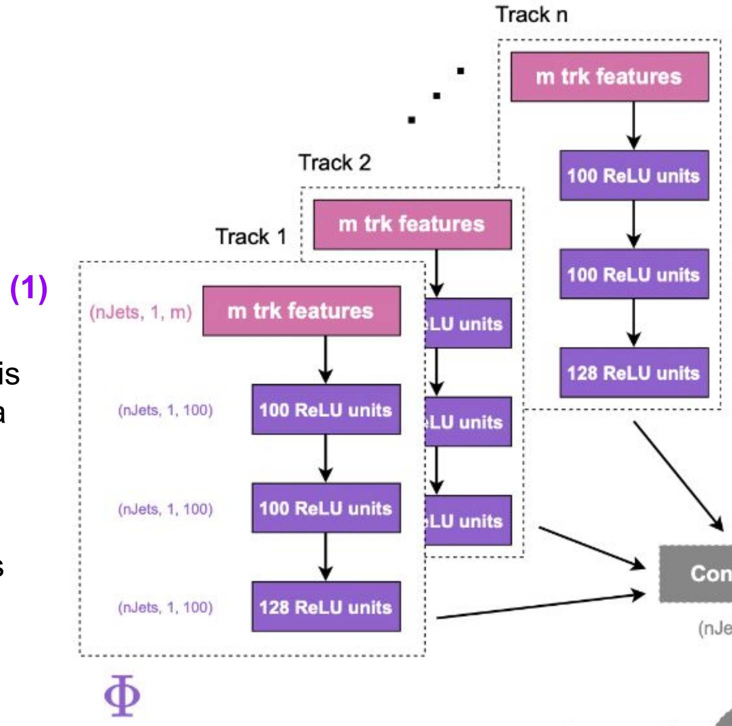
- Depending on which baseline tagger is used, we can distinguish different algorithms
  - **DL1** (28 input features as MV2)
  - **DL1r** (44 input features, RNNIP added)

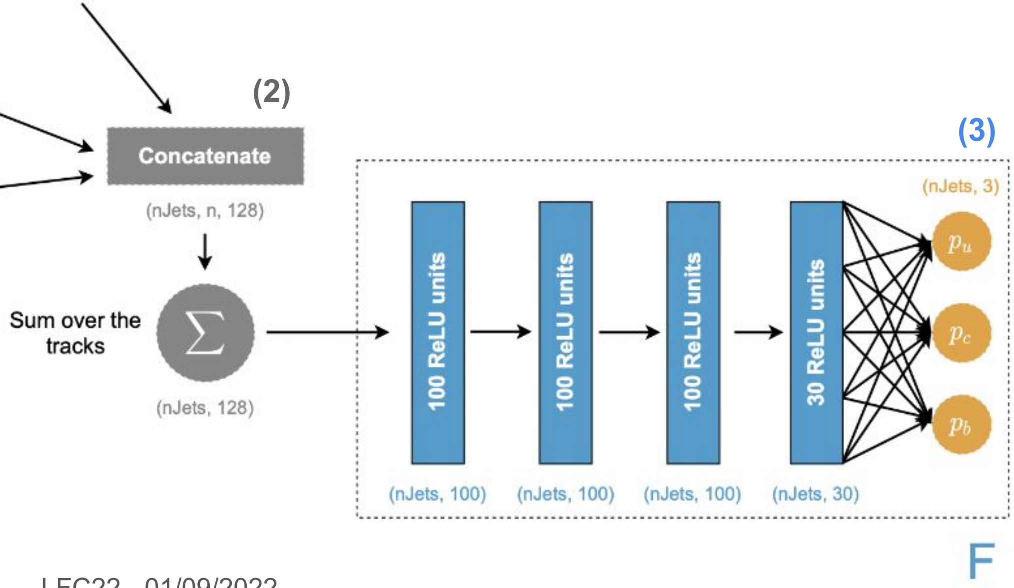




DIPS

The **Deep Impact Parameter Sets** (DIPS) tagger uses a **new architecture for flavour tagging** which treats all tracks of one jet as an unordered, variable-sized set to identify jets originating from heavy flavour decays.





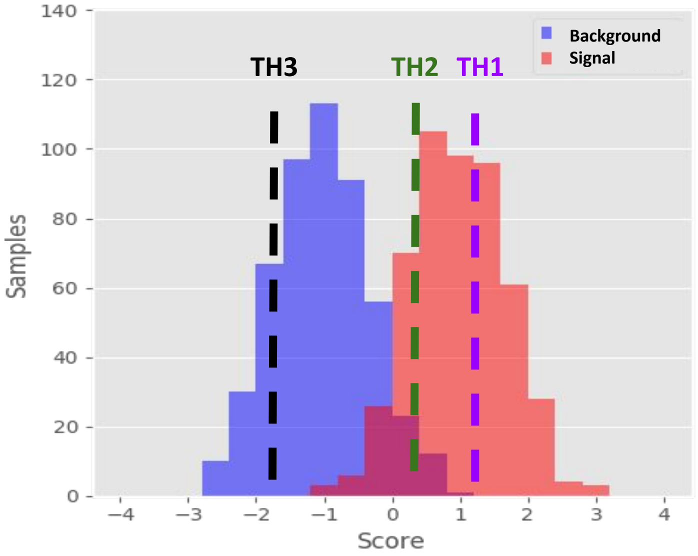
(1) Each track is processed by a neural network with shared weights between tracks (Φ network)

(2) Outputs of the neural network for the tracks are pooled for further processing

(3) Pooled outputs are used for classification by a subsequent neural network (F network)

Courtesy of Andrea Di Luca

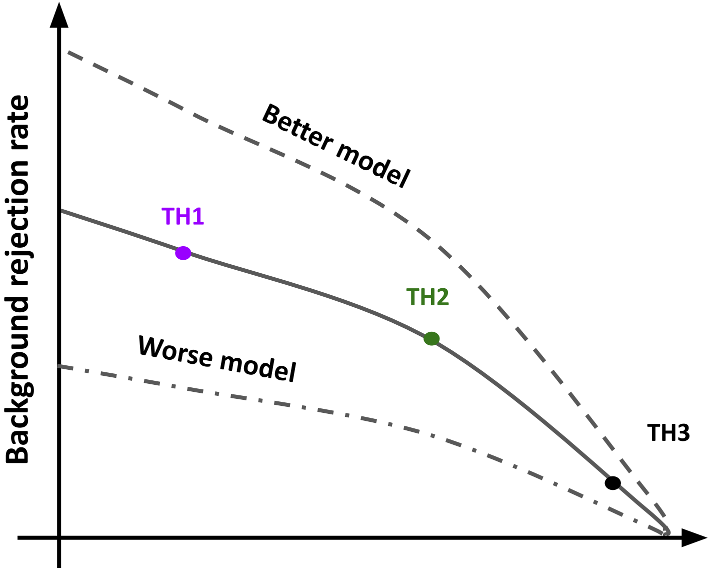
How to evaluate classifier performance



$$\text{Signal efficiency} = \frac{Signal > TH}{Signal}$$

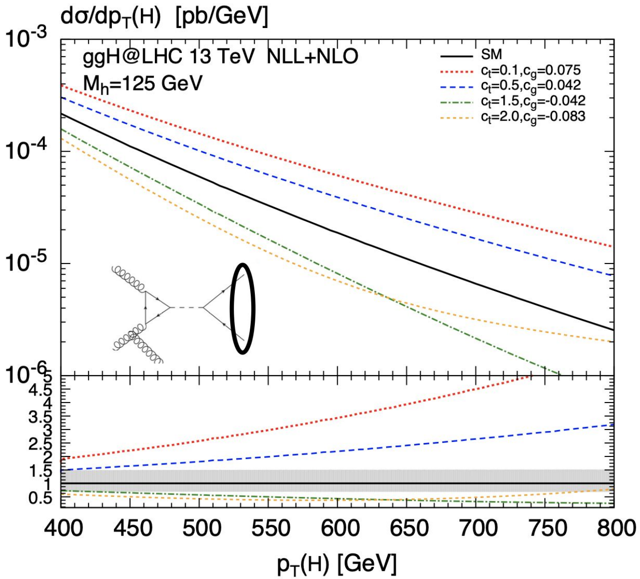
$$\text{Background rejection rate} = \frac{Background}{Background > TH}$$

Receiver Operating Characteristic (ROC) curves can be used to **compare performance of different models**.



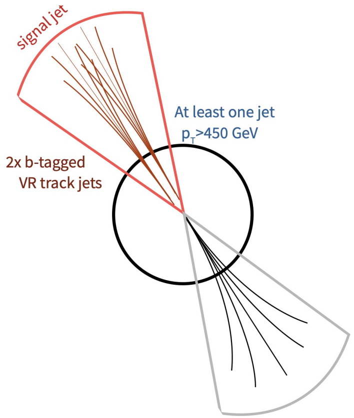
Based on Francesco Maria Follega's work
Courtesy of Andrea Di Luca

All-had H(bb) analysis (Full Run2)



Analysis Goals (V→qq, H→bb)
- Inclusive measurement
- pT differential measurement (STXS)
- fiducial measurement (pT,truth>450GeV)

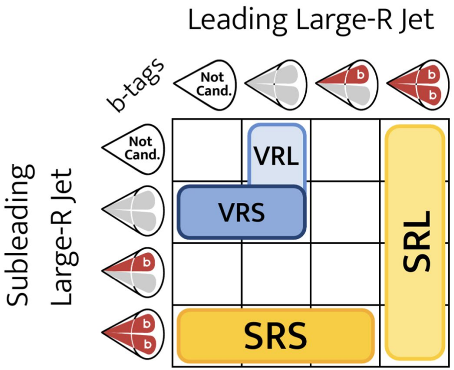
Event Selection
- Trigger, GRL, event and jet cleaning
- >= 1 large-R jet with pT>450 GeV, mJ> 60 GeV
- At least 2 large-R jets with pT> 200 GeV
- At least one signal candidate:
  - pT>450 GeV, mJ> 60 GeV
  - 2mJ/pT<1(boosted regime)
  - ΔR(VR1,VR2)/VR1>1
  - categorized in SR/VR based on VRjets



Event categorization

<u>Signal Regions</u>
- Inclusive SRL = [450,∞], SRS = [250,∞]
- pT bins: [250, 450], [450,650], [650, 1000]

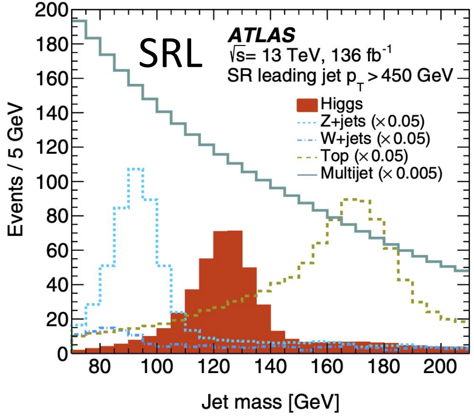
<u>Validation Regions</u>
- Same pT range as the inclusive
- used for QCD bkg modelling

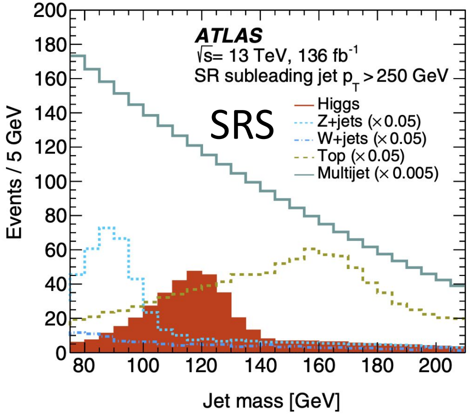


Results

Based on Francesco Maria Follega's work
Courtesy of Andrea Di Luca

Signal and backgrounds





- The dominant background process is **multijet production**, which exhibits a monotonically decreasing jet mass distribution.
- **Hadronically decaying vector bosons**, produced in association with jets (V + jets) and **events with one or two top quarks** (jointly referred to as Top) populate the jet mass regions below and above m_H respectively.

Recently (11th May 2022) published a paper (PRD link)

Inclusive

Result	μ_H	μ_Z	$\mu_{t\bar{t}}$
Expected	1.0 ± 3.2	1.00 ± 0.17	1.00 ± 0.07
Observed	0.8 ± 3.2	1.29 ± 0.22	0.80 ± 0.06

Fiducial volume
pT > 450 GeV
$|y_H| < 2$

Result	μ_H	μ_Z	$\mu_{t\bar{t}}$
Expected	1.0 ± 3.4	1.00 ± 0.18	1.00 ± 0.08
Observed	-0.1 ± 3.5	1.30 ± 0.22	0.75 ± 0.06

pT binned result