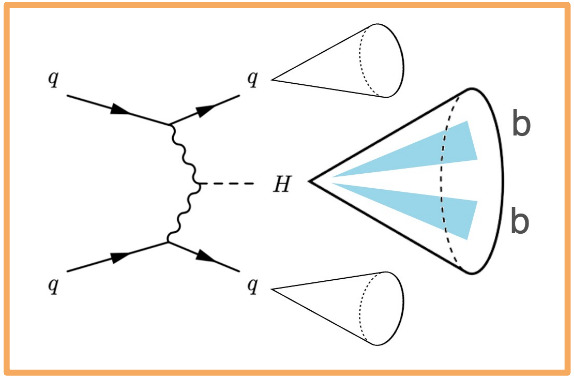
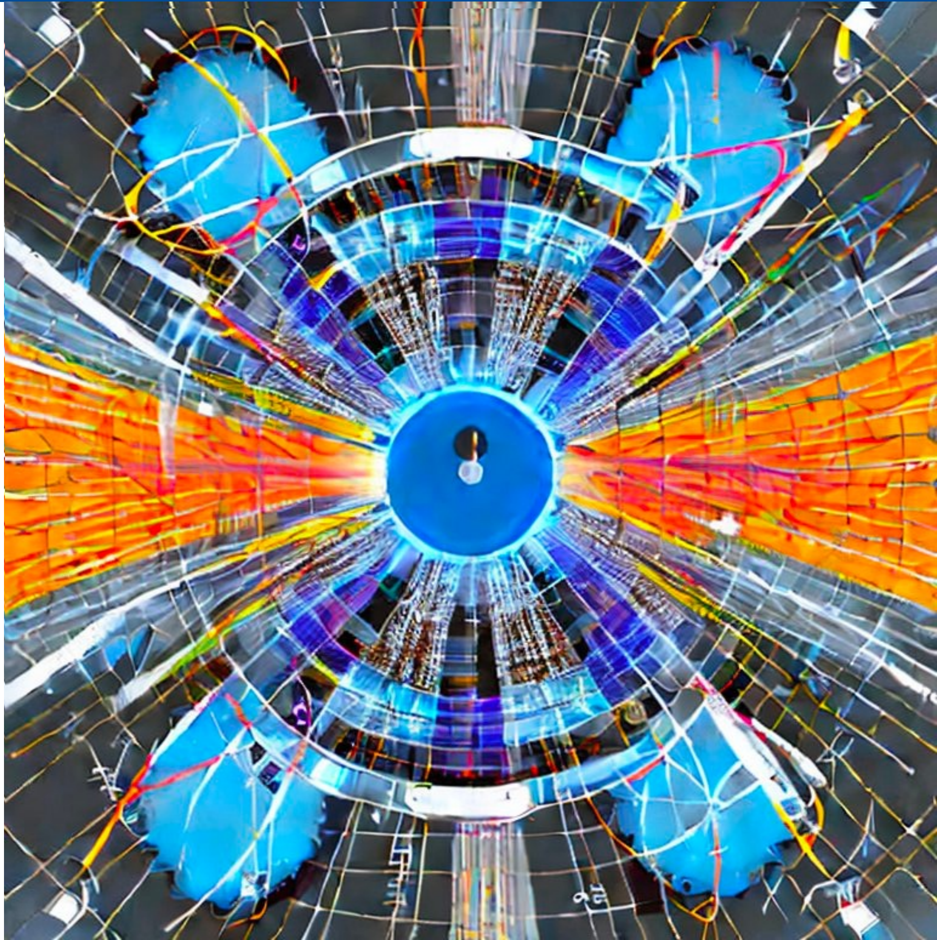
Boosted Higgs production via vector boson fusion with the CMS experiment

Jennet Dickinson
on behalf of the CMS Collaboration

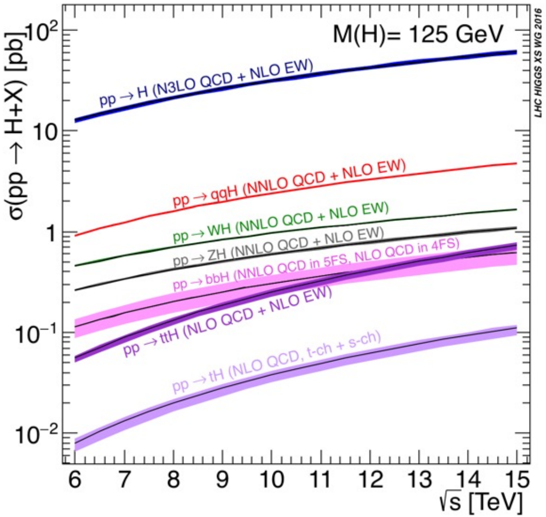
20th Workshop of the LHC Higgs Working Group





Higgs production in pp collisions

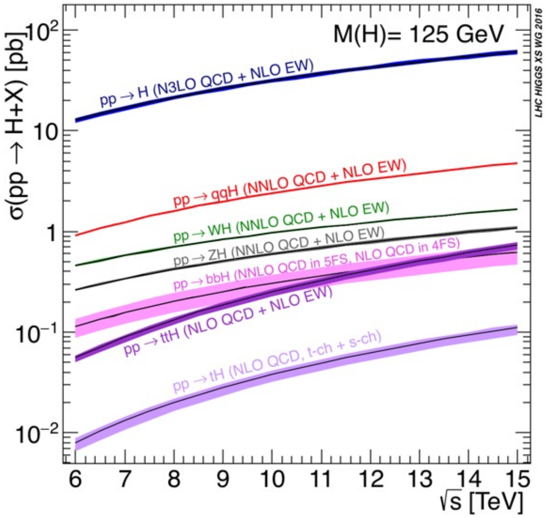
- Gluon fusion accounts for 90% of the Higgs boson cross section at 13 TeV

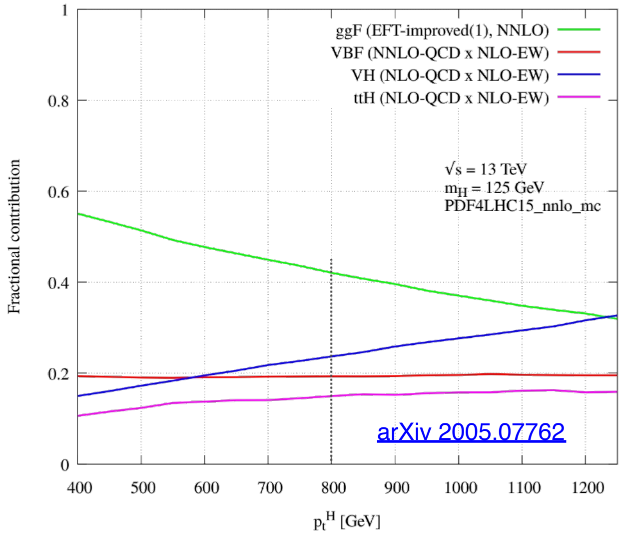


Higgs production in pp collisions

- Gluon fusion accounts for 90% of the Higgs boson cross section at 13 TeV …

if you measure inclusively in p_T





[arXiv 2005.07762](https://arxiv.org/abs/2005.07762)

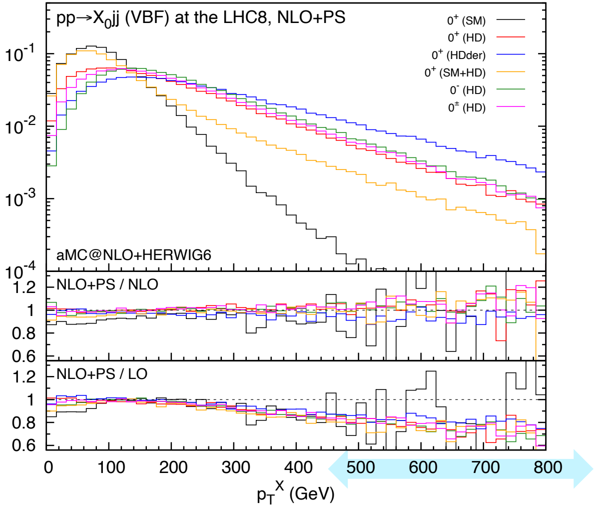
Why VBF at high p_T?

- ggF becomes less dominant at high p_T

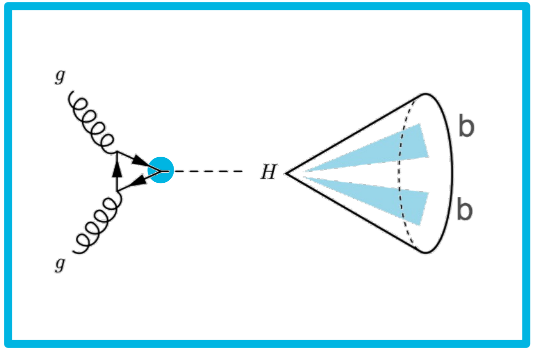
  And we have precise predictions for other production modes (link)

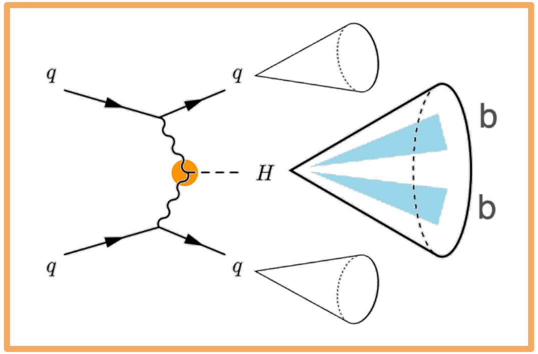
- High p_T tails are sensitive to new physics at high energy scales

  Different production modes probe different BSM operators



arXiv 1311.1829

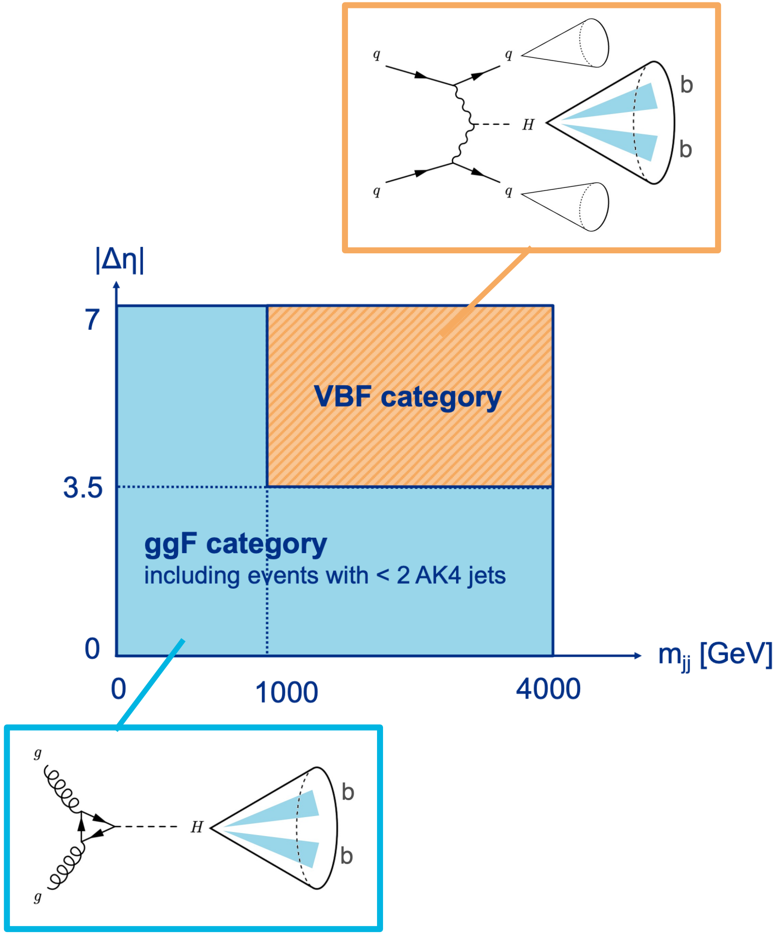




Fermilab

Analysis overview

- Apply selection targeting boosted Higgs candidates, rejecting backgrounds
- Add tailored cuts to target the VBF process
  - Define orthogonal ggF and VBF categories
- Divide into b-tag passing and failing regions using the **DeepDoubleB** (DDB) tagger
  - Use DDB fail for data-driven QCD background estimate
- Fit to the **soft drop** mass of the Higgs candidate jet in both b-tag regions
  - Simultaneously extract signal strength for ggF and VBF



Event selection

- Start with events passing ≥1 trigger selecting for H_T, jet p_T, jet mass, b-tagging

  Fully efficient for leading jet $p_T > 500$ GeV

- Require at least one **large radius jet**

  AK8 jet with $p_T > 450$ GeV, $|\eta| < 2.5$

  Must have **two-prong substructure:** N_2 variable decorrelated with mass $N_2^{DDT} < 0$

  If more than one jet qualifies, select the one with **highest DDB score**

- Lepton veto

- Top veto: MET < 140 GeV, no b-jet in the hemisphere opposite candidate jet

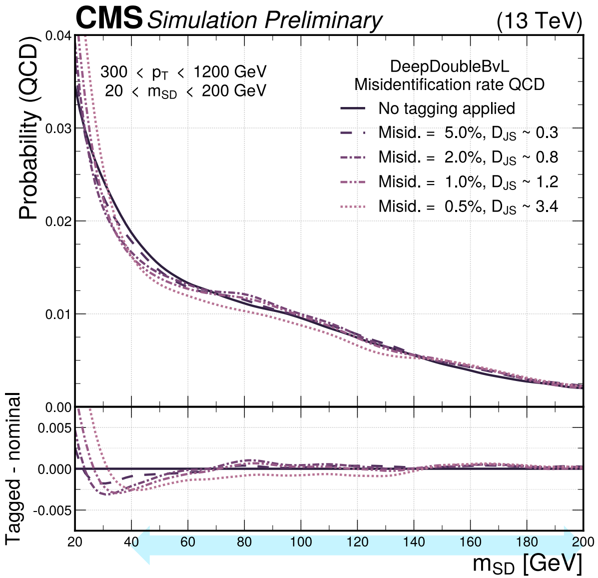
- If event has ≥ 2 more thin jets with **$\Delta\eta_{jj} > 3.5$ and $m_{jj} > 1$ TeV** ⟶ VBF category

  Otherwise ⟶ ggF category

DeepDoubleBvL-v2 tagger (DDB)

Congqiao's talk
CMS-DP-2022-041

- **CNN architecture** trained on simulation to separate QCD and scalar X ⟶ bb decays
  - Signal generated for m_X from 20-200 GeV
- Input features include:
  - Particle flow candidates (up to 40 charged, 60 neutral)
  - Secondary vertices
  - High-level jet variables
- DDB threshold chosen to optimize VBF sensitivity
  - Events below DDB threshold (DDB fail) are used to estimate QCD background
- **Tagger efficiency** is constrained in-situ by the Z ⟶ bb peak
  - One of the dominant experimental systematics



Fermilab

Signal Monte Carlo

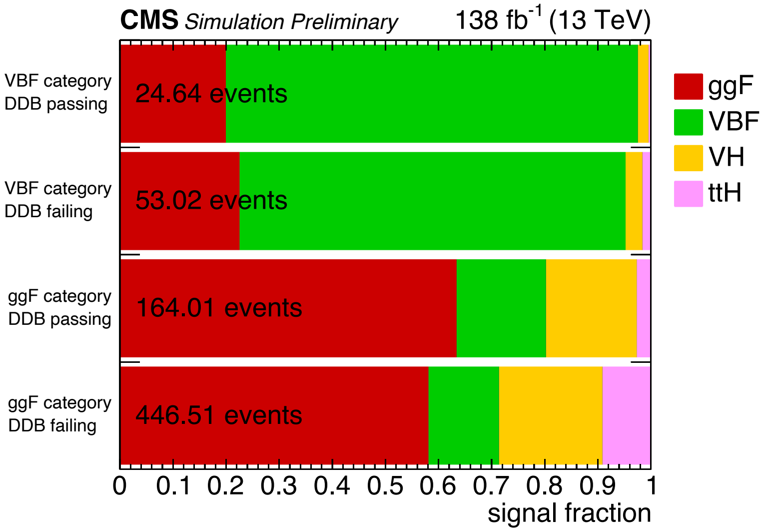
- **ggF**: POWHEG HJMINLO

  Good agreement with LHC XS WG recommendations

- **VBF**: POWHEG re-weighted for EW and N^3LO corrections

  Good agreement with LHC XS WG recommendations

- **Other Higgs** (WH, ZH, ttH, ggZH): POWHEG reweighted for EW corrections



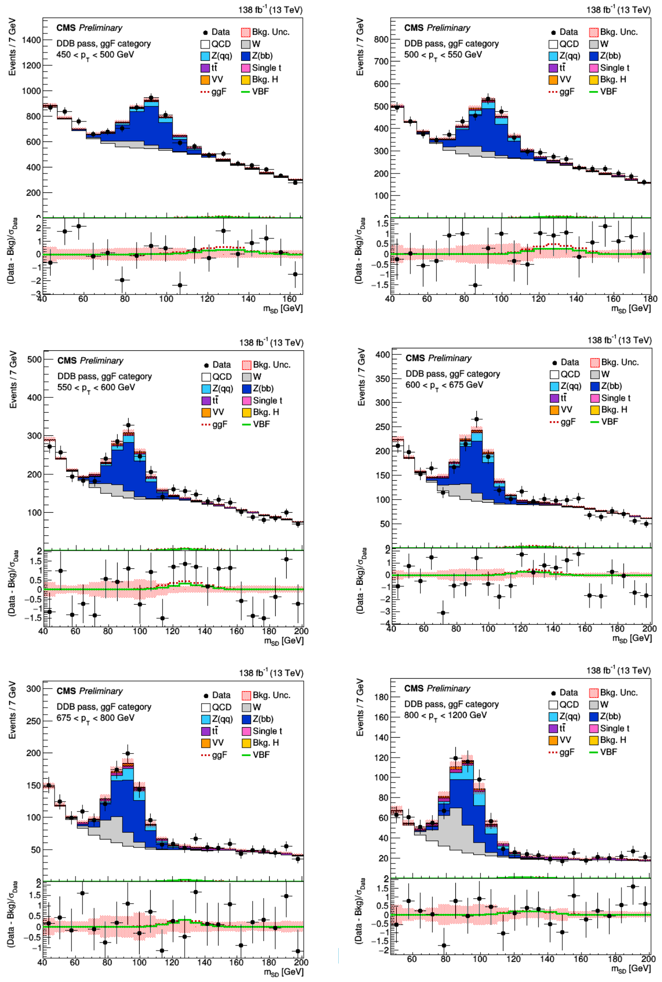
- Renormalization/factorization scale, PDF and parton shower uncertainties included on all Higgs samples

  Scale uncertainty on ggF (~20%) and VBF (~5%) is the dominant theory systematic

Differential bins

- Combining multiple bins with different signal purity gives better sensitivity
- **ggF category**: 6 bins in Higgs candidate p_T

  [450, 500, 550, 600, 675, 800, 1200] GeV



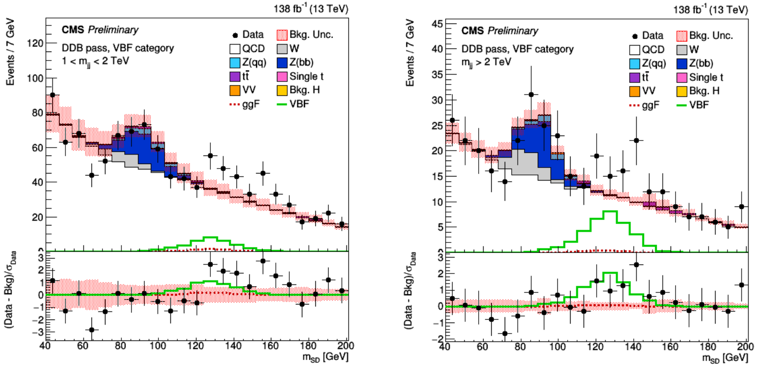
Differential bins

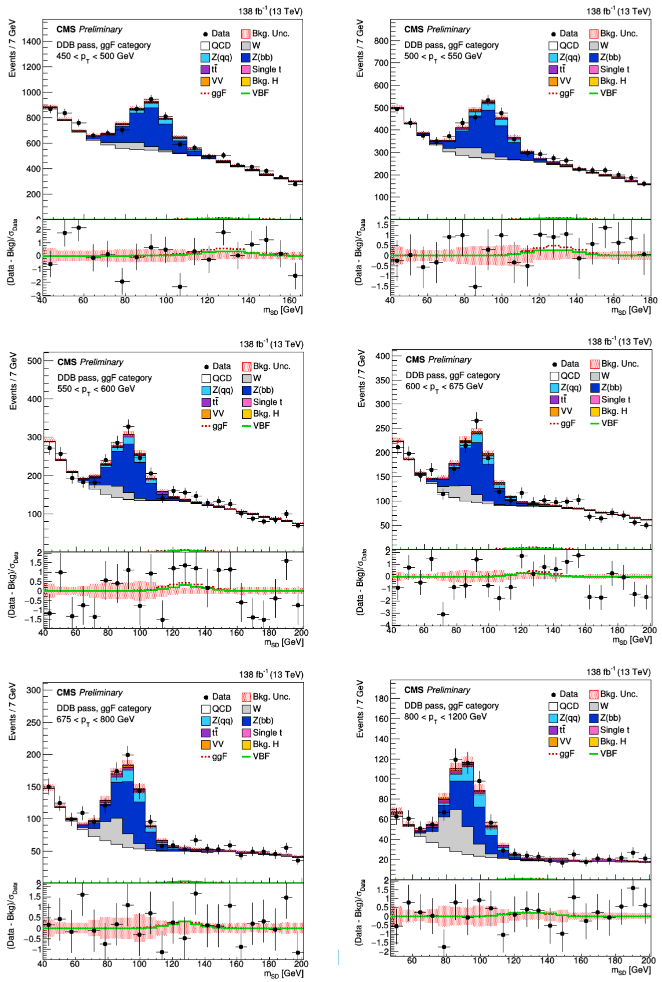
- Combining multiple bins with different signal purity gives better sensitivity
- **ggF category**: 6 bins in Higgs candidate p_T

  [450, 500, 550, 600, 675, 800, 1200] GeV

- **VBF category**: 2 bins in the invariant mass of the forward jets, m_{jj}

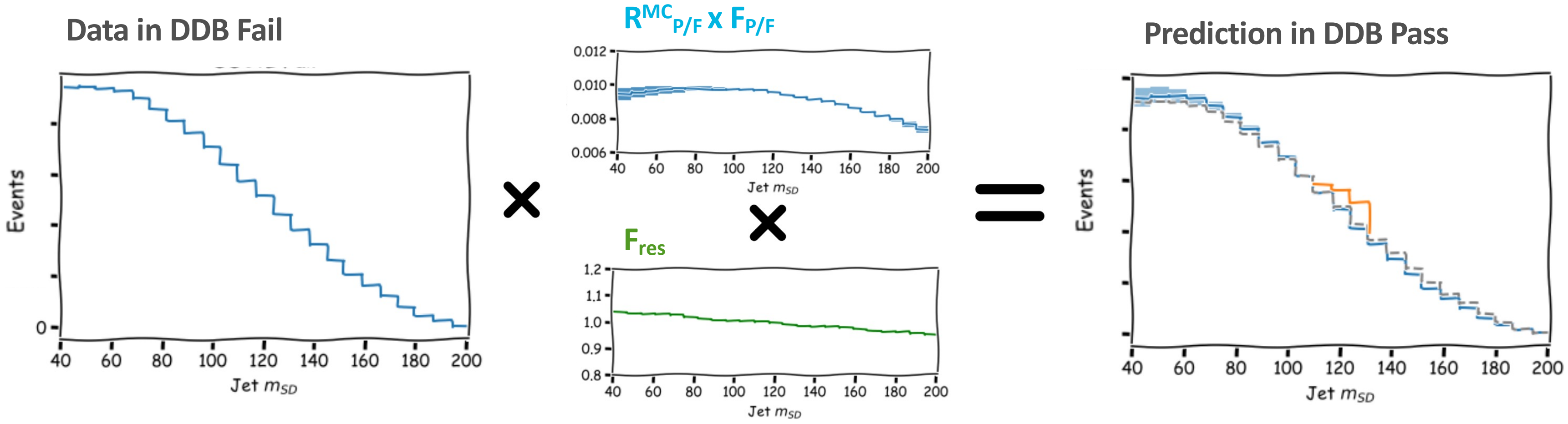
  [1000, 2000, ∞] GeV





QCD background estimation

- Goal: predict the QCD distribution in the DDB pass region
- Use data in the DDB fail region as a starting point and apply two polynomial **transfer factors**



Fermilab

First transfer factor: $F_{P/F}$

- Accounts for differences in the m_{SD} shape in the DDB pass / fail regions due to tagger selection
- Coefficients extracted from a standalone fit to the DDB pass / fail ratio in QCD MC only

Overall normalization is treated as a separate factor, $R_{P/F}^{MC}$

Uncertainties are propagated to the final fit

$$\frac{N_P^{\mathrm{MC},i}}{N_F^{\mathrm{MC},i}} = R_{\mathrm{P/F}}^{\mathrm{MC}}\, F_{\mathrm{P/F}}^{i}$$

First transfer factor: $F_{P/F}$

- Accounts for differences in the m_{SD} shape in the DDB pass / fail regions due to tagger selection
- Coefficients extracted from a standalone fit to the DDB pass / fail ratio in QCD MC only

Overall normalization is treated as a separate factor, $R_{P/F}^{MC}$

Uncertainties are propagated to the final fit

$$\frac{N_P^{\mathrm{MC},i}}{N_F^{\mathrm{MC},i}} = R^{\mathrm{MC}}_{\mathrm{P/F}}\, F^i_{\mathrm{P/F}}$$

Second transfer factor: F_{res}

- Accounts for any additional differences the m_{SD} shape in the DDB pass / fail regions
- Coefficients extracted from simultaneous fit to DDB pass and fail regions

Uncertainty on fitted polynomial coefficients is a dominant systematic

$$N_P^i = R^{\mathrm{MC}}_{\mathrm{P/F}}\, F^i_{\mathrm{P/F}}\, F^i_{\mathrm{res}}\, N_F^{\mathrm{data},i}$$

Transfer factor polynomials

- **ggF category**

  1 x 2D Bernstein polynomial in jet p_T and $\rho = \ln(m_{SD}^2/p_T^2)$

- **VBF category**

  2 x 1D Bernstein polynomial in jet ρ only (one per m_{jj} bin)

$$F_{P/F}(p_T, \rho) = \sum_{k=0}^{n_\rho} \sum_{l=0}^{n_{p_T}} a_{k,l} \left[b_{k,n_\rho}(\rho) b_{l,n_{p_T}}(p_T) \right]$$

$$b_{\nu,n} = \binom{n}{\nu} x^\nu (1-x)^{n-\nu}$$

- Determining polynomial order

  Start with a low order polynomial, which is nested within higher order polynomials

  Systematically increase polynomial order until the goodness of fit no longer increases significantly

- Independent fits performed per category, per data-taking period

Control regions

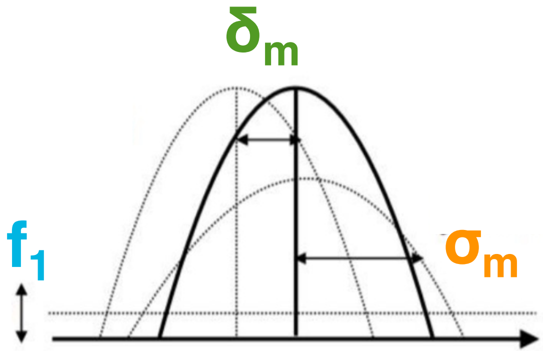
- **Top control region**: derive normalization and DDB efficiency on top background processes from data

  Nominal selection, but 0 $\mu \longrightarrow$ 1 loose μ and require an additional b-jet

  Treated as a single bin counting experiment per data taking period in the final fit

Control regions

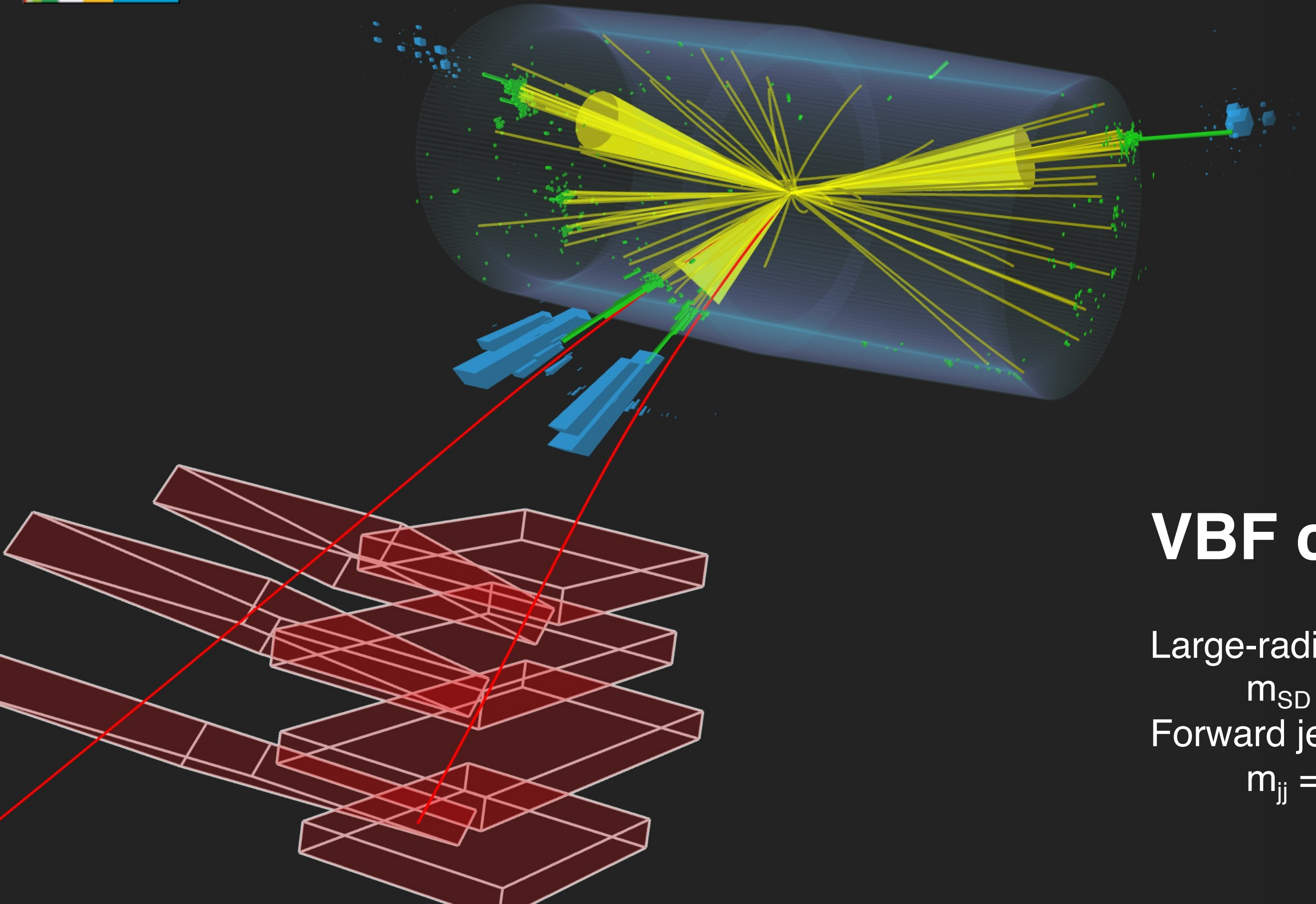
- **Top control region**: derive normalization and DDB efficiency on top background processes from data
  - Nominal selection, but 0 μ ⟶ 1 loose μ and require an additional b-jet
  - Treated as a single bin counting experiment per data taking period in the final fit
- **W-tag control region**: derive scale factors for substructure selection, jet mass scale & resolution
  - Require μ and MET ⟶ reco W = (μ+MET) with $p_T > 200$ GeV
  - Split each MC sample into truth W-matched and unmatched
  - Fit regions $N_2^{DDT} > 0$ and < 0 simultaneously for **substructure scale factor**, **jet mass resolution** and **jet mass scale**



CMS Experiment at the LHC, CERN

Data recorded: 2018-Sep-29 22:54:37.754176 GMT

Run / Event / LS: 323727 / 488169591 / 262



VBF candidate event

Large-radius jet:

m_{SD} = 125.2 GeV, p_T = 613.5 GeV

Forward jets:

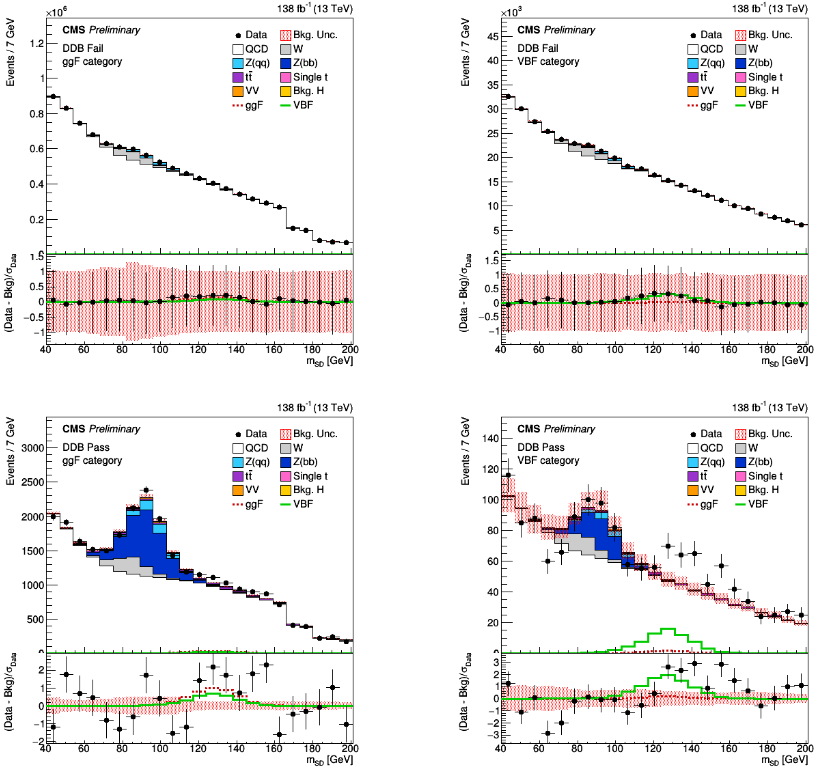
m_{jj} = 2220.7 GeV, $\Delta\eta_{jj}$ = 4.2

Results

- Observed significance is calculated with other process freely floating
- VBF: 3.0σ (0.9σ expected)
- ggF: 1.2σ (0.9σ expected)

	Lumi [fb^{-1}]	μ_{VBF}	μ_{ggF}
Early 2016	19.5	$2.9^{+5.8}_{-4.5}$	$4.3^{+5.5}_{-5.4}$
Late 2016	16.8	$5.8^{+6.3}_{-4.7}$	$-0.9^{+4.7}_{-5.1}$
2017	41.5	$-0.7^{+2.8}_{-2.6}$	$6.7^{+4.0}_{-3.1}$
2018	59.8	$10.0^{+4.4}_{-3.4}$	$-0.6^{+2.8}_{-3.1}$
Combined	137.6	$5.0^{+2.1}_{-1.8}$	$2.1^{+1.9}_{-1.7}$

ggF category　　　VBF category

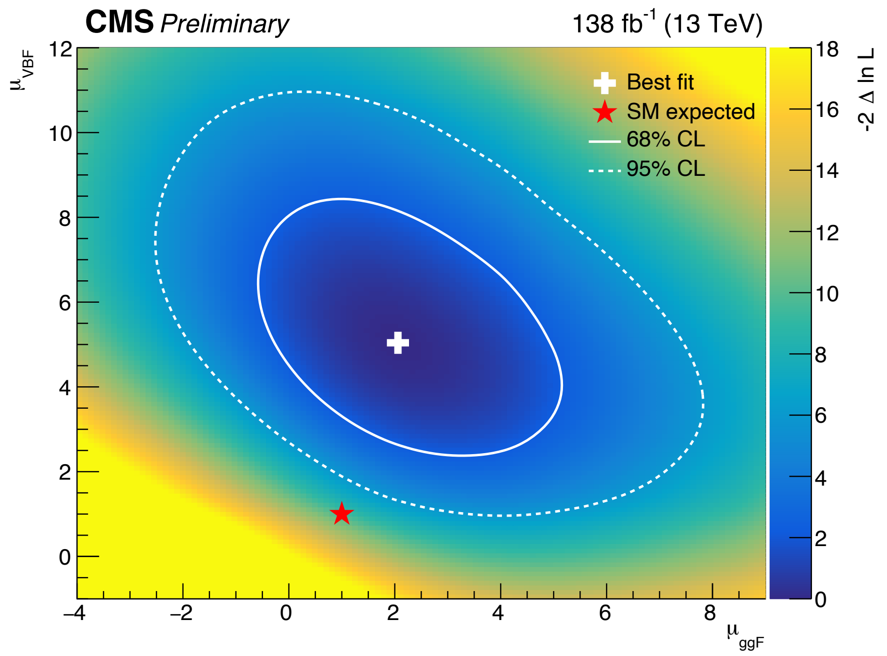


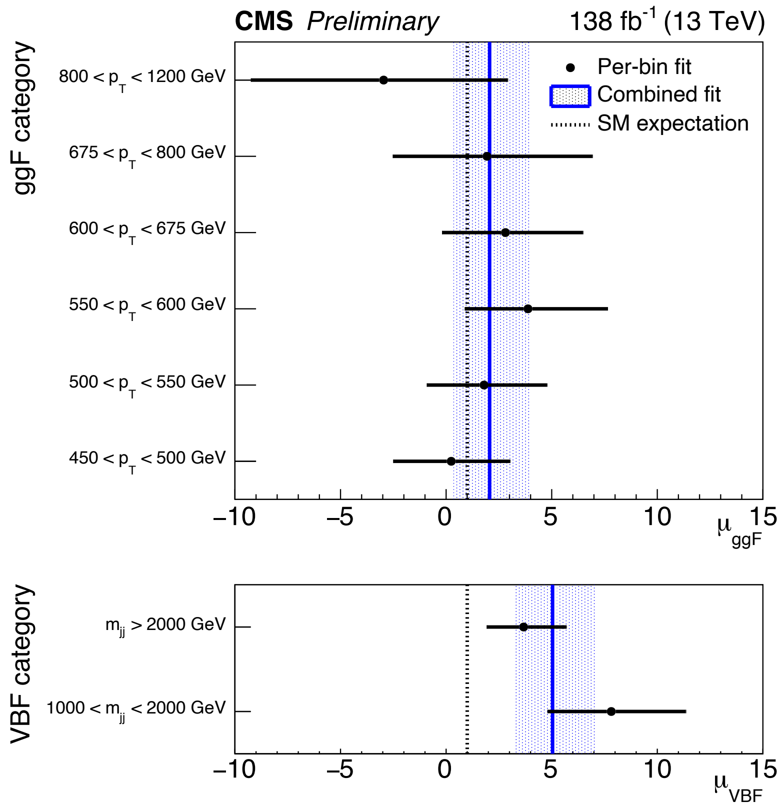
Fermilab

Results

- Best fit differs from SM by 2.6σ

  and from (0,0) by 3.9σ





Fermilab

Summary

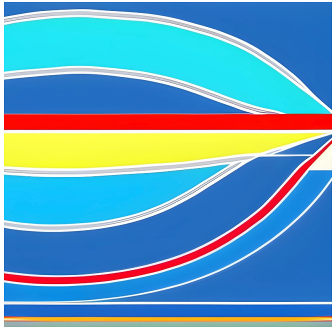
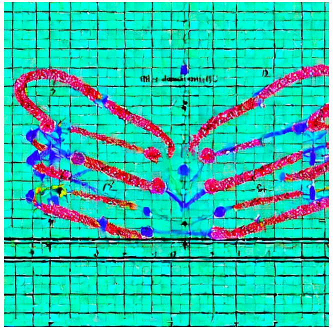
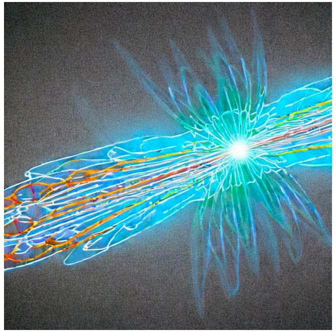
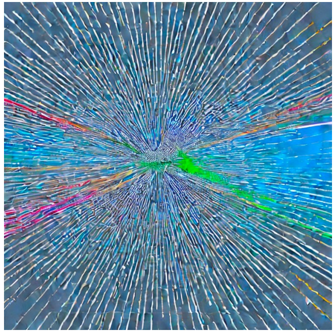
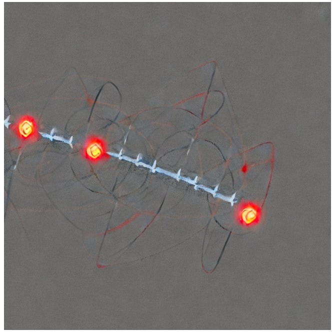
- We have presented the first search for **VBF in the boosted H(bb) channel**
- Simultaneous measurement of ggF and VBF signals is performed

$\mu_{VBF} = 5.0^{+2.1}_{-1.8}$

$\mu_{ggF} = 2.1^{+1.9}_{-1.7}$

- Observed results differ from SM expectation by **2.6σ**
- Further details in HIG-21-020

Additional material

Background simulation

- **V+jets**:

  Madgraph LO corrected to NLO gen-level p_T spectrum

  NNLO QCD, EW corrections applied following "mono-jet" prescription
- **Electroweak V**: Madgraph LO
- **Diboson**: Pythia LO corrected to NNLO with MCFC
- **ttbar, single top**: POWHEG NLO
- **QCD**: p_T sliced Pythia8

  Estimation mostly from data

Fermilab

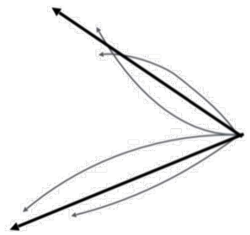
Substructure selection

- Variable N_2 (N_2^1) identifies two-prong jets using IRC safe energy correlation functions

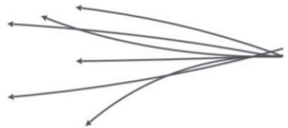
$$e_2^\beta = \sum_{1\le i<j\le n_J} z_i z_j \Delta R_{ij}^\beta$$

$$e_3^\beta = \sum_{1\le i<j<k\le n_J} z_i z_j z_k \Delta R_{ij}^\beta \Delta R_{ik}^\beta \Delta R_{jk}^\beta$$

$$\longrightarrow \quad N_2^\beta = \frac{{}_2e_3^\beta}{({}_1e_2^\beta)^2}$$



- Find the cut value on N_2 that has 26% efficiency on QCD MC, as a function of p_T and ρ: $c_{0.26}(p_T, \rho)$
- Resulting variable is decorrelated from jet p_T and mass



$$N_2^{1,\mathrm{DDT}} = N_2^1 - c_{0.26}(p_\mathrm{T}, \rho)\ .$$

Fermilab

W-tag control region

- Derive scale factors for substructure selection, jet mass scale & resolution

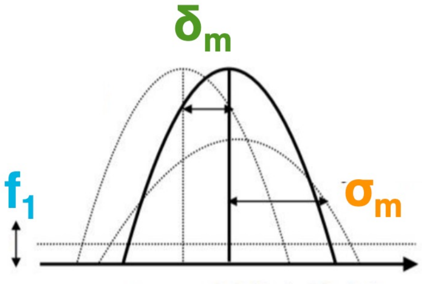
  Require μ and MET ⟶ reco W = (μ+MET) with $p_T > 200$ GeV

  Split each MC sample into truth W-matched and unmatched

  Fit regions $N_2^{DDT} > 0$ and < 0 simultaneously for **substructure scale factor**, **jet mass resolution** and **jet mass scale**

-

$$f_1 n^{\text{P-sub}}_{\text{match}}(\delta_m,\sigma_m) + \left[(1-f_1)\frac{\sum N^{\text{P-sub}}_{\text{match}}}{\sum N^{\text{F-sub}}_{\text{match}}} + 1\right] N^{\text{F-sub}}_{\text{match}}(\delta_m,\sigma_m) +$$
$$f_2 N^{\text{P-sub}}_{\text{unmatch}} + \left[(1-f_2)\frac{\sum N^{\text{P-sub}}_{\text{unmatch}}}{\sum N^{\text{F-sub}}_{\text{unmatch}}} + 1\right] N^{\text{F-sub}}_{\text{unmatch}}$$



	Substructure (f_1)	Mass scale (δ_m) [GeV]	Mass resolution (σ_m)
Early 2016	0.85 ± 0.14	-1.50 ± 0.45	0.98 ± 0.04
Late 2016	0.68 ± 0.18	$+1.13 \pm 0.41$	1.26 ± 0.04
2017	1.18 ± 0.14	$+0.49 \pm 1.16$	1.18 ± 0.08
2018	0.90 ± 0.10	-0.84 ± 0.24	1.14 ± 0.04