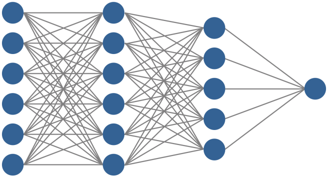
Machine Learning: Lessons Learned

Higgs Pairs Mini-Workshop, 30th September 2021

Elliot Reynolds



BERKELEY LAB

ATLAS EXPERIMENT

Run-2 dataset resonant and non-resonant (NR) *HH* publications

Experiment	Search	ML use case
ATLAS	$HH \to b\bar{b}b\bar{b}$ (VBF, NR+res.)	b-jet energy correction
ATLAS	$HH \to b\bar{b}\tau\tau$ (NR+res.)	Classification
ATLAS	$HH \to b\bar{b}\gamma\gamma$ (NR+res.)	Classification
ATLAS	$HH \to b\bar{b}\ell\nu\ell\nu$ (NR)	Classification
CMS	$HH \to b\bar{b}b\bar{b}$ (ggF+VBF, NR)	Classification & bkd estimation
CMS	$HH \to b\bar{b}b\bar{b}$ (res., boosted)	Classification
CMS	$HH \to b\bar{b}b\bar{b}$ (VBF, NR, boosted)	Classification & $m_{b\bar{b}}$ regression
CMS	$HH \to b\bar{b}\tau\tau$ (res. hh_S)	Classification
CMS	$HH \to b\bar{b}\gamma\gamma$ (ggF+VBF, NR)	Classification
CMS	$HH \to b\bar{b}\ell\ell\nu\nu$ (res.)	Classification
CMS	$HH \to b\bar{b}\ell\ell\ell\ell$ (NR)	Classification

- Neural networks (NNs) are **optimal** for a single signal (backup) ✓
- Many problems more complex...

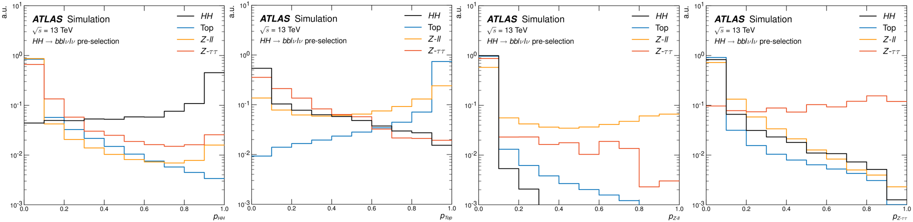
- MVAs can be used to target multiple signal processes

- This can be seen in CMS non-resonant (NR) $HH \to b\bar{b}b\bar{b}$ search, where 26–28% of ggF HH events contain additional jets which satisfy the VBF HH category selection

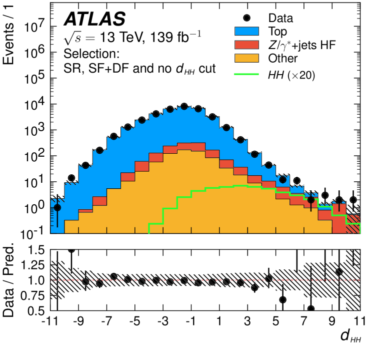
- A boosted decision tree (BDT) is used to **separate ggF and VBF *HH* signal events**
  - $\sim$ 96–97% of ggF HH events are categorised correctly ✓
  - 60% (80%) of SM ($\kappa_{2V} = 2$) VBF HH events are categorised correctly ✓

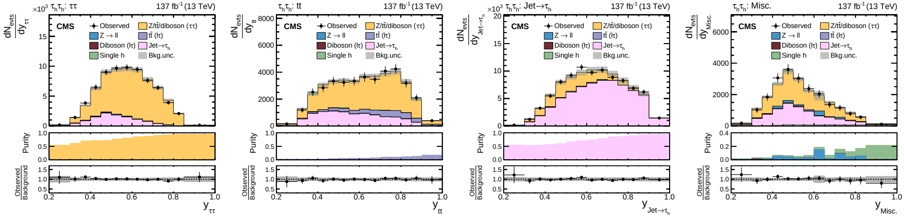
- This can be achieved in addition to background rejection using a **multi-output MVA** ✓

ATLAS $H \to b\bar{b}\ell\nu\ell\nu$ Search Classifier

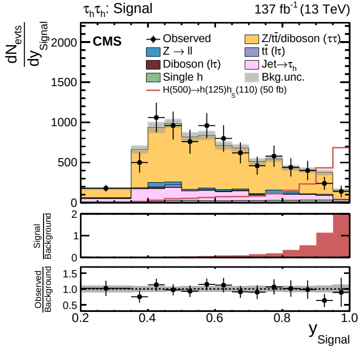


- ATLAS $HH \to b\bar{b}\ell\nu\ell\nu$ search uses a multi-output NN with 4 output nodes:
  - HH signal events
  - Top-quark
  - $Z \to \ell\ell$
  - $Z \to \tau\tau$
- 35 input variables
- Outputs combined: $d_{HH} = \ln[p_{HH}/(p_{\text{Top}} + p_{Z-\ell\ell} + p_{Z-\tau\tau})]$

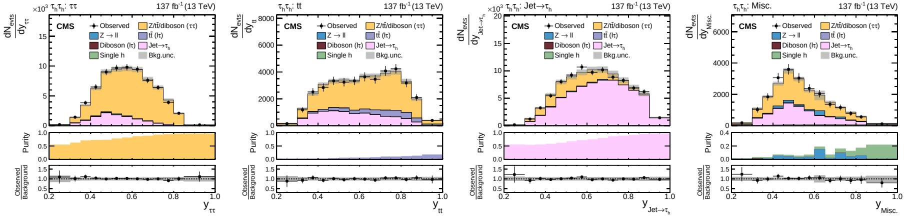




- [CMS $H \to hh_s \to b\bar{b}\tau\tau$ search]() uses a NN with 5 output nodes:
  - *HH* signal events
  - $\tau\tau$
  - fake-τ
  - $t\bar{t}$
  - Other smaller backgrounds
- Fit performed to the maximum of the NN outputs
  - Background categories **constrain systematic uncertainties** ✓



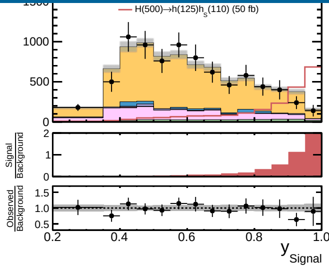
CMS $H \to hh_s \to b\bar{b}\tau\tau$ Search Classifier



Lesson 1: Multi-class or multiple MVAs can isolate multiple signals and control background systematics

- hh signal events
- $\tau\tau$
- fake-τ
- $t\bar{t}$
- Other smaller backgrounds

- Fit performed to the maximum of the NN outputs
  - Background categories **constrain systematic uncertainties** ✓



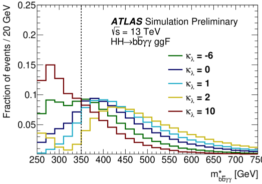
ATLAS $HH \to b\bar{b}\gamma\gamma$ Dual-MVAs for κ_λ Sensitivity

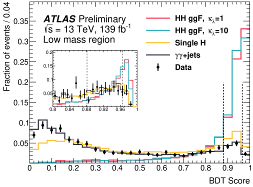
- Neural networks are optimal for a single signal hypothesis, but m_{HH} is highly dependant on κ_λ

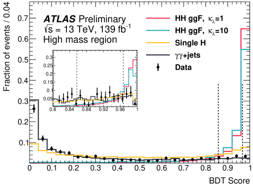
  $\rightarrow$ MVA trained on a single signal hypothesis will not be optimal for other signals ✗

- ATLAS $HH \to b\bar{b}\gamma\gamma$ search uses BDTs to reject $\gamma\gamma$, $t\bar{t}H$, ggH and ZH backgrounds, in $m^*_{b\bar{b}\gamma\gamma}$ categories trained on:
  - $\kappa_\lambda = 10$ HH signal for $m^*_{b\bar{b}\gamma\gamma} < 350$ GeV
  - $\kappa_\lambda = 1$ HH signal for $m^*_{b\bar{b}\gamma\gamma} > 350$ GeV

- This maintains **good sensitivity to both SM and BSM signals** ✓

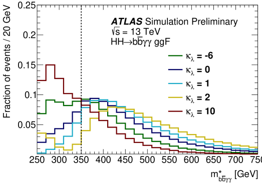




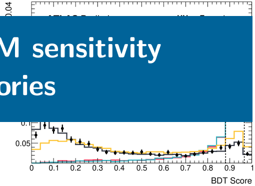


ATLAS $HH \to b\bar{b}\gamma\gamma$ Dual-MVAs for κ_λ Sensitivity

- Neural networks are optimal for a single signal hypothesis, but m_{HH} is highly dependant on κ_λ
  - $\rightarrow$ MVA trained on a single signal hypothesis will not be optimal for other signals ✗

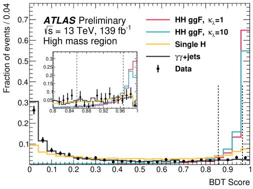


Lesson 2: Can maintain good SM and BSM sensitivity using MVAs trained in m_{HH} categories



$m^*_{b\bar{b}\gamma\gamma}$ categories trained on:

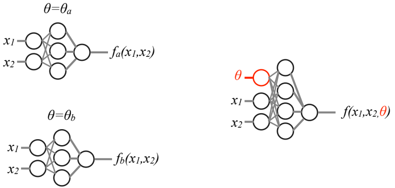
- $\kappa_\lambda = 10$ HH signal for $m^*_{b\bar{b}\gamma\gamma} < 350$ GeV
- $\kappa_\lambda = 1$ HH signal for $m^*_{b\bar{b}\gamma\gamma} > 350$ GeV



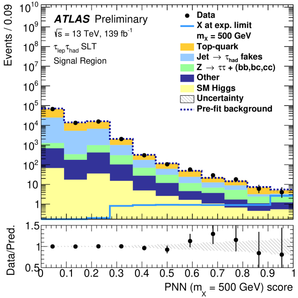
- This maintains **good sensitivity to both SM and BSM signals** ✓

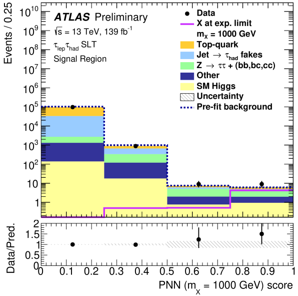
Parameterised Neural Networks

arXiv:1601.07913



- PNN: family of NNs, connected by continuous input parameter
- Often superior performance to classic NN ✓
- Strong interpolation performance ✓
- Only have to train one MVA for multiple signal regions ✓
- Parameterised neural networks used for resonance-mass-dependant classification in ATLAS $HH \to b\bar{b}\tau\tau$ search





PNN-Based κ_λ Sensitivity

- **PNNs could be parameterised in $\boldsymbol{\kappa_\lambda}$** to maximise sensitivity to all κ_λ scenarios ✓

- Many $\sigma(HH)$ limits finely-spaced in κ_λ can then be set by fitting the PNN outputs, these could then be compared with the expected cross section to set κ_λ constraints

- PNNs can also be used to profile systematic uncertainties ✓
  - This should prove useful as we collect more data!

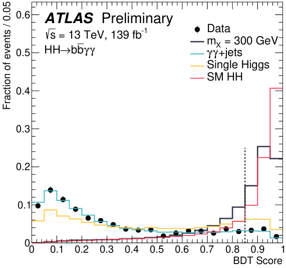
- **PNNs could be parameterised in $\boldsymbol{\kappa_\lambda}$** to maximise sensitivity to all κ_λ scenarios ✓

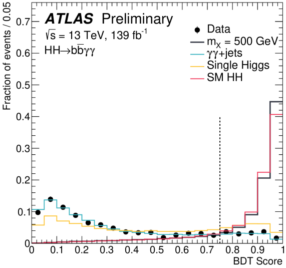
Lesson 3: PNNs can provide optimal sensitivity to a range of resonant masses (or $\kappa_\lambda/\kappa_{2V}$ hypotheses)

- PNNs can also be used to profile systematic uncertainties ✓
  - This should prove useful as we collect more data!

ATLAS $HH \to b\bar{b}\gamma\gamma$ Resonant Search BDTs

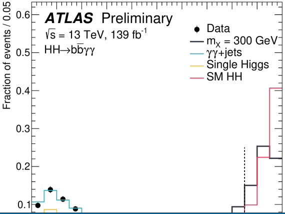
- ATLAS resonant $HH \to b\bar{b}\gamma\gamma$ search uses BDTs for signal-background separation
- **Avoid sculpting** by reweighting signal in $m^*_{b\bar{b}\gamma\gamma}$ to match background during training
- Separate BDTs trained to reject
  - $\gamma\gamma$ and $t\bar{t}\gamma\gamma$
  - Single Higgs boson events
- Weighted quadrature sum of two BDT outputs used, and weight optimised
  - This could be used to reduce systematic uncertainties on total background ✓
  - Systematic reduction could also be achieved by up-weighting high-systematics backgrounds in training ✓





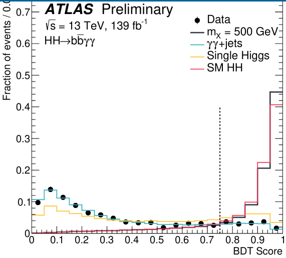
ATLAS $HH \to b\bar{b}\gamma\gamma$ Resonant Search BDTs

- ATLAS resonant $HH \to b\bar{b}\gamma\gamma$ search uses BDTs for signal-background separation
- **Avoid sculpting** by reweighting signal in $m^*_{b\bar{b}\gamma\gamma}$ to match background during training



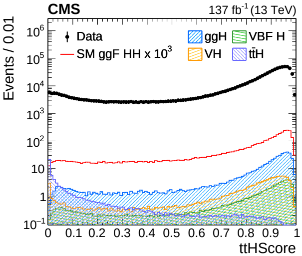
Lesson 4: Systematics can be mitigated using multiple MVA outputs (or by weighting backgrounds in MVA training)

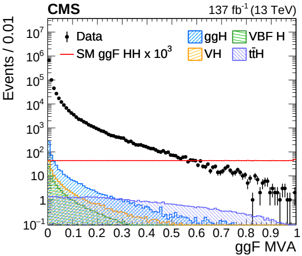
- Weighted quadrature sum of two BDT outputs used, and weight optimised
  - This could be used to reduce systematic uncertainties on total background ✓
  - Systematic reduction could also be achieved by up-weighting high-systematics backgrounds in training ✓



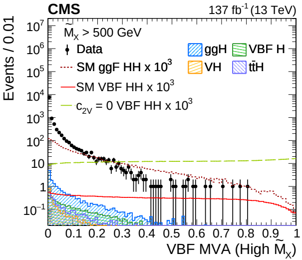
CMS Non-Resonant $HH \to b\bar{b}\gamma\gamma$ Search (1/2)

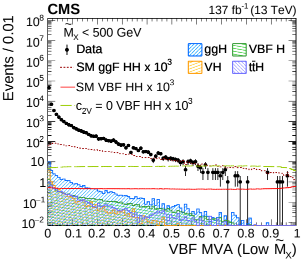
- CMS NR $HH \to b\bar{b}\gamma\gamma$ search is a **ML tour de force**!
- Dedicated NN to reject $t\bar{t}H$, training SM and BSM HH signals against $t\bar{t}H$ background, based on topology-classifier architecture NN
  - Topology-classifier architecture uses feed-forward and LSTM layers
  - Hyperparameters optimised using **Bayesian method**
- BDTs used to classify ggF HH signal against NR backgrounds
  - Mass dependence mitigated by using **dimensionless kinematic variables**
  - Signal events weighted using inverse mass resolutions to use information about resonant nature of Higgs boson decays





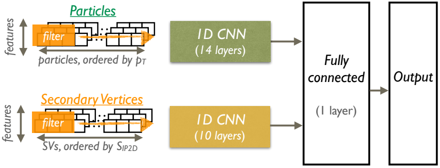
- BDTs used to classify VBF HH signal events
  - Similar to ggF BDT in many ways
  - Separate BDTs for $\tilde{M}_X < (>)$ 500 GeV $\rightarrow$ sensitivity to (B)SM signals
  - Multi-class BDT to separate VBF HH signal from $\gamma\gamma$+jets, γ+jets and SM ggF HH backgrounds
- BDTs used to isolate $t\bar{t}H$ process to simultaneously constrain κ_t
  - $t\bar{t}H$ NN output and BDT-based top-quark tagger input to this BDT





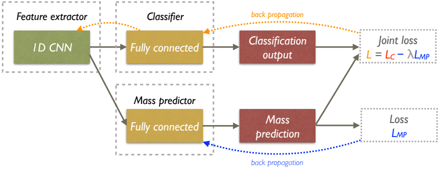
- CMS uses the DeepAK8-MD **mass-decorrelated** $X \to b\bar{b}$ tagger, e.g. in the $HH \to b\bar{b}\ell\nu\ell\nu$ search
- Multi-class $X \to b\bar{b}$ taggers, targeting: $X = W/Z/H/t/\text{other}$

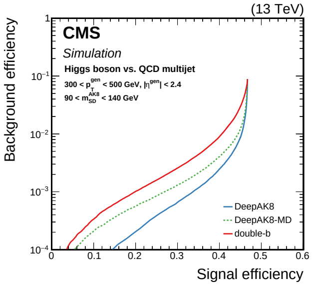
DeepAK8



JINST 15 (2020) P06005

DeepAK8-MD





CMS Boosted VBF NR $HH \to b\bar{b}b\bar{b}$: ParticleNet

- CMS VBF $HH \to b\bar{b}b\bar{b}$ search uses ParticleNet to identify $H \to b\bar{b}$ candidates and estimate $m_{b\bar{b}}$
  - ParticleNet is a permutation-invariant graph-convolutional-NN
  - The classification algorithm is multi-class, and rejects multijet and $t\bar{t}$ events
  - $> 2\times$ background rejection compared to DeepAK8-MD ✓
- Trained using dedicated samples with flat m_X distribution and reweighted $m_{b\bar{b}}$ and $p_{\mathrm{T}}^{b\bar{b}}$ distributions to ensure ParticleNet is **$m_{X/b\bar{b}}$ and $p_{\mathrm{T}}^{b\bar{b}}$ independent**
  - **Avoids sculpting background**, allowing m_X estimate to be used in background estimation ✓
- Samples with flat distribution in m_X and $\ln(p_{\mathrm{T}}^{b\bar{b}})$ are used to train m_X regression
  - **Avoids biasing m_X estimate** ✓
  - Generator-level soft drop mass is calculated for background processes

- CMS VBF $HH \to b\bar{b}b\bar{b}$ search uses ParticleNet to identify $H \to b\bar{b}$ candidates and estimate $m_{b\bar{b}}$
  - ParticleNet is a permutation-invariant graph-convolutional-NN
  - The classification algorithm is multi-class, and rejects multijet and $t\bar{t}$ events
  - $> 2\times$ background rejection compared to DeepAK8-MD ✓

Lesson 5: Reweighted training data, dimensionless inputs and adversarial trainings decorrelate MVA outputs from other variables

and p_{T} independent

  - **Avoids sculpting background**, allowing m_X estimate to be used in background estimation ✓

- Samples with flat distribution in m_X and $\ln(p_{\mathrm{T}}^{b\bar{b}})$ are used to train m_X regression
  - **Avoids biasing m_X estimate** ✓
  - Generator-level soft drop mass is calculated for background processes

- CMS $HH \to b\bar{b}b\bar{b}$ search uses ML in the background estimation
- The multijet and $t\bar{t}$ backgrounds are estimated using data, using events in a 3 b-tag control region
- Differences in several variables are addressed by **reweighting the 3 b-tag events to match the 4 b-tag events using a BDT**
  - The BDT is trained in data in a nearby $m_{H_1} - m_{H_2}$ region
  - Separate BDTs are trained to test performance of reweighting by separating reweighted and target events, and the area under the ROC curve was always 0.5 ✓
- The uncertainties deriving from the limited number of events in the 3 b-tag control regions are among the dominant uncertainties
  - The statistics in these regions could conceivably be enhanced, e.g. using Generative Adversarial Networks

- CMS $HH \to b\bar{b}b\bar{b}$ search uses ML in the background estimation
- The multijet and $t\bar{t}$ backgrounds are estimated using data, using events in a 3 b-tag control region
- Differences in several variables are addressed by **reweighting the 3**

Lesson 6: ML can be used for background estimation (and much more)

separating reweighted and target events, and the area under the ROC curve was always 0.5 ✓

- The uncertainties deriving from the limited number of events in the 3 b-tag control regions are among the dominant uncertainties
  - The statistics in these regions could conceivably be enhanced, e.g. using Generative Adversarial Networks

Summary: Lessons Learned

1. Multi-class or multiple MVAs can isolate multiple signals and control background systematics
2. Can maintain good SM and BSM sensitivity using MVAs trained in m_{HH} categories
3. PNNs can provide optimal sensitivity to a range of resonant masses (or $\kappa_\lambda/\kappa_{2V}$ hypotheses)
4. Systematics can be mitigated using multiple MVA outputs (or by weighting backgrounds in MVA training)
5. Reweighted training data, dimensionless inputs and adversarial trainings decorrelate MVA outputs from other variables
6. ML can be used for background estimation (and much more)

Backup Slides

Why Use MVAs for Classification?

- Output of neural network (NN) trained with a binary cross-entropy (BCE) loss approximates the signal probability
  - Monotonically related to density ratio: $f_S(\boldsymbol{x}_i)/f_B(\boldsymbol{x}_i)$

- Shape component of likelihood a standard mixture-model:

$$
\begin{aligned}
L(\mu;\{\boldsymbol{x}\}) &\propto \prod_{i=1}^{N}\left[\frac{\mu S}{\mu S+B}f_S(\boldsymbol{x}_i)+\frac{B}{\mu S+B}f_B(\boldsymbol{x}_i)\right] \\
&= \prod_{i=1}^{N}\left[f_B(\boldsymbol{x}_i)\right]\times\prod_{i=1}^{N}\left[\frac{\mu S}{\mu S+B}\frac{f_S(\boldsymbol{x}_i)}{f_B(\boldsymbol{x}_i)}+\left(1-\frac{\mu S}{\mu S+B}\right)\right]
\end{aligned}
$$

- This satisfies Fisher-Neyman factorisability criterion

  $\rightarrow$ NN output is a **sufficient statistic** ✓

What about Systematics?

- Mixture model in the presence of systematics:

$$
\begin{aligned}
L(\mu, \boldsymbol{\nu}; \{\boldsymbol{x}\}) &\propto \prod_{i=1}^{N} \left[\frac{\mu S}{\mu S + B} f_S(\boldsymbol{x}_i, \boldsymbol{\nu}) + \frac{B}{\mu S + B} f_B(\boldsymbol{x}_i, \boldsymbol{\nu}) \right] \\
&= \prod_{i=1}^{N} [f_B(\boldsymbol{x}_i, \boldsymbol{\nu})] \times \prod_{i=1}^{N} \left[\frac{\mu S}{\mu S + B} \frac{f_S(\boldsymbol{x}_i, \boldsymbol{\nu})}{f_B(\boldsymbol{x}_i, \boldsymbol{\nu})} + \left(1 - \frac{\mu S}{\mu S + B}\right) \right]
\end{aligned}
$$

- First term no longer constant in model parameters
  $\rightarrow$ NN output is no longer a **sufficient statistic** ✗