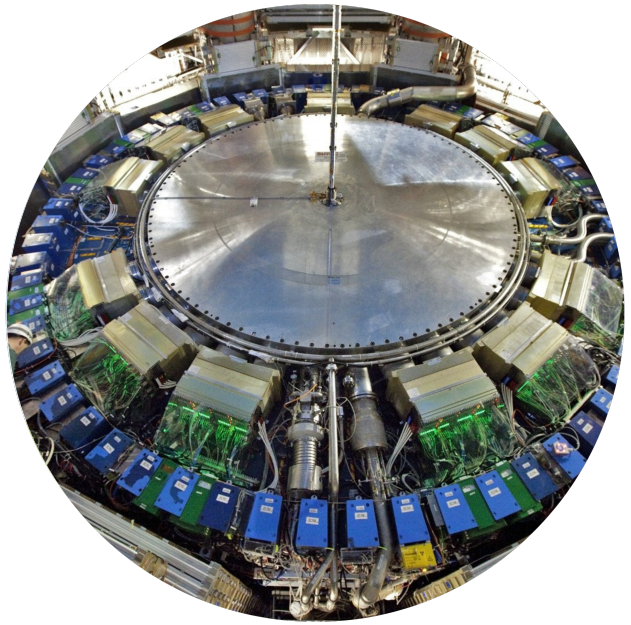
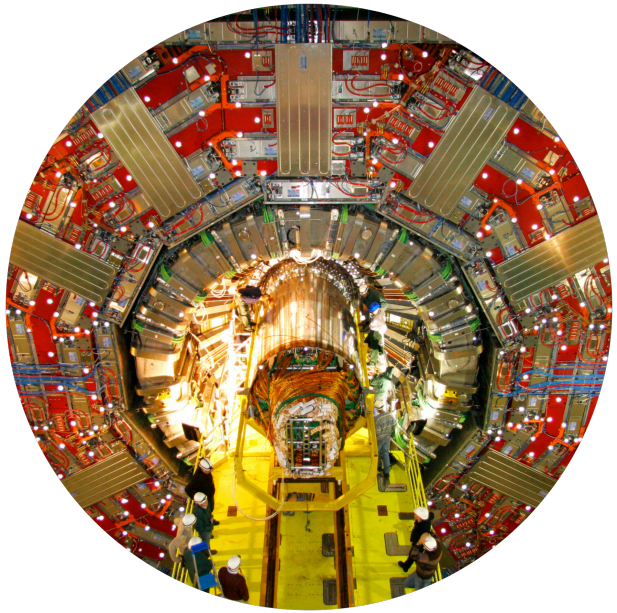




Machine learning for top quark physics at the edge in LHC pp collisions with ATLAS and CMS

Clara Nellist (ATLAS)

Radboud University and NIKHEF

On behalf of ATLAS and CMS

With input from Nicholas Tonon (CMS)

TOP2021

17th September 2021

Nikhef

Radboud University

Machine learning for top quark physics at the edge in LHC pp collisions with ATLAS and CMS

Overview

Purple flag = Example of how ML is used outside of HEP

Reference: A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow.

What is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

-- Arthur Samuel, 1959

Great for:

- Problems which require a lot of fine-tuning.
- Complex problems where traditional approaches don't yield a solution.
- Situations where simulations don't match the data.
- Fluctuating environments.
  - Adapting to new input data.
- Gaining insights about a complex problem and large datasets.

Sound familiar?

Can be split into:

- Supervised vs unsupervised vs reinforcement learning.
- Online vs batch learning.
- Instance-based vs model-based.

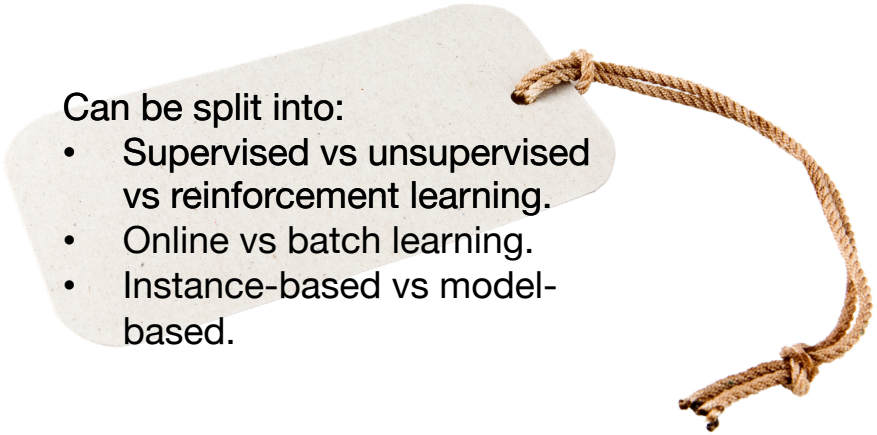
What is machine learning?

The basic concept:

Have a flexible algorithm

- **E.g. a Neural Network with weights**

In a learning phase, the weights or parameters are trained with an objective, which will depend on different learning methods.



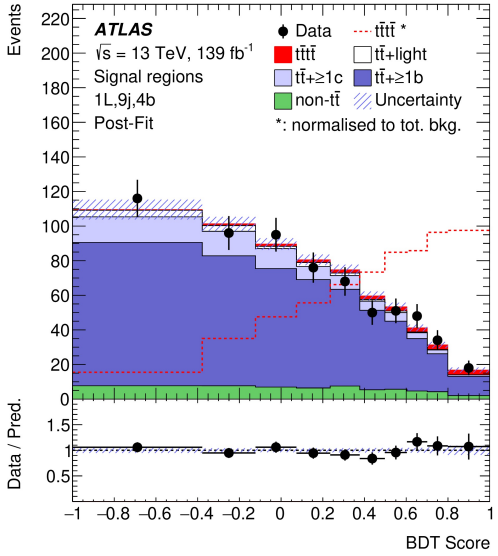
Can be split into:

- Supervised vs unsupervised vs reinforcement learning.
- Online vs batch learning.
- Instance-based vs model-based.

Reinforcement Learning

System observes, selects and performs actions and gets rewards or penalties for these actions.

Example: DeepMind's AlphaGo



Supervised

Training data contains labels indicating desired output.

Example: classification using a BDT (familiar in HEP).

Unsupervised

Training data doesn't contain labels and the system attempts to learn without this input. It learns the underlying function of the data.

Example: clustering or anomaly detection.

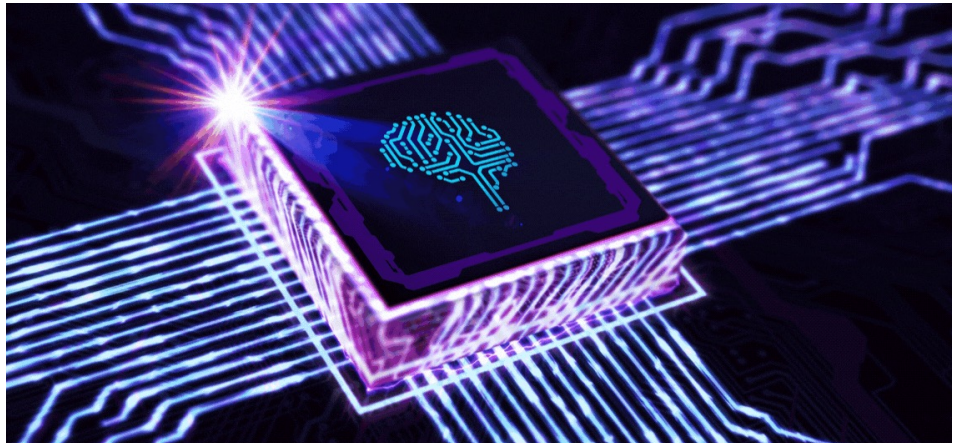
A history of machine learning in LHC particle physics
(in one slide)

There are 40 million collisions per second in the LHC. Most of these are not interesting.

Crucial to **distinguish interesting events** from overwhelming number of non-interesting events.

What makes machine learning in high-energy physics different from other typical cases?
- Quantum mechanical nature of particle production.
- Highly accurate simulation tools.

Allows us to exploit all of the data and increase the potential of our search



Been used in HEP for decades – it's older than the running of the LHC!

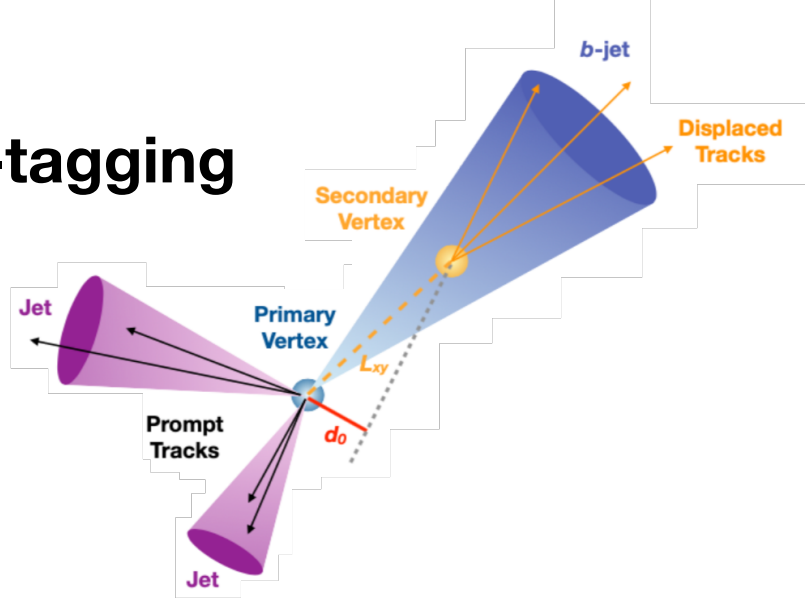
Previously, well-understood high-level information was given to a multivariate analysis for classification.

Move now to using low-level data with minimal processing into deep neural networks and other more modern ML algorithms.

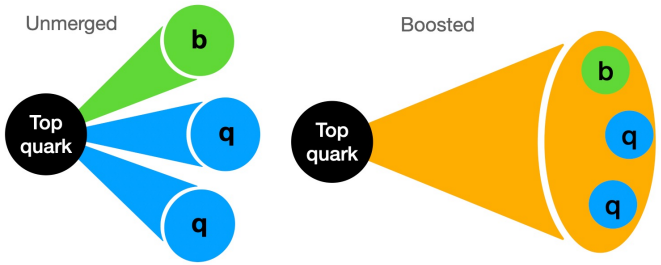
References: M. Schwartz, https://arxiv.org/pdf/2103.12226.pdf and D. Guest, K. Cranmer and D. Whiteson, https://arxiv.org/abs/1806.11484

Machine learning in top quark research

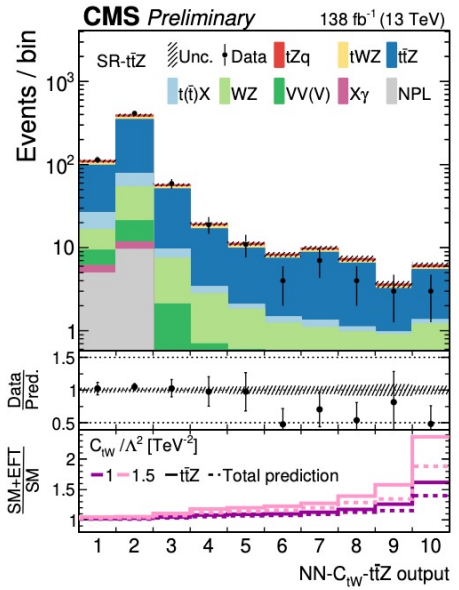
b-tagging



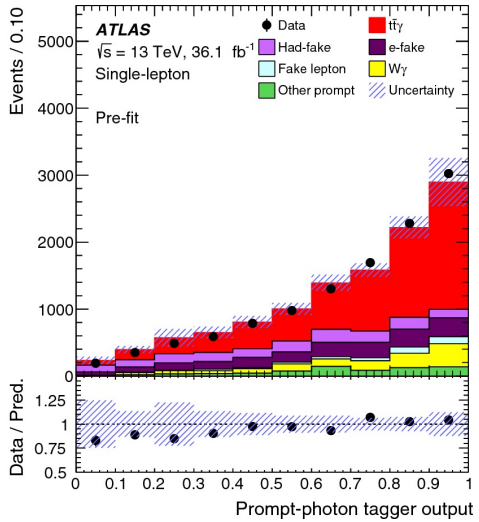
Top Reconstruction & Top Jet Tagging



Searches for new physics



Signal-to-background rejection methods

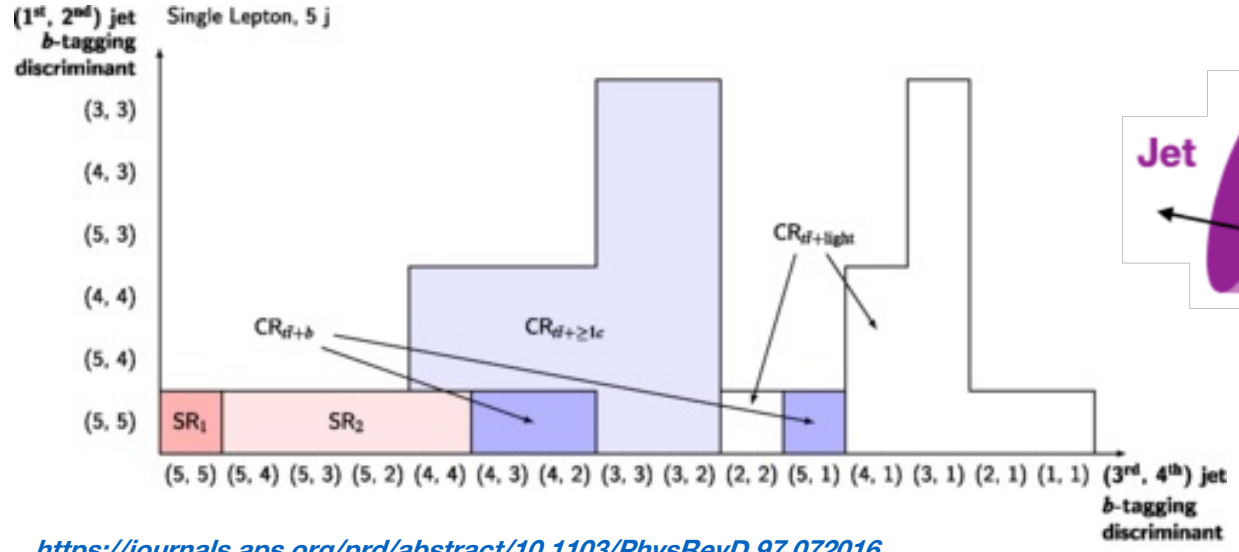


b-tagging

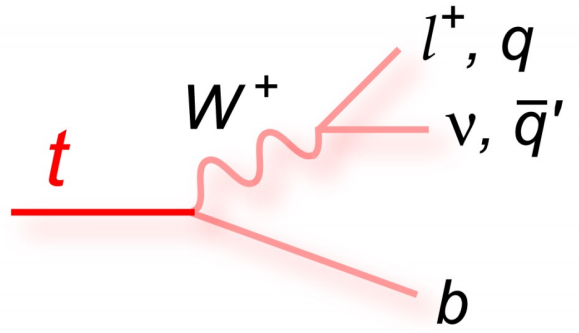
We know that b-tagging is an essential tool for top-quark measurements.

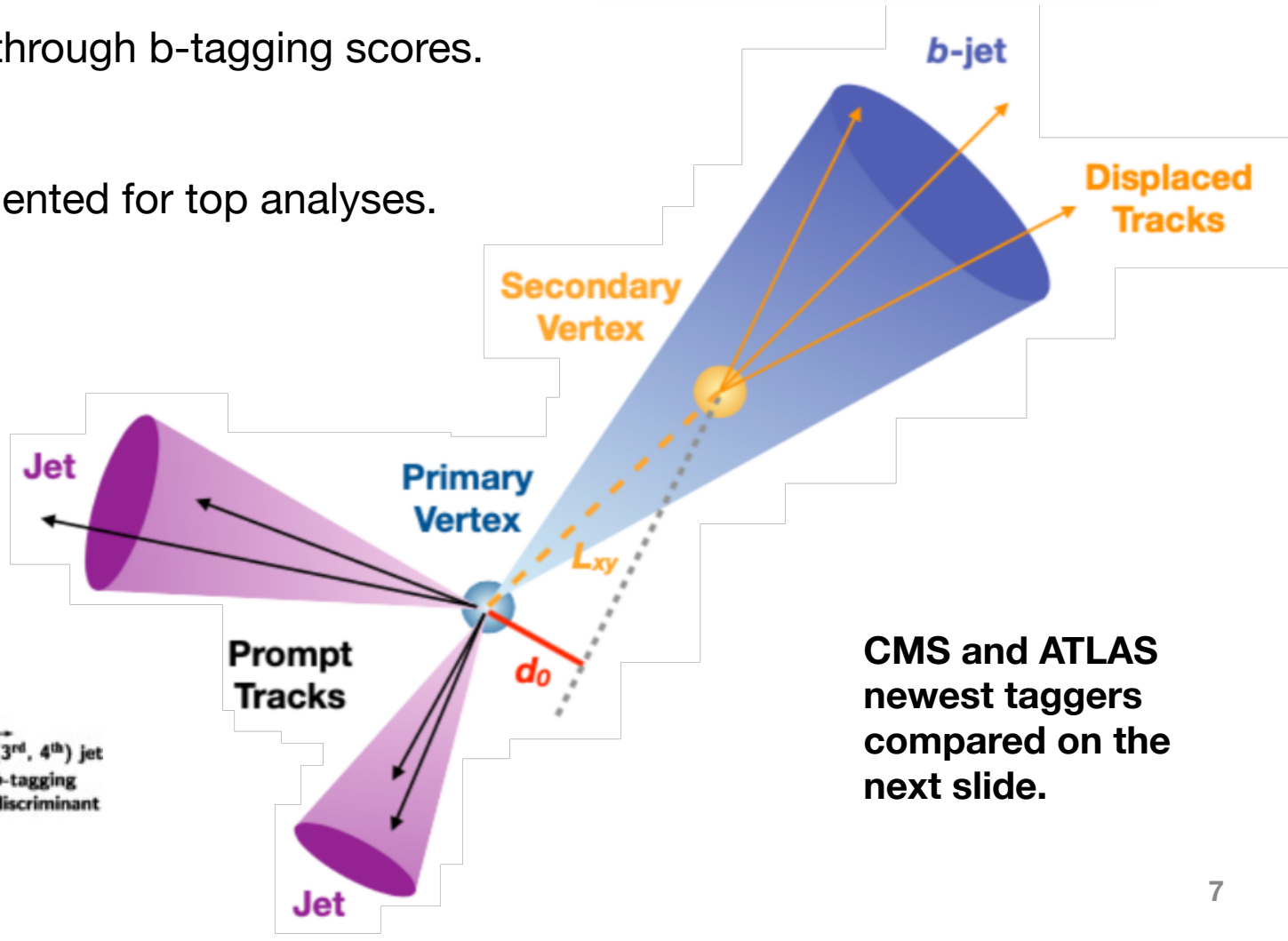
Many (if not all) top signal regions are defined through b-tagging scores.
Want to exploit the topology of the decay.

This is a key place where ML has been implemented for top analyses.



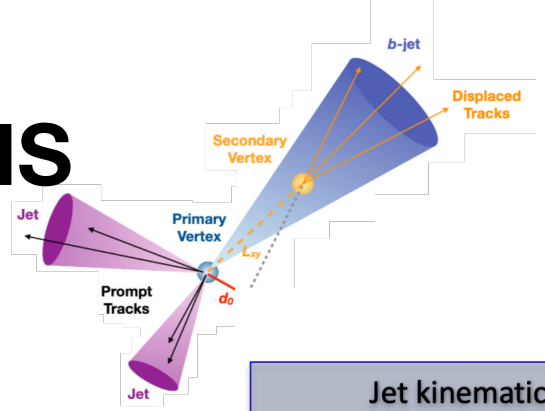
https://journals.aps.org/prd/abstract/10.1103/PhysRevD.97.072016





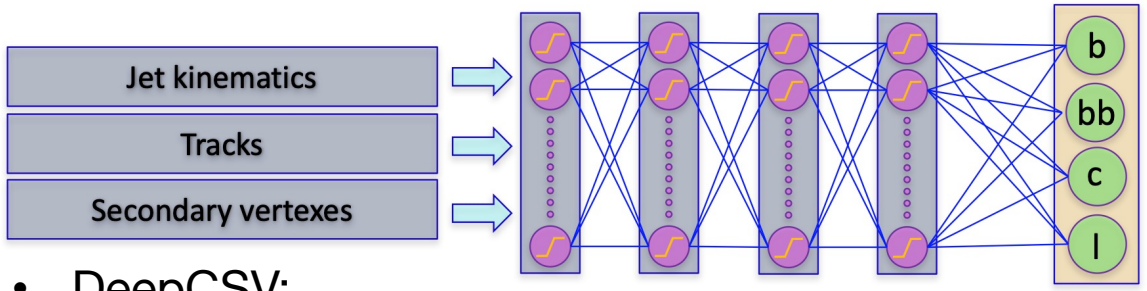
CMS and ATLAS newest taggers compared on the next slide.

b-tagging: ATLAS and CMS

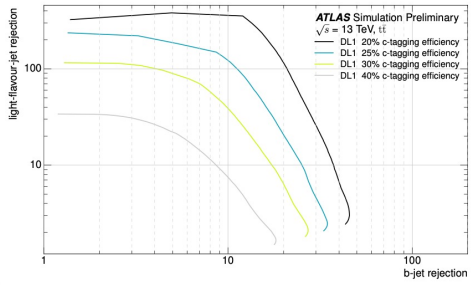


Recent CMS b-tagging includes:

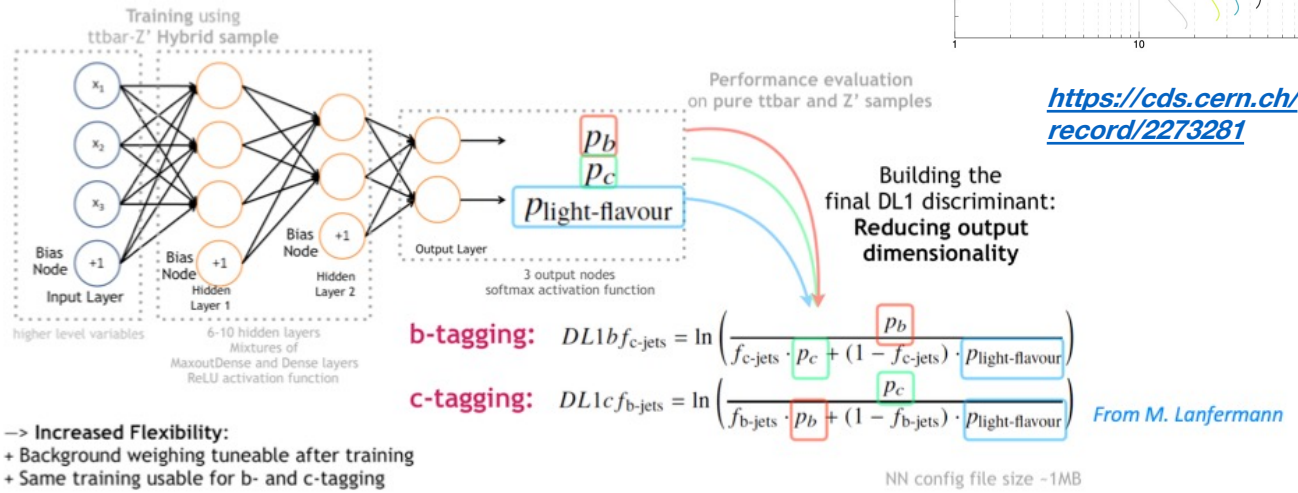
- DeepCSV and DeepFlavour (both DNNs)



- ATLAS has recently included a Deep Feed-Forward Neural Network for b-tagging in addition to a BDT algorithm.
  - Combine outputs from low-level tagging algorithms as input.
  - Probability assigned for jets to be b-,c-, or light-flavoured.
- DNN used other MVAs as input. Optimised:
  - the number of training epochs,
  - the learning rates,
  - and training batch size.
  - Includes batch normalisation.



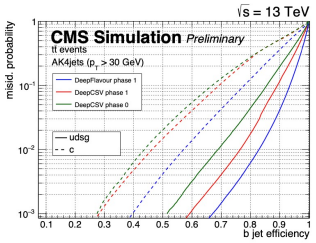
https://cds.cern.ch/record/2273281



–> Increased Flexibility:
+ Background weighing tuneable after training
+ Same training usable for b- and c-tagging

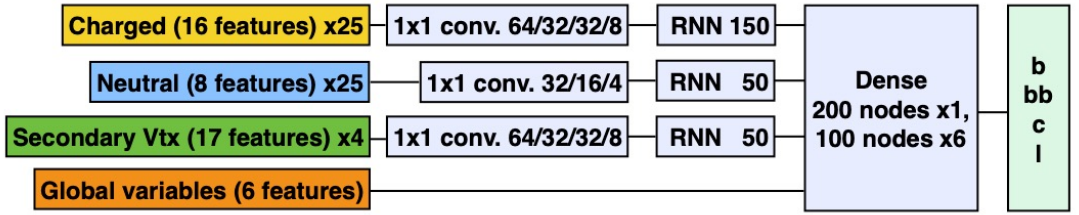
- DeepCSV:
  - 4 fully-connected layers with 100 nodes each
  - Multiclassification
  - KerasDL-library interfaced with Tensorflow

https://cds.cern.ch/record/2627468



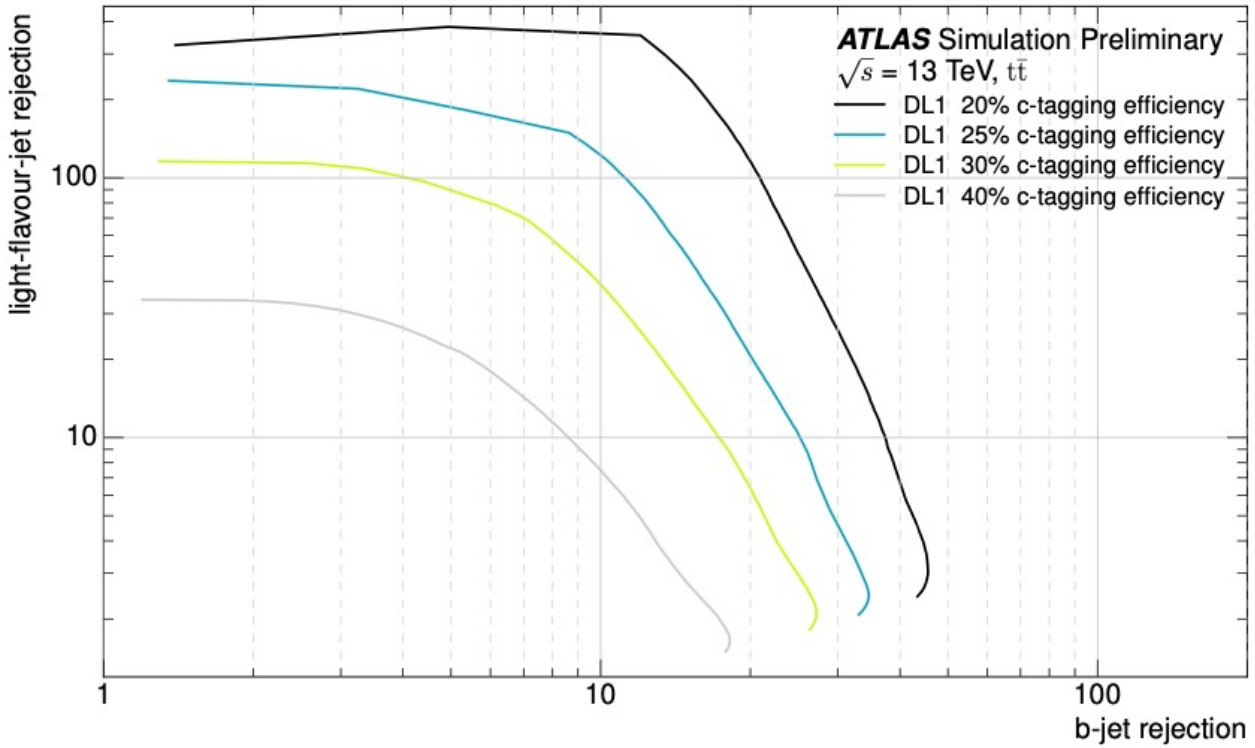
DeepFlavour:

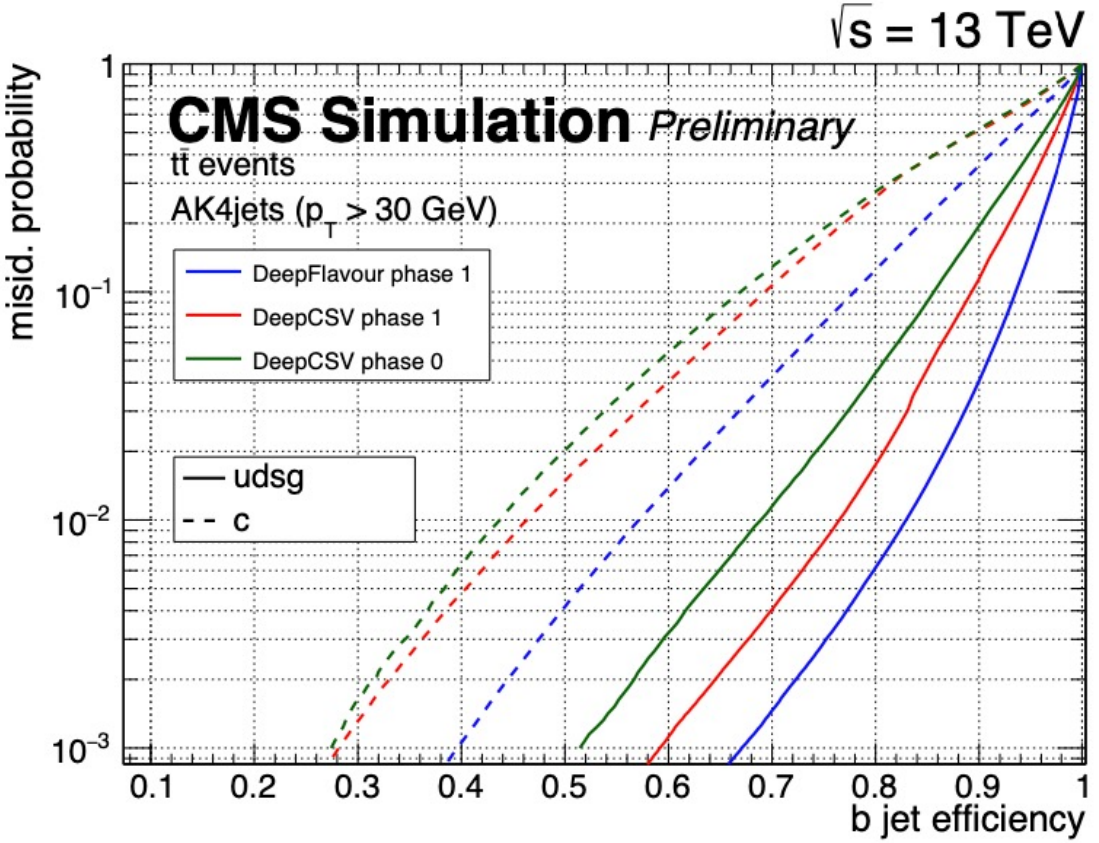
- Input: charged, neutral, secondary vertex and global variables.
- Separate 1x1 convolutional layers trained.
- Output into 3 LSTMs, then into single dense layer.



More info: https://indico.cern.ch/event/731450/contributions/3090101/attachments/1711227/2758918/Chicago_6_09_18_v2.pdf

b-tagging: ATLAS and CMS





https://cds.cern.ch/record/2273281

https://cds.cern.ch/record/2627468

More info: https://indico.cern.ch/event/731450/contributions/3090101/attachments/1711227/2758918/Chicago_6_09_18_v2.pdf

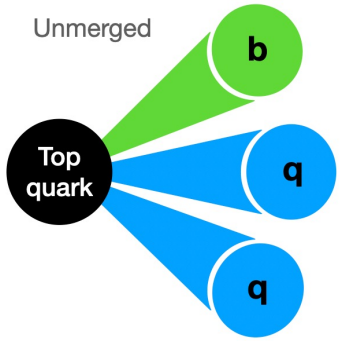
Top Jet tagging and reconstruction

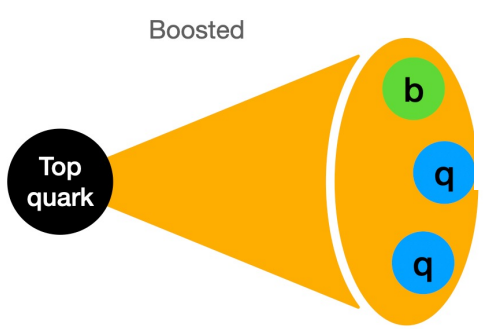
Aim: tagging algorithms to identify boosted jets originating from top quark decays

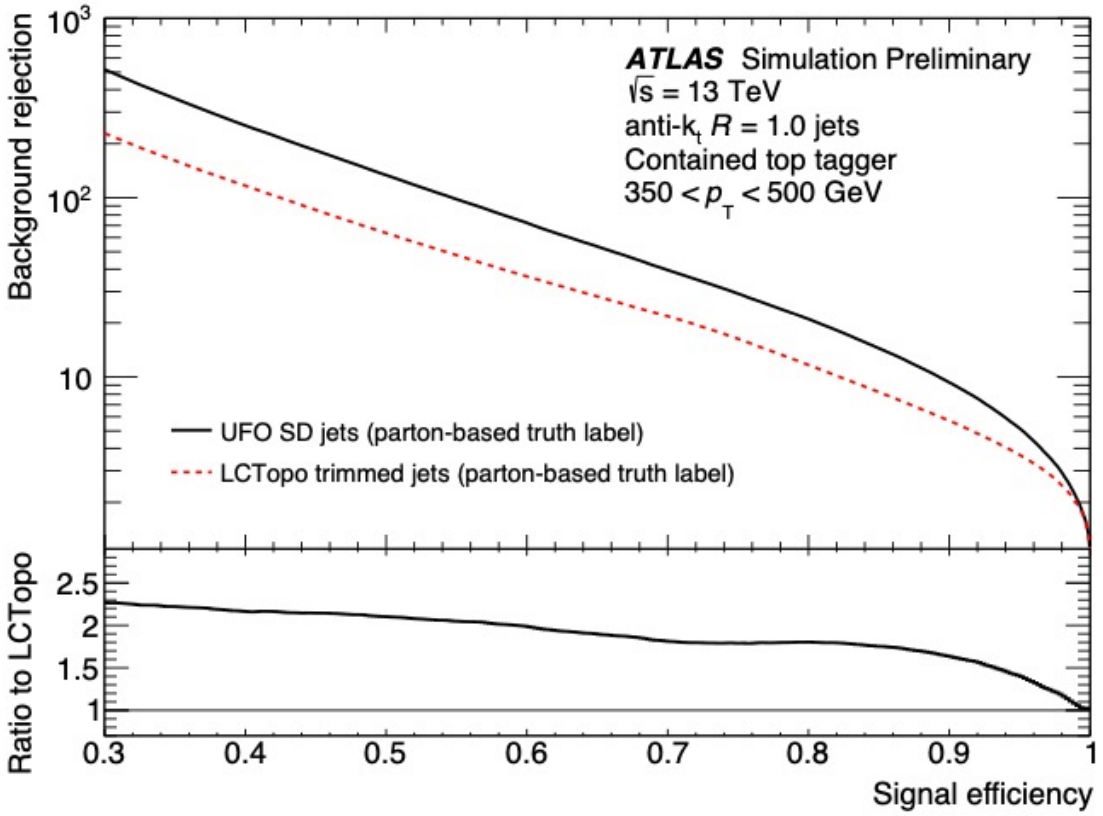
Identification of hadronically-decaying top quarks using UFO jets with ATLAS in Run 2

Two taggers based on deep neural-network (DNN) using hadronic jet properties as inputs, including various jet substructure variables, were optimised to identify jets.

- **Performance of top taggers is improved** especially for 50% WP compared to previous top taggers developed for jets reconstructed solely from topoclusters.
- Inclusive top tagger, the multijet background rejection was improved.







https://cds.cern.ch/record/2776782

Boosted taggers

How it's used outside of HEP: Food recommendation

Aim: tagging algorithms to identify boosted jets
Challenge: Particles have different sizes

Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques

A new algorithm based on ParticleNet, a graph neural network using an unordered set of jet constituent particles as the input.
Shows significantly improved performance.
Two new methods were investigated:

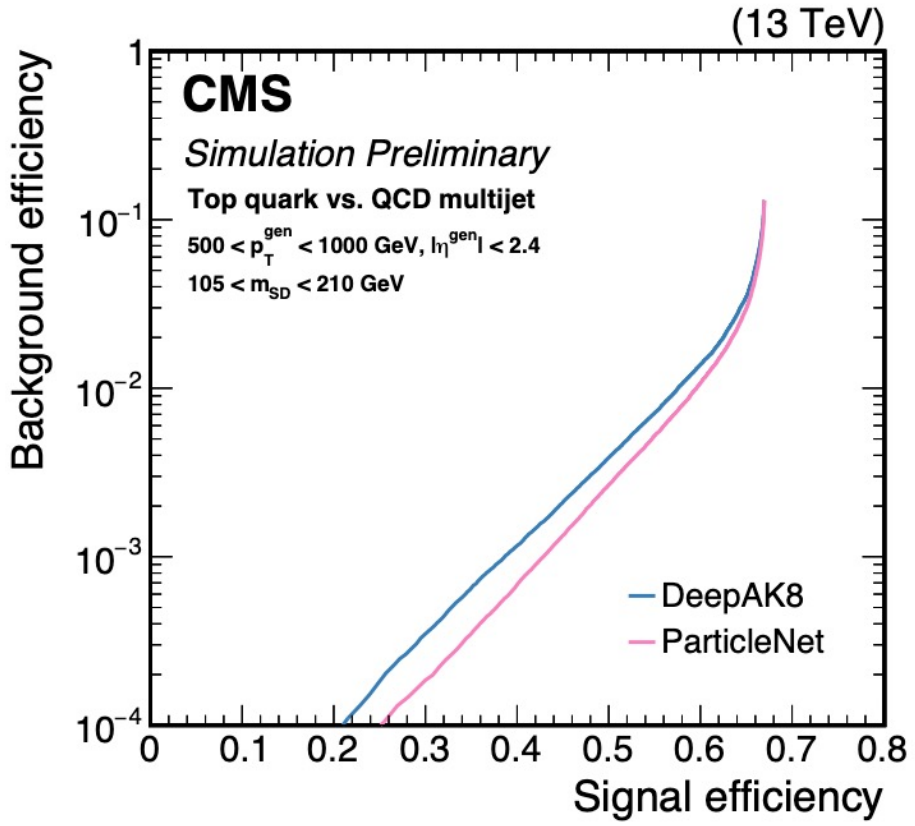
1. Designing Decorrelated Taggers (DDT) approach.
2. Training on an artificial signal sample generated with a flat mass spectrum for the signal particle.

Methods were compared to the existing (adversarial training based) one, and show better performance.

https://cds.cern.ch/record/2707946/files/DP2020_002.pdf

https://arxiv.org/abs/2007.13681

A graph network:
A graph has a set of vertices and a set of edges (the relationship between the vertices).



Signal-to-Background rejection

One of the more common areas to see machine learning applied in top analyses.

Measurements of inclusive and differential fiducial cross-sections of $tt\bar{}\gamma$ production in leptonic final states at $s\sqrt{}$=13 TeV in ATLAS

Challenge: discriminate the $tt\gamma$ signal from backgrounds

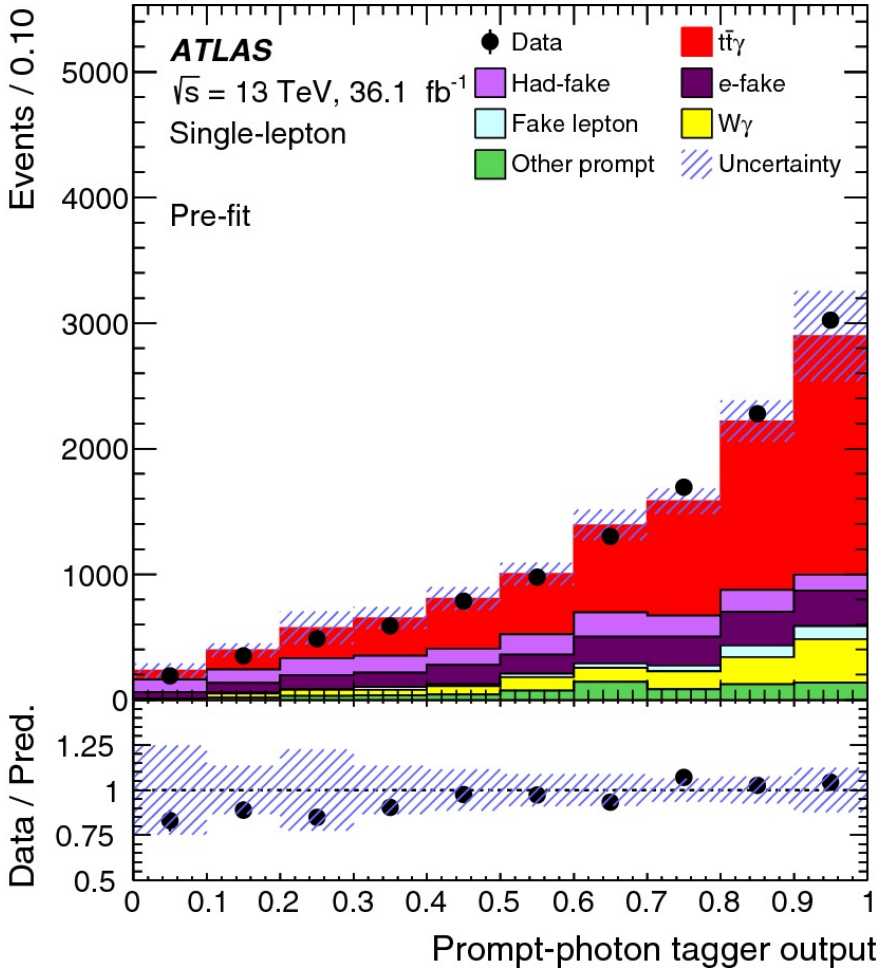
Two Neural Network algorithms - feedforward binary classifiers.

- Event-level discriminator (ELD), trained to discriminate the signals.
- Prompt-photon tagger (PPT), was trained to discriminate prompt photons & hadronic-fake photons.
  - The PPT was used as input to the ELD.

Training for both:

- Normalised the input variables to have standard deviation 0->1.
- Reduced over training with dropout and batch normalisation.
- K-fold cross-validation was used.

- First use of a NN for photon identification/fake photon rejection.



Signal-to-Background rejection

Classification without labels (CWoLa) – a weakly supervised approach.

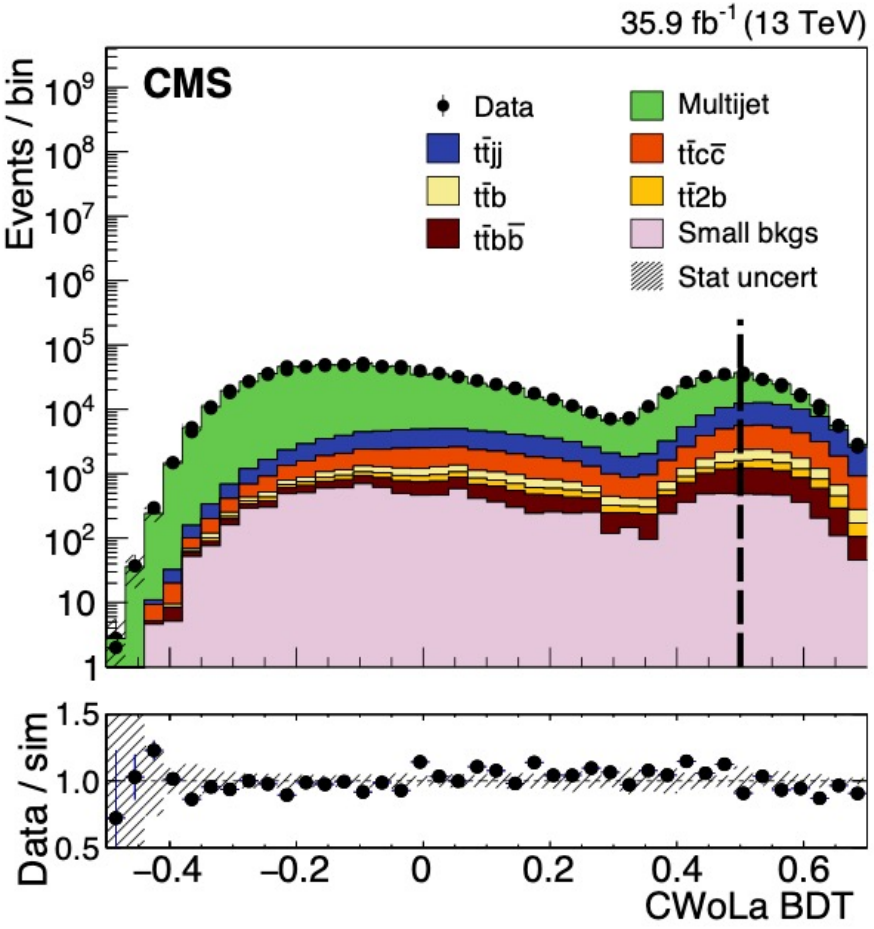
Measurement of the ttbb production cross section in the all-jet final state in pp collisions at $\sqrt{s}$ = 13 TeV

A combination of MVA techniques used to:

- reduce large background from multijet events not containing a top quark pair
- help discriminate between jets originating from top quark decays and other additional jets.

- Used CWoLa to mitigate the poor modelling of multijet production & insufficient size of available simulated samples.
  - Classifier trained on data split into two regions.
  - Two conditions required:
  1. relative rates of actual signal and background processes should be different in each region.
  2. distributions of the variables entering CWoLa classifier should be independent of the quantity used to define the two regions, for both the signal and background processes.

Performance found to be comparable to that of a supervised classifier trained using simulated samples.



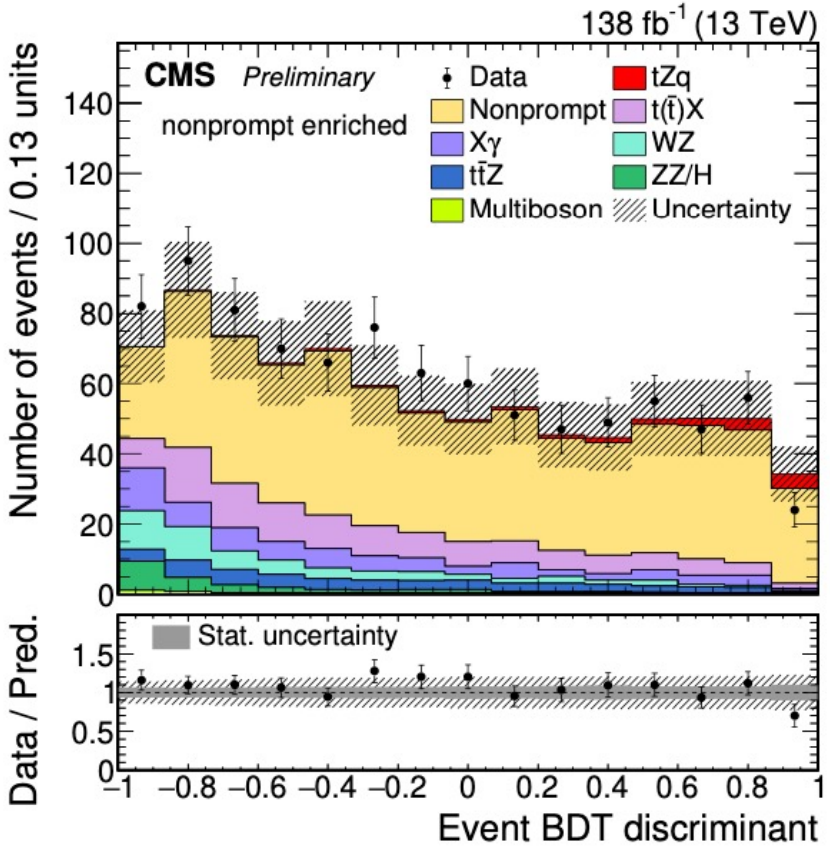
Distinguish objects to reduce systematics

Example: want to distinguish prompt from non-prompt leptons.
Challenge: minimise systematics

Variables are combined into a multivariate-analysis (MVA) based discriminant.

- A Boosted Decision Tree algorithm is trained on large sample of:
  - simulated prompt leptons originating from tZq, ttZ, and ttW processes.
  - Non-prompt leptons taken from simulated tt events.
- Requirement on the lepton MVA value corresponds to selection efficiencies of **prompt (95%)** and non-prompt (2%) leptons.
- Leptons that pass are labelled as tight.
- Led to improvements with respect to the previous result.

https://cds.cern.ch/record/2771809/files/TOP-20-010-pas.pdf

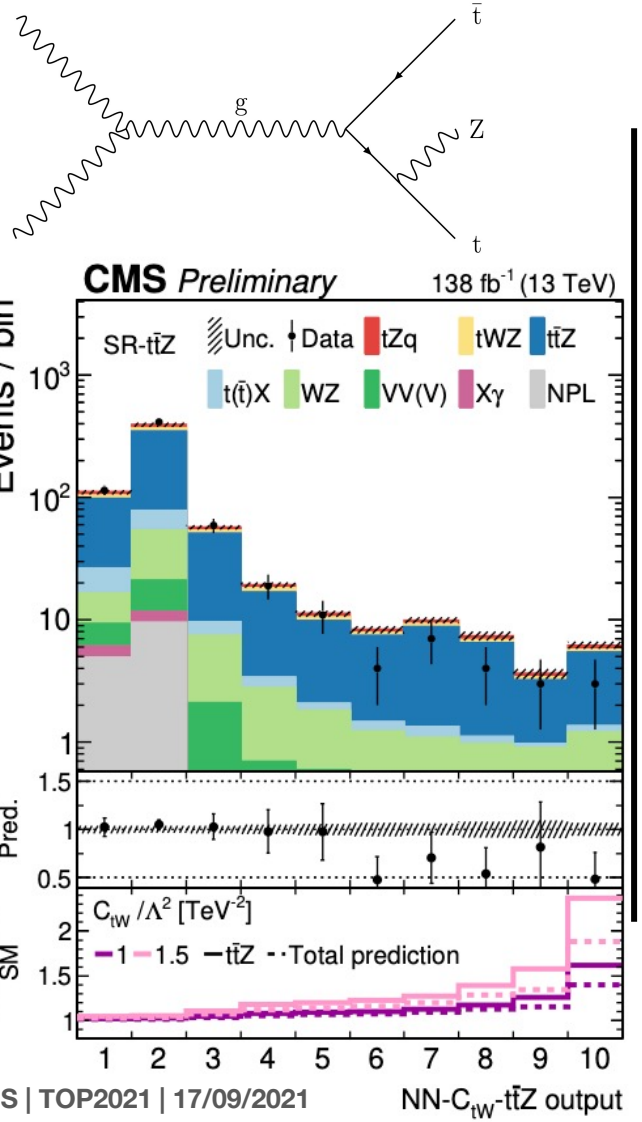


Searches for new physics

Probing effective field theory operators

https://cds.cern.ch/record/2771677/files/TOP-21-001-pas.pdf

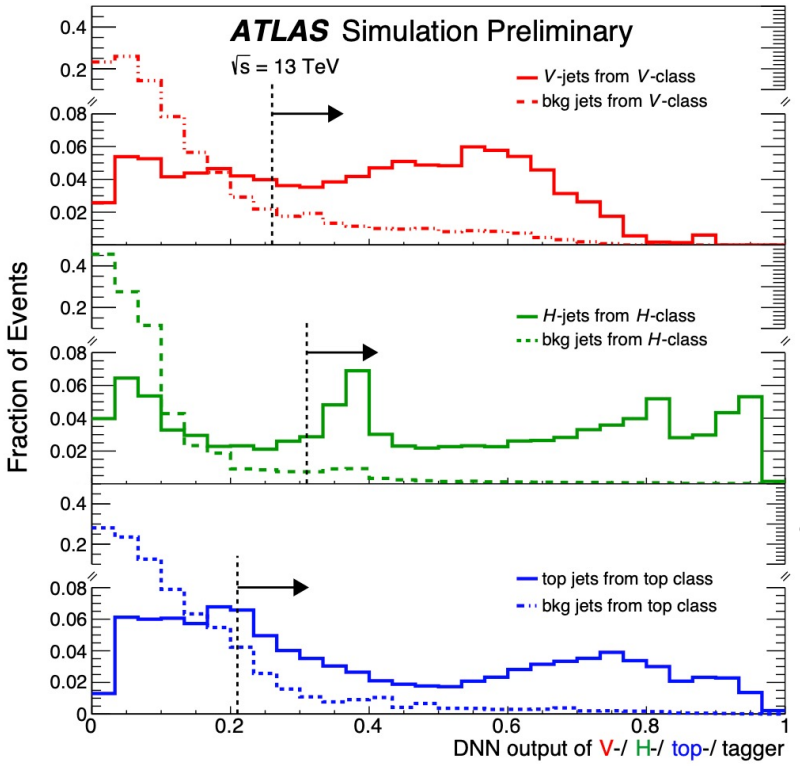
- Direct EFT measurement targeting ttZ+tZ in multilepton final states.
- ML used extensively:
  1. isolate the "signal processes"
     - 3 hidden layers, each with 100 rectified linear units, and 3 output nodes: “tZq “, ”ttZ”, and “Others”.
  2. Taught a Deep Neural Networks what EFT effects look like (binary output).
- Allowed tighter constraints to be set on EFT parameters.



Search for pair-production of vector-like quarks with > 1 leptonically-decaying Z boson and a 3rd-generation quark

https://cds.cern.ch/record/2773300

Used RC jets as an input to Multi-Class Boosted Object Tagger (MCBOT) to identify origin of each RC jet. Either: hadronically-decaying V boson, H boson, or top quark.



- MCBOT is based on a multi-class DNN using the Keras & TensorFlow.
  - Four label classes.
  - 18 input variables.
  - 4 fully-connected hidden layers and 4D output layer which define WPs.
- For the doublet configuration, the excluded T mass limits were extended by 90 GeV.

Anomaly Detection Methods

How it's used outside of HEP:
Fraud detection

What is anomaly detection?

"Finding patterns that do not conform to expected behaviour"

- This is a useful strategy to look for new physics without having to specify the model that new physics will come from (**model independent**).
- Often training is **unsupervised**.
- BDTs are good for very specific regions, but with AD we are looking over a much larger phase space.
- Just as with a BDT – can define a signal region with an anomaly score.

Anomaly Detection example

How it's used outside of HEP:
Generation of art, or deep fakes.

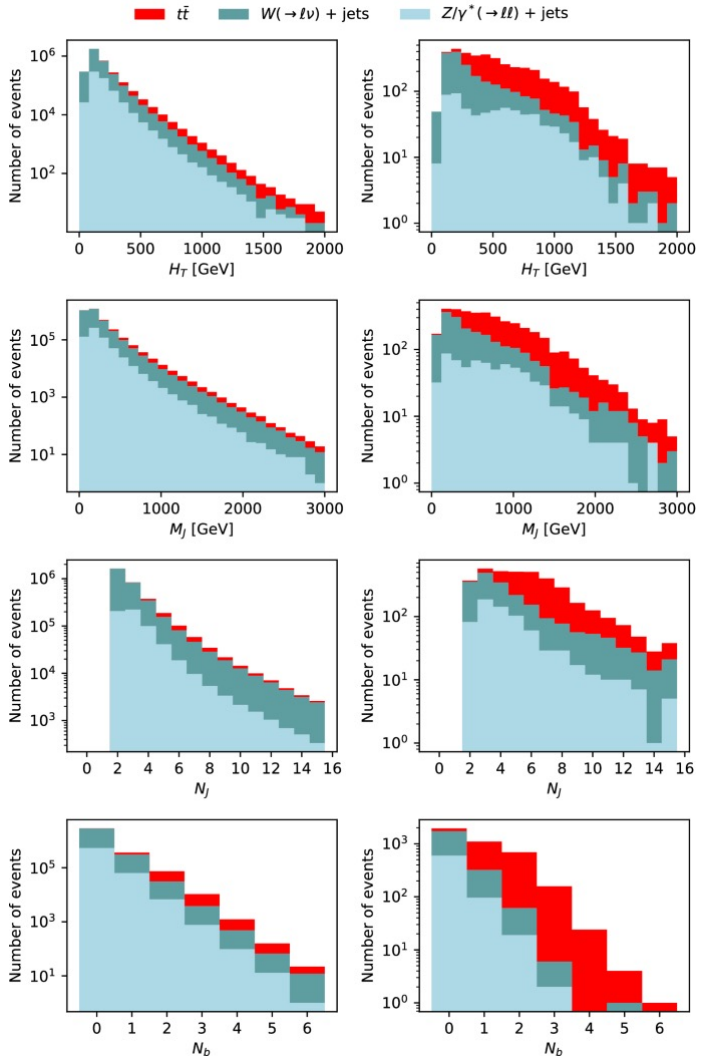
Reference: https://doi.org/10.1140/epjp/s13360-021-01109-4

Adversarially Learned Anomaly Detection on CMS open data to rediscover the top quark

Trained an Adversarially Learned Anomaly Detection (ALAD) algorithm to search for 'new physics' using 4.4 fb^{-1} of 8 TeV CMS Open data.

- Algorithm is a type of Generative Adversarial Network (GAN) where two neural networks compete against each other during the training phase.
- Senario: LHC collision data was available, but there was **no knowledge of the top quark.**
  - ~4% of the dataset after selection requirements.
- Using anomaly score, defined a post-selection procedure that produced an almost pure 'anomalous' sample.
- **First use of a data-driven anomaly detection process on LHC data.**
- **Shows that rare events can be searched for with this method.**

Next steps: Implement this in a search?



Above: A generated image by StyleGAN based on input portraits.

Left: Event distribution before and after anomaly selection.

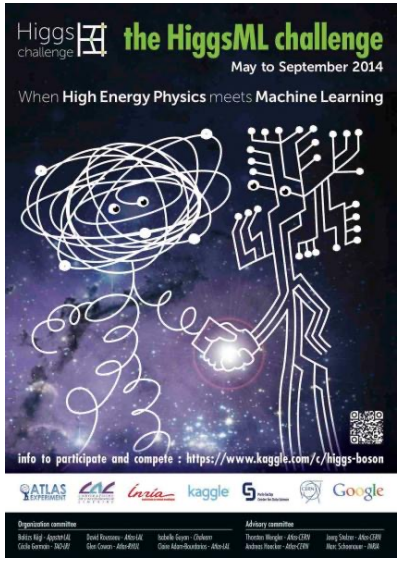
Challenge events

Follows the very popular Kaggle style of setting a challenge to inspire and innovate new ideas within a community,

There have been a number of challenge events, which are opportunities to test out new applications on a specially selected dataset and compare the results directly to each other.

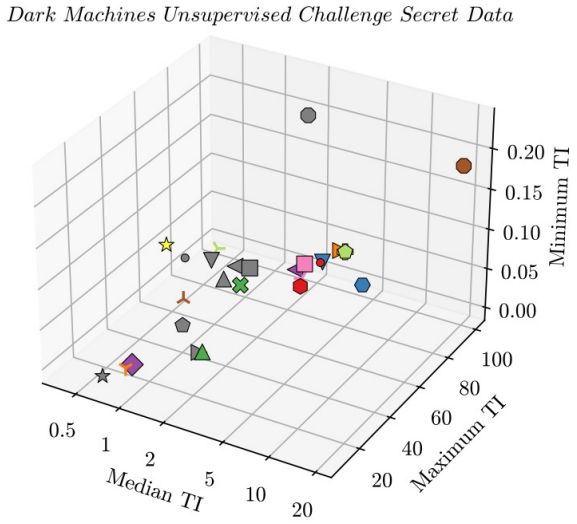
The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics: [hep-ph/2101.08320](hep-ph/2101.08320)

The ATLAS Higgs Machine Learning Challenge in 2015
High-scoring submissions we invited to a workshop at CERN to discuss their methods.

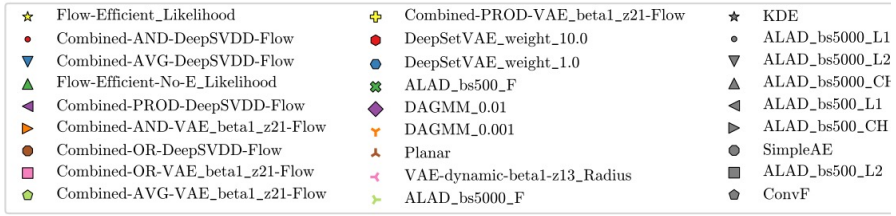


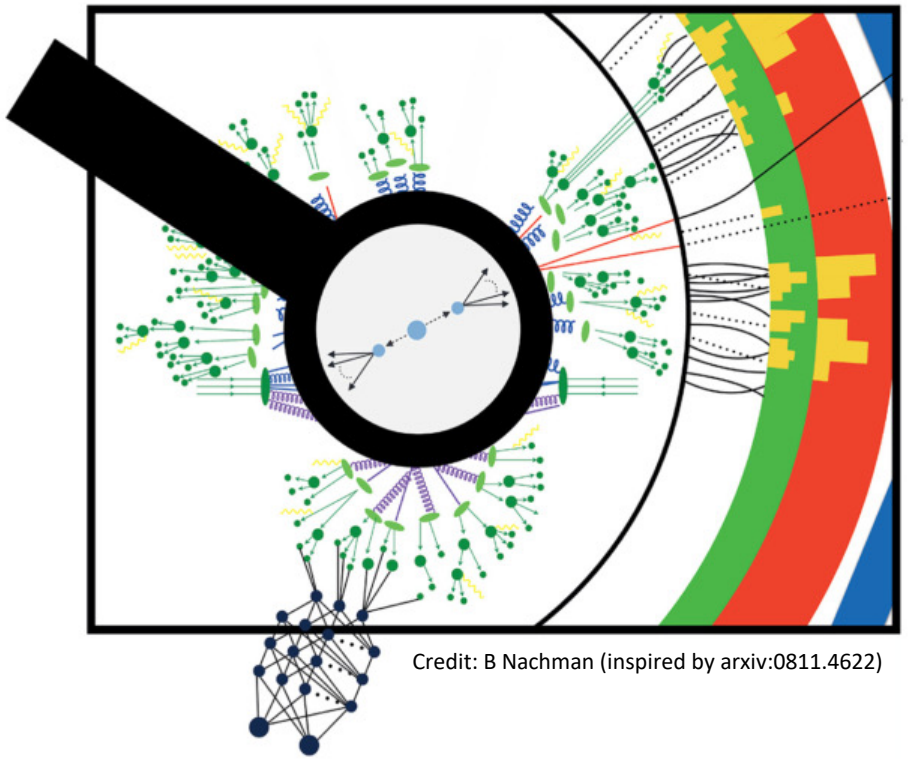
https://atlas.cern/updates/news/machine-learning-wins-higgs-challenge

The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider
https://arxiv.org/abs/2105.14027



The minimum, median, and maximum best total improvements for each technique applied on each of the signals in the secret dataset.





Credit: B Nachman (inspired by arxiv:0811.4622)

B. Nachman https://arxiv.org/pdf/1909.03081.pdf

Uncertainties when evaluating ML algorithms

It is often not immediately apparent how to determine the uncertainty of using a ML algorithm in an analysis.

Should decide as a community what we want to see and **determine community standards**.

Some additional things to think about:

- Simulated data might not fully replicate the real situation and ML algorithms could pick up on something that isn't actually there.
- No labels for real data.
  - Not a problem for unsupervised searches.
- **Active area of study**, and many things can be done to ensure proper training (overtraining checks, verify input modelling, ...)

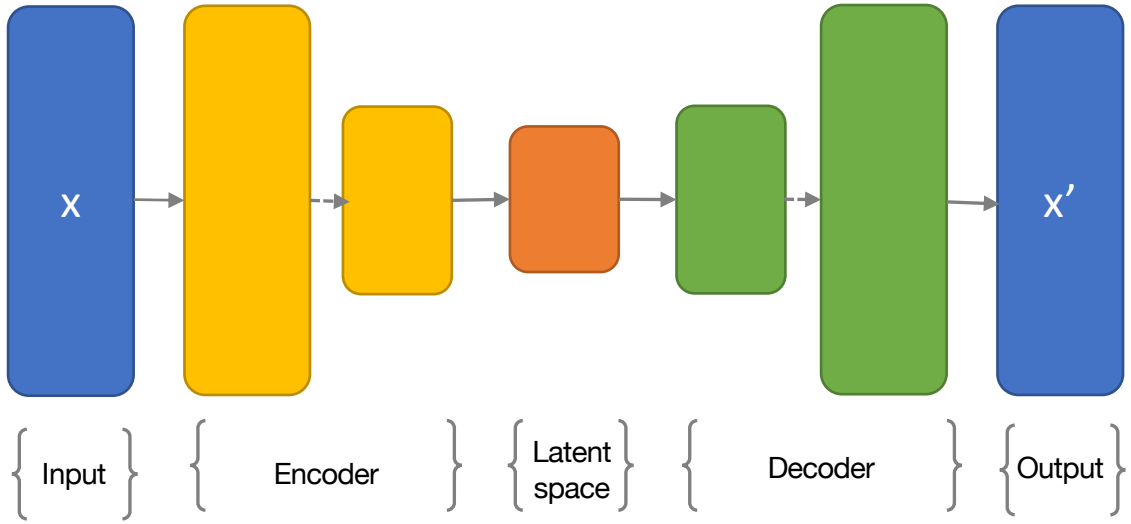
In high speed to finish within time (if I am still within time at this point...)

Variational Autoencoders for New Physics Mining

Reference: https://cerncourier.com/a/hunting-anomalies-with-an-ai-trigger/ and https://arxiv.org/pdf/1811.10276.pdf

Tl;dr : too long; didn't read

Tl;dr: Can we use anomaly detection methods in the trigger to search for new physics in a model independent and unsupervised way?

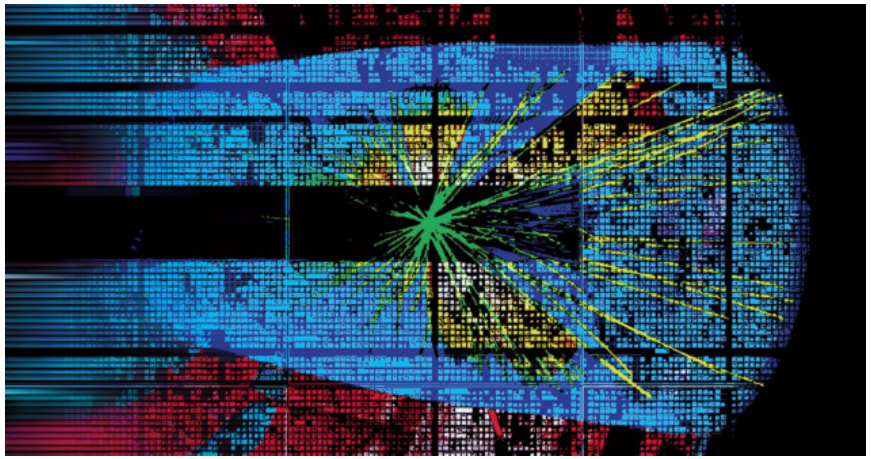


Using variational autoencoders trained on known physics to search for anomalous outlier events which are model-independent.

Paper suggests deploying an unsupervised algorithm in the online trigger to be stored in a special stream.

- Experts would evaluate the stream to determine if anomaly tag was for a detector-related reason.
- Event topologies repeating in this dataset could inspire new-physics model building and new experimental searches.

- **Challenge:** could these be run on FPGAs?



Credit: S Sioni/CMS-PHO-EVENTS-2021-004-2/M Rayner

Invertible Networks

Reference: https://arxiv.org/pdf/2006.06685.pdf

Typically, neural networks lose some information when proceeding through layers. Therefore they are not (generally) invertible.

Tl;dr: Can reversible networks help with unfolding?

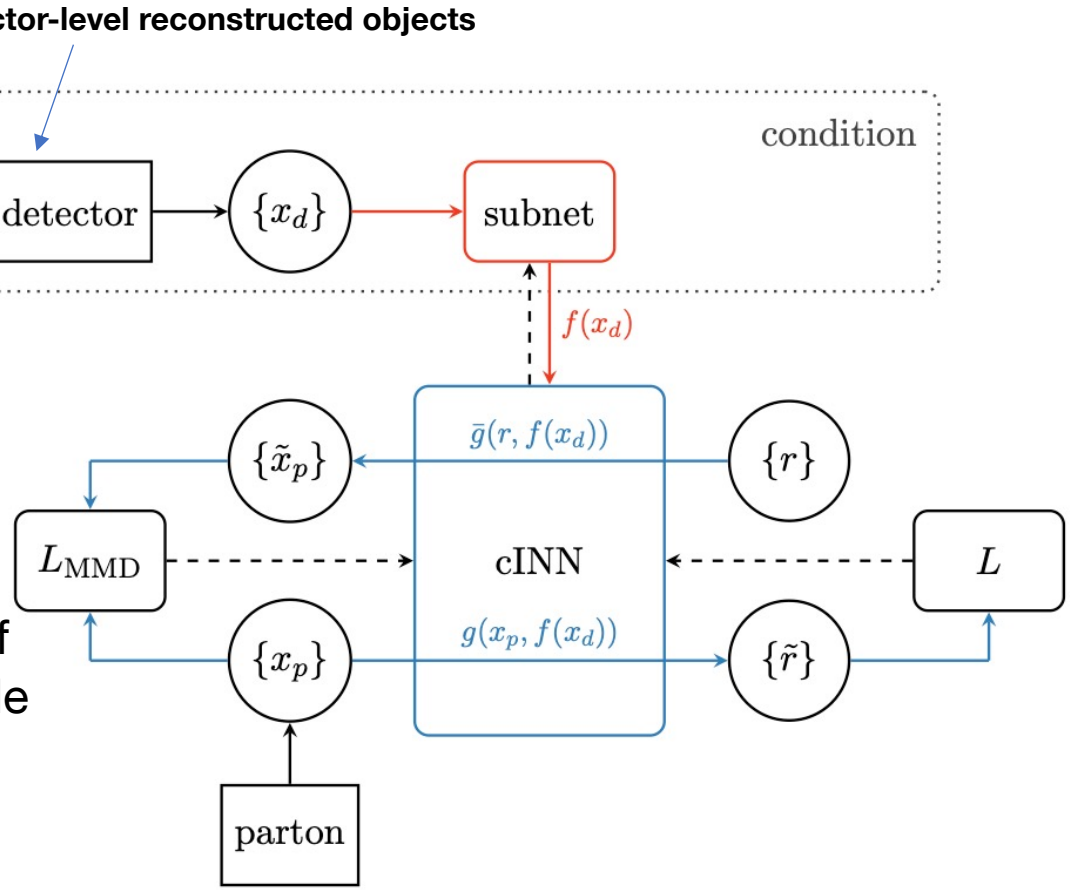
Idea: Invert the detector simulation chain in terms of high-level observables.

Requirements:

1. Mapping from input -> output is invertible.
2. Both directions can be evaluated efficiently.

- Unfolded detector effects for the simple example process of ZW → lljj production at the LHC using a conditional Invertible Network (cINN).
- After, allowed for a variable number of QCD jets..

Paper demonstrated that this application is feasible.
=> Potential to improve unfolding methods.



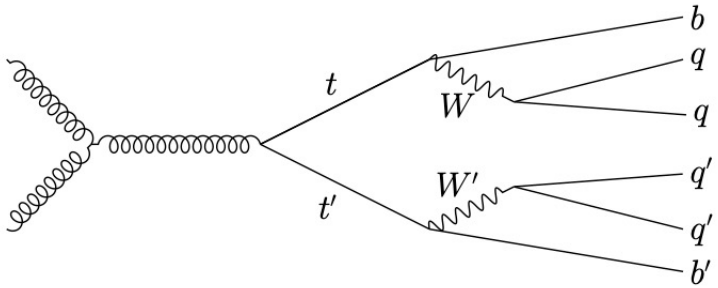
Flow models can also be reversible.

SPA-NET for all-hadronic top quark events

Reference: https://arxiv.org/pdf/2010.09206.pdf and https://arxiv.org/pdf/2106.03898.pdf

How attention networks are used outside of HEP: Natural language processing for translation

Tl;dr: Can attention networks help solve the problem of combinatorics for decay-product-assignment for top events?



Idea: Jet assignment for hadronic tops using Symmetry Preserving Attention NETworks (Spa-Net)

Attention networks focus on important input data through the use of gating.

- activities of a set of neurons are multiplied, component-wise, by activities of another set

Aim here: to reduce the problem of combinatorics in ttbar events.

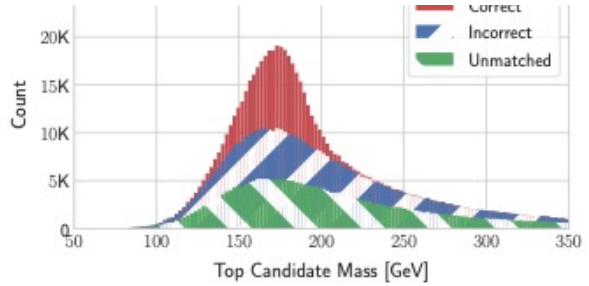
Network output should be invariant under permutations of the input jet order.

Authors state that fraction of events that are well reconstructed:

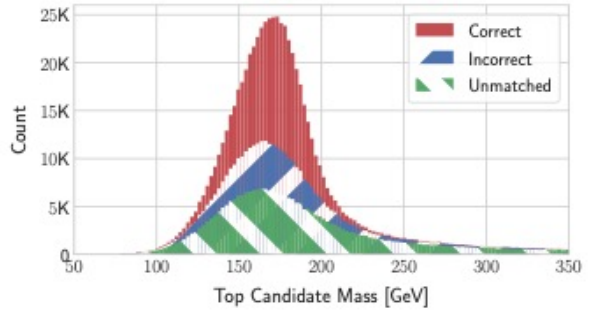
- **37.7%** (existing methods)
- **64.1%** (this new technique)

TABLE I: Performance of the χ^2 and SPA$t\bar{t}$ER assignments assessed by per-event efficiency ϵ^{event} and per-top efficiencies ϵ^{top}, inclusively and by jet multiplicity N_{jets}.

	χ^2 Method			SPA$t\bar{t}$ER		
N_{jets}	ϵ^{event}	ϵ_2^{top}	ϵ_1^{top}	ϵ^{event}	ϵ_2^{top}	ϵ_1^{top}
6	61.8%	65.0%	24.2%	80.7%	84.1%	56.7%
7	40.8%	50.4%	24.6%	66.8%	75.7%	56.2%
≥8	23.2%	35.5%	20.2%	52.3%	66.2%	52.9%
Inclusive	**37.7%**	**47.0%**	**23.0%**	**63.7%**	**73.5%**	**55.2%**



(a) χ^2



(b) SPA$t\bar{t}$ER

FIG. 4: Stacked distributions of reconstructed m_{top} using (a) the χ^2, and (b) SPA$t\bar{t}$ER.

But really, there are many possibilities!

Auto-encoders

Deep Neural Networks

Graph Networks
https://arxiv.org/abs/2007.13681

Attentive Networks

Generative Adversarial Networks

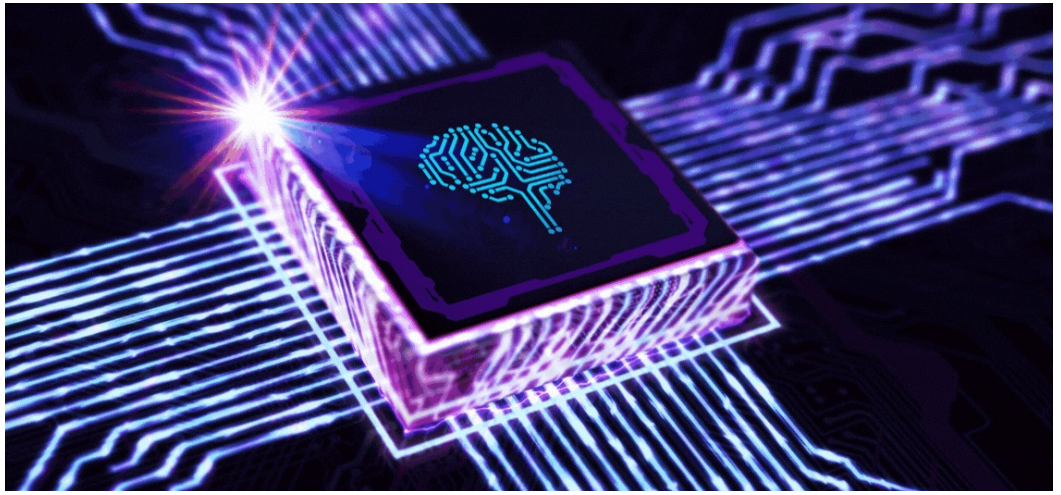
Long short-term memory

And more!!

Conclusion

- Have shown a number of applications of modern machine learning techniques already used in top quark research at the LHC to great effect.

- ATLAS and CMS analyses have made heavy use of BDTs in TMVA the past.
  - This was successful and led to improvements.
- Great scope for Deep Learning for Run 3 and beyond.



There are many more exciting ways we can implement new techniques to fully utilise and explore the data collected at the LHC!

Thanks to the following people for providing additional input for this talk:
- Nicolas Tonon
- Johnny Raine
- Sascha Caron
- Leonid Serkin