Machine Learning and Multiclassifiers for Improved Measurements of 2-Lepton Final States in the Higgs to *WW* Decay Channel in High Energy Physics Analyses

Virtual CAP Congress 2020
Particle Physics Session – June 9, 2020

Matthew Basso[1] (University of Toronto)

With help from lots of people!

1 – mbasso@physics.utoronto.ca

UNIVERSITY OF TORONTO

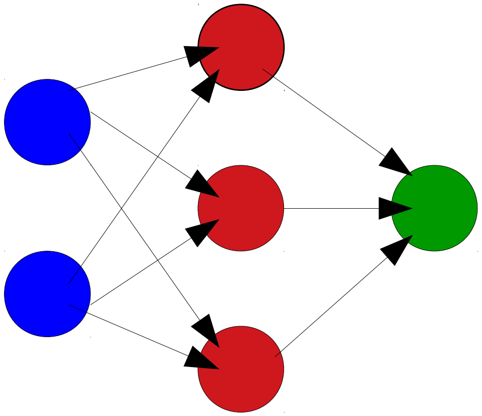
NSERC CRSNG

Canadian Association of Physicists

Association canadienne des physiciens et physiciennes

Overview

- <u>Introduction</u>:
  - Physics problem: measuring (*V->qq*)(*H->WW->lνlν*)
    - Rare process with large background contributions
- <u>Machine learning</u>:
  - Motivation with a discussion of neural networks in particular
  - Training and implementation: Keras+TensorFlow in Python
- <u>Results</u>:
  - Performance and validation
  - Binning optimization and regularization algorithm for statistical significance



Introduction and Multiclassifier Development

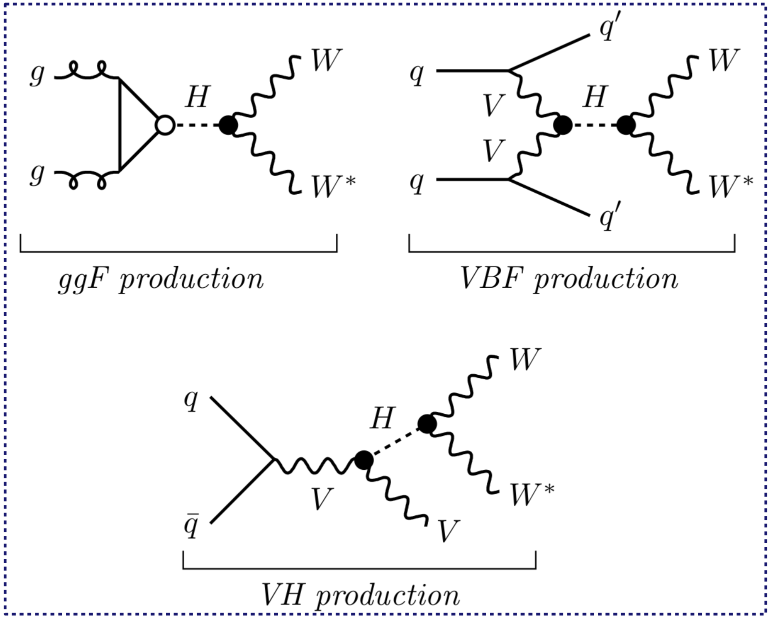
Introduction

- Discovery of scalar consistent with Higgs boson in 2012 by ATLAS and CMS at the LHC was a critical test of the Standard Model (SM) [1,2]
  - Important to verify the SM in individual decay channels, e.g., *H->ZZ*, *H->WW*, *H->γγ*, etc. – deviations would point to Beyond the SM (BSM) physics
- Here, we consider *H->WW->lνlν*
  - *H->WW* 2nd highest branching fraction at $\sqrt{s}$ = 13 TeV (after *H->bb*)
  - Can't reconstruct Higgs mass due to missing transverse energy (MET) carried away by neutrinos
  - Observed (using threshold in *p*-value of 5σ) using a combination of the gluon-gluon fusion (ggF), vector boson fusion (VBF), and associated (VH) production channels in ATLAS Run 1 [3] and ggF+VBF in ATLAS Early Run 2 [4]

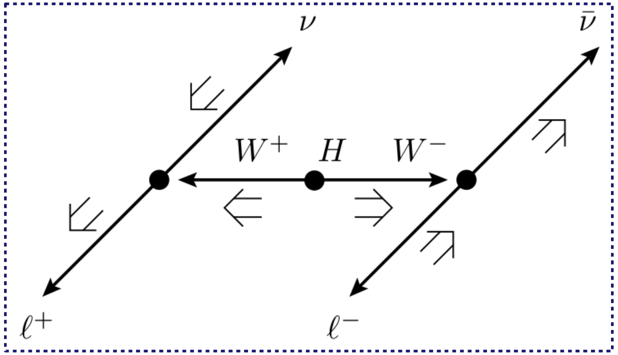
Introduction (2)

- Measurements of *V*(*H*->*WW*) (abbr. VH) have been made independently of ggF+VBF HWW for Run 1 [5] and Early Run 2 [6]
  - Never reached 5σ "observation" – may be possible with 139 ifb of data from LHC Full Run 2
- Consider 2-lepton VH channel: *V*(->*qq*)*H*(->*lνlν*)
  - Looking for 2 different-flavour, opposite-sign (DFOS) leptons + MET and 2 jets
  - Small signal (cross section ~100, ~10 times smaller than ggF, VBF) with large backgrounds:
    - Top: top-antitop quark pair production (ttbar), single top production (*Wt*)
    - Drell-Yan or Z+jets: *Z*->$\tau\tau$ in association with jets where the taus decay to DFOS leptons
    - Diboson *WW*: irreducible background matching signal decay
  - Analyzed in Run 1 using a cut-based analysis in 1 signal bin [5] – **can we use machine learning to perform a better measurement (using Run 1 cuts as the baseline)?**

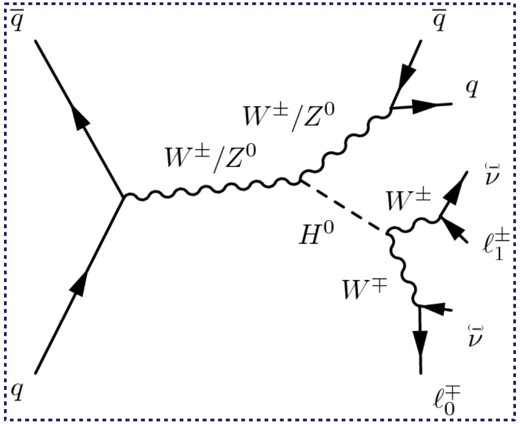
Relevant Feynman Diagrams



Higgs production modes considered. Obtained from Fig. 1 of Ref. [3].



Direction/spin of particles in HWW decay. Obtained from Fig. 3 of Ref. [3].



DFOS VH channel. Obtained from Fig. 1 of Ref. [5].

Machine Learning and Neural Nets

- Multivariate analysis (MVA) techniques such as machine learning (ML) have seen widespread use in high-energy physics
  - ML: attempts optimize a set of free parameters (“training”) to best map inputs to desired outputs (i.e., supervised learning)
  - Exploits correlations between input variables in ways (rectangular) cut-based analyses cannot
- <u>Neural nets (NNs)</u>: map input vectors to output vectors via a series linear matrix operations (“layers”) with (possibly) nonlinear functions applied to the output of each (“activations”)
  - $\mathbf{x}' = f(\mathbf{A}.\mathbf{x} + \mathbf{b})$ for a single layer, where f(...) is applied element-by-element
  - Free parameters: choice of f(...), kernel (matrix) **A**, bias (vector) **b**

Multiclassifiers and Samples

- Multiclassifiers: MVAs mapping inputs to >2 output classes (e.g., physics processes) – natural choice for our problem!
  - For *N* output classes, get *N* output discriminants describing the probability of an event belonging to each class (Σoutputs = 1)
- Balance to be struck between too many classes and too few
  - Signal-like: *V*(*qq*)*H*(*lνlν*), ggF(*lνlν*), VBF(*lνlν*)
  - Background-like: top (ttbar + *Wt*), Z+jets, *WW*
    - **Exclude** subleading backgrounds: other *VV*, W+jets (“fakes”), ...
- Use ATLAS Run 2 (2015-18) Monte Carlo (MC) samples for the processes of interest as well as Run 2 data for validation (**work-in-progress!**)

Input Variable Definitions

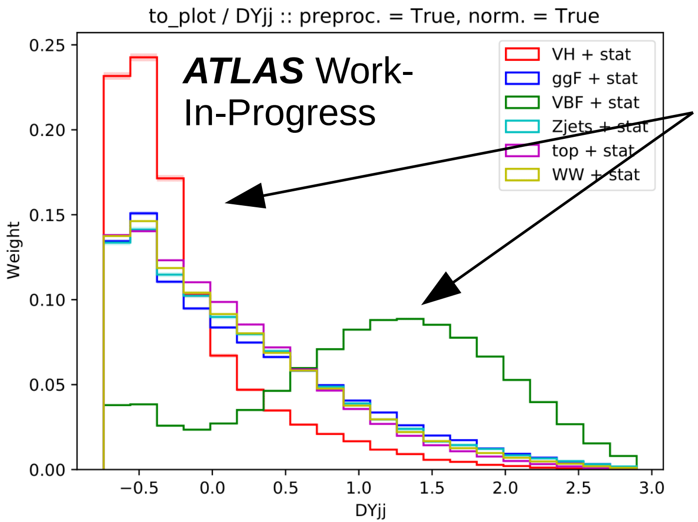
- ***Choice of inputs is important!*** – based partly on DFOS VH Run 1 cuts [5] and Early Run 2 VBF MVA variables [4]
  - Lepton variables:
    - Leading/subleading lepton p_T's, dilepton angular separation $\Delta\varphi_{ll}$, rapidity difference ΔY_{ll}, and mass M_{ll}
  - Jet/MET variables:
    - Leading/subleading jet p_T's, dijet angular separation $\Delta\varphi_{jj}$, rapidity difference ΔY_{jj}, and mass M_{jj} ($|M_{jj} - 85$ GeV$|$)
    - Tau-tau mass using collinear approximation $M_{\tau\tau}$ ($|M_{\tau\tau} - M_Z|$, M_Z = mass of Z boson) [7]
    - Transverse mass $M_T = \sqrt{((E_T^{ll} + E_T^{miss})^2 - |\mathbf{p}_T^{ll} + \mathbf{E}_T^{miss}|^2)}$
    - Track-based MET $E_T^{miss,track}$
    - Sum of all p_T-hard objects + soft (track+calorimeter) contributions H_T^{soft} [8]
    - MET-based significance $E_T^{miss} / \sigma(E_T^{miss})$ [8]
    - Sum of lepton/jet p_T's + track-based MET $\Sigma p_T^{total,track}$ and all lepton-jet mass combinations ΣM_{lj}

17 in total – could maybe be reduced by removing highly-correlated variables

Preprocessing and Training

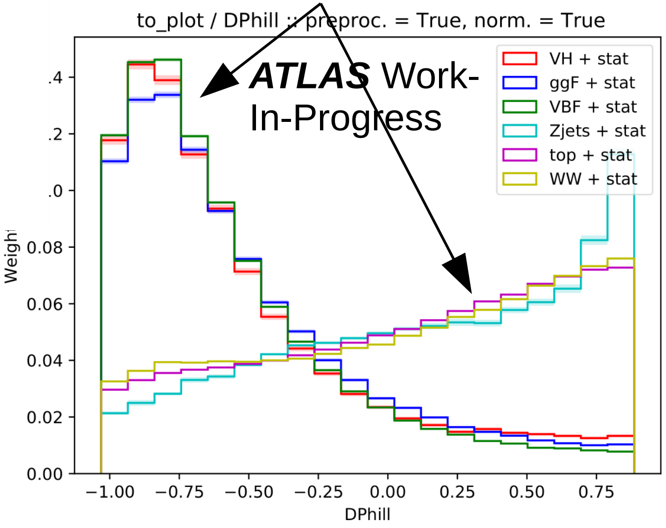
- Apply preselection (2 DFOS leptons with min p_T, M_{ll} cuts), $N_{jets} \geq 2$, and $N_{b\text{-}jets} = 0$ (to reject top) to our samples
  - Raw yields: 127K VH, 115K ggF, 390K VBF, 2.00M top, 998K Z+jets, 1.59M *WW*
- Preprocessing performed with the help of scikit-learn [9]
  - Median is subtracted from input variables and scaled to interquartile range (robust scaling)
  - Sum of weights/class is scaled to 1 (to account for differing raw yields)
  - Gaussian noise applied to inputs and throughout network to minimize overtraining
- Use Keras [10] with the TensorFlow [11] backend for training
  - Hyperparameters (# of nodes/layer, learning rate, activation function) optimized using Ray [12] + Tune [13] with space defined using Hyperopt [14] (see Backup)
  - Final training is performed using 80-20% train-test splitting with 5-way K-folding (see Backup)

Example Shape Distributions



ΔY_{jj} is **good discriminator** for VH and VBF

$\Delta\varphi_{ll}$ is a **good discriminator** for Higgs signal vs. background



All input shape distributions included in Backup!

- Shapes:
  - *x*-axis is median scaled (*unitless*)
  - Sum of weights per class is normalized to 1 (to show shapes)

Optimized Network Structure

- Optimization metric: validation receiver-operator characteristic (ROC) curve area (AUC) – plots signal efficiency vs. background rejection (ideally AUC ~ 1)
- Network yielding **highest** ROC AUC (~0.87):
  - # of nodes/layer (10 in total) = 20, 20, 40, 70, 20, 50, 40, 30, 40, 30
  - Activation function = exponential linear unit (ELU)
  - Learning rate (i.e., gradient descent step size) = 0.00222

***ATLAS* work-in-progress**

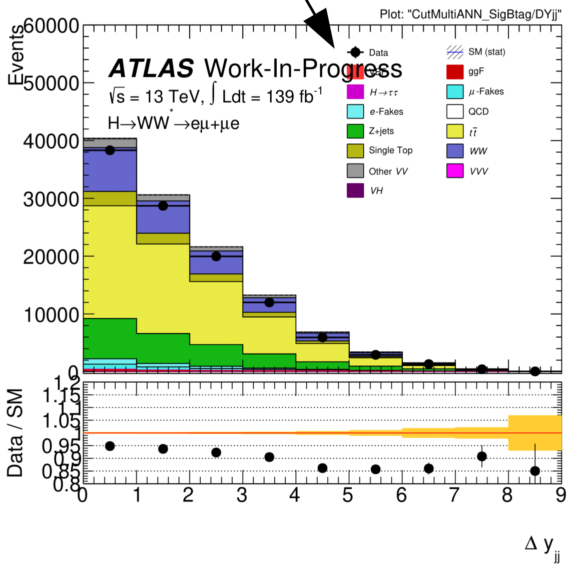
- Other (*static*) network parameters/features:
  - Batch size = 1024 and optimizer Adam
  - Categorical crossentropy loss function (typical for multiclassifiers)
  - Softmax output activation (typical for multiclassifiers, ensures Σoutputs = 1)
  - Batch normalization at the output of each layer (to guard against large output values at specific nodes), early stopping during training (to guard against overtraining)

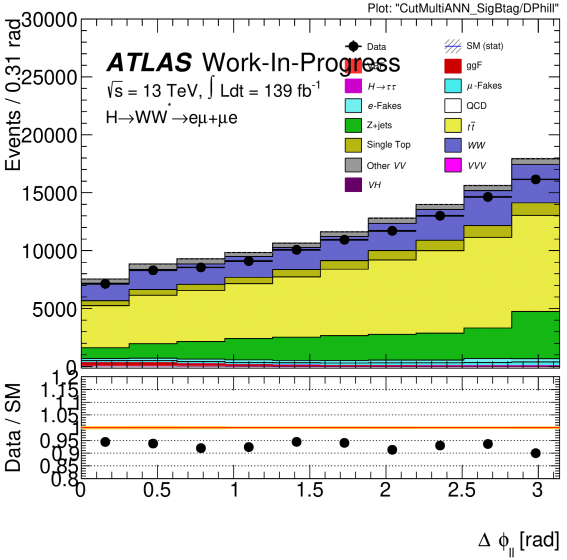
Multiclassifier Training/Validation Metrics and Performance

Data-MC Comparisons @ Level of Inputs

- See all plots in the Backup
  - Include additional processes: *ZZ*, *WZ*, *V*+*γ*, W+jets, ...
- Generally **OK** (0.9-0.95) data/MC agreement
  - Disagreement is **not** completely understood, but it's also not a showstopper
  - We know we have imperfect Z+jets modelling
    - Potentially fixed by applying a normalization factor derived in a Z+jets control region (e.g., see Run 1 [5])
    - Not applied here

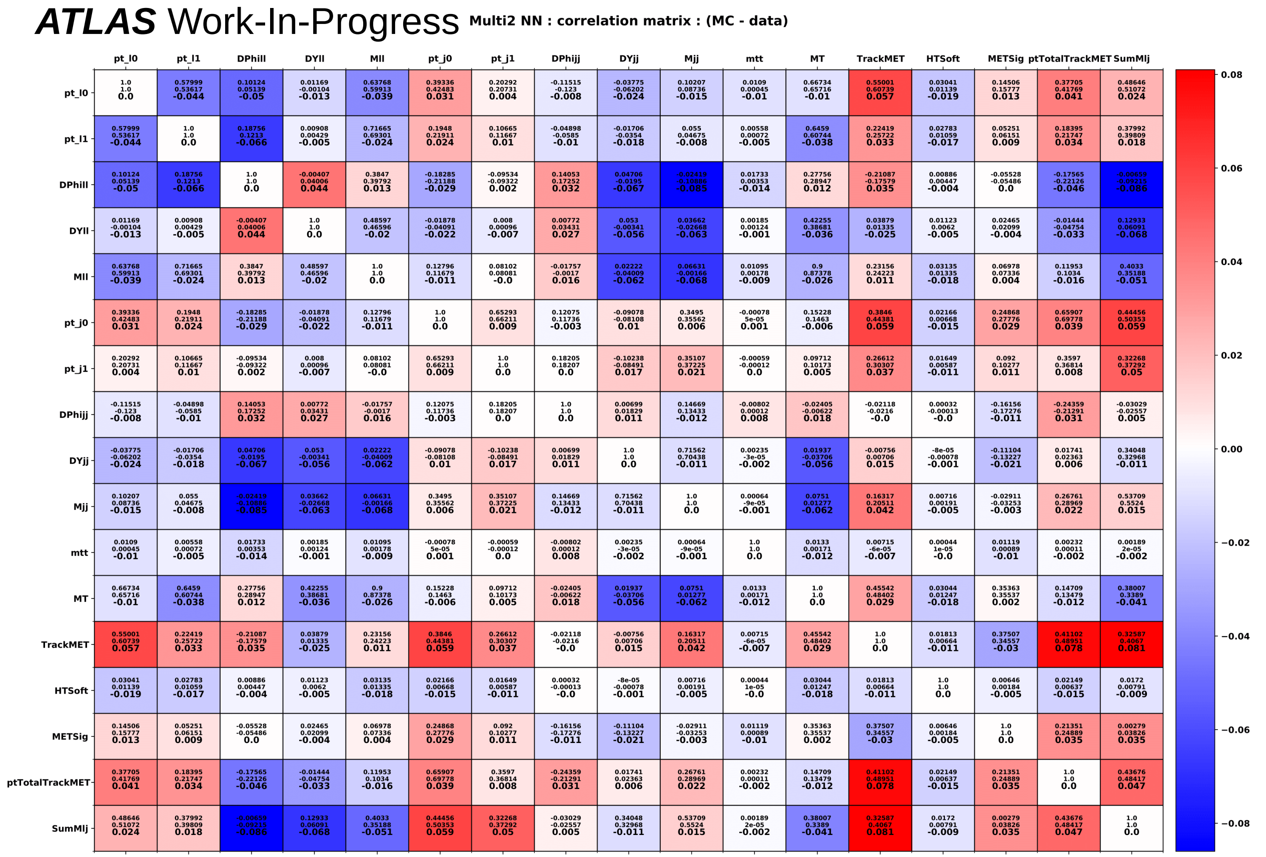
Covered in plots, but **salmon** histogram colour refers to **VBF**





Correlation Matrix: Difference (MC – data)

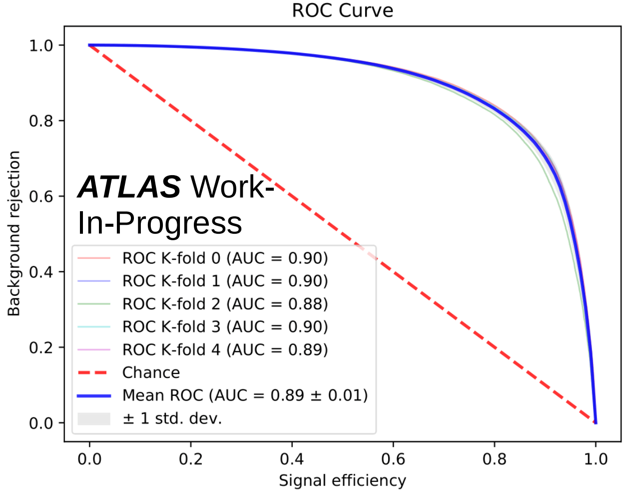
- NNs rely on *correlations* between input variables:
  - Expect NN to perform well on data if correlations are well modelled in MC
  - Element-by-element difference in correlation matrices **doesn’t exceed ±0.08** – modelling seems satisfactory
  - MC correlation matrix included in Backup

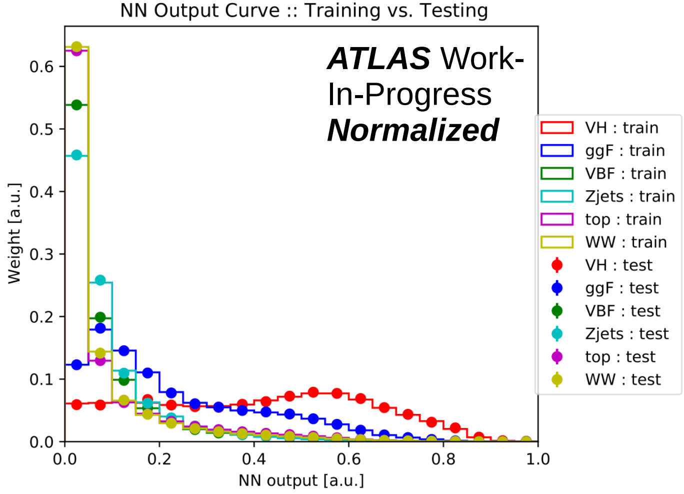


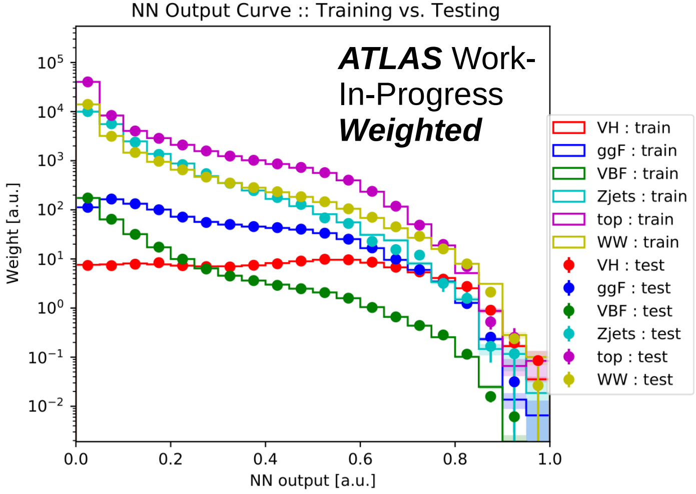
2020/06/09

VH Discriminant Performance

Validation metrics **and** **performance plots for all outputs** **in Backup**







Good ROC AUC

ROC calculation only considers signal to be VH (different from optimization definition)

Good training (training+validation) **and testing agreement** on linear and log scales (here, we show weighted average in output over all 5 NNs)

Optimizing Discriminant Binning

- Ultimately, we want to fit our MVA in our VH discriminant
  - Total significance of discovery ~ sum in quadrature of bin-by-bin significances
- We want to maximize total significance of discovery for our binning, use asymptotic formula (assuming Asimov data) [16] – it can be shown (assuming Poisson counting in signal region with Gaussian auxiliary measurement on background) that the significance of discovery goes as:

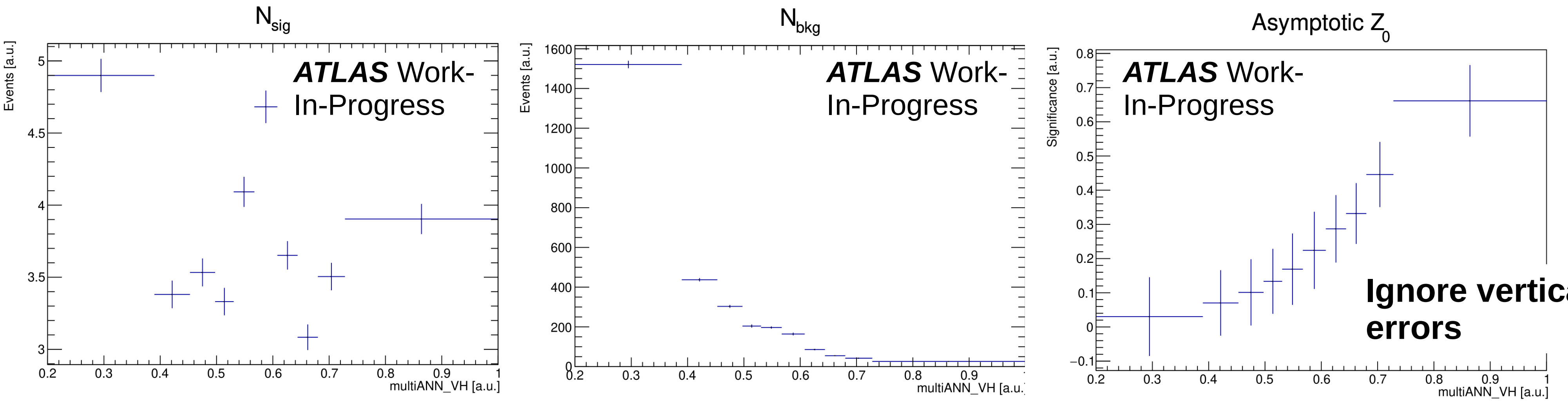
$$Z_0 = \sqrt{2 \cdot \left((s+b) \cdot \ln\left(\frac{s+b}{\hat{\hat{b}}}\right) + \hat{\hat{b}} - s - b + \frac{(b-\hat{\hat{b}})^2}{2\sigma^2} \right)},\ \hat{\hat{b}} = \frac{1}{2} \cdot \left(b - \sigma^2 + \sqrt{(b-\sigma^2)^2 + 4 \cdot (s+b) \cdot \sigma^2} \right)$$

- In the above, *s* := MC signal, *b* := MC background, and *σ* = 10% as a (very!) rough estimation of background uncertainty (to penalize background in a bin)
  - Statistics-only fit might also be a practical solution

Optimizing Discriminant Binning (2)

- Consider binning our VH discriminant in [0.2, 1.0]
  - Doesn't need to be strictly bounded by 0.2, but [0.0, 0.2] is overwhelmingly background
- Progressively cut the range into 2 parts: if Z_0 would **decrease**, **don't apply** the cut; if Z_0 would **increase**, **apply** the cut
  - We don't keep a bin if it doesn't **pass regularization**: N_{VH} , N_{bkg} > 3 for each bin
    - Protect against upward fluctuations in data from 0
  - Additional optimization: split each bin 5%-95%, 10-90%, 15-85%, ..., 90%-10%, 95%-5% and pick the splitting which increases Z_0 the most
- Also apply $|M_{jj} – 85| < 15$ GeV, $\Delta Y_{jj} < 1.2$ (Run 1 orthogonality cuts with ggF+2jets analysis [5])

Optimized Binning



Ignore vertical errors

- **Optimized binning** (rounded to reasonable sigfigs): [0.20, 0.39, 0.45, 0.50, 0.53, 0.57, 0.61, 0.64, 0.68, 0.73, 1.00] – $\boldsymbol{Z_0 = 0.97}$ **(*ATLAS* work-in-progress)**

For reference: network deployed in C++ environment using lwtnn [15]

Discussion

- Nominally, using Run 1 cuts (right) [5] on Run 2 MC (normalized to 139 ifb) using the same significance formula yields ***Z_0* = 0.29 (*ATLAS* work-in-progress)**
  - *Not an entirely fair comparison*, Run 1 used a single bin – maybe more appropriate to bin in M_T in e.g. [50, 125] GeV using the same optimization algorithm
  - In any case, **a considerable improvement** from what we started with!

$$E_T^{miss,track} > 20 \text{ GeV}$$
$$n_{jets,tight} \geq 2$$
$$n_{bjets} = 0$$
$$m_{\tau\tau} < (m_Z - 25 \text{ GeV})$$
$$m_{\ell\ell} < 50 \text{ GeV}$$
$$\Delta\varphi_{\ell\ell} < 1.8$$
$$\Delta Y_{jj} < 1.2$$
$$|m_{jj} - 85| < 15 \text{ GeV}$$
$$m_T < 125 \text{ GeV}$$

Summary

- Presented a (hopefully informative!) talk on the use of multiclassifiers for improved measurements of *V*(*H*->*WW*) in LHC physics analyses
  - Summarized training and validation – nothing too suspect
  - Asimov significance of discovery $\boldsymbol{Z_0}$ **= 0.97** for the MVA over **0.29** for the cut-based analysis **(*ATLAS* work-in-progress)**
  - Mostly care about *V*(*qq*)*H*(*lνlν*), but multiclassifier also performs **well** for ggF and **very well** for VBF in the 2 jet bin

Next Steps

- Defining control regions (CRs): possible to use other multiclassifier output discriminants to define CRs for constraining particular backgrounds
  - Careful thought is needed in order to make the signal bins and the CRs orthogonal
  - Alternatively, if non-MVA-discriminant cuts are used to define the CRs, we could *merge* classes like top and *WW*
    - Or even merge VH+ggF and use the Run 1 orthogonality cuts to separate the two
- Binning optimization with some theory systematics and a fit
  - e.g., 2-point theory systematics would be the most straightforward, e.g., parton shower uncertainty
- Simplified template cross-section (STXS) framework dictates VH hadronic be measured in $|M_{jj} - 90| < 30$ GeV [17]
  - Eventually we will need to drop the M_{jj} input (unless we fit the entire VH discriminant distribution)

References

- [1] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. **B716** (2012) 1, arXiv: 1207.7214 [hep-ex].
- [2] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. **B716** (2012) 30, arXiv: 1207.7235 [hep-ex].
- [3] ATLAS Collaboration, *Observation and measurement of Higgs boson decays to WW⁎ with the with the ATLAS detector*, Phys. Rev. **D92** (2015) 012006, arXiv: 1412.2641 [hep-ex].
- [4] ATLAS Collaboration, *Measurements of gluon-gluon fusion and vector-boson fusion Higgs boson production cross-sections in the H->WW*->eνμν decay channel in pp collisions at √s = 13 TeV with the ATLAS detector*, Phys. Lett. **B789** (2019) 508, arXiv: 1808.09054 [hep-ex].
- [5] ATLAS Collaboration, *Study of (W/Z)H production and Higgs boson couplings using H->WW* decays with the ATLAS detector*, J. High Energy Phys. **08** (2015) 137, arXiv: 1506.06641 [hep-ex].
- [6] ATLAS Collaboration, *Measurement of the production cross section for a Higgs boson in association with a vector boson in the H->WW⁎->lνlν channel in pp collisions at √s = 13 TeV with the ATLAS detector*, Phys. Lett. **B798** (2019) 134949, arXiv: 1903.10052 [hep-ex].

References (2)

- [7] T. Plehn, D. Rainwater, and D. Zeppenfeld, *A method for identifying* $H \to \tau^+\tau^- \to e^\pm\mu^\mp p_T^{miss}$ *at the CERN LHC*, Phys. Rev. **D61** (2000) 093005, arXiv: hep-ph/9911385 [hep-ph].
- [8] ATLAS Collaboration, *Object-based missing transverse momentum significance in the ATLAS detector*, ATLAS-CONF-2018-038, CERN (2018) https://cds.cern.ch/record/2630948.
- [9] F. Pedregosa *et al.*, *Scikit-learn: Machine Learning in Python*, J. Mach. Learn. Res. **12** (2011) 2825, arXiv: 1201.0490 [cs.LG].
- [10] F. Chollet *et al.*, *Keras*, (2015) https://keras.io.
- [11] M. Abadi *et al.*, *TensorFlow: A System for Large-Scale Machine Learning*, Proc. 12th USENIX Conf. Operating Syst. Design Implementation (2016) 265.
- [12] P. Moritz *et al.*, *Ray: A Distributed Framework for Emerging AI Applications*, Proc. 13th USENIX Conf. Operating Syst. Design Implementation (2018) 561, arXiv: 1712.05889 [cs.DC].

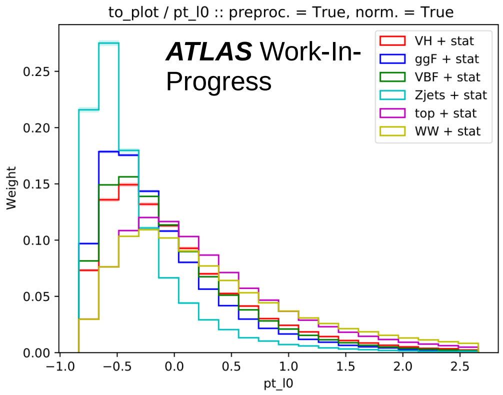
References (3)

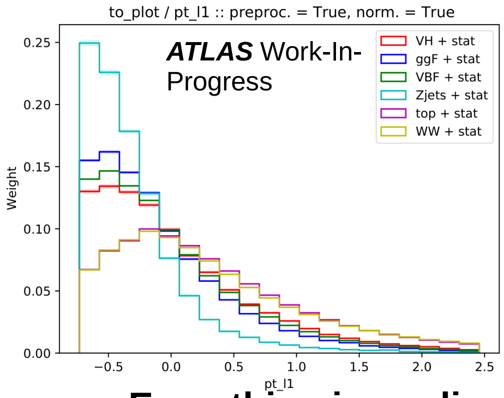
- [13] R. Liaw *et al.*, *Tune: A Research Platform for Distributed Model Selection and Training*, (2018), arXiv: 1807.05118 [cs.LG].
- [14] J. Bergstra *et al.*, *Hyperopt: a Python library for model selection and hyperparameter optimization*, Comput. Sci. Disc. **8** (2015) 014008.
- [15] D. Guest *et al.*, *Lightweight Trained Neural Network*, Github (2019) https://github.com/lwtnn/lwtnn.
- [16] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. **C71** (2011) 1554, arXiv: 1007.1727 [physics.data-an].
- [17] M. Duehrssen-Debling, P. Francavilla, F. J. Tackmann, and K. Tackmann, *Simplified template cross sections*, LHCHXSWG-DRAFT-INT-2016-006, CERN (2016) https://cds.cern.ch/record/2138079.

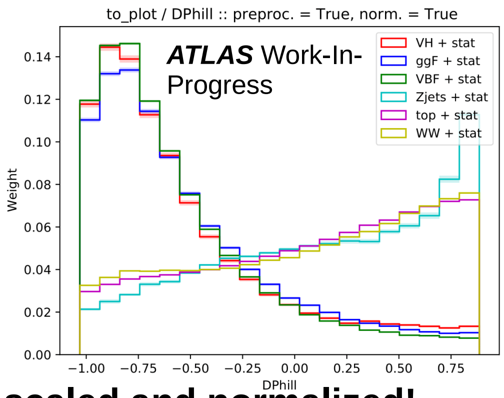
Questions?

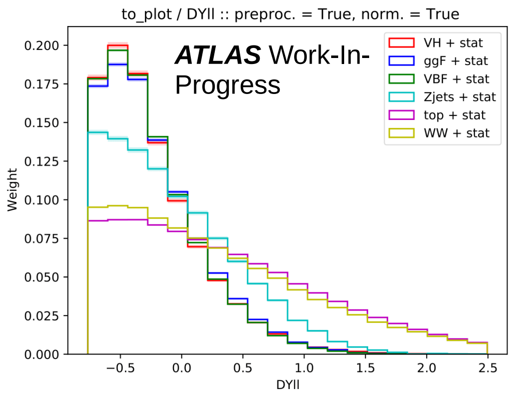
Backup

Input Shape Distributions

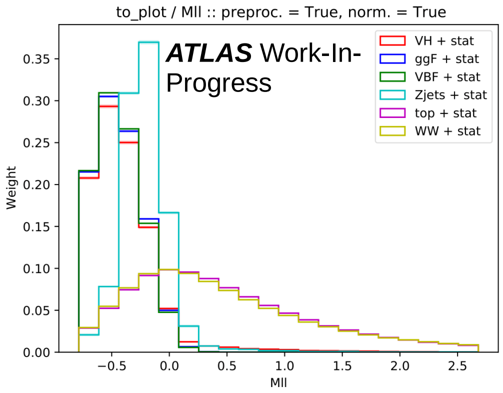


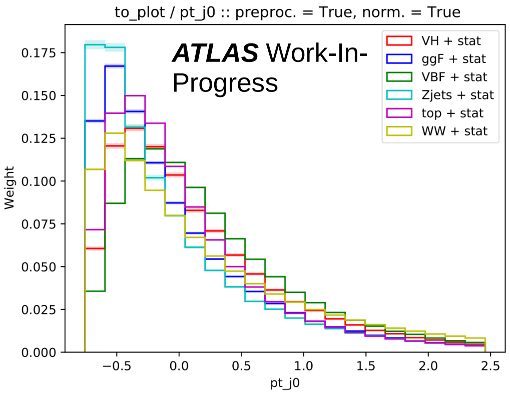


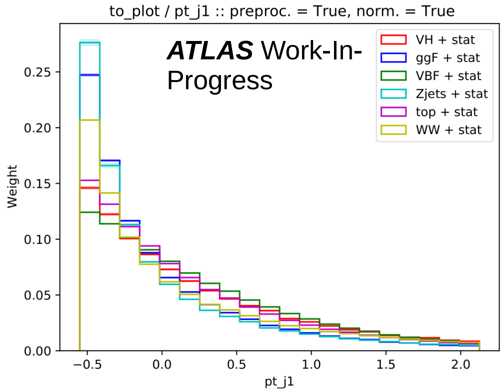


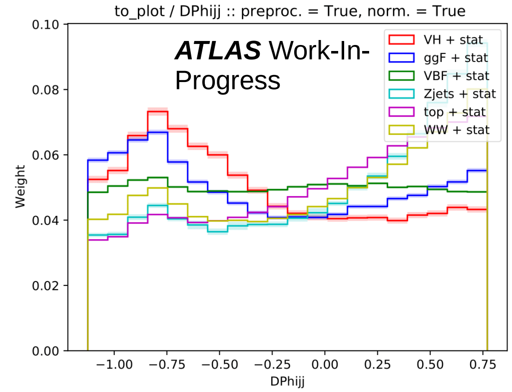


Everything is median-scaled and normalized!

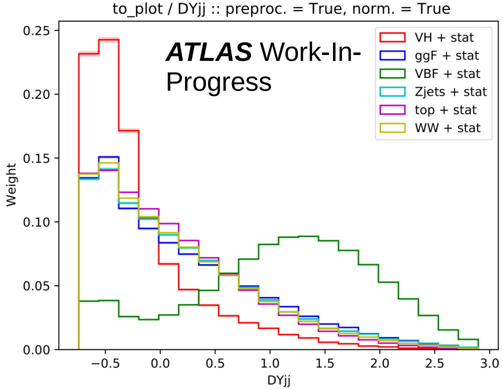


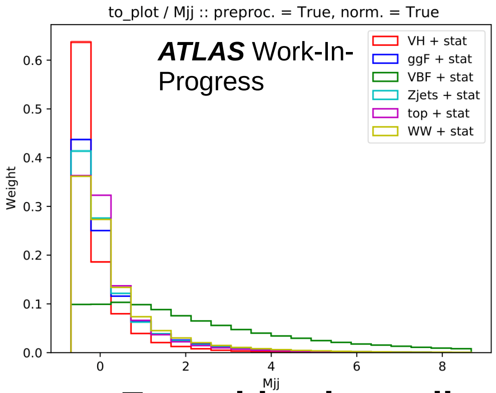


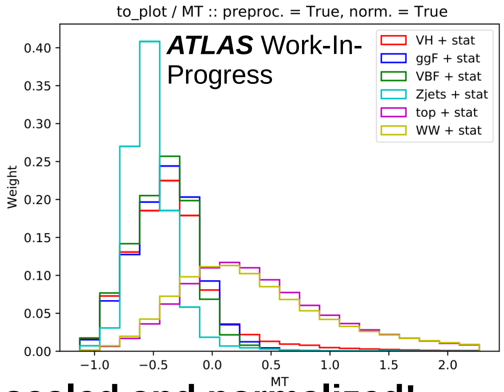


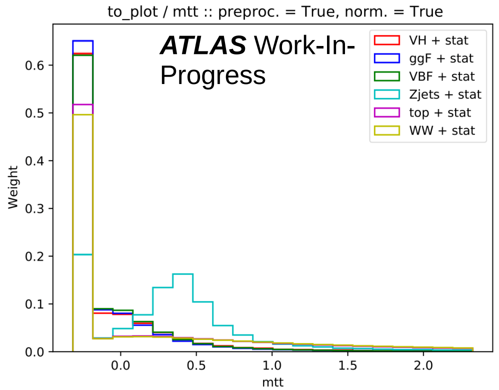


Input Shape Distributions (2)

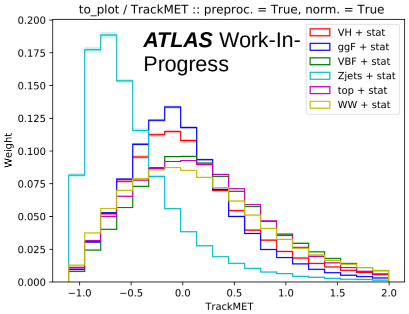


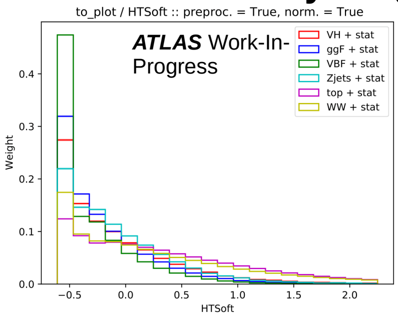


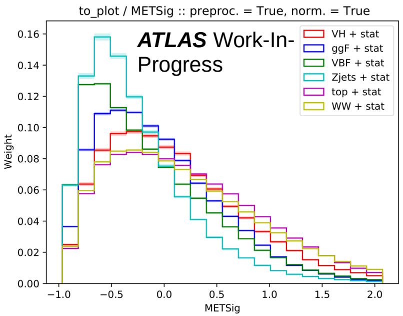


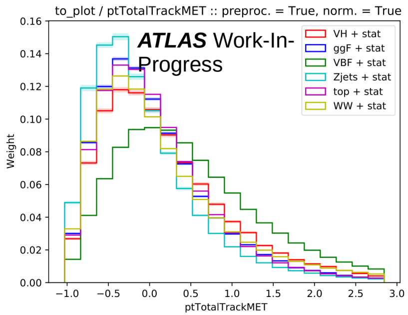


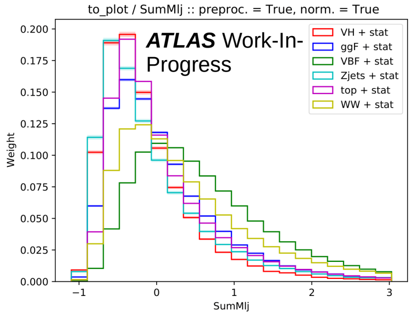
Everything is median-scaled and normalized!











K-Folding Demonstration

	All Data (100%)					
	Training (80%)					Testing (20%)
K-fold 1	Validation	Training				
K-fold 2	Training	Validation	Training			
K-fold 3	Training		Validation	Training		
K-fold 4	Training			Validation	Training	
K-fold 5	Training				Validation	

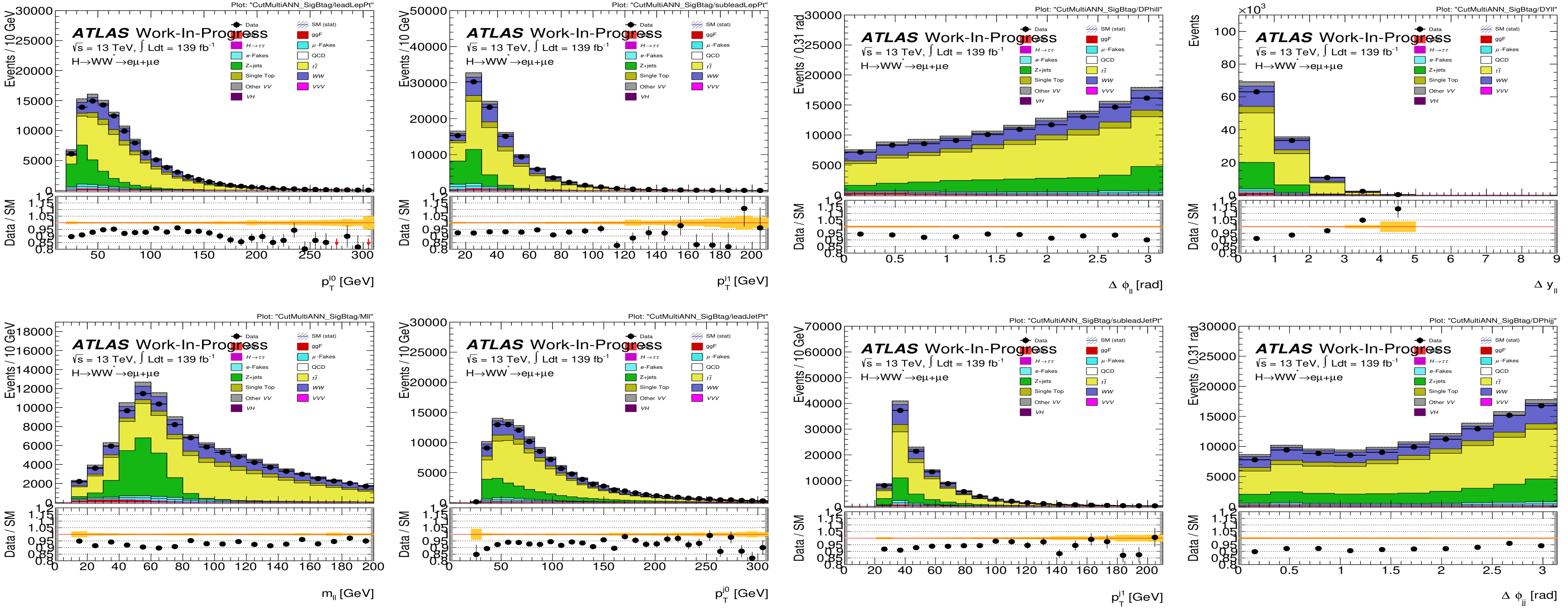
K-folding – break training data into 5 equal partitions where we are always training on 4 of the partitions and validating on 1 (to get feedback during training)

Reserved for final testing

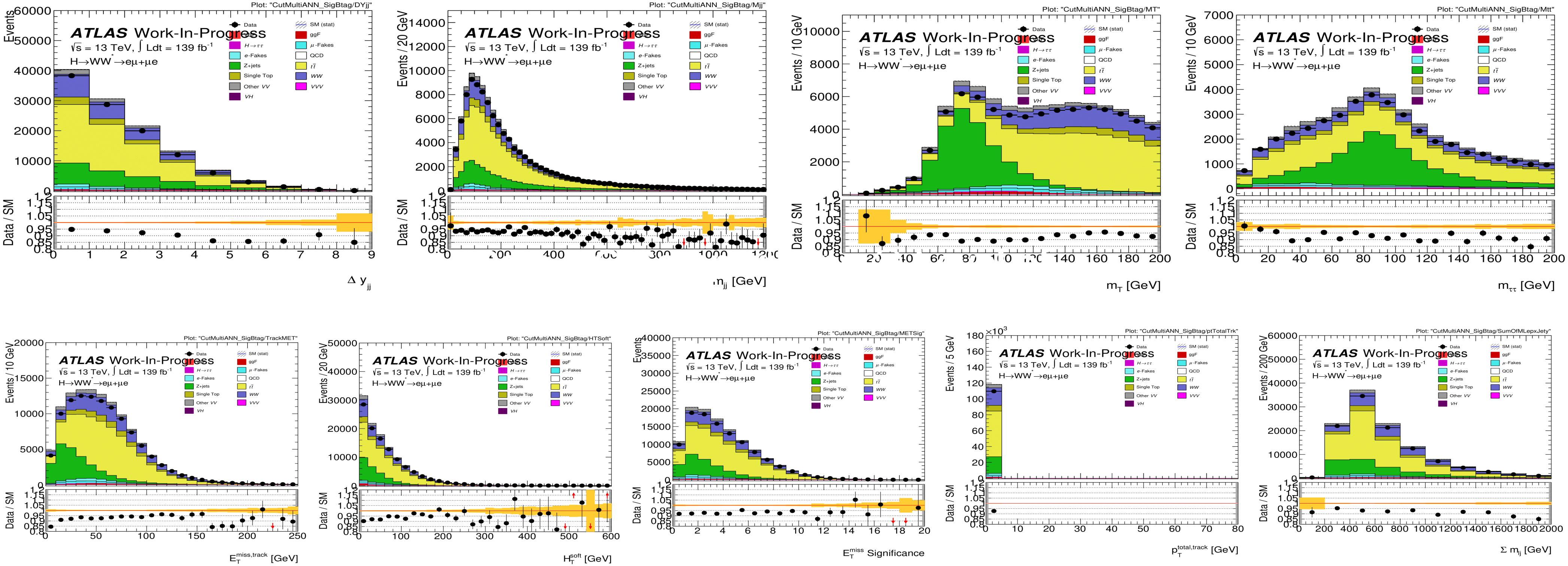
Hyperparamter Space

- Hyperparameter space defined by:
  - Number of nodes/layer: 10 to 100 in steps of 10
  - Activation function: exponential linear unit (ELU), rectified ELU (RELU)
  - Learning rate: log-uniform sampling from 1e-7 to 1e-2
- CPU scheduling performed by Tune's asynchronous HyperBand scheduler with HyperOpt search algorithm

Data-MC Comparisons @ Level of Inputs



Data-MC Comparisons @ Level of Inputs (2)



Correlation Matrix: MC Raw

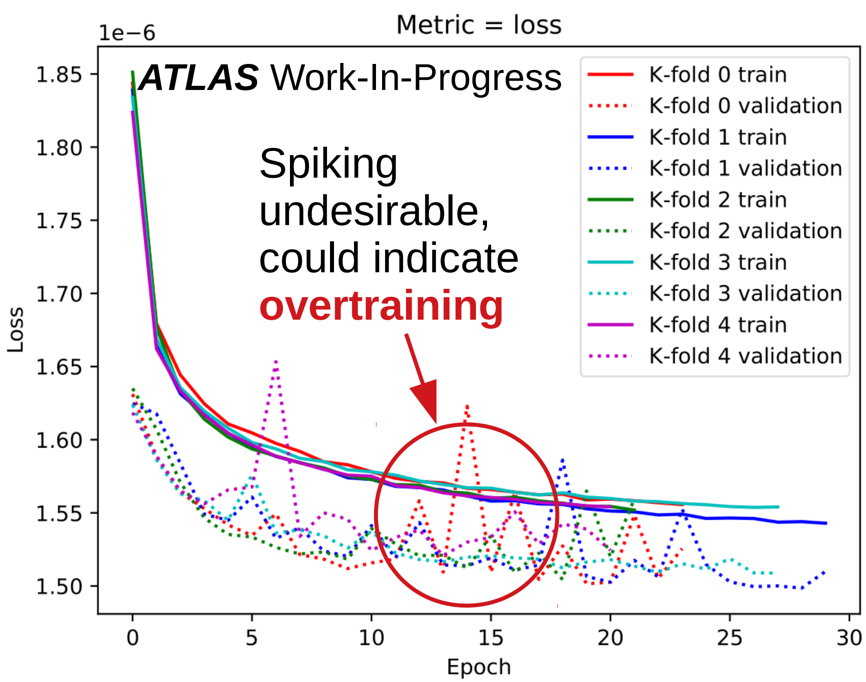
- Matrix of correlation factors in MC at the level of inputs
  - Larger correlations (e.g., ~>±0.7) could be indicative of variables which contain overlapping information

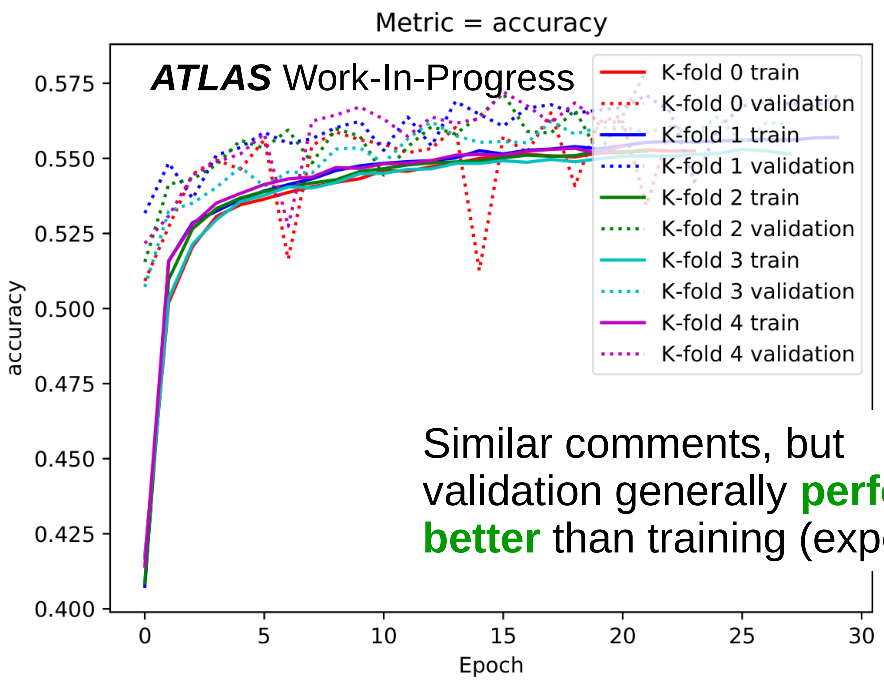
ATLAS Work-In-Progress

Multi2 NN : correlation matrix : MC

	pt_l0	pt_l1	DPhill	DYll	Mll	pt_j0	pt_j1	DPhijj	DYjj	Mjj	mtt	MT	TrackMET	HTSoft	METSig	ptTotalTrackMET	SumMlj
pt_l0	1.0	0.53617	0.05139	-0.00104	0.59913	0.42483	0.20731	-0.123	-0.06202	0.08736	0.00045	0.65716	0.60739	0.01139	0.15777	0.41769	0.51072
pt_l1	0.53617	1.0	0.1213	0.00429	0.69301	0.21911	0.11667	-0.0585	-0.0354	0.04675	0.00072	0.60744	0.25722	0.01059	0.06151	0.21747	0.39809
DPhill	0.05139	0.1213	1.0	0.04006	0.39792	-0.21188	-0.09322	0.17252	-0.0195	-0.10886	0.00353	0.28947	-0.17579	0.00447	-0.05486	-0.22126	-0.09215
DYll	-0.00104	0.00429	0.04006	1.0	0.46596	-0.04091	0.00096	0.03431	-0.00341	-0.02668	0.00124	0.38681	0.01335	0.0062	0.02099	-0.04754	0.06091
Mll	0.59913	0.69301	0.39792	0.46596	1.0	0.11679	0.08081	-0.0017	-0.04009	-0.00166	0.00178	0.87378	0.24223	0.01335	0.07336	0.1034	0.35188
pt_j0	0.42483	0.21911	-0.21188	-0.04091	0.11679	1.0	0.66211	0.11736	-0.08108	0.35562	5e-05	0.1463	0.44381	0.00668	0.27776	0.69778	0.50353
pt_j1	0.20731	0.11667	-0.09322	0.00096	0.08081	0.66211	1.0	0.18207	-0.08491	0.37225	-0.00012	0.10173	0.30307	0.00587	0.10277	0.36814	0.37292
DPhijj	-0.123	-0.0585	0.17252	0.03431	-0.0017	0.11736	0.18207	1.0	0.01829	0.13433	0.00012	-0.00622	-0.0216	-0.00013	-0.17276	-0.21291	-0.02557
DYjj	-0.06202	-0.0354	-0.0195	-0.00341	-0.04009	-0.08108	-0.08491	0.01829	1.0	0.70438	-3e-05	-0.03706	0.00706	-0.00078	-0.13227	0.02363	0.32968
Mjj	0.08736	0.04675	-0.10886	-0.02668	-0.00166	0.35562	0.37225	0.13433	0.70438	1.0	-9e-05	0.01277	0.20511	0.00191	-0.03253	0.28969	0.5524
mtt	0.00045	0.00072	0.00353	0.00124	0.00178	5e-05	-0.00012	0.00012	-3e-05	-9e-05	1.0	0.00171	-6e-05	1e-05	0.00089	0.00011	2e-05
MT	0.65716	0.60744	0.28947	0.38681	0.87378	0.1463	0.10173	-0.00622	-0.03706	0.01277	0.00171	1.0	0.48402	0.01247	0.35537	0.13479	0.3389
TrackMET	0.60739	0.25722	-0.17579	0.01335	0.24223	0.44381	0.30307	-0.0216	0.00706	0.20511	-6e-05	0.48402	1.0	0.00664	0.34557	0.48951	0.4067
HTSoft	0.01139	0.01059	0.00447	0.0062	0.01335	0.00668	0.00587	-0.00013	-0.00078	0.00191	1e-05	0.01247	0.00664	1.0	0.00184	0.00637	0.00791
METSig	0.15777	0.06151	-0.05486	0.02099	0.07336	0.27776	0.10277	-0.17276	-0.13227	-0.03253	0.00089	0.35537	0.34557	0.00184	1.0	0.24889	0.03826
ptTotalTrackMET	0.41769	0.21747	-0.22126	-0.04754	0.1034	0.69778	0.36814	-0.21291	0.02363	0.28969	0.00011	0.13479	0.48951	0.00637	0.24889	1.0	0.48417
SumMlj	0.51072	0.39809	-0.09215	0.06091	0.35188	0.50353	0.37292	-0.02557	0.32968	0.5524	2e-05	0.3389	0.4067	0.00791	0.03826	0.48417	1.0

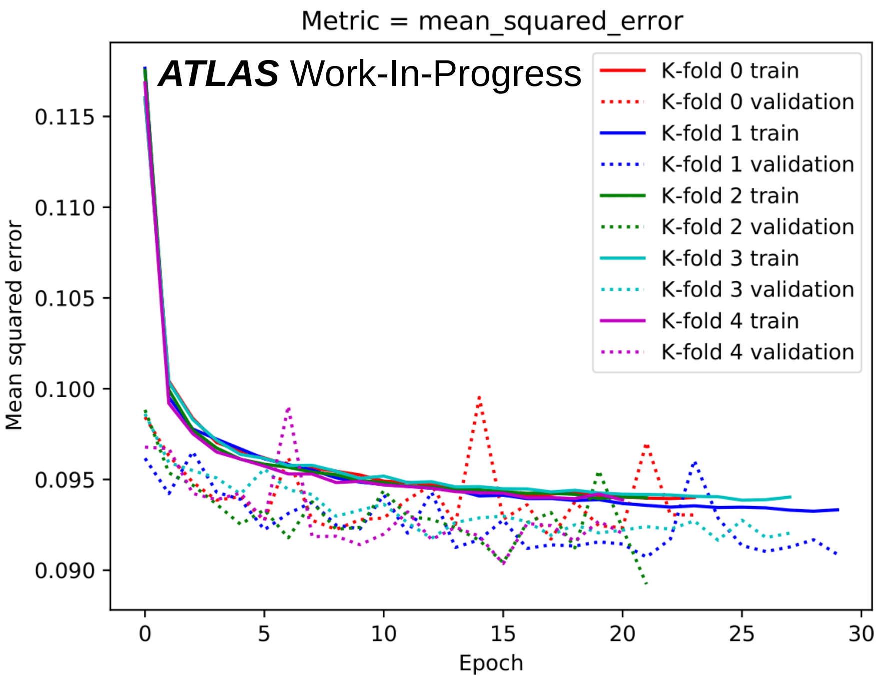
Validation Metrics

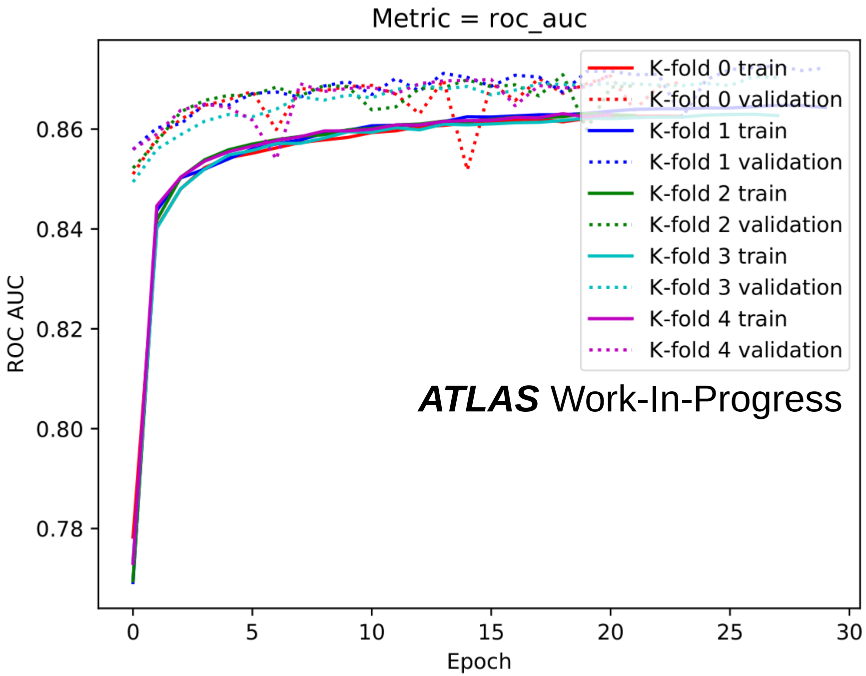




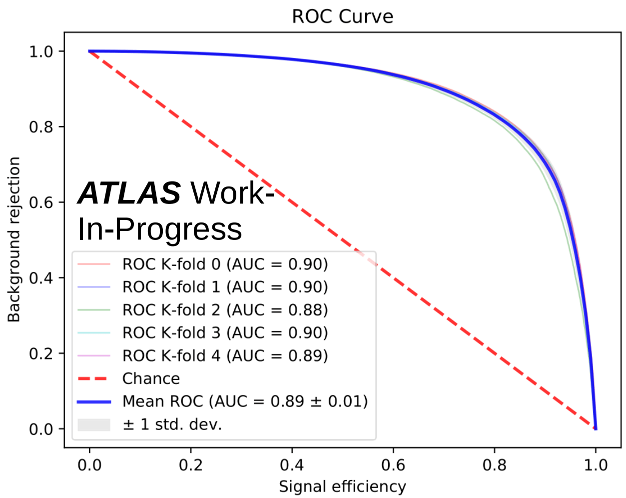
Similar comments, but validation generally **performs better** than training (expected)

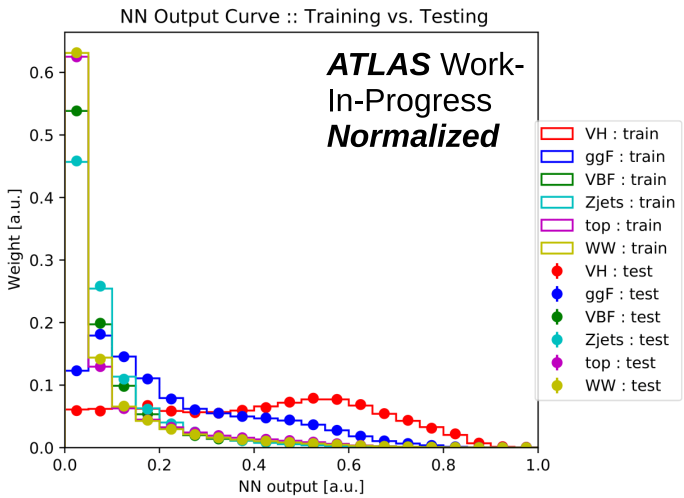
Validation Metrics (2)

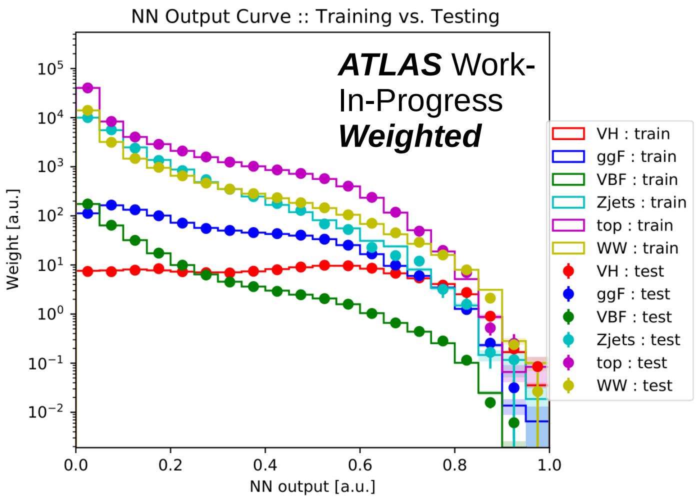




VH Discriminant Performance

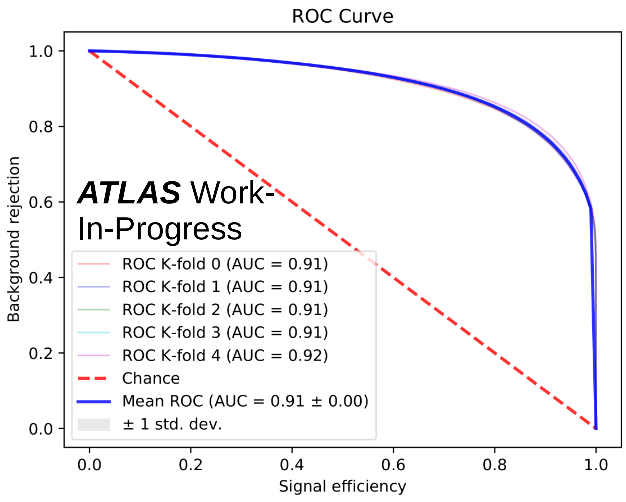


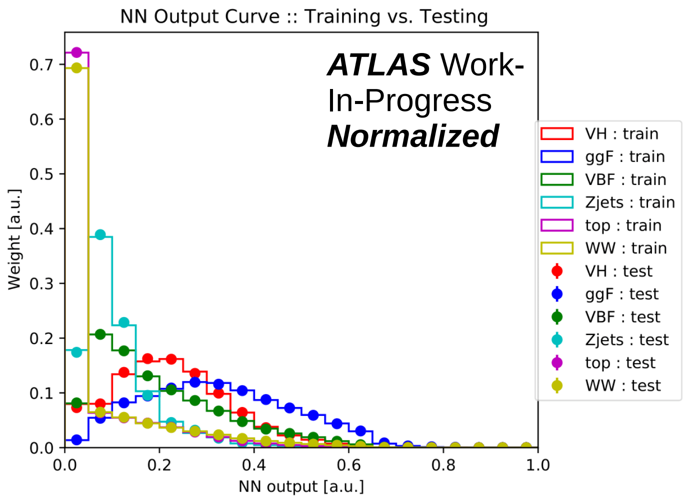


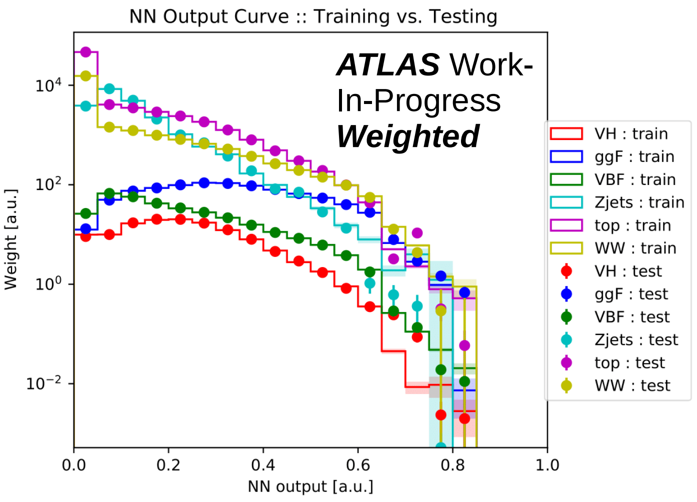


Good ROC AUC, **good train-test agreement** (**worsens** in highest bins due to lack of stats)

ggF Discriminant Performance

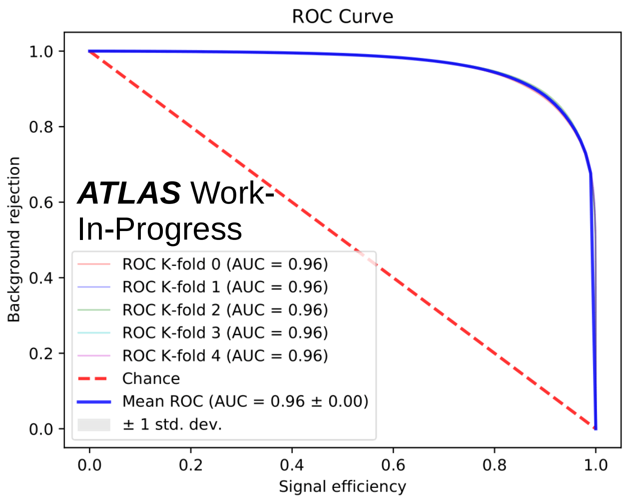


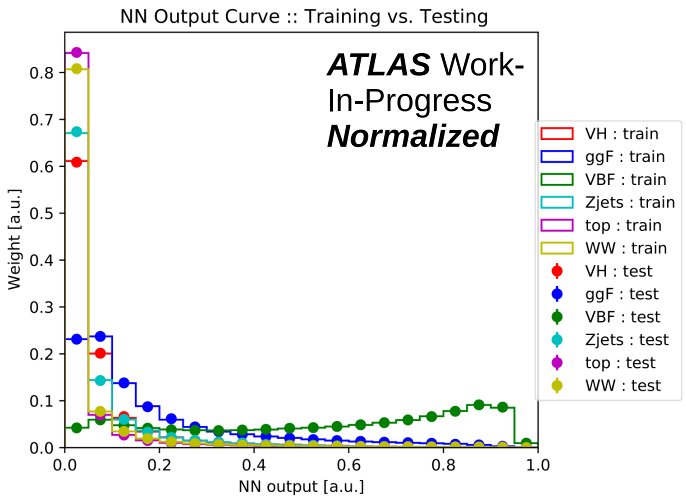


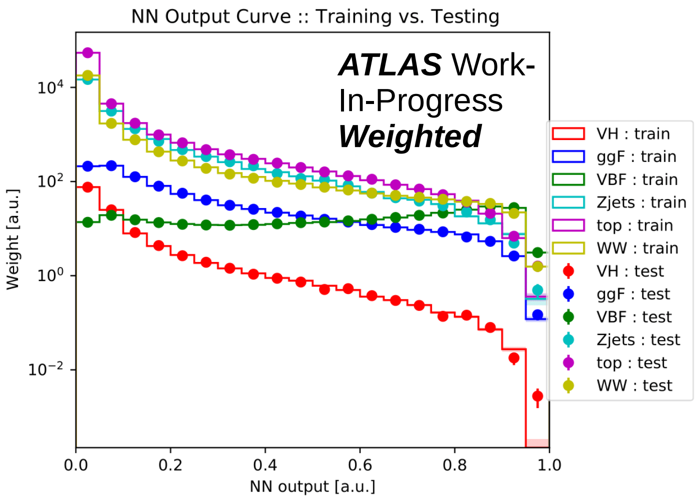


Good ROC AUC, **decent train-test agreement** (**tension** in top in highest bins, **disagreement** in Z+jets due to poor statistics)

VBF Discriminant Performance

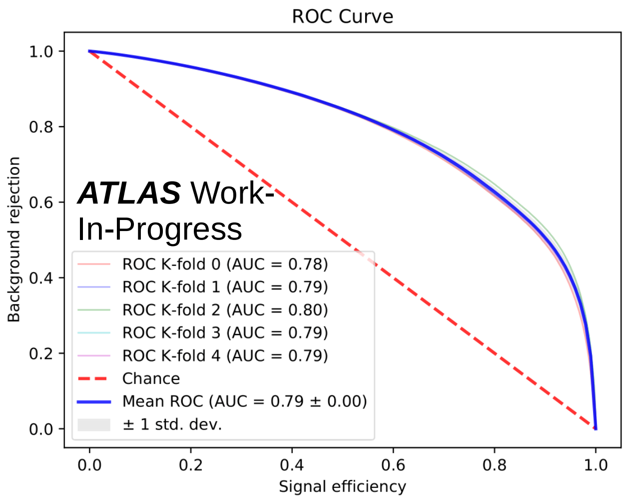


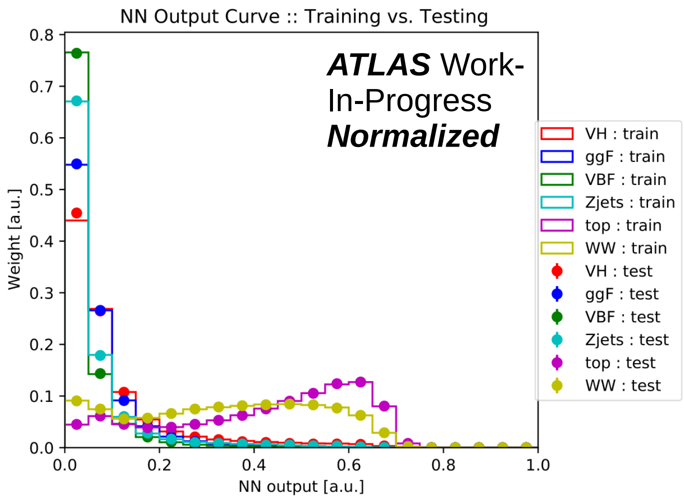


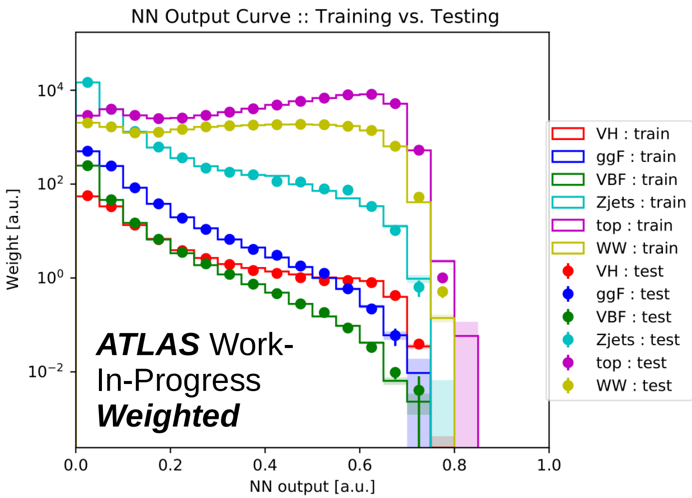


Excellent ROC AUC, excellent train-test agreement!

Top Discriminant Performance

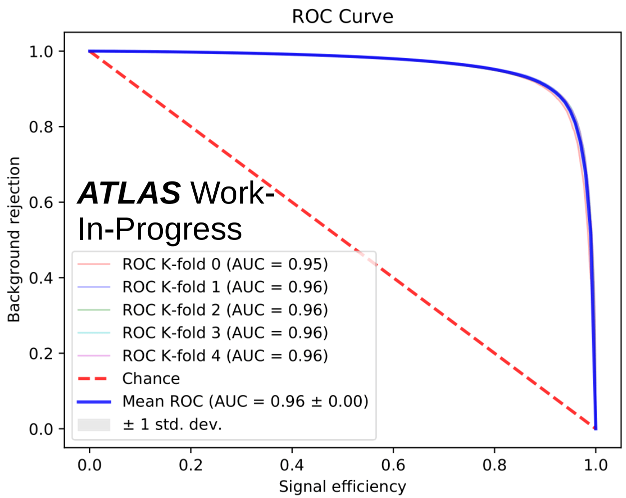


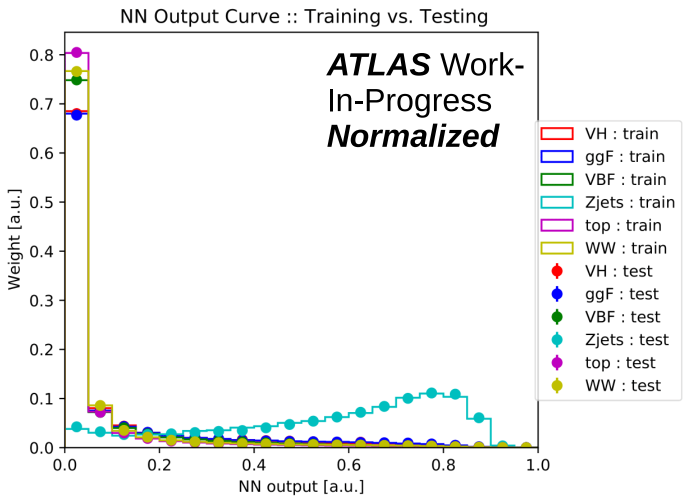


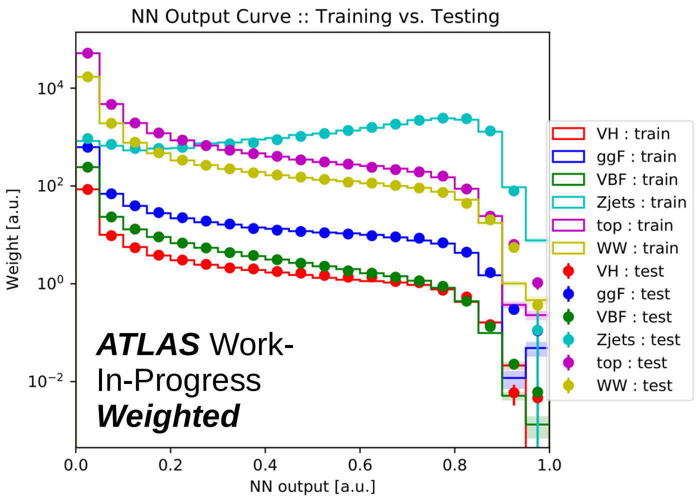


Decent ROC AUC, good train-test agreement. Hard to separate from *WW*

Z+jets Discriminant Performance

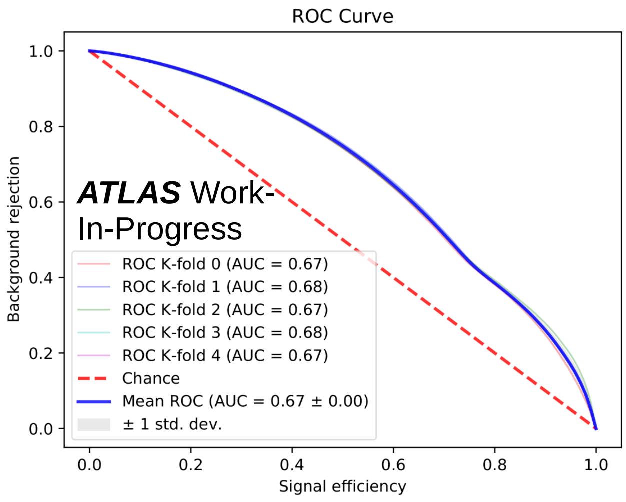


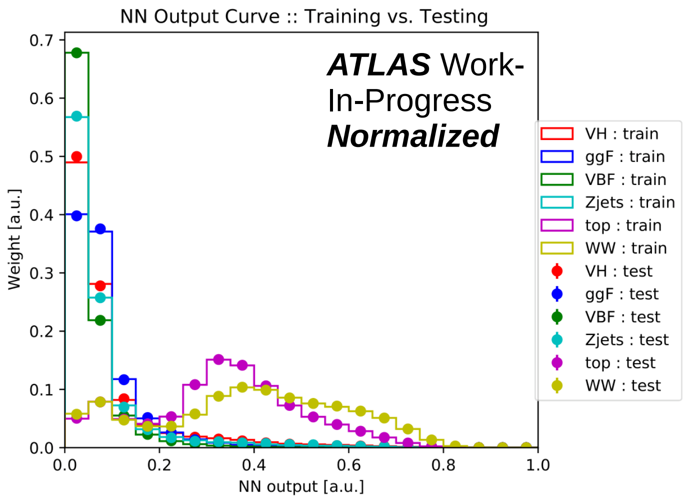


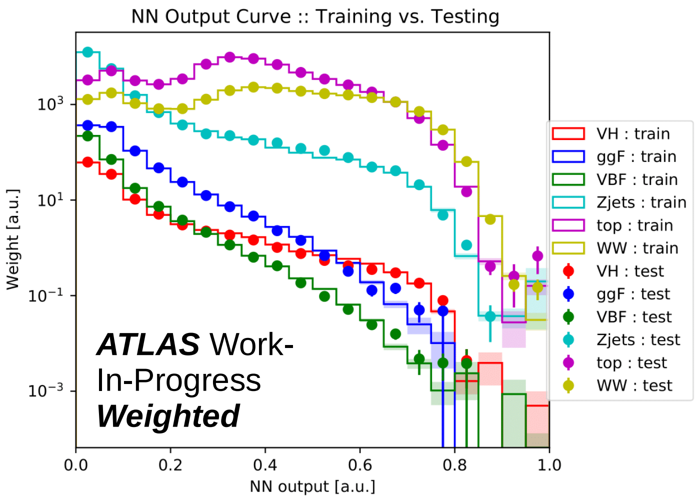


Excellent ROC AUC, excellent train-test agreement! However, we do see in **tension in Z+jets, top, *WW* in highest bins** (yields are dominated by Z+jets, however)

WW Discriminant Performance







Poor ROC AUC, good train-test agreement. Hard to separate from top