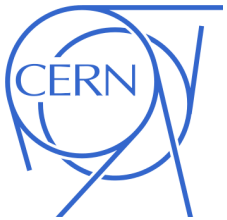
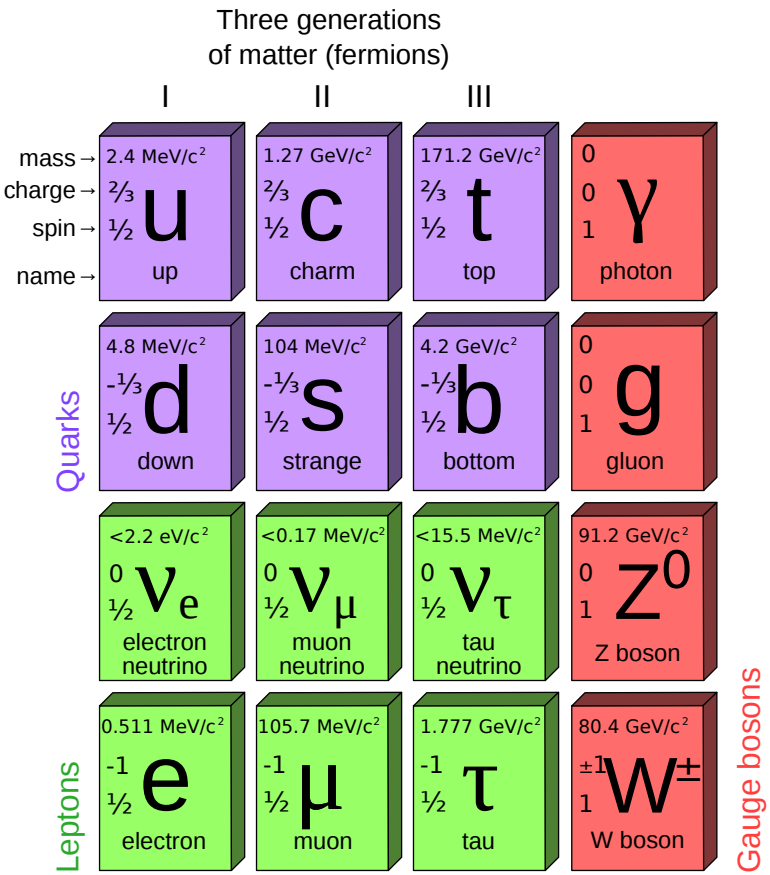
Machine Learning and the Higgs search and discovery at CMS

Josh Bendavid
(CERN)



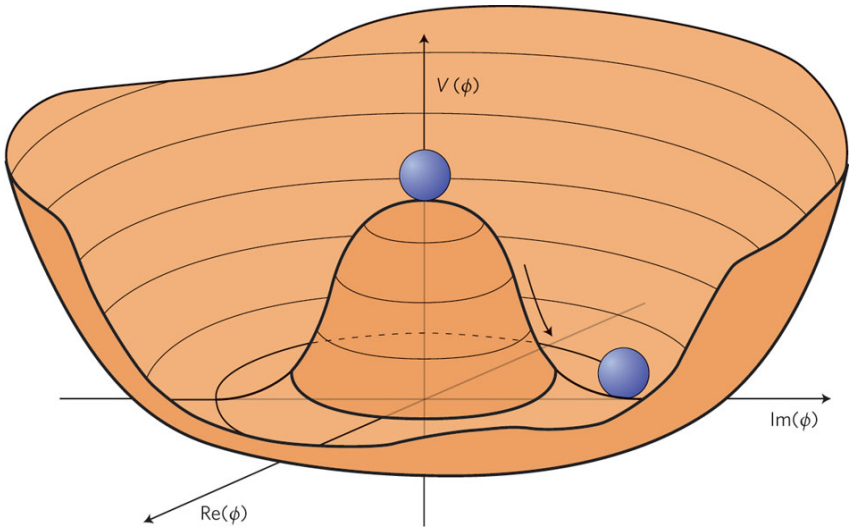
Summer school on Machine Learning in High Energy Physics
Saint-Petersburg, Russia
Aug 30, 2015

The Standard Model



- The Standard Model of Particle Physics describes all known physics (except for gravity)
- Relativistic quantum field theory of fermions and gauge bosons describing strong and electroweak interactions
- With this alone, theory predicts massless fermions and bosons...

Higgs Mechanism



- Complex scalar doublet

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_0 + i\phi_3 \end{pmatrix}$$

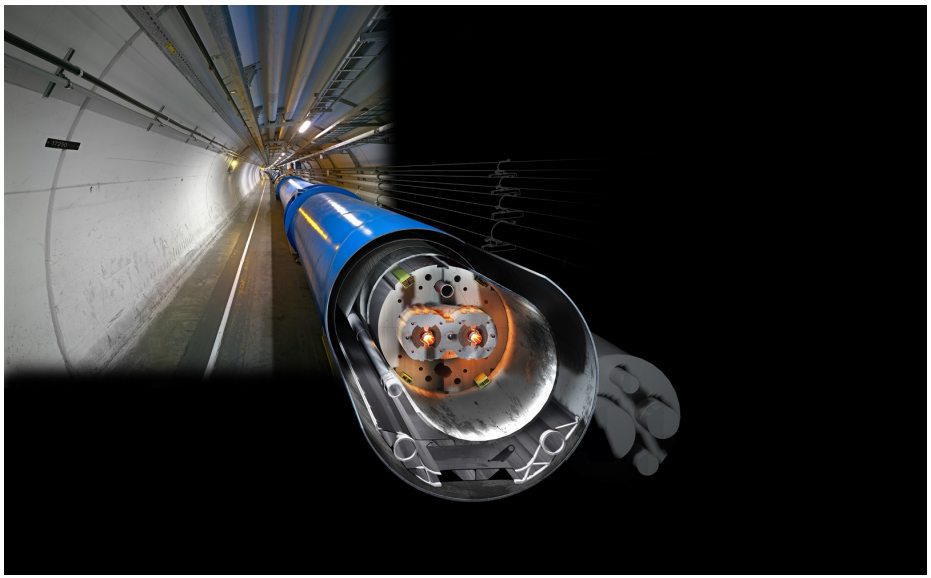
- Symmetry of vacuum is spontaneously broken leading to a ground state $\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H \end{pmatrix}$, where H is a massive scalar field
- The other three degrees of freedom become the longitudinal polarizations of the W^+, W^-, and Z bosons, giving them mass
- Additional Yukawa couplings between the Higgs field and the fermions lead to fermion mass terms due to v

Higgs Boson

- Remaining component of the Higgs field behaves as a massive scalar boson, last particle to be observed in the standard model
- Coupling of the Higgs boson to the standard model particles predicted by the theory:
  - Coupling to W and Z bosons
  - Coupling to fermions proportional to their mass
- Mass of the Higgs is a free parameter of the theory (but unitarity of WW scattering in the SM requires $m_H <\sim 1$ TeV)
- Entire region $m_H < 114.4$ GeV ruled out by previous experiments
- Indirect electroweak constraints prefer a relatively light Higgs mass ($m_H < 152$ GeV at 95% C.L)
- Higgs discovered by ATLAS and CMS collaborations in summer 2012

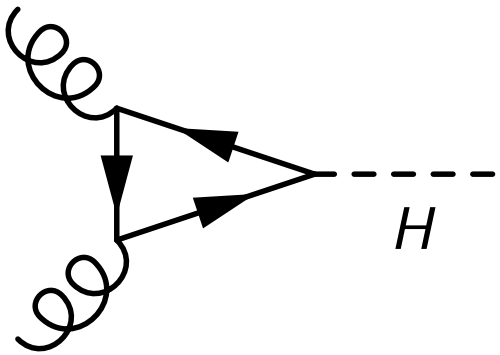
The Large Hadron Collider

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

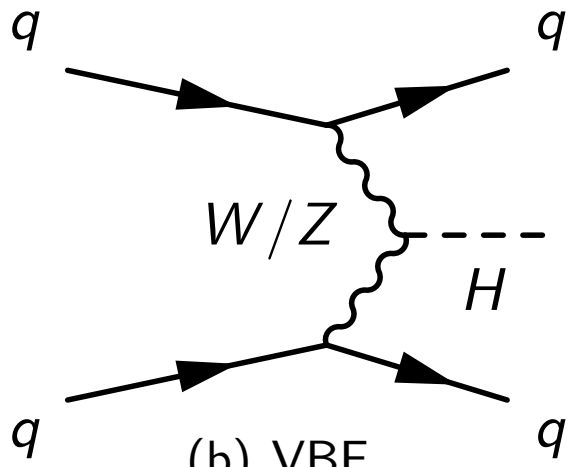


- Superconducting dipole magnets with a design field of 8.3 T, cooled to 1.9 K using superfluid helium

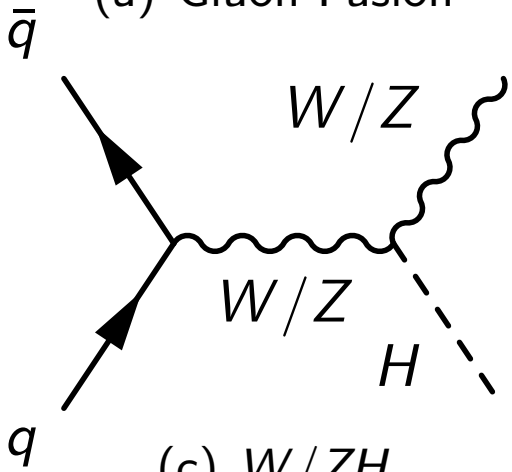
Higgs Production Processes at LHC



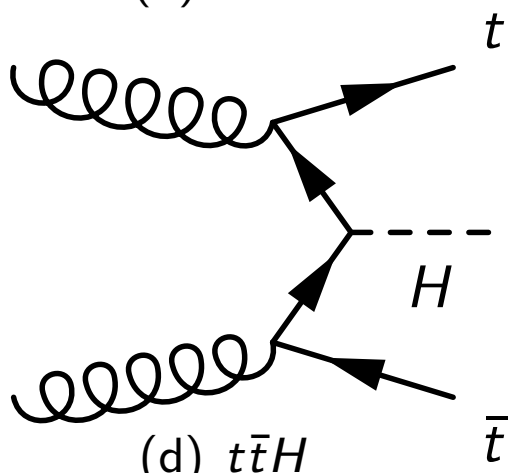
(a) Gluon Fusion



(b) VBF

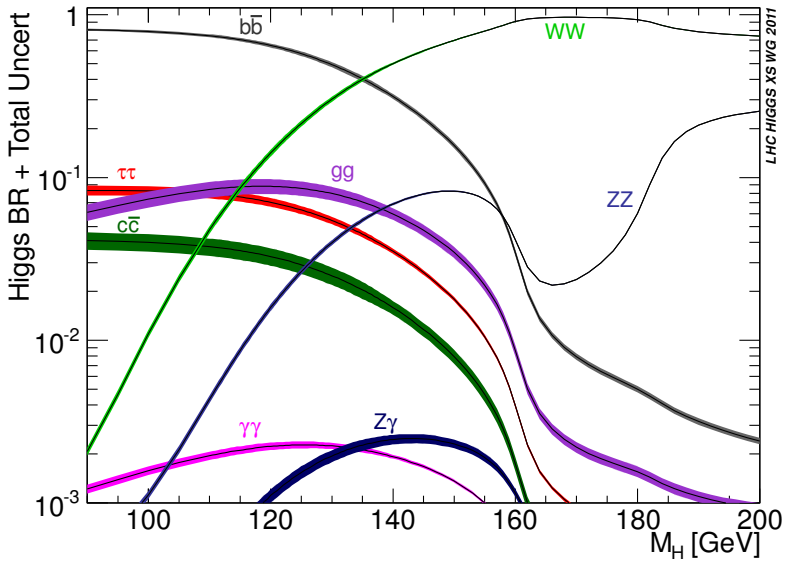


(c) W/ZH

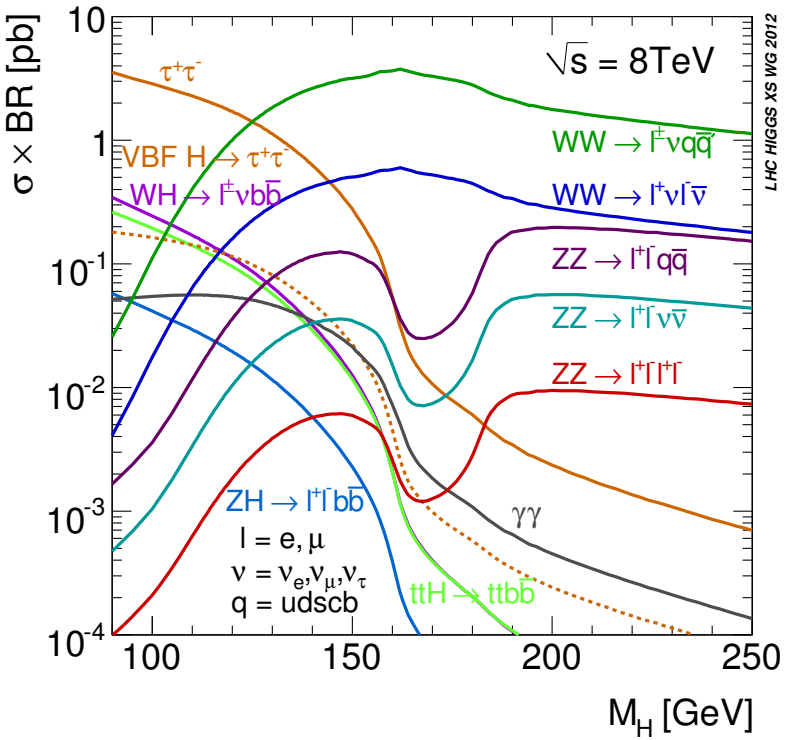


(d) $t\bar{t}H$

Higgs Production and Decay at LHC



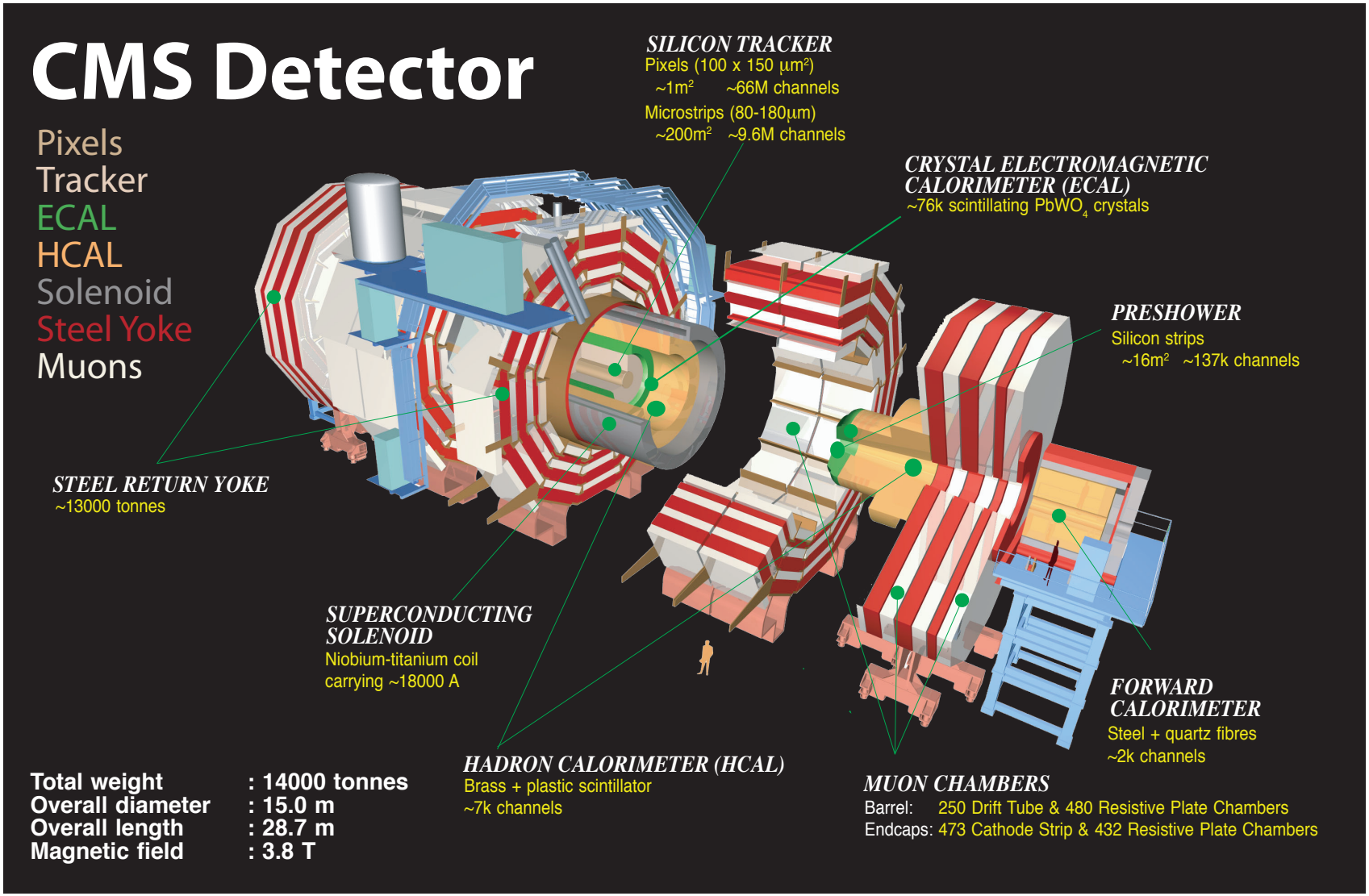
(e) Branching Ratios



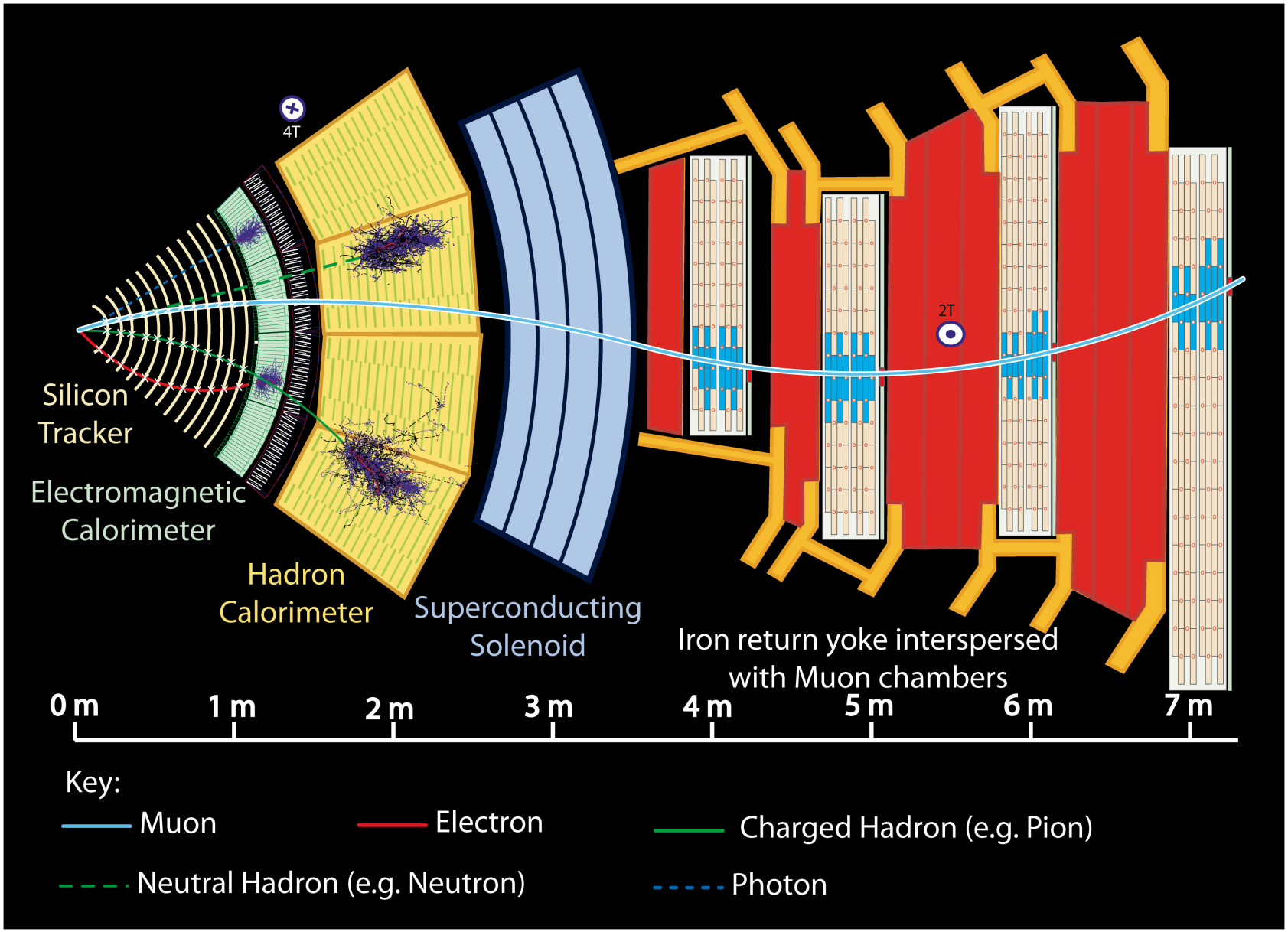
(f) Cross Sections

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

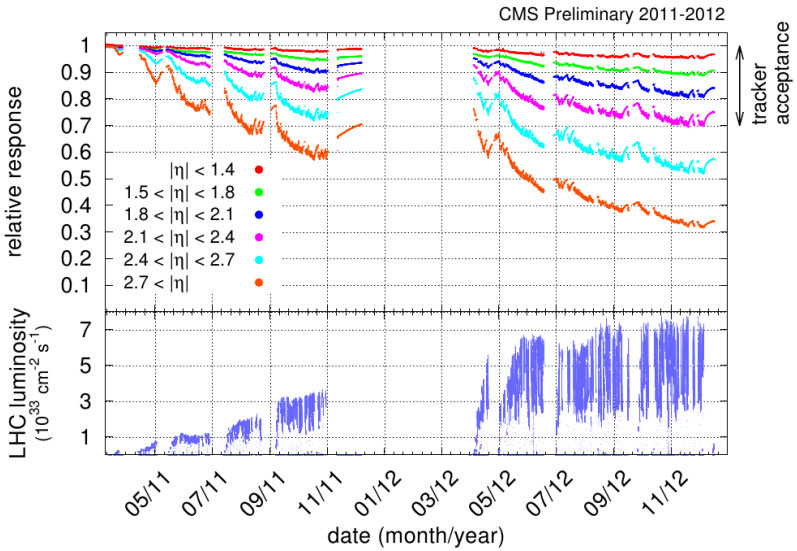
The CMS Detector



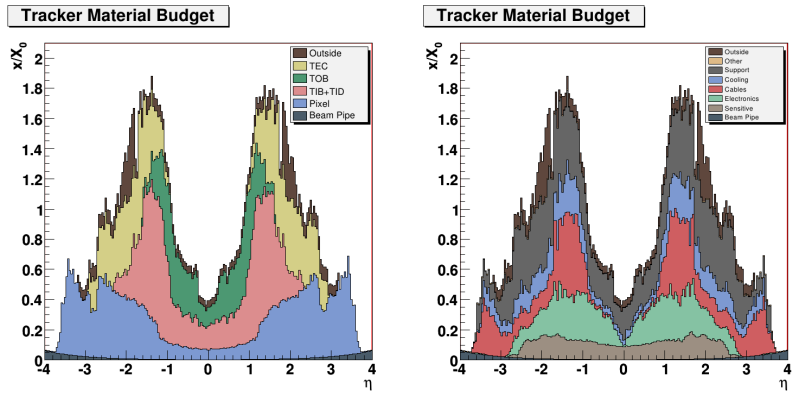
The CMS Detector



The CMS Detector: Some Challenges



(g) ECal Transparency Loss



(h) Tracker Material Budget

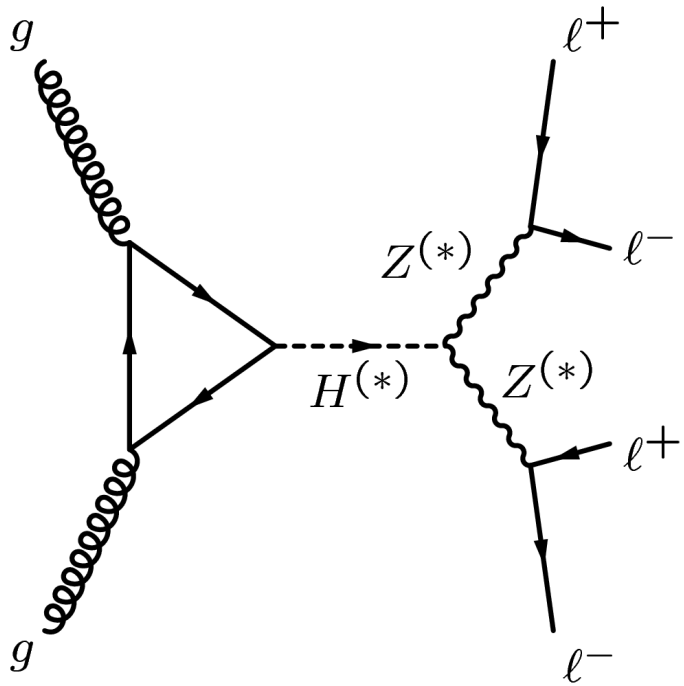
- ECal crystals lose and recover transparency under exposure to radiation
- Monitored in situ with LED/laser monitoring system, but still a major challenge for calibration
- Lots of material in front of the ECal

Machine Learning Techniques in Higgs Analyses

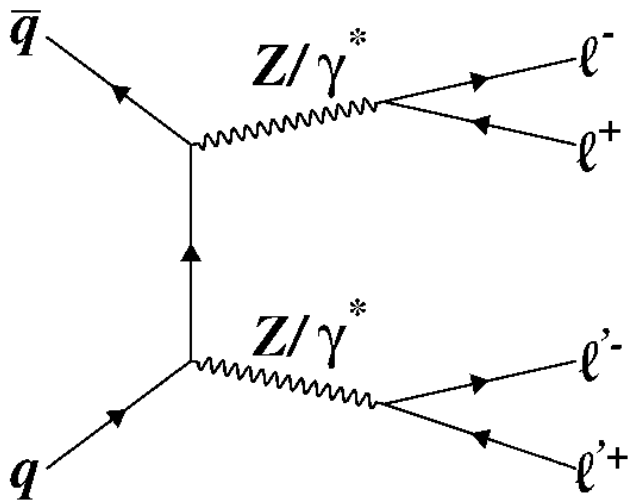
- Machine learning techniques extensively used in Higgs search/discovery, will cover a few specific cases and their particularities
  - $H \to ZZ \to 4\ell$: Use of Matrix Element likelihood techniques for well understood and well measured background
  - $H \to b\bar{b}$: Use of BDT classifiers for complex set of backgrounds with large systematic uncertainties
  - $H \to \gamma\gamma$: Use of BDT regression/classifiers for photon reconstruction/selection, BDT classifiers to distinguish different signal components, interplay with mass fits

$H \to ZZ \to 4\ell$

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



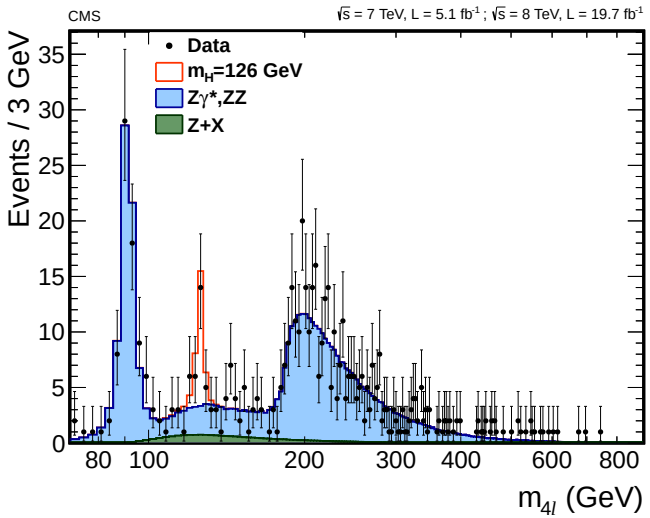
(i) Main signal

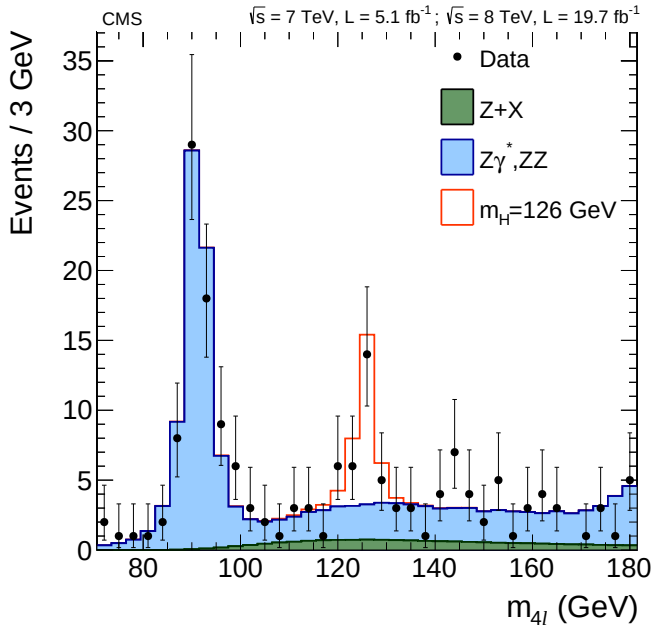


(j) Main background

$H \to ZZ \to 4\ell$

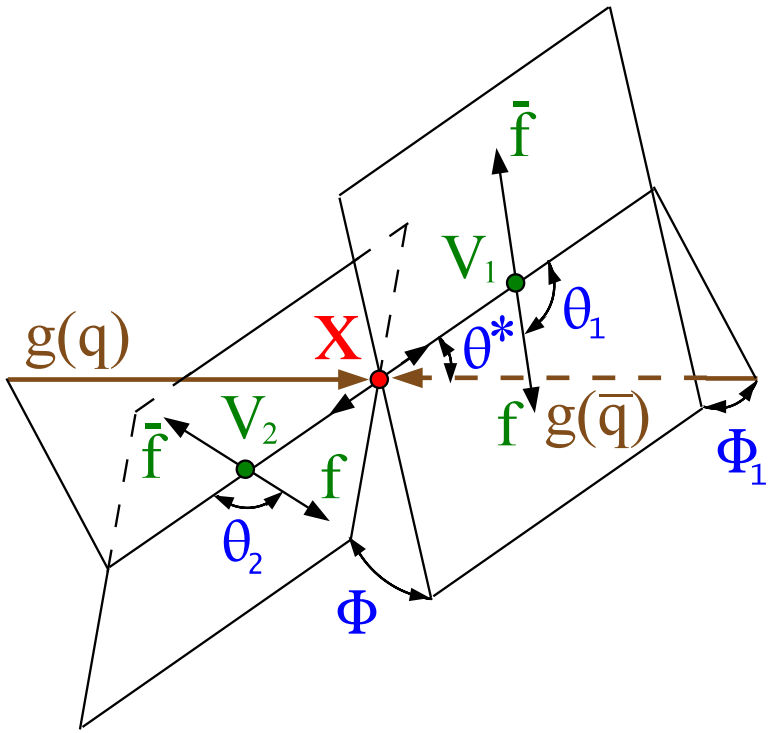
- Select 4 leptons of appropriate charge and flavour combinations (+FSR recovery) with $40 < m_{Z1} < 120$ GeV, $12 < m_{Z2} < 120$ GeV
- Electron acceptance: $|\eta| < 2.5$, $p_T > 7$ GeV, Muon acceptance: $|\eta| < 2.4$, $p_T > 5$ GeV
- Irreducible $ZZ \to 4\ell$ continuum background estimated from 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





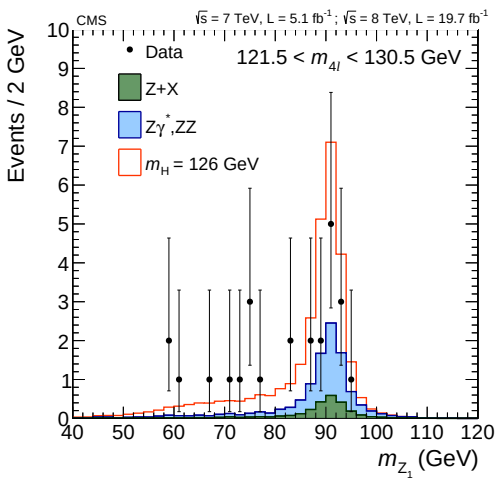
$H \to ZZ \to 4\ell$: Beyond the mass distribution

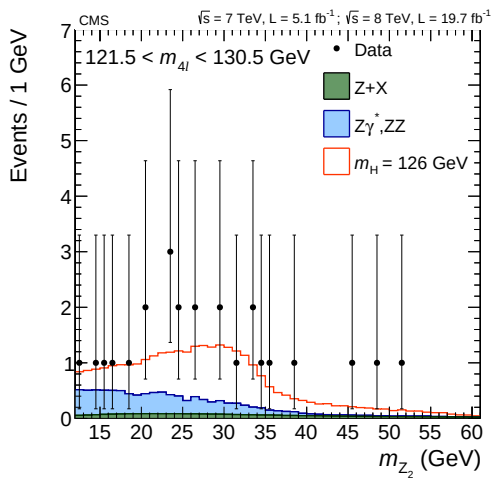
- Higgs is a scalar $\to$ decay angles θ_1,θ_2,Φ, and lepton pair masses m_{Z1},m_{Z2} provide additional discrimination against continuum background

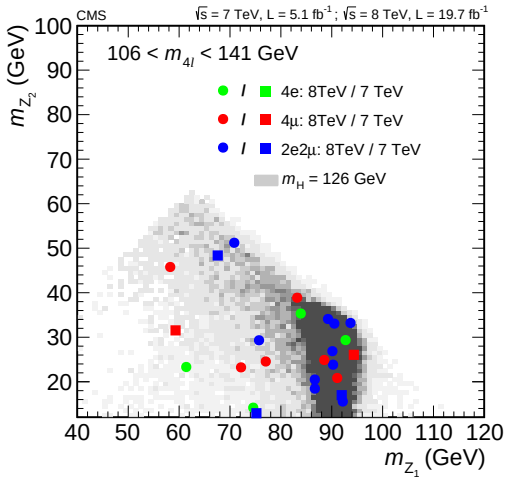


$H \to ZZ \to 4\ell$: Beyond the mass distribution

- 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

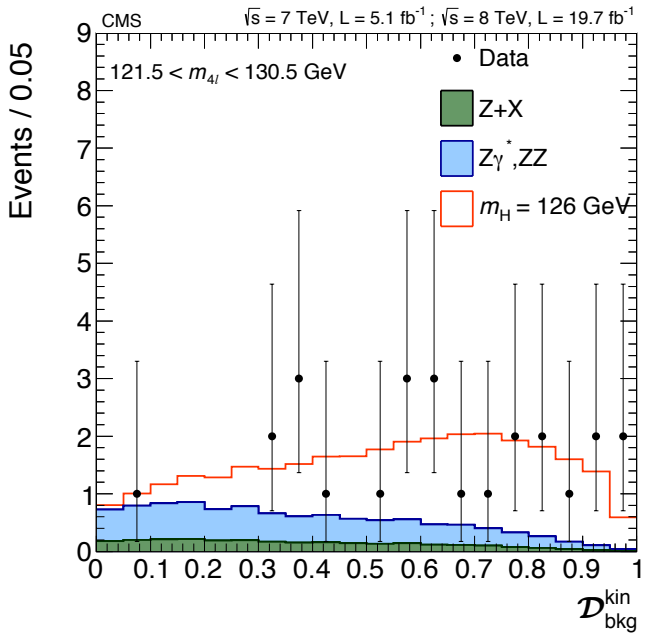
- For $H \to ZZ \to 4\ell$, final state is fully reconstructed, and charged leptons have excellent momentum resolution in CMS ($\mathcal{O}(\%)$)
- Matrix element likelihood discriminator constructed directly from dilepton pair masses, plus decay angles as:

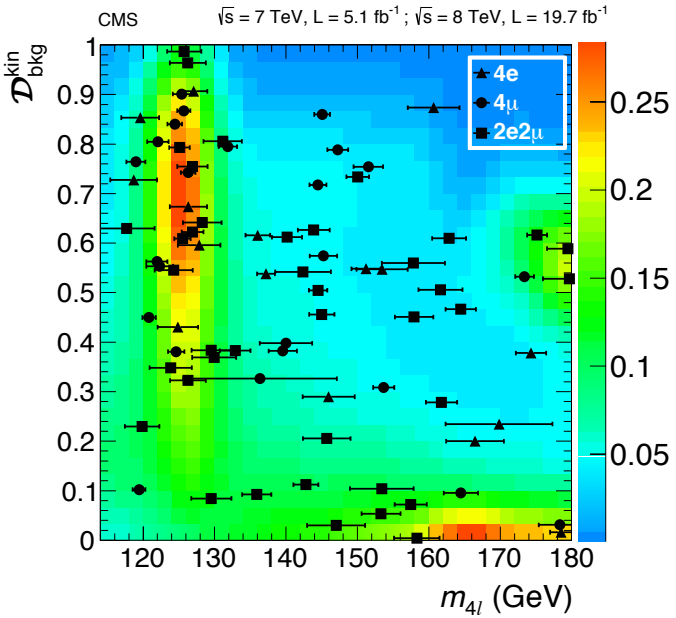
$$D = \frac{p_{sig}(m_{Z1}, m_{Z2}, \theta_1, \theta_2, \Phi | m_{4\ell})}{p_{sig}(m_{Z1}, m_{Z2}, \theta_1, \theta_2, \Phi | m_{4\ell}) + p_{bkg}(m_{Z1}, m_{Z2}, \theta_1, \theta_2, \Phi | m_{4\ell})}$$

- 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

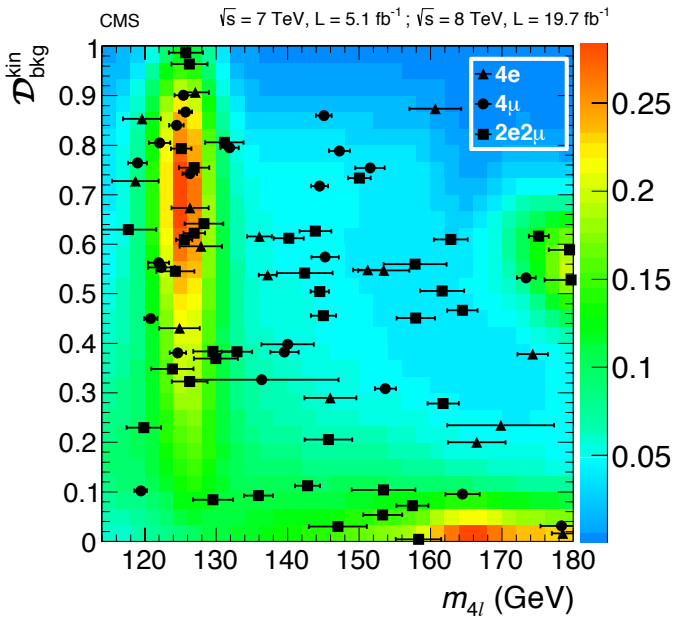
- Signal strength results extracted from 3d unbinned maximum likelihood fit to $m_{4\ell}$ distribution with matrix element likelihood discriminant and $p_T^{4\ell}$

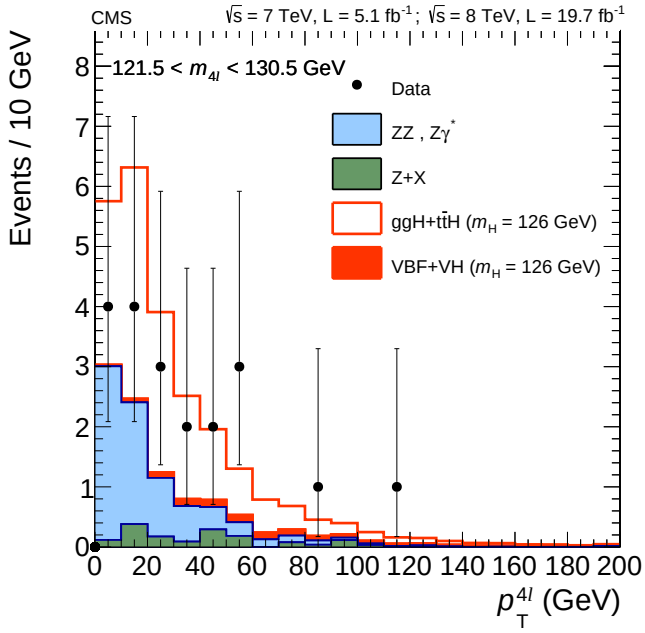




$H \to ZZ \to 4\ell$ Results

- 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

$W/Z + H \to bb$

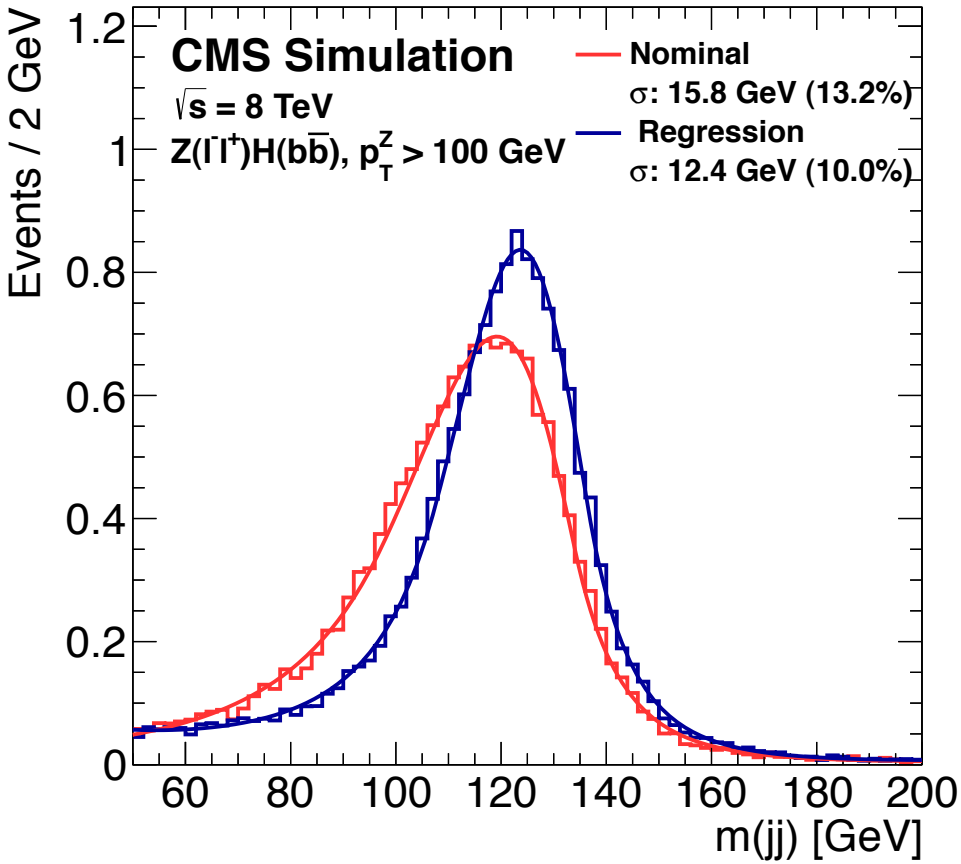
- $H \to bb$ has high branching ratios but huge QCD backgrounds
- To achieve reasonable S/B, select $W/Z + H \to \ell\nu\ \ell\ell\ \nu\nu + bb$ events with significant W/Z boost ($p_T^{W}/Z > 50$ or 100 GeV depending on the channel, with additional categories for higher pt regions)
- Events selected 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)

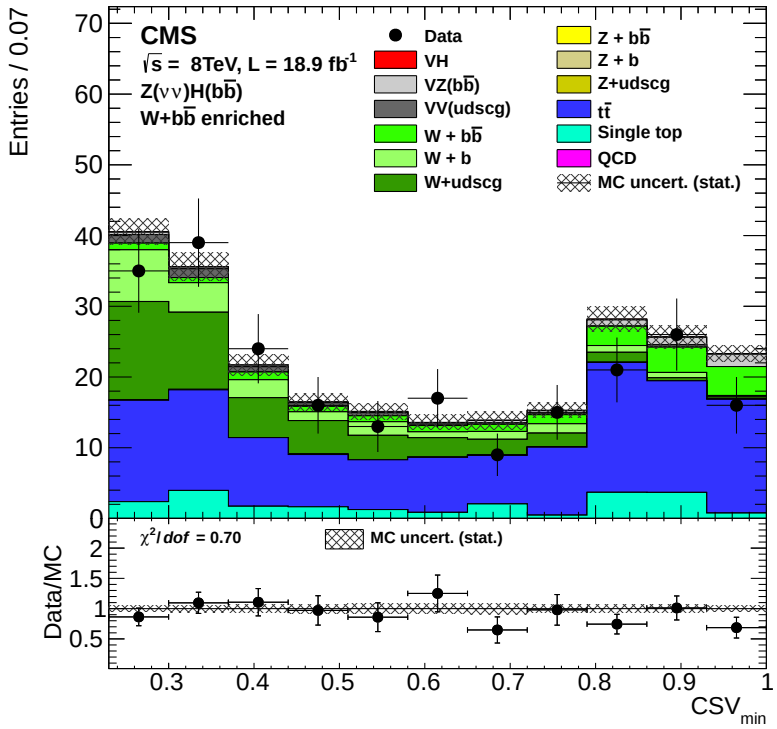


$W/Z + H \to bb$ Background scale factors

- 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

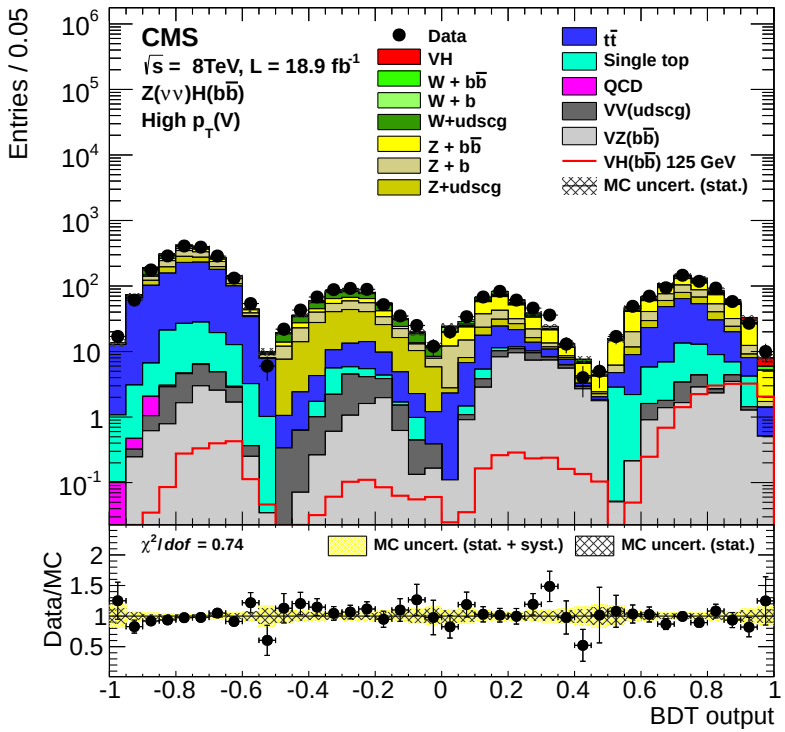
- Example shown here for high p_T (MET$>$ 170 GeV) $H + Z \to \nu\nu$ control region (m_{bb} sidebands) enriched in $W + bb$ by requiring an additional lepton
- Use of mass sidebands ensures this control region is orthogonal not just to $H + Z \to \nu\nu$ signal region, but also to $H + W \to \ell\nu$ 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)

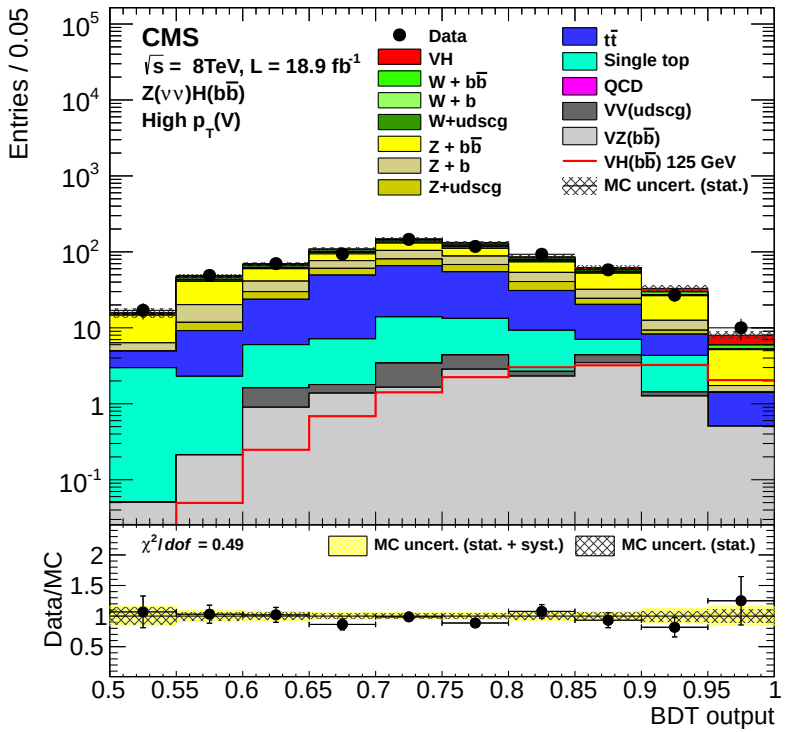


$W/Z + H \to bb$ Signal Extraction

- 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

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



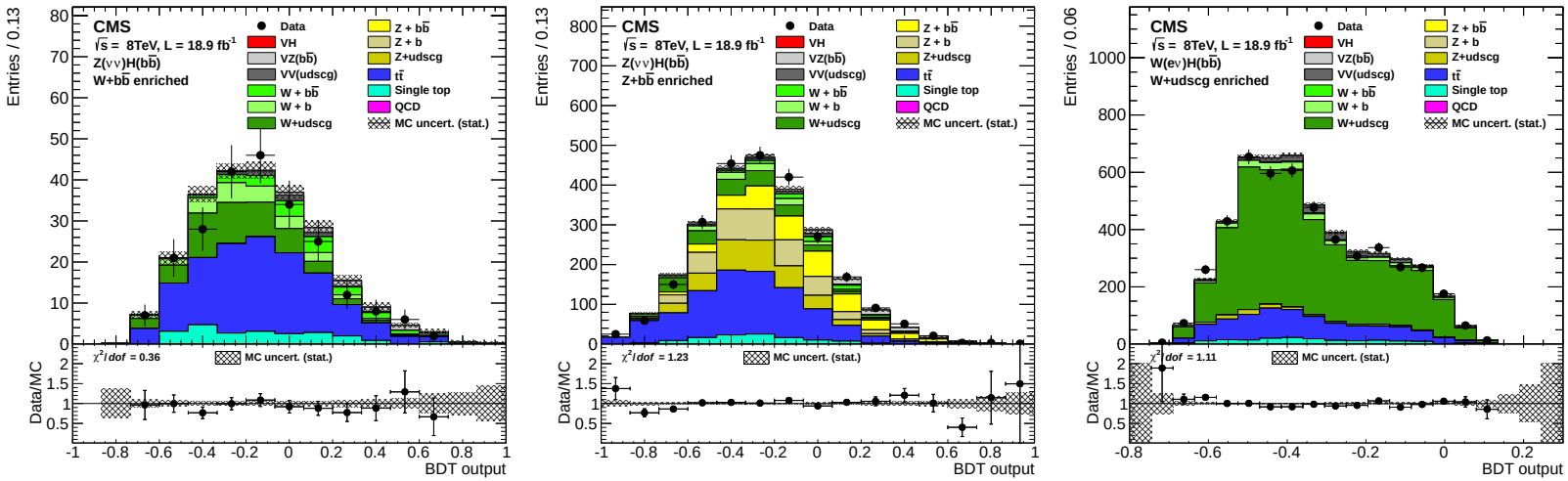


- Results extracted from fit to final BDT distribution, partitioned using dedicated BDT's into individual background and signal-enriched regions

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

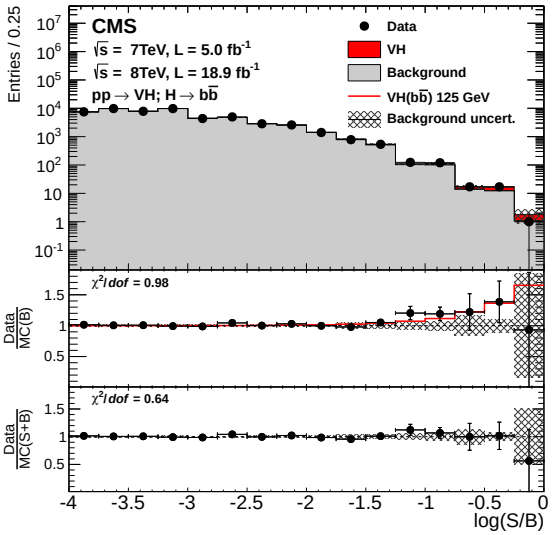
- Input variables for BDT's:
  - Several kinematic variables for selected jets (including dijet mass) and W/Z candidate (lepton, missing transverse momentum kinematics)
  - Number of additional jets
  - b-tag discriminant value for selected and additional jets
- Jet energy scale and b-tag discriminant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

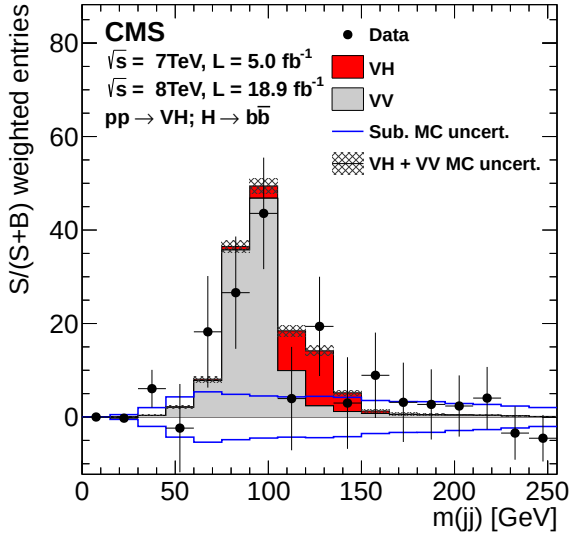
$W/Z + H \to b\bar{b}$ Signal Extraction Controls

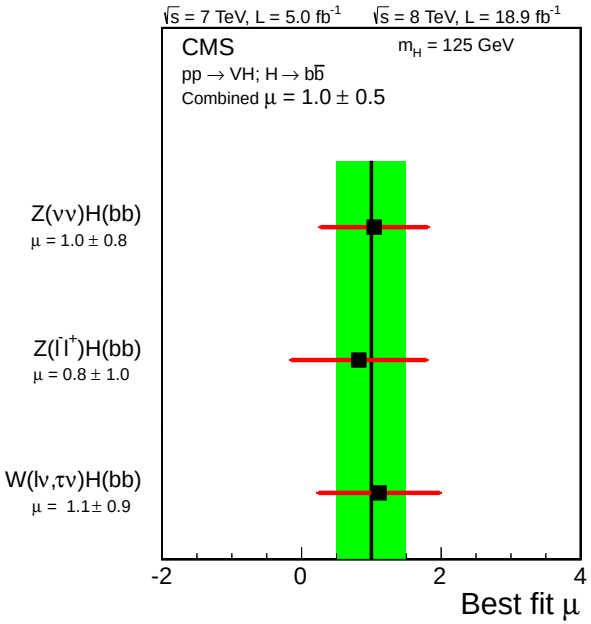


- Final BDT distribution also validated in control regions

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



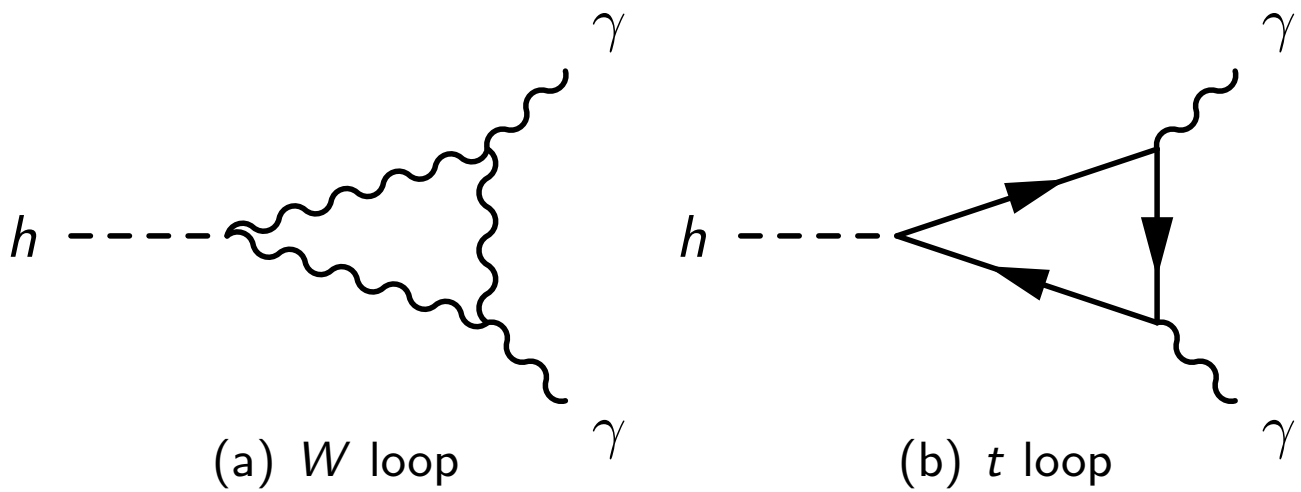




- $\sigma/\sigma_{SM} = 0.84 \pm 0.44$ (including also contribution from $tt + H \to b\bar{b}$ and taking into account gluon-induced $Z + H$ production)
- Observed Significance 2.0σ (Expected 2.6σ)

Higgs$\to \gamma\gamma$ Decay

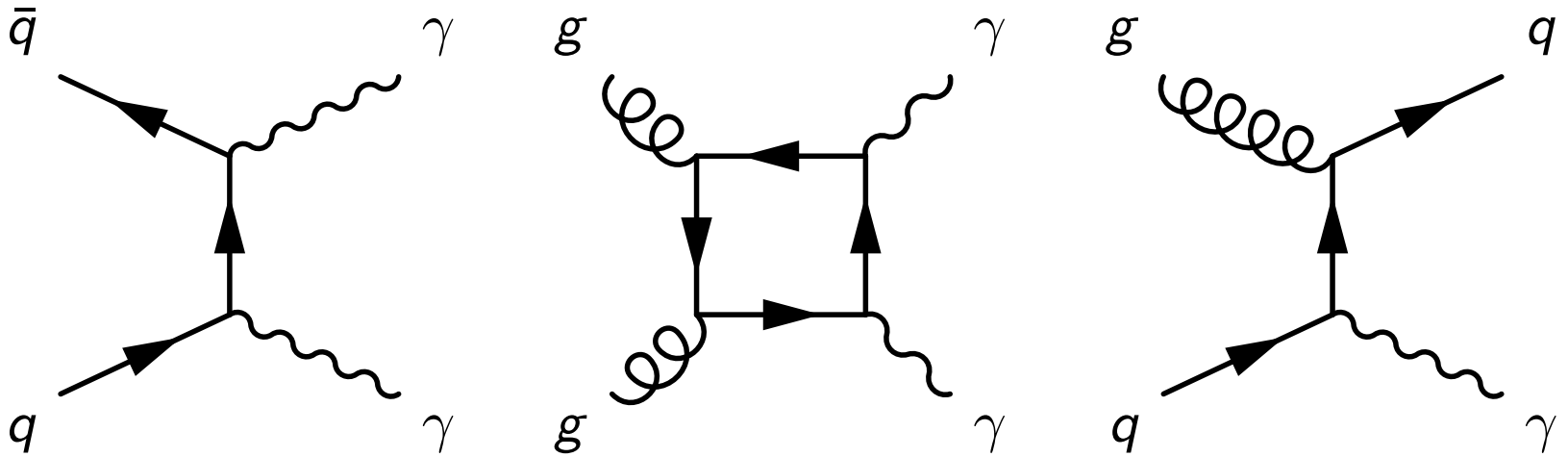
- No tree-level $h\gamma\gamma$ vertex, decay proceeds through W and fermion (top) loops which interfere destructively
- Branching ratio to two photons very sensitive to fermion vs boson couplings and possible new particles in the loop



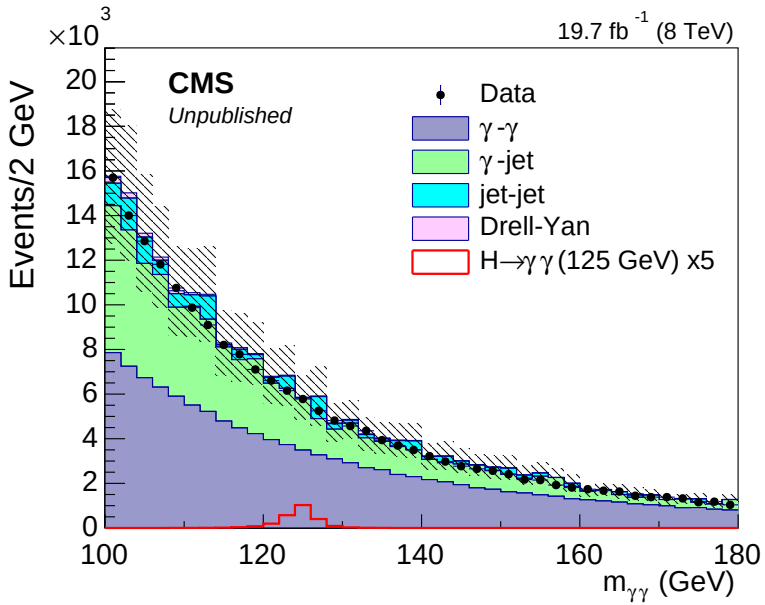
(a) W loop (b) t loop

Higgs$\to \gamma\gamma$ Analysis Overview

- Higgs$\to$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

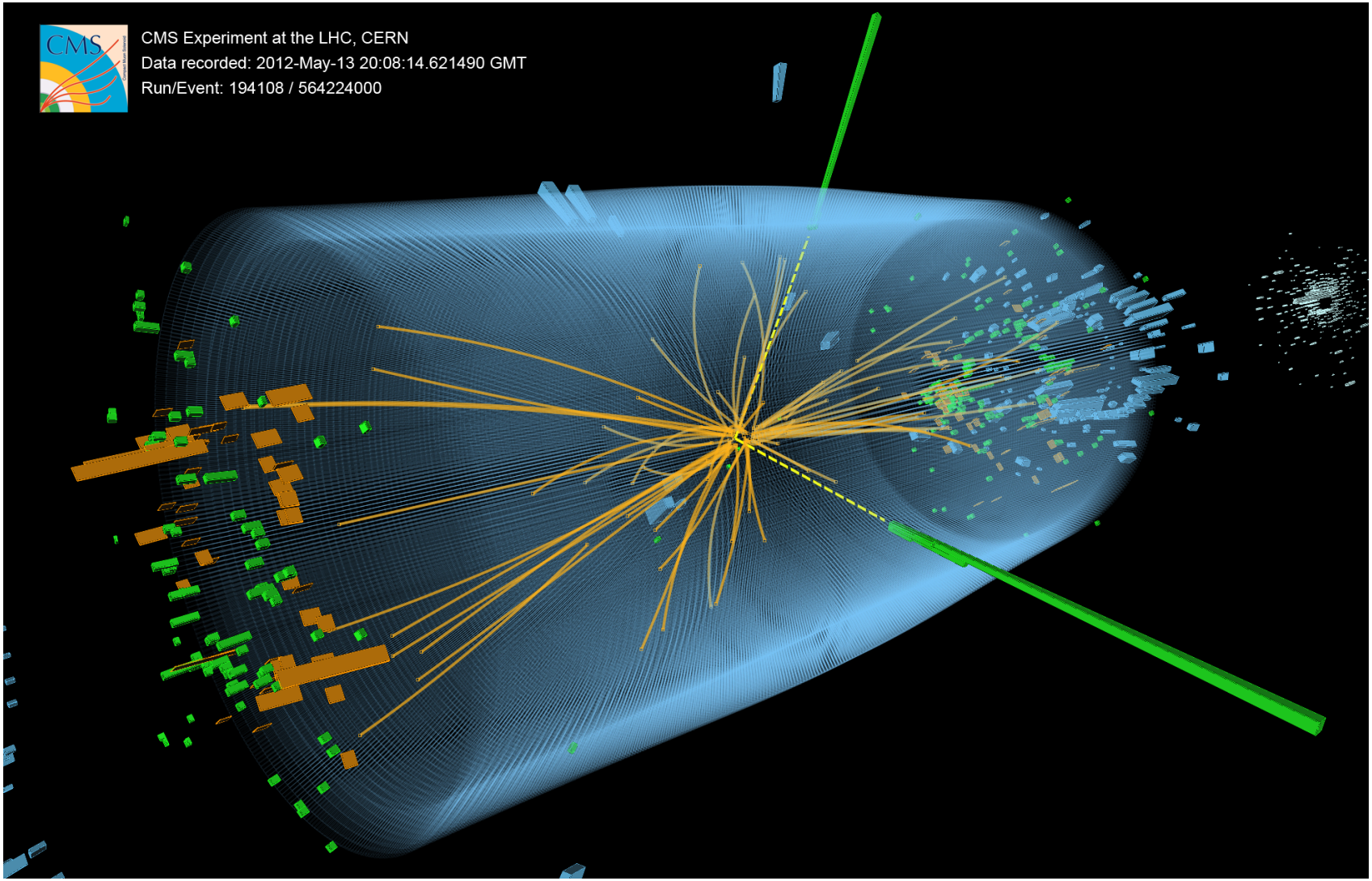


Inclusive selection with coarse binning

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

- 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

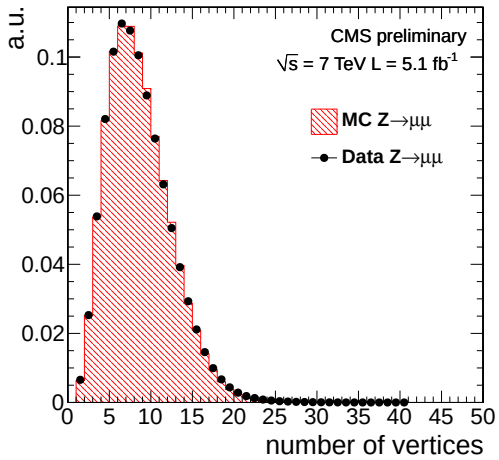
Higgs$\to \gamma\gamma$ Analysis Overview



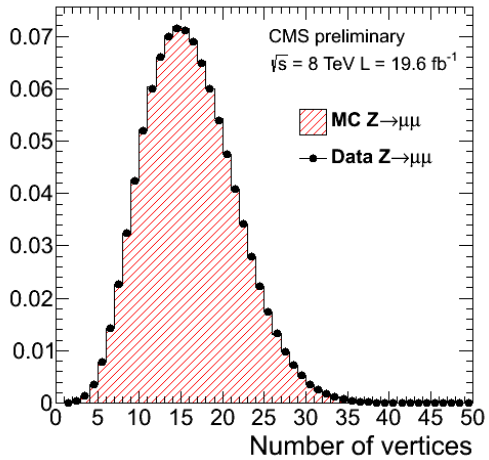
$$m_{\gamma\gamma} = 125.9 \text{ GeV}$$

Dataset/Pileup Conditions

- 5.1 fb^{-1} of 7 TeV data from 2011, 19.7 fb^{-1} of 8 TeV data from 2012



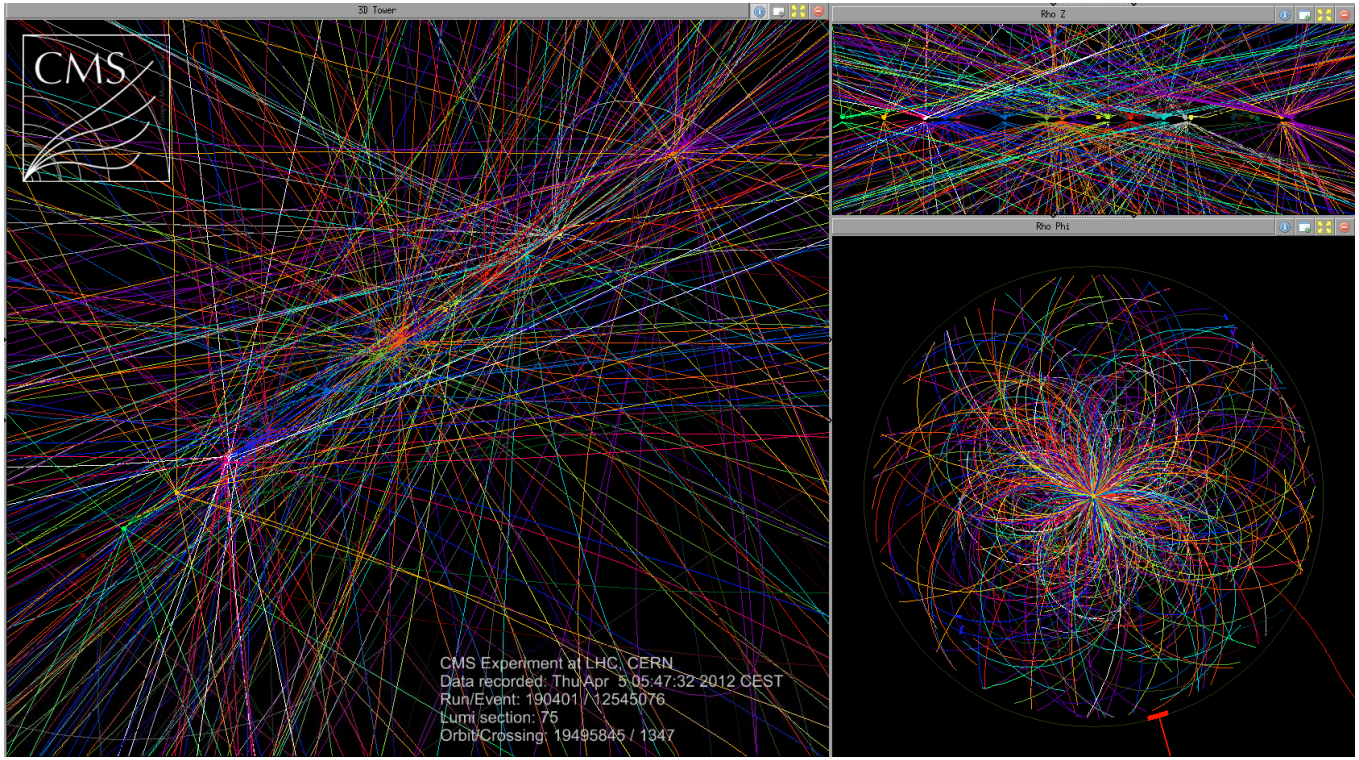
(c) 7 TeV, $< \text{NPU} >= 9.5$



(d) 8 TeV, $< \text{NPU} >= 19.9$

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

Pileup in CMS



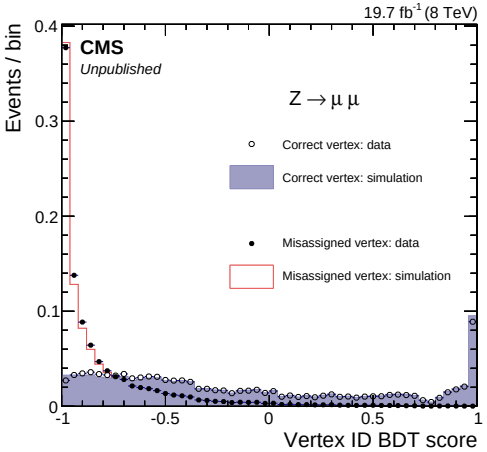
- An event with 29 reconstructed primary vertices

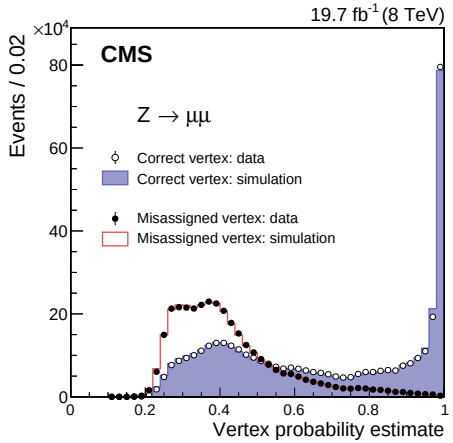
Higgs$\to \gamma\gamma$ Analysis Overview

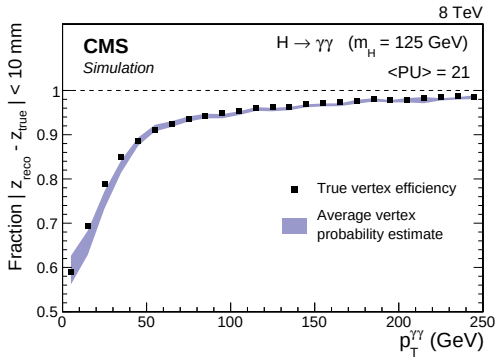
1. Primary Vertex Selection (Vertex Selection MVA)
2. Photon Selection (Preselection + Photon-jet MVA discriminator)
3. Multivariate Regression for EM Cluster corrections with per-photon resolution estimate
4. Energy Scale and Resolution corrections from $Z \to ee$
5. Event Categorization (MVA Discriminator)
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

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







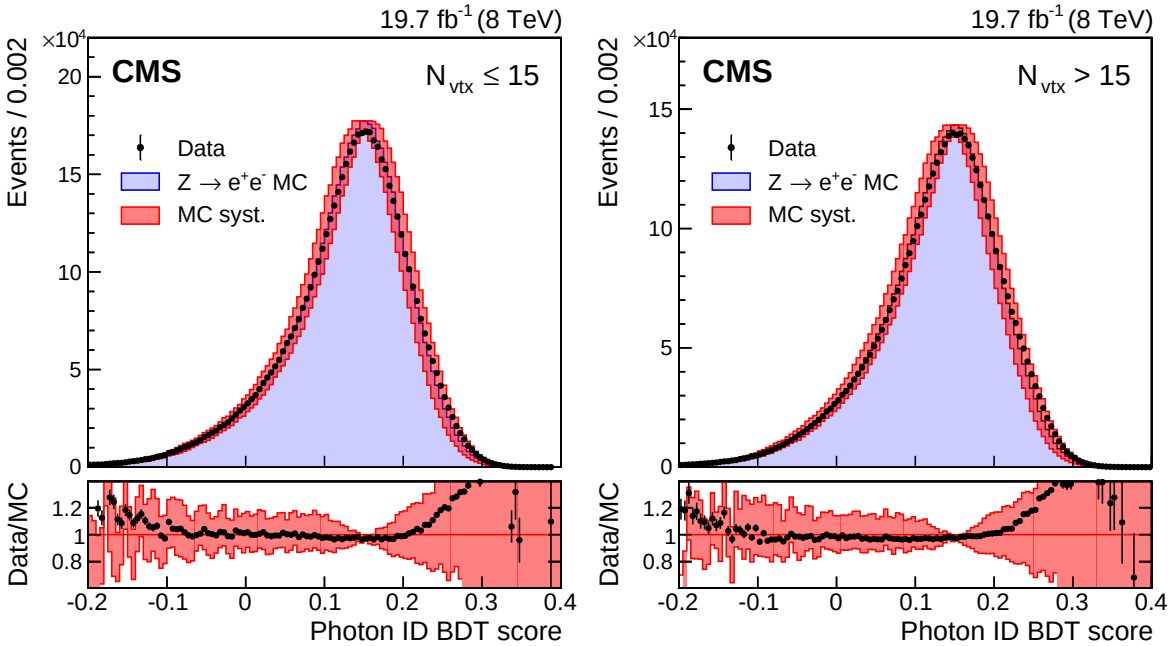
- Inclusive vertex selection efficiency ∼80 %, but strong dependence on Higgs p_T

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)

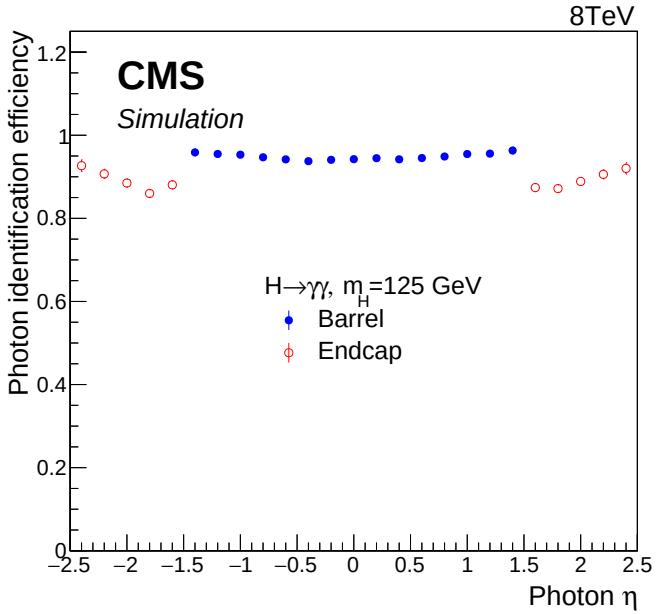
Photon Identification: MVA

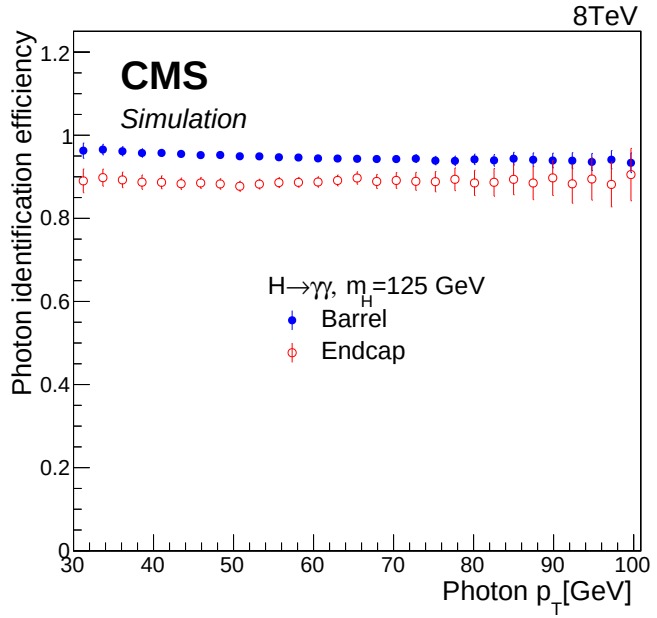
- Start with a very loose pre-selection matching trigger requirements
- Construct a multivariate discriminator using a BDT trained on prompt photons vs fakes from jets in MC, using shower and isolation variables as input
- Only a loose cut on the ID MVA value, which is fed forward to the final di-photon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



Photon Identification: MVA

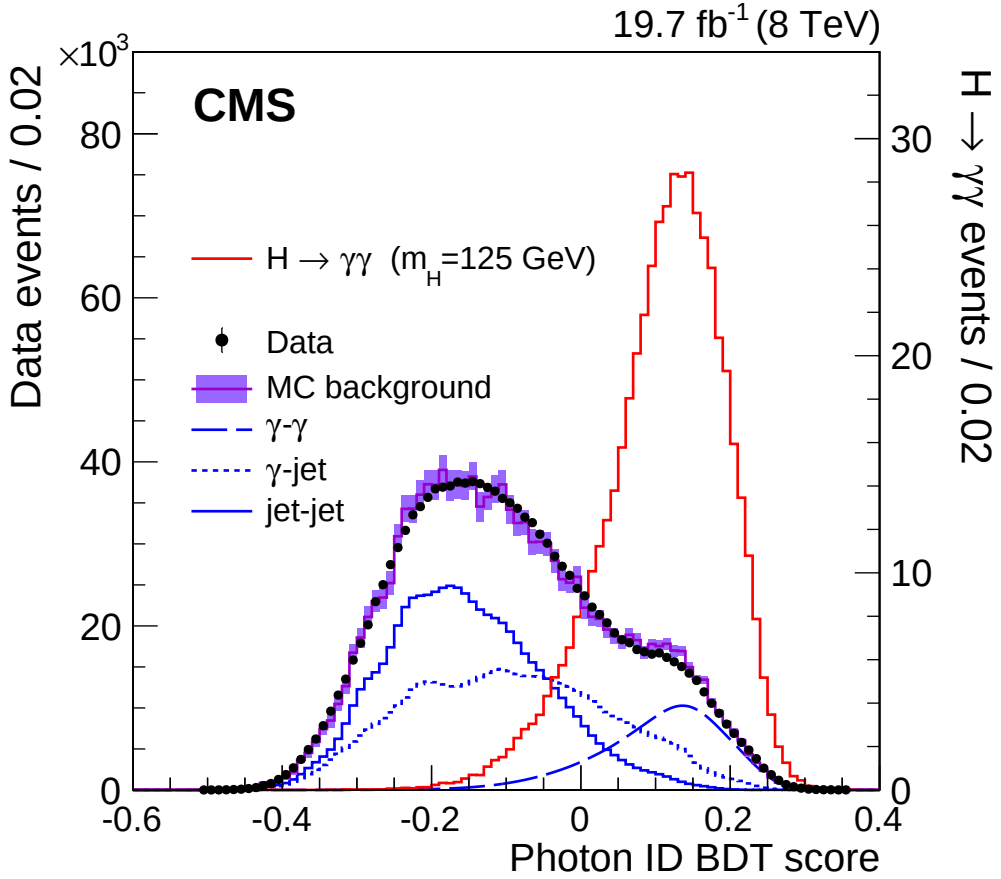
- Photon identification intended to be uncorrelated with photon kinematics (p_T and rapidity), in order to avoid shaping the mass distribution and allow kinematics to be optimally exploited by event level BDT
- Signal training sample reweighted in two-dimensions (p_T,η) to match background training sample 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)





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)



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 (Shower containment, crack/gap corrections, PU effects)
- 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 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
- 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 $y = |E_{Cor}/ERaw - E_{True}/E_{Raw}|$

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.

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

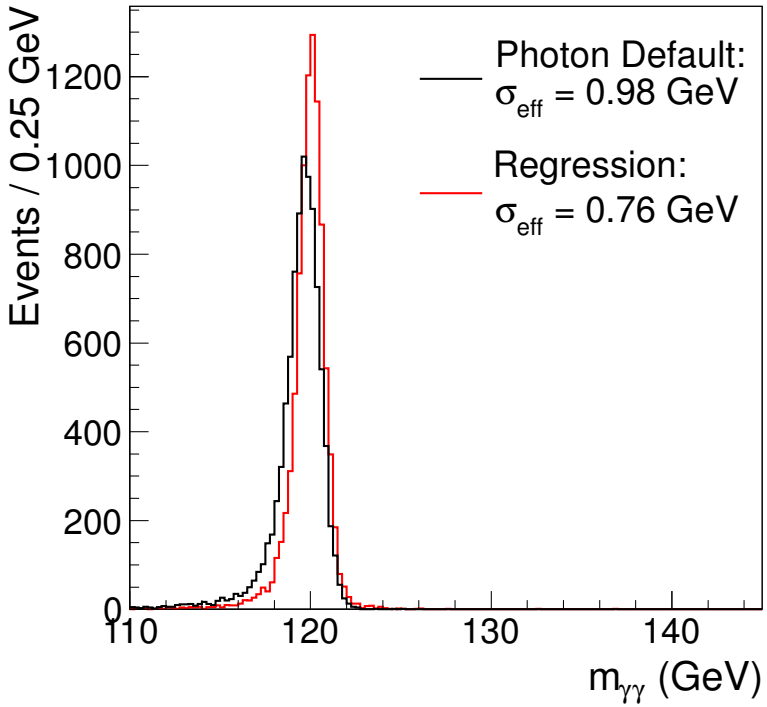
Semi-parametric Regression

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

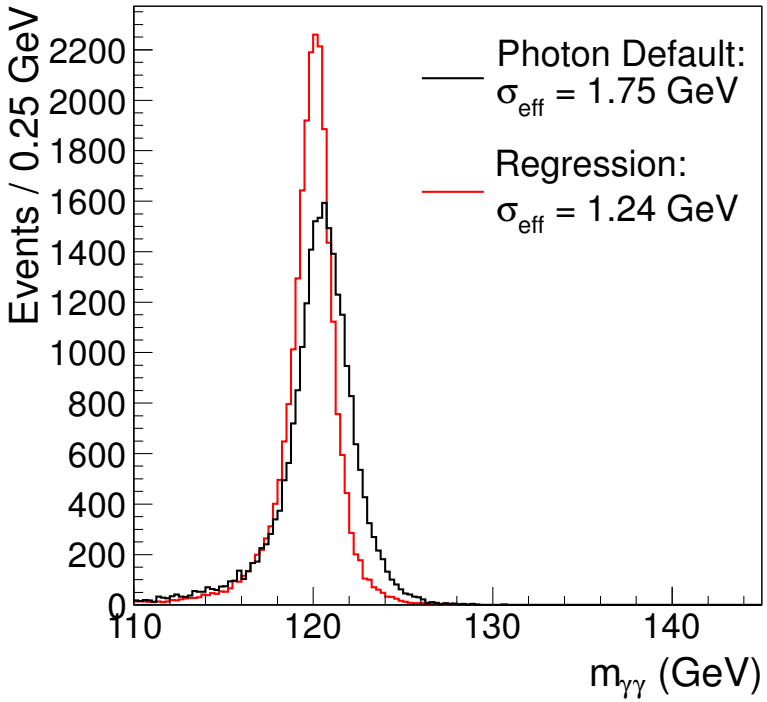
$$L = -\sum_{MCPhotons} \ln p(y|\bar{x})$$

- 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



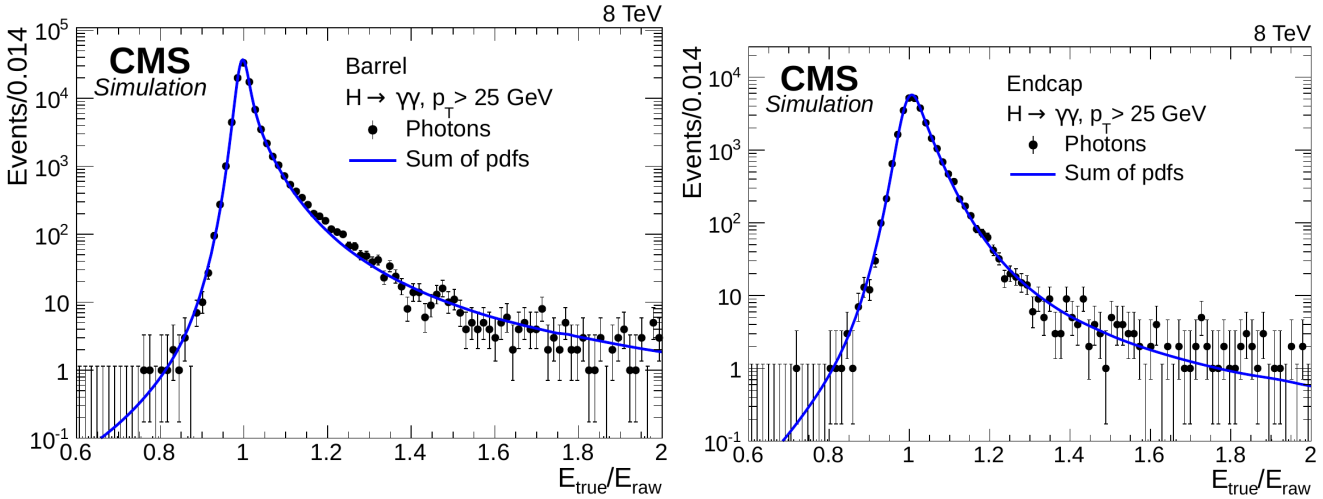
(e) Barrel Unconverted



(f) Barrel 1/2 Converted

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

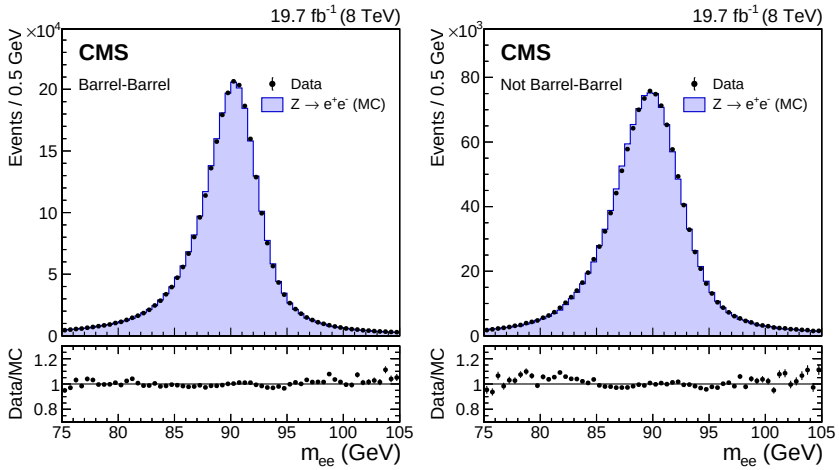
Energy Regression: Predicted Response Distribution



- 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 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)



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

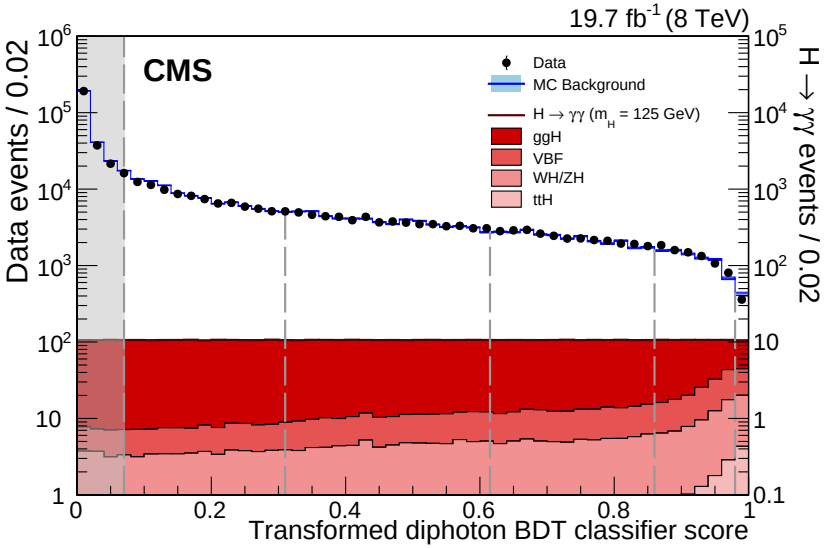
- Input variables cover kinematics (sans mass), per-event resolution and vertex probability, and photon ID
- Input Variables:
  1. $p_T^1/m_{\gamma\gamma}$
  2. $p_T^2/m_{\gamma\gamma}$
  3. η_1
  4. η_2
  5. $\cos\Delta\phi_{\gamma\gamma}$
  6. $\sigma_m/m_{\gamma\gamma}$ (Right Vtx Hypothesis)
  7. $\sigma_m/m_{\gamma\gamma}$ (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

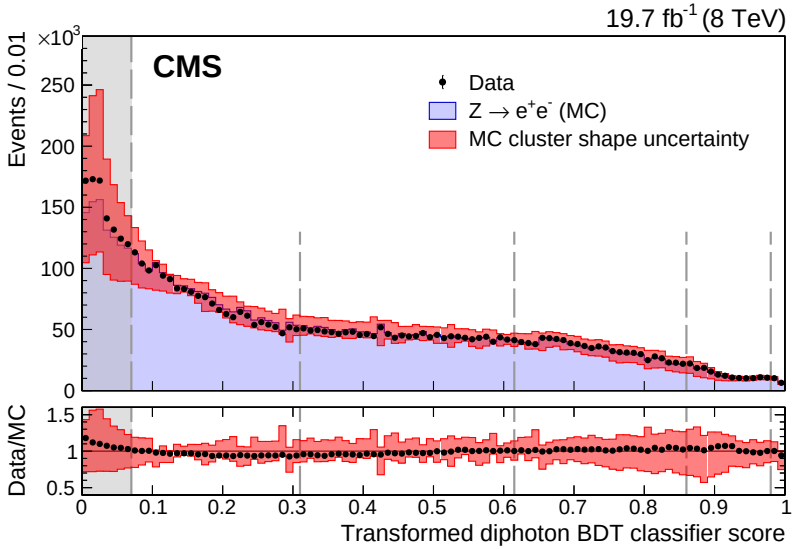
- Since input variables are mass-independent, MVA is not sensitive to mass resolution (since inclusive S/B in full mass range does not change with resolution)
- Correct this by weighting the signal events during training by 1/resolution, taking into account right and wrong primary vertex hypotheses weighted 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 $z \rightarrow ee$ events by inverting electron veto in the pre-selection
- Analysis does not rely on MVA shape of Monte Carlo background

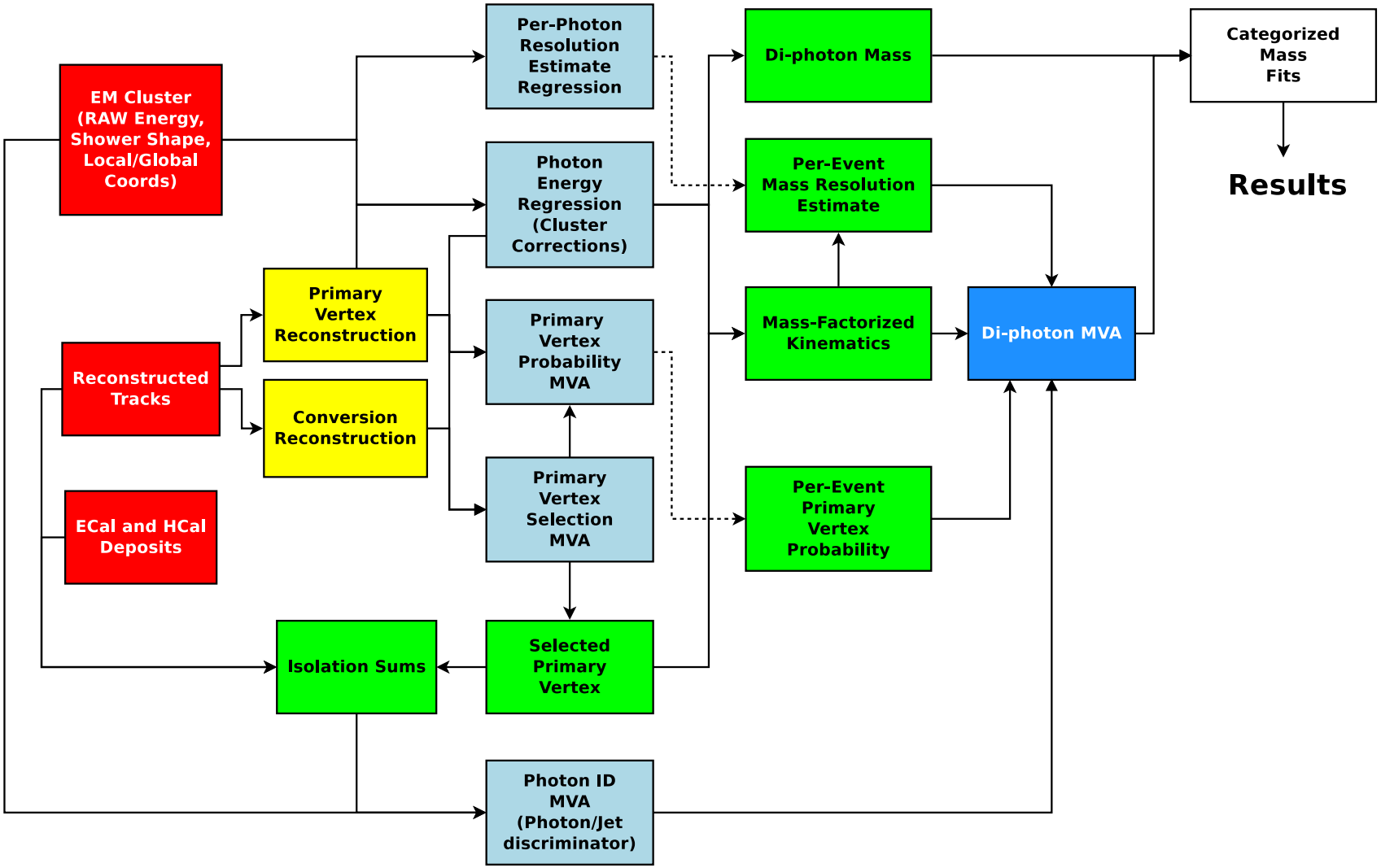


(g) Full Selection



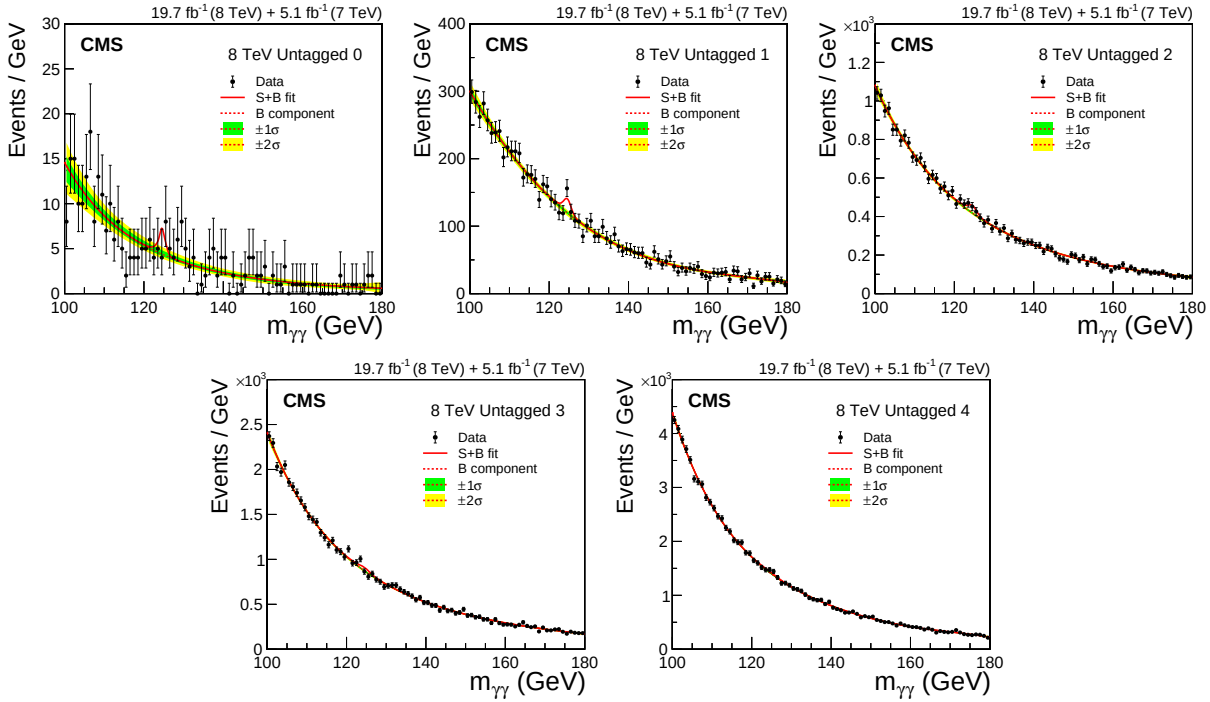
(h) Inverted e-Veto

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



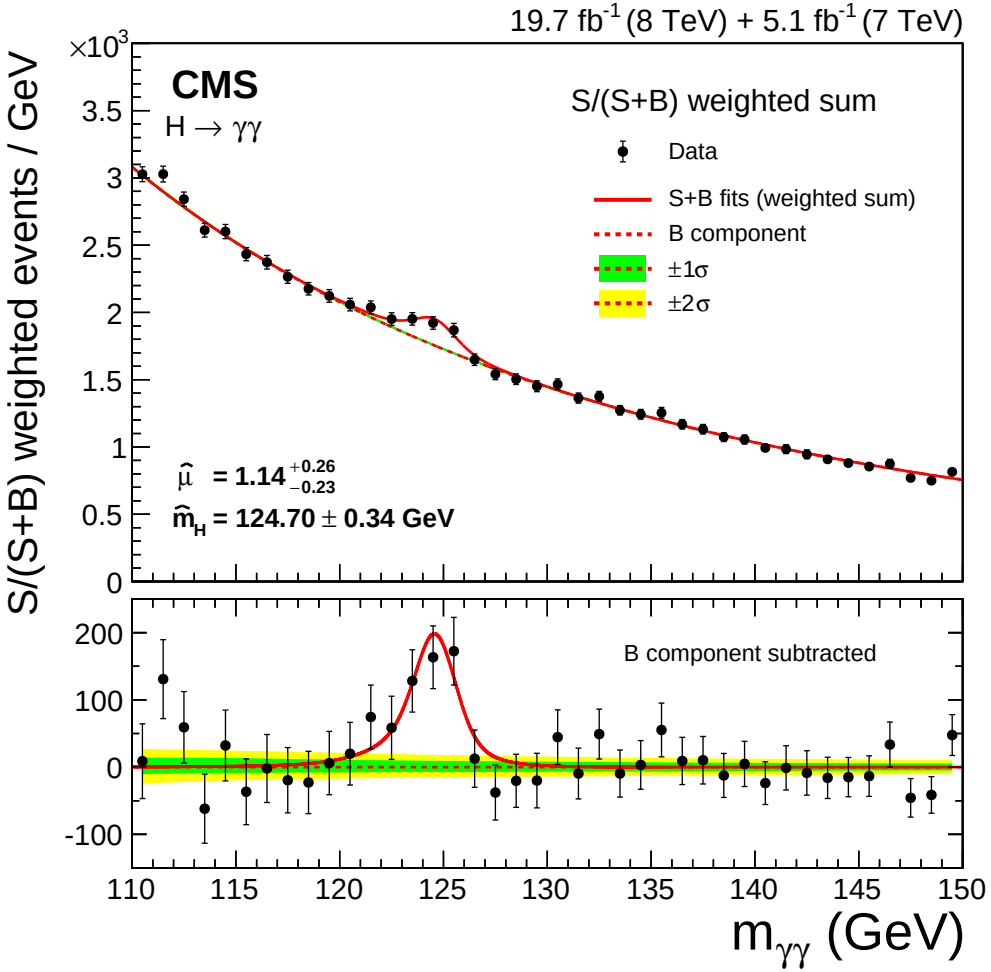
- Strategy: Process available information into quantities with straightforward physical interpretations in order to combine per-event knowledge of expected mass resolution and S/B into a single "Diphoton MVA" variable

S+B Fits - 8 TeV



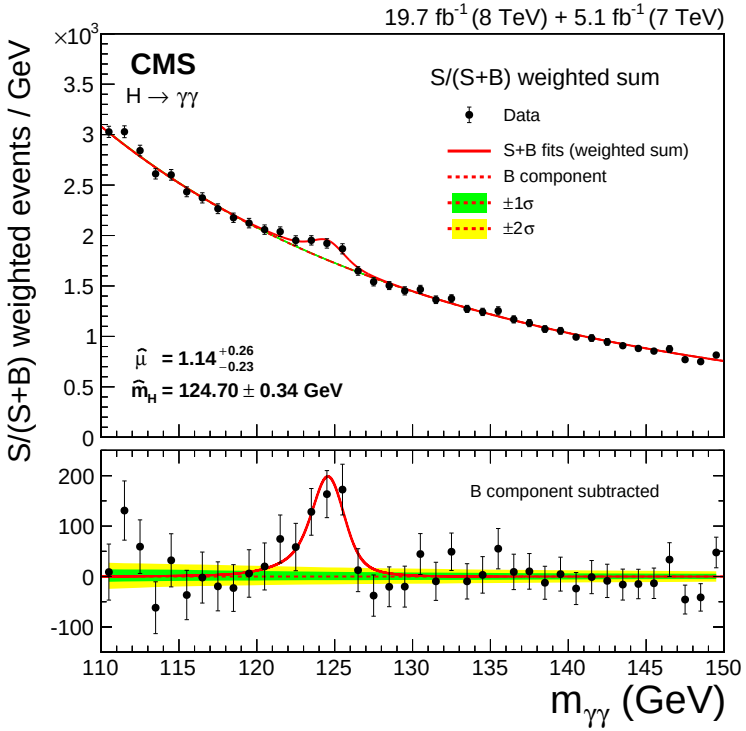
- Plus 20 more distributions for exclusive-tagged modes and 7 TeV

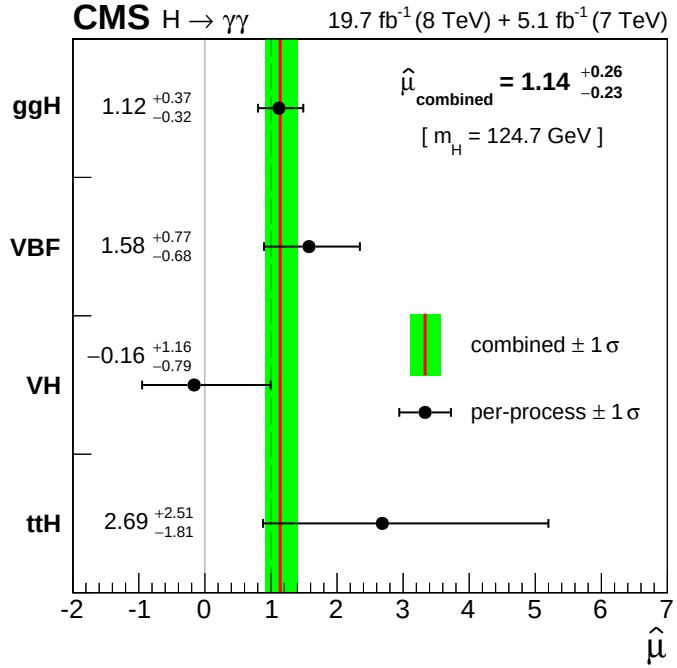
S+B Fit - Weighted Combination



- Results extracted from simultaneous fit to 25 event classes, but combined mass spectrum useful for visualisation
- Combination of all 25 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

Higgs→ $\gamma\gamma$ Results



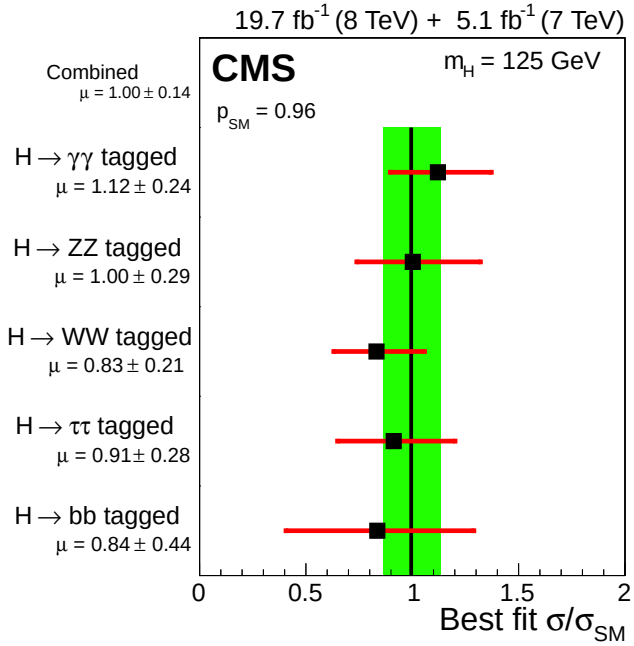


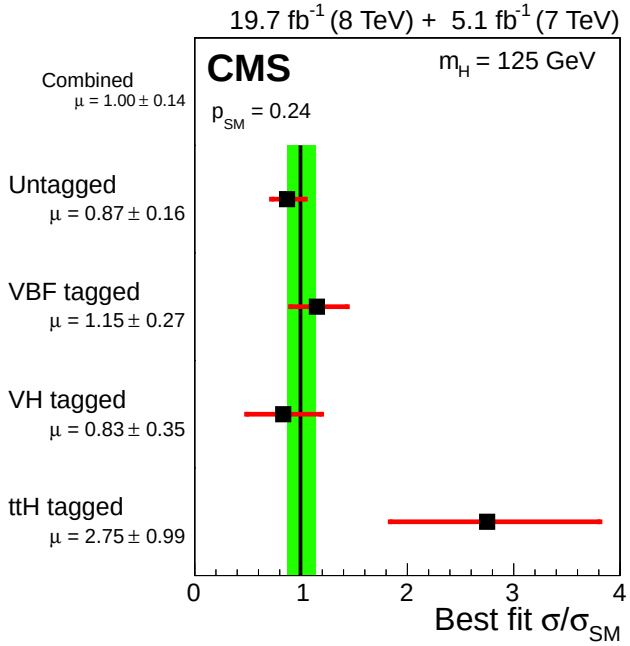
- Overall $\sigma/\sigma_{SM} = 1.14 \pm 0.21(\text{stat.})^{+0.09}_{-0.05}(\text{syst.})^{+0.13}_{-0.09}(\text{th.})$
  5.7σ observed significance (5.2σ expected)
- $m_H = 124.70 \pm 0.31(\text{stat.}) \pm 0.15(\text{syst.})$ GeV

Conclusions

- Machine learning techniques extensively used in Higgs analyses in CMS (and ATLAS) at all levels of the analysis
- Often analysis design and design of Machine Learning models/tools are deeply intertwined
- Have discussed a (non-exhaustive) set of examples which illustrates some of the more interesting use cases/issues

CMS Higgs Results: Run 1





- Overall $\sigma/\sigma_{SM} = 1.00 \pm 0.14$
- Combining $H \to ZZ, \gamma\gamma$: $m_H = 125.02^{+0.26}_{-0.27}(\text{stat.})^{+0.14}_{-0.15}(\text{syst.})$
- Measured signal strengths broadly consistent with SM expectations
- Tests of angular distributions indicate particle is indeed a scalar

Prospects for Run 2

- Gluon fusion Higgs cross section increases by ~ 2.3 from 8 TeV to 13 TeV
- $tt + H$ cross section increases by ~ 4
- Background cross sections of course also increase
- Up to ~ 100 fb^{-1} expected for Run 2
- "signal strength" measured so far: model-dependent cross section extrapolated to full phase space
- Run 2: **Fiducial Cross Sections, Differential Cross Sections**
- Complete the transition from discovery to precision physics
- Maintain object and analysis performance with 25ns bunch spacing