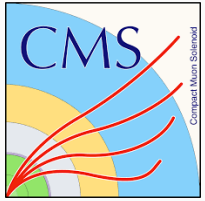


Machine Learning in CMS

ICNFP 2020

Samuel May
On behalf of the CMS Collaboration

Introduction

- Traditional machine learning (ML) methods like boosted decision trees (BDTs) have been used for many years in HEP, typically for classification and regression tasks.
  - Variety of BDTs used in the discovery of the Higgs boson: Phys. Lett. B 716 (2012) 30
- Recently, deep learning (DL) has shown great promise in multiple applications:
  1. Improving the performance of classification and regression tasks.
     - Examples: jet flavor classification and jet energy regression.
  2. Automating tasks typically done "by hand" or optimizing solutions to problems that have been traditionally solved with deterministic algorithms.
     - Examples: data quality monitoring with deep autoencoders, using DL for charged track reconstruction or particle flow.

Machine Learning in CMS

- Which tasks in HEP benefit most from the use of ML?
- Which ML algorithms provide the best solutions?

Outline

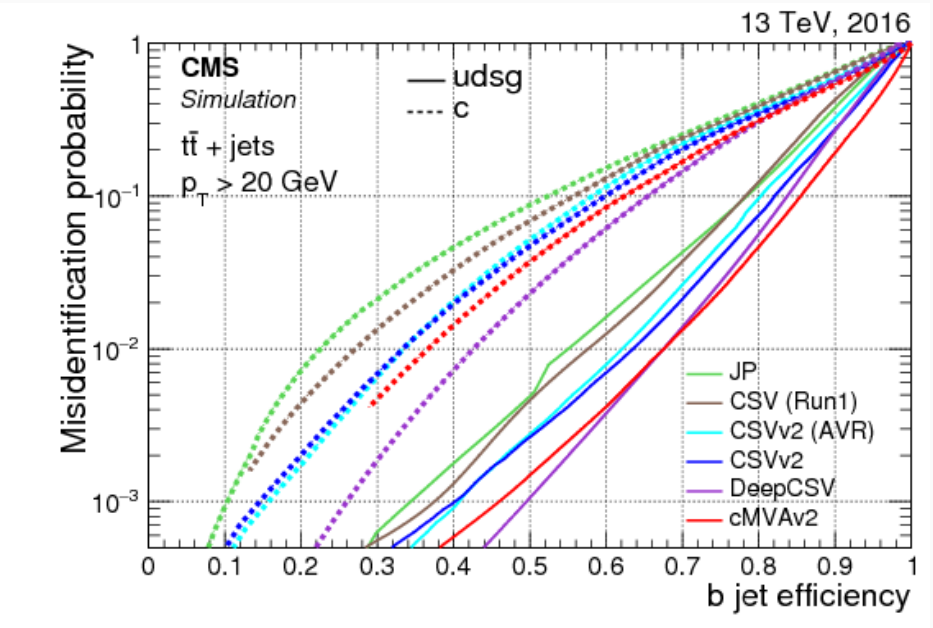
1. **Deep learning: where and why?**
    - Where can we expect DL to provide an advantage over "traditional" ML methods (e.g. BDTs)?
    - What are the challenges and risks associated with DL?
2. **Improving physics performance in CMS**
    - Which CMS analyses have benefitted substantially from extensive use of ML?
        - Collaboration-wide tools: DNN-based jet flavor algorithm
        - Analysis-specific algorithms: broad survey of ML in CMS & focus on measurements of the Higgs boson's properties in the $H \to \gamma\gamma$ channel
3. **Future prospects & conclusions**
    - What are the promising applications of ML in Run 3 of the LHC and beyond?
        - Searching for new physics with representation learning
        - DL for charged track reconstruction and particle flow

Deep learning: where and why?

Deep learning: where and why?

- Where can we expect DL methods to provide a significant advantage over "traditional" ML methods (e.g. BDTs)?

- To date, one of the most successful applications of DL in CMS is jet flavor classification:
  - DeepCSV algorithm (JINST 13 (2018) P05011)
- Jet flavor identification: use variables describing a jet's kinematics, constituent particles, and secondary vertices to identify the quark flavor from which a jet originates.
- Deep neural networks (DeepCSV) provides a huge increase in performance over likelihood-based methods (JP).



Deep learning: where and why?

- Why does DL provide such an improvement in the task of jet classification?
- First, a hand-wavy argument. Later I will argue more rigorously in studies of DNNs used in ▸ first single-channel observation of $t\bar{t}H$.

1. Low-level features
   - DNNs excel in domains where forming high-level representations is difficult to do manually.
     - Example: facial recognition in images
   - Can form high-level "summary" variables of a jet, but lose some of the original information!
   - Successors of DeepCSV, such as DeepFlavour, use even lower-level training inputs (full lists of particle flow candidates, secondary vertices) and achieve even better performance.
2. Large number of training events
   - DNNs are notoriously subject to overfitting and poor generalization – typically only an issue in the case of limited training cases.

Pitfalls of DL: Domain Adaptation

- DNNs seem great, why don't we train on the lowest-level input features possible for every classification and regression task?

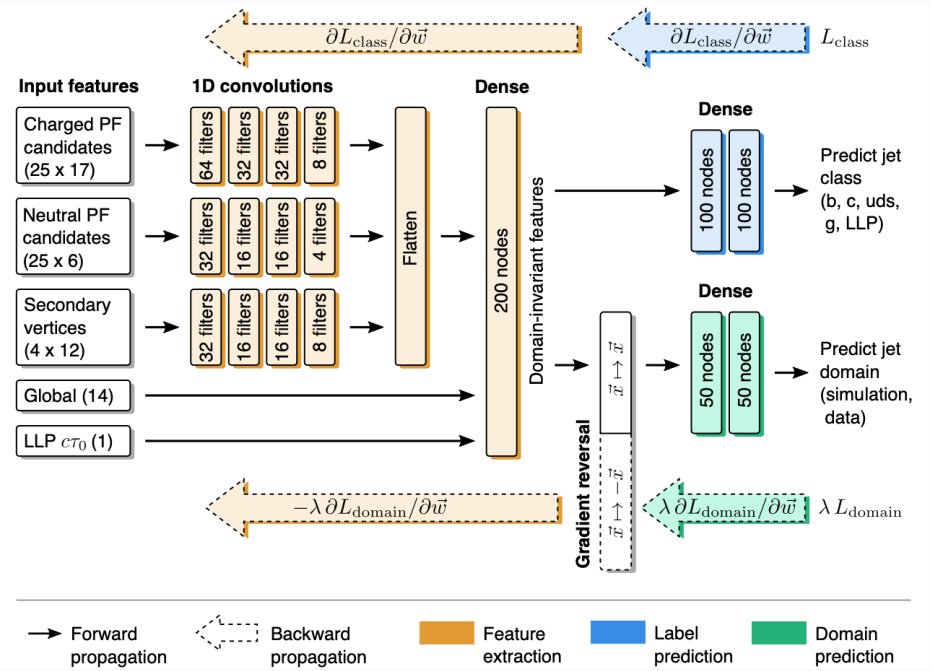
- Typically, ML algorithms in HEP are trained on simulation but applied on actual data. Problematic because:
  1. Simulations of the underlying physics are not perfect!
  2. Simulations of detector responses are not perfect!
- Blindly applying DNNs may be suboptimal:
  - Discriminants likely show disagreement between data and simulation.
    - Large systematic uncertainties associated with this difference $\Longrightarrow$ degrade analysis sensitivity.
  - A potentially more responsible approach:
    - Rather than feeding all possible low-level features to a DNN is using high-level "summary" features which can be studied individually and better understood.
- More formally, this is known as the problem of **domain adaptation**: arXiv:1409.7495

Addressing Domain Adaptation

- Can we directly address the problem of domain adaptation?

- CMS analysis searching for new long-lived particles decaying to jets uses a jet classification DNN which builds upon the DeepJet algorithm:
  Submitted to Machine Learning: Science and Technology.

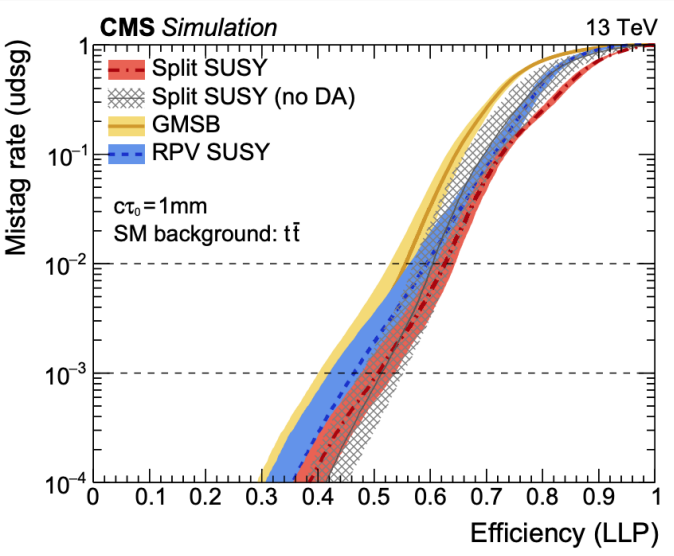
- Loss function introduces a domain adaptation component in addition to classification component.
- DNN predicts not 1, but 2 labels per jet:
  - Jet flavor (as in DeepJet)
  - Whether the event is from simulation or data
- A **gradient reversal layer** *discourages* the DNN from learning features which allow it to distinguish between jets in data and simulation.

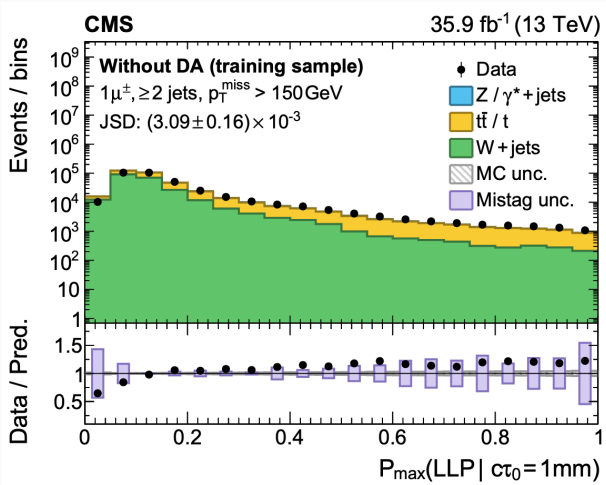


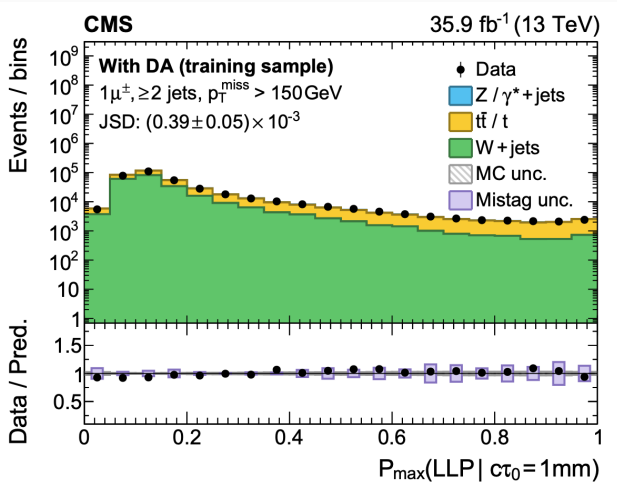
Addressing Domain Adaptation

- Can we directly address the problem of domain adaptation?

- Does domain adaptation component improve data/MC agreement?
  - Yes! Moreover, systematic uncertainties on the classifier's output are decreased.







- Performance training with DA component is comparable to training without!

Improving physics performance in CMS

Improving physics performance in CMS

- Where have ML algorithms provided significant improvement to the quality of physics results delivered by CMS?

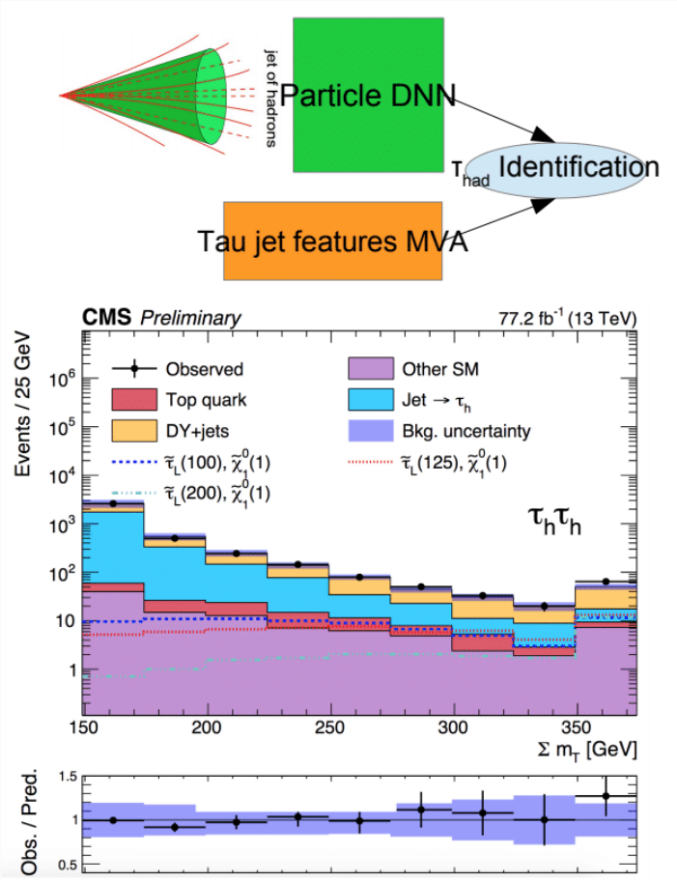
- **Survey of ML in CMS:** broad overview of applications of ML in CMS
  - Non-exhaustive list – far too many successful applications of ML in CMS to cover in a single talk!
- **Case study:** focus on measurements of the properties of the Higgs boson in the H $\rightarrow \gamma\gamma$ decay channel:
  1. First single-channel observation of $t\bar{t}H$ production in H $\rightarrow \gamma\gamma$:
  2. Measurements of Higgs boson properties in the diphoton decay channel:

Survey of ML in CMS

DNN for Tau Identification

- Eur. Phys. J. C 80 (2020) 189

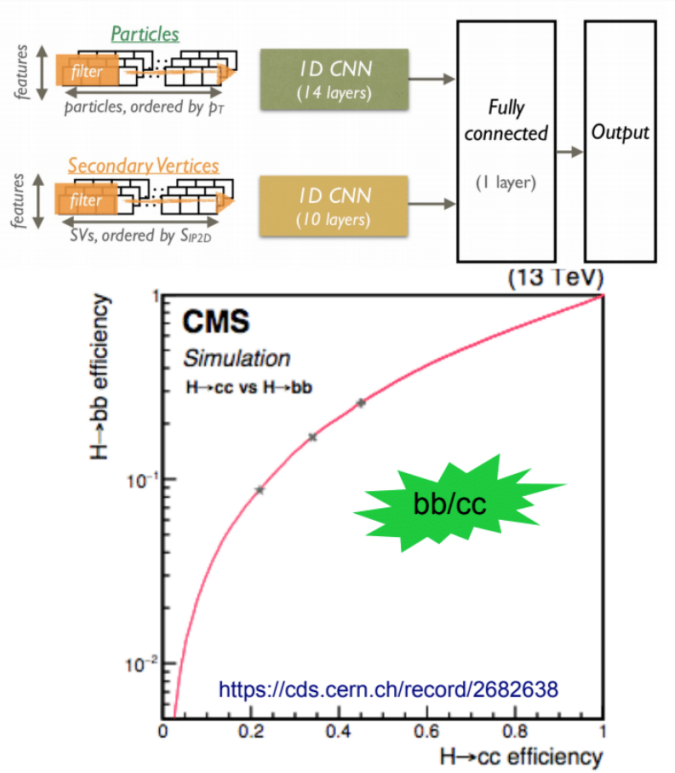
- Train a DNN "DeepPF" with information about the constituent particles inside a $\Delta R < 0.5$ cone around jet in order to identify hadronic taus, τ_h.
- Features for each particle include $p_{\mathrm{T}}^{\mathrm{particle}}/p_{\mathrm{T}}^{\mathrm{jet}}$, $\Delta R(\mathrm{particle},\tau_h)$, track quality information, impact parameter, etc.
- Trained on simulation, with signal taken as jets matched to a τ_h, background taken as jets from multi-jet and W+jets events.



H → cc Tagging

- JHEP 2003 (2020) 131

- Search for H → cc decays using large-R ($\Delta R < 1.5$) jets.
- Also featured in 10.1088/1748-0221/15/06/P06005.
- DNN trained on large-jet constituent particles and secondary vertices, using 1D CNN.
- Special focus on boosted ($p_T \geq 200$) H → cc events, which are more likely to contain both c quarks within the large-R jet.
- Observed (expected) upper limit on $\sigma_{VH} \times \mathcal{B}(H \to cc)$ of 70 (37) times SM expectation.

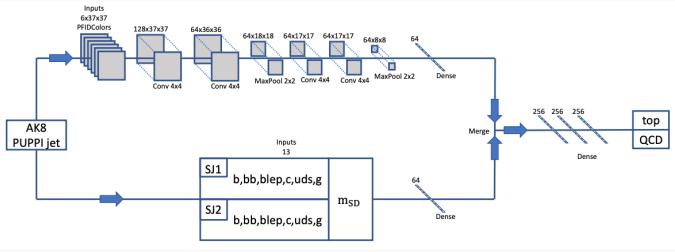


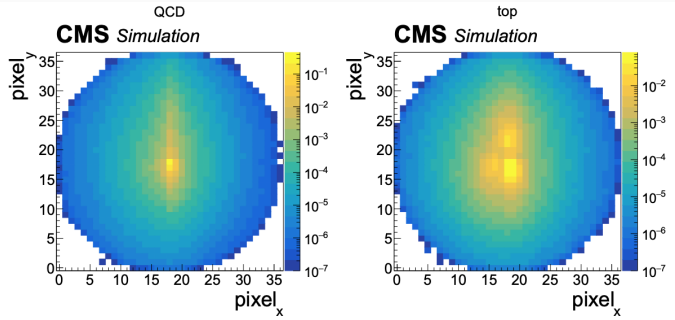
Heavy object tagging with DNNs

- 10.1088/1748-0221/15/06/P06005

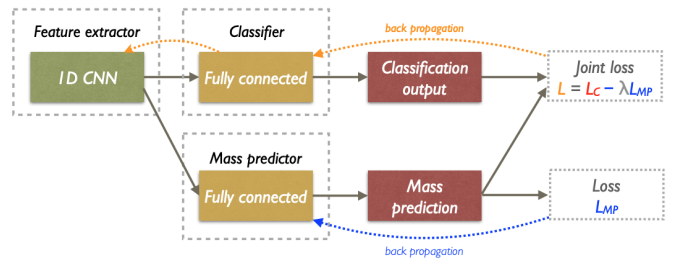
- Variety of ML solutions to the problem of heavy-object (t, W, Z, H) tagging, including comparison with standard methods.
  - 2D CNN approach to top-tagging through representing the jet as an image.
- Different analyses have different needs:
  - Train mass-decorrelated version of jet taggers for analyses which wish to use mass differences between signal and background directly.
  - NN predicts soft-drop mass and is penalized for accuracy $\Longrightarrow$ learn to distinguish jets without using mass information.

ImageTop Network





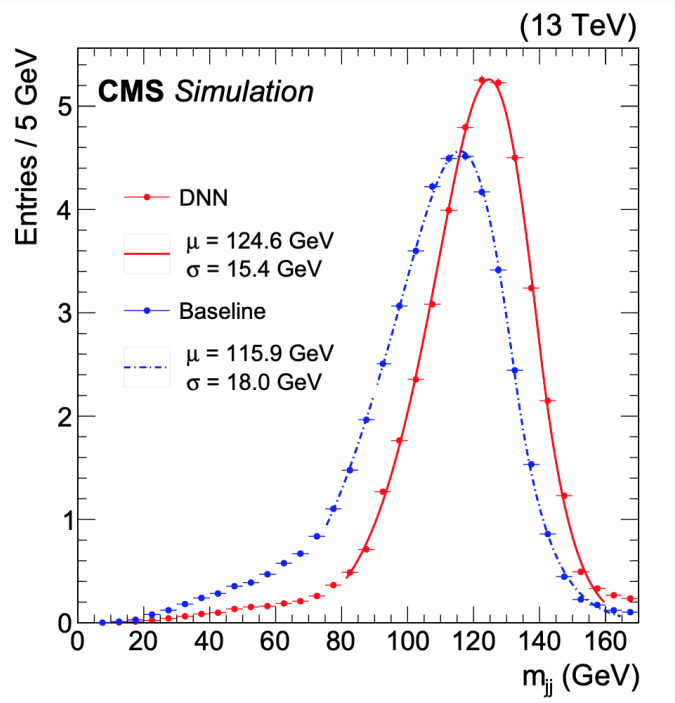
Mass-decorrelated jet tagger architecture



DNN for b-jet energy regression

- arXiv:1912.06046

- DNN trained to simultaneously estimate the energy & uncertainty of b-jets.
  - Trained with jet kinematics, information about event pileup & energy density, constituent particles of the jet, etc.
  - Custom loss function allows for estimation of 25th and 75th percentile energy values as well.
- Significant improvement over baseline method (correction factors derived from momentum balance in di-jet, γ/Z + jet events).

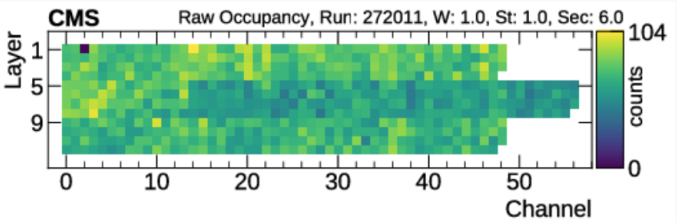


$$\text{loss}(\hat{y}, \hat{y}_{25\%}, \hat{y}_{75\%}) = E_{(x,y)\sim p(x,y)}[H_1(y - \hat{y}(x)) + \rho_{0.25}(y - \hat{y}_{25\%}(x)) + \rho_{0.75}(y - \hat{y}_{75\%}(x))], \quad (4)$$

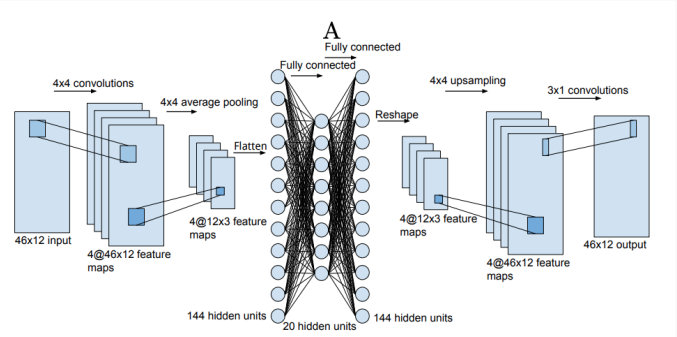
Data quality monitoring with DL

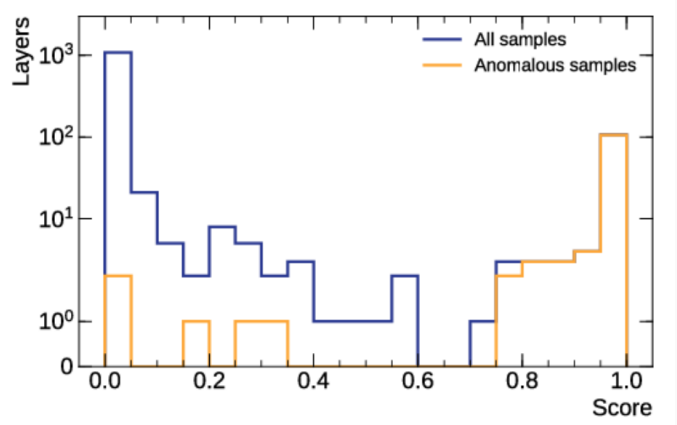
- arXiv:1808.00911

- DQM traditionally framed as supervised classification problem: "normal" vs. anomalous detector performance.
- Deep autoencoders are trained on image-like representations of the muon drift tube occupancy plots.
- Autoencoder approach offers several benefits:
  - Global approach: don't simply predict normal vs. anomalous, but localize the origin of the anomaly.
  - Previous algorithms produce a chamber-wise goodness estimate, while autoencoder approach can point to a specific problematic layer.
  - Performance gains over previous statistical algorithms, especially in the case of low-stats (i.e. beginning of data taking).



Convolutional Autoencoder architecture



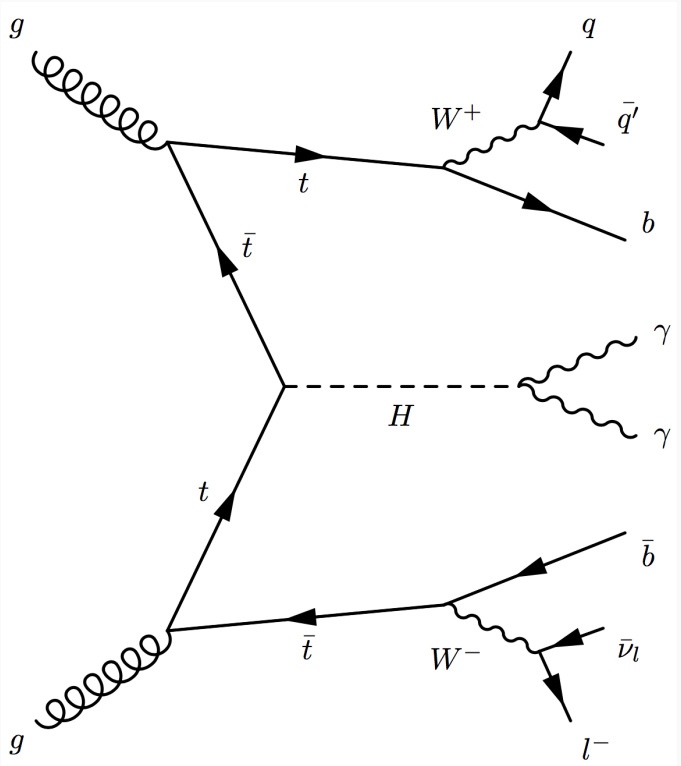


Case Study: $\mathbf{H} \to \gamma\gamma$

Observation of $t\bar{t}H$ ($H \to \gamma\gamma$)

Overview of analysis strategy for first single-channel observation of $t\bar{t}H$ production: Phys. Rev. Lett. 125, 061801 (2020)

- **Preselection:** Select for two high p_T, isolated photons and additional jets and leptons from top decays.
  - Split into two orthogonal channels: hadronic (0 leptons) and leptonic ($\geq$ 1 leptons).
- **MVAs:** for each channel, construct an MVA, "BDT-bkg", trained to separate $t\bar{t}H$ ($H \to \gamma\gamma$) from relevant SM backgrounds.
  - MVAs trained on simulation of backgrounds but applied on data $\implies$ challenges with **domain adaptation**.
- **Signal Strength Extraction:** cut on MVA score to define signal regions in each channel, constrain $\mu_{t\bar{t}H}$ through simultaneous fit in all signal regions to the diphoton invariant mass spectrum ($m_{\gamma\gamma}$).



BDT-bkg

- For each channel train a binary classification BDT ("BDT-bkg") to distinguish between $t\bar{t}H$ and other SM processes.
  - Signal: simulation of $t\bar{t}H$
  - Background: simulation of $\gamma\gamma$ + jets, $t\bar{t}$ + up to 2 photons, $Z + \gamma$, $W + \gamma$, etc and *data-driven description* of multi-jet and γ + jets.

- Features shown in red are inputs only to the Hadronic channel BDT-bkg, features shown in blue are inputs only to the Leptonic channel BDT-bkg.
- Limited description of photon/diphoton kinematics to prevent BDT from learning $m_{\gamma\gamma}$.

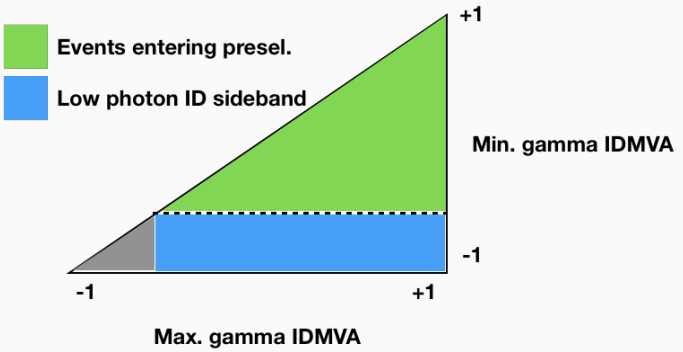
Input Features to BDTs

Category	Features		
Photon Kinematics	γ_1 $p_T/m_{\gamma\gamma}$	γ_1 η	γ_1 Pixel Seed Veto
	γ_2 $p_T/m_{\gamma\gamma}$	γ_2 η	γ_2 Pixel Seed Veto
	Max γ ID MVA	Min γ ID MVA	
Jet Kinematics	Jet 1 p_T	Jet 1 η	Jet 1 b-tag score
	Jet 2 p_T	Jet 2 η	Jet 2 b-tag score
	Jet 3 p_T	Jet 3 η	Jet 3 b-tag score
	Jet 4 p_T	Jet 4 η	Jet 4 b-tag score
	Max b-tag score	2nd max b-tag score	
	N_{jets}	H_T	
DiPhoton Kinematics	$p_T^{\gamma\gamma}/m_{\gamma\gamma}$	$Y_{\gamma\gamma}$	$\lvert\cos(\Delta\phi)_{\gamma\gamma}\rvert$
	$\Delta R_{\gamma\gamma}$	$\lvert\cos(\text{helicity angle}(\theta))\rvert$	
Lepton Kinematics	lepton p_T	lepton η	$N_{\text{leptons (tight ID)}}$
Event-level Kinematics	E_T^{miss}		

Data-Driven (γ) + jets Description

- **Challenge:** multi-jet and γ + jets events are main backgrounds ($> 50\%$) in the hadronic channel preselection, but poorly described by simulation.
- **Solution:** replace their simulation description with a data-driven description.
  - Addressing problem of domain adaptation.

- Use events which fail the preselection cut on photon ID: "low photon ID sideband".
- Low photon ID sideband dominated ($> 95\%$) by multi-jet and γ + jets events.
- Replace minimum photon ID score for each event with a new value generated from a pdf for the photon ID of fake photons.
- Scale normalization appropriately and use in place of simulation samples.

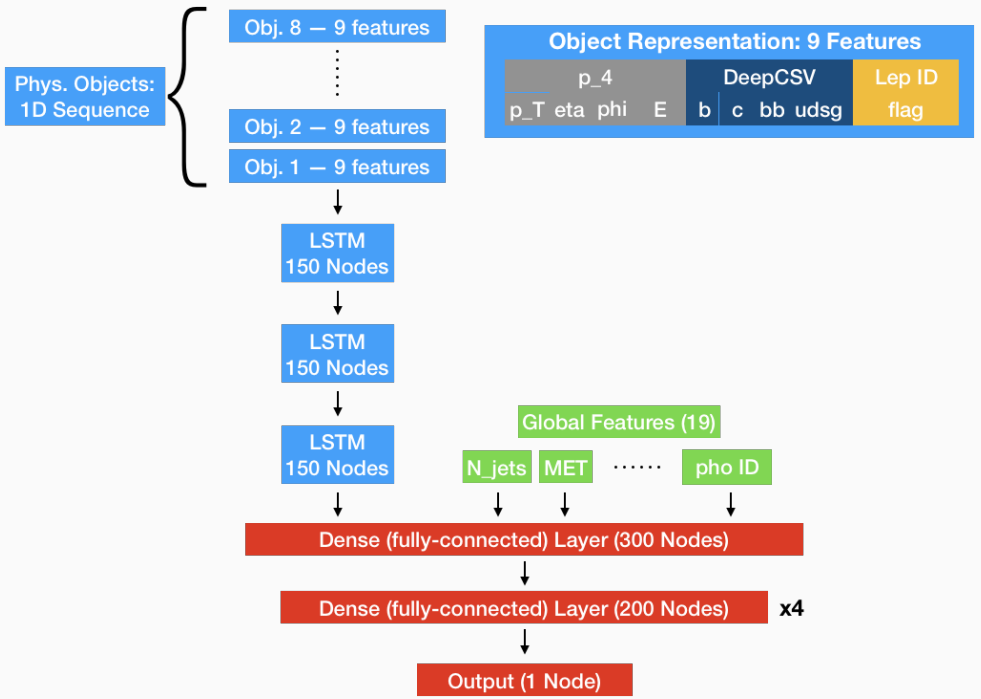


  - Improves MVA performance $\implies$ ~5% improvement in expected significance for hadronic channel.

DNNs for Specific Backgrounds

- Each event is summarized into a set of (▸ high-level input features) – these form the basis for BDT training.
- Some information is lost in summarizing.
  - Can we exploit directly the low-level information in each event with a DNN?
  - Low-level information: four vectors of leading 8 jets and leptons.

- Consider jets and leptons as 1d sequence and use LSTM architecture.
- But, DNN only outperformed BDT when enough training events were available ($\geq\approx 100$k).
- Train DNN on high-stats samples ($t\bar{t}H$ vs. $t\bar{t}+\gamma\gamma$, $t\bar{t}H$ vs. $\gamma\gamma$+jets) and use as additional input features to BDT.
- Improves MVA performance $\implies$ ∼5% improvement in expected significance for hadronic channel.



$t\bar{t}H(H \to \gamma\gamma)$: Results

- Observation of $t\bar{t}H$ production recently announced by CMS, using combination of multiple channels and Run 1 + Run 2 data [1].
- CMS [2] and ATLAS [3] recently announced measurements of signal strength and CP structure of $t\bar{t}H$ in the $H \to \gamma\gamma$ decay channel.

$$\text{signal strength} = \mu_{t\bar{t}H} = \frac{\sigma^{\text{obs}}_{t\bar{t}H}}{\sigma^{\text{SM}}_{t\bar{t}H}} \tag{1}$$

Summary of recent $t\bar{t}H$ results

Result	$\mathcal{L}$ (fb^{-1})	Obs. Signal Strength ($\mu_{t\bar{t}H}$)	Obs. (Exp.) Significance	Obs. (Exp.) CP-Odd Exclusion
CMS [2]	137	$1.38^{+0.36}_{-0.29}$	6.6 (4.7) σ	3.2 (2.6) σ
ATLAS [3]	139	1.4 ± 0.4	5.2 (4.4) σ	3.9 (2.5) σ

- $\mathcal{O}(5\%)$ improvements from domain adaptation (e.g. data-driven γ + jets description) and DL (e.g. DNNs targeting specific backgrounds) contribute to the analysis's competitive sensitivity.

[1] CMS Collaboration, "Observation of $t\bar{t}H$ Production." Physical Review Letters 120.23 (2018)

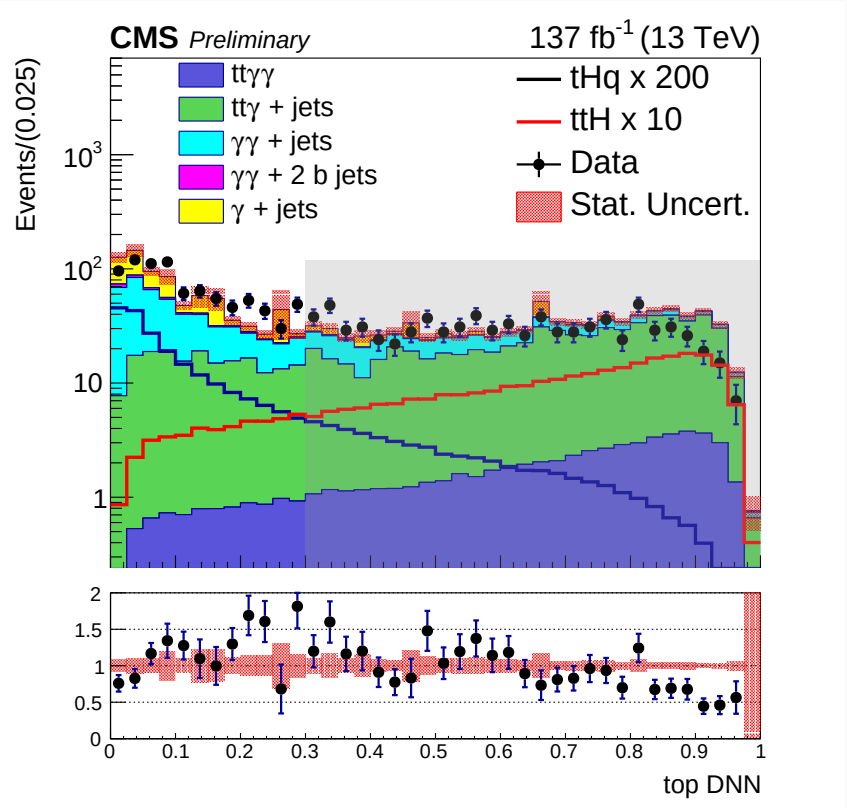
[2] CMS Collaboration, "Measurements of $t\bar{t}H$ production and the CP structure of the Yukawa interaction between the Higgs boson and top quark in the diphoton decay channel", Phys. Rev. Lett. 125, 061801 (2020).

[3] ATLAS Collaboration, "Study of the CP properties of the interaction of the Higgs boson with top quarks using top quark associated production of the Higgs boson and its decay into two photons with the ATLAS detector at the LHC", Phys. Rev. Lett. 125, 061802 (2020).

Measurements of Higgs boson properties in H $\to \gamma\gamma$: Top DNN

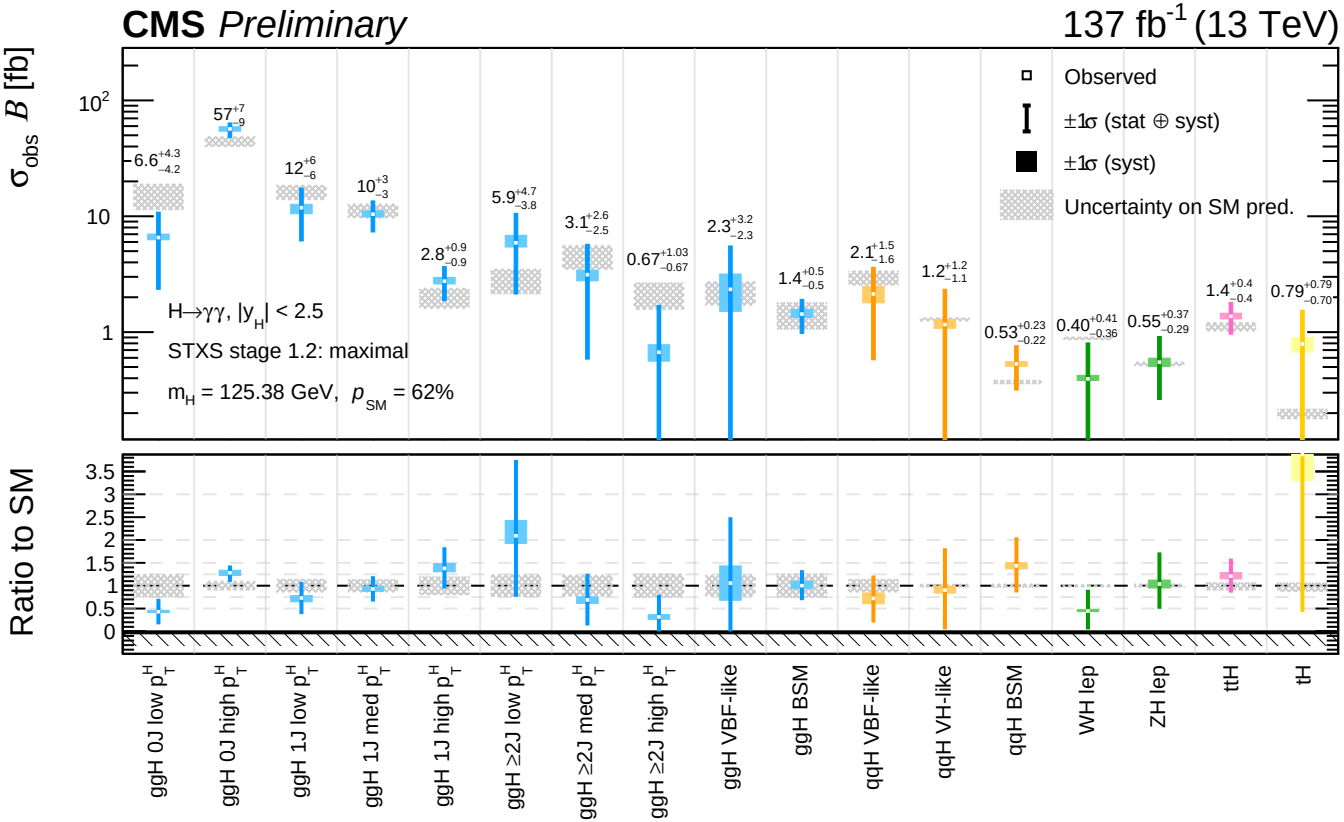
- The magnitude top Yukawa coupling y_t can be constrained through measurements of the $t\bar{t}H$ cross section.
  - But, not sensitive to the sign of y_t.
- Studying tHq production allows us to constrain the sign as well: tHq production cross section greatly enhanced if $y_t = -y_t^{SM}$.

- CMS-PAS-HIG-19-015 employs dedicated signal regions for both ttH and tHq.
- Similar final states between these two processes make them very difficult to distinguish experimentally.
- Dedicated "Top DNN" is trained to separate between ttH and tHq.
  - Same architecture as DNNs used in $t\bar{t}H$ analysis.
  - Shown to significantly outperform a BDT trained for the same task.



Measurements of Higgs boson properties in H $\rightarrow \gamma\gamma$: Results

- Upper limit on cross section for Higgs boson production in association with a single top quark is constrained to 12 times the SM prediction at 95% CL.



Future Prospects & Conclusions

Future Prospects

1. Detector reconstruction with graph neural networks (GNNs): Eur. Phys. J. C, 79 7 (2019) 608
   - Although detector components like calorimeters bear much similarity to images, their irregular geometry provides limitations to approaches using convolutional neural networks (CNNs).
2. Implementation of DNNs on firmware (FPGAs) for trigger application – reduction of backgrounds, trigger rate: JINST 13 P07027 (2018)
3. Variety of solutions to trackML kaggle challenge: NeurIPS '18 Competition
4. Particle reconstruction with GNNs: arXiv:2003.11603
   - Combine information from calorimeters, tracker, and muon system to produce a list of candidate particles in each event.
5. Anomaly detection with variational autoencoders (VAEs): J. High Energ. Phys. (2019) 2019: 36
   - Search for new physics with a model-agnostic approach: VAE creates a lower-dimensional "latent" representation of each event $\implies$ new physics would present itself as an outlier among the SM distribution in the latent space.

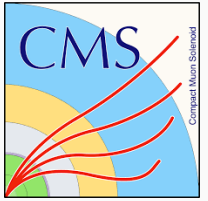
- **Disclaimer:** not explicitly CMS results, though many CMS members are involved.

Conclusion

- Recent results from CMS demonstrate the power of advanced ML methods like deep learning with an increasing number of DL algorithms being employed.
  - Case study: measurements of Higgs boson properties in the $H \to \gamma\gamma$ channel benefit from a wide variety of ML methods, including many DL algorithms
  - Significant improvements to analysis sensitivity brought through these developments.
- Deep learning tends to provides an advantage over "traditional" ML algorithms like BDTs in the regimes of (1) large number of training events and (2) absence of easily constructable high-level summary features.
- However, deep learning presents challenges:
  - Problem of domain adaptation: labeled events are often only available in simulation. Differences between data and simulation can then result in discrepancies between algorithms' behavior on simulation and data.

- A variety of innovative solutions to typical problems in HEP are currently underway, showing promise to improve physics results in Run 3 and beyond.


UNIVERSITY OF CALIFORNIA
SAN DIEGO
LET THERE BE LIGHT

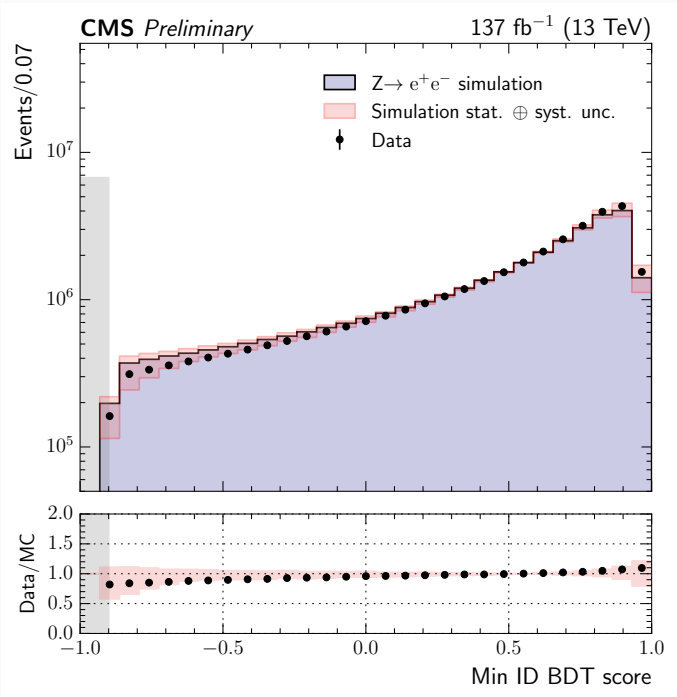

CMS
Compact Muon Solenoid

Backup

Photon ID in $H \to \gamma\gamma$

- Two recent $H \to \gamma\gamma$ results from CMS each utilize a photon ID BDT trained to separate between "prompt" and "fake" photons.
  - Observation of $t\bar{t}H$ production: Phys. Rev. Lett. 125, 061801 (2020)
  - Measurements of Higgs boson properties in diphoton decay channel: CMS-PAS-HIG-19-015

- "Fake" photons are hadronic jets which are misidentified as photons (mainly $\pi_0 \to \gamma\gamma$).
- Inputs include shower shape variables, isolation variables, etc. List (▸ here).
  - However, shower shape variables are not perfectly modeled in simulation.
- Need for domain adaptation: correct input features with a chained quantile regression (CQR) method: 10.1007/s10994-016-5546-z.
  - Set of BDTs which morph the CDFs of shower shape variables in simulation to match those in data. Good agreement observed post-CQR!



Photon ID MVA

- Inputs to the photon ID MVA include: (red = endcap only)
  1. Full 5x5 R_9
  2. Full 5x5 $\sigma_{i\eta i\eta}$
  3. η width
  4. ϕ width
  5. Covariance ($i\eta i\phi$)
  6. S4 ratio (E2x2 / E5x5)
  7. PF Photon Isolation
  8. Charged isolation wrt chosen vertex
  9. Charged isolation wrt worst vertex
  10. Photon supercluster η
  11. Photon supercluster E
  12. ρ
  13. ES effective sigma (preshower spread)
  14. ES energy / supercluster raw energy