Machine Learning and JetMET towards 100 fb^{-1}

Steven Schramm

JetMET Workshop - Physics at 100 fb^{-1}

Helsinki, Finland

May 10, 2017

Introduction

UNIVERSITÉ DE GENÈVE

- Machine Learning (ML) is a field of growing interest in HEP
- This is likely to continue as we approach the $100\,\mathrm{fb}^{-1}$ regime
  - ML is aimed at extracting information from large datasets
- While one can predict growing interest, the future is never clear
- This is my *personal* view on where the field will go
  - I have my biases, but will try to be as fair as possible

- My institute (Université de Genève) is very active in deep learning
- I have been dedicated to jetMET trigger, perf, analysis since 2012
  - I've done much more for jets than $\mathrm{E_T^{miss}}$, but have a background in both
- I am from ATLAS, not CMS
  - I will only be speaking about public results
  - Some parts may already be well-known within CMS

Introduction to the IML

UNIVERSITÉ DE GENÈVE

- The Inter-experimental Machine Learning (IML) group [website]
  - An LPCC group recognized by all four LHC experiments
  - Five coordinators: ALICE, ATLAS (me), CMS, LHCb, LPCC/EP-SFT
  - Frequent participation from non-LHC experiments, ML, and industry
- Full scope is defined in the mandate
  - General idea: enhance communication between HEP and ML communities, share ML expertise between experiments, support software development and maintenance, organize ML tutorials, and provide a forum for general ML discussions relating to HEP needs
- The group hosts ∼monthly meetings and an annual workshop
  - More than 450 people signed up to the group mailing list
  - The first workshop was in March, with almost 300 registrants
  - Next meeting - May 24: deep learning, workshop challenge follow-up

What does $100\,\mathrm{fb}^{-1}$ mean for ML?

UNIVERSITÉ DE GENÈVE

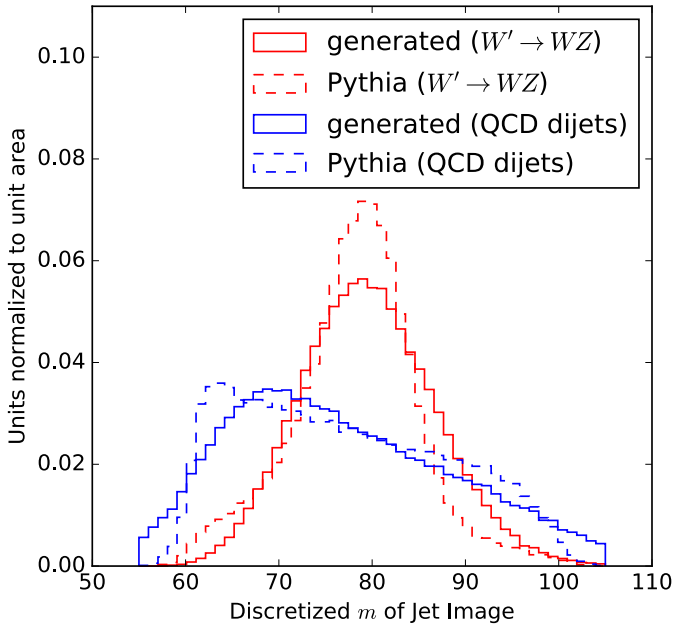
1. Larger datasets means that more MC is needed to have data-like stats
   - Simulation costs will continue to rise
2.
3.
4.

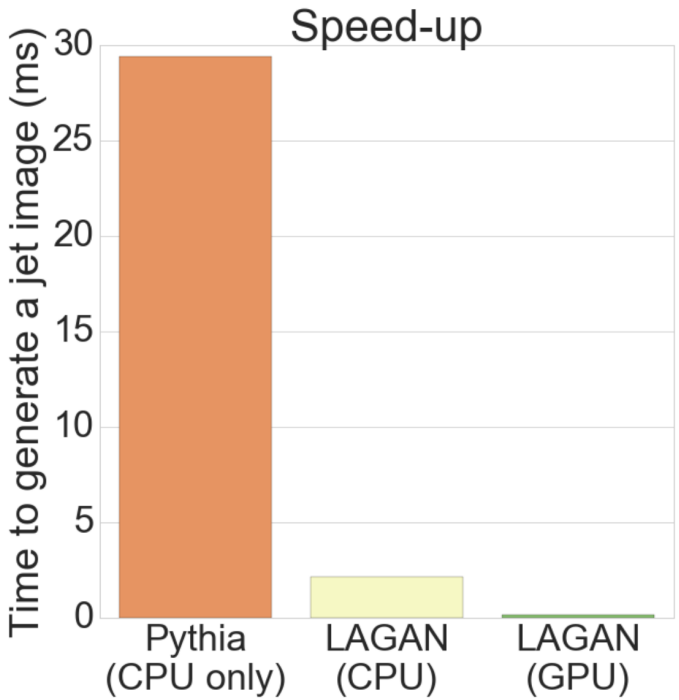
1. Generative adversarial networks

[Paganini et al]

UNIVERSITÉ DE GENÈVE

- Idea: two networks - a generator and an adversary
  - The first tries to turn random values into a desired result (image, etc)
  - The second compares the output of the first and real results
  - Train until the second cannot differentiate generated from real results
- Example of jet images generation below from IML workshop



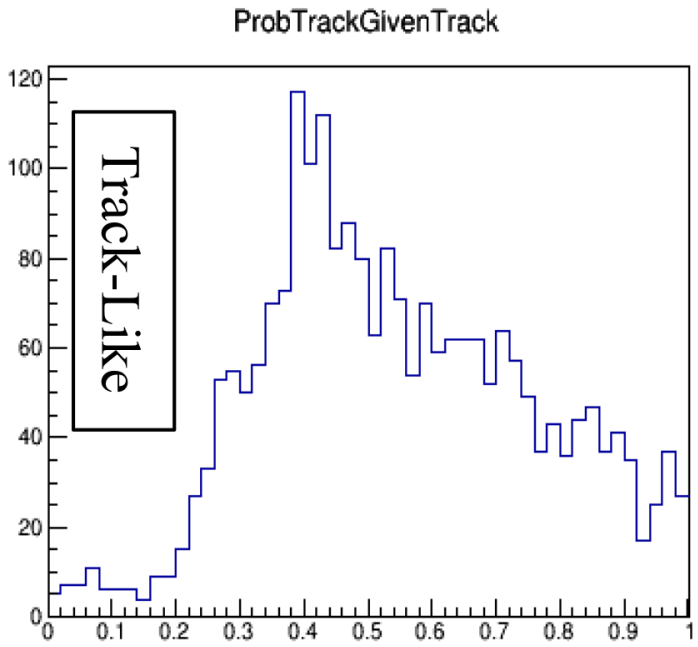


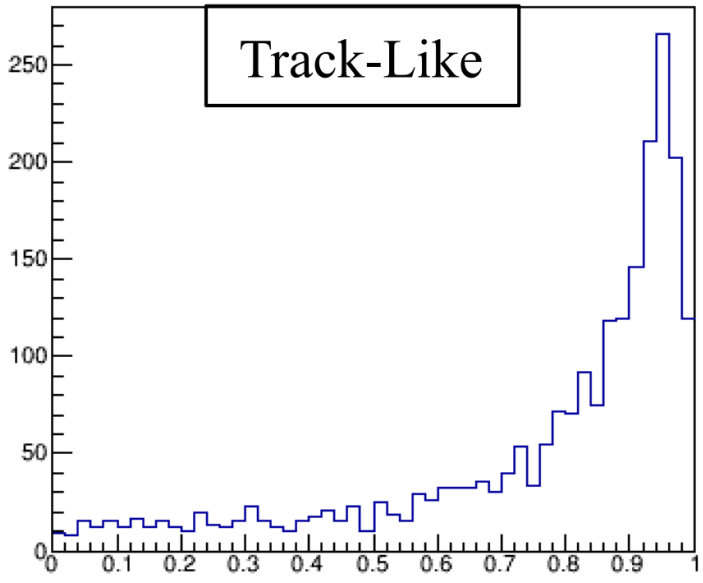
1. Generative adversarial networks

[Smith et al]

UNIVERSITÉ DE GENÈVE

- Idea: two networks - a generator and an adversary
  - The first tries to turn random values into a desired result (image, etc)
  - The second compares the output of the first and real results
  - Train until the second cannot differentiate generated from real results
- Example of fixing LArIAT MC/data differences from last IML meeting
  - Track-like probability given a track input, left is before, right is after





What does $100\,\mathrm{fb}^{-1}$ mean for ML?

UNIVERSITÉ DE GENÈVE

1. Larger datasets means that more MC is needed to have data-like stats
   - Simulation costs will continue to rise
   - Possible solution: generative adversarial networks
2. Larger datasets means higher precision and better taggers
   - Small data/MC and MC1/MC2 differences become more important
3.
4.

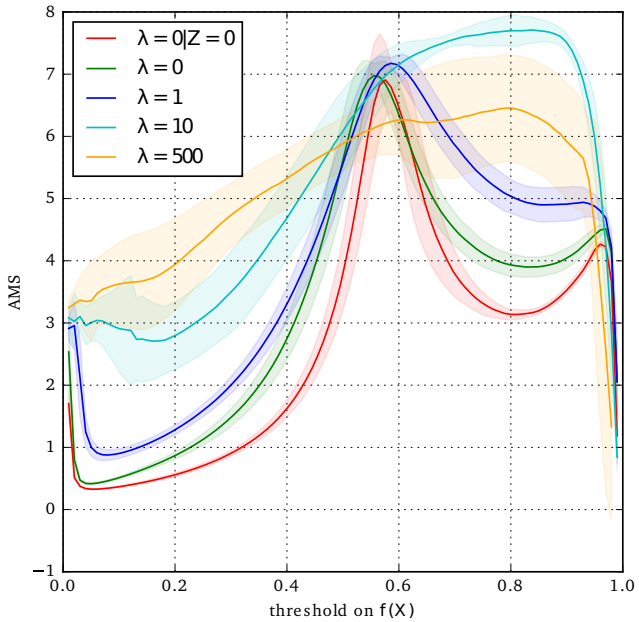
2. Adversarial networks

[Louppe et al]

UNIVERSITÉ DE GENÈVE

- Similar approach to generative adversarial networks
  - Train classifier 1 as usual (b-tagger, etc)
  - Train classifier 2 to predict a given parameter using output of #1
- If the second classifier can predict the parameter, then first classifier is dependent on that parameter. We can block/tweak this.

- Example: reduce the impact of systematics (Z), aka data/MC
- $\lambda = 0|Z = 0$: no systematics
- $\lambda = 0$: no adversary
- $\lambda = 10$: trade precision for robustness, results in net gain of statistical significance



What does 100 fb^{-1} mean for ML?

UNIVERSITÉ DE GENÈVE

1. Larger datasets means that more MC is needed to have data-like stats
   - Simulation costs will continue to rise
   - Possible solution: generative adversarial networks
2. Larger datasets means higher precision and better taggers
   - Small data/MC and MC1/MC2 differences become more important
   - Possible solution: adversarial networks
3. Larger datasets mean that we have more well-known events
   - More data-driven control regions for evaluating taggers
4.

3. Low-level taggers

UNIVERSITÉ DE GENÈVE

- Most taggers so far have dealt with *high-level* variables
  - This excludes flavour tagging, where it's a bit different
- High-level variables are easier to understand
  - This is usually true both conceptually and in terms of uncertainties
  - Example: jet mass, number of matched tracks, n-subjettiness, etc
- What if we just throw all PFlow objects into a network?
  - This would then be a low-level tagger
  - These are typically difficult to validate, but can improve performance
- Large datasets will support these low-level taggers
  - They need *lots* of data to find all useful features
  - They need to be evaluated in well-understood regions, like $t\bar{t}$
- Some examples follow later in this talk

What does 100 fb^{-1} mean for ML?

UNIVERSITÉ DE GENÈVE

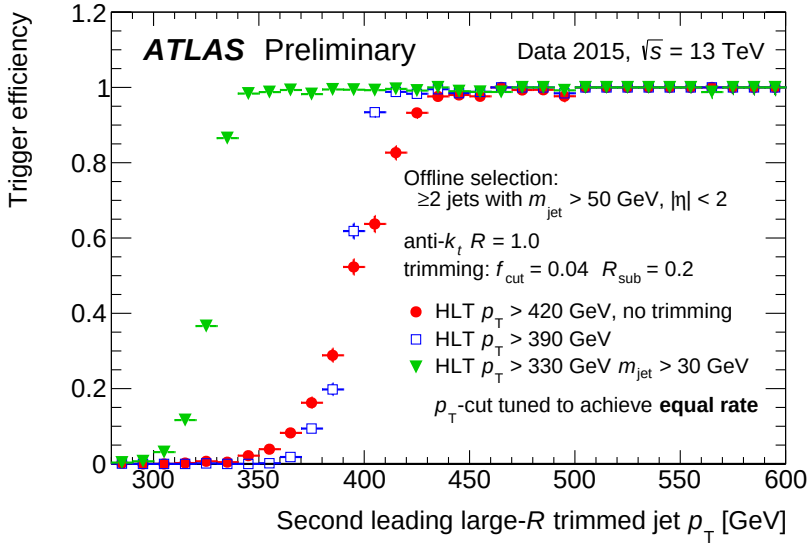
1. Larger datasets means that more MC is needed to have data-like stats
   - Simulation costs will continue to rise
   - Possible solution: generative adversarial networks
2. Larger datasets means higher precision and better taggers
   - Small data/MC and MC1/MC2 differences become more important
   - Possible solution: adversarial networks
3. Larger datasets mean that we have more well-known events
   - More data-driven control regions for evaluating taggers
   - Possible benefit: increased ability to validate low-level taggers in data
4. Higher inst. luminosity makes it harder to keep all interesting events
   - Simple p_{T} cuts and similar may no longer be desired

4. ML in the trigger

[ATLAS] UNIVERSITÉ DE GENÈVE

- ML is already used within the trigger, but this is likely to grow
  - LHCb in particular has done a lot of work in this direction
- Anomaly detection can help to reduce rate lost to detector problems
  - See IML meeting on anomaly detection for some examples
- May be useful to train classifiers to *reject* the main process
  - Dijets is a perfect example, as it hides all hadronic decays
  - Build anti-dijets classifier with very high "generic signal" efficiency
  - Split rate between recording dijets and "anti-dijets"

- Non-ML approach: mass cut
- Already done in ATLAS
- Lines are all for equal rates
- Easily extensible to ML
  - Should be much better!



What does $100\,\mathrm{fb}^{-1}$ mean for ML?

UNIVERSITÉ DE GENÈVE

1. Larger datasets means that more MC is needed to have data-like stats
   - Simulation costs will continue to rise
   - Possible solution: generative adversarial networks
2. Larger datasets means higher precision and better taggers
   - Small data/MC and MC1/MC2 differences become more important
   - Possible solution: adversarial networks
3. Larger datasets mean that we have more well-known events
   - More data-driven control regions for evaluating taggers
   - Possible benefit: increased ability to validate low-level taggers in data
4. Higher inst. luminosity makes it harder to keep all interesting events
   - Simple p_{T} cuts and similar may no longer be desired
   - Possible solution: more ML in the trigger

What else may be useful in the future to JetMET

UNIVERSITÉ DE GENÈVE

- Deep Learning and other more recent tagging techniques
  - Jet substructure taggers are becoming increasingly diverse
- Multi-class classification
  - Flavour tagging and jet substructure are clear targets
- ML calibrations
  - Jet calibrations in particular are a natural application
- Hardware considerations
  - CPUs and GPUs

- Note on $\mathrm{E_T^{miss}}$: for ATLAS, I can think of direct uses for ML
  - Our ambiguity resolver could benefit from multiple of the above
  - However, I have limited knowledge of CMS $\mathrm{E_T^{miss}}$ reconstruction
    - Likely some of these methods can be applied on your $\mathrm{E_T^{miss}}$ inputs

What is Deep Learning?

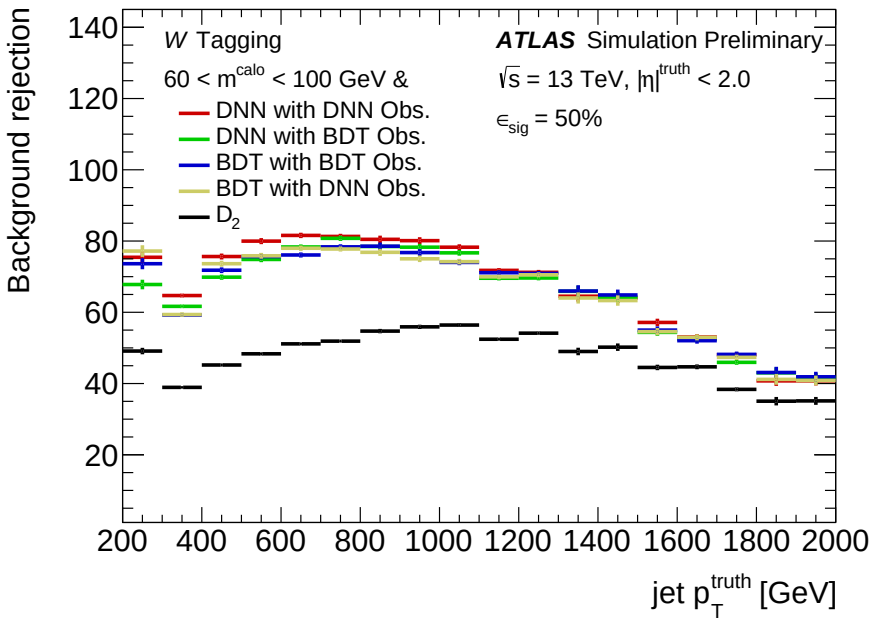
UNIVERSITÉ DE GENÈVE

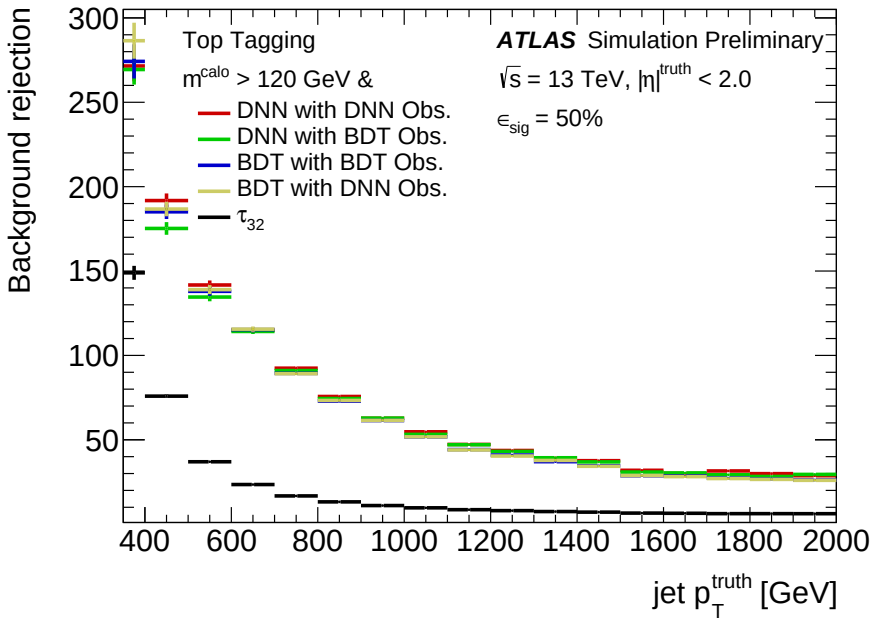
- Nice intro by M. Schwartz this morning in the context of q/g tagging
- Deep Learning (DL) is mostly just a fancy name for a Neural Network (NN) with multiple hidden layers
  - DL has become a bit of a buzz-phrase
  - It is not a magical solution, and is often not even the best choice
  - In pursuit of DL, shallow NNs have been significantly improved
- Idea: any transformation can be built from a set of nonlinear functions
  - However, this requires $\sim \infty$ functions and $\sim \infty$ training data
  - We already have hard examples of this (next slide)
- In general, deep networks provide enhanced non-linear discrimination
  - Typically comes at the cost of larger training sets and longer training
  - Also more susceptible to over-training (requires regularization)

Deep Learning vs BDTs, high-level

[ATLAS] UNIVERSITÉ DE GENÈVE

- Recent ATLAS publication on DNNs vs BDTs for W and top tagging
- BDTs and DNNs both performed well compared to simple taggers
  - Not surprising: more variables and non-linear correlations
- BDTs and DNNs perform ∼identically (this study, high-level features)
- BDT easier to train, easier to understand, provide same discrimination
  - BDT evaluation time is ∼ 10× slower [T. Nitta]
  - Likely BDT is still the "better" choice, unless it's used in the trigger

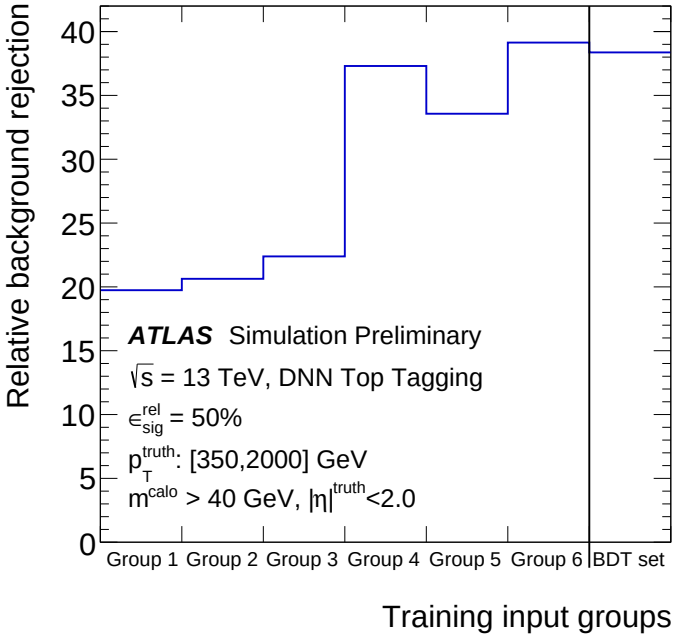




A comment on DL/DNNs

[ATLAS] UNIVERSITÉ DE GENÈVE

- This is an example of top-tagging DNN training
- Focus on groups 4, 5, 6
  4. τ_1, τ_2, τ_3, ECF{1,2,3}, ...
  5. {Group 4} - ECF{1,2,3} + $\{\tau_{21}, \tau_{32}, C_2^{\beta=1}, D_2^{\beta=1}\}$
  6. Superset of 4 and 5



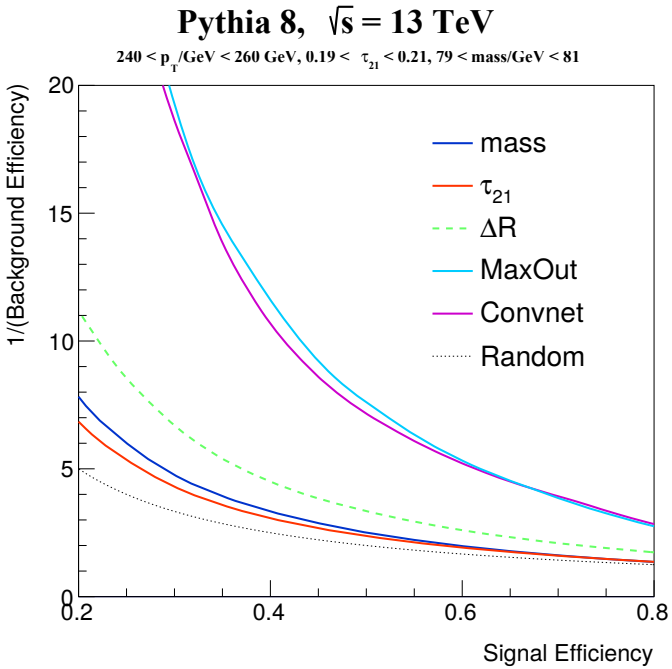
- DNN learned more from the input variables (4) than their ratios (5)
  - DL can learn more from lower-level variables
    - Not surprising in this case, given what ECF{1,2,3} are
  - Higher-level variables have lost information
- However, it didn't learn everything, as (6) did a bit better
  - DL is not perfect, this is not an infinite sum of non-linear functions

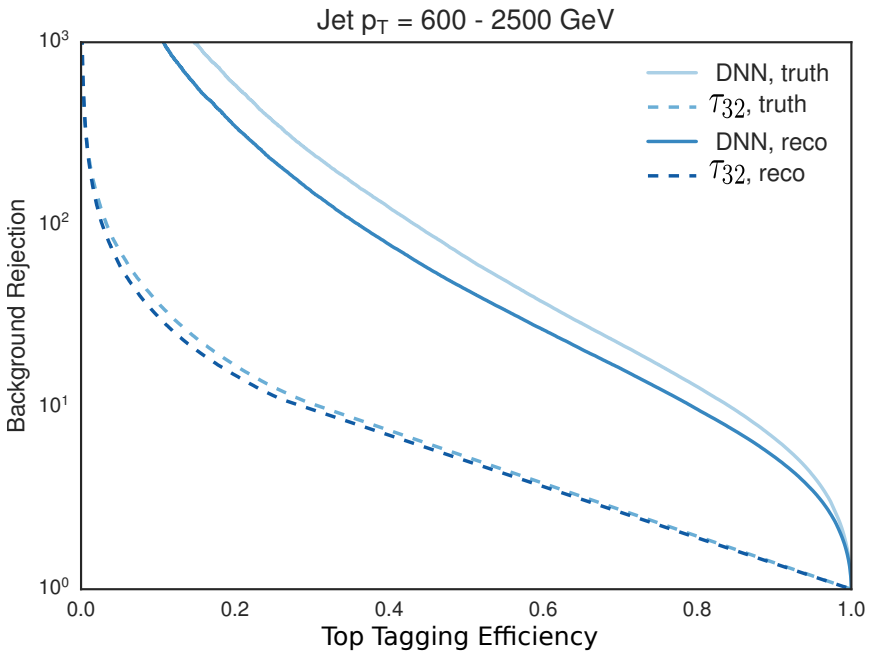
Deep Learning, low-level

[Oliveira et al, Pearkes et al]

UNIVERSITÉ DE GENÈVE

- Jet images (left)
  - Build a calorimeter grid and apply standard DNN image recognition
- Sequences of jet constituents (right)
  - Build DNNs using p_{T}, η, ϕ of leading 120 constituents
  - If there are fewer than 120 constituents, then zero-pad the list
- Two promising ways to go beyond ML trained on high level vars

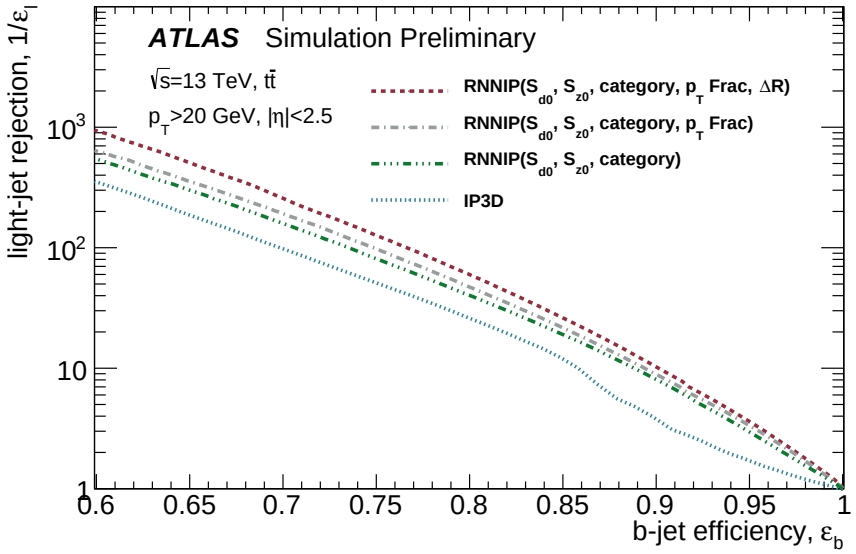


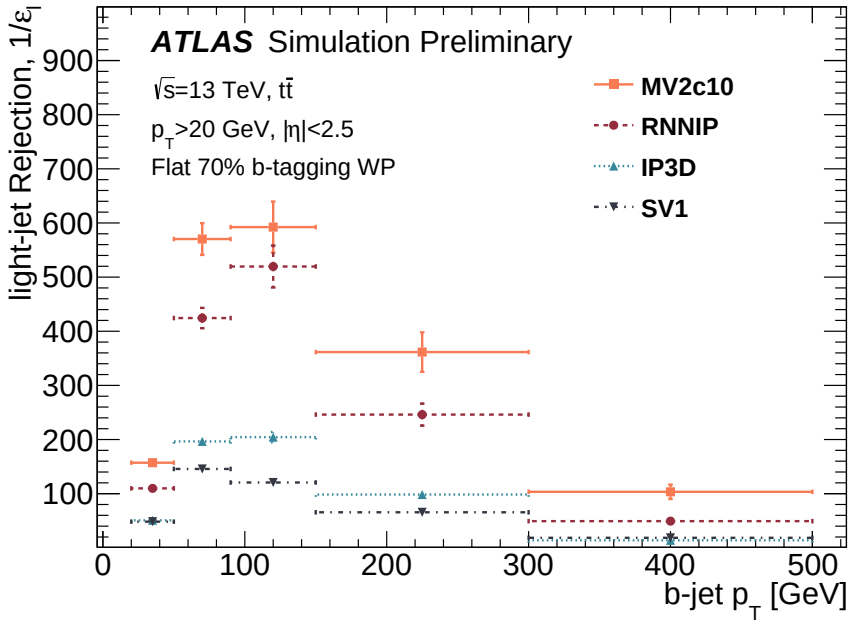


Recurrent neural networks (RNNs)

[ATLAS] UNIVERSITÉ DE GENÈVE

- RNNs are a type of DNN with directed cycles
  - They can thus be applied to (for example) dynamic numbers of tracks
- Comparing the RNN to IP3D (likelihood tagger)
  - With same variables, RNN learns more
  - Adding more variables increases the power further, as expected
- RNNs can be used in more contexts, ie dynamic number of jet inputs





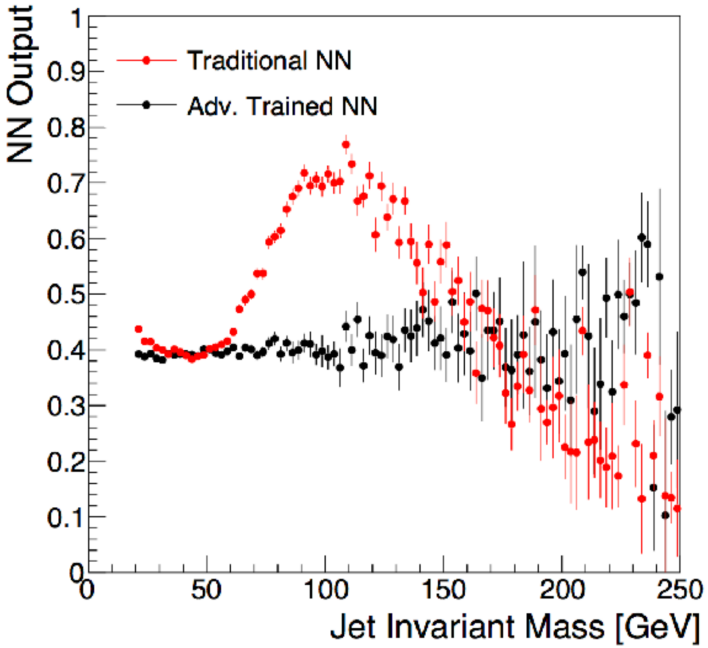
Decorrelating taggers using ML

[Shimmin et al]

UNIVERSITÉ DE GENÈVE

- Again, making use of adversarial networks
  - Train classifier 1 to reject QCD (for example)
  - Train classifier 2 to predict jet mass using the output of classifier 1
  - If this is possible, penalize classifier 1, we don't want to sculpt the mass

- Similar intent to DDT for τ_{21}
- More complex than DDT, so easier to use DDT if possible
- Want to extend to any variable or combinations of variables
  - Same method will work



Multi-class tagging

UNIVERSITÉ DE GENÈVE

- Multi-class tagging means allowing for more than a yes/no answer
- Example from RNNIP: light, charm, b, or tau (4-class output)
  - When calculating ROC curves, the tau fraction was set to zero
  - However, training the network to understand more categories still helps!
- The next talk (DeepFlavour) is another great example
  - I don't want to spoil the results, so I won't provide any plots
- The same could be applied to hadronic decays
  - Multi-class outputs for: W, Z, H, top, q, g

ML calibrations

UNIVERSITÉ DE GENÈVE

- Calibrations (at least in ATLAS) depend on standard variables
  - E, η, N_{tracks}, etc
  - In each case, binned numerical inversions are performed
- ML can help: minimize the difference from truth p_{T}
  - Exploits correlations, handles many variables
- Need to be careful to not pick up on MC features
  - Many shower variables are not modelled as well as we'd like
  - Recall adversarial networks from earlier, this can help
- Needs sufficient validation in data control regions
  - Standard Z/γ+jet balance, multi-jet balance, dijet balance, etc
- Lots of work, but should be able to reduce jet uncertainties
  - From this morning, it looks like you already may be doing so

CPUs and GPUs

UNIVERSITÉ DE GENÈVE

- CPUs are great, but GPUs are better in the land of Deep Learning
  - Typical DNNs use enormous number of floating point operations
  - For ML purposes, single-point precision is more than sufficient
    - Cheap GPUs are better than expensive CPUs
    - Expensive DL-oriented GPUs are of course still the best, but are pricey
- Comparison for a high-performance DNN: [M. Lanfermann]
  - Standard GPU (Titan X Pascal), not DL-dedicated GPU
  - $\sim$5M training jets, $\sim$1M validation jets, $\sim$6M testing jets
  - 3.5 min/epoch (GPU) vs 16+ min/epoch (CPU), $\sim 5\times$ speedup
  - Performance will vary with usage, but this trend is the standard
- Other considerations: architecture
  - Interesting talk at last IML meeting by IBM
  - IBM added a direct CPU-GPU link at $5\times$ the speed of PCIe for NVIDIA
  - Impressive benchmarks, maybe the future of large-scale DL

Summary

UNIVERSITÉ DE GENÈVE

- Generative and adversarial networks are increasingly popular
  - Huge simulation speed-ups, make simulation look more like data, systematic-aware classifiers, de-correlate taggers and jet mass, ...
- Low-level deep learning is increasingly used for tagging
  - Jet images, DNNs with zero-padding, RNNs with dynamic inputs, ...
  - Given sufficient information, deep learning can often pick out features we don't yet understand but which do provide better discrimination
    - See talk by M. Schwartz this morning for an example
  - Don't blindly use DNNs, with high-level vars BDTs often just as good
- To remain at the forefront of ML in HEP, you're probably going to need to invest in GPUs (either at the institute or collaboration level)
- ML is likely to continue to become more relevant and more powerful
  - We will definitely benefit from ML at $100\,\mathrm{fb}^{-1}$
  - We may strictly require it at $1000\,\mathrm{fb}^{-1}$