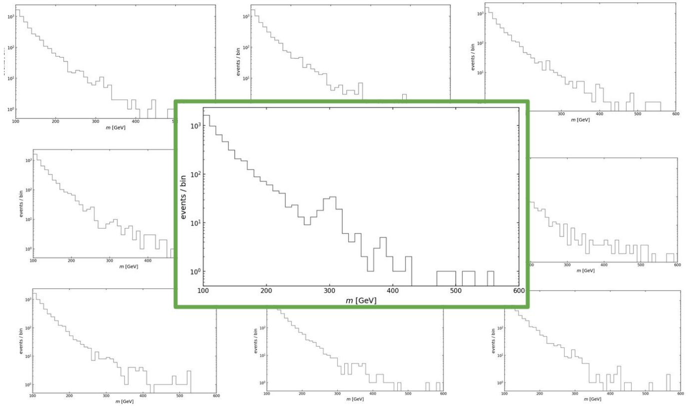
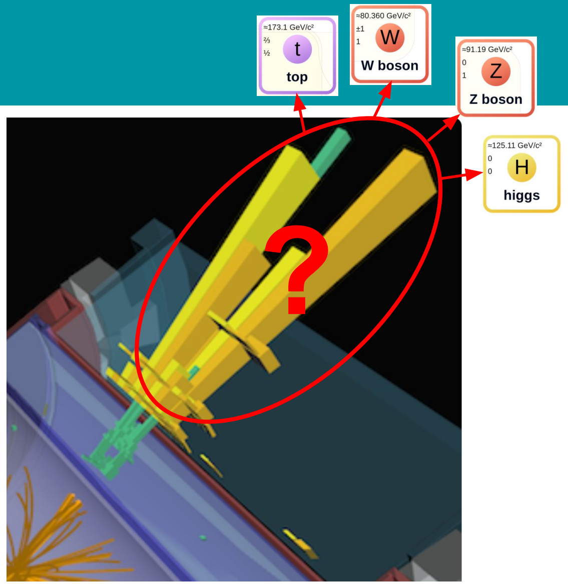
Searching for bumps and classifying jets in ATLAS



Eva Mayer

LPCA seminar

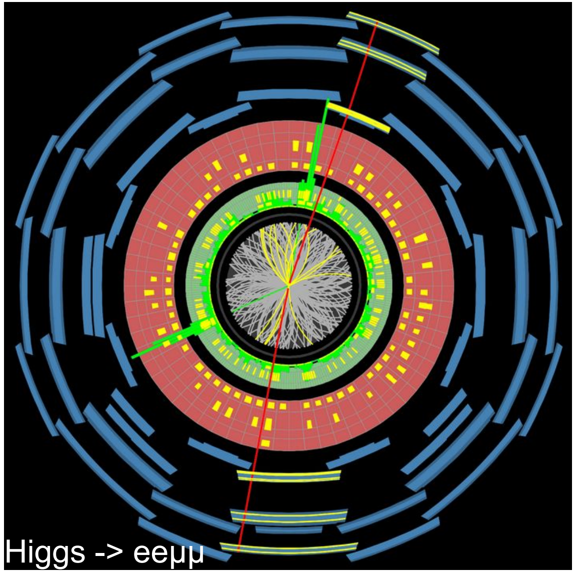


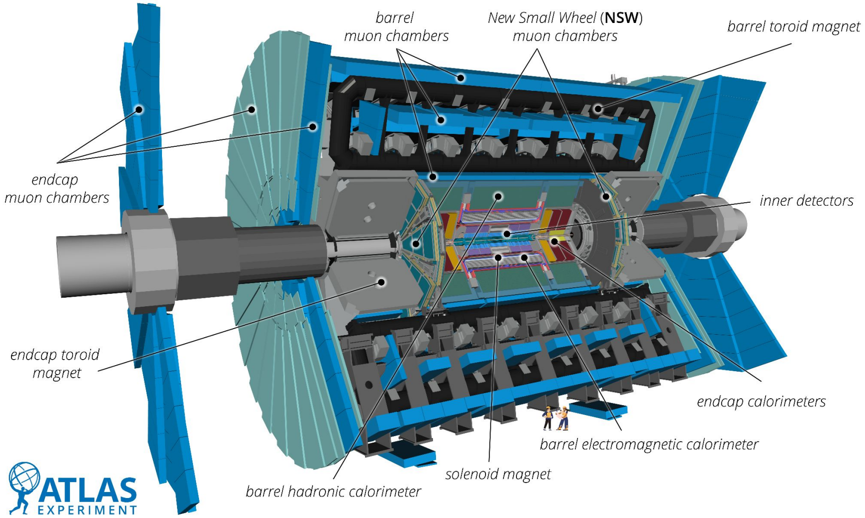
The Large Hadron Collider

- ★ 27km long ring
- ★ Accelerates protons
- ★ Collides them at 13.6TeV center-of-mass energy
- ★ 4 main experiments record the collisions: ATLAS, ALICE, CMS, LHCb

The ATLAS experiment

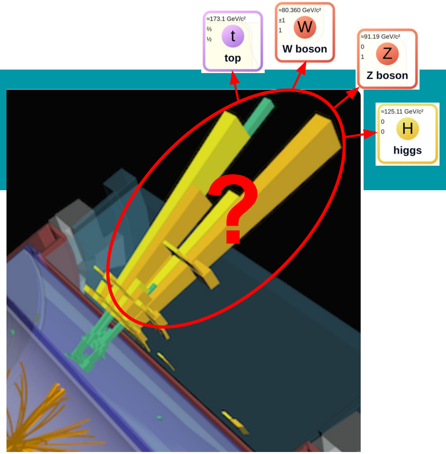
★ General-purpose detector:
  - testing the Standard Model (SM)
  - Searching for BSM physics



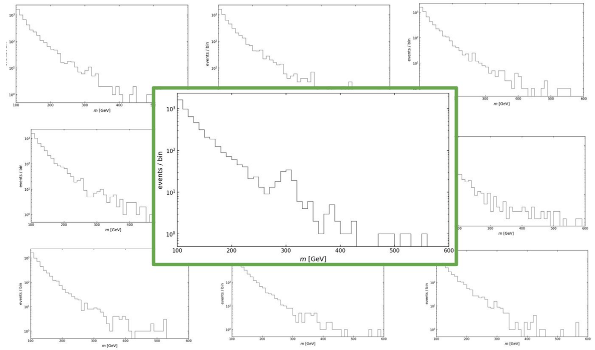


Reconstructs: photons, electrons, muons, hadronic decays

Topics of this talk

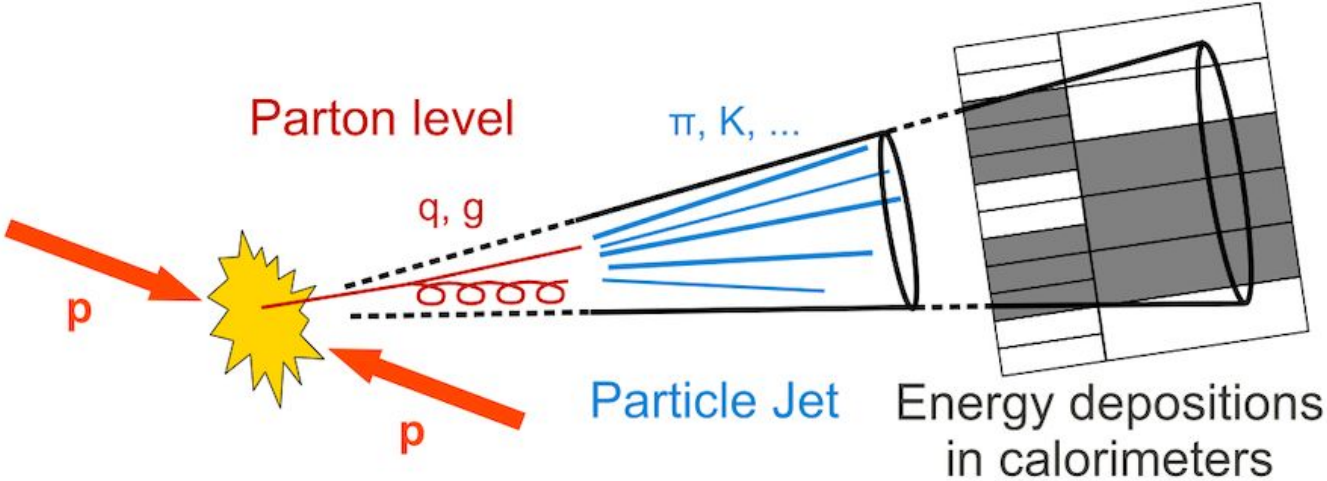


- Two parts:
    - How can we reconstruct the physics objects from the detector response?
    - How can we speed up the analysis of the data using ML ?



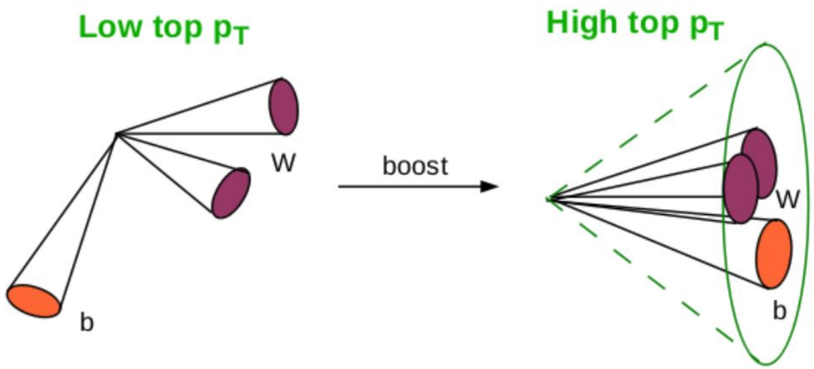
Zooming in on the reconstruction of jets

★ Quarks and gluons hadronize (=form compound particles) in detector
★ Resulting parton shower leaves group of energy deposits in calorimeter
★ Reconstructed cone representing the shower originating from one particle: “jet”



Large-Radius jets

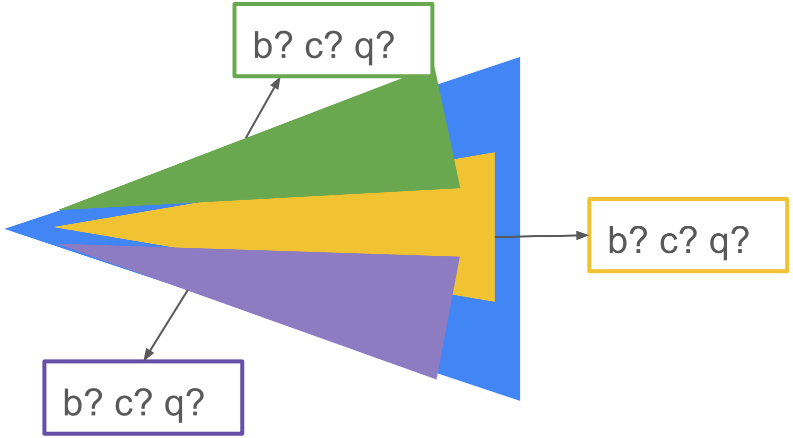
★ "Large-R jet": Jet with large radius
★ Can contain smaller jets (subjets)
★ Can come from:
  ○ QCD (=light quarks and gluons)
  ○ boosted (->high pT) top quark
  ○ boosted W boson
  ○ boosted Z boson
  ○ Higgs boson



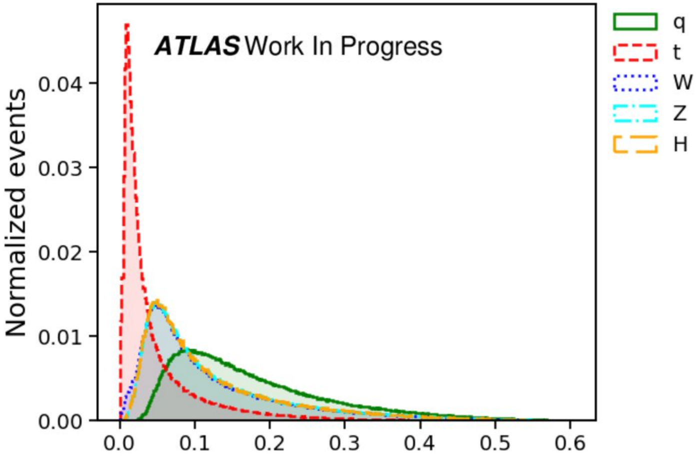
★ Typical approach: Create a tagger for one class vs QCD
★ Can we create a tagger that can simultaneously classify for all 5 classes?

Features

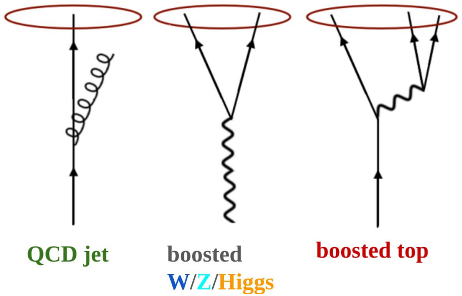
- **Jet-level features:**
  - Describe the jet as a whole
  - e.g. kinematic properties, substructure features,..
- **Subjet features:**
  - smaller, variable-radius jets can be reconstructed within the large-R jet
  - per subjet (up to 3): kinematic properties + flavor tag

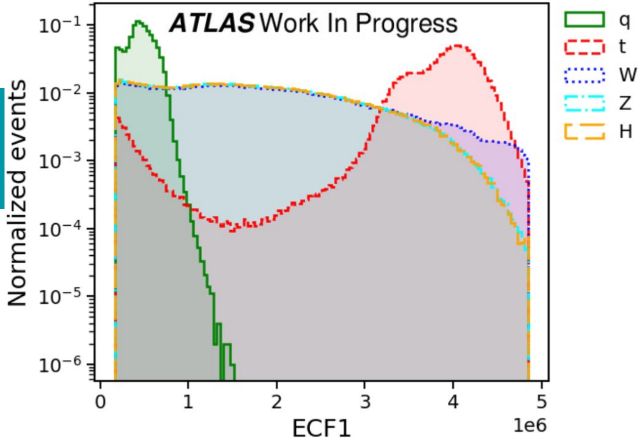


Features

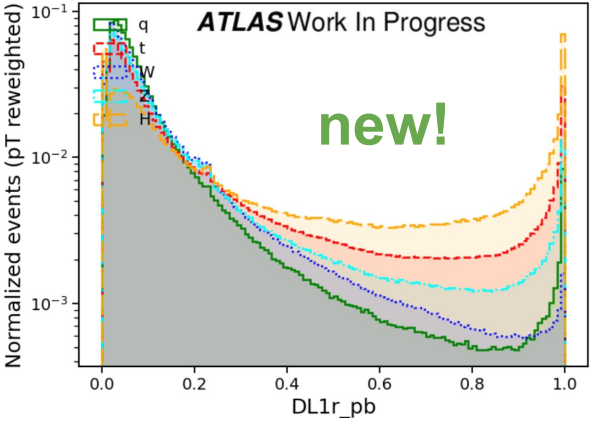


n-subjettiness variable





Scalar sum of pT of constituents

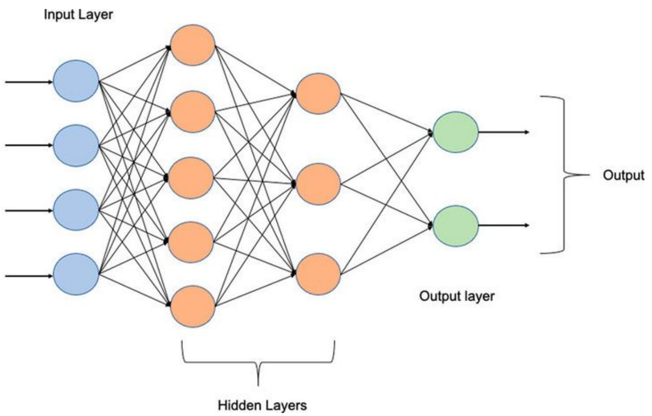


Subjet b-tagging score

… and many more jet- and subjet-features

Neural Network

★ Fully connected neural network
★ Input: 33 features, up to 3 subjets
→ input shape (3,33)
★ Output: 5 scores, one probability score per class
★ 1.8 million simulated jets per class for training



Schematic view of a neural network taken from source.
Actual details of the architecture are shown to the right

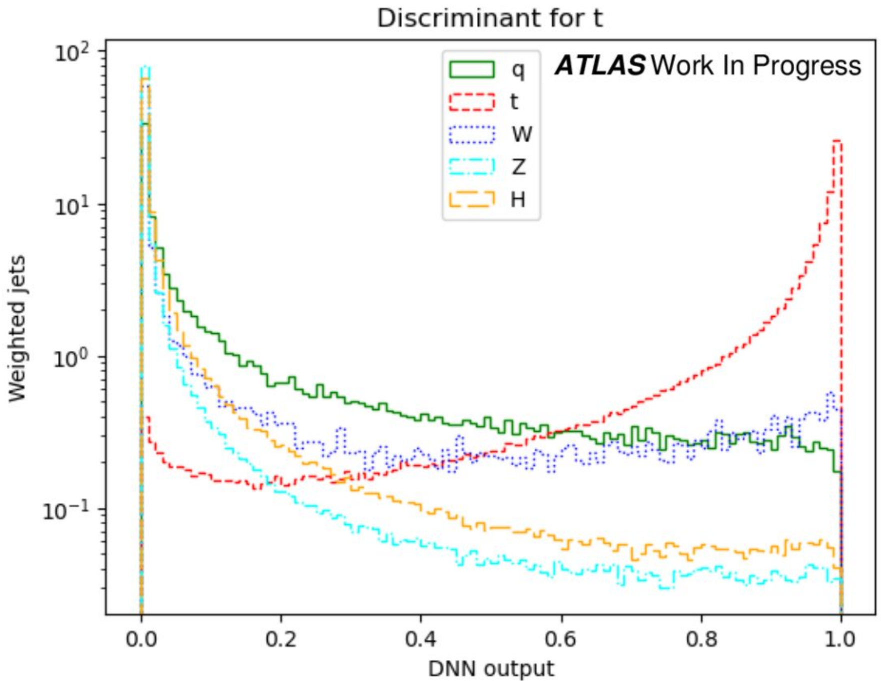
Input shape: (3,33)

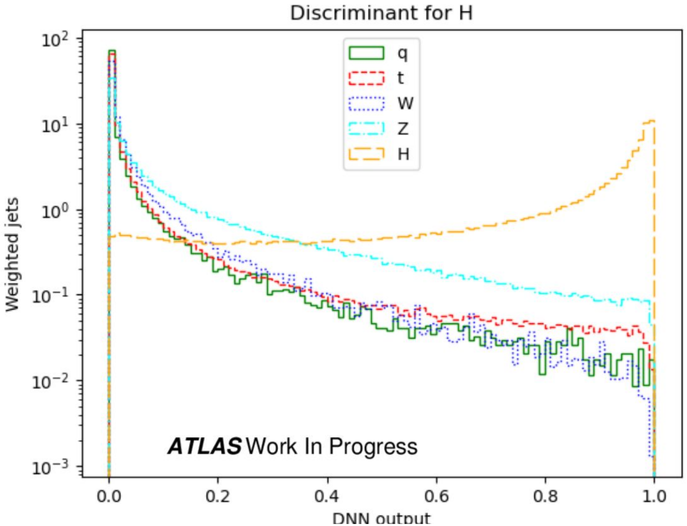
	activation	output shape
Dense(200)	gelu	(3,200)
Dropout		(3,200)
Flatten		(200)
Dense(200)	gelu	(600)
Dropout		(200)
Dense(200)	gelu	(200)
Dropout		(200)
Dense(5)	softmax	(5)

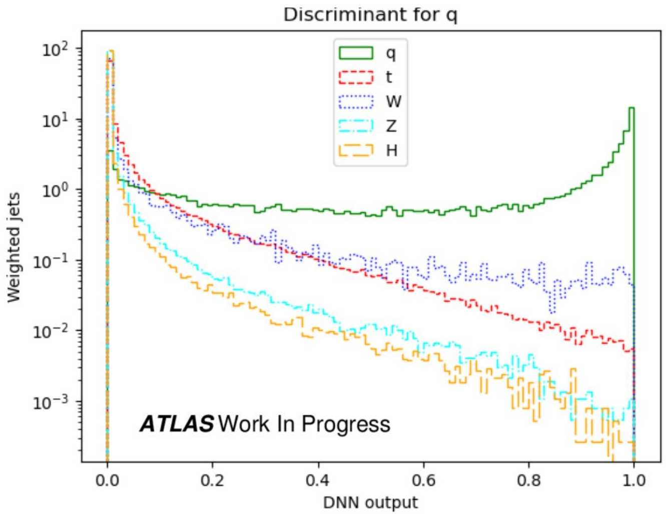
Output shape: (5)

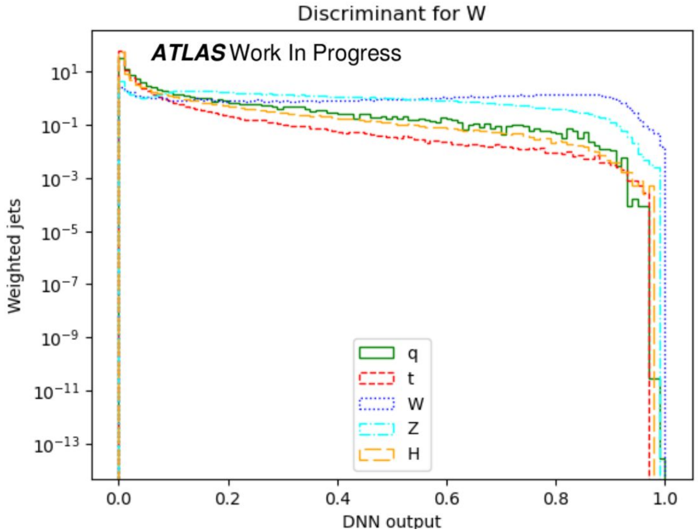
Output scores

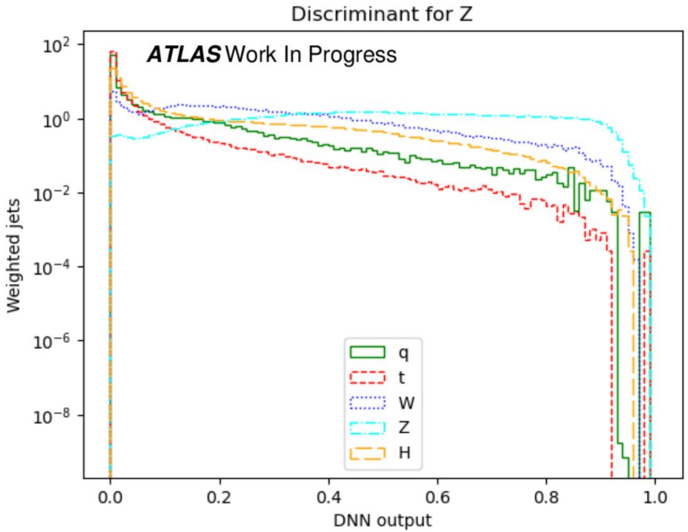
★ 5 scores: 1 for each class
★ the scores are normalized
→ the 5 scores sum up to 1 for each jet
★ Can be used to define selections that favor one specific class



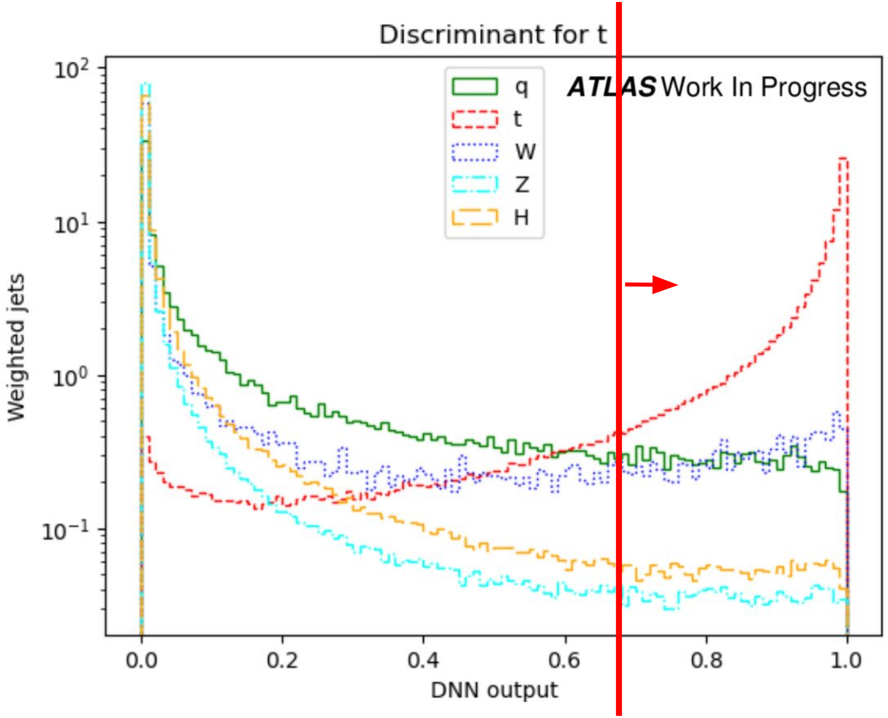

Discriminant for H
q
t
W
Z
H
ATLAS Work In Progress
Weighted jets
DNN output

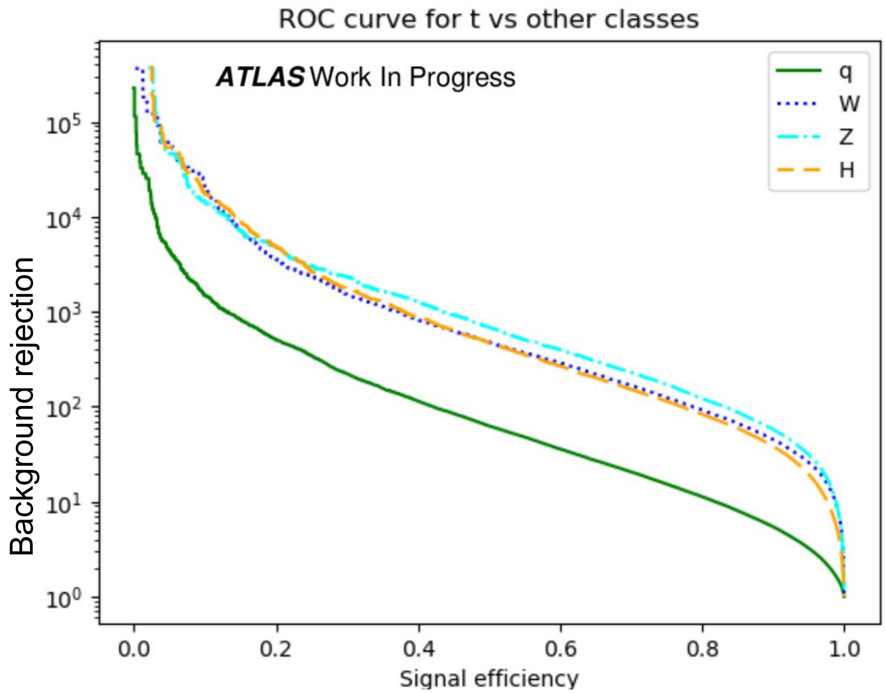

Discriminant for q
q
t
W
Z
H
ATLAS Work In Progress
Weighted jets
DNN output


Discriminant for W
ATLAS Work In Progress
q
t
W
Z
H
Weighted jets
DNN output


Discriminant for Z
ATLAS Work In Progress
q
t
W
Z
H
Weighted jets
DNN output

Defining cuts for single-class tagging

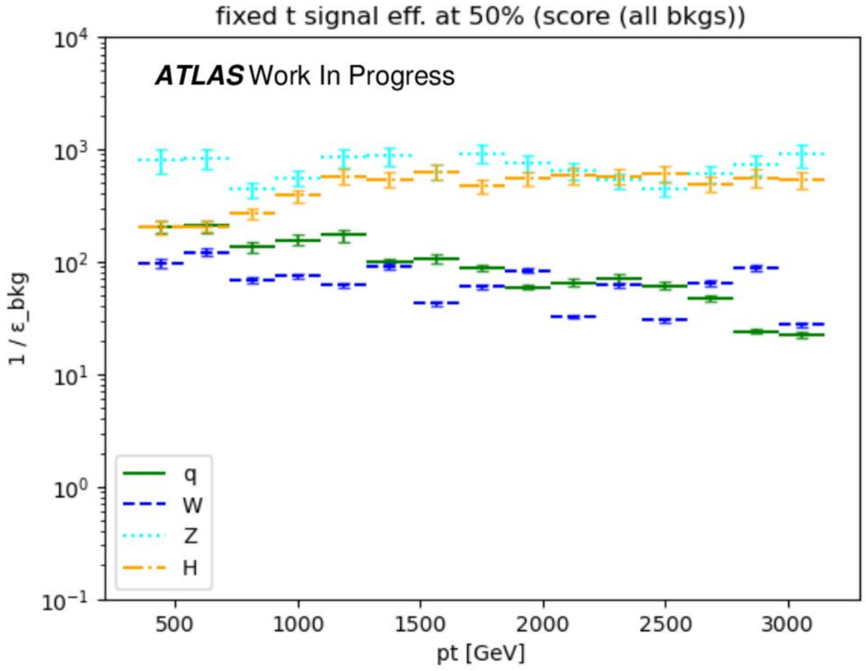




Different thresholds lead to different signal efficiency and background rejection

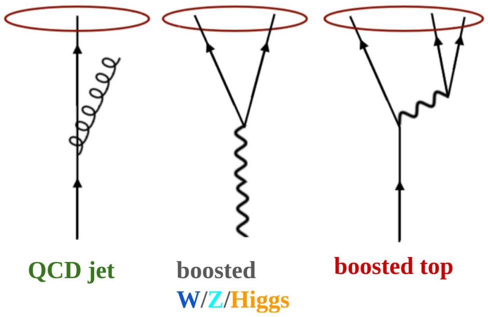
Background rejection vs pT

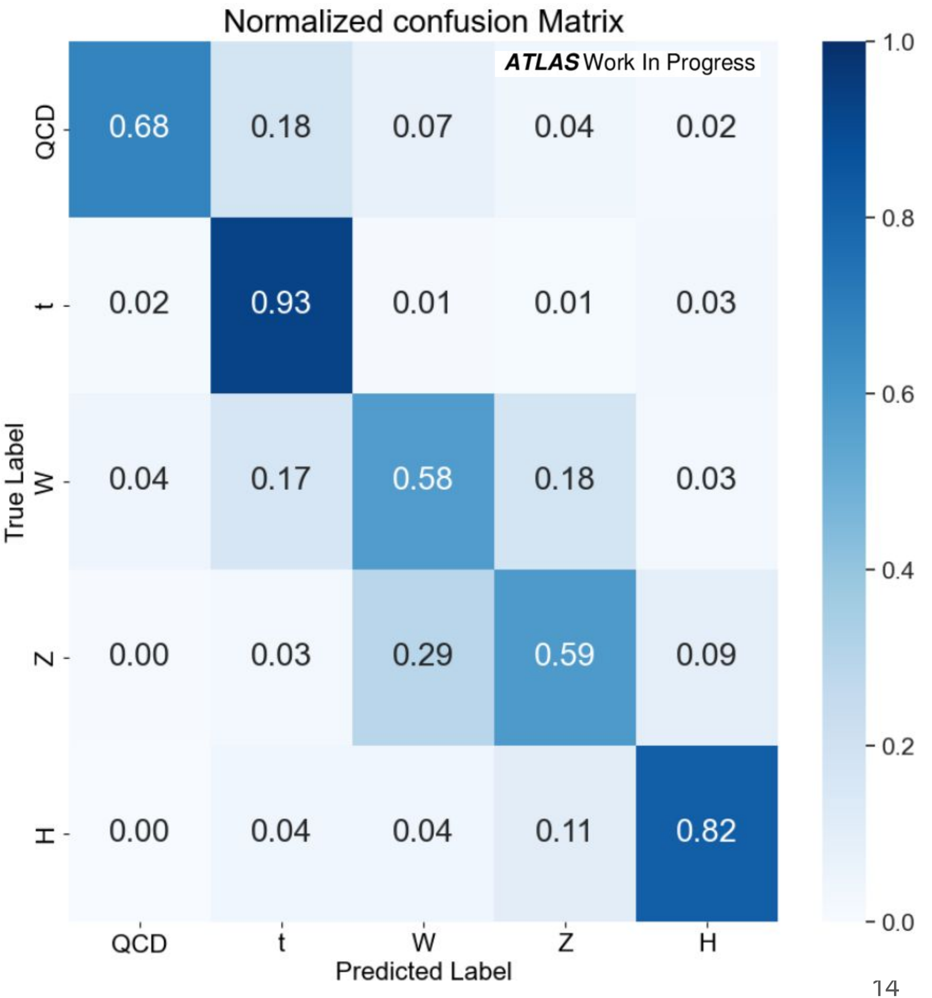
- ★ Signal efficiency is fixed across pT range → different thresholds per pT bin
- ★ Stability of background rejection across pT range tested



Multi-class tagging

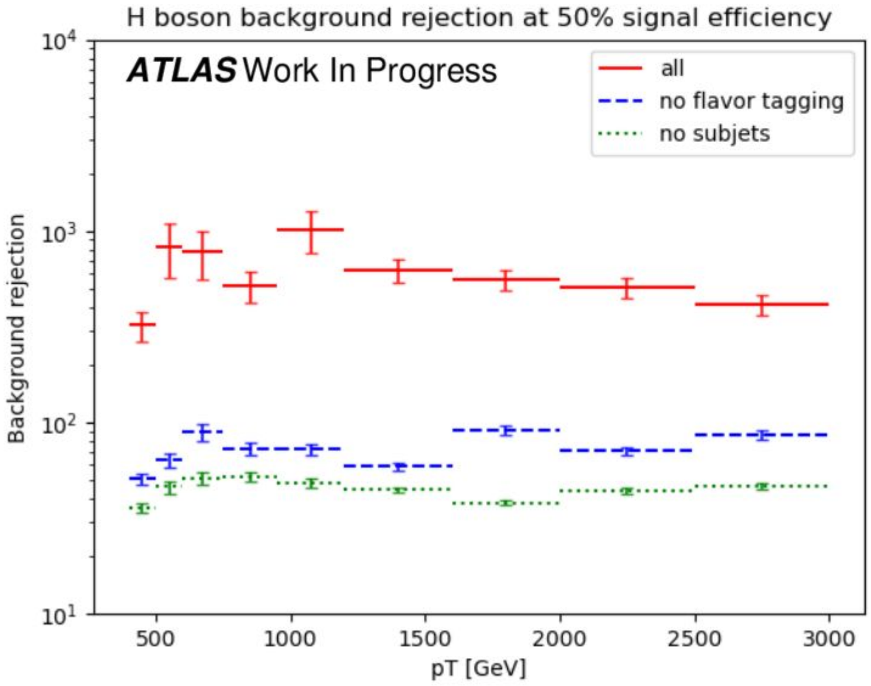
- ★ **Multi-class decision rule:** choose class with maximum output score per jet
- ★ Main confusion: W vs Z
  - ○ similar mass (80 vs. 91 GeV)
  - ○ similar decay





Impact of the subjet features on performance

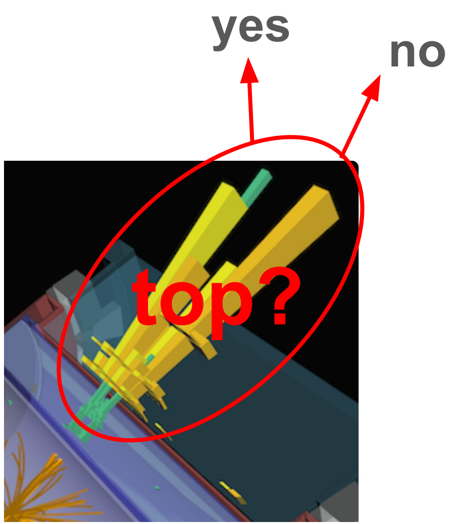
★ The subjet features, especially their flavour tag, are a novel input feature for large-R taggers
★ Their impact was studied in detail
★ 3 trainings:
  ○ training with all features
  ○ training with all except subjet flavor tag
  ○ training with only jet-level features



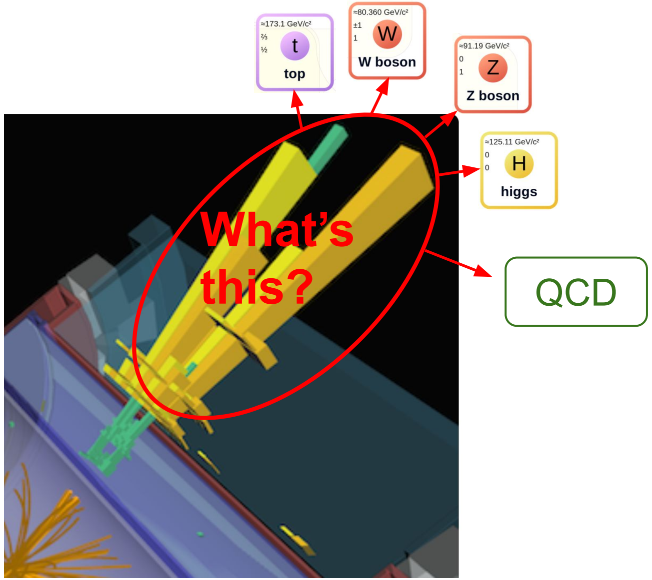
In this plot, the background only contains QCD

Part 1 summary

We went from:



To:

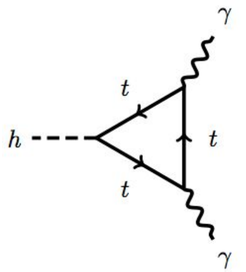


Part 1 summary

- ★ We can tag all classes at the same time
- ★ Good performance with very simple architecture
- ★ Very simple input features
- ★ Subjet flavor tag improves performance
- ★ This tagger opens the door for analyses to tag several boosted objects at once

  → **Stay tuned for possible use case!**

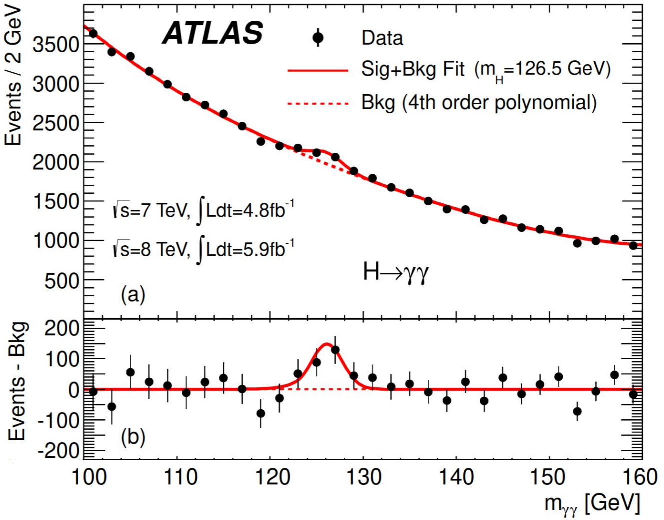
Zooming out of jet tagging



Objects are reconstructed.. Now what do we do with the data?

We look for New Physics:

- ★ Select a theory prediction for a new particle
- ★ Choose best final state to find it in
- ★ Select decay products of predicted particle
- ★ Simulate background+signal processes of inv. mass
- ★ Compare data to expectation

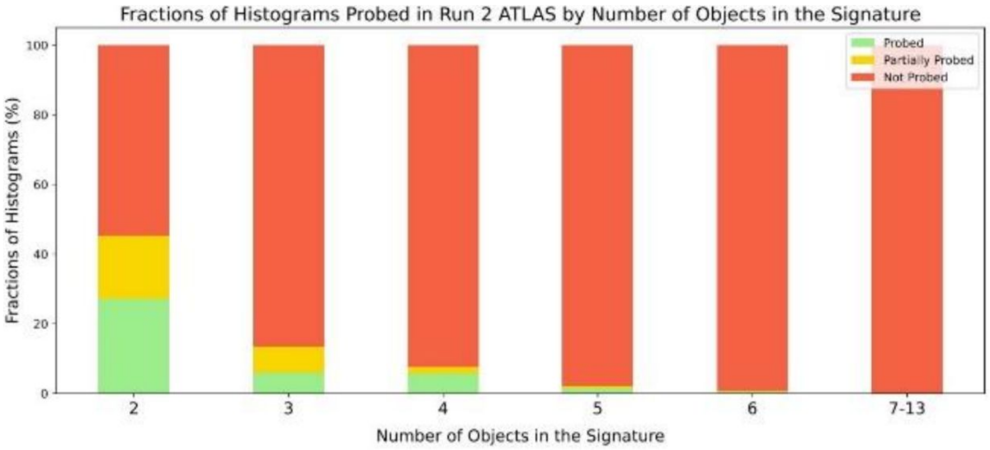


reference: ATLAS Higgs discovery: https://arxiv.org/abs/1207.7214

Status of ATLAS Run 2 searches

★ One search takes a few years with many people involved
★ Large majority of histograms of Run 2 (data taken 2015-2018) unprobed
★ No New Physics found so far
→ We need a more efficient approach:
Can we create a smart ML tool to scan hundreds of histograms* for excesses?

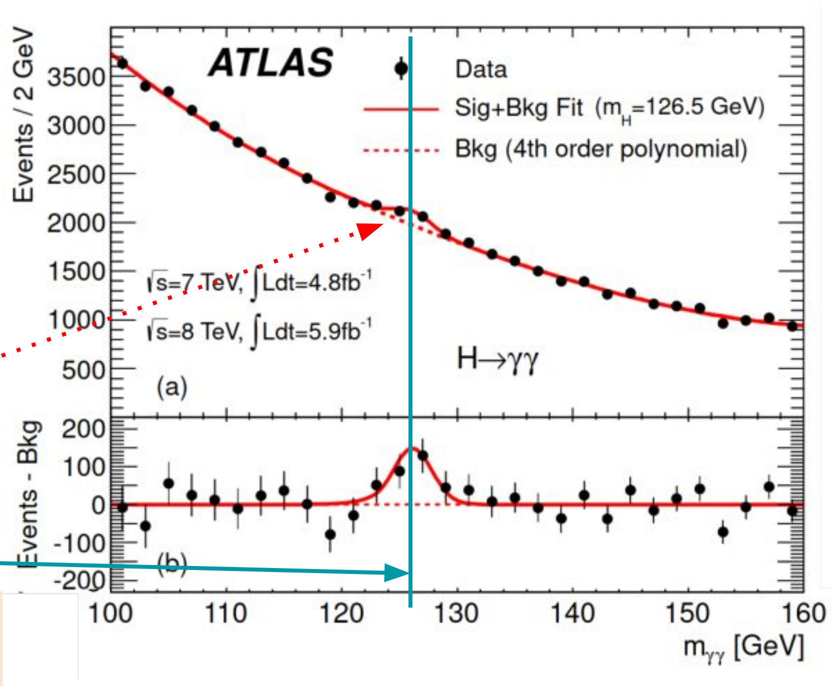
***Histograms:**
E.g. final state with 3 objects (2 electrons,1 jet)
→ inv mass histograms for
(e_0e_1), (e_0j_0), (e_1j_0),$(e_0e_1j_0)$



BumpNet

★ Instead of probing each histogram individually: **Train a network to scan all histograms at once**

★ How?
  - New particles appear as bumps in mass histograms
  - BumpNet needs to find bumps **without** knowing
    - the background shape
    - the signal mass

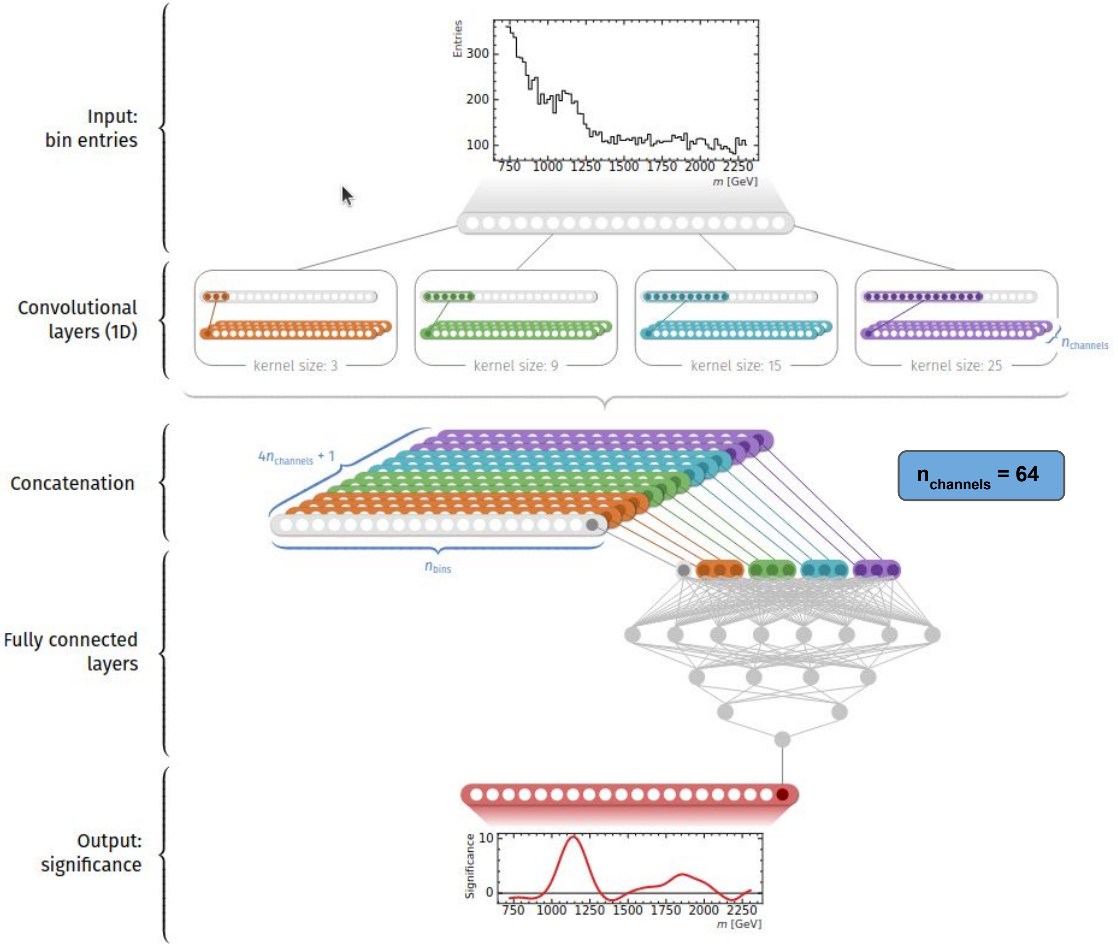


-> Rely on most general common properties:
Background is smoothly falling and
signal appears as bump

https://arxiv.org/abs/1207.7214

BumpNet

- ★ Convolutional Neural Network (CNN)
- ★ Layers of different kernel sizes to learn **local** and **global** patterns
- ★ MLP processes CNN output bin-by-bin
- ★ **Supervised learning:** target significance provided in training



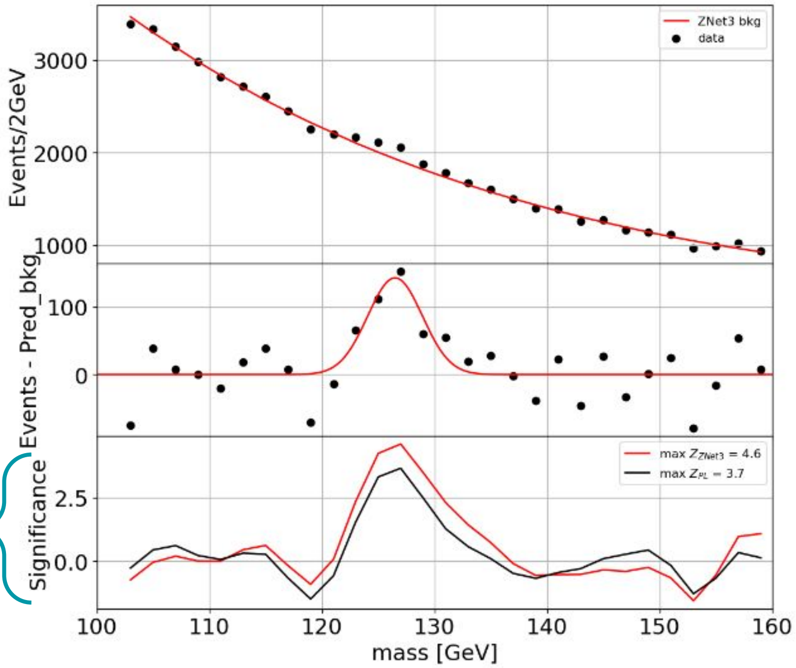
BumpNet

- ★ To predict the presence of bumps, the network learns to predict the statistical significance z_i of a bump at each mass m_i
- ★ Significance z: calculated with likelihood-ratio test [1]
    - ○ ratio of likelihood of background hypothesis to likelihood of gaussian signal:

$$z_i = \sqrt{-2\ln\left[\frac{L(0)}{L(\hat{\mu})}\right]}$$

Likelihood of **background** hypothesis

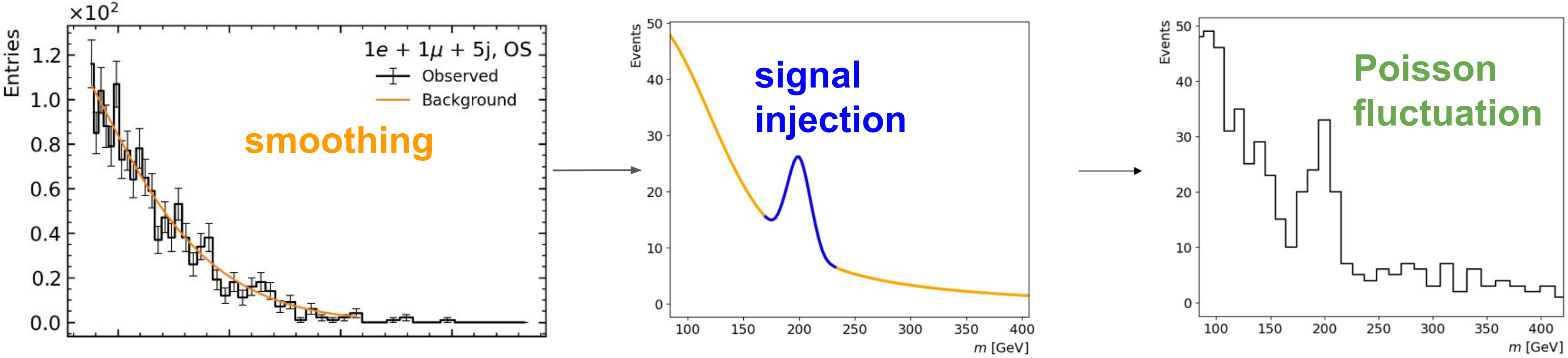
Likelihood of best fit for **signal**



[1] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Eur. Phys. J. C 71, 1554 (2011), [Erratum:Eur.Phys.J.C 73, 2501 (2013)], arXiv:1007.1727 [physics.data-an]

Training samples

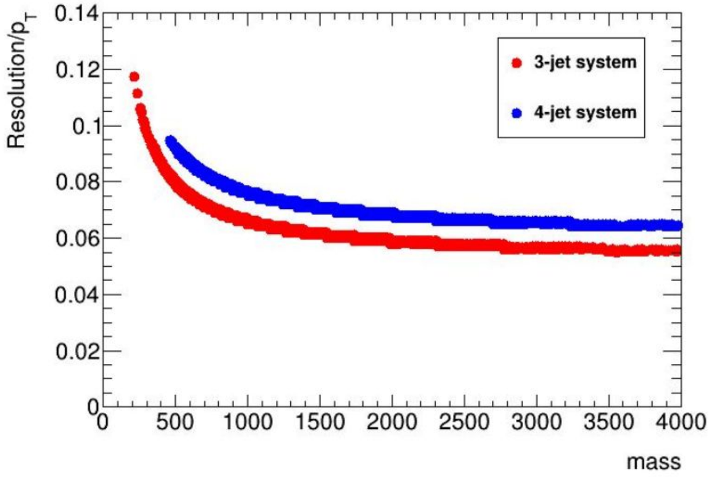
- ★ Analytical functions:
  - ○ Set of 20 smoothly falling functions
- ★ Histograms from MC simulations:
  - ○ Samples of MC simulated LHC-like data (Dark Machines[1] in proof of concept, ATLAS simulations later)
  - ○ Simulated samples are **smoothed**
- ★ Signal injection
  - ○ **Gaussian signal** injected (width = 1 bin)
  - ○ **Poisson fluctuated**

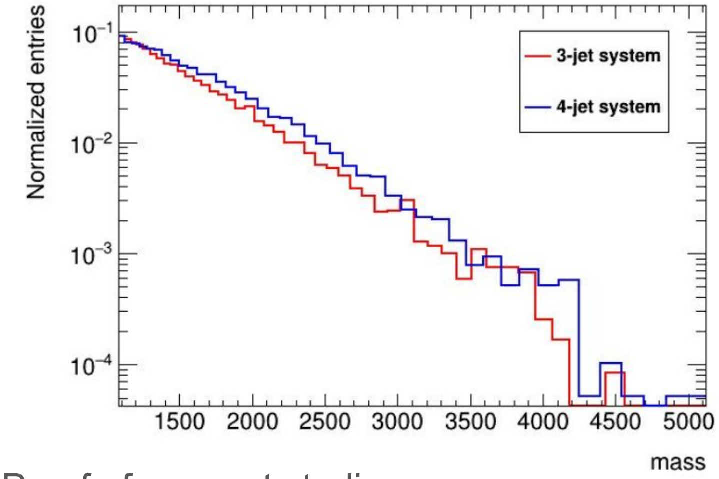


[1]Dark Machines: Aarrestad, T. et al. SciPost Phys. 12, 043 (2022). doi:10.21468/SciPostPhys.12.1.043

Object resolution, signal width, and bin size

- ★ Optimization of the search for **signal width of 1 bin**
- ★ Bin size needs to be adapted to **experimental resolution of the invariant mass**
- ★ Each object type has individual resolution that depends on its momentum and rapidity
- ★ Bin size varies between histograms and even inside each histogram depending on mass



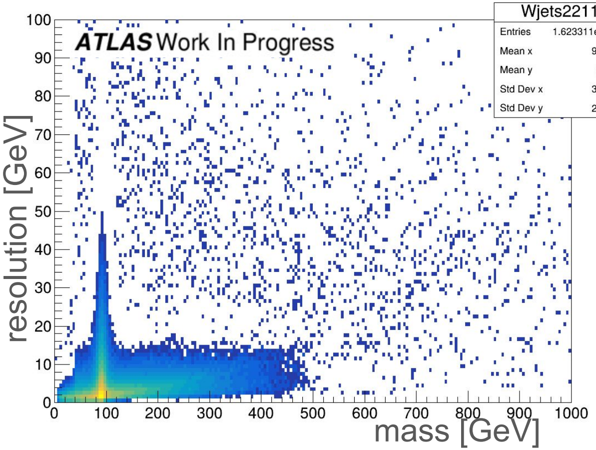


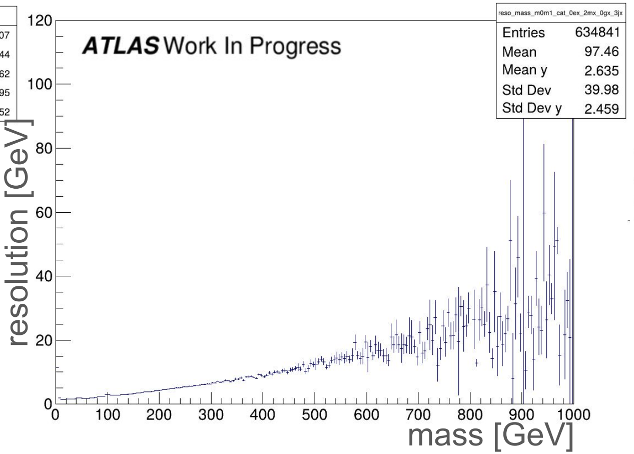
From the Dark Machines Proof of concept studies

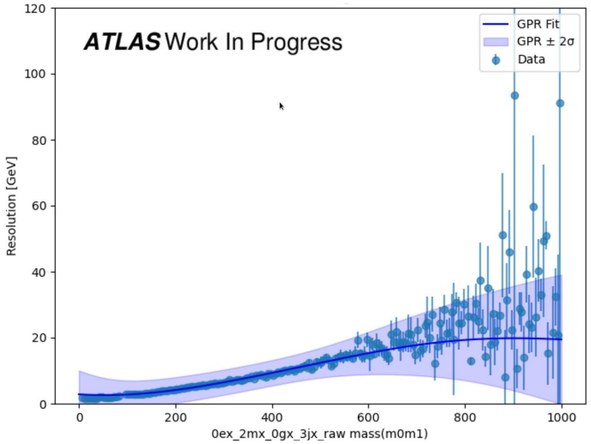
Rebinning strategy in ATLAS

Detector resolution per event and object is given as part of the simulated samples.

Workflow for each histogram:







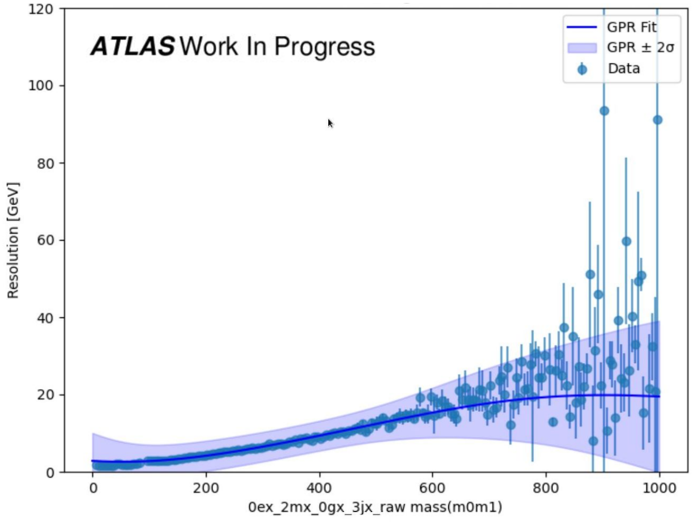
build 2D histogram resolution vs mass

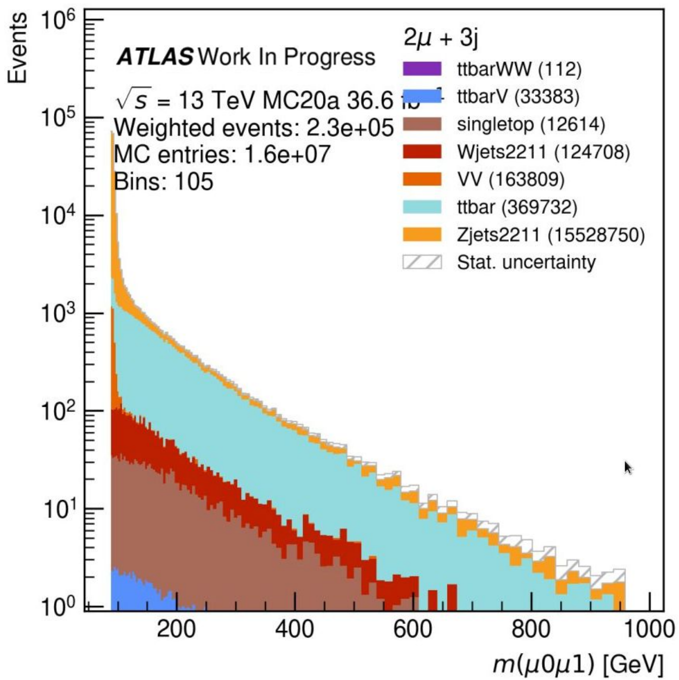
extract average resolution vs mass

perform Gaussian Process Regression (GPR) fit

Binning and detector resolution (ATLAS)

Fit values are directly translated into bin sizes

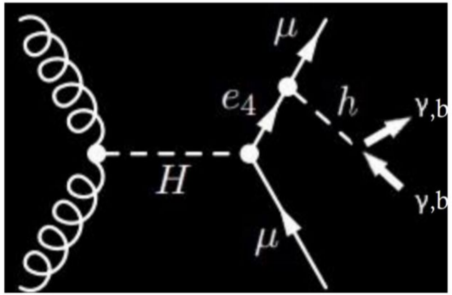




Lots of effort into finding a good fitting method that works reliably for hundreds of different histogram with different levels of statistics and different ranges of mass&resolution

Validation of bin sizes using BSM signal simulations

- ★ Usage of different ATLAS simulations of Beyond Standard Model (BSM) candidates
- ★ Gaussian fits to their bumps
- ★ Comparison of bump width to bin size extracted through profile fit

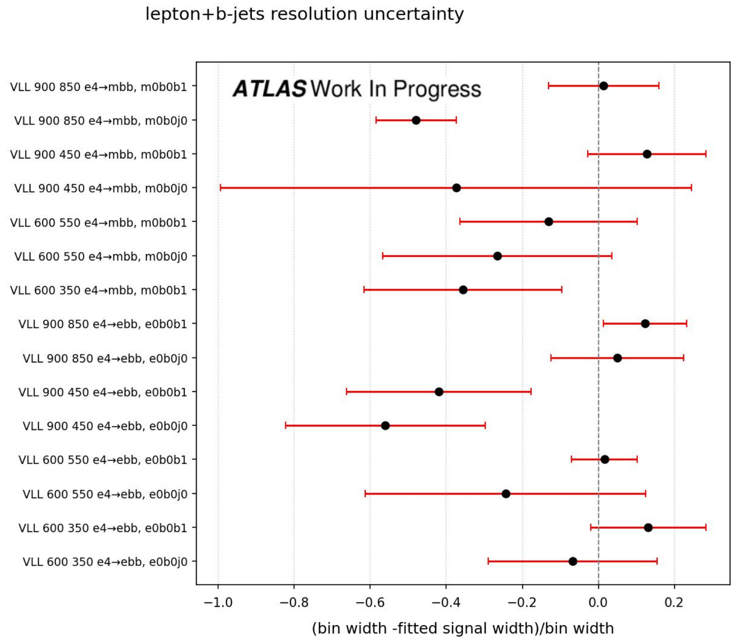


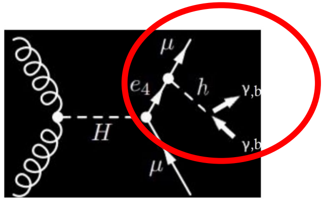
Heavy Higgs → l + VLL → llbb

Validation of bin sizes using BSM signal simulations

★ Comparison of the observed signal width in the simulation and the bin width extracted from profiles

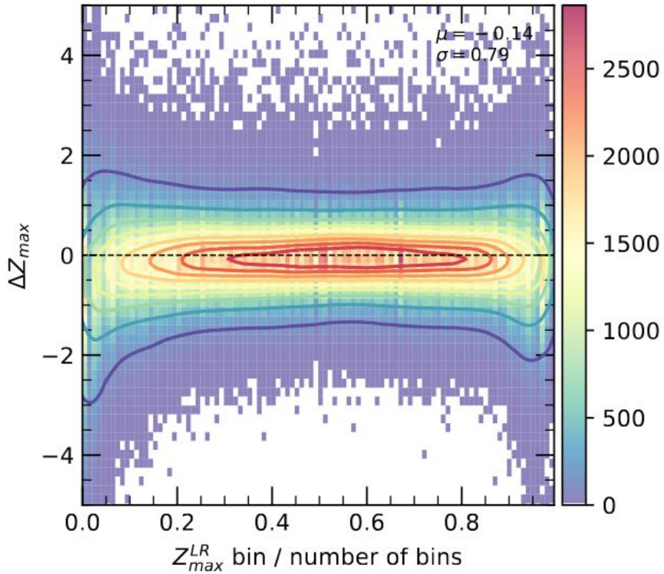
★ Signal simulations at different mass points and different decay modes to get some statistics

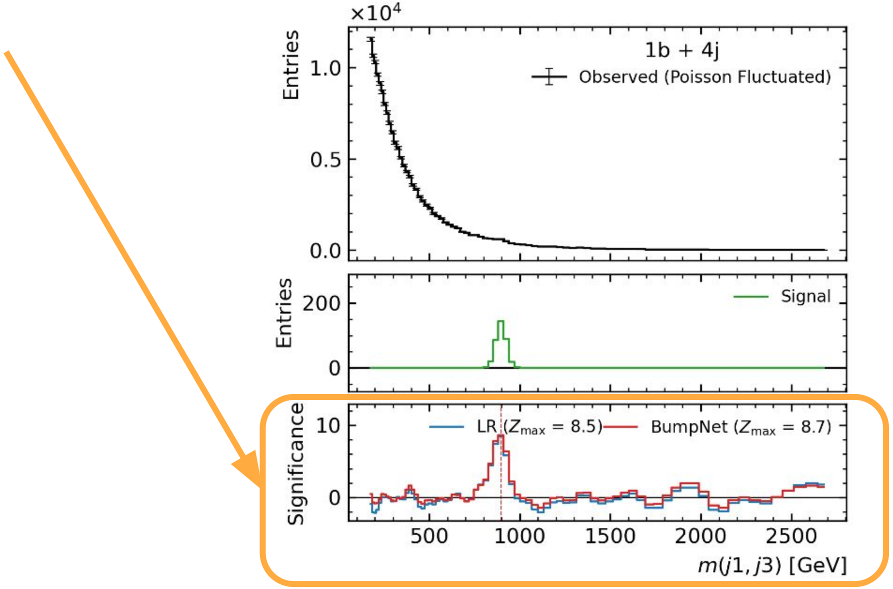




BumpNet Performance

- Main metric: difference between predicted significance z_{pred} and true significance z_{LR}
- $\Delta z_{max} = z_{LR} - z_{pred}$ at $\max(z_{LR})$

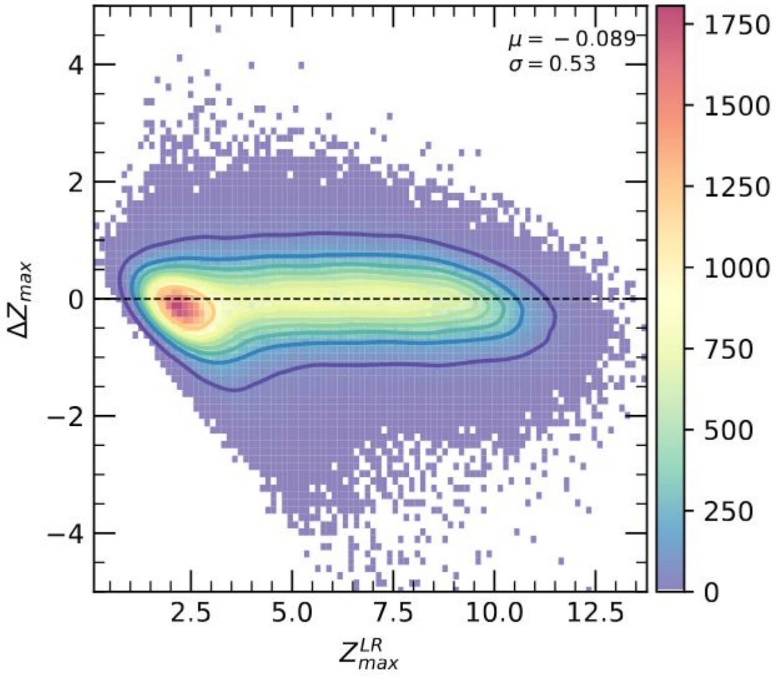




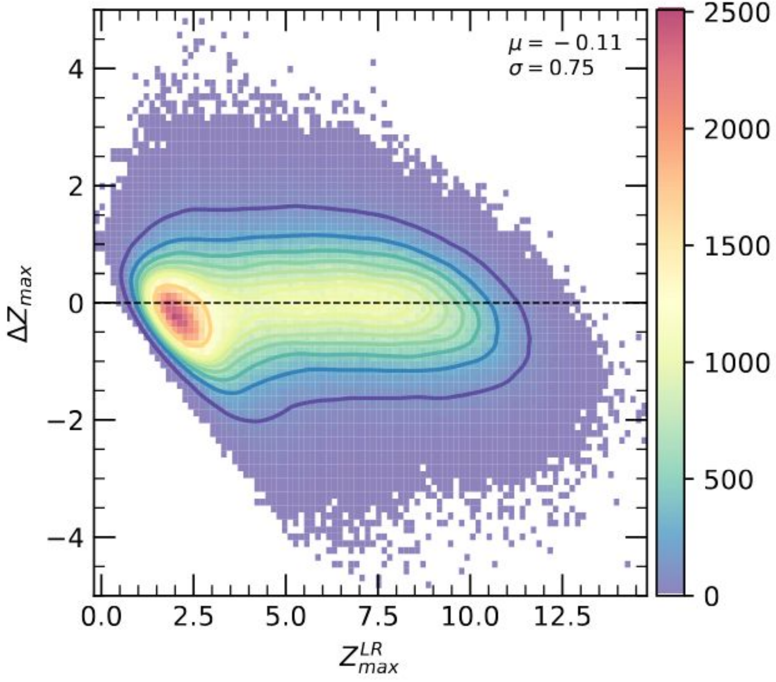
Results from Proof of concept: https://arxiv.org/abs/2501.05603

Performance with respect to signal strength

tested on functions



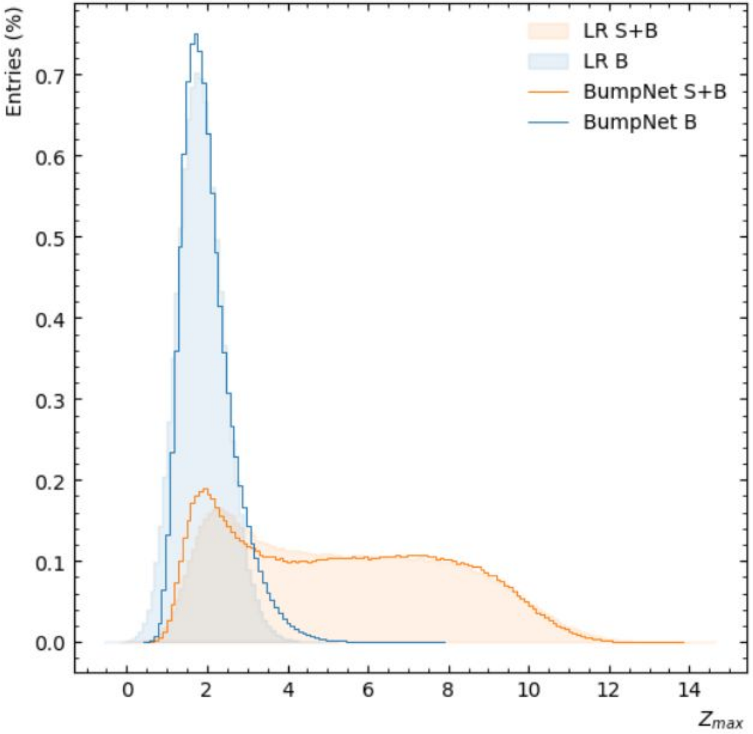
tested on Dark Machines samples



Results from Proof of concept: https://arxiv.org/abs/2501.05603

Performance

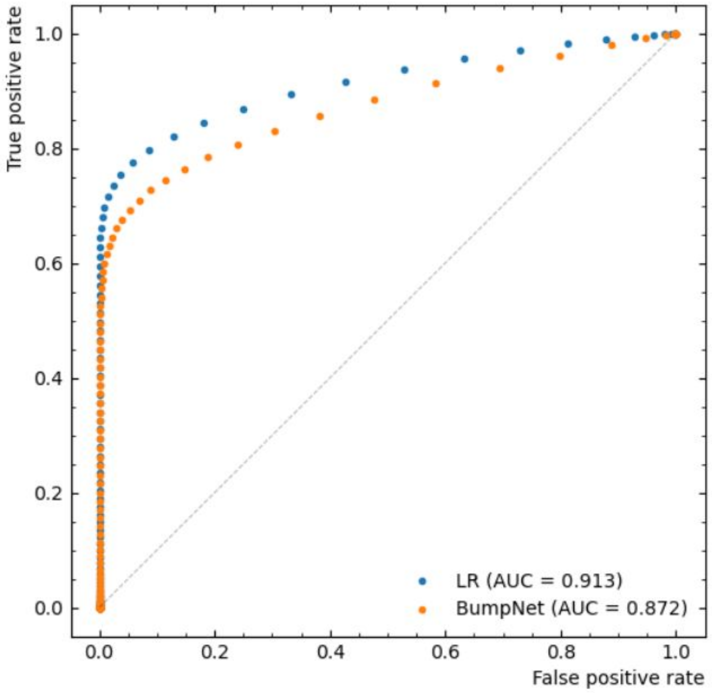
★ Colors: background only vs signal injected
★ Lines vs shaded: BumpNet vs Likelihood-Ratio (LR)
★ BumpNet predictions and LR follow similar distributions



Results from Proof of concept: https://arxiv.org/abs/2501.05603

Performance

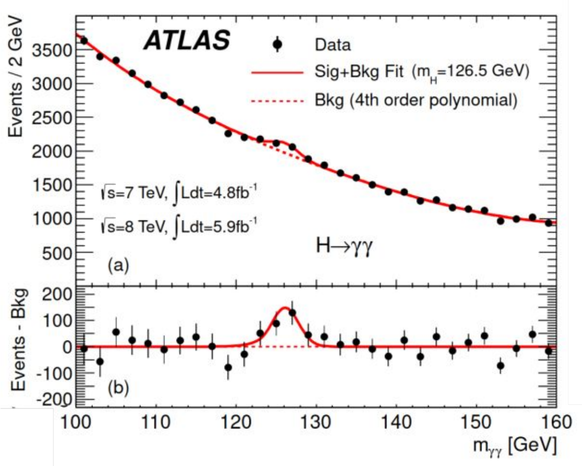
★ ROC curve: each point represents different threshold on zmax to distinguish signal/no signal
★ LR performance is the best theoretically possible performance

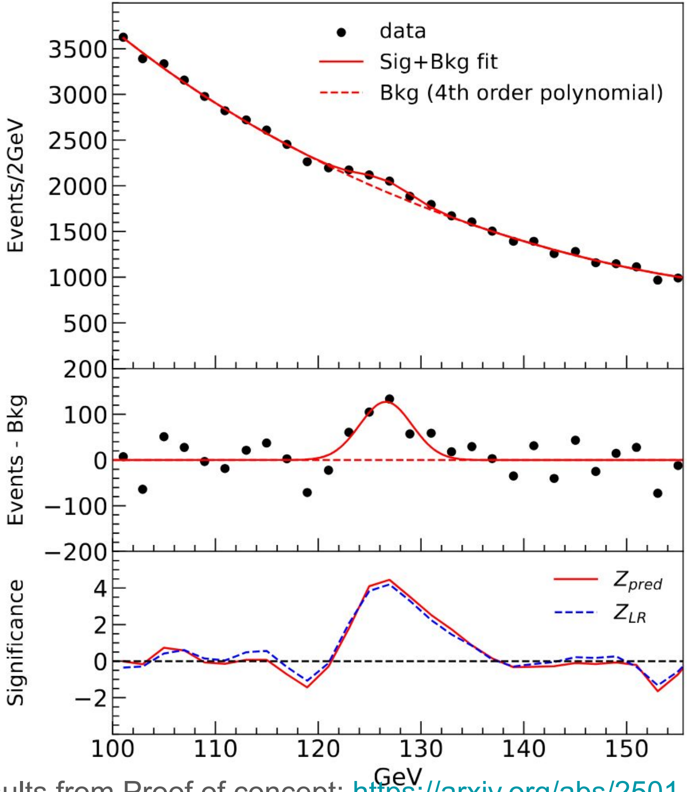


Results from Proof of concept: https://arxiv.org/abs/2501.05603

Performance over data: Higgs

- ★ Tested BumpNet on real data: H->$\gamma\gamma$
- ★ BumpNet succeeds to find Higgs bump at correct mass value
- ★ Very good agreement between z_{LR} = 4.2 and z_{pred} = 4.5



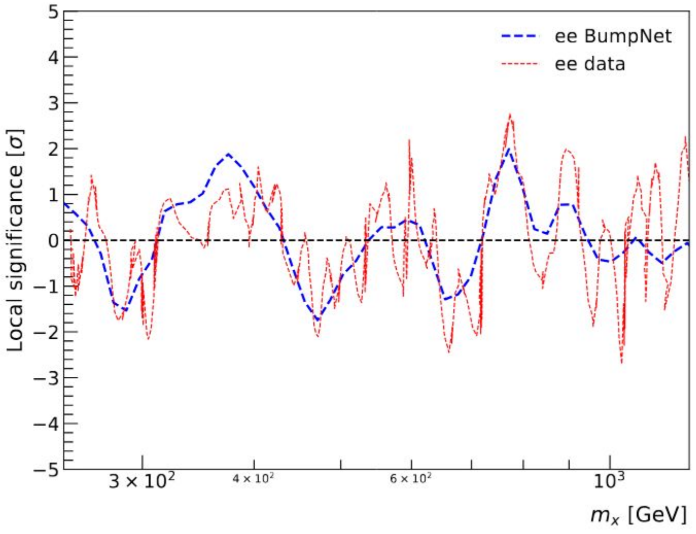


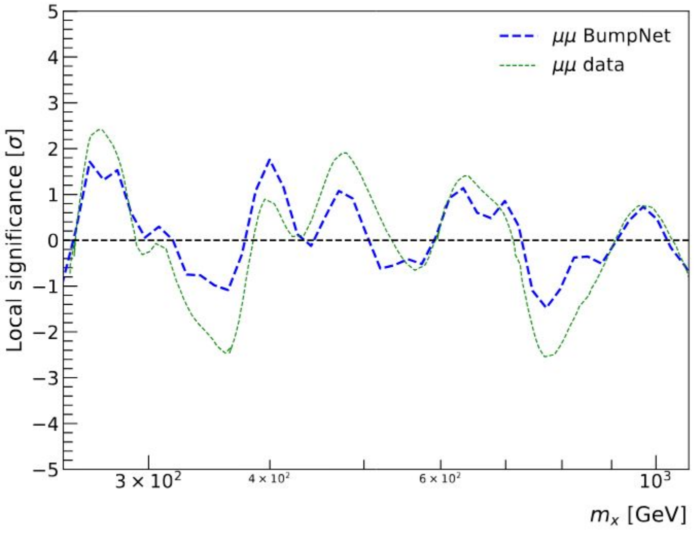
Results from Proof of concept: https://arxiv.org/abs/2501.05603

Performance over data: dilepton search

★ BumpNet applied to data extracted from recently published dilepton-search arXiv:1903.06248 [hep-ex]

★ BumpNet predicted significance aligns well with paper significance

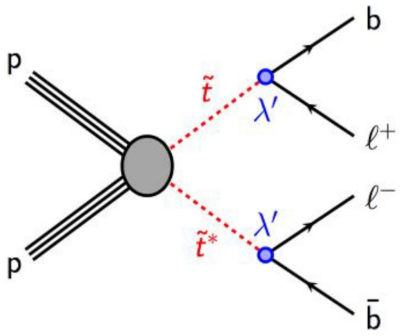




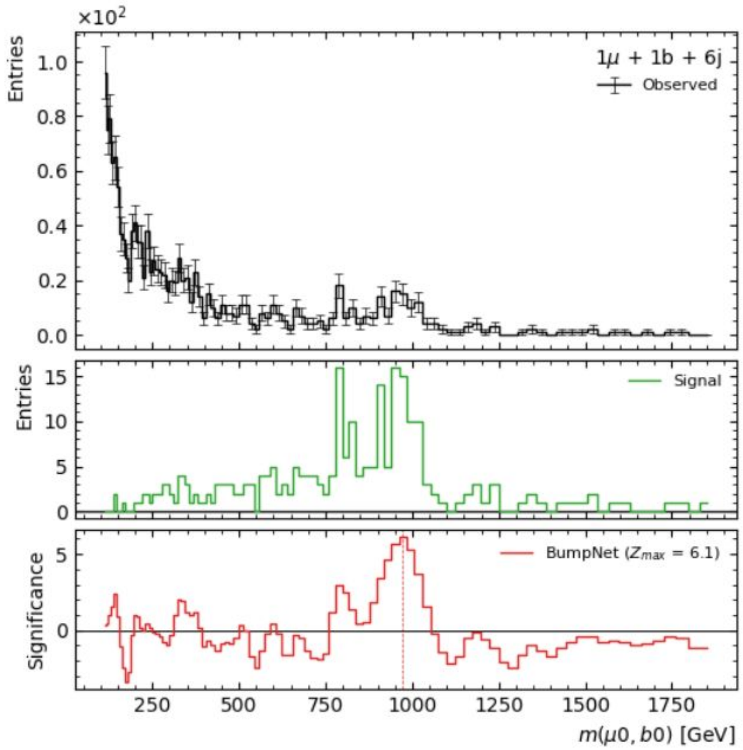
Results from Proof of concept: https://arxiv.org/abs/2501.05603

Performance over New Physics simulations

- Dark Machines samples include New Physics samples
- Tests on DM backgrounds with simulated signal bumps added
- Example: RPV stop -> bl



m_{stop} = 1TeV



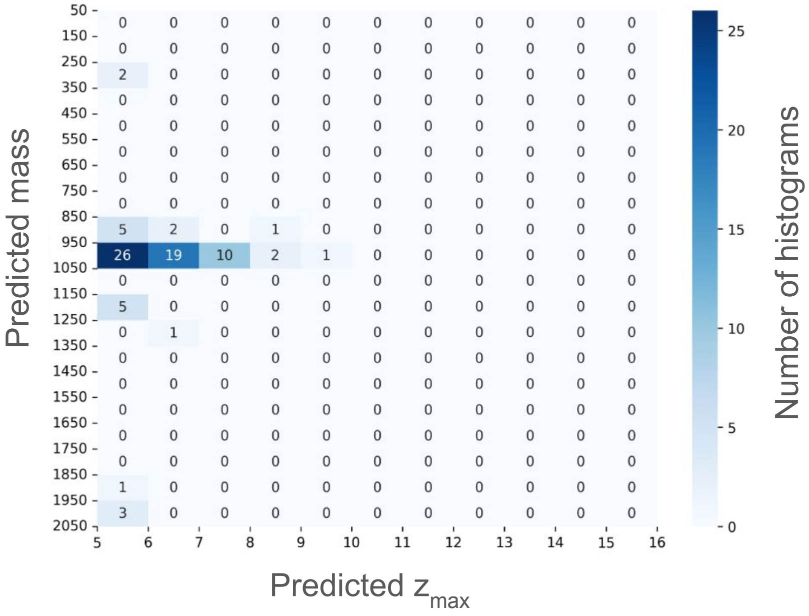
T. Aarrestad et al., SciPost Phys. 12, 043 (2022), arXiv:2105.14027 [hep-ph]

Results from Proof of concept: https://arxiv.org/abs/2501.05603

Performance over new physics simulations

- ★ False-positive rate of 0.1%
- ★ New physics particle likely to appear in multiple histograms
- ★ False positives appear, but unlikely to cluster at specific mass values over different histograms

Position of maximum vs significance



(f) Two RPV Stop$\rightarrow b\ell$

with m_{stop} = 1TeV

Results from Proof of concept: https://arxiv.org/abs/2501.05603

Part 2: Conclusion and outlook

- Proof of concept published (JHEP02 (2025) 122):
  confirmed that BumpNet performs well on functions and Dark Machines simulations
- Now: training on ATLAS-simulations, planning on applying to ATLAS data
- This analysis will be the first of this kind in ATLAS -> many open questions:
  - How can the results be interpreted?
  - Can we set limits on models?
  - Can we extract a global significance?
  - Systematic uncertainty of our prediction?

Bonus!

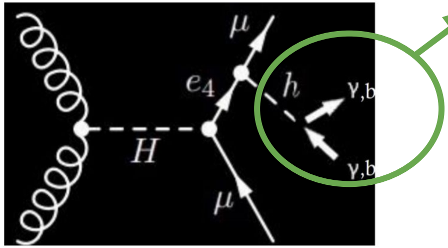
★ Can we use the multi-class tagger to increase sensitivity to signals in BumpNet?

★ Very preliminary work in progress:

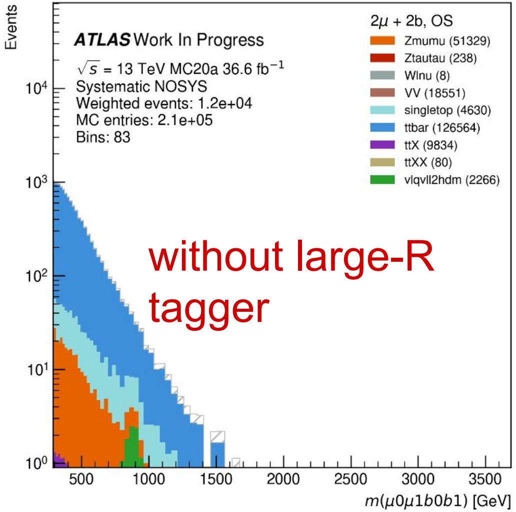
- Large-R tagging can be used to create histograms with hadronic W/Z/t/Higgs
- Multi-class tagger allows final states with more than one tagged object

  → **More histograms to probe**
- Example usage:

replace with Higgs-tagged large-R jet

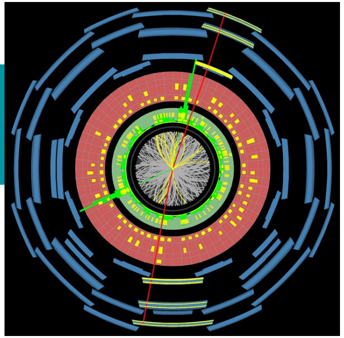


Heavy Higgs to VLL signal process

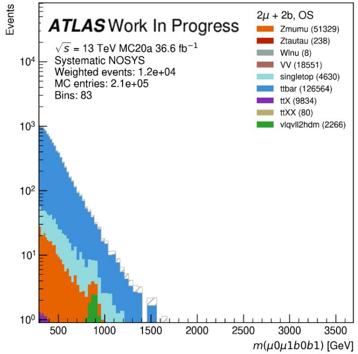


Can the tagger increase sensitivity? Find out in my defense!

Summary

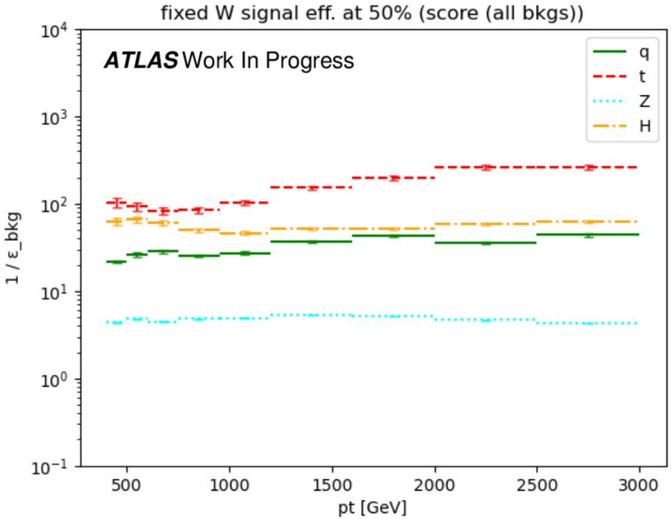


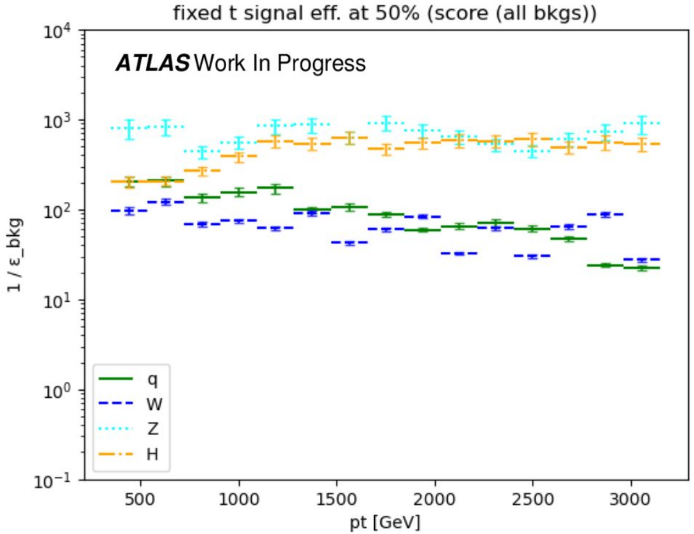
- ★ Machine Learning methods improve efficiency in all areas of ATLAS
- ★ In my work:
  - ○ Jet classification
  - ○ Data analysis for New Physics searches
- ★ ML methods allow us to redefine the problems and how they approached:
  - ○ single-class → multi-class tagging
  - ○ Searching 1 histogram → Scanning hundreds of histograms

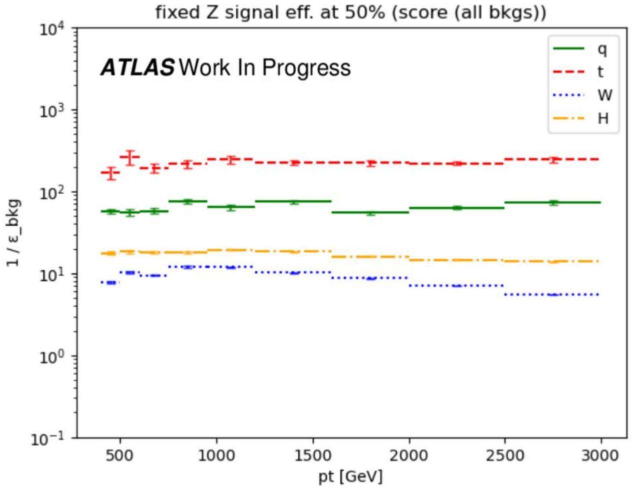


Backup

Background rejection vs pT

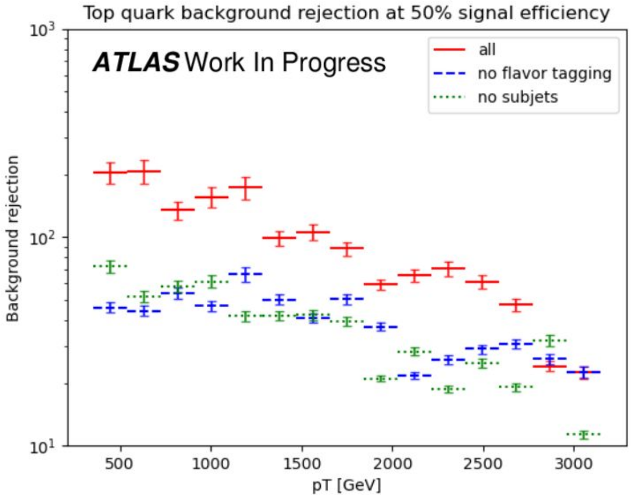


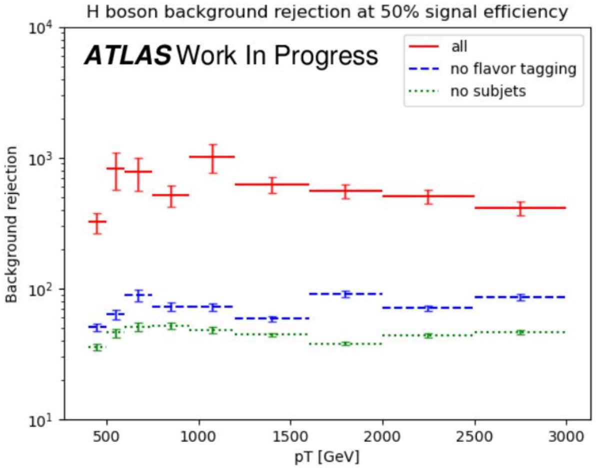




Impact of the subjet features on performance

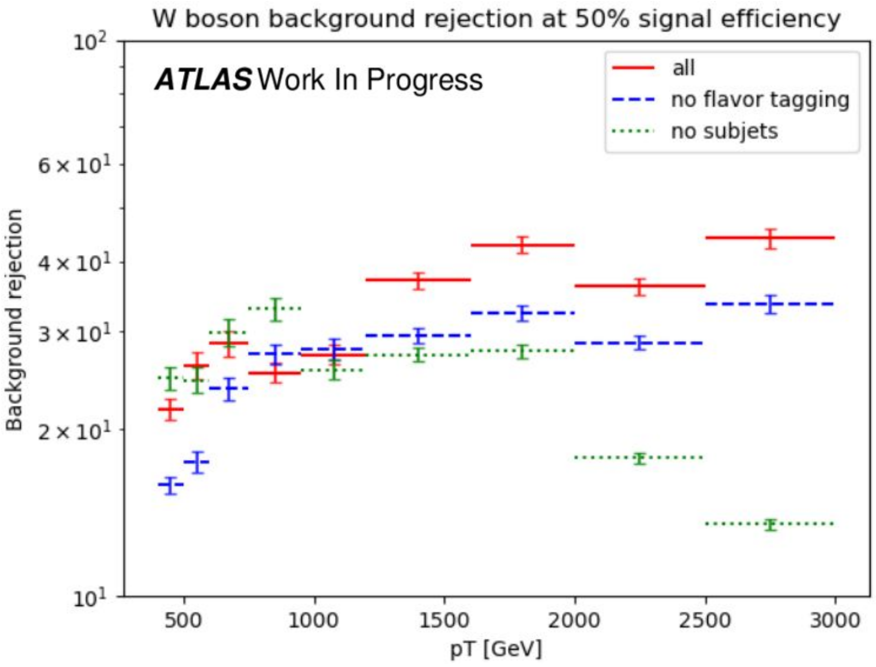
The subjet features, especially their flavour tag, are a novel input feature for large-R taggers, therefore their impact on the performance was studied explicitly.

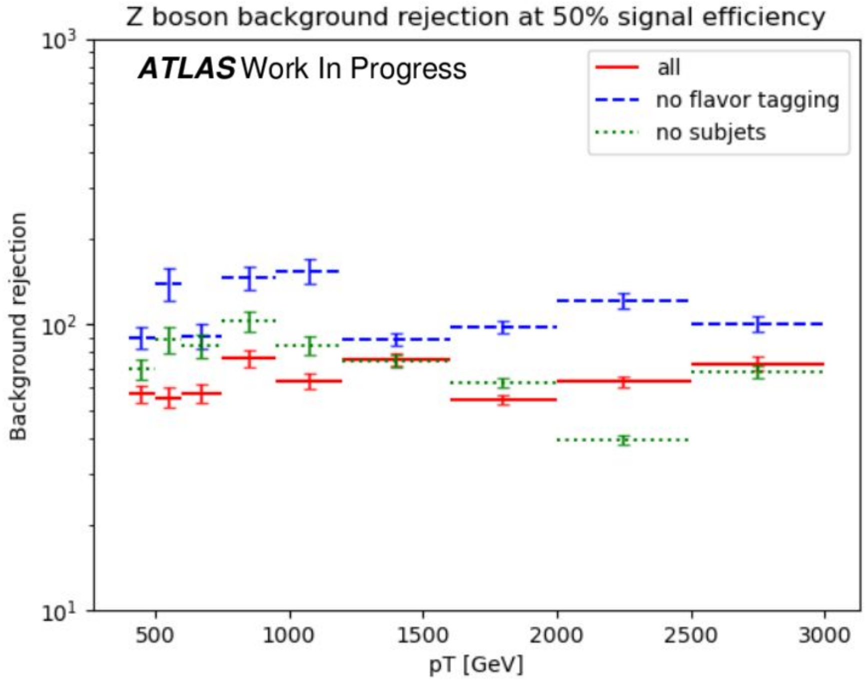




Higgs samples only consist of H→bb events, therefore big impact

Impact of the subjet features on performance



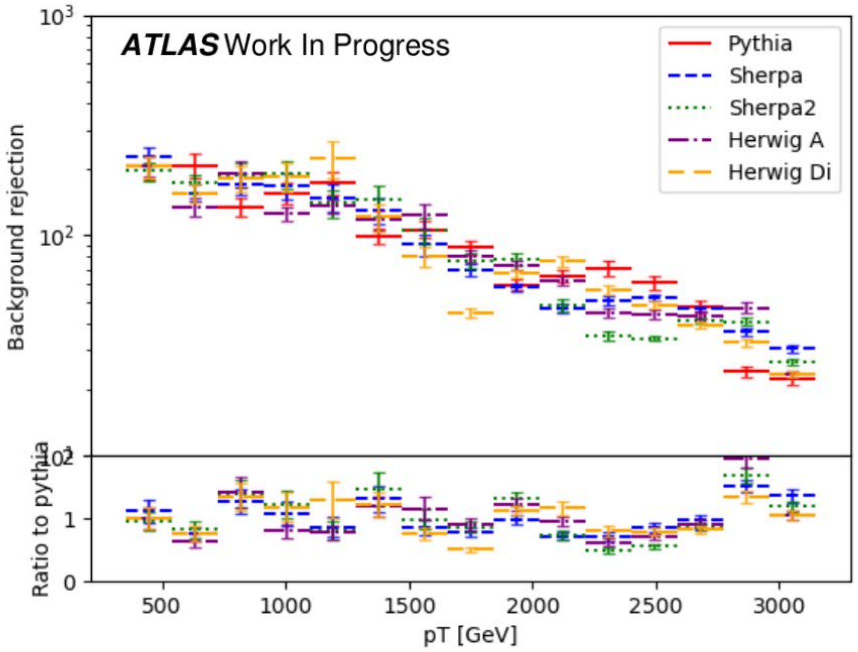


Z samples are mixed: Z→bb/cc/qq, therefore flavor tag can add confusion

Does the network learn physics or MC modelling?

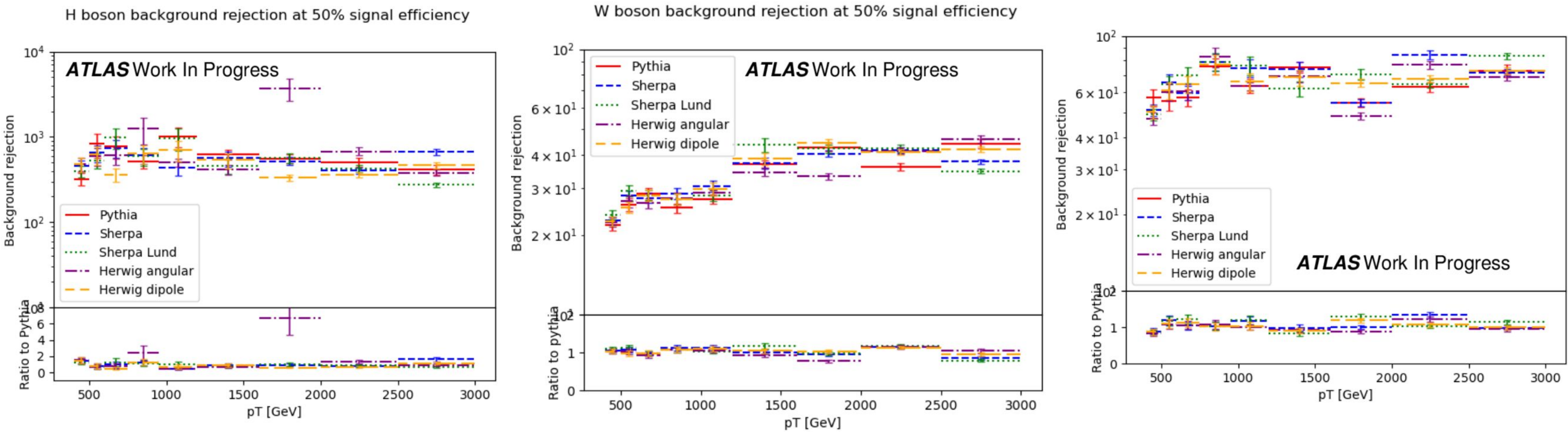
- ★ We cannot test performance on real data because we don't have labels
- ★ Best available test:
  - ○ train on samples from one MC generator
  - ○ test on samples from other MC generators which use different hadron showering algorithms

Top quark background rejection at 50% signal efficiency

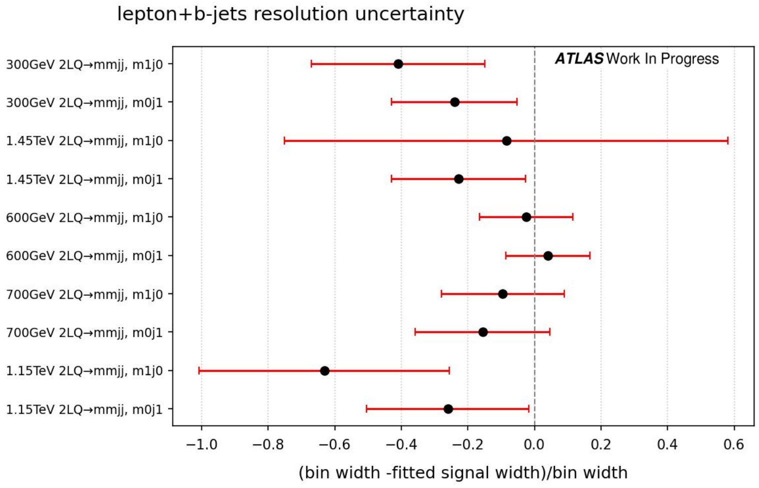


Does the network learn physics or MC modelling?

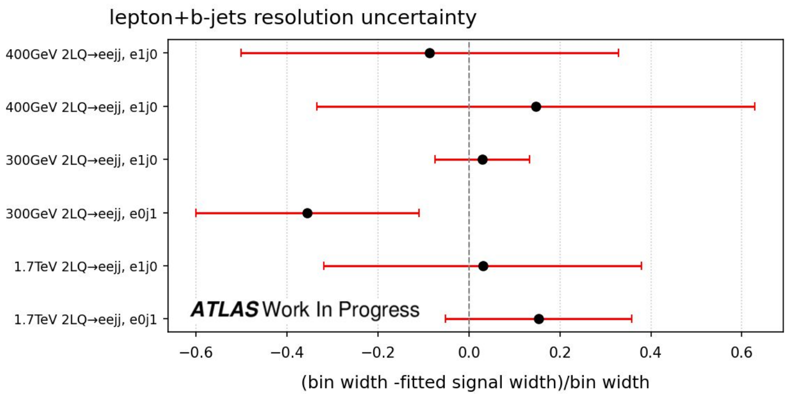
No clear bias towards any generator can be observed



Validation of bin sizes using BSM signal simulations



LQ → mu + b



LQ → e + b

Resolution tends to be slightly underestimated, usually within 0-60% of signal width