Machine learning (ML) activities in ATLAS

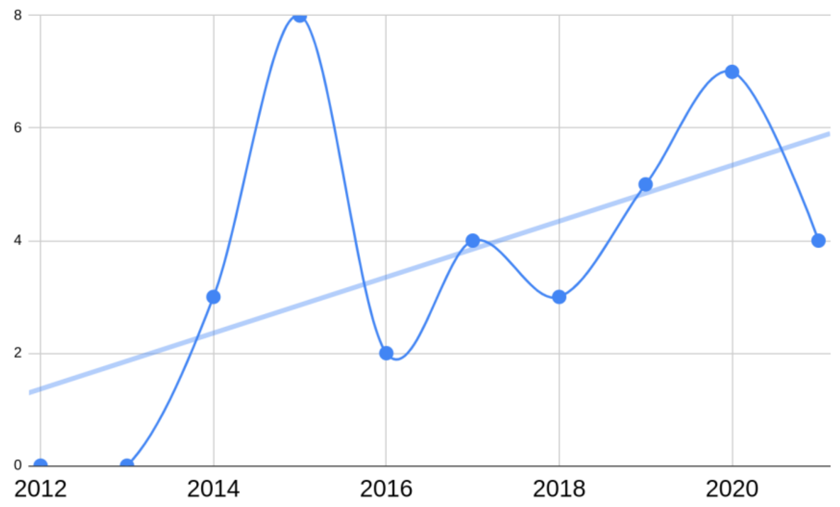
Steffen Mæland

Assoc. prof., HVL

NORCC

Overview

- «Traditional» multivariate analysis techniques well established
- Shift towards more modern, computationally intensive methods
  - Nontrivial implications for both workflow and interpretation
- Shift from in-house software frameworks (TMVA, etc) to external ones
- ***Different approaches***:
  - ML as «drop-in» replacement for cut-based analysis
  - ML operating on low-level data, w/o a traditional counterpart
  - Unsupervised ML, learning abstract representations rather than labels



ATLAS journal papers tagged ML/MVA

Applications

- Event classification
  - Diverse portfolio of BSM searches (*see Physics session*)
  - Both signal vs background and multiclass approaches

Analyses with standard or nonstandard input data types

- Object reconstruction and identification
  - Tau reconstruction
  - Jet identification

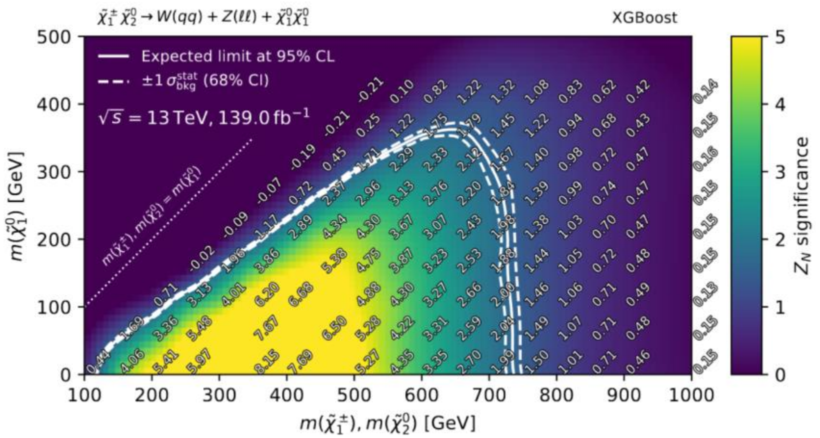
Physics performance

- Regression & anomaly detection

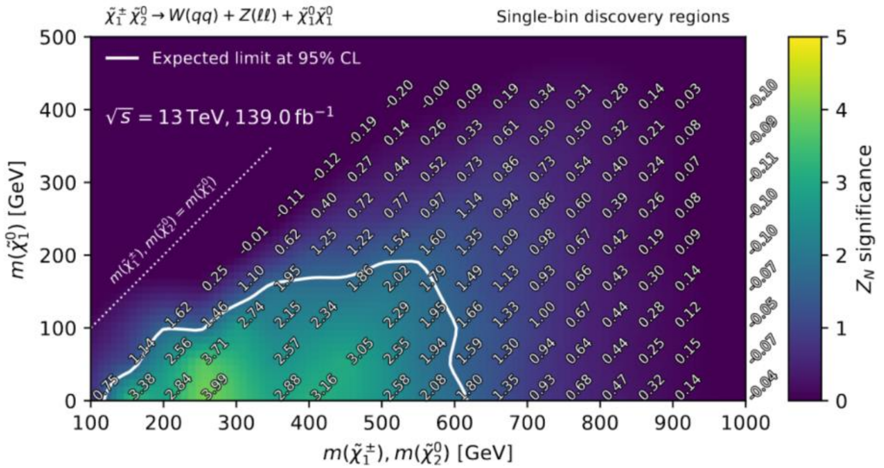
Analyses with nonstandard outputs

Applications: Event classification

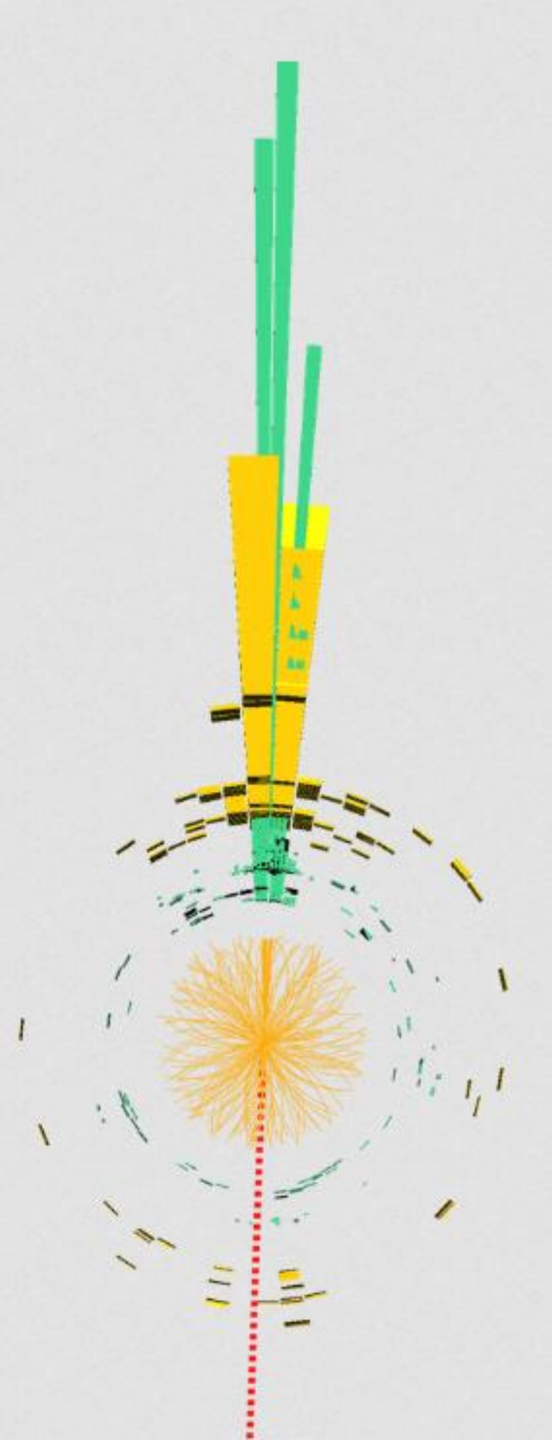
- Boosted decision trees (BDTs) common replacement for cut-based (CB) analysis
  - XGBoost most popular, backed by open-source I/O and data manipulation libraries (`UpROOT, Pandas, Awkward Array`)
  - Typical performance relationship:
    Single-bin CB < multi-bin CB ≈ single-bin BDT
- Neural networks (NNs) gaining traction, but performance on columnar data often not outweighing technical complexity
  - Most published results have used shallow (<3 layers) feed-forward architectures



BDT



Cut-based



Rethinking event classification

Several avenues being explored:

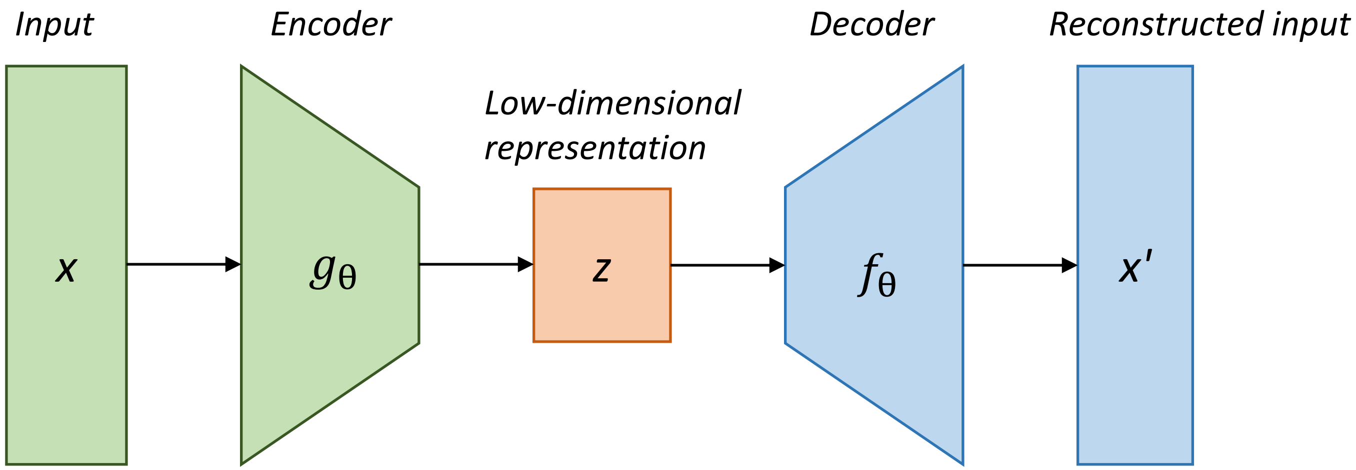
- New methods and architectures applied to columnar data
    - *Discovery significance (or exclusion limits) used as the performance metric*
- Unsupervised approaches
    - *Anomaly detection with autoencoder NNs*
- Representing parts of, or the entire detector, as an image
    - *Opens for modern image recognition methods*

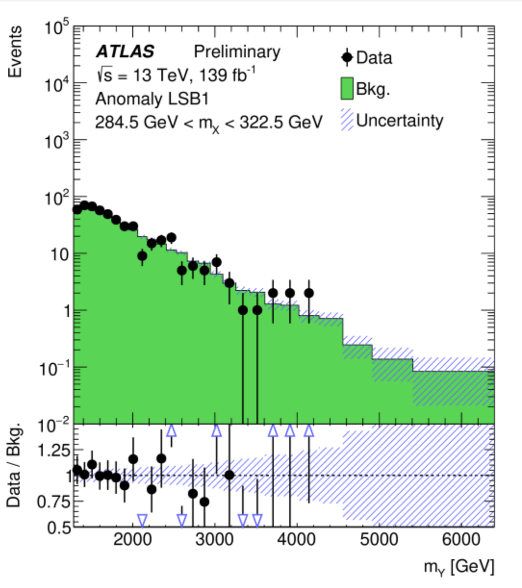
Unsupervised learning

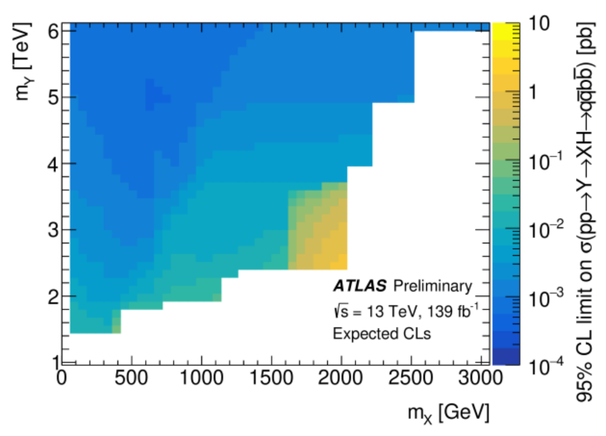
Anomaly detection:

Reframe BSM detection task

- SM background constitutes the majority of data
  - Model trained to reduce dimensionality, then reconstructs its input
- BSM signal is the ‘anomaly’
  - Unknown to trained model, yields high reconstruction error



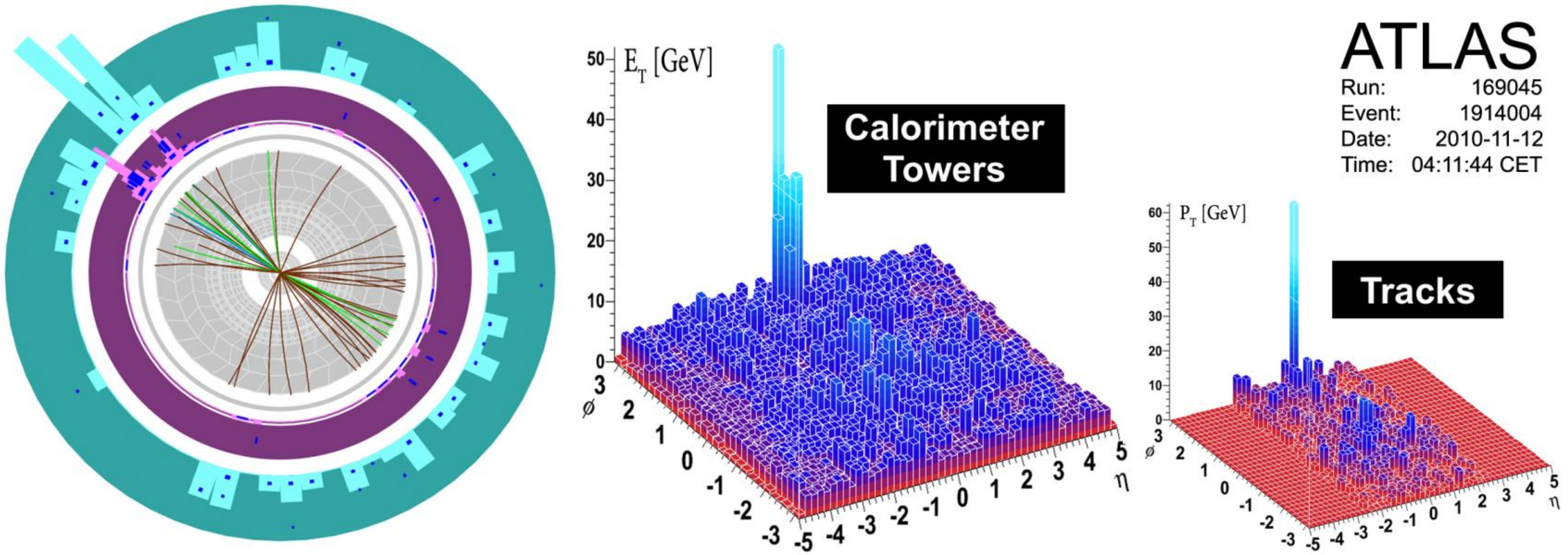




ATLAS-CONF-2022-045

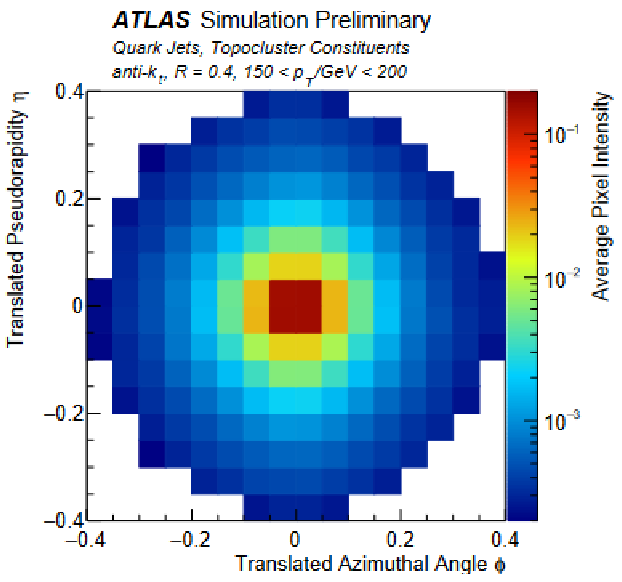
ATLAS as a camera

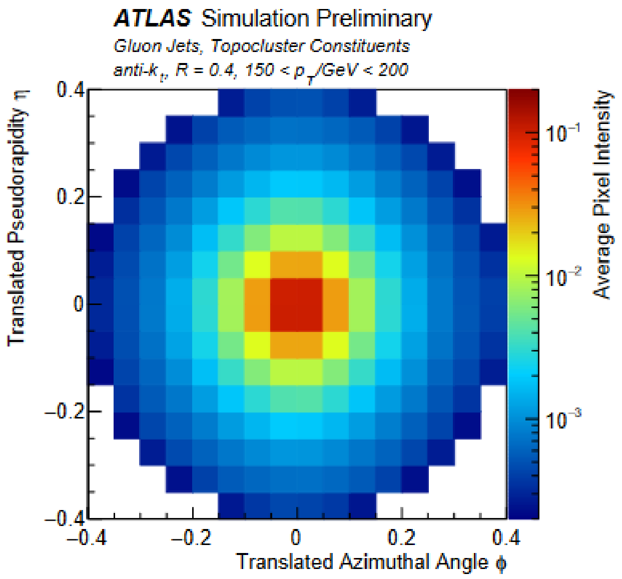
- Calorimeter response in eta-phi represents a 2D (or 3D) image
- Can tap into modern computer vision techniques, while skipping or replacing reconstruction algorithms
- Two different scopes:
  - ***Local*** – e.g. neighborhood around a jet / object level
  - ***Global*** – entire detector / event level

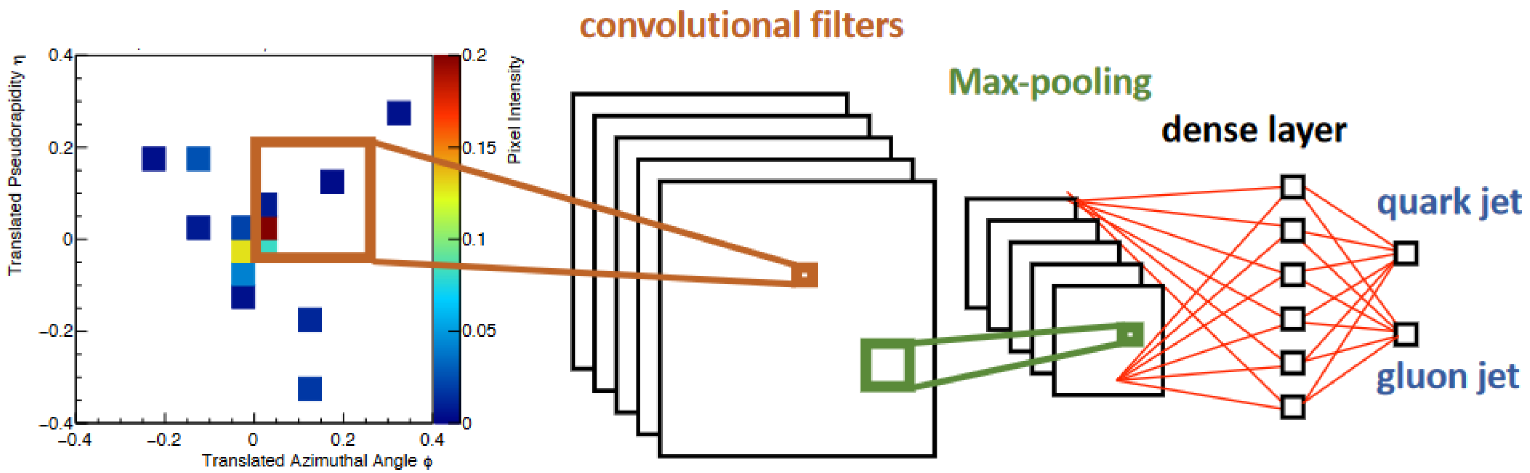


Jet image classification

- Use convolutional neural networks (CNNs) to classify jets by shape
- Can use sophisticated pre-trained models
  - Weights optimised on separate data -> less training data required
  - Jet images typically sparse, so transfer learning may not be a good solution?

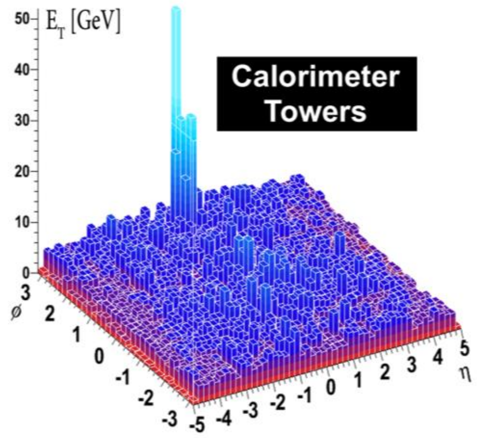


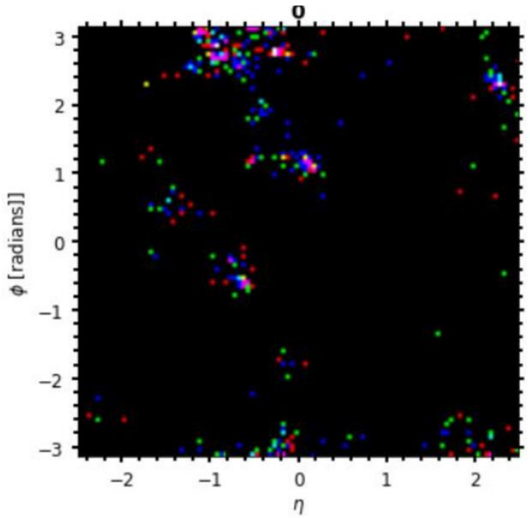


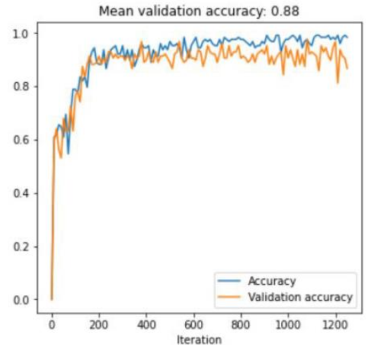


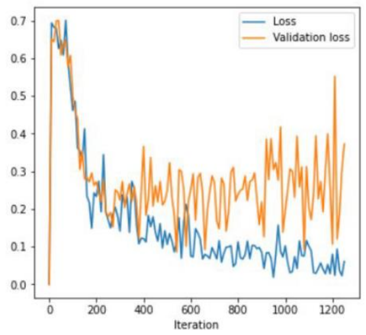
Event image classification

- Represent calorimeter cells as image pixels
  - Extendable by adding tracks, calo depth
- Image recognition models well established, but:
  - Data wraps around φ
  - Sparse images
  - Nontrivial spatial structure
- Loads of interesting extensions:
  - Segmentation → object identification
  - Generative networks :o





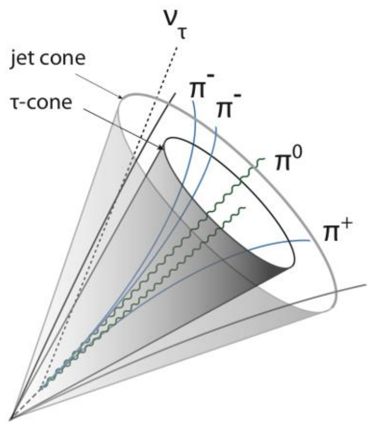


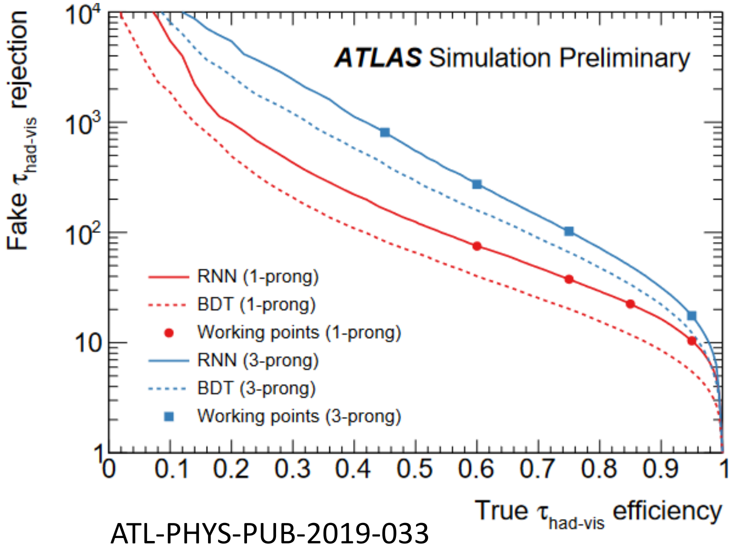


ML-driven tau reconstruction

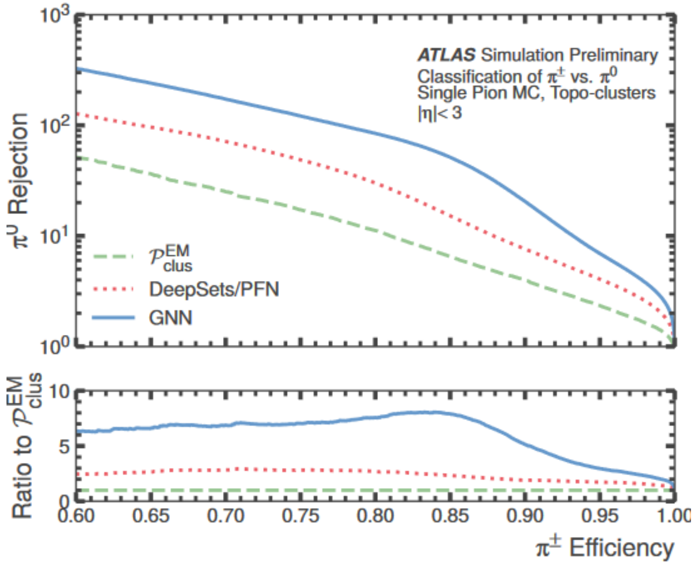
- *ML success story*: tau identification
- Improved jet rejection by use of a recurrent neural network (RNN)
    - Single model can take an arbitrary number of inputs:

      1-prong / 3-prong / X neutrals
- Pion identification improved by graph neural networks
- Strategy milestone commitments RE-4.1 and RE-4.2





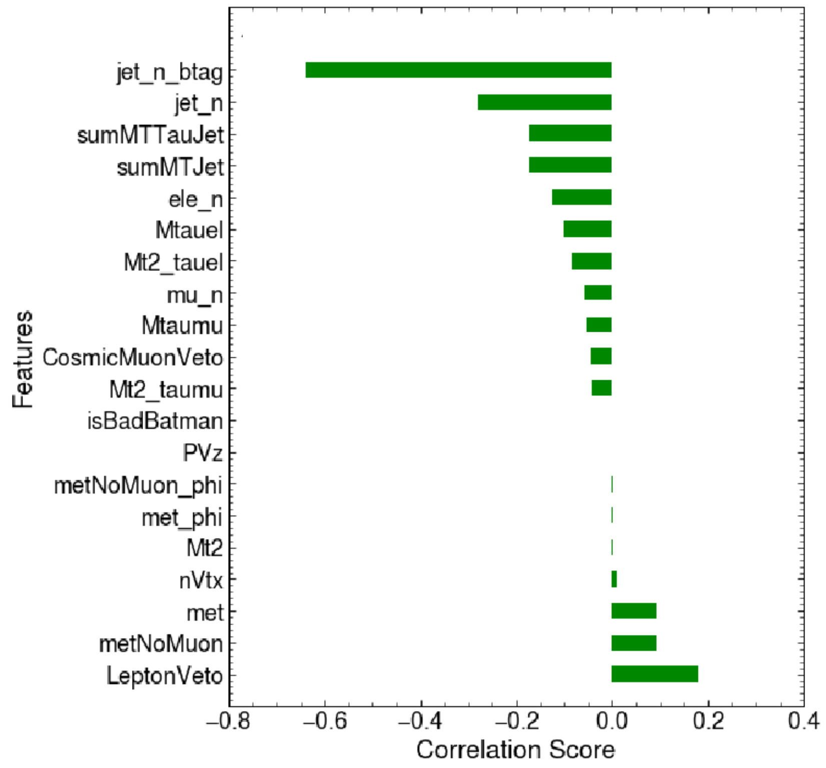
ATL-PHYS-PUB-2019-033



ATL-PHYS-PUB-2022-040

Interpretation, robustness & explainability

- Some models are more understandable than others
  - Decision trees offer feature ranking
  - No equivalent for NNs, but approaches exist (e.g. from game theory)
- Yet all models are influenced by distribution shifts, reconstruction errors, etc
  - Nonlinear nature of modern ML may yield nontrivial response to such shifts
- Increased interest in model interpretation, both statistical and on per-event level



Framework development

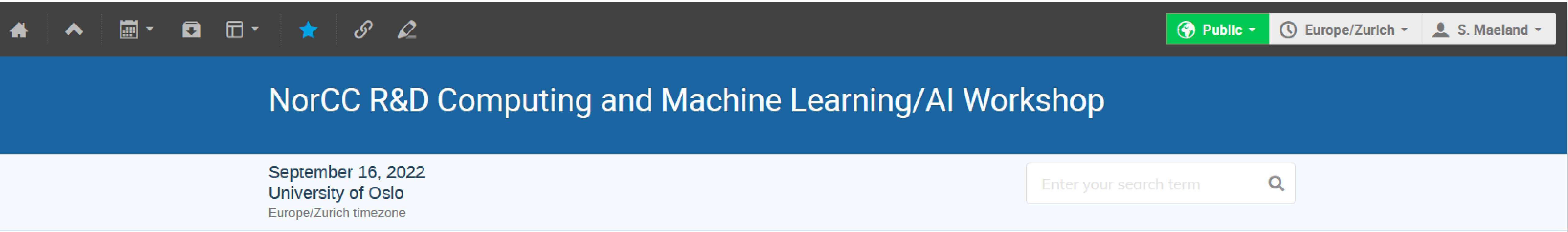
- External ML software forces use of external data formats
  - Conversion adds to work and turnaround time
  - Data formats often suboptimal for detector data
- Considerable work done on format interoperability
  - ROOT's `RDataFrame` intended to replace external columnar formats
  - Interoperable with xAODs w/o ATLAS libraries
  - Preparations for Run4 PHYSLITE format ongoing

```
#Create dataframe from PHYSLITE
df = ROOT.RDataFrame("CollectionTree", "DAOD_PHYSLITE.stringNoTrig.pool.root")
```

```
# Apply cuts
# Two muons with pt > 20 GeV
df = df.Filter('AnalysisMuonsAuxDyn.pt.size() == 2', 'Exactly two muons')
df = df.Filter('AnalysisMuonsAuxDyn.pt[0] > 20000 && AnalysisMuonsAuxDyn.pt[1] > 20000', 'Muons have pt > 20 GeV')
```

Collaboration

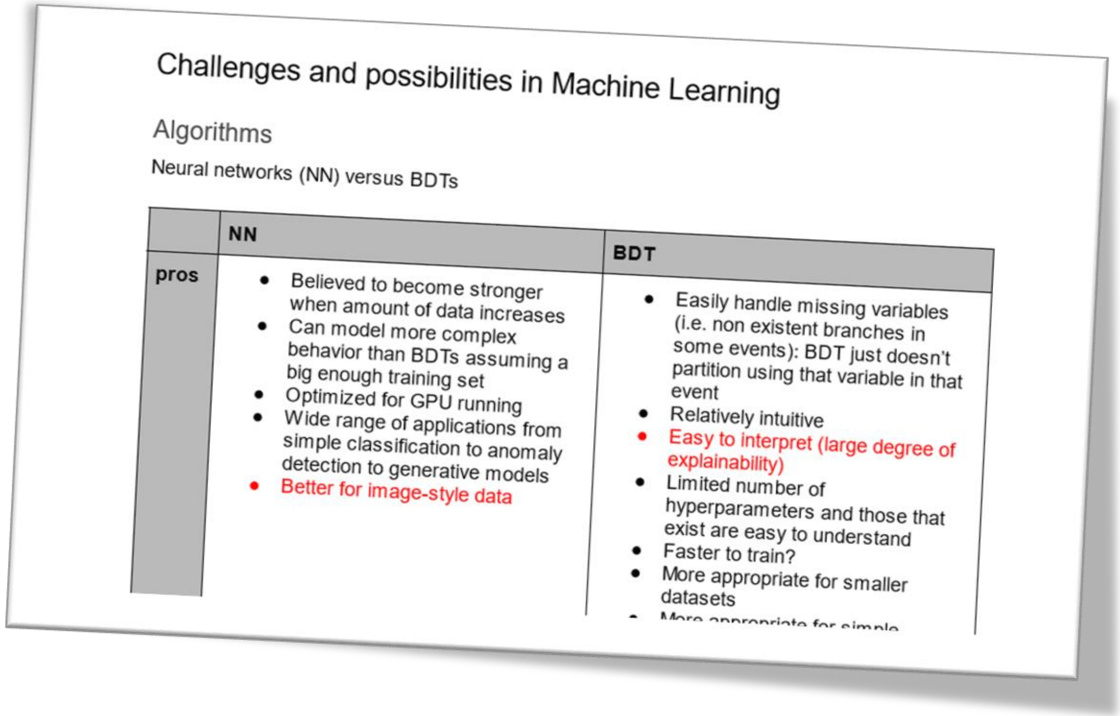
- R&D ML/AI network holds workshops twice yearly
  - Last one in May 2022
  - Next one tomorrow
- ATLAS Machine Learning Forum: semiweekly meetings
- Inter-experimental ML Working Group (IML): monthly meetings
- Institute-level collaboration w/ computing



Collaboration on technical challenges

- Lack of industry standards leads analysis teams to do redundant work on ML challenges
  - Choice of model/architecture
  - Choice of baseline hyperparameters
  - Treating jagged arrays
  - … and more

- Intiative started at prev. workshop to form recommendations
  - Intended as general starting points
  - Potentially reducing not only work/GPU cycles, but also erroneous results



Challenges and possibilities in Machine Learning

Algorithms

Neural networks (NN) versus BDTs

	NN	BDT
pros	• Believed to become stronger when amount of data increases • Can model more complex behavior than BDTs assuming a big enough training set • Optimized for GPU running • Wide range of applications from simple classification to anomaly detection to generative models • Better for image-style data	• Easily handle missing variables (i.e. non existent branches in some events): BDT just doesn't partition using that variable in that event • Relatively intuitive • Easy to interpret (large degree of explainability) • Limited number of hyperparameters and those that exist are easy to understand • Faster to train? • More appropriate for smaller datasets • More appropriate for simple

R&D ML/AI workshop tomorrow:
https://indico.cern.ch/event/1152542

NORCC