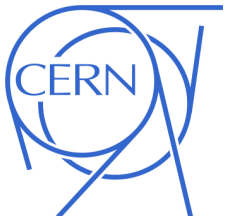
Machine Learning for Physics Analysis at the LHC (Selected Topics)

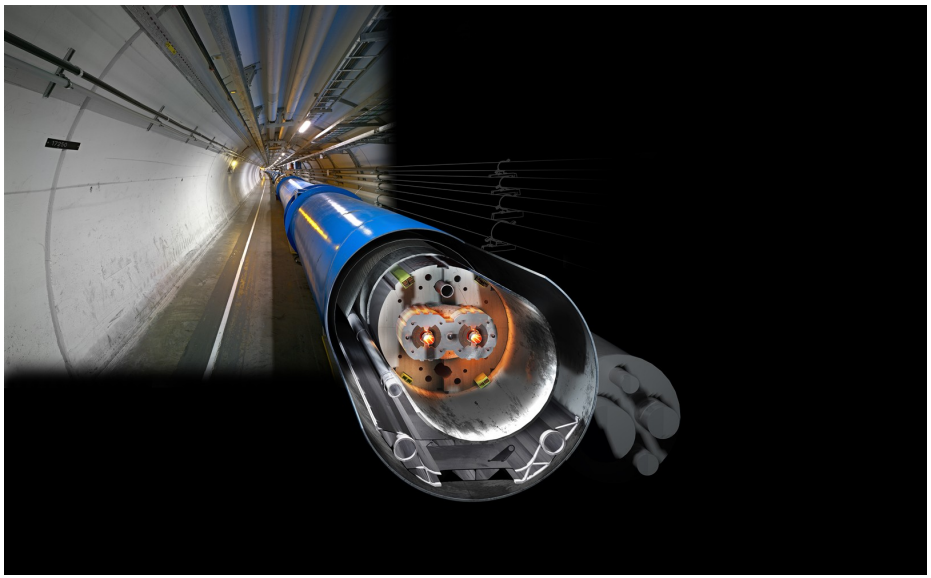
Josh Bendavid (CERN)



Feb. 22, 2018

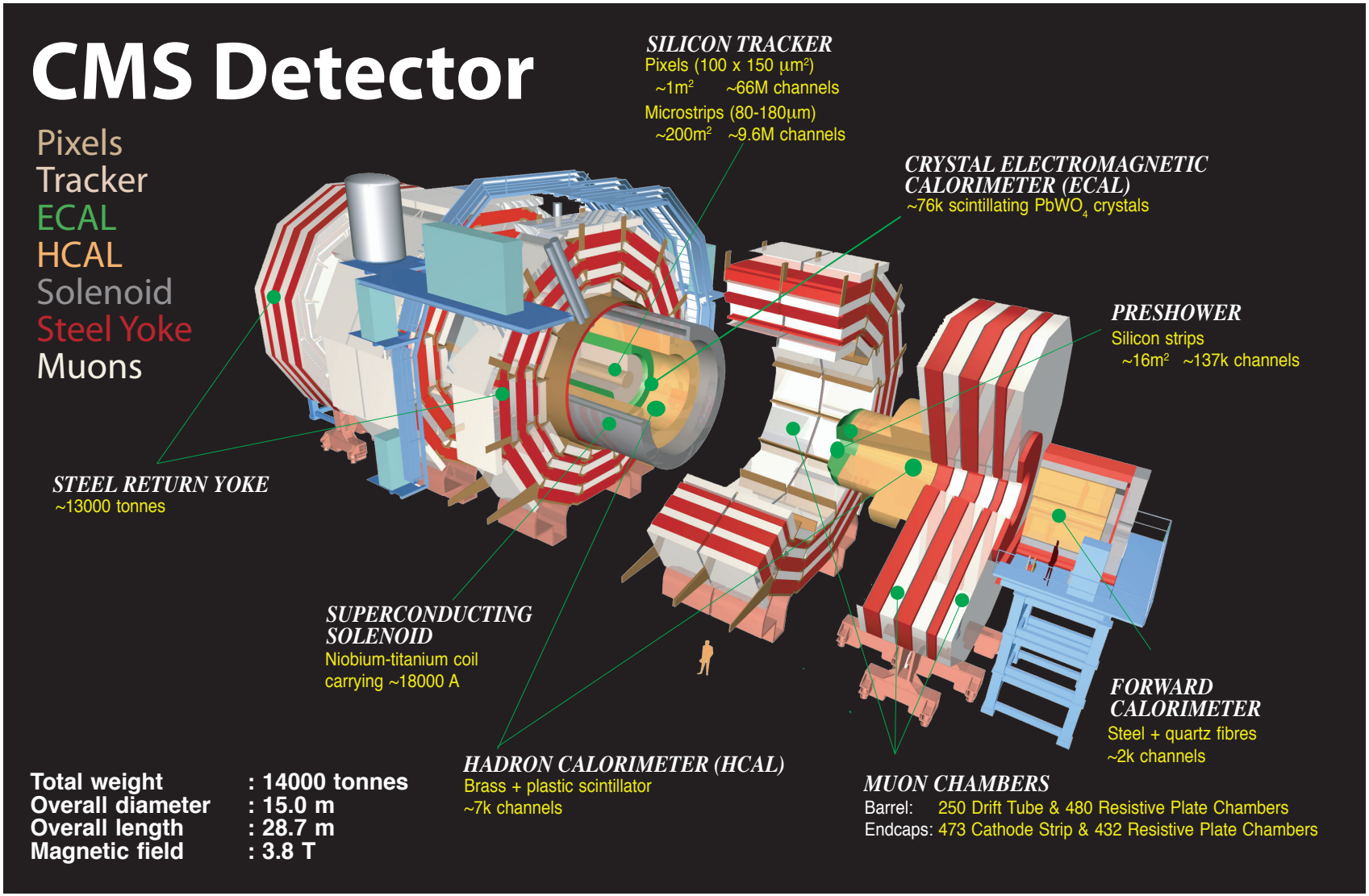
The Large Hadron Collider

- Proton-proton collider
  27 km in circumference,
  located at CERN in Geneva
- Design energy of 14 TeV

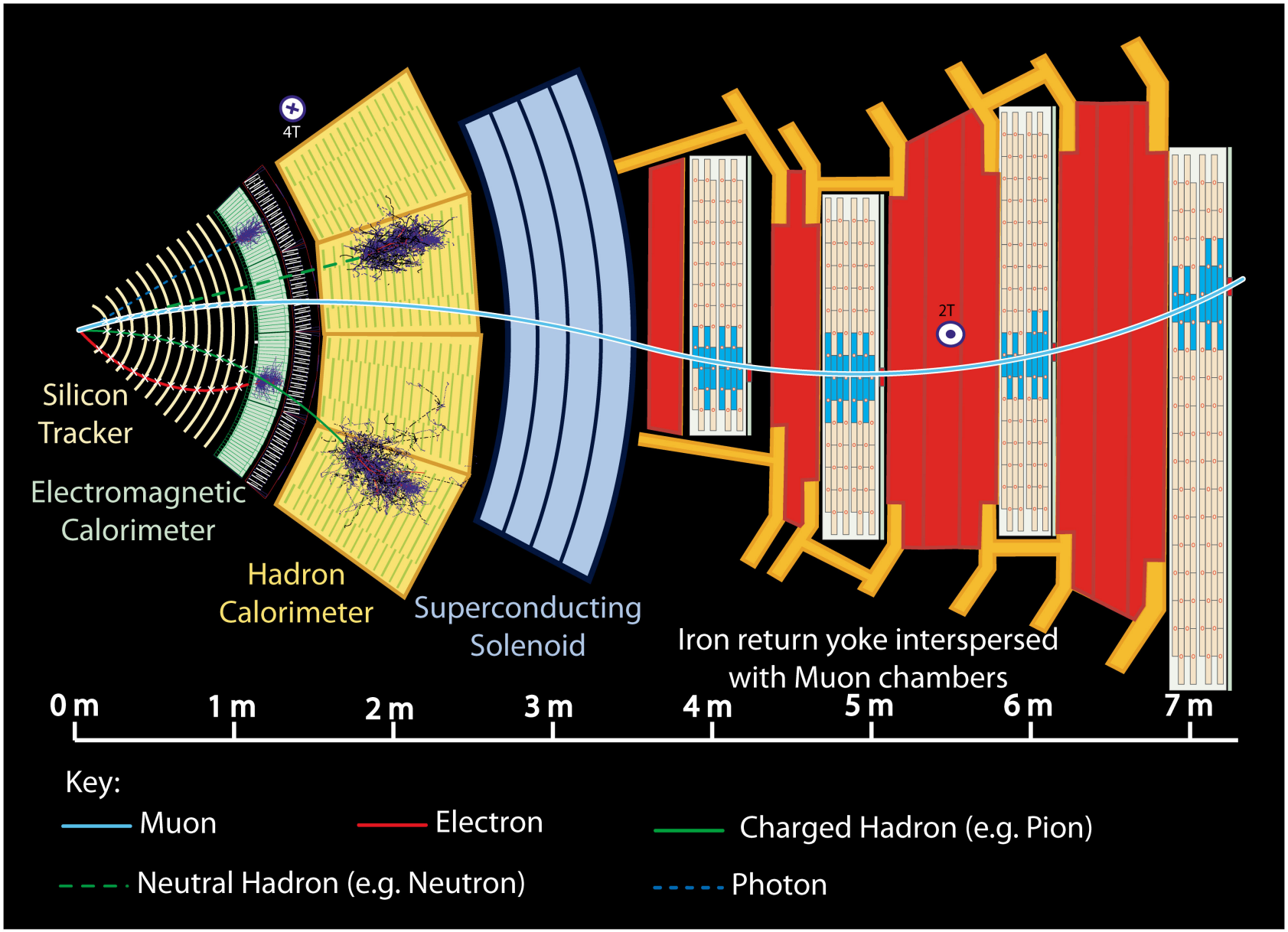


- Superconducting dipole magnets with a design field of 8.3 T, cooled to 1.9 K using superfluid helium

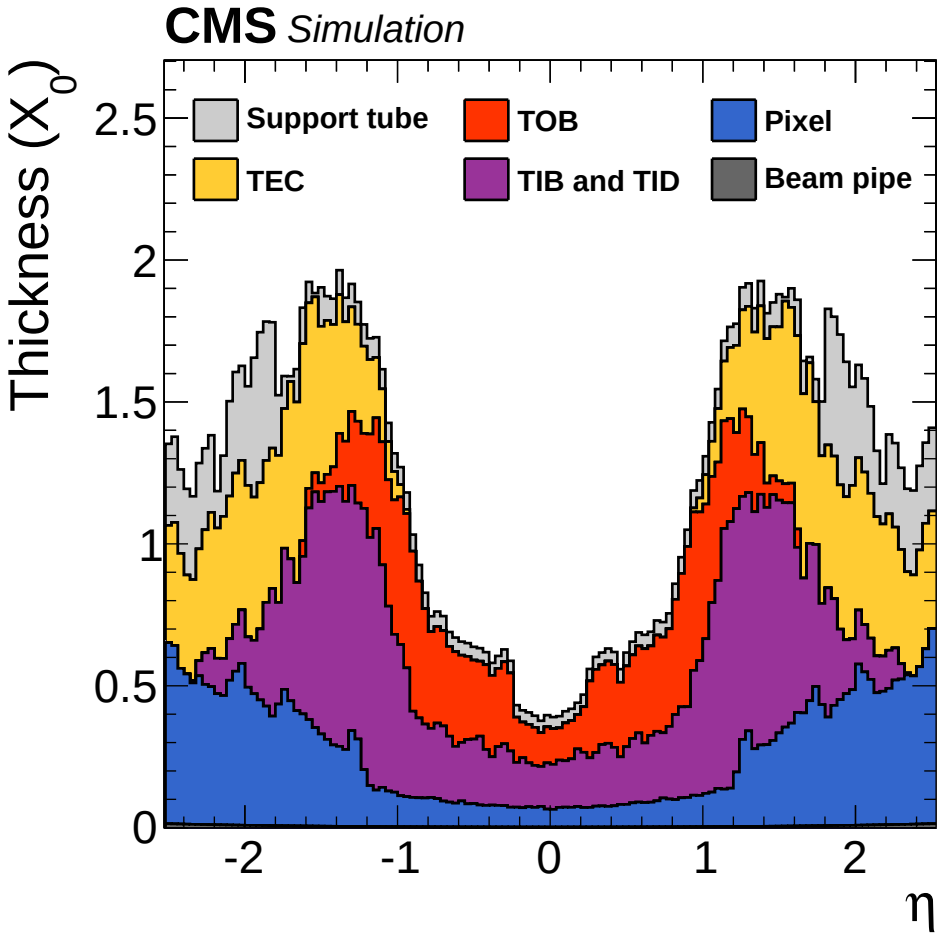
The CMS Detector



The CMS Detector



The CMS Detector: Some Challenges



(a) Tracker Material Budget (pre-2016)

- Lots of material in front of the ECAL
- Induces e.g. bremsstrahlung and conversions for electrons and photons and therefore non-trivial correlations of object properties with η

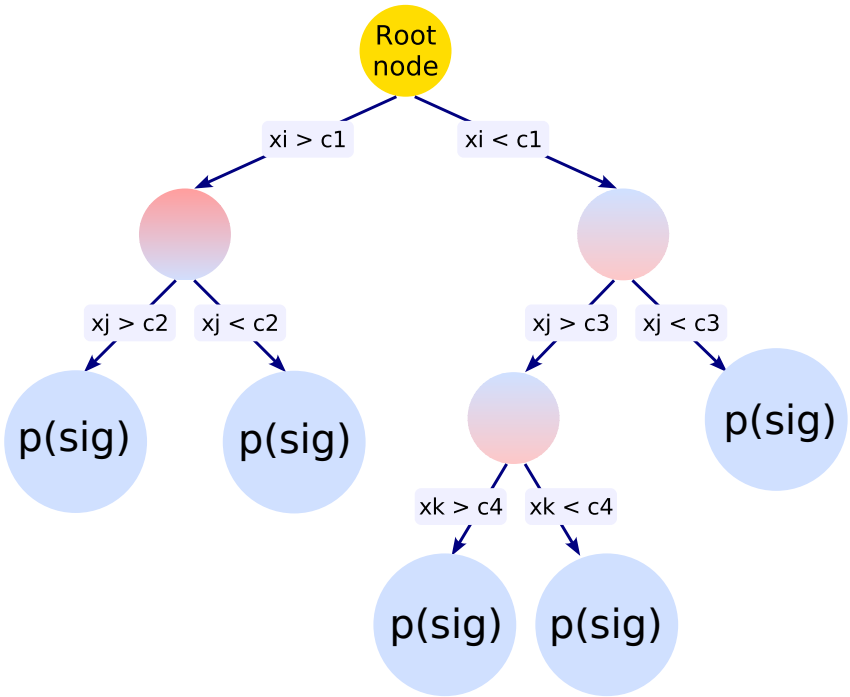
Multivariate Analysis Techniques/Machine Learning at the LHC

- LHC data is valuable and finite
- Physics processes have non-trivial and multi-dimensional underlying kinematics of the produced particles
- Space of observables is even further expanded by the interaction with and measurement by the detectors
- Need to maximally exploit the large amount of information in each collision event
- Example: Optimal discrimination between signal and background from full multidimensional log-likelihood ratio $L_R = \frac{\mathcal{L}_s(\bar{x})}{\mathcal{L}_s(\bar{x})+\mathcal{L}_b(\bar{x})}$
- Not known analytically in general, need to estimate from finite Data or Monte Carlo "training" samples
- Machine learning classifier typically implemented with Boosted Decision Tree (BDT) or Artificial Neural Networks

Overview

- Brief intro to basic machine learning techniques
- Illustrative examples of machine learning used in several contexts:
  - Physics object identification in CMS
  - Physics object reconstruction in CMS
  - High-level analysis (Higgs search/observation/measurements in CMS)
  - Monte Carlo generation/simulation
- Future prospects

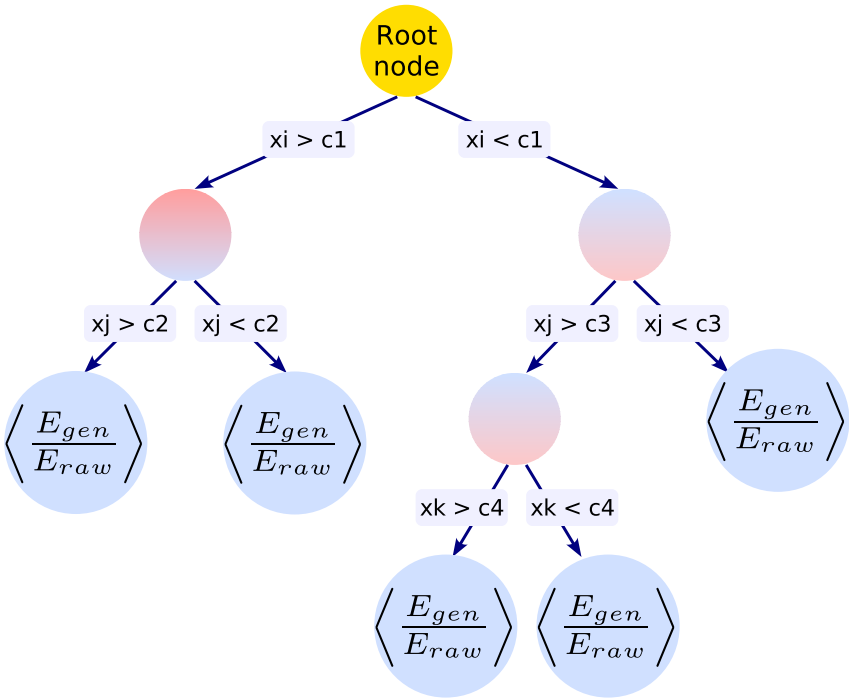
Boosted Decision Trees for Classification



- Decision Tree is a simple structure consisting of a set of connected "nodes"

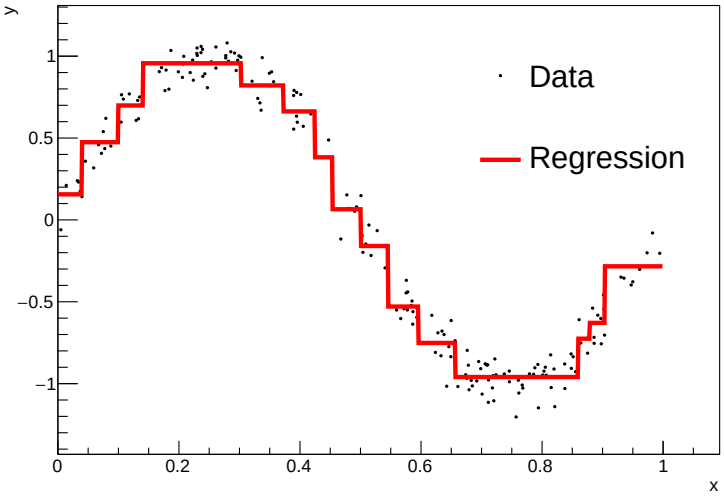
- Intermediate nodes where a variable and cut value is selected to split events into two subsets
- Terminal nodes are assigned a response, in this case the relative signal probability $\frac{\mathcal{L}_s(\vec{x})}{\mathcal{L}_b(\vec{x})}$
- Multidimensional likelihood ratio is therefore approximated by a piecewise-continuous function over the multivariate input space
- **Boosting**: Construct an iterative series of decision trees to improve the overall response

Boosted Decision Trees for Regression

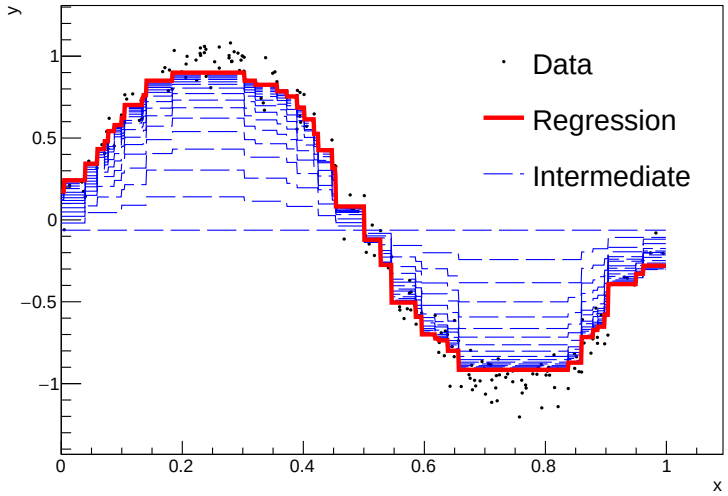


- Boosted Decision Trees can also be used for a multivariate regression problem
- Replace log likelihood ratio with generic function $f(\bar{x})$
  e.g $f(\bar{x}) \equiv \left\langle \frac{E_{True}}{E_{Raw}} \right\rangle (\bar{x})$
- Minimize deviation between training sample and regression function
- Decision trees form a series of piecewise continuous approximations for the function $f(\bar{x})$ in the multidimensional input space

Gradient Boosting



(a) Single Tree

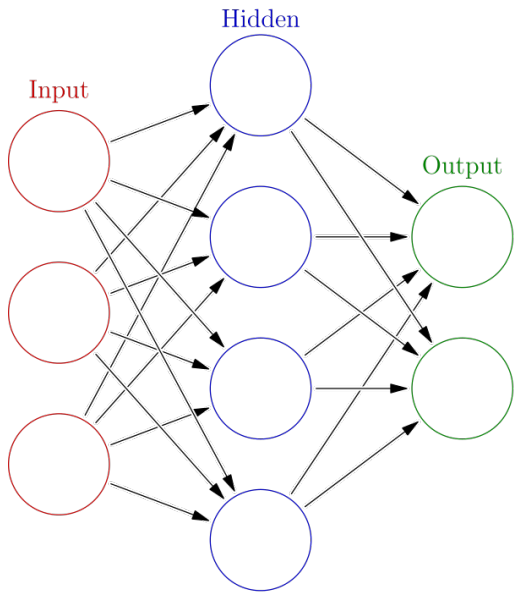


(b) Gradient Boosted ($\sim$ 20 trees)

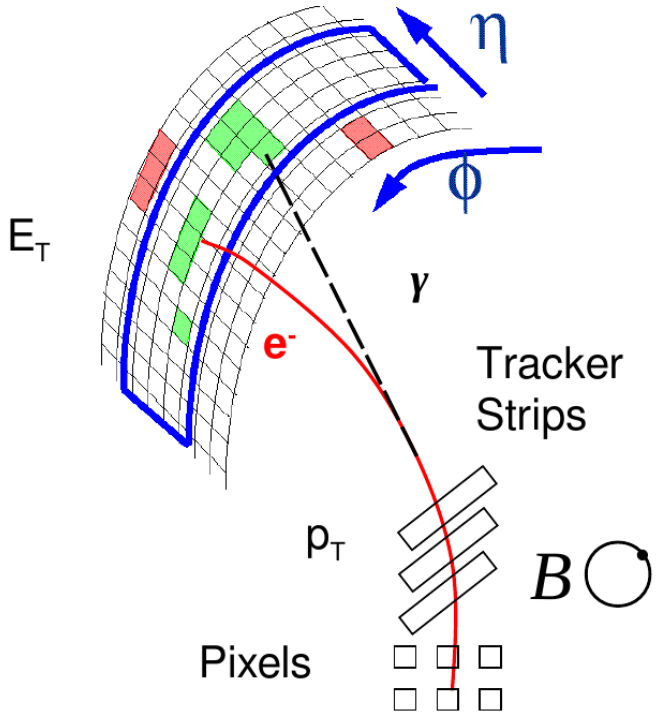
- Decision trees form an additive series of piecewise continuous approximations for the function $f(\bar{x})$ in the multidimensional input space
- Additive series can represent more complex functions than single tree with a given number of nodes
- Trivial example of Sine in 1d with relatively few trees

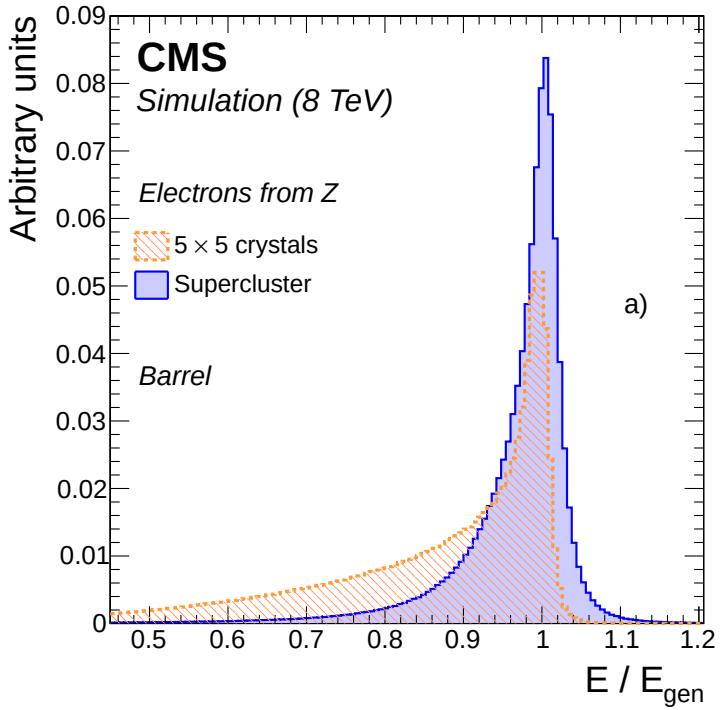
Artificial Neural Networks

- Inspired by biology, artificial neural networks comprise one or more layers of artificial neurons with **weight**, **bias**, and **activation function** with many possible architectures for how the neurons/layers are connected
- Already the simple "densely connected" neural network with non-linear activation functions can serve as a universal function approximator in a similar manner to BDT's
- Such neural networks can be trained for classification or regression problems with the appropriate loss function
- Training = finding optimal values for weights and biases to minimize the loss function using some variation of Stochastic Gradient Descent



Electron Identification

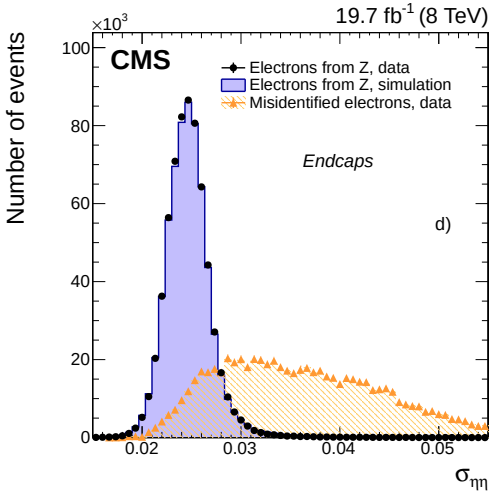




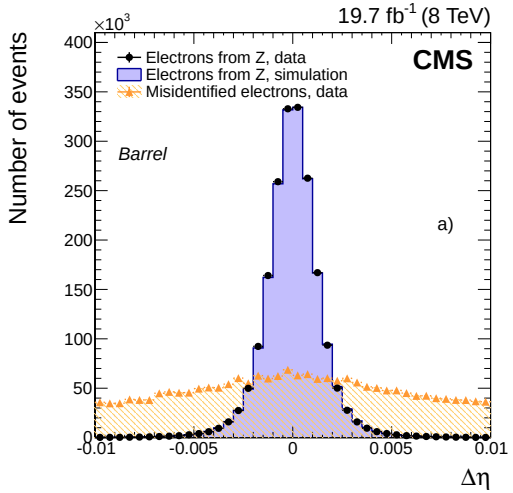
- Reconstruction forms Superclusters extended in ϕ to collect conversion legs/bremsstrahlung spread out by magnetic field
- Soft conversion legs and associated bremsstrahlung may not reach calorimeter or arrive too far to be included in Supercluster

Electron Identification

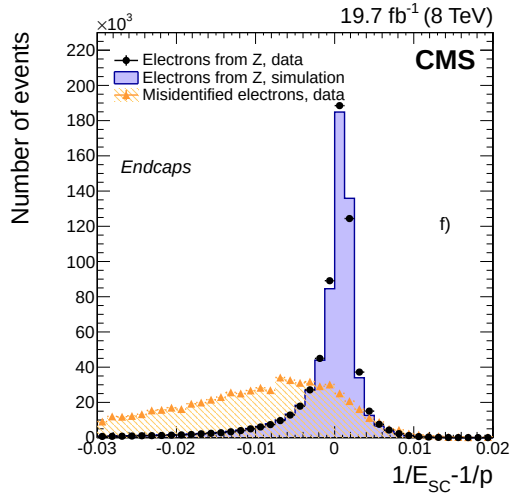
- **Main sources of mis-identified electrons:** $\pi^{\pm}$ inelastic charge exchange, γ conversions (prompt or from $\pi^0 \to \gamma\gamma$), semi-leptonic heavy flavour decays
- Distinguish with isolation, and with electromagnetic shower profile, track properties, and track-cluster compatibility
- Many variables with non-trivial correlations $\to$ BDT classifier on shower/track/compatibility variables



(a) Transverse Shower Width

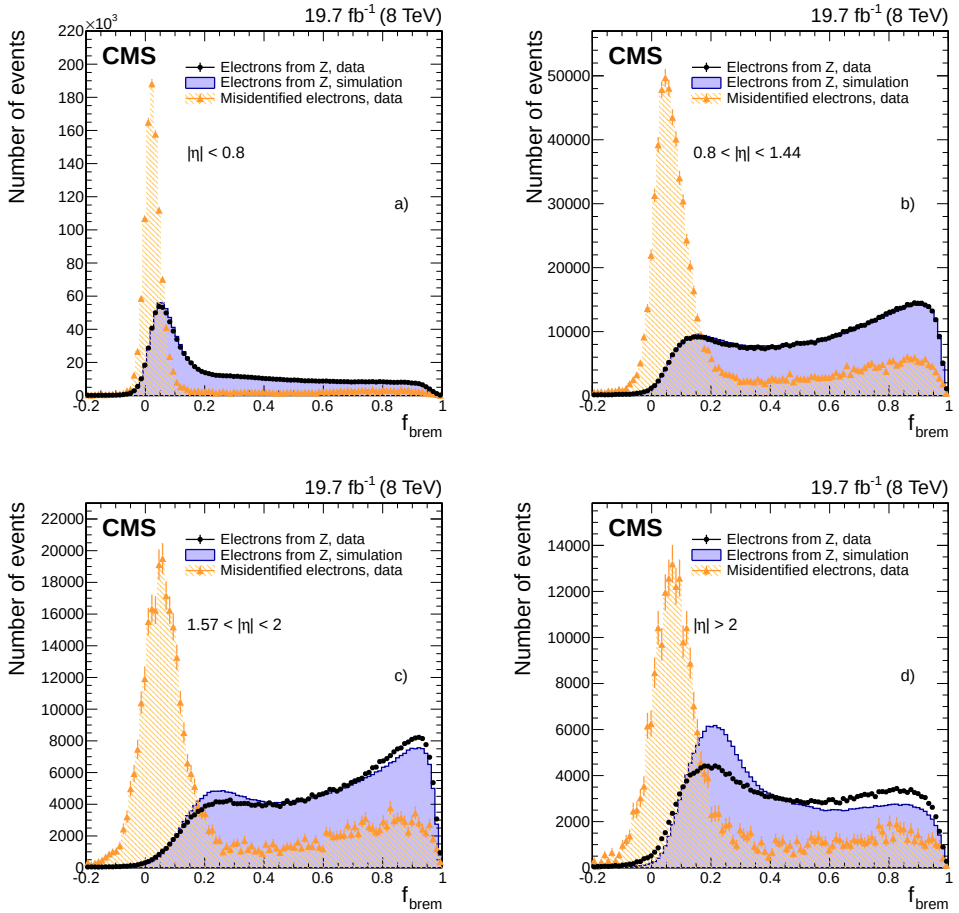


(b) Track-Cluster Match



(c) Energy-Momentum Match

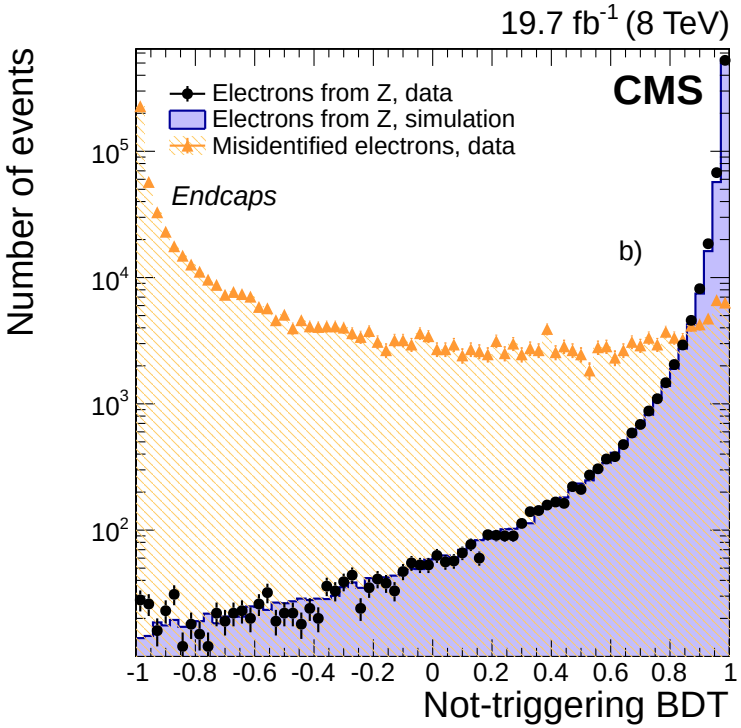
Electron Identification



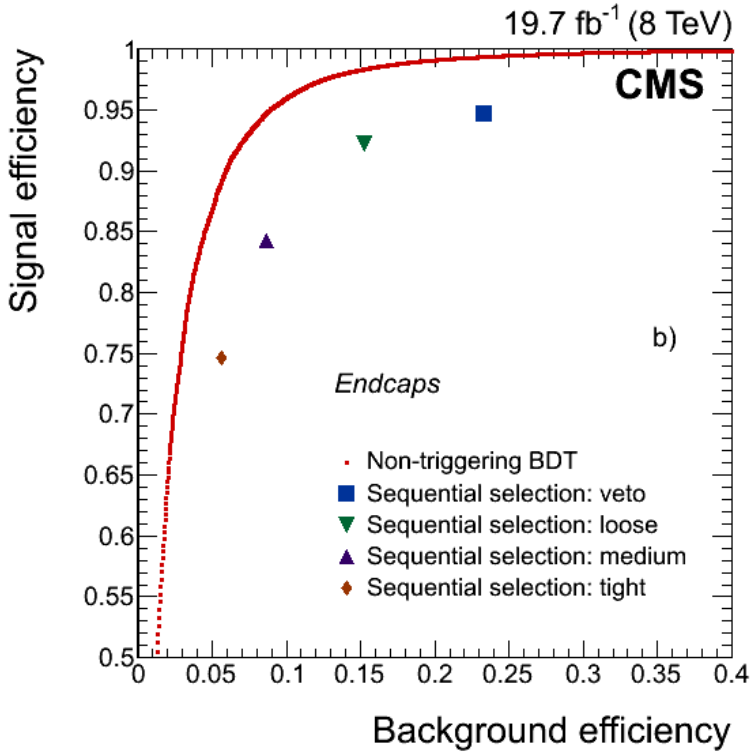
- Non-trivial correlations such as material-induced evolution of discriminating variables with η
- f_{brem}: Fraction of initial momentum radiated as measured by track reconstruction shown in different slices of η

Electron Identification

- Many variables with non-trivial correlation
- BDT classifier on shower/track/compatibility variables greatly improves signal-background discrimination compared to rectangular cuts



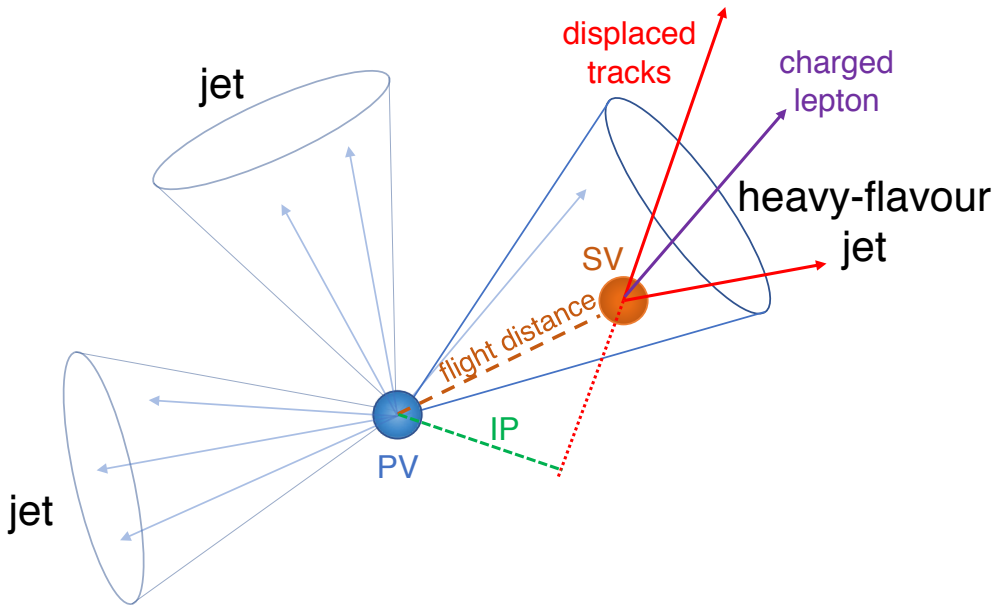
(a) BDT Output Distribution



(b) ROC Curve

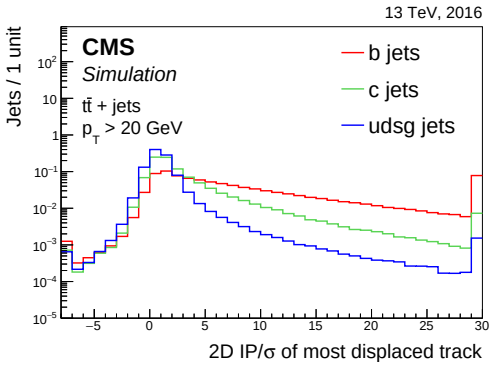
b-jet identification

- Discriminating b-jets from light flavour (or c-jets) crucial for top physics, many Higgs final states, and many BSM searches
- b-jets are characterized by displaced tracks and possible reconstructed secondary vertices

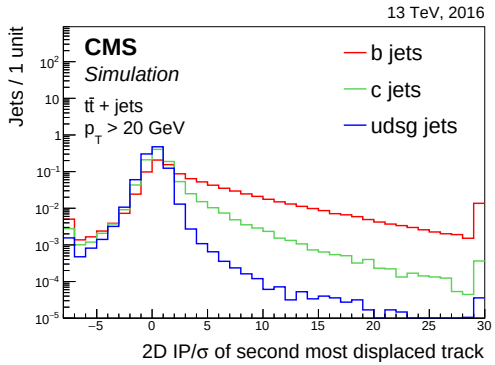


b-jet identification

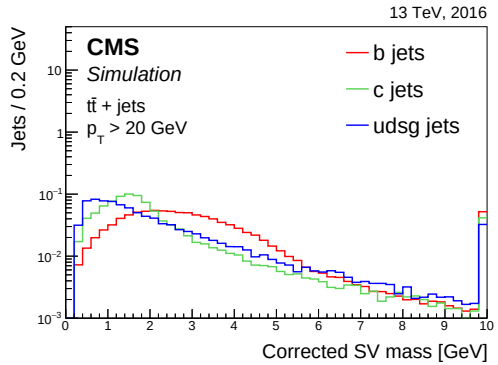
- Relevant information on kinematics, displacement, etc is in principle available for **each** reconstructed particle or secondary vertex



(a) Highest 2d IP/σ



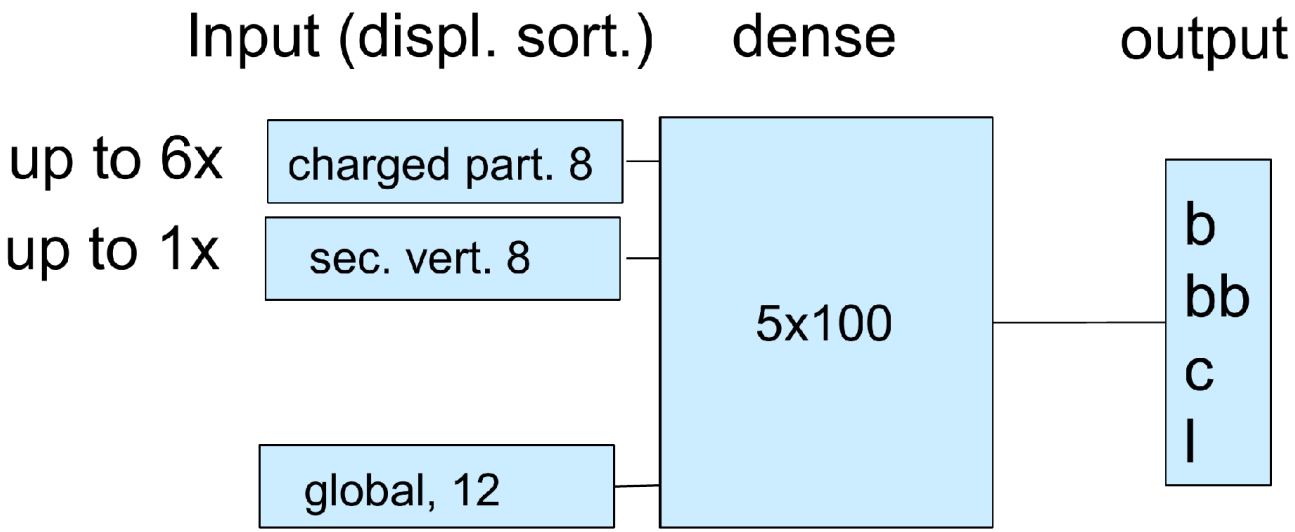
(b) 2nd Highest 2d IP/σ



(c) SV Mass

b-jet identification: "DeepCSV"

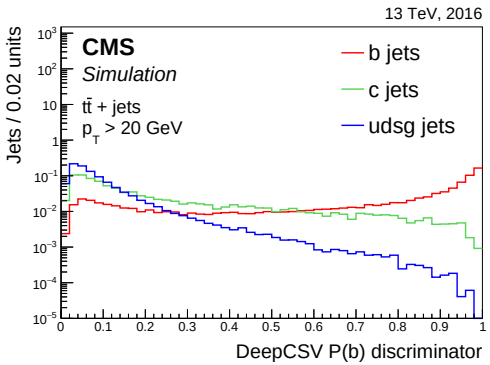
- Most advanced fully commissioned b-tagging in CMS uses densely connected deep (5 layers * 100 node) neural network with some global information, plus detailed info from up to 6 tracks and 1 secondary vertex



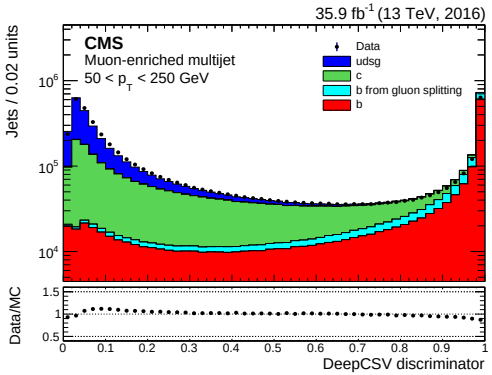
M. Stoye, DS@HEP 2017

b-jet identification: "DeepCSV"

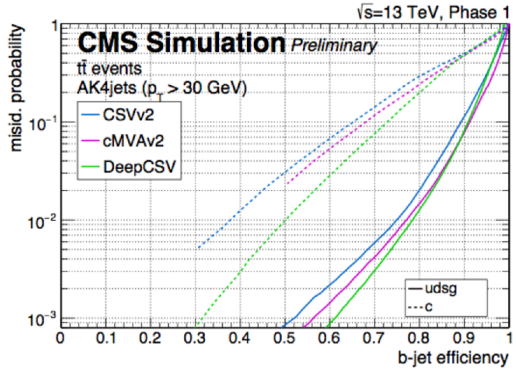
- All b-taggers validated in data, with efficiencies and mistag rates measured from b/light-flavor enriched control regions



(a) DNN output



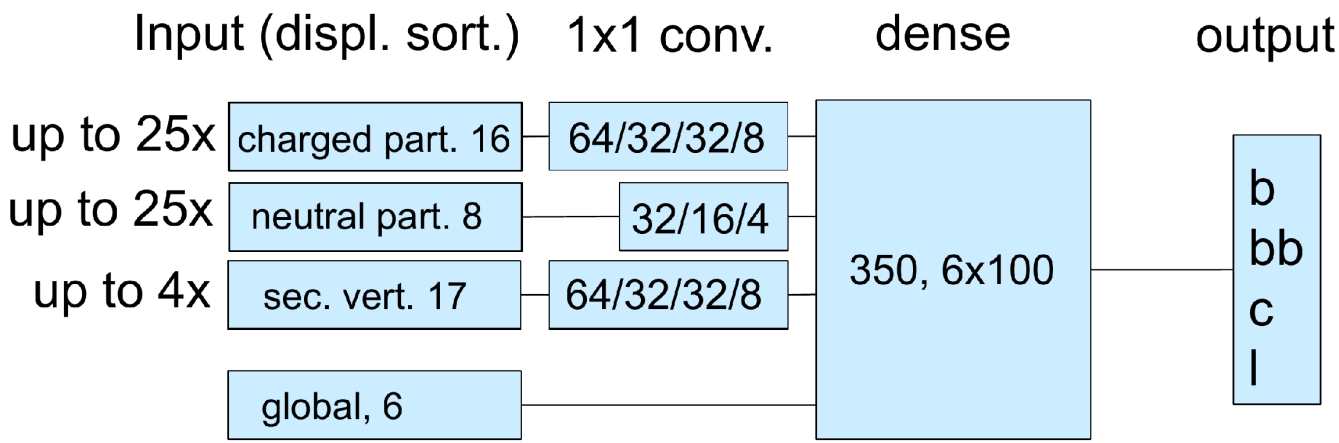
(b) DNN output



(c) ROC Curves

b-jet identification: "DeepFlavor"

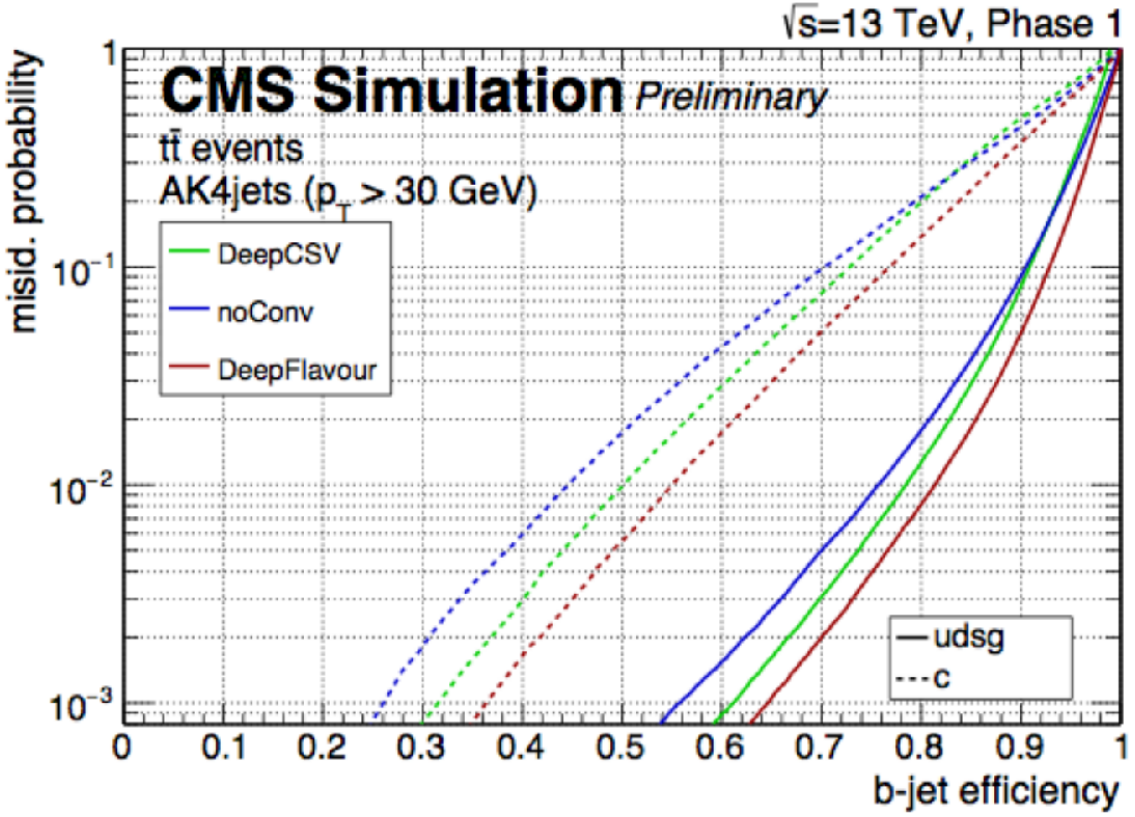
- More recent developments incorporate more advanced network architecture
- Convolutional layers employed at the particle and secondary vertex level to significantly increase the amount of available information and number of particles/SVs



M. Stoye, DS@HEP 2017

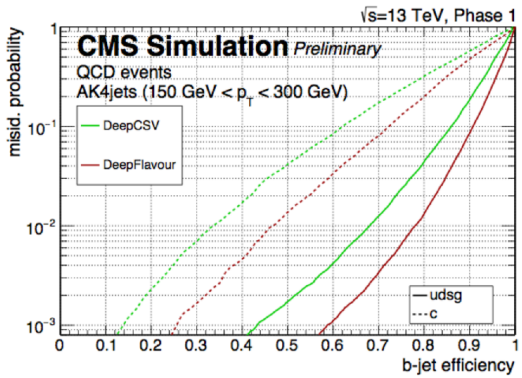
b-jet identification: "DeepFlavor"

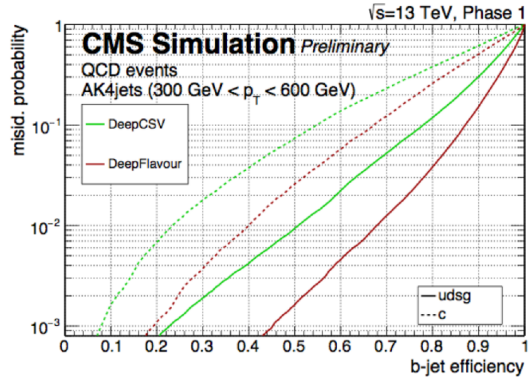
- More recent developments incorporate more advanced network architecture
- Convolutional layers employed at the particle and secondary vertex level to significantly increase the amount of available information and number of particles/SVs

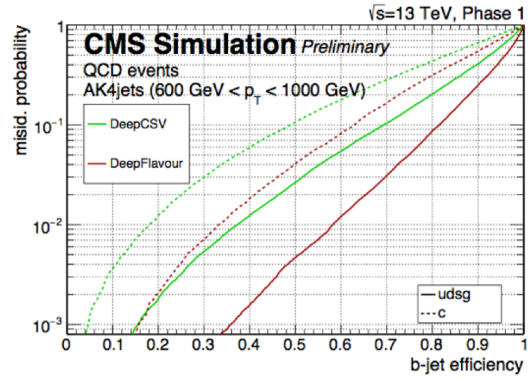


b-jet identification: "DeepFlavor"

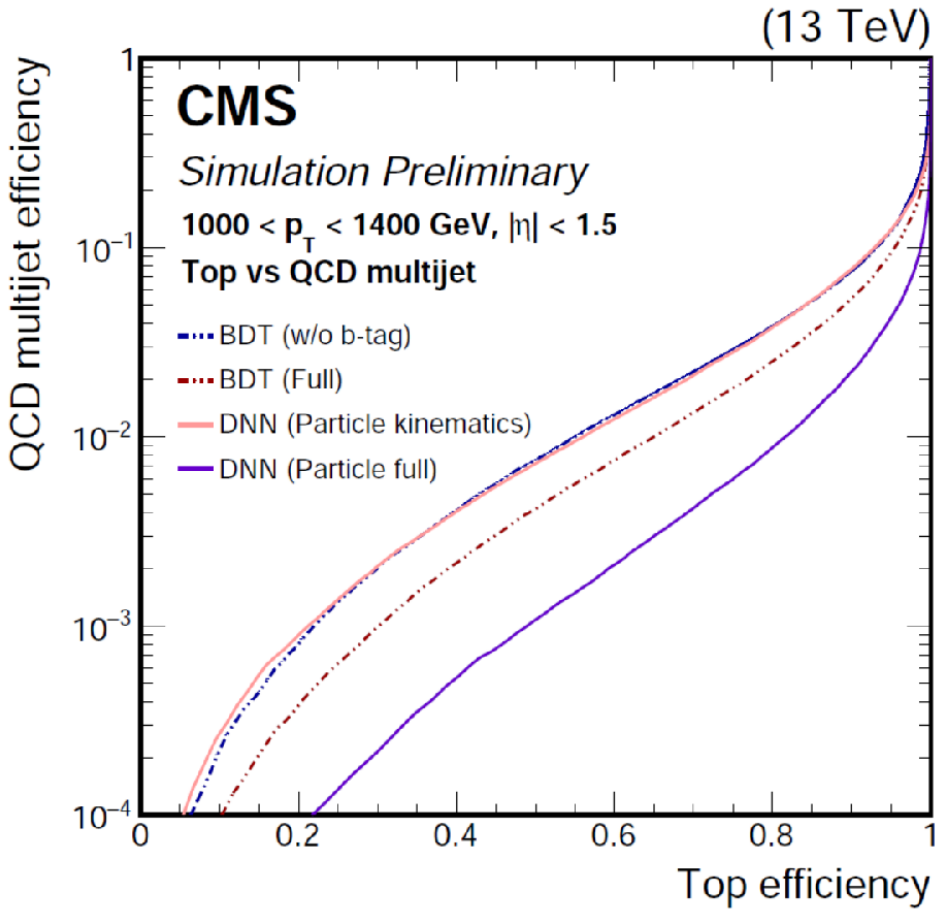
- More recent developments incorporate more advanced network architecture
- Convolutional layers employed at the particle and secondary vertex level to significantly increase the amount of available information and number of particles/SVs
- Relative gain increases with the p_T of the b-jet (existing algorithms loose discriminating power faster)







Boosted top-jet identification



- DeepFlavor architecture used for b-tagging adapted to the identification of boosted top jets (collimated top decay products at $p_T >> m_T$)
- Incorporating convolutional layers over a larger number of particles and secondary vertices within each jet

Photon (Electron) Regression Energy Corrections

- Photon energy reconstruction in CMS:

$$E_{e/\gamma} = F_{e,\gamma}(\bar{x}) \times \sum_{i}^{N_{crystals}} G(GeV/ADC) \times S_i(t) \times c_i \times A_i$$

- Two main components to photon energy resolution which at least partly factorize:
  1. Crystal level calibration (ADCtoGEV, Intercalibration, transparency corrections)
  2. Higher level reconstruction (**local containment, global containment, PU contamination**)
- Shower containment is complex and not clear if/how different contributions factorize
- Best performance is obtained with multivariate regression using BDT with cluster η, ϕ, shower shape variables, local coordinates, and number of primary vertices/median energy density as input
- Regression is trained on real electrons/photons in Monte Carlo, using the ratio of the generator level energy to the raw cluster energy, also provides a per photon estimate of the energy resolution

Evolution of Regression Energy Corrections in CMS

- Photon energy regression in CMS initially trained using TMVA BDT implementation
- Physics performance was ok, but serious problems with size on disk and memory consumption (1GB xml files!)
- CMS has an in-house BDT storage format, persistable in root file or conditions database, disk/memory/cpu efficient (tree structure represented in flattened arrays, one inlined while loop for evaluation). Can convert weights from TMVA or produce with native BDT training tool written to exploit parallelization, speed up training with large datasets, produce more compact trees
- Later CMS moved to "semi-parametric" regression

Evolution of Regression Energy Corrections in CMS: "Traditional" Regression

- Multivariate techniques used in general to overcome lack of knowledge of multidimensional likelihood using finite event samples
- Traditional regression as used so far based on minimization of Huber loss function for target prediction $F(\bar{x})$ given target variable $y = E_{True}/E_{Raw}$ for a set of input variables $\bar{x}$ (in our case cluster position, shower profile and pileup variables)

$$L = \begin{cases} \frac{1}{2}(F-y)^2 & |F-y| \leq \delta \\ \delta\left(|F-y| - \delta/2\right) & |F-y| > \delta \end{cases}$$

- Minimized the square deviation out to some cutoff (by default $\pm 1\sigma$) and the linear deviation beyond that
- No built-in estimate of the per-photon resolution, accomplished with a second training on an independent subset of the training sample with target $y = |E_{Cor}/E_{Raw} - E_{True}/E_{Raw}|$

Semi-parametric Regression

- Start with ansatz that in any infinitesimal slice of phase space in $\bar{x}$, the energy response distribution is given by a double crystal ball (ie gaussian core with power law tails on both sides)
- In terms of E_{True}/E_{Raw} the **right** tail (undermeasurement of the energy) corresponds to the usual radiative losses, etc, whereas the **left** tail (overmeasurement of the energy) comes from pileup, etc.

$$p(y|\bar{x}) = \text{DoubleCrystalBall}\left(y|\mu(\bar{x}), \sigma(\bar{x}), \alpha_{left}(\bar{x}), n_{left}(\bar{x}), \alpha_{right}(\bar{x}), n_{right}(\bar{x})\right)$$

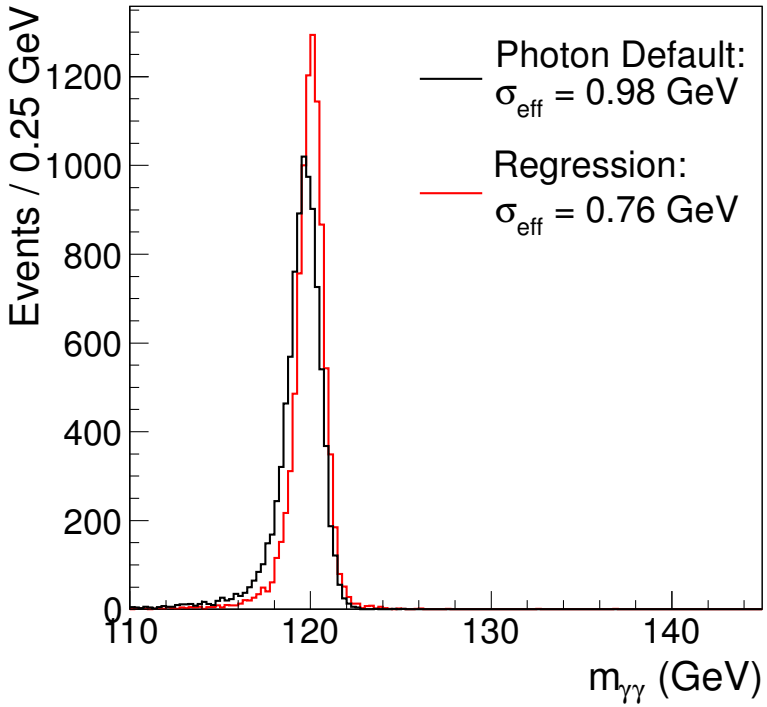
Semi-parametric Regression

- The log likelihood ratio for a training sample can be written simply as

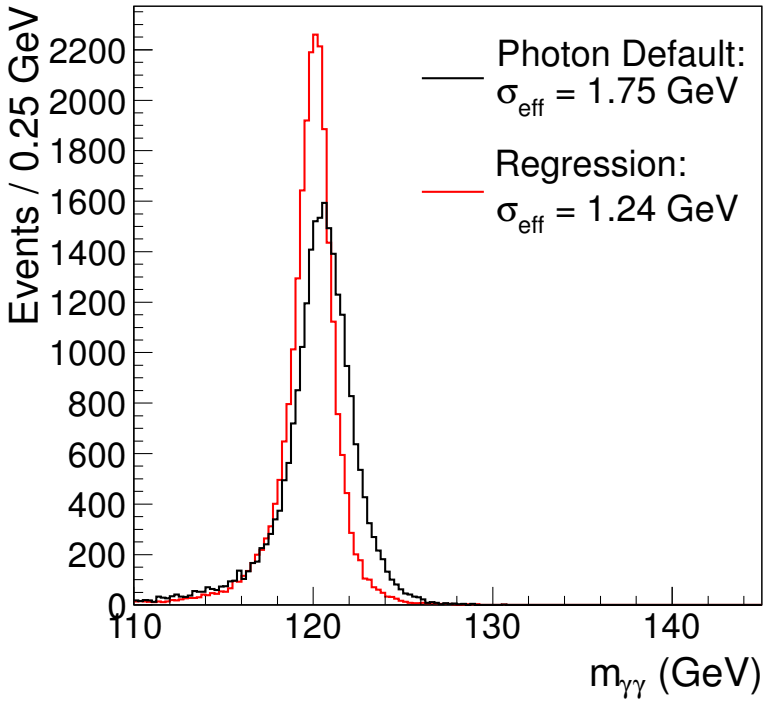
$$L = - \sum_{MCPhotons} \ln p(y|\bar{x})$$

- Minimize this loss function directly with gradient boosting, where $\mu(\bar{x}), \sigma(\bar{x}), n_{left}(\bar{x}), n_{right}(\bar{x})$ are regression outputs estimated by BDT's (using RooFit-based bdt-training tool, which ensures proper pdf normalization, etc)
- This gives a simultaneous estimate for energy correction and resolution among other things

Regression Performance: Simulation



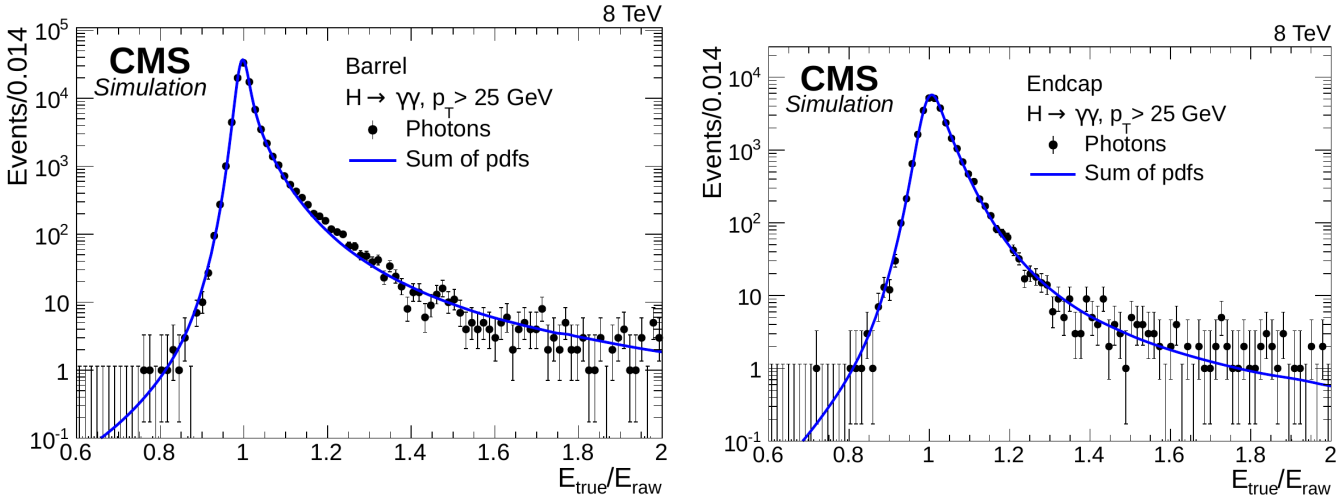
(a) Barrel $\sim$ Unconverted



(b) Barrel $\sim$ at least one converted

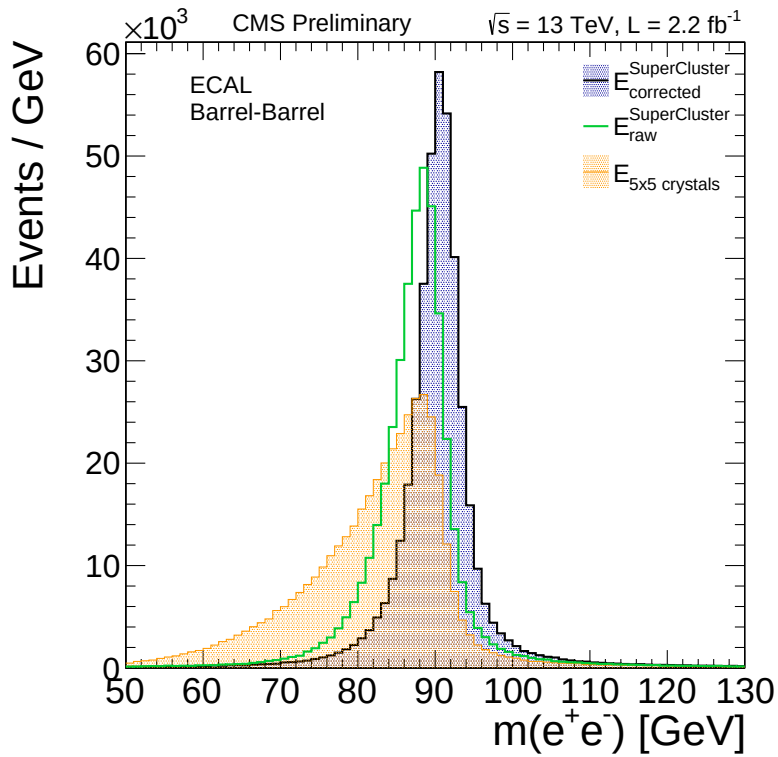
- Substantial improvement in diphoton mass resolution in simulation compared to simpler parameterized corrections (representative plots here)

Energy Regression: Predicted Response Distribution

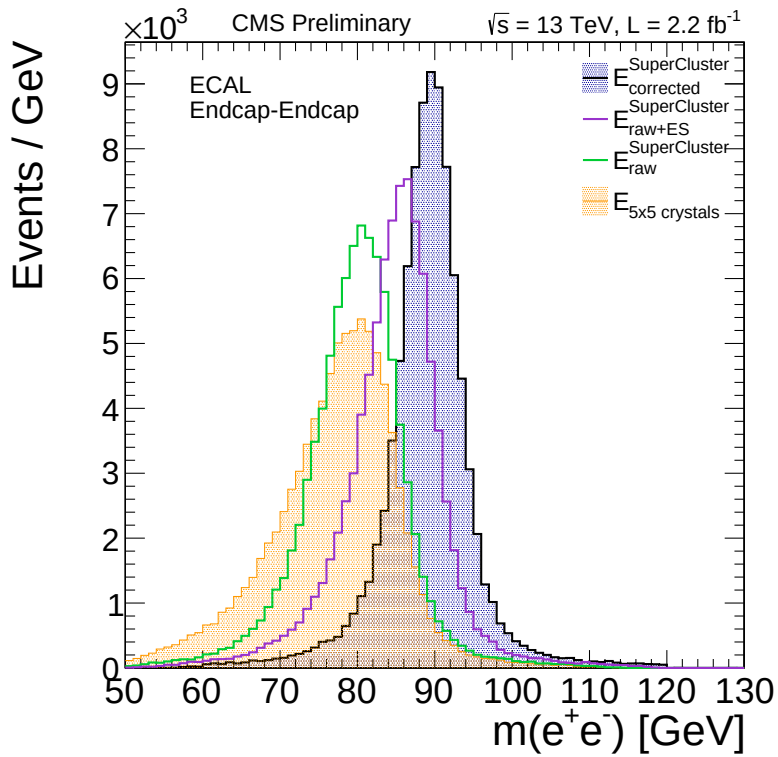


- Semi-parametric regression provides a prediction for the full lineshape (here showing simulation vs regression-prediction for target variable E_{True}/E_{Raw})
- Total predicted pdf is given by sum of predicted lineshape for each simulation event

Energy Reconstruction: Data



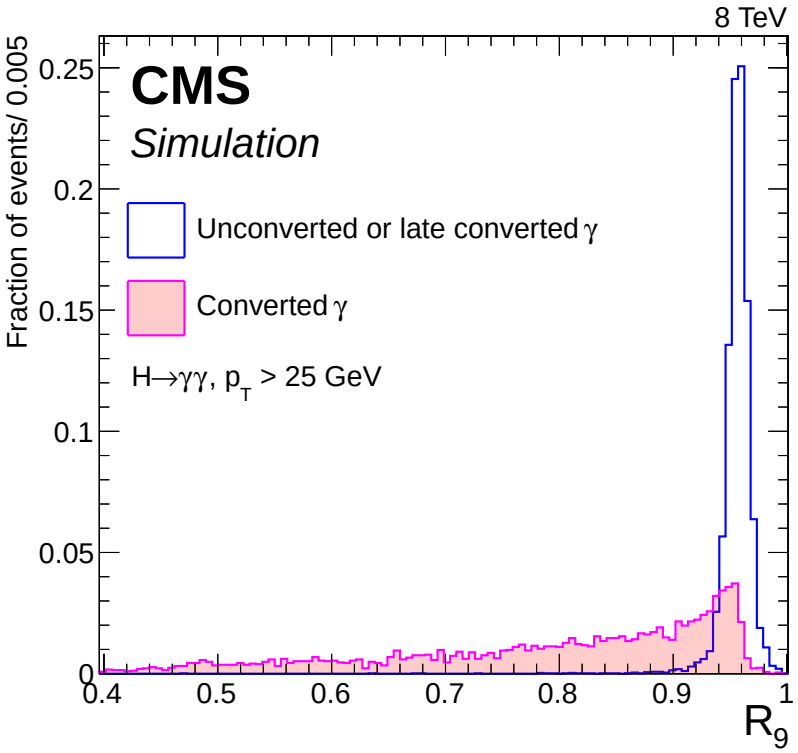
(a) Barrel

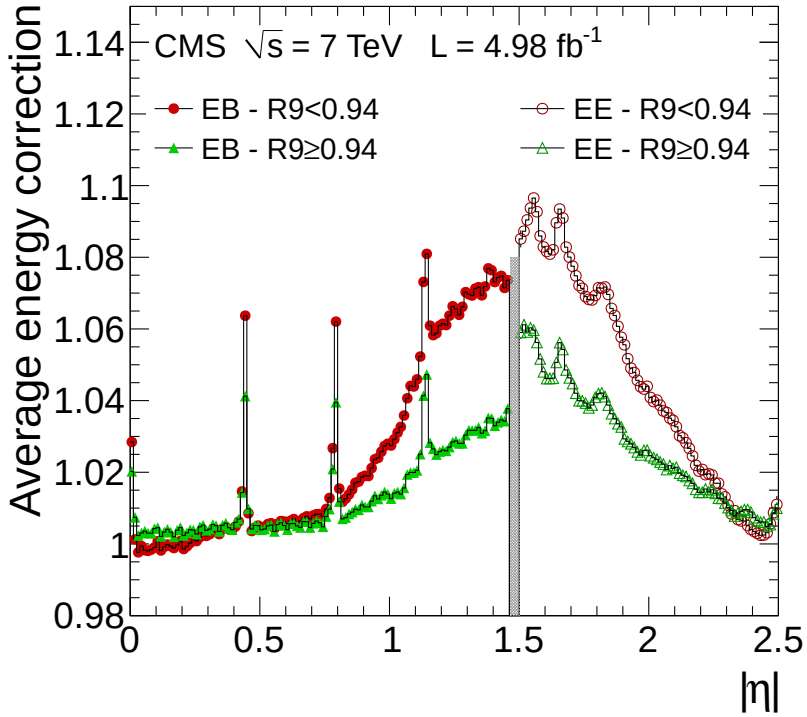


(b) Endcap

- Reconstructed Z mass in data with different levels of energy reconstruction and corrections
- Progression clearly visible even with 2.5 GeV natural Z width

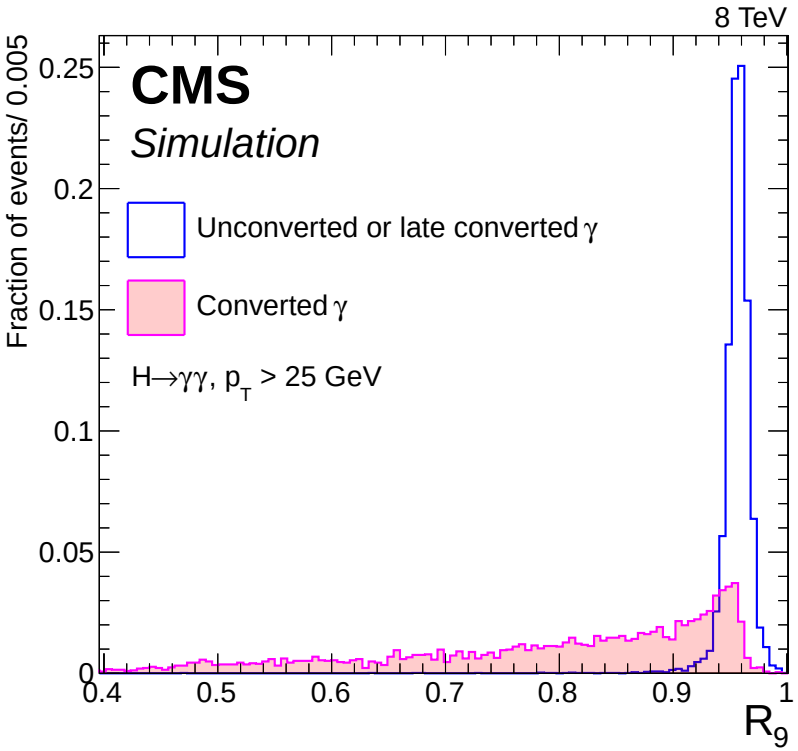
Inside the corrections

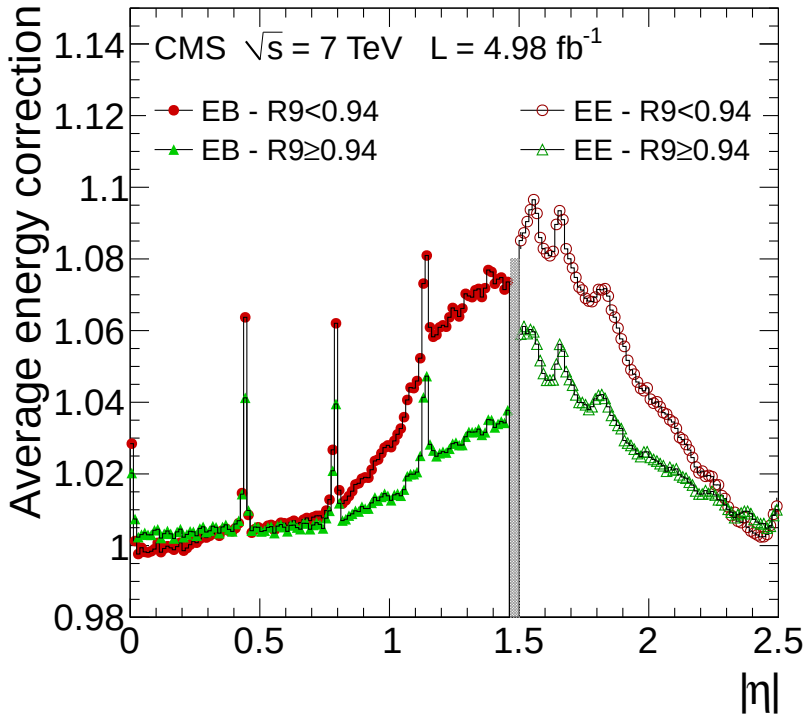




- $R_9 = E_{3x3}/E_{SC}$ is an effective, but not 100% pure conversion tagging variable (electrons and photons treated separately, no explicit converted vs unconverted distinction)
- Correction vs η has a non-trivial correlation with R_9 (and other shower profile variables)

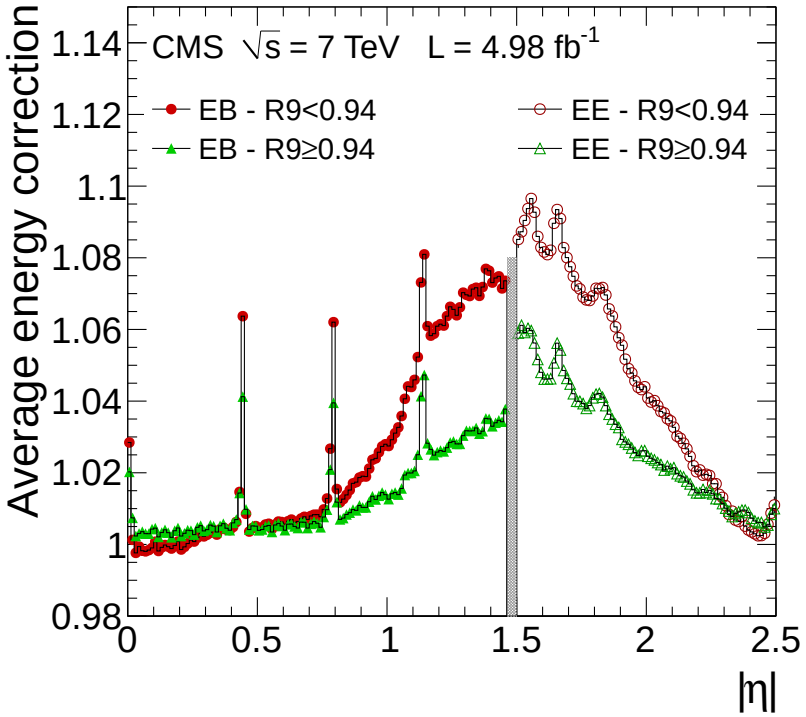
Inside the corrections



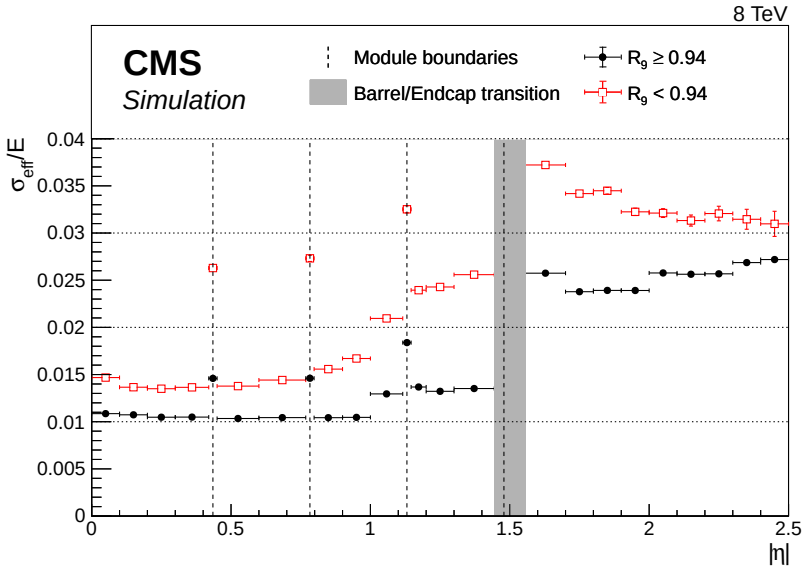


- Correction is parametrized and plotted with respect to the supercluster energy, but the corrected energy can also be considered a non-trivial weighting of supercluster, 3x3, 5x5, and other energy sums/ratios in input (dynamic noise/pileup vs containment tradeoff as a function of shower energy, inferred impact position, etc)

Per-photon Resolution Estimate



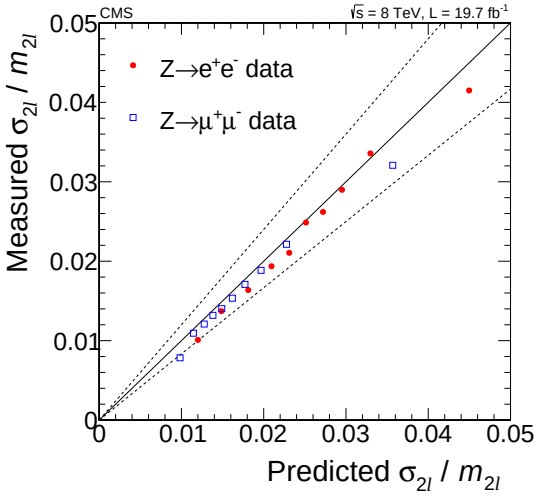
(a) Correction



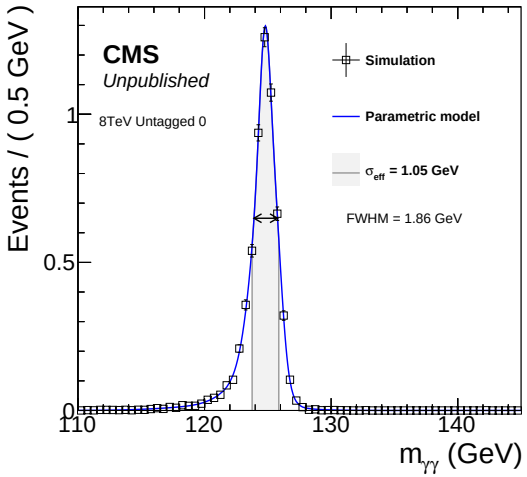
(b) post-correction resolution

- Strong, but non-trivial relationship between size of correction and post-correction resolution (size of effect vs photon-to-photon fluctuations)
- Per-photon resolution estimate mapped with the full granularity of the multidimensional space used to derive the corrections

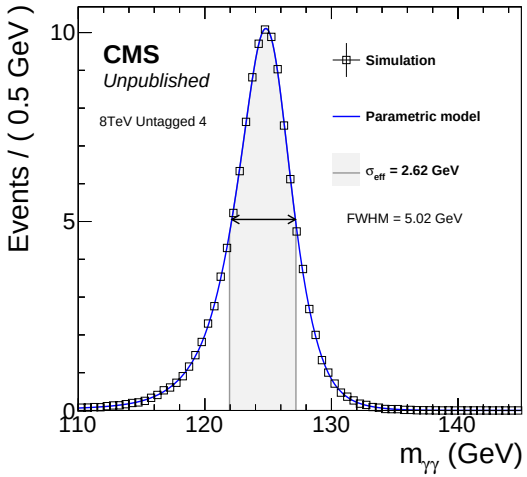
Per-photon Resolution Estimate



(a) Observed vs predicted σ_m



(b) $H \to \gamma\gamma$ Best Category

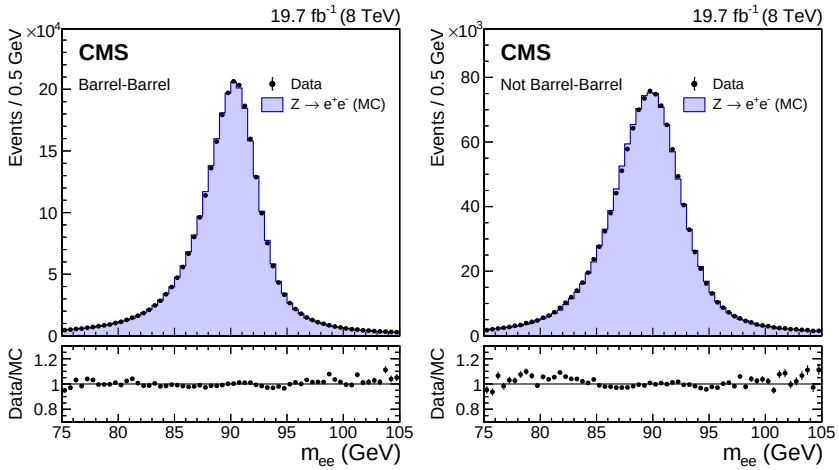


(c) $H \to \gamma\gamma$ Worst Category

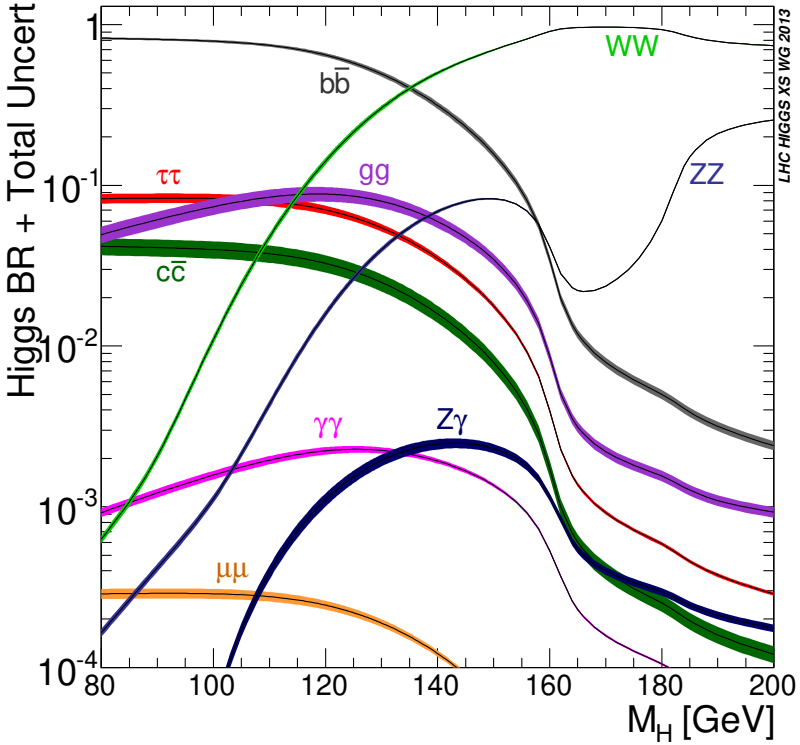
- In a resonance search, per-photon resolution estimate can be used to construct a per-event mass resolution estimate $\frac{\sigma_m}{m_{\gamma\gamma}} = \frac{1}{2}\sqrt{\frac{\sigma_{E1}^2}{E_1^2} + \frac{\sigma_{E2}^2}{E_2^2}}$
- Can be used to select or categorize events to make optimal use of highest resolution events (two unconverted photons in the center of the detector, incident on the center of the crystal, far from module boundaries)

Energy Scale and Resolution

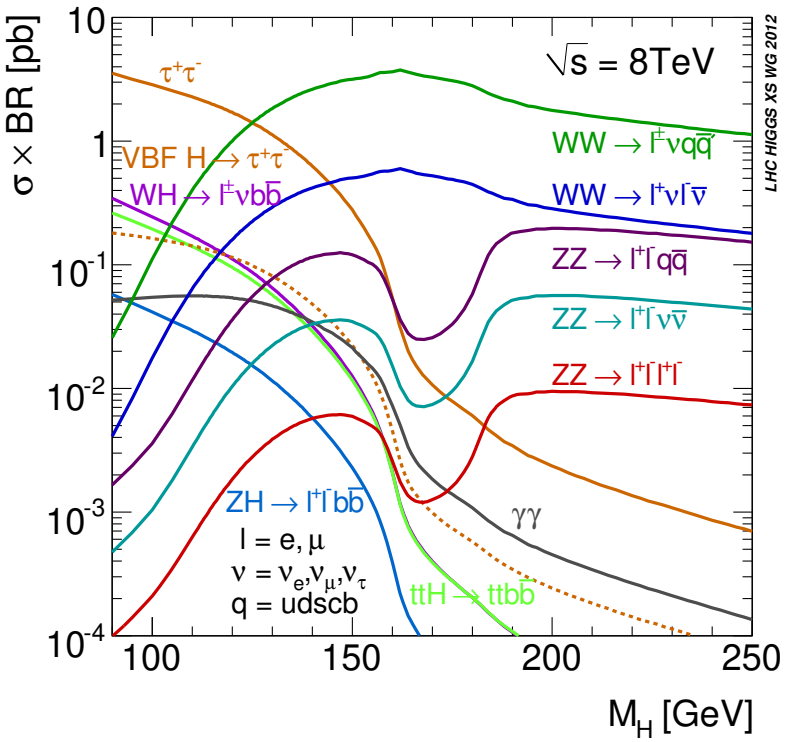
- Photon Energy Scale and Resolution in data measured with $Z \to ee$ events, applying either final photon-trained regression corrections, or equivalent electron-trained version
- Monte Carlo is smeared to match data resolution
- Data energy scale is adjusted to match Monte Carlo
- Energy scale is determined very precisely from (millions of) $Z \to ee$ events, remaining systematic uncertainties from electron-photon extrapolation and extrapolation in energy
- Overall systematic uncertainty on higgs mass measurement (dominated by energy scale uncertainty) 0.12% (but per-photon energy scale uncertainty varies according to detector region and photon quality)



Higgs Production and Decay at LHC



(a) Branching Ratios

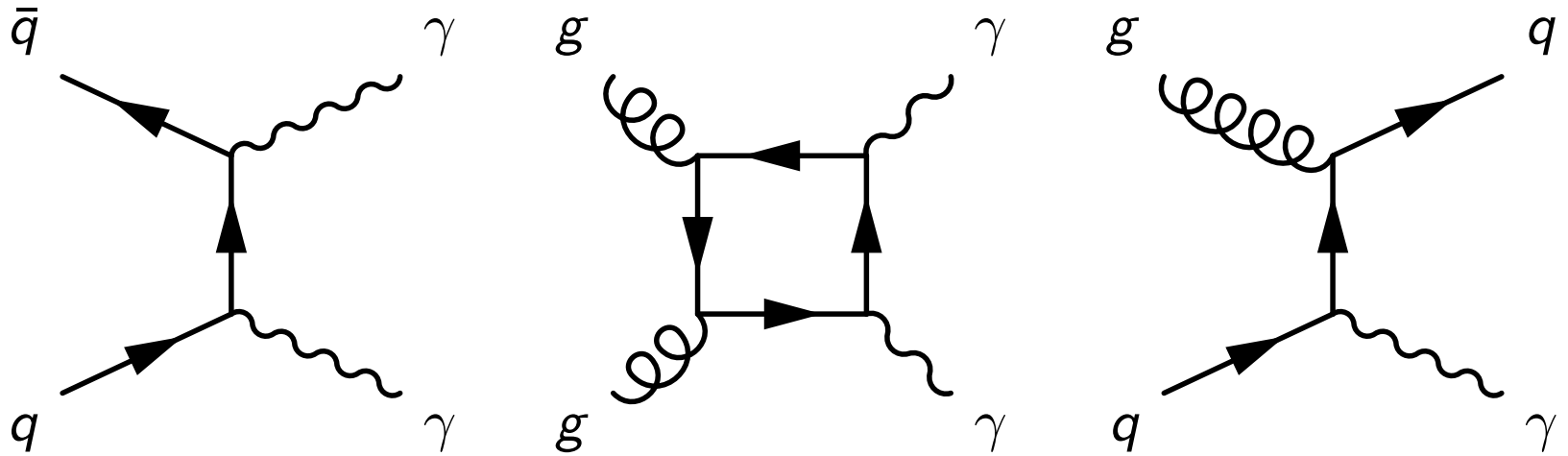


(b) Cross Sections

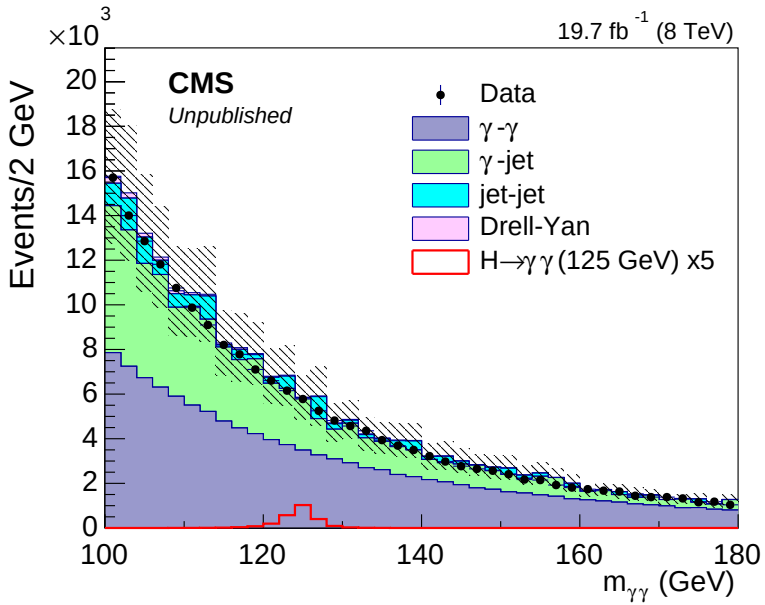
- Variety of final states, would like to extract Higgs signal from as many as possible

Higgs→ $\gamma\gamma$ Analysis Overview

- Higgs→diphoton search at CMS simple in principle: Search for a small but narrow mass peak on a large, smoothly falling background
- Irreducible background from QCD di-photon production, reducible background from QCD γ+jets and multi-jet production with one or more jets faking a photon



Higgs$\to \gamma\gamma$ Analysis Overview

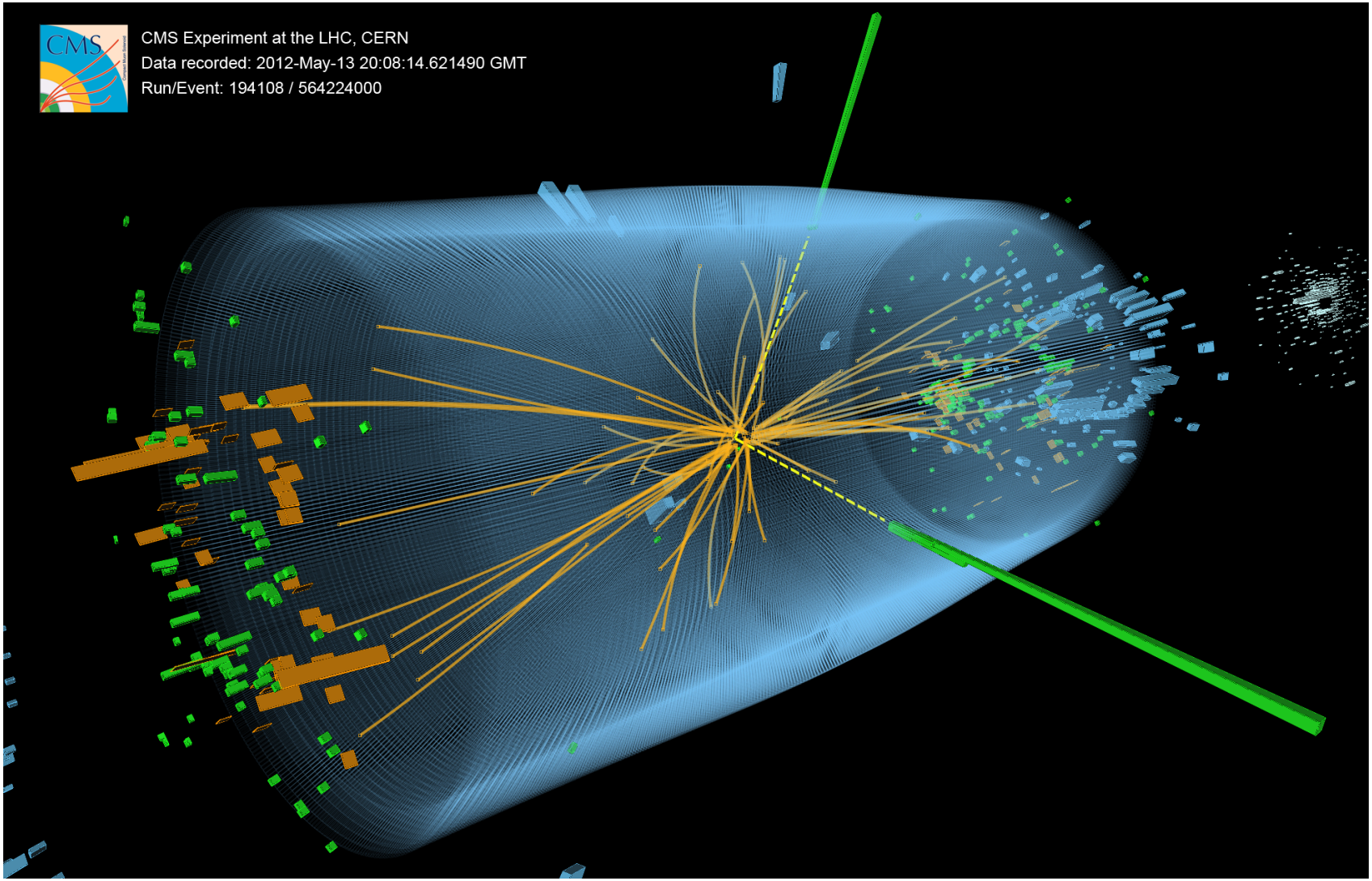


Inclusive selection with coarse binning

$$m_{\gamma\gamma} = \sqrt{2E_1E_2(1-\cos\theta_{12})}$$

- Standard Model search is carried out in inclusive, vector-boson-fusion tagged, W/Z, and $t\bar{t}$ associated production tagged channels
- Analysis makes extensive use of multivariate techniques to optimize the sensitivity, but basic principle of "bump hunt" is preserved

Higgs$\to \gamma\gamma$ Analysis Overview



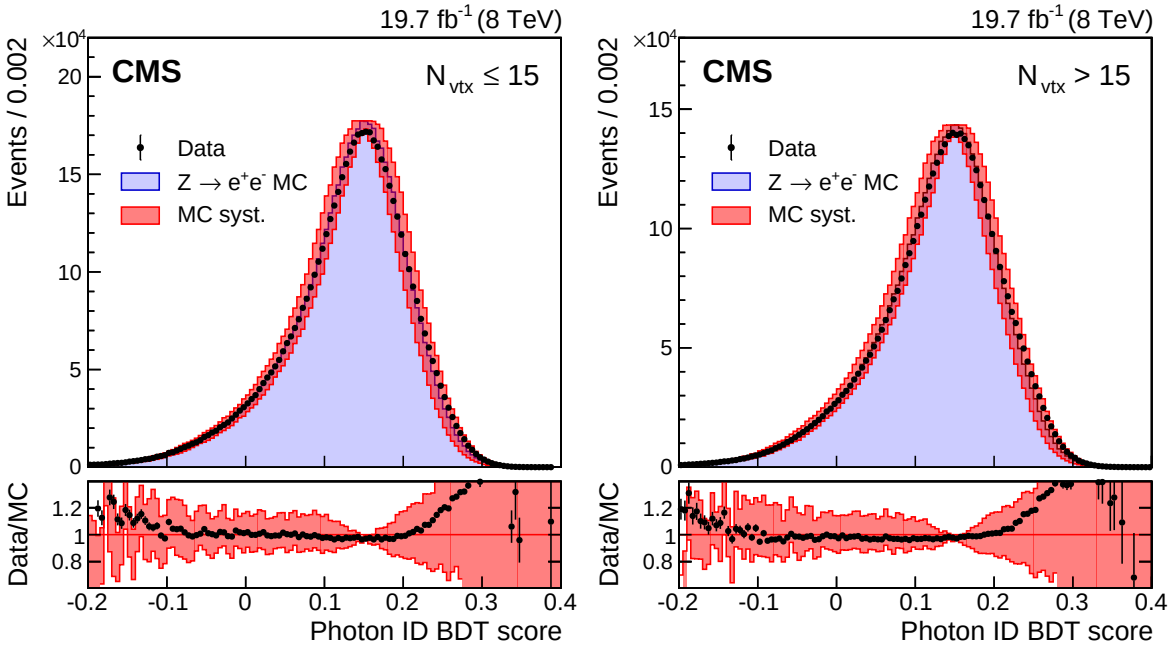
$$m_{\gamma\gamma} = 125.9 \text{ GeV}$$

Higgs$\to \gamma\gamma$ Analysis Overview

1. Primary Vertex Selection (Vertex Selection MVA)
2. Photon Selection (Preselection + Photon-jet MVA discriminator)
3. Multivariate Regression for EM Cluster corrections with per-photon resolution estimate
4. Energy Scale and Resolution corrections from $Z \to ee$
5. Event Categorization (MVA Discriminator)
6. Signal modeling from Monte Carlo with smearing and scale factors applied
7. Background modeling from fit to data
8. Statistical Interpretation: Limits/Significance using maximum likelihood fit to $m_{\gamma\gamma}$ distribution in event categories categories

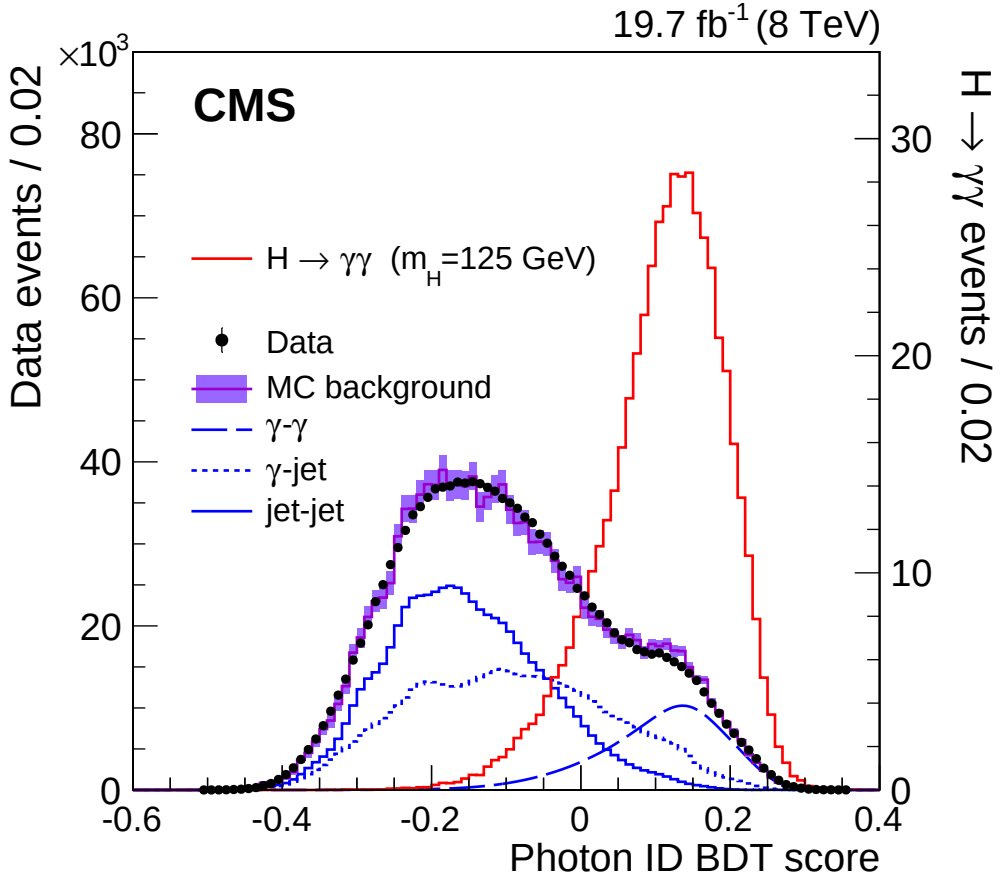
Photon Identification: MVA

- Start with a very loose pre-selection matching trigger requirements
- Construct a multivariate discriminator using a BDT trained on prompt photons vs fakes from jets in MC, using shower and isolation variables as input
- Only a loose cut on the ID MVA value, which is fed forward to the final di-photon MVA
- MVA output shown for $Z \to ee$ events (electron-veto inverted)



Photon Identification: MVA

- Different background components clearly visible in the ID MVA output distribution (though knowledge of the relative fractions is not required for the analysis)



Di-Photon MVA

- Basic Strategy: Train di-photon mva on Signal and Background MC with input variables which are to 1st order independent of $m_{\gamma\gamma}$
- Goal is to encode all relevant information on signal vs background discrimination (aside from $m_{\gamma\gamma}$ itself) into a single variable
- Can then simply categorize on Diphoton MVA output (5 categories, with cut values optimized against expected limit/significance using MC background, plus additional VBF/VH/ttH tagged categories with loose cut on di-photon MVA)
- Input variables cover kinematics (sans mass), per-event mass resolution and vertex probability, and photon ID

Di-Photon MVA Input Variables

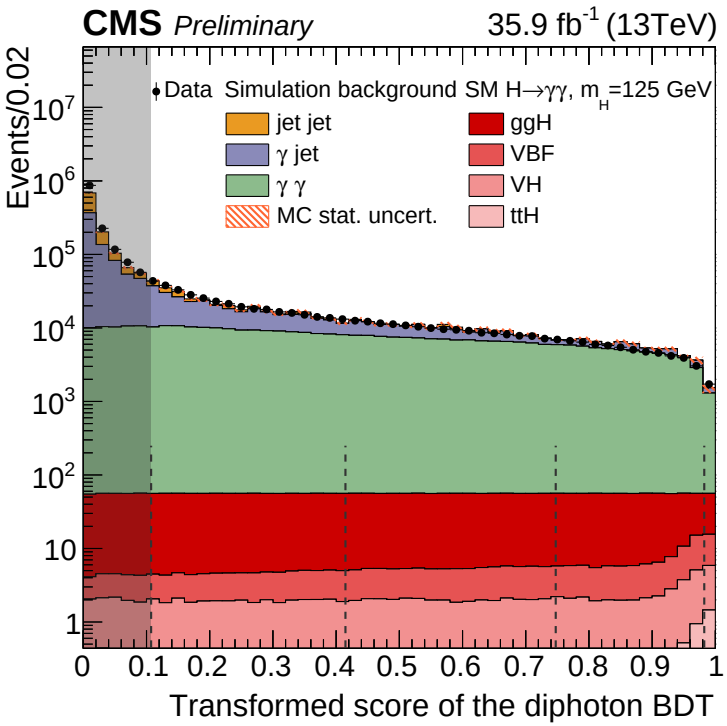
- Input variables cover kinematics (sans mass), per-event resolution and vertex probability, and photon ID
- Input Variables:
  1. $p_T^1/m_{\gamma\gamma}$
  2. $p_T^2/m_{\gamma\gamma}$
  3. η_1
  4. η_2
  5. $\cos\Delta\phi_{\gamma\gamma}$
  6. $\sigma_m/m_{\gamma\gamma}$ (Right Vtx Hypothesis)
  7. $\sigma_m/m_{\gamma\gamma}$ (Wrong Vtx Hypothesis)
  8. p_{vtx}
  9. $IDMVA_1$
  10. $IDMVA_2$
- σ_m constructed from per-photon σ_E estimate from regression, adding also beamspot width contribution for wrong vtx hypothesis
- Per-event primary vertex selection probability p_{vtx} comes from per-event vertex MVA

Di-Photon MVA: Resolution

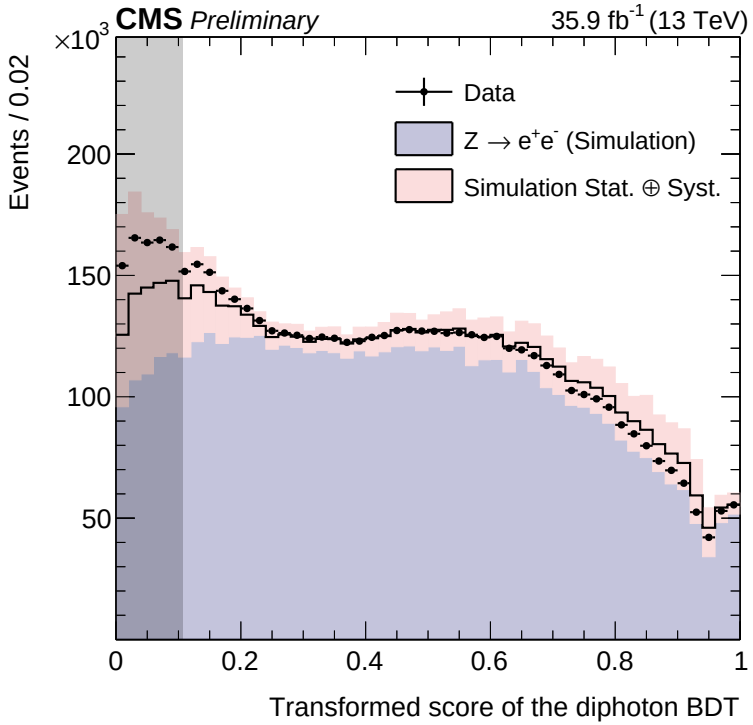
- Since input variables are mass-independent, MVA is not sensitive to mass resolution (since inclusive S/B in full mass range does not change with resolution)
- Correct this by weighting the signal events during training by 1/resolution, taking into account right and wrong primary vertex hypotheses weighted by the per-event probability
- $w_{sig} = \frac{p_{vtx}}{\sigma_m^{right}/m_{\gamma\gamma}} + \frac{1-p_{vtx}}{\sigma_m^{wrong}/m_{\gamma\gamma}}$
- $\frac{\sigma_m^{right}}{m_{\gamma\gamma}} = \frac{1}{2}\sqrt{\frac{\sigma_{E1}^2}{E_1^2} + \frac{\sigma_{E2}^2}{E_2^2}}$
- $\frac{\sigma_m^{wrong}}{m_{\gamma\gamma}} = \sqrt{\left(\frac{\sigma_m^{right}}{m_{\gamma\gamma}}\right)^2 + \left(\frac{\sigma_m^{vtx}}{m_{\gamma\gamma}}\right)^2}$
- With σ_m^{vtx} computed analytically from beamspot width and calorimeter positions of the photons

Di-Photon MVA Output

- Lowest score region not included in the analysis
- Diphoton MVA output for signal-like events can be validated with $z \to ee$ events by inverting electron veto in the pre-selection
- Analysis does not rely on MVA shape of Monte Carlo background

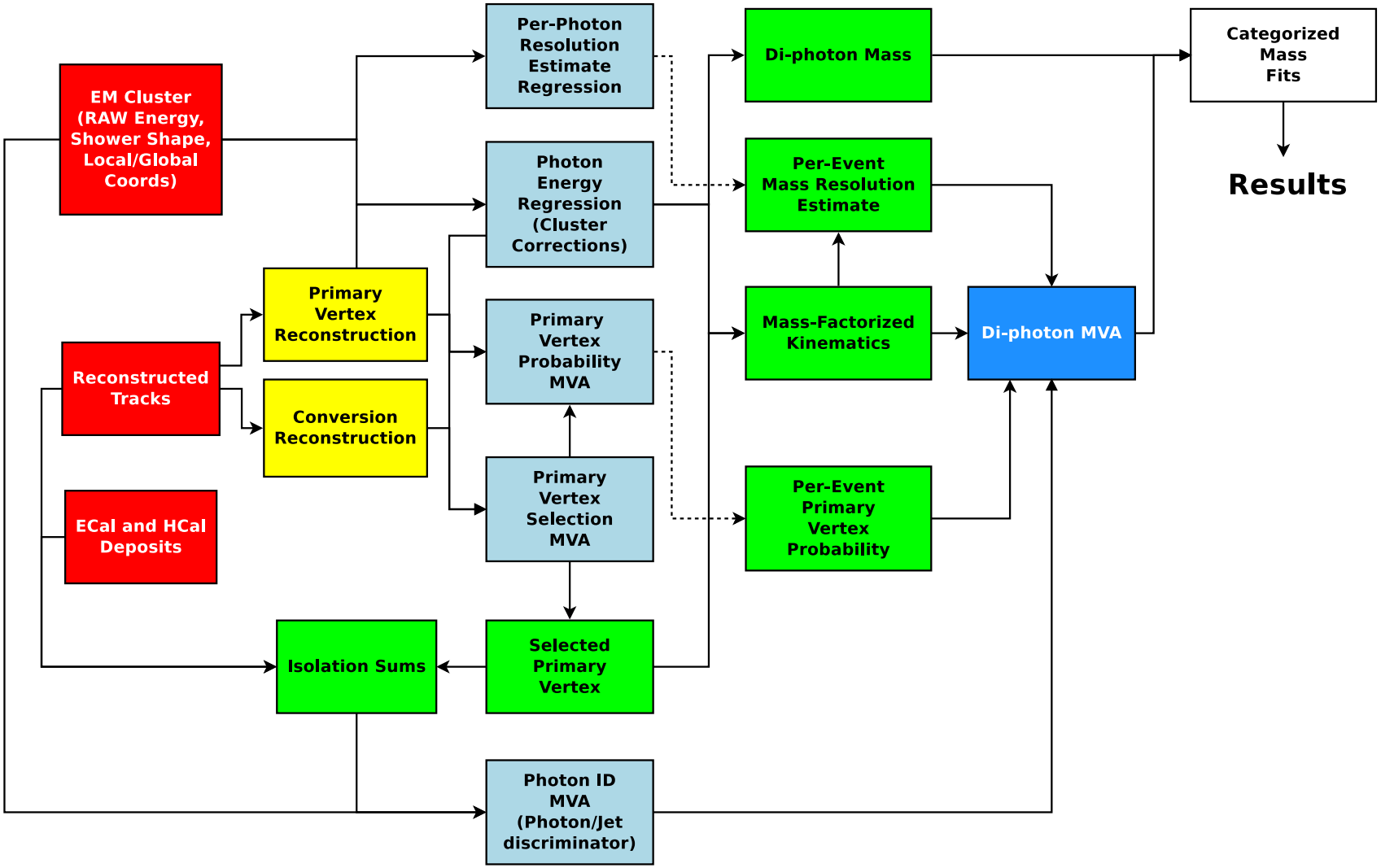


(a) Full Selection



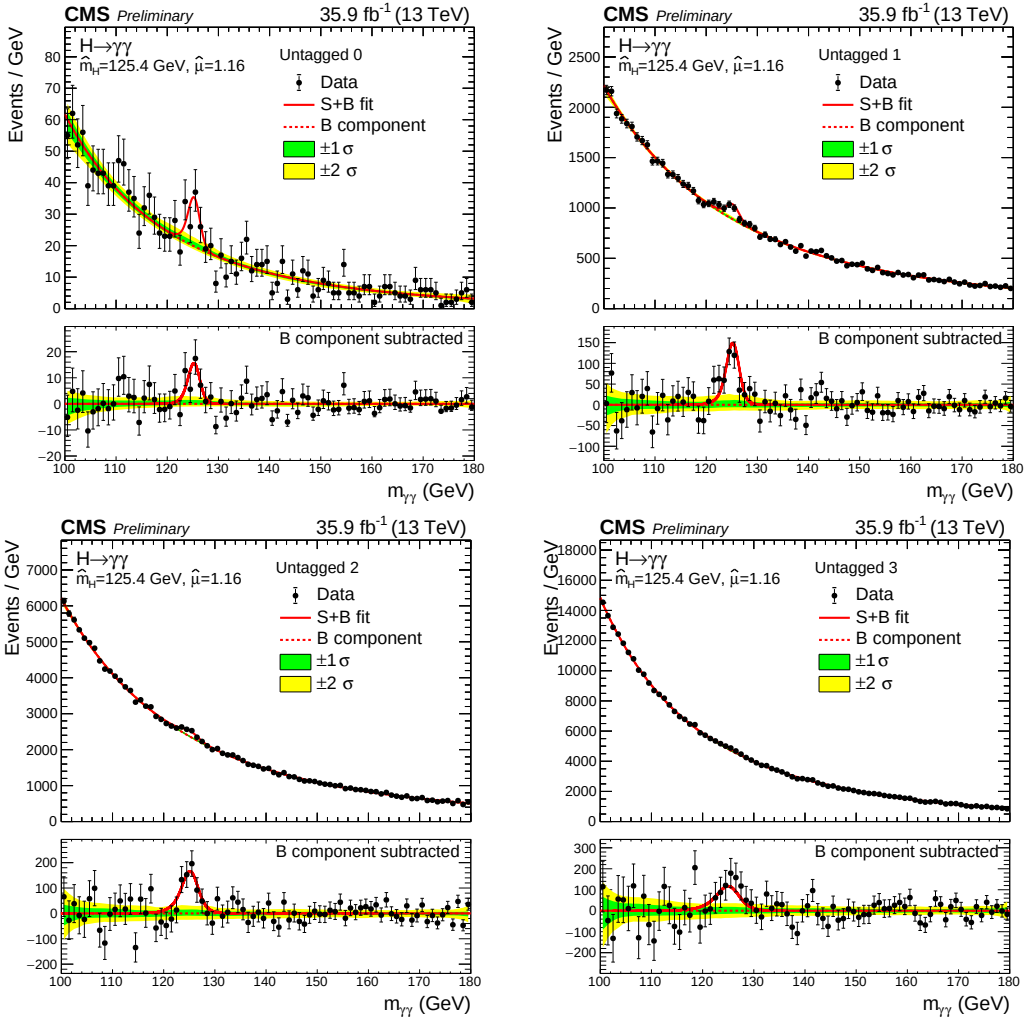
(b) Inverted e-Veto

Higgs$\to \gamma\gamma$: All Together



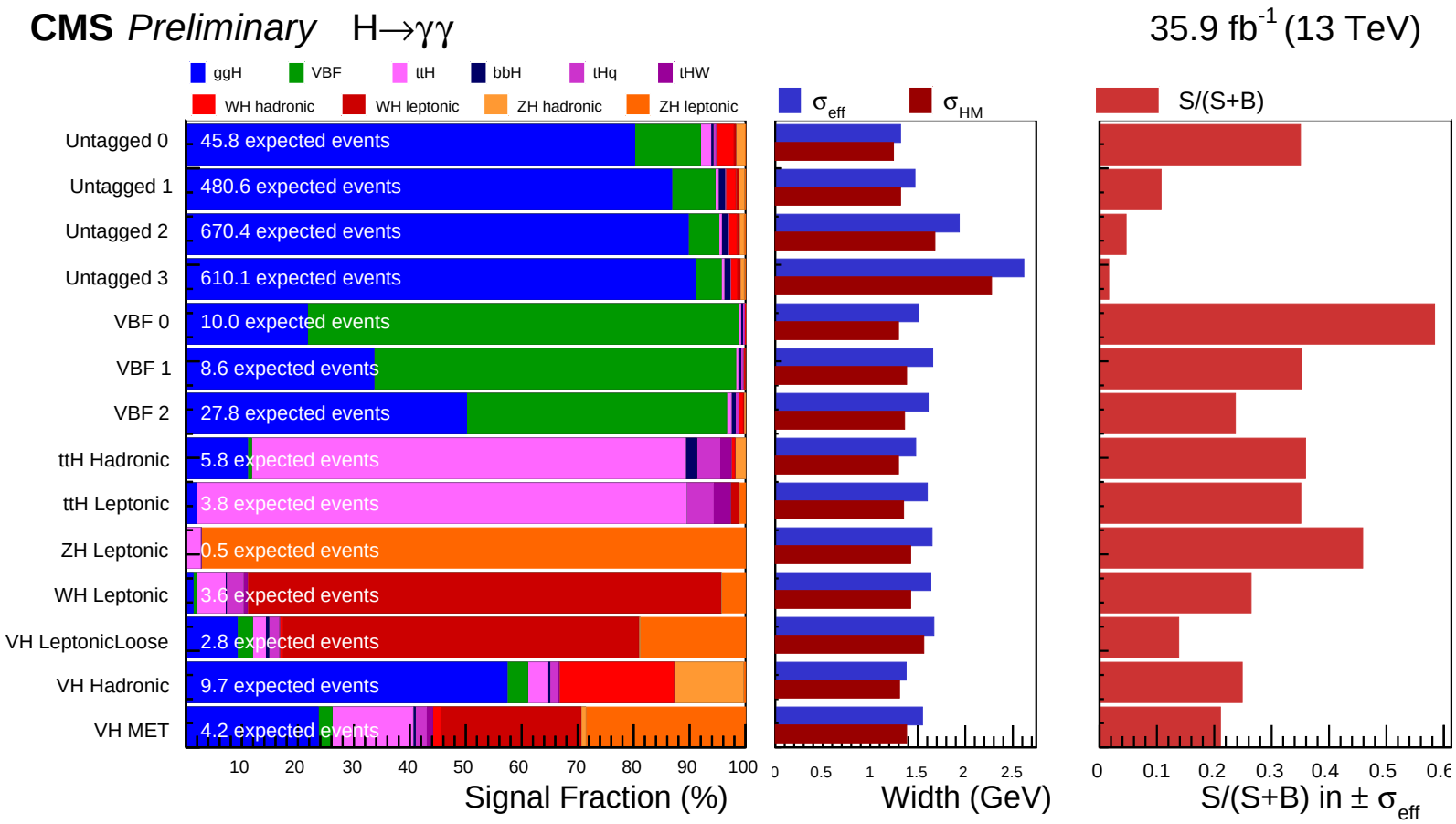
- Strategy: Process available information into quantities with straightforward physical interpretations in order to combine per-event knowledge of expected mass resolution and S/B into a single "Diphoton MVA" variable

S+B Fits - 13 TeV

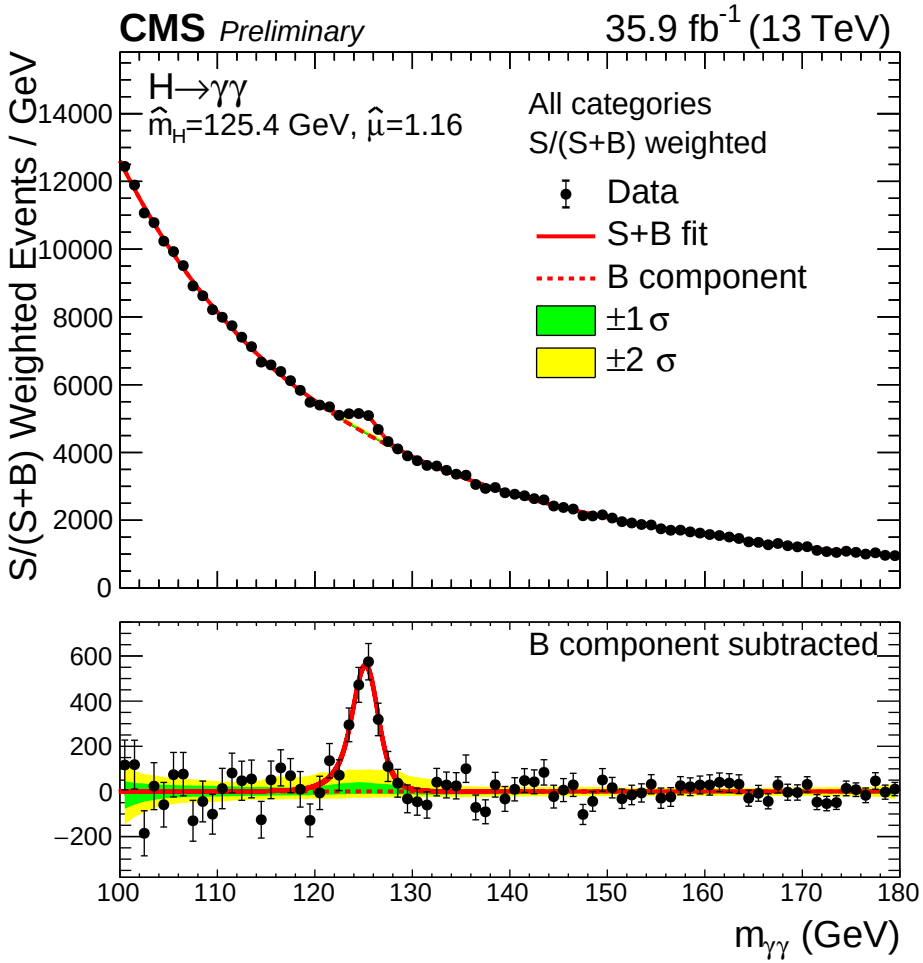


- Plus 10 more distributions for exclusive-tagged modes

Event Categorization



S+B Fit - Weighted Combination

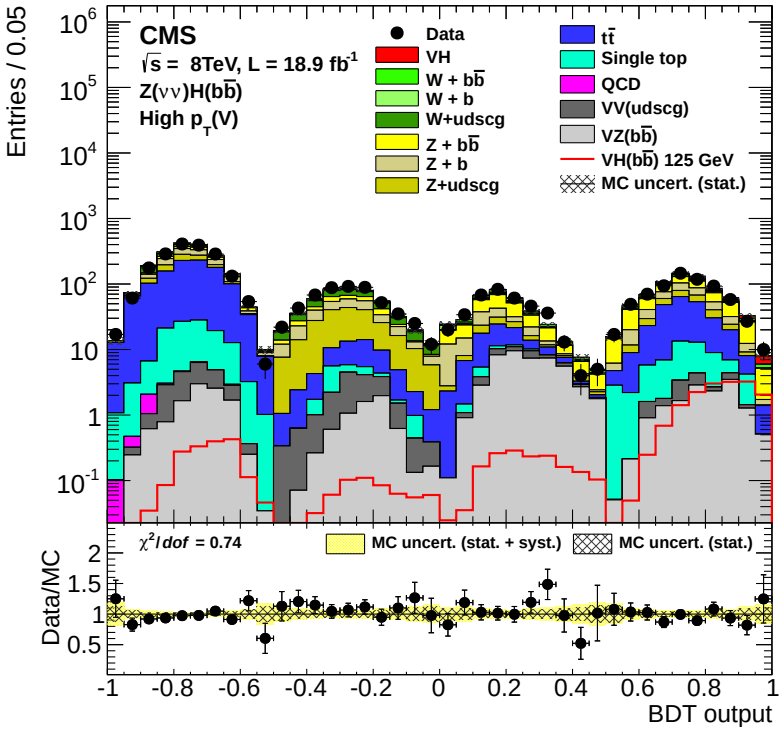


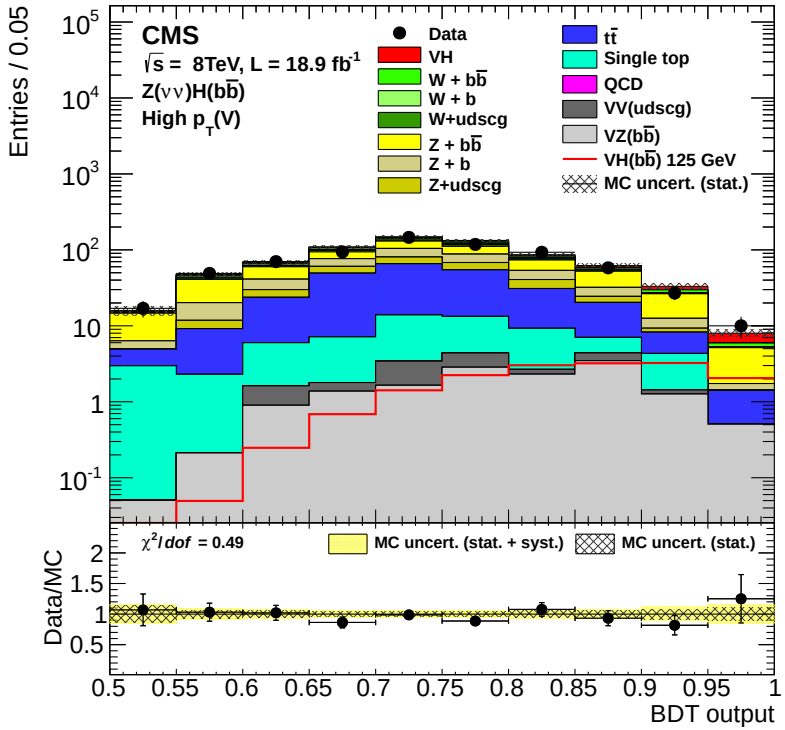
- Results extracted from simultaneous fit to 14 event classes, but combined mass spectrum useful for visualisation
- Combination of all 14 event classes, weighted by $S/(S+B)$ for a $\pm\sigma_{eff}$ window in each event class
- Weights are normalised to preserve the fitted number of signal events

$W/Z + H \to bb$ Signal Extraction

- Even after determination of scale factors from control regions, backgrounds have non-negligible uncertainty
- Final sensitivity benefits from being able to further constrain background normalizations in the final fit
- Procedure:
  1. Train four BDT's for each channel: signal vs $t\bar{t}$, signal vs W/Z+jets, signal vs dibosons, signal vs (all) background
  2. Cuts on background-specific BDT's are used to **partition** final signal vs (all) background distribution into four subsets

$W/Z + H \to b\bar{b}$ Signal Extraction





- Results extracted from fit to final BDT distribution, partitioned using dedicated BDT's into individual background and signal-enriched regions

$W/Z + H \to bb$ Signal Extraction

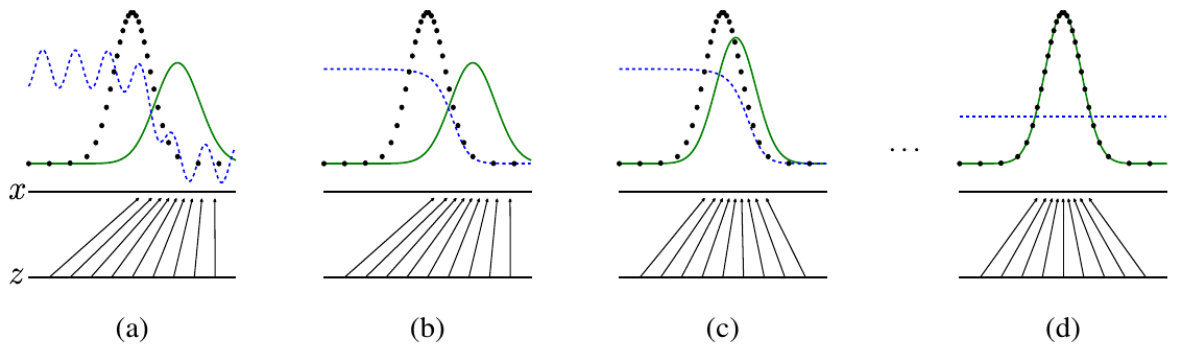
- Input variables for BDT's:
  - Several kinematic variables for selected jets (including dijet mass) and W/Z candidate (lepton, missing transverse momentum kinematics)
  - Number of additional jets
  - b-tag discriminant value for selected and additional jets
- Jet energy scale and b-tag discriminant uncertainties enter as **shape uncertainties** for final BDT distributions

Generative Deep Neural Networks

- Significant recent work on generative deep neural networks in the data science community, with image processing/generation as a common use case
  e.g arXiv:1406.2661
- Typical existing use cases:
  - Have a fixed set of data, or a black box generator
  - Train a generative model to produce samples following the distribution of the training data (or in high dimensional cases such as images, to produce "similar" images to those in the training set)
- Various architectures and training procedures: Variational auto-encoders, auto-regressive models, generative adversarial networks

Generative Adversarial Networks

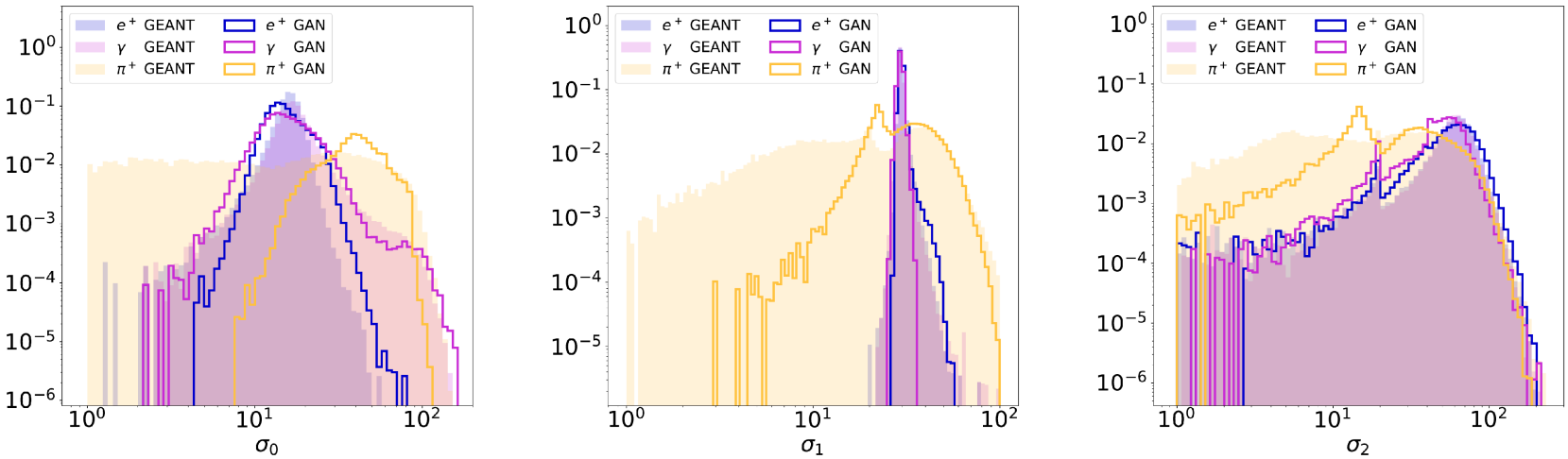
- Generative adversarial networks train a deep neural network to generate samples starting from a known prior distribution $p(\bar{z})$ which is easy to sample from (e.g. an N-dimensional normal distribution)
- The generative network $\bar{G}$ transforms the input samples to the output space $\bar{x}$, ie $G(\bar{z}) = \bar{x}$
- A discriminator network D (e.g. a standard DNN classifier) is trained to distinguish the generated samples from the training samples
- Training proceeds iteratively such that the D is trained to maximally discriminate and G is trained to minimize the discrimination power of D until the generated samples follow the $\sim$ same distribution as the training set (MINIMAX problem/saddle point, difficult to train)



arXiv:1406.2661

Simulating Calorimeter Showers

- Accurate, physics-based simulation of calorimeter showers is available with GEANT, but computationally intensive
- Generative Deep Neural networks could be used to simulate showers
- **Goal is not a better simulation, but a computationally faster one**

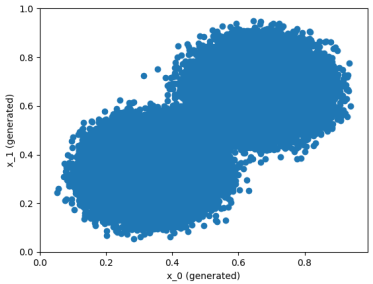


- CaloGAN work in arXiv:1705.02355 achieves approximate modelling of shower profiles, but 100,000x speedup for DNN on GPU vs Geant on CPU

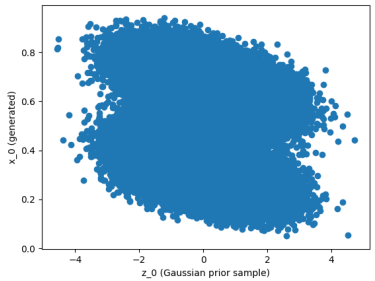
Machine Learning Monte Carlo Integration

- Use machine learning to improve Monte Carlo integration efficiency in generators beyond what is achievable with VEGAS
- J. Bendavid, "Efficient Monte Carlo Integration Using Boosted Decision Trees and Generative Deep Neural Networks" `https://arxiv.org/abs/1707.00028`

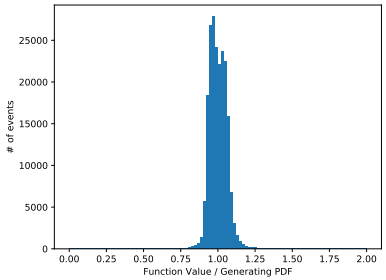
DNN 4D Camel Function Example



(a) Generated (2D Slice)



(b) Generated vs Prior (1D pair)



(c) Integration Weight

- 3x smaller weight variance to foam with 10x less function evaluations

Algorithm	# of Func. Evals	$\sigma_w / < w >$	σ_I / I (2e6 add. evts)
VEGAS	300,000	2.820	$\pm 2.0 \times 10^{-3}$
Foam	3,855,289	0.319	$\pm 2.3 \times 10^{-4}$
Generative DNN	300,000	0.082	$\pm 5.8 \times 10^{-5}$
Generative DNN	294,912	0.083	$\pm 5.9 \times 10^{-5}$
Generative DNN (staged)	294,912	0.030	$\pm 2.1 \times 10^{-5}$

Some results - 9D Camel Function Integration

- Comparing Vegas, GBRIntegrator, Generative DNN for 9-dimensional camel function

Algorithm	# of Func. Evals	$\sigma_w / < w >$	σ_I / I (2e6 add. evts)
VEGAS	1,500,000	19	$\pm 1.3 \times 10^{-2}$
GBRIntegrator	3,200,000	0.63	$\pm 4.5 \times 10^{-4}$
GBRIntegrator (staged)	3,200,000	0.31	$\pm 2.2 \times 10^{-4}$
Generative DNN	294,912	0.15	$\pm 1.1 \times 10^{-4}$
Generative DNN (staged)	294,912	0.081	$\pm 5.7 \times 10^{-5}$

- 50x smaller weight variance to Vegas with 2x function evaluations (BDT)
- DNN approach scales much better with dimensionality ($>$ 100x smaller weight variance than Vegas with 5x **fewer** function evaluations

Outlook: Machine Learning Monte Carlo Integration

- Large improvements with novel algorithms already demonstrated on test cases
- Exploring alternative DNN architectures including auto-regressive models and convolutional elements
- Integration into Madgraph_aMC@NLO and tests with QCD matrix elements in progress

Conclusions and Outlook

- Machine learning used extensively already in LHC Run 1, typically BDT's and simple ANN's taking **high level features as input**, for classification and also regression in some cases
- Important to have regression/classification **accuracy estimates/uncertainties** to make optimal use of events (weighting or classification)
- Integration into analysis and extraction of results as important as underlying machine learning techniques
- Underyling machine learning techniques transitioning to **Deep Neural Networks** with a range of architectures, benefiting from active research and technical implementations from broader data science community and industry
- For HEP: Enables much-higher dimensional problems: Use of **lower level or even detector-level inputs**. Lots of work already underway
- Personal interests: Likelihood-based/in-situ uncertainty estimates for DNN's, calibration of simulations, interplay of ML with unfolding and parameter extraction at high level of the analysis

Backup

Conclusions (Energy Regression)

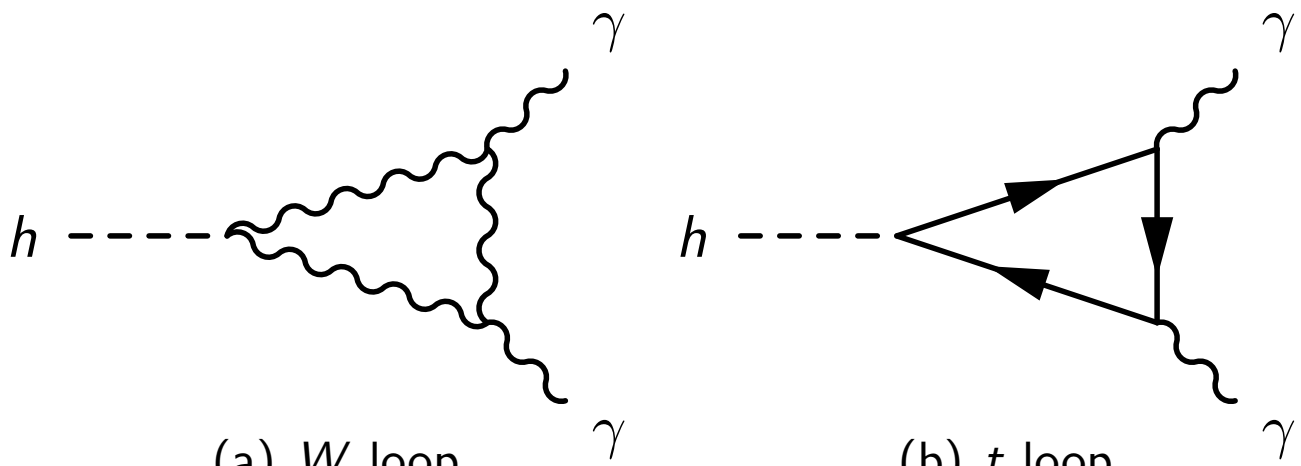
- Many effects (local containment, material interactions, pileup) lead to both shifts in energy scale and additional event-to-event fluctuations, with non-trivial correlations between them
- Transversely segmented calorimeters provide potentially large amount of information about the shower properties which can be used to construct in-situ corrections
- Multivariate/machine learning techniques very effective for high dimensionality problems with no (fully) parametric model but large training datasets available
- For optimal use of data, important to have fine-grained resolution estimate along with optimized energy corrections
- Has achieved significant improvements in energy resolution for electrons/photons in CMS electromagnetic calorimeter with large payoff for the Higgs discovery and properties measurements
- Semi-parametric extension of existing algorithms developed as part of this effort

Further Considerations (Energy Regression)

- In CMS we have so far trained multivariate corrections in Monte Carlo simulation
- Better simulation for training $\rightarrow$ better performance of corrections on data
- Test-beam data with known beam energy could also be used for training in principle (and may be very useful as part of detector development and characterization)
- Ongoing work to use $Z \rightarrow ee$ collision data to train corrections in-situ (but more complicated due to correlation between two electrons in each event as well as natural Z width)
- Corrections can always be staged, e.g., simulation-trained correction + data trained "residual" corrections
- Effective use cases for prediction of full lineshape in addition to Gaussian resolution?
- So far pulse reconstruction/out-of-time pileup has been treated as a separate factorized piece. Algorithmic solutions are effective albeit cpu-expensive, machine learning opportunities here?
- Opportunities for further improvements with more modern machine learning techniques like deep neural networks?

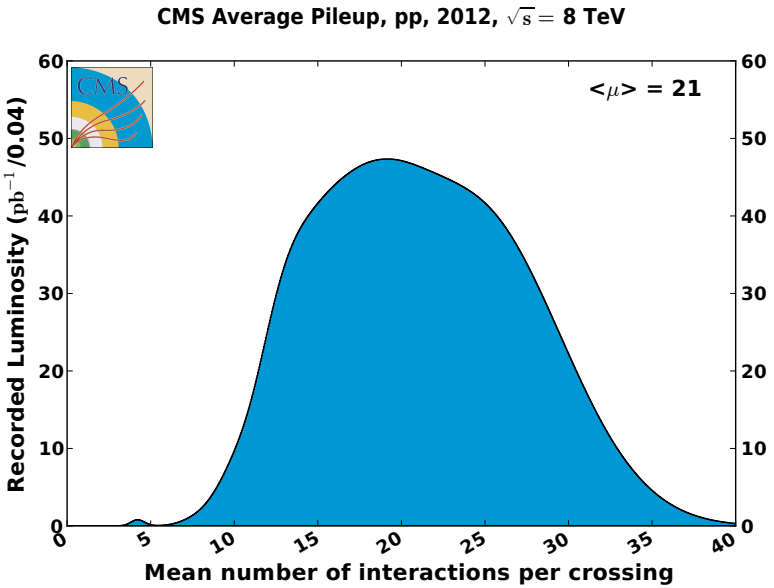
Higgs$\to \gamma\gamma$ Decay

- No tree-level $h\gamma\gamma$ vertex, decay proceeds through W and fermion (top) loops which interfere destructively
- Branching ratio to two photons very sensitive to fermion vs boson couplings and possible new particles in the loop

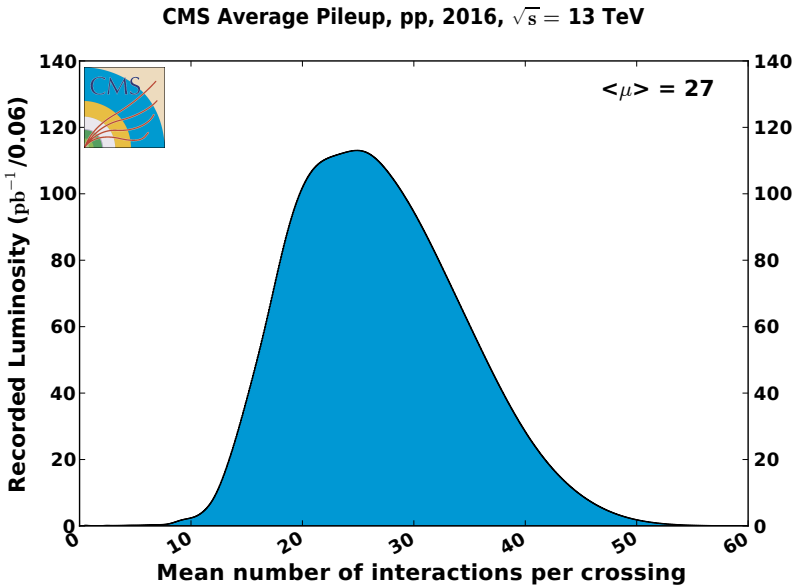


(a) W loop (b) t loop

Pileup Conditions



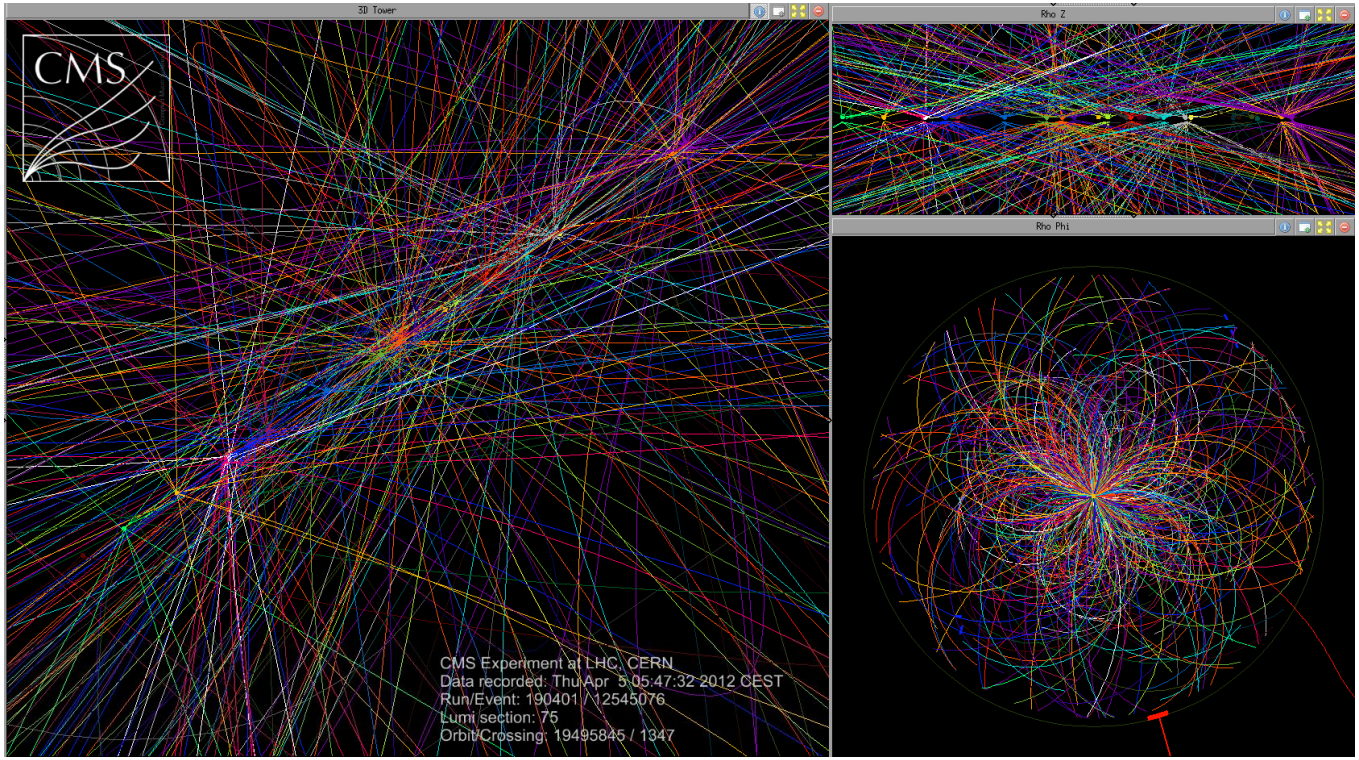
(a) 2012 8 TeV,< NPU >= 21



(b) 2016 13 TeV,< NPU >= 27

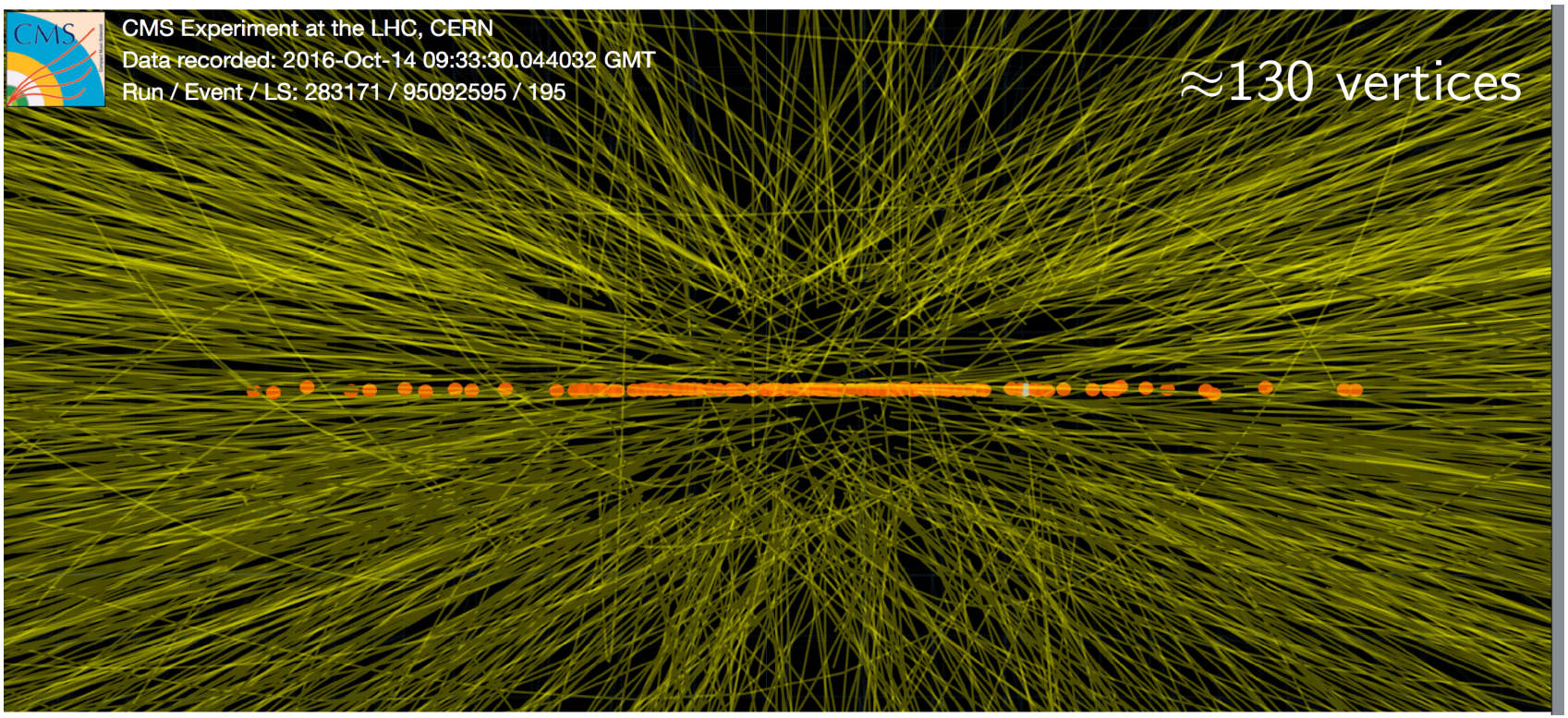
- Large number of pileup interactions, interaction region extended in z direction with σ = 5-6 cm

Pileup in CMS



- An event with 29 reconstructed primary vertices

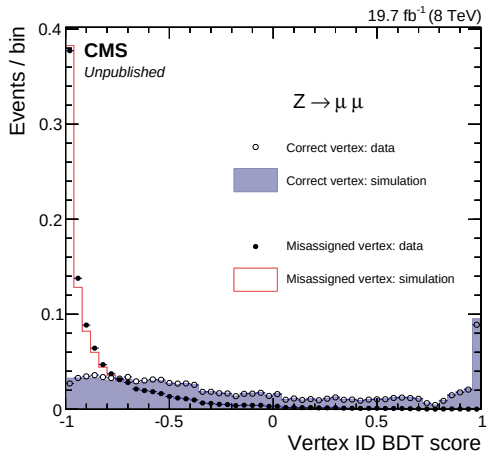
Pileup in CMS

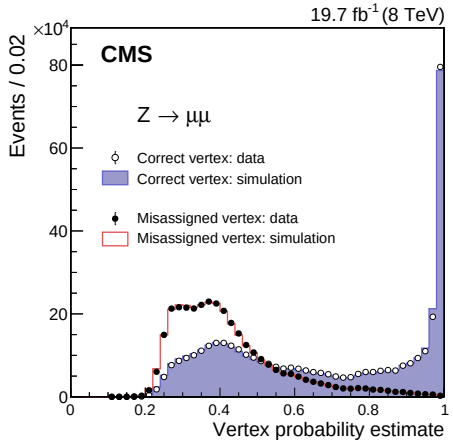


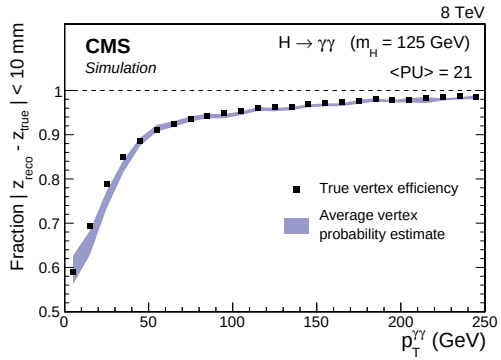
- Real-life event with HL-LHC-like pileup from special run in 2016 with individual high intensity bunches

Primary Vertex Selection

- Opening angle needed to calculate diphoton mass: need to know production vertex location
- No charged particles in general, primary vertex selection ambiguous with large pileup
- Per-vertex MVA to select hard interaction from pileup vertices, using hadronic recoil balancing with diphoton system, and tracks from converted photons
- A second MVA is trained to estimate for each event the probability that the vertex choice is correct







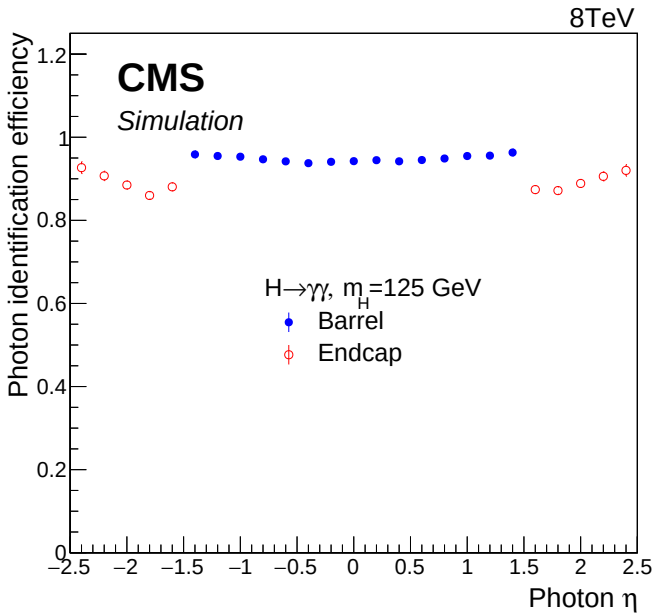
- Inclusive vertex selection efficiency ∼80 %, but strong dependence on Higgs p_T

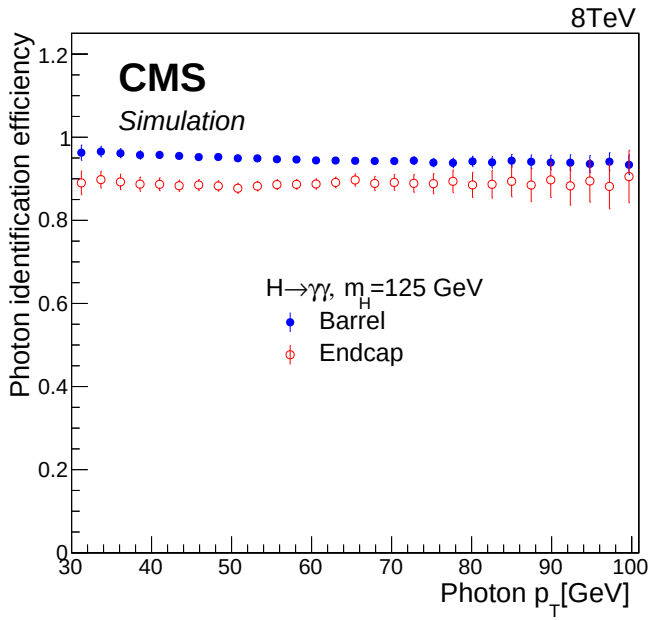
Photon Selection

- Geometric and (scaled) transverse momentum pre-selection cuts driven by detector acceptance and trigger requirements
- Veto electrons
- Need to discriminate between prompt isolated photons, and fakes from jets (mainly collimated $\pi^0/\eta^0 \to \gamma\gamma$ decays)
- Two handles:
  - Shower Shape: Two photons from π^0/η^0 produce a wider EM cluster on average.
  - Isolation: Select against additional particles produced in the jet alongside the leading π^0/η (some complications from pileup)

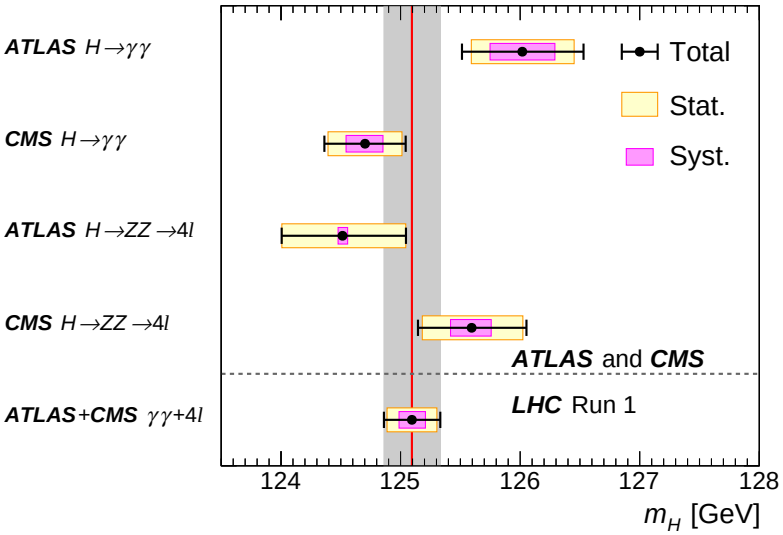
Photon Identification: MVA

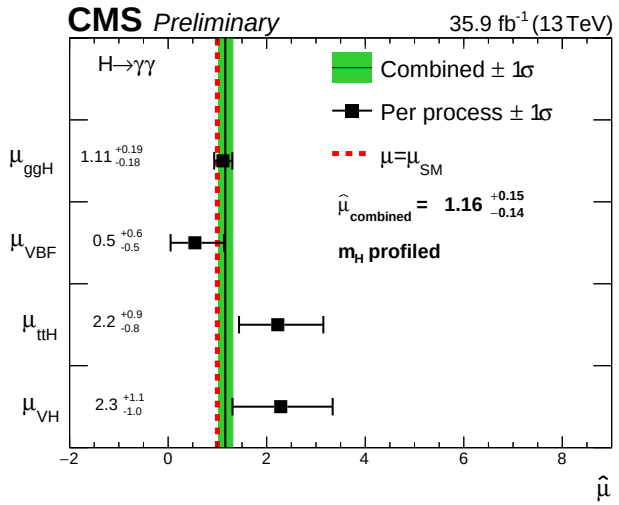
- Photon identification intended to be uncorrelated with photon kinematics (p_T and rapidity), in order to avoid shaping the mass distribution and allow kinematics to be optimally exploited by event level BDT
- Signal training sample reweighted in two-dimensions (p_T,η) to match background training sample at preselection level
- Results not perfect (some residual η dependence in endcaps), but sufficient (may investigate uboost/flatness boosting/multivariate decorrelation or similar techniques in the future)





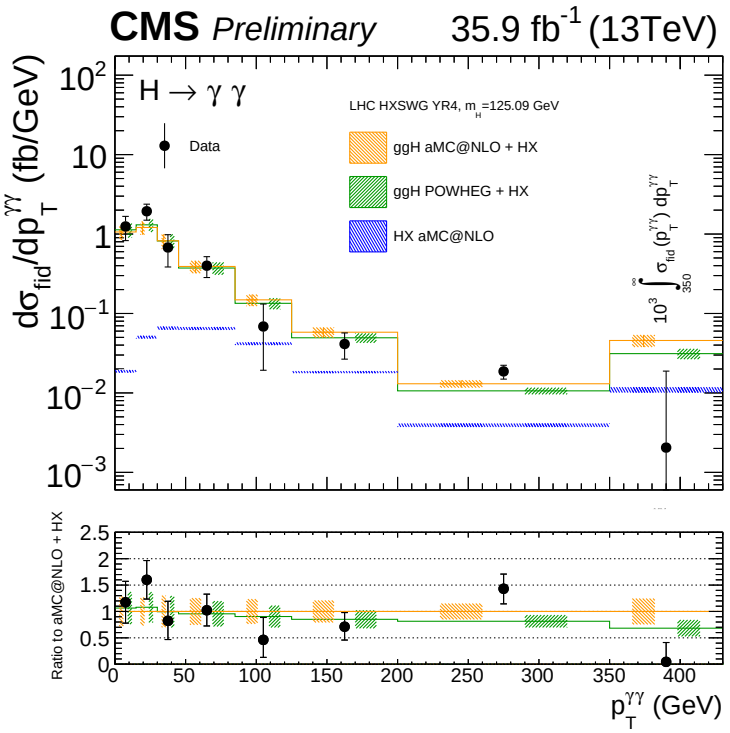
Higgs→ $\gamma\gamma$ Results

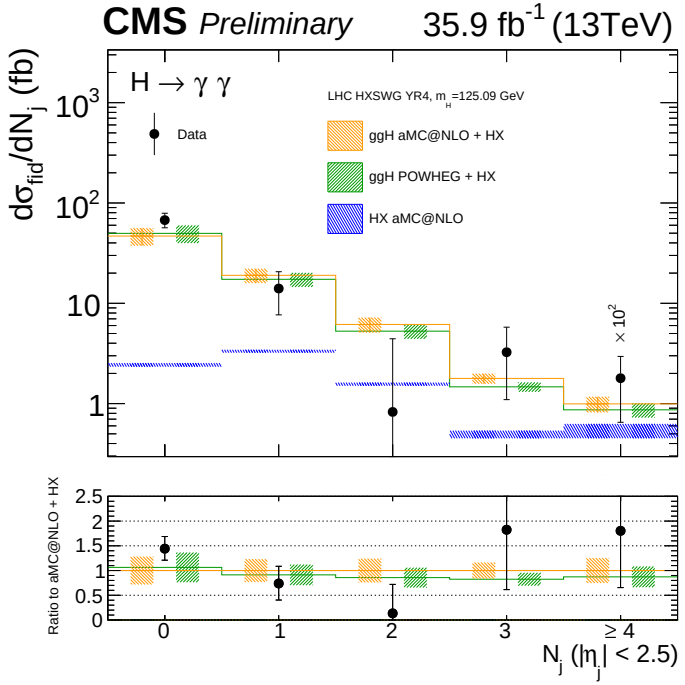




- (Run 1) Mass Measurement:
  $m_H = 124.70 \pm 0.31(\text{stat.}) \pm 0.15(\text{syst.})$ GeV
- Overall $\sigma/\sigma_{SM} = 1.16^{+0.11}_{-0.10}(\text{stat.})^{+0.09}_{-0.08}(\text{syst.})^{+0.06}_{-0.05}(\text{th.})$

Higgs$\to \gamma\gamma$ Fiducial/Differential Cross Sections





- No event-level kinematics used for categorization to reduce model dependence, categerization based on σ_m/m

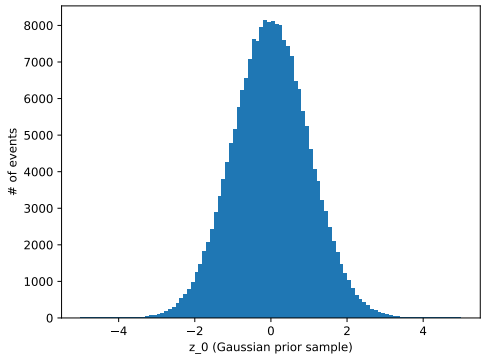
Monte Carlo Integration and Generation

- **Monte Carlo integration:** Given an arbitrary/black box multidimensional function $f(\bar{x})$, find the integral $\int f(\bar{x})d\bar{x}$
- **Monte Carlo generation:** Given an arbitrary/black box multidimensional function $f(\bar{x})$, generate an unweighted set of vectors $\bar{x}$ with a probability density $p(\bar{x}) = f(\bar{x})/\int f(\bar{x})d\bar{x}$
- Typical HEP use case: Given a numerical implementation for a matrix element fully differential in incoming/outgoing four-vectors, compute the total cross section (integral), and generate a set of unweighted events

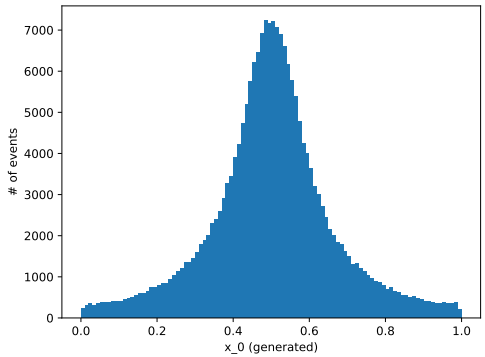
Monte Carlo Integration

- Canonical approach: **Importance Sampling:** Construct an easily sampled from approximation to the target function
  - VEGAS: Product of adaptively-binned 1D histograms
  - FOAM: Sampling from a (single) binary decision tree $\rightarrow$ phase space divided into hyper-rectangles with optimized boundaries
  - BDT: Sampling from an additive series of decision trees $\rightarrow$ Gradient boosting used to improve performance wrt FOAM just like in classification and regression problems
  - DNN: Sampling from a generative deep-neural network

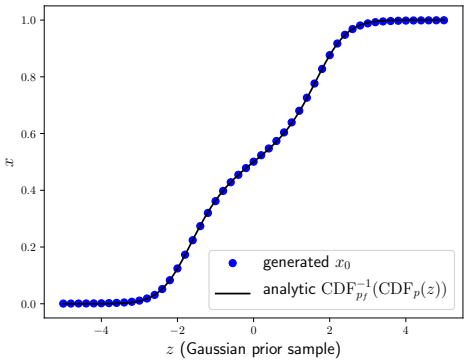
1D DNN Example with Analytic Solution



(a) Prior



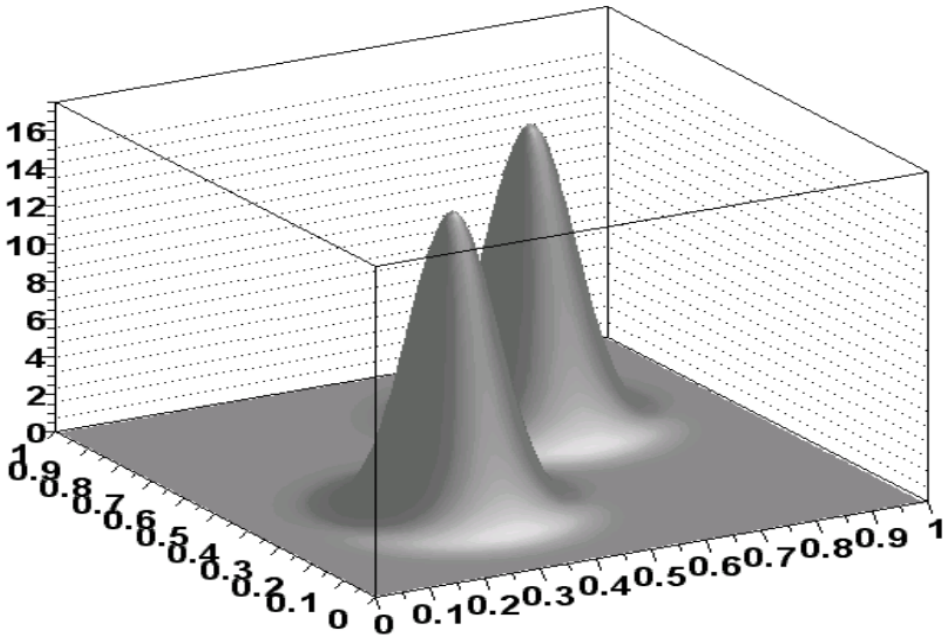
(b) Generated



(c) Generated vs Prior

- In 1D the generative network is essential just learning the inverse CDF of the target distribution (numerically)
- Technically the function is $x = \mathrm{CDF}_{p_f}^{-1}(\mathrm{CDF}_p(z))$
- For Cauchy distribution in this example, this can be computed analytically and compared to the trained DNN result

Monte Carlo Integration and Generation: Example Function



S. Jadach, physics/0203033

- This is the "camel" function from the original VEGAS paper, which can be generalized to N dimensions
- Factorized approach will not work well
- Significant low-density regions which cannot be easily excluded a-priori

Some results - 4D Camel Function Integration

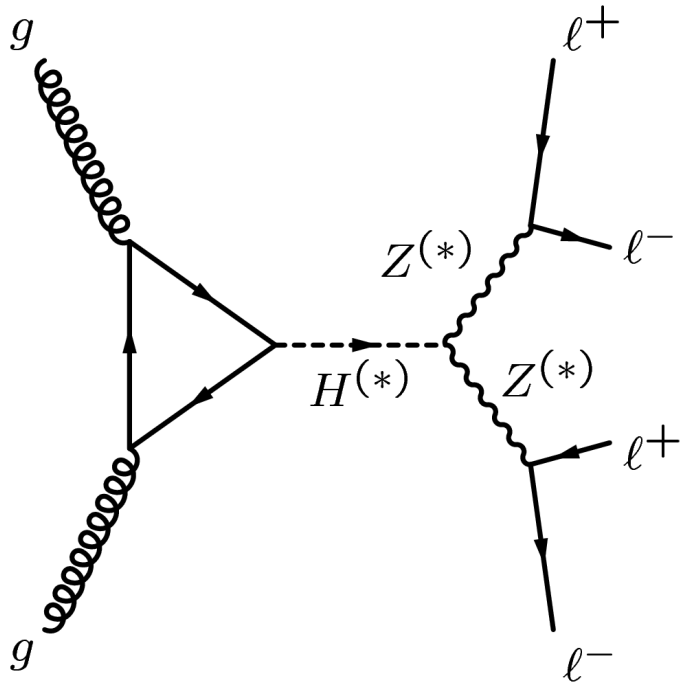
- Comparing Vegas,Foam, GBRIntegrator, Generative DNN for 4-dimensional camel function (since this appears in both VEGAS and Foam papers).
- Given relative weight variance $\sigma_w / < w >$ after training/grid building, relative uncertainty on integral evaluated with N additional events is $\sigma_I / I = \frac{1}{\sqrt{N}} \sigma_w / < w >$

Algorithm	# of Func. Evals	$\sigma_w / < w >$	σ_I / I (2e6 add. evts)
VEGAS	300,000	2.820	$\pm 2.0 \times 10^{-3}$
Foam	3,855,289	0.319	$\pm 2.3 \times 10^{-4}$
Generative BDT	300,000	0.082	$\pm 5.8 \times 10^{-5}$
Generative BDT (staged)	300,000	0.077	$\pm 5.4 \times 10^{-5}$
Generative DNN	294,912	0.083	$\pm 5.9 \times 10^{-5}$
Generative DNN (staged)	294,912	0.030	$\pm 2.1 \times 10^{-5}$

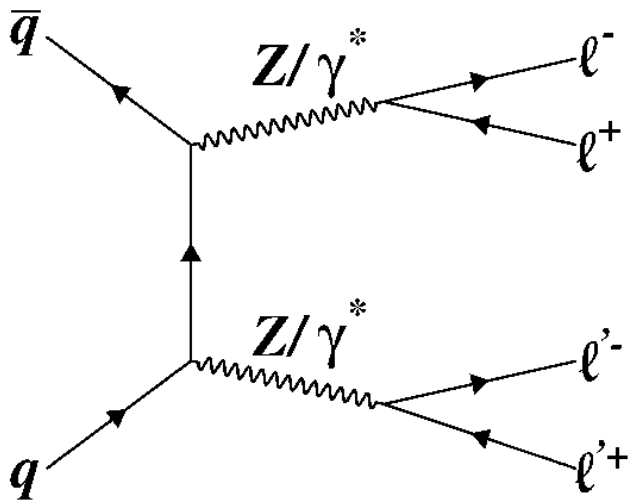
- 3x smaller weight variance to foam with 10x less function evaluations
- For this particular function VEGAS performance saturates at relatively poor weight variance

$H \to ZZ \to 4\ell$

- "Golden channel" - Narrow mass peak on small background
- Irreducible $ZZ \to 4\ell$ continuum background small and well understood



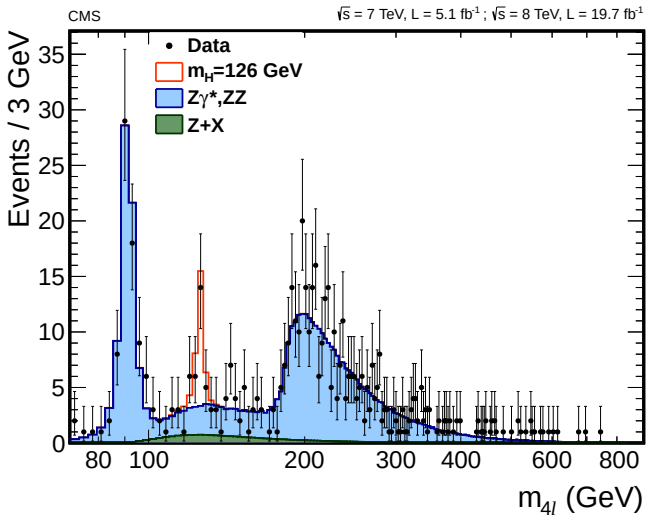
(a) Main signal

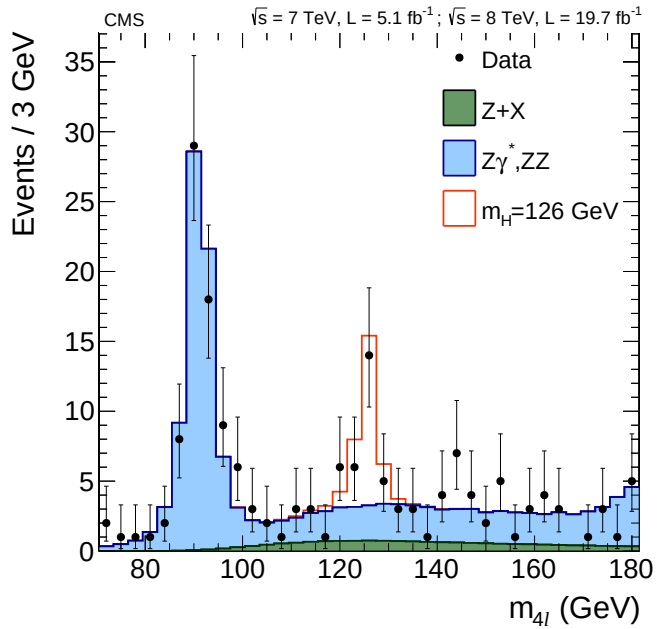


(b) Main background

$H \to ZZ \to 4\ell$

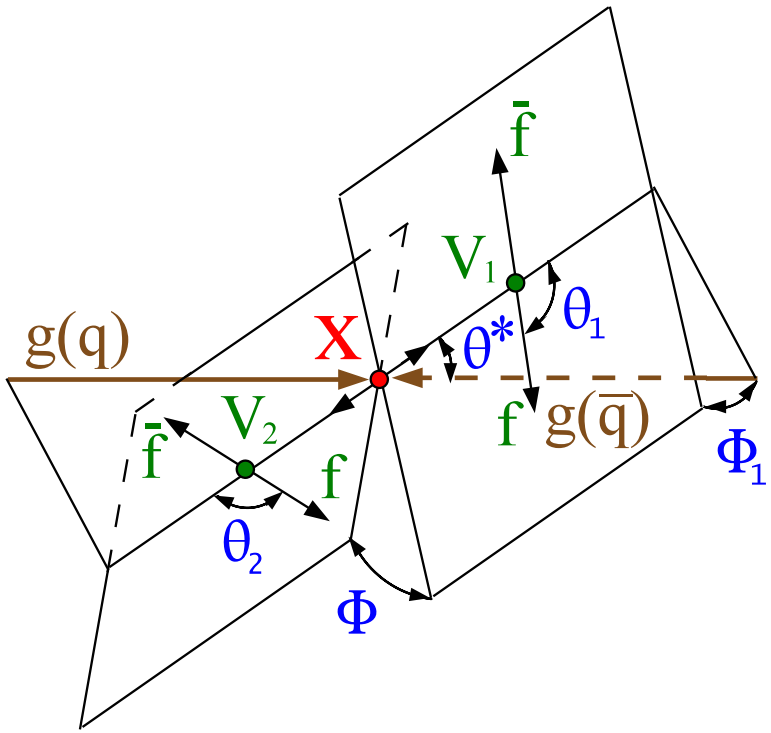
- Select 4 leptons of appropriate charge and flavour combinations (+FSR recovery) with $40 < m_{Z1} < 120$ GeV, $12 < m_{Z2} < 120$ GeV
- Electron acceptance: $|\eta| < 2.5$, $p_T > 7$ GeV, Muon acceptance: $|\eta| < 2.4$, $p_T > 5$ GeV
- Irreducible $ZZ \to 4\ell$ continuum background estimated from MC
- Reducible $Z + b\bar{b}$ and $t\bar{t}$ backgrounds estimated from Z + same-sign dilepton/Z + loose dilepton samples, with fake rates from Z + loose ℓ sample





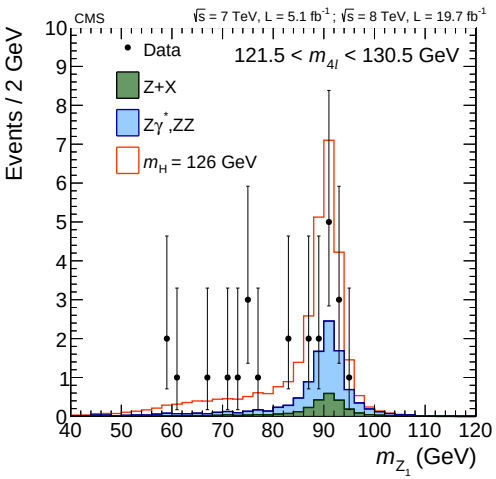
$H \to ZZ \to 4\ell$: Beyond the mass distribution

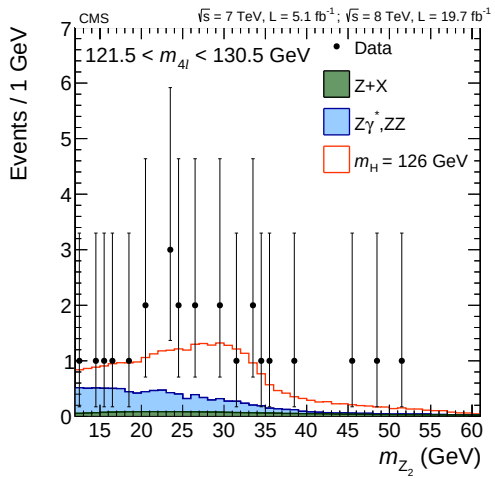
- Higgs is a scalar $\to$ decay angles θ_1,θ_2,Φ, and lepton pair masses m_{Z1},m_{Z2} provide additional discrimination against continuum background

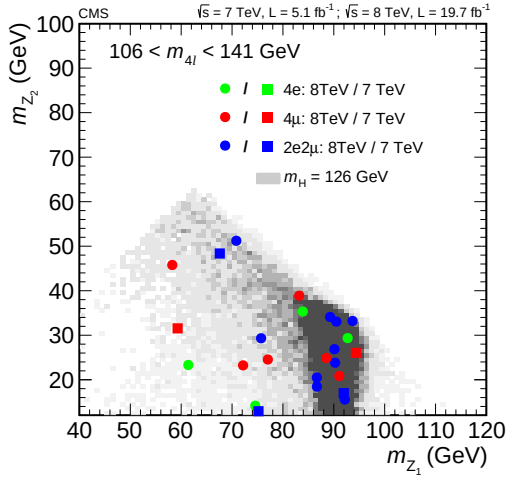


$H \to ZZ \to 4\ell$: Beyond the mass distribution

- Higgs is a scalar $\to$ decay angles θ_1,θ_2,Φ, and lepton pair masses m_{Z1},m_{Z2} provide additional discrimination against continuum background







Matrix Element Likelihood Techniques

- Common problem in machine learning: build a classifier to distinguish signal from background given labeled training samples with features $\bar{x}$
- If probability densities for signal and background $p_{sig}(\bar{x})$, $p_{bkg}(\bar{x})$ are known a priori, then no machine learning is needed, can construct an optimal classifier for hypothesis testing as eg $\frac{p_{sig}(\bar{x})}{p_{sig}(\bar{x})+p_{bkg}(\bar{x})}$
- In high energy physics, often the probability density is known **at the level of the theoretical calculation** and in terms of all initial/final state kinematics
- Can be used directly in cases where final state is fully reconstructed (eg. no neutrinos), detector resolution effects can be neglected, and all/dominant fraction of background is theoretically well-known
- Otherwise painful analytic/numerical integration is needed to convert the matrix element into a pdf relevant for detector-level quantities $\rightarrow$ use Monte Carlo simulation + machine learning as an alternative

$H \to ZZ \to 4\ell$: Matrix Element Likelihood Discriminator

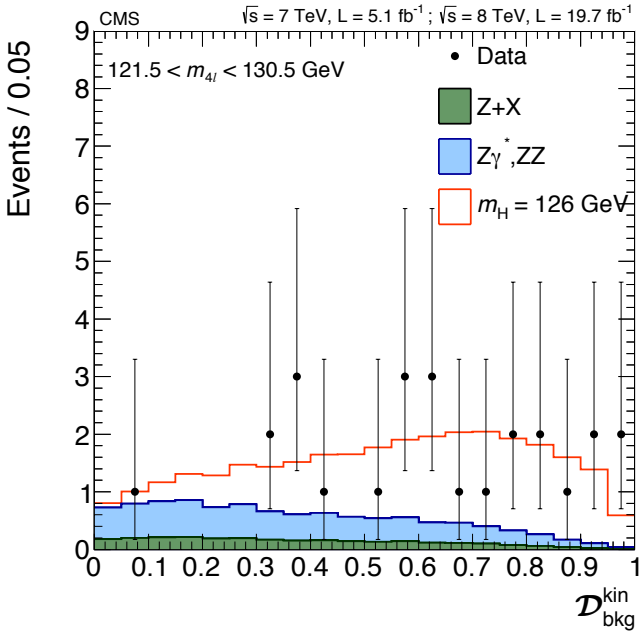
- For $H \to ZZ \to 4\ell$, final state is fully reconstructed, and charged leptons have excellent momentum resolution in CMS ($\mathcal{O}(\%)$)
- Matrix element likelihood discriminator constructed directly from dilepton pair masses, plus decay angles as:

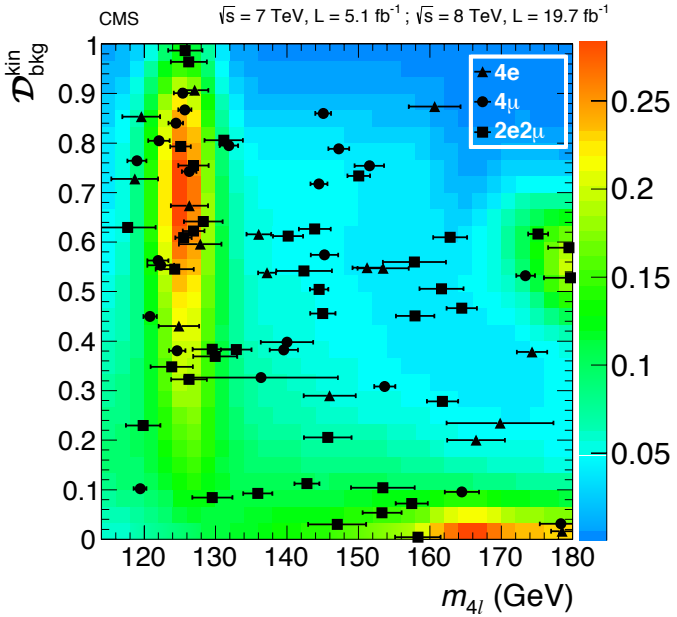
$$D = \frac{p_{sig}(m_{Z1}, m_{Z2}, \theta_1, \theta_2, \Phi | m_{4\ell})}{p_{sig}(m_{Z1}, m_{Z2}, \theta_1, \theta_2, \Phi | m_{4\ell}) + p_{bkg}(m_{Z1}, m_{Z2}, \theta_1, \theta_2, \Phi | m_{4\ell})}$$

- Properly normalized conditional probability densities ensure that D does not bias the four-lepton mass $m_{4\ell}$

$H \to ZZ \to 4\ell$: Matrix Element Likelihood Discriminator

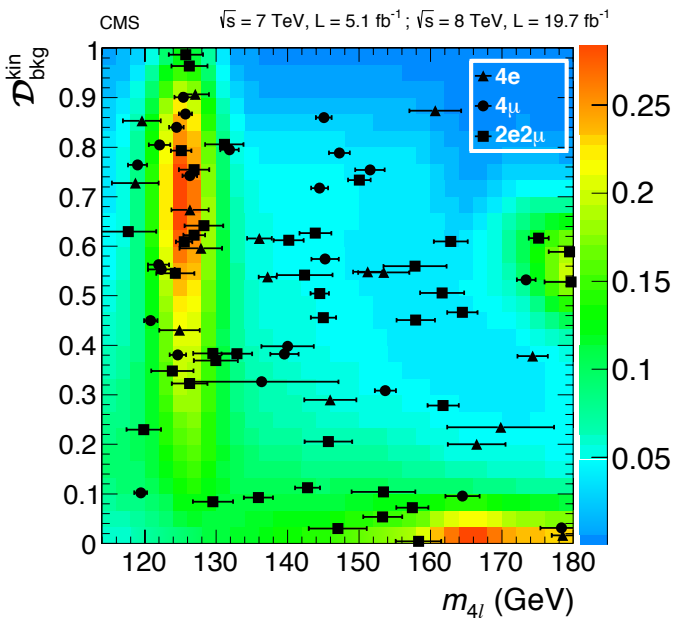
- Signal strength results extracted from 3d unbinned maximum likelihood fit to $m_{4\ell}$ distribution with matrix element likelihood discriminant and $p_T^{4\ell}$

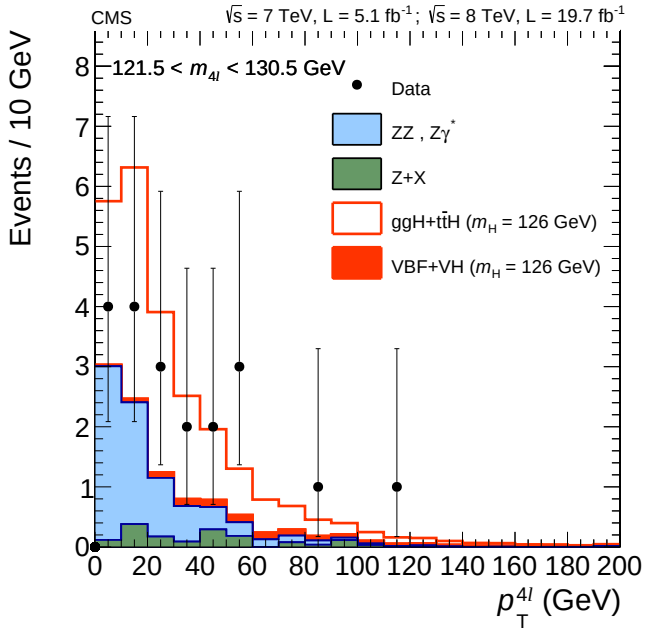




$H \to ZZ \to 4\ell$ Results

- Signal strength results extracted from 3d unbinned maximum likelihood fit to $m_{4\ell}$ distribution with matrix element likelihood discriminant and $p_T^{4\ell}$





- Multidimensional fit more sensitive than $m_{4\ell}$ alone
- $\sigma/\sigma_{SM} = 0.93^{+0.26}_{-0.23}(\text{stat.})^{+0.13}_{-0.09}(\text{syst.})$, 6.8σ observed significance (6.7σ expected)
- ML techniques also used for electron energy reconstruction/per event mass resolution estimate

$W/Z + H \to bb$

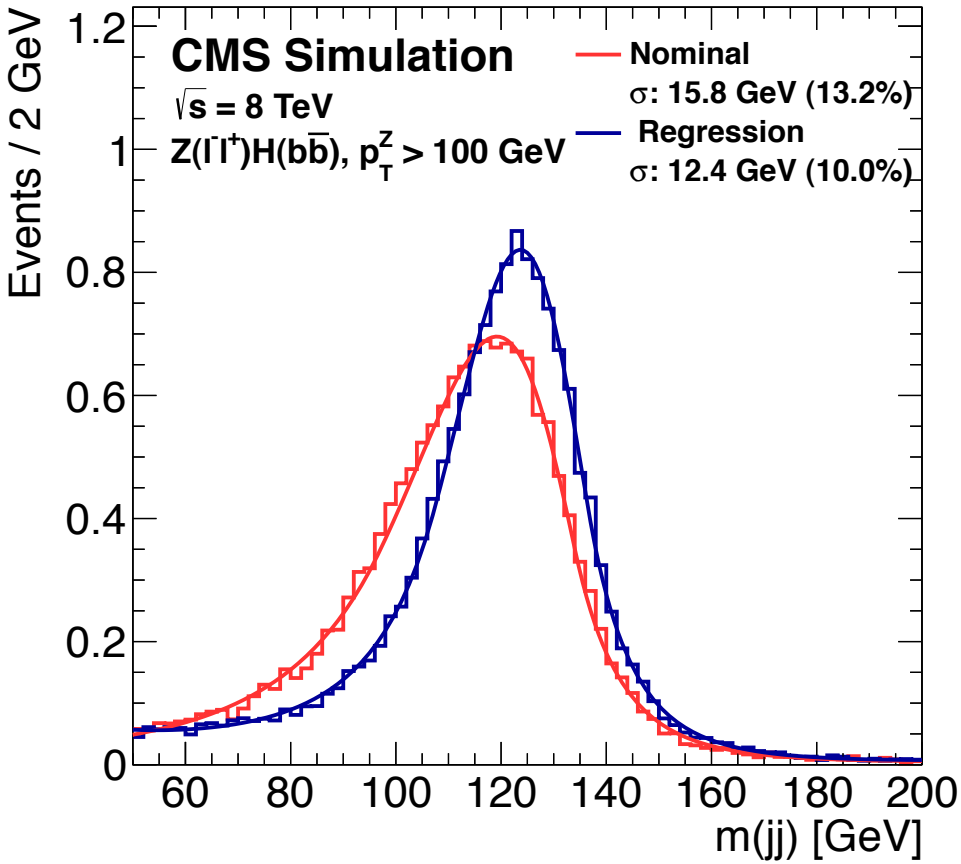
- $H \to bb$ has high branching ratios but huge QCD backgrounds
- To achieve reasonable S/B, select $W/Z + H \to \ell\nu\ \ell\ell\ \nu\nu + bb$ events with significant W/Z boost ($p_T^{W}/Z > 50$ or 100 GeV depending on the channel, with additional categories for higher pt regions)
- Events selected with two b-tagged jets (secondary-vertex-based b-tag discriminant)
- Significant backgrounds still remain from W/Z+jets, $t\bar{t}$, and diboson processes ($WW/ZZ/WZ$)
- Complex mixture of backgrounds with real b-jets and mistagged gluon/light quark jets

$W/Z + H \to bb$ mass reconstruction

- Energy of jets less precisely measured than charged leptons
- b-jet energy reconstruction improved using BDT regression
- Input variables included information on the relative charged/neutral hadron/electromagnetic fraction of the jet, details on the tracks and secondary vertex to correct for variations in the energy response from fluctuations in jet fragmentation, variation in track reconstruction efficiency and resolution with secondary vertex position, etc
- Additional variables on lepton kinematics included in case of semileptonic b-decays (regression infers/corrects for missing energy from the neutrino)
- Missing transverse momentum directly included in regression **only in** $H + Z \to \ell\ell$ **channel** (additional neutrinos from W/Z decays break correlation with neutrinos from b-decays)

$W/Z + H \to bb$ mass reconstruction

- After regression, dijet mass resolution is about 10%
- Mass resolution/signal purity not sufficient for simple bump hunt
- m_{bb} is instead used directly as input to subsequent BDT (ie the BDT is intentionally strongly correlated with the mass)

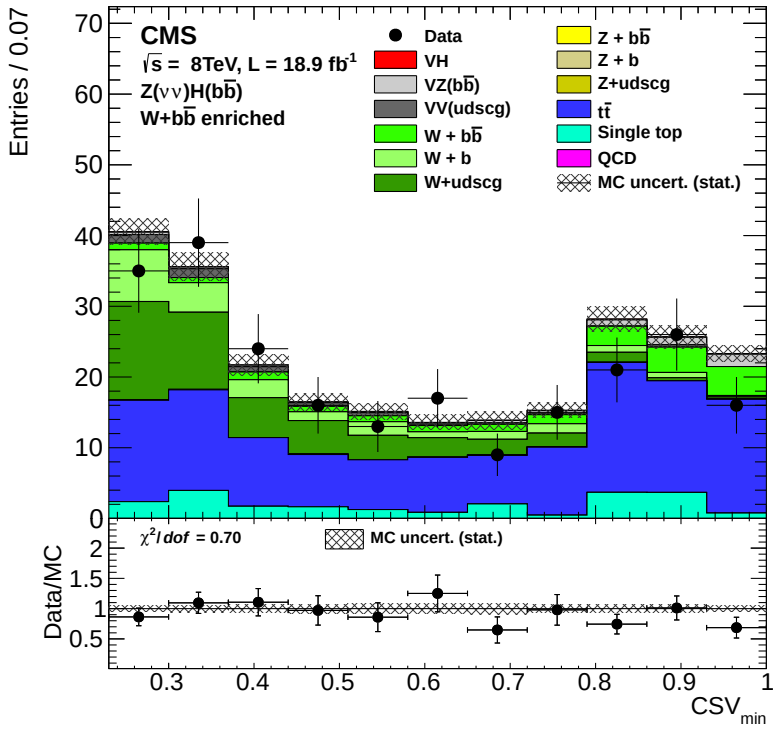


$W/Z + H \to bb$ Background scale factors

- Various background components are not well-predicted by simulation
- Fit data/mc scale factors for different background components in dedicated control regions for each channel
- Background yields scaled from inverted b-tagging (W/Z+light flavour), tighter b-tagging plus extra jets ($t\bar{t}$), M_{jj} sidebands (W/Z+$b\bar{b}$)
- W/Z+jets split into light flavour, light + 1 b, and 2 b components since relative fractions are not well-predicted

$W/Z + H \to bb$ Background scale factors

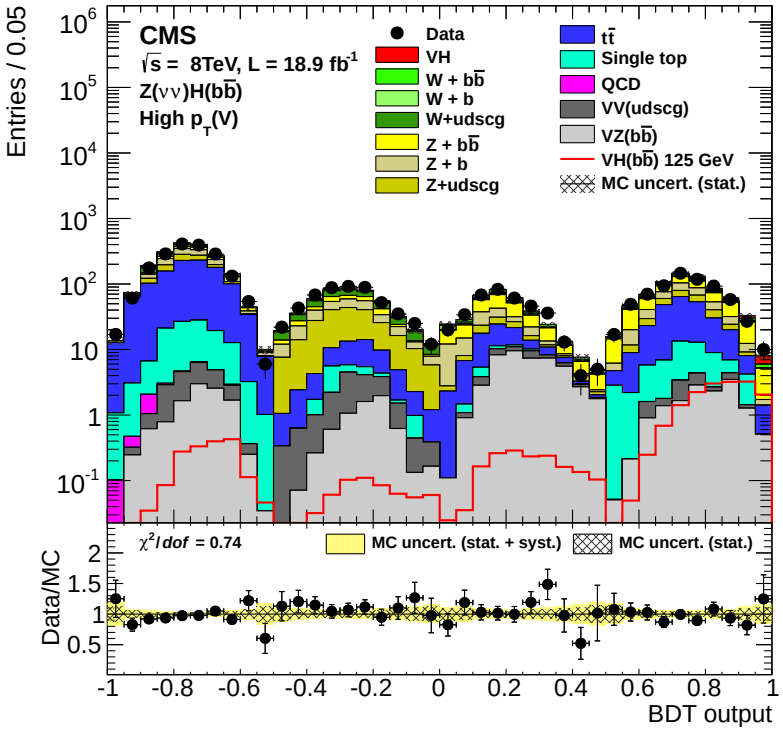
- Example shown here for high p_T (MET$>$ 170 GeV) $H + Z \to \nu\nu$ control region (m_{bb} sidebands) enriched in $W + bb$ by requiring an additional lepton
- Use of mass sidebands ensures this control region is orthogonal not just to $H + Z \to \nu\nu$ signal region, but also to $H + W \to \ell\nu$ signal region
- Scale factors extracted from simultaneous fits to b-tag discriminant distributions in different control regions
- Background normalizations are shown post-fit ($V + b$ has a scale factor close to 2, resulting form poor modeling of gluon spliting in the simulation)

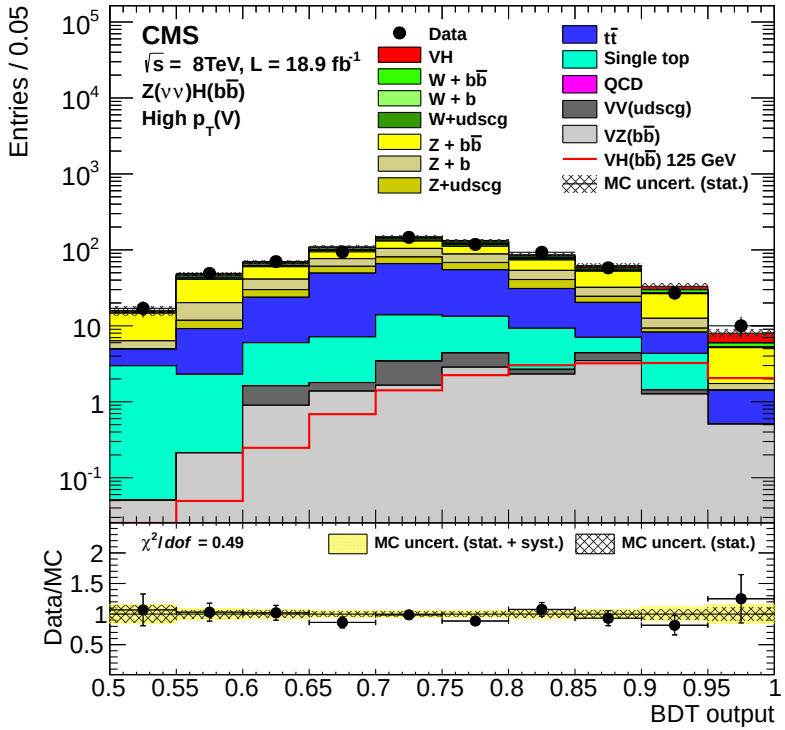


$W/Z + H \to bb$ Signal Extraction

- Even after determination of scale factors from control regions, backgrounds have non-negligible uncertainty
- Final sensitivity benefits from being able to further constrain background normalizations in the final fit
- Procedure:
  1. Train four BDT's for each channel: signal vs $t\bar{t}$, signal vs W/Z+jets, signal vs dibosons, signal vs (all) background
  2. Cuts on background-specific BDT's are used to **partition** final signal vs (all) background distribution into four subsets

$W/Z + H \to b\bar{b}$ Signal Extraction



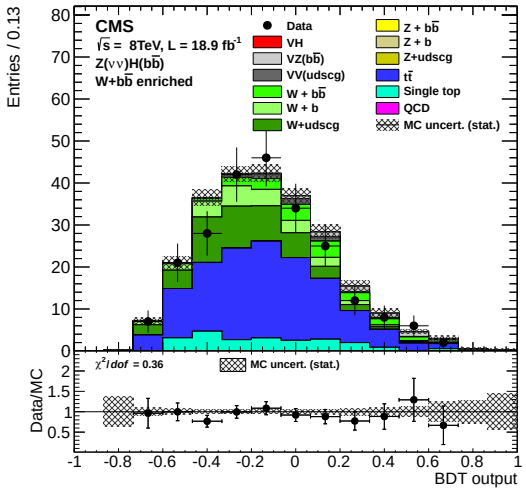


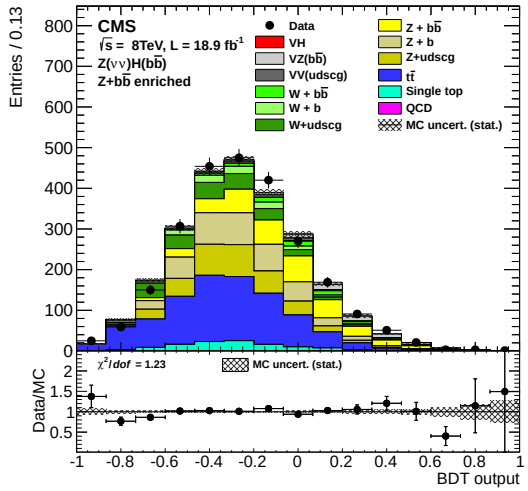
- Results extracted from fit to final BDT distribution, partitioned using dedicated BDT's into individual background and signal-enriched regions

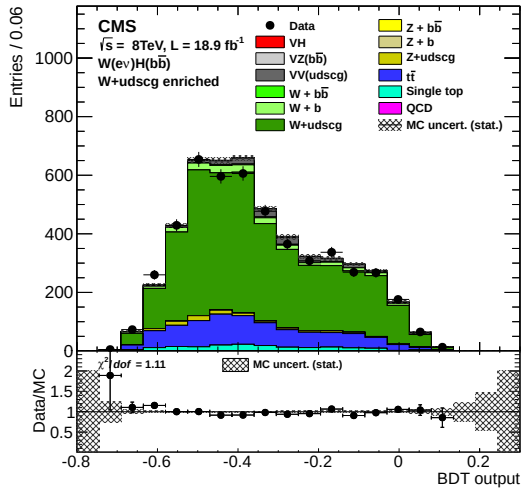
$W/Z + H \to bb$ Signal Extraction

- Input variables for BDT's:
  - Several kinematic variables for selected jets (including dijet mass) and W/Z candidate (lepton, missing transverse momentum kinematics)
  - Number of additional jets
  - b-tag discriminant value for selected and additional jets
- Jet energy scale and b-tag discriminant uncertainties enter as **shape uncertainties** for final BDT distributions

$W/Z + H \to b\bar{b}$ Signal Extraction Controls







- Final BDT distribution also validated in control regions