XX FRASCATI SUMMER SCHOOL

“BRUNO TOUSCHEK”

IN NUCLEAR, SUBNUCLEAR AND ASTROPARTICLE PHYSICS

LNF, July 11-15, 2022 Frascati (Italy)

INFN LNF
Istituto Nazionale di Fisica Nucleare
Laboratori Nazionali di Frascati

7th Young Researchers’ Workshop, Frascati 11 July 2022

Pruning Deep Neural Networks for LHC Challenges

Daniela Mascione

UNIVERSITÀ DI TRENTO

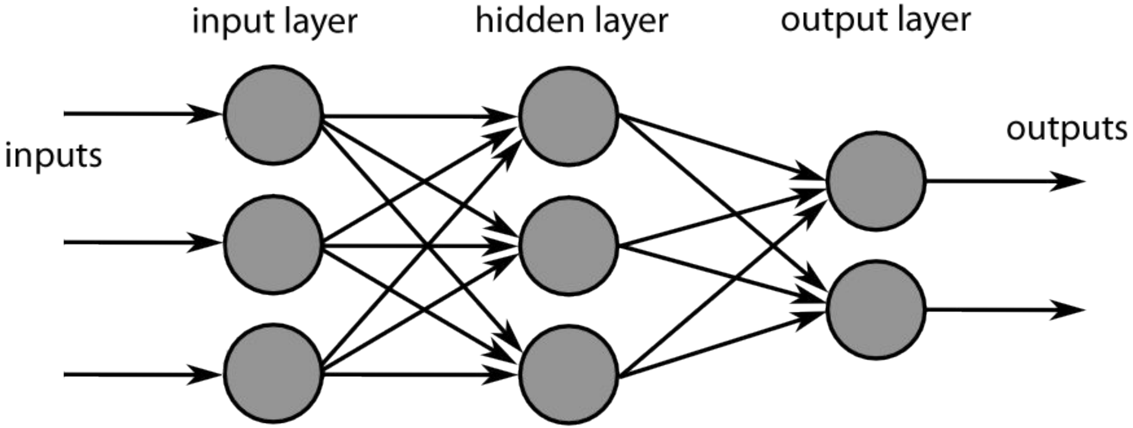
FONDAZIONE BRUNO KESSLER

deeppp

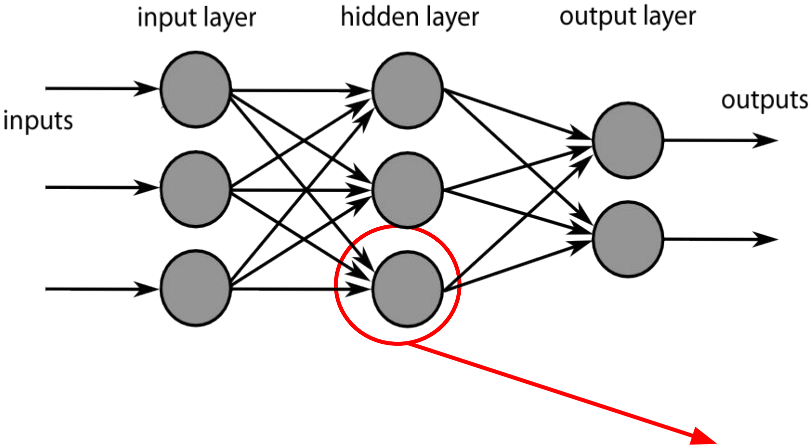
TIFPA Trento Institute for Fundamental Physics and Applications

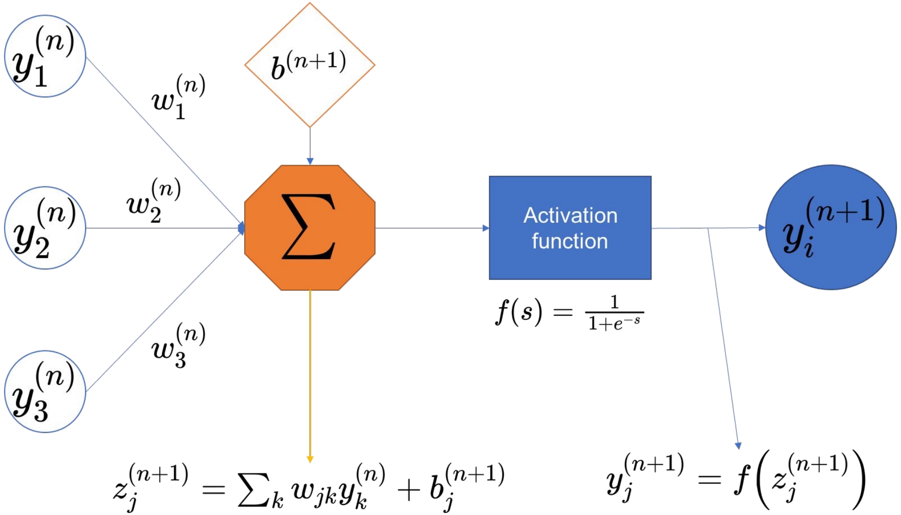
Deep Neural Networks

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.




$y_1^{(n)}$
$y_2^{(n)}$
$y_3^{(n)}$
$w_1^{(n)}$
$w_2^{(n)}$
$w_3^{(n)}$
$b^{(n+1)}$
Σ
Activation function
$f(s) = \frac{1}{1+e^{-s}}$
$y_i^{(n+1)}$
$z_j^{(n+1)} = \sum_k w_{jk} y_k^{(n)} + b_j^{(n+1)}$
$y_j^{(n+1)} = f\left(z_j^{(n+1)}\right)$

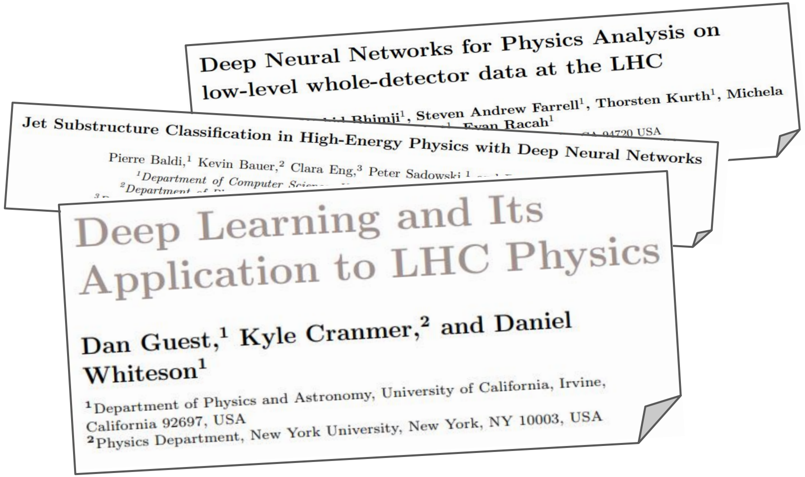
Nodes convert weighted inputs to outputs. The **weights keep getting updated** in the process of learning.

Image: F. Di Clemente

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

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

Evan Racah[1]

94720 USA

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

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

[1]*Department of Computer Scien*

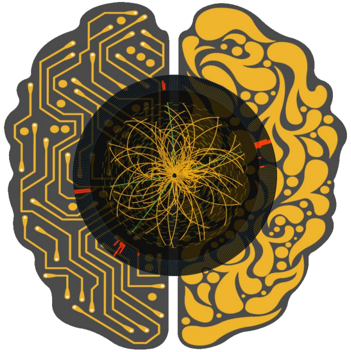
[2]*Department*

Deep Learning and Its Application to LHC Physics

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

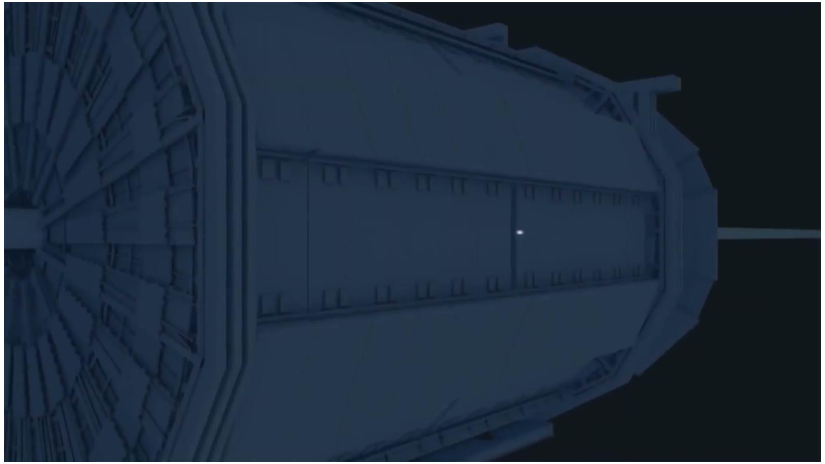
[1]Department of Physics and Astronomy, University of California, Irvine, California 92697, USA

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



The triggering challenge at LHC

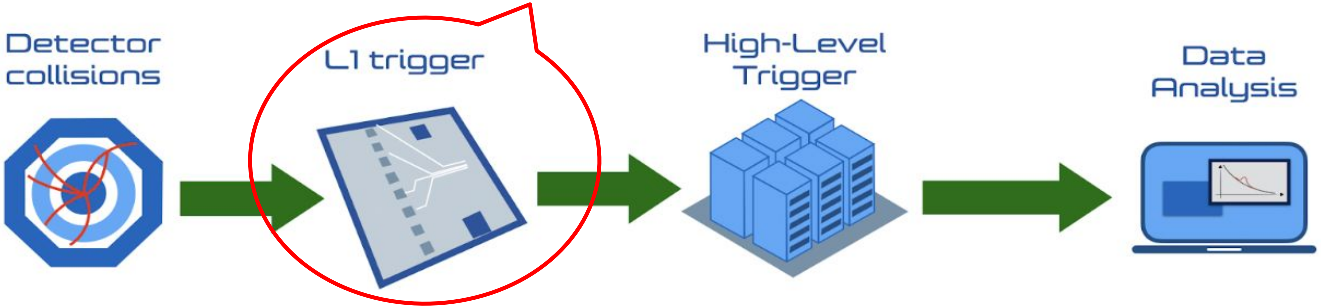
Triggering = **filter events** to reduce data rates to manageable levels



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

Idea

L1 of data processing typically uses custom hardware with FPGAs



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.

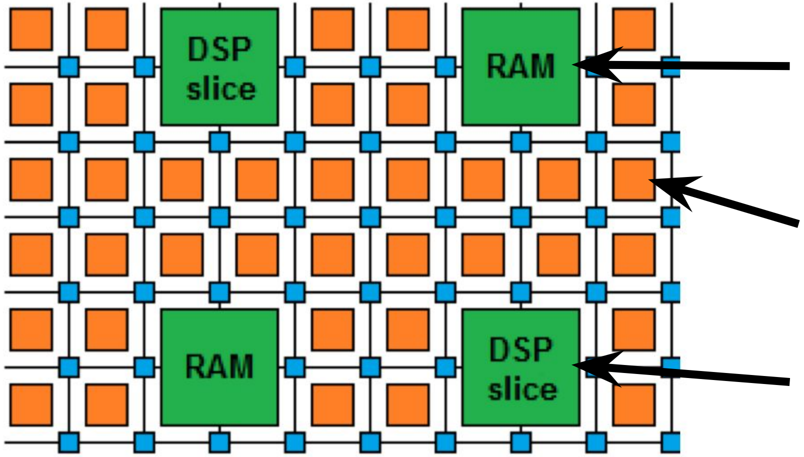


Image source

Random Access Memories to store constant values

Logic cells for simple arithmetic operations

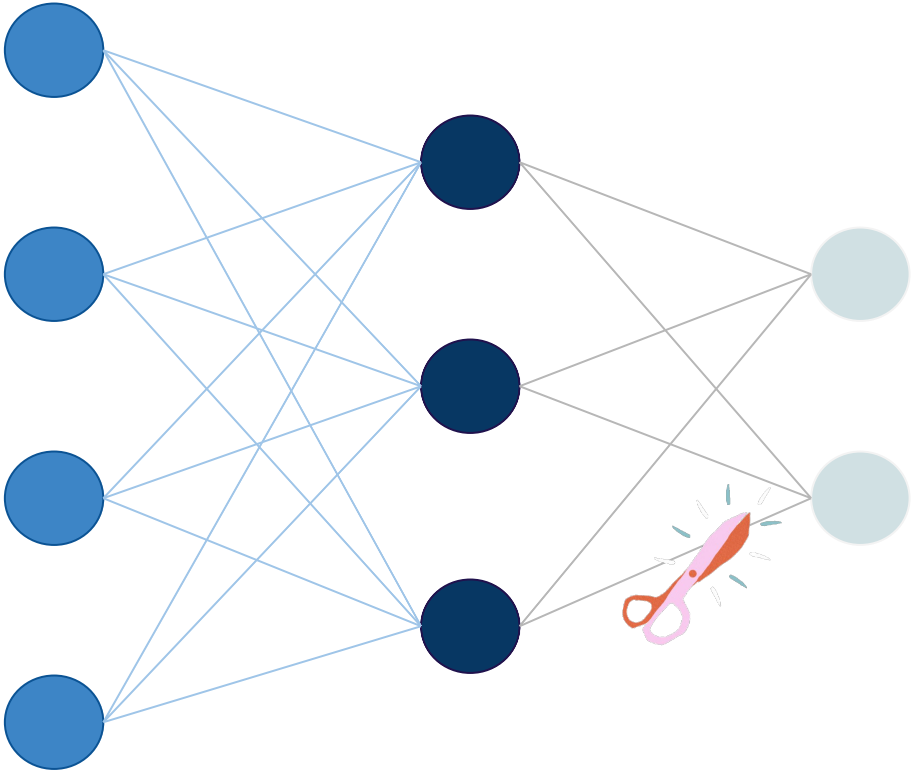
Digital Signal Processors to perform multiplications

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**.

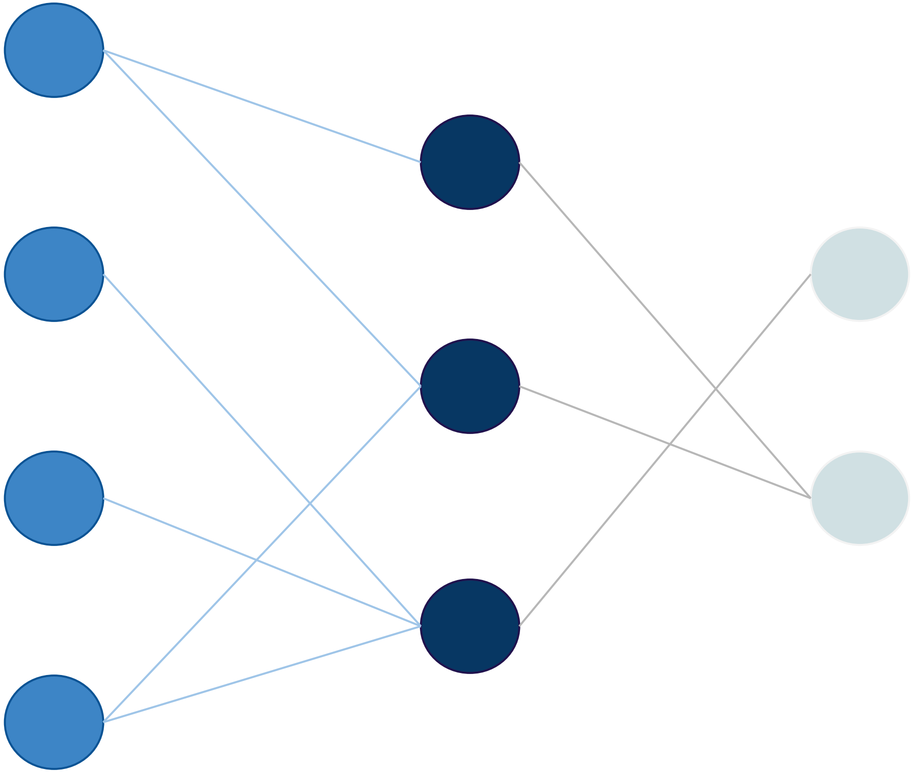
Pruning = **removing** superfluous structure



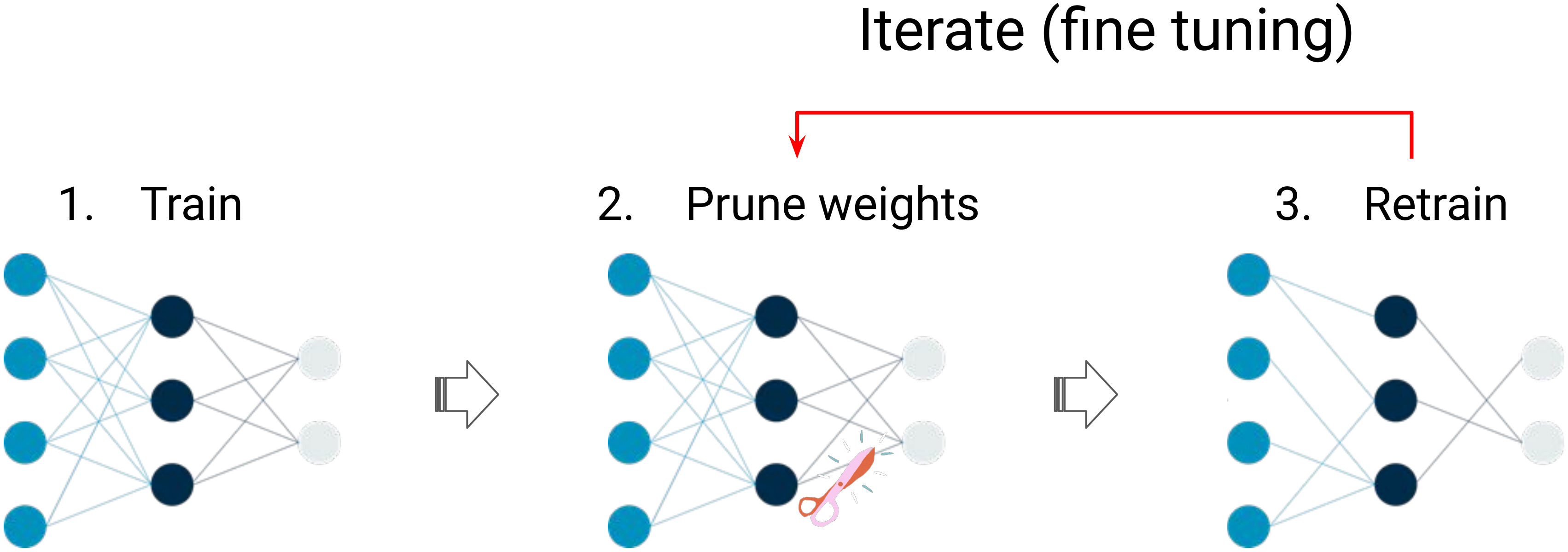
Pruning

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

Pruning = **removing** superfluous structure



Usual pruning scheme



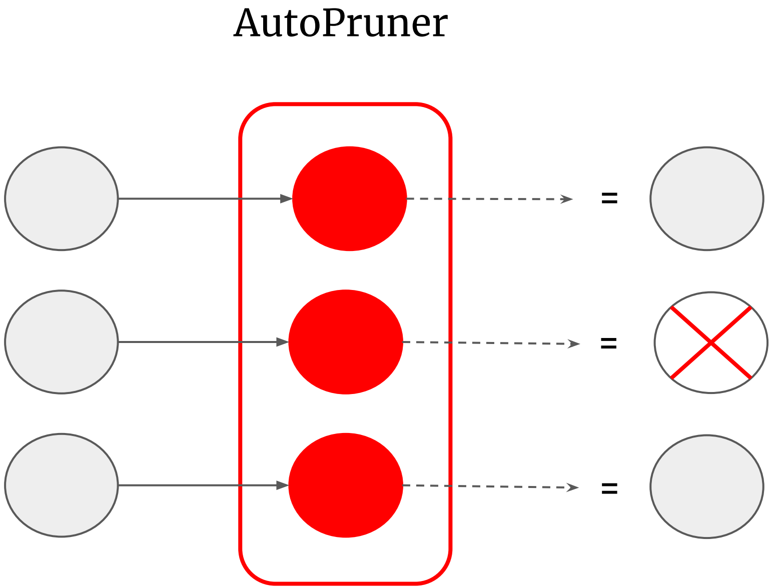
Iterate (fine tuning)

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

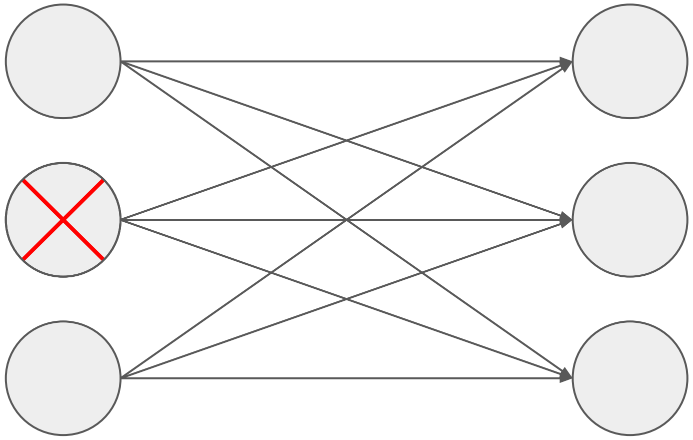
AutoPruner: a novel pruning strategy

- 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**



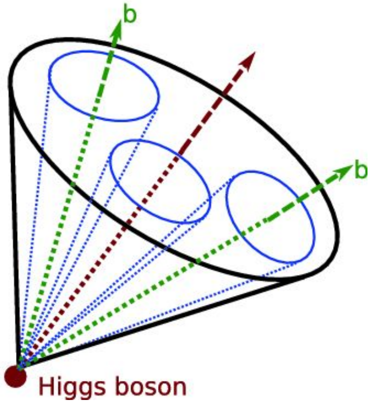
AutoPruner: a novel pruning strategy

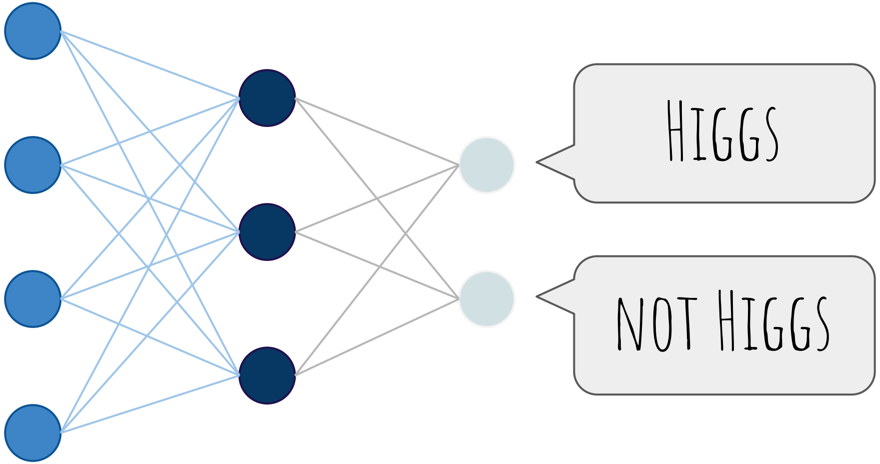
- it can prune **nodes**
- it prunes **during training**
- the number of nodes to be pruned can be determined by the **user**
- it can determine the most suitable **network architecture**



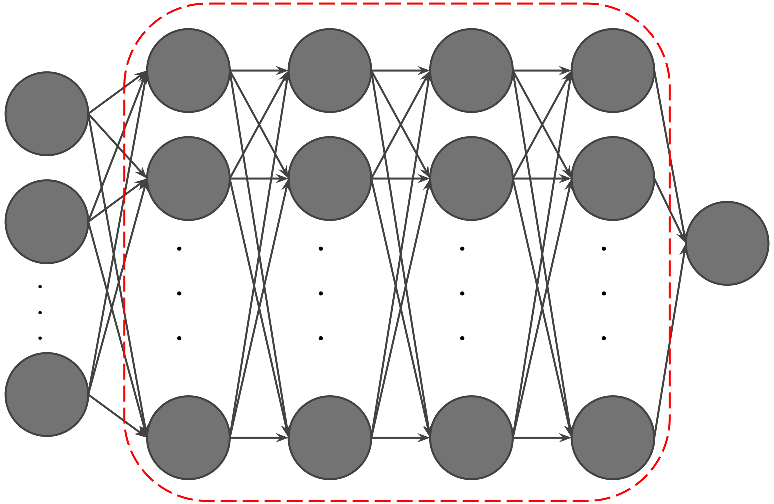
Use case

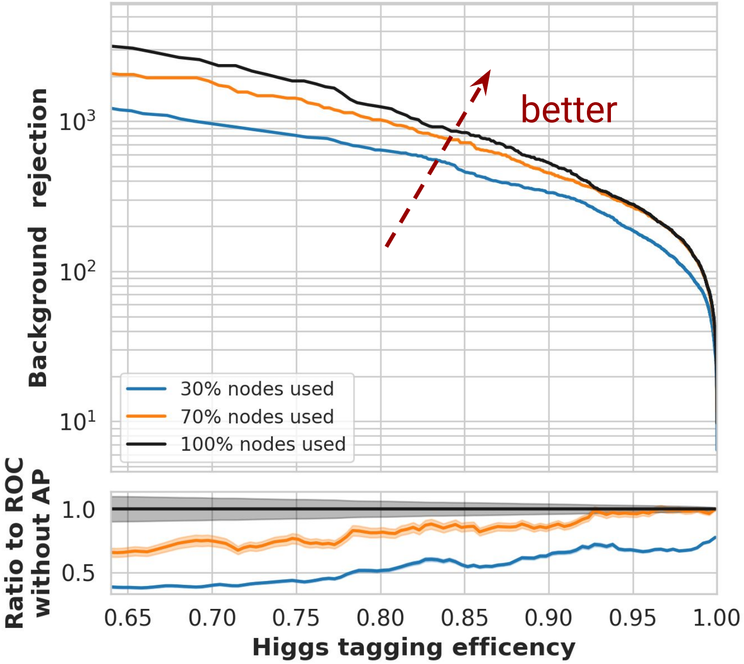
Identify jets that contain both the *b* quarks from boosted Higgs decay in *pp* collision experiments using Deep Neural Networks





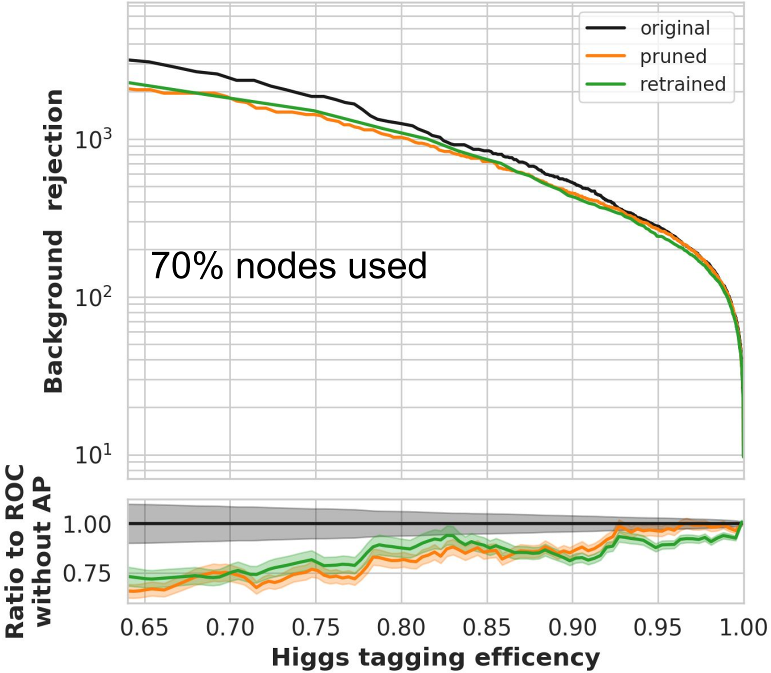
Results





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

Results



After finding the **optimal network layout** with AutoPruner, the reduced network can be retrained as a new independent model, with **performance compatible** with the pruned one within the uncertainties.

➔ The performance of the pruned networks reflects the performance of the reduced networks to be implemented on FPGAs.

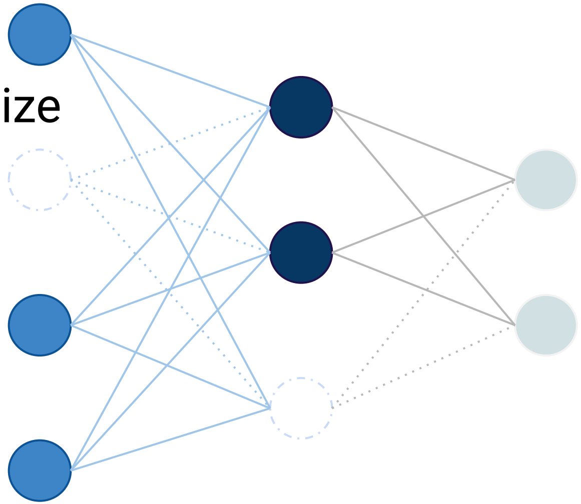
Conclusions

AutoPruner proved to be:

- **simple** to incorporate
- **effective** and **successful** in reducing the networks’ size
- **fast** (pruning during training, no need to fine tune)
- very **understandable**

Further developments are focusing on:

- quantify stability against initial conditions
- characterize optimality



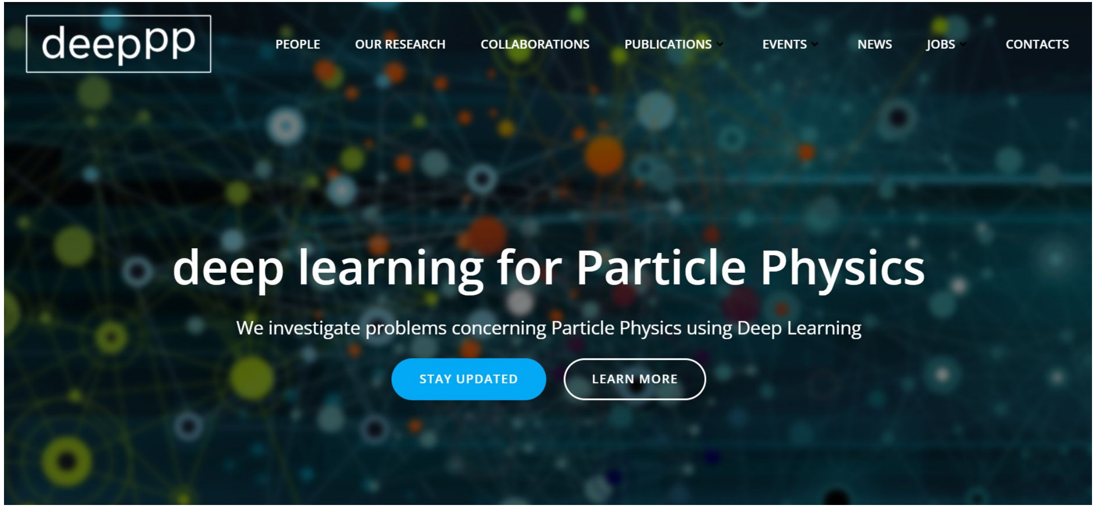
Thanks!

Want to know more about Deep Learning applications in Particle Physics?

Awesome!

Visit https://www.deeppp.eu/

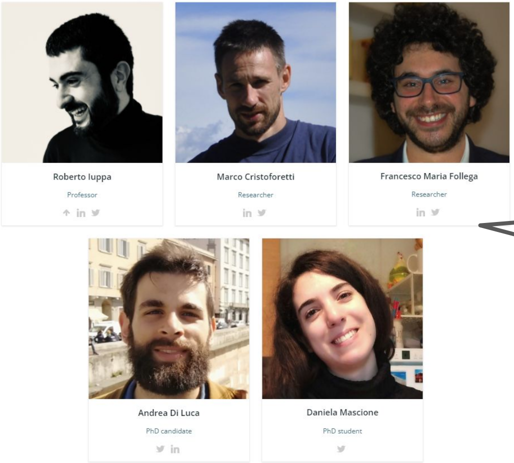
deepPP



deep learning for Particle Physics

We investigate problems concerning Particle Physics using Deep Learning

Our work



Roberto Iuppa
Professor

Marco Cristoforetti
Researcher

Francesco Maria Follega
Researcher

Andrea Di Luca
PhD candidate

Daniela Mascione
PhD student

Us

Who we work with

TIFPA

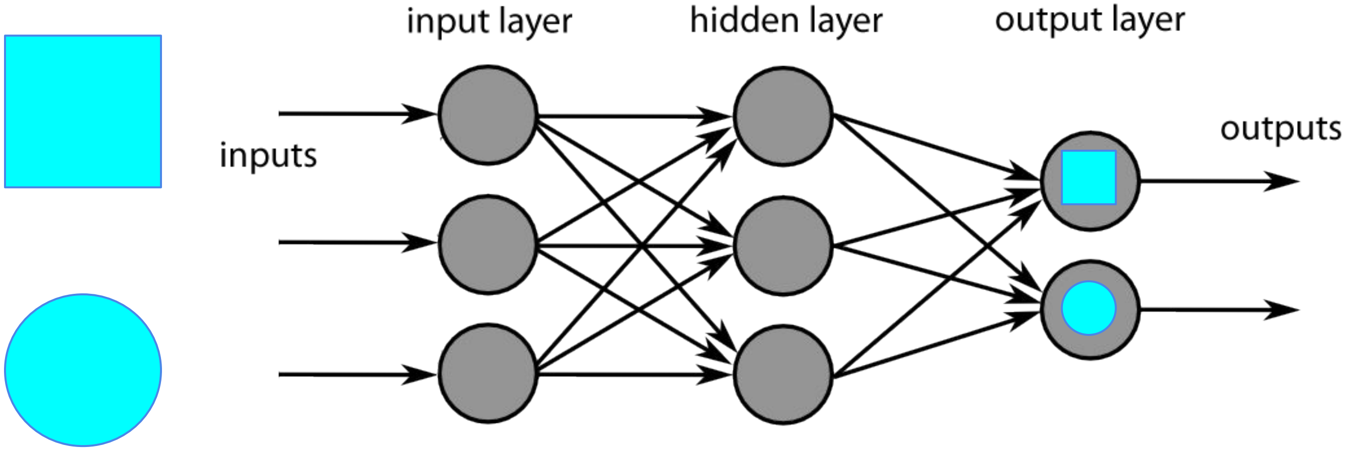
Trento Institute for Fundamental Physics and Applications

ATLAS EXPERIMENT

ATLAS Flavour Tagging Working Group

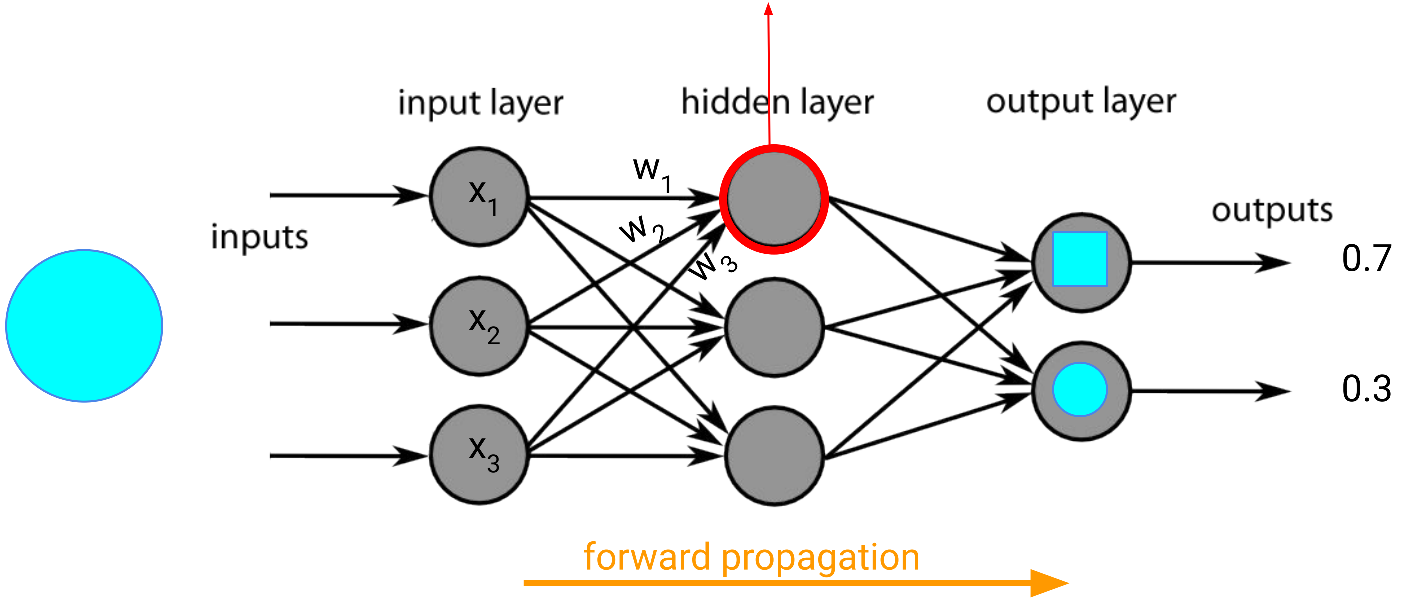
Backup

Simple neural network: an example



Example

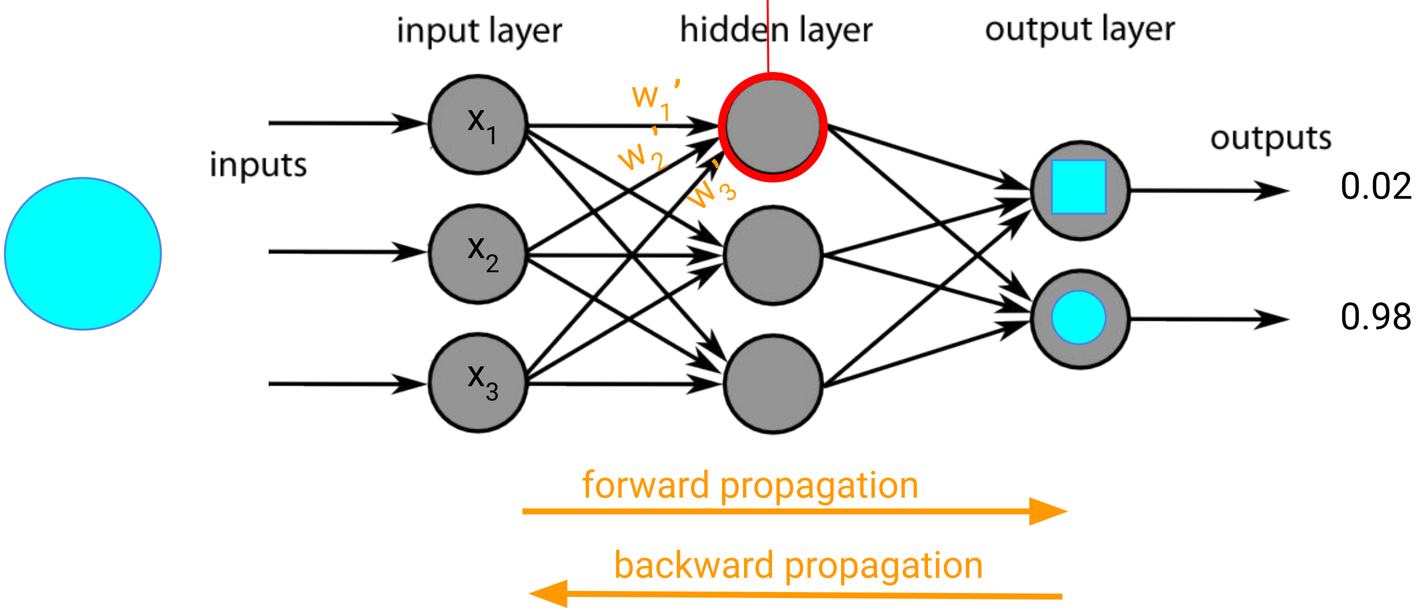
$(x_1 * w_1 + x_2 * w_2 + x_3 * w_3) + b_1$ ⇒ activation function



actual output
0
1

Example

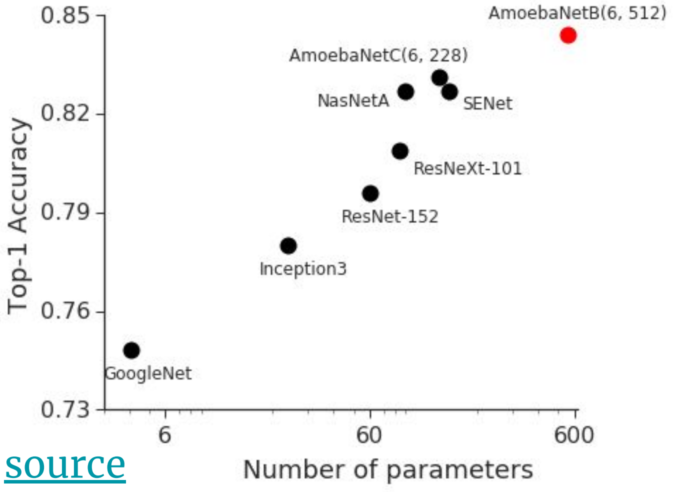
$(x_1 * w_1' + x_2 * w_2' + x_3 * w_3') + b_1$ → activation function



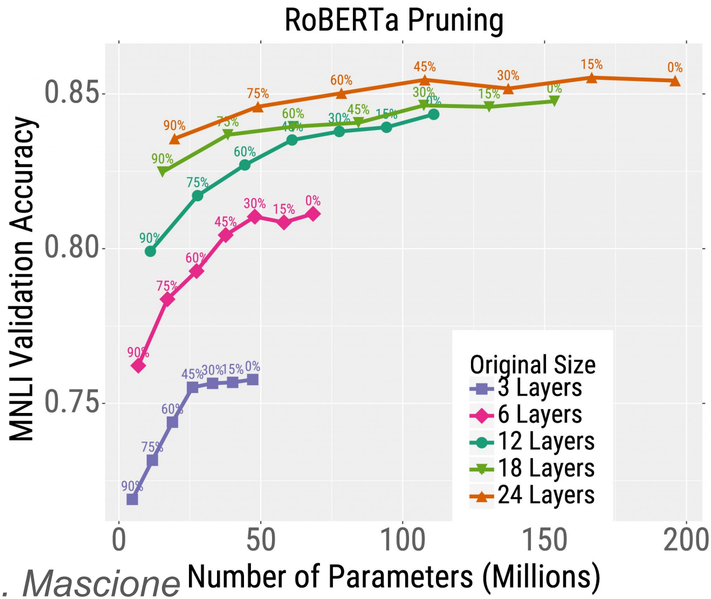
actual output
0
1

Why pruning?

Bigger networks are usually more **accurate**



source



source

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

D. Mascione

Pruning for applied research

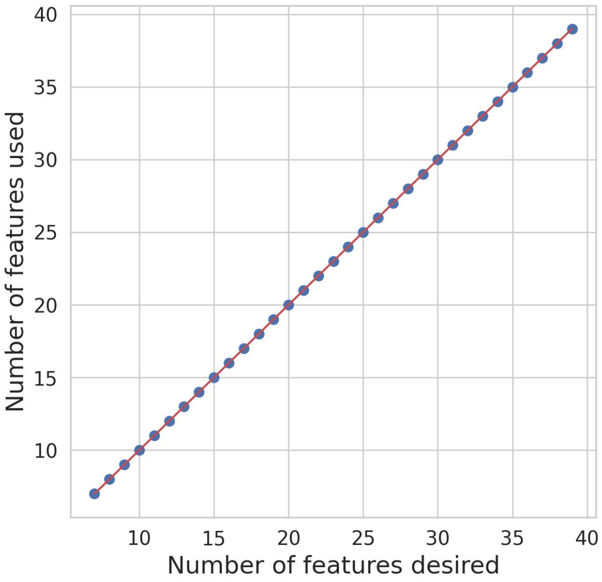
Relevance to the outside world:

- Reduction in storage requirements
- Private on-device computation (mobile, VR, IoT)
- Power savings
- Reduced heat dissipation in wearable devices
- Way to test neuron importance assumptions

Michela Paganini, *Neural Network Pruning: from over-parametrized to under-parametrized networks*, 4th IML Workshop, CERN

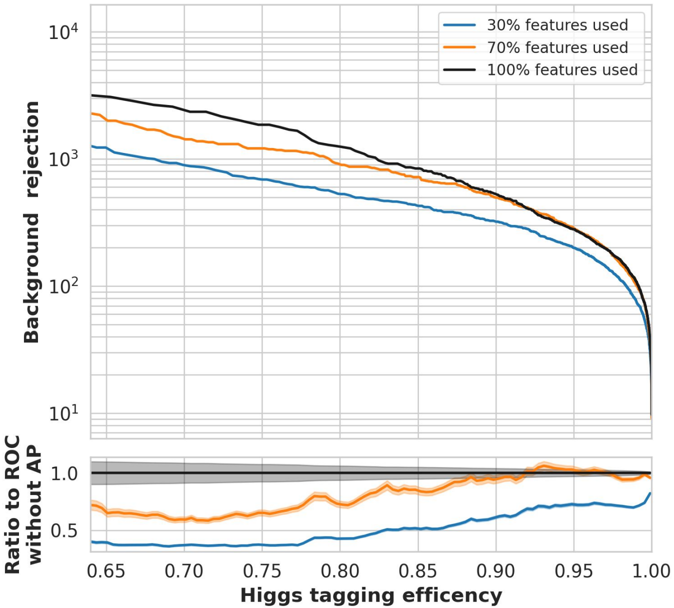
AutoPruner for feature selection

One AutoPruner layer following the input layer can be used also to **select relevant features**



The number of features used is equal to the required number

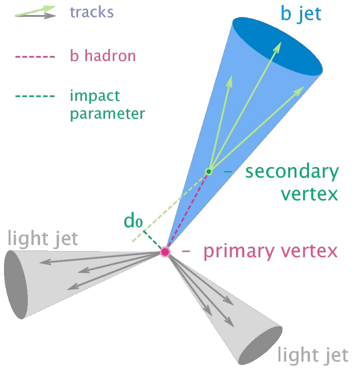
The performance increases with the number of features used



D. Mascione

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**