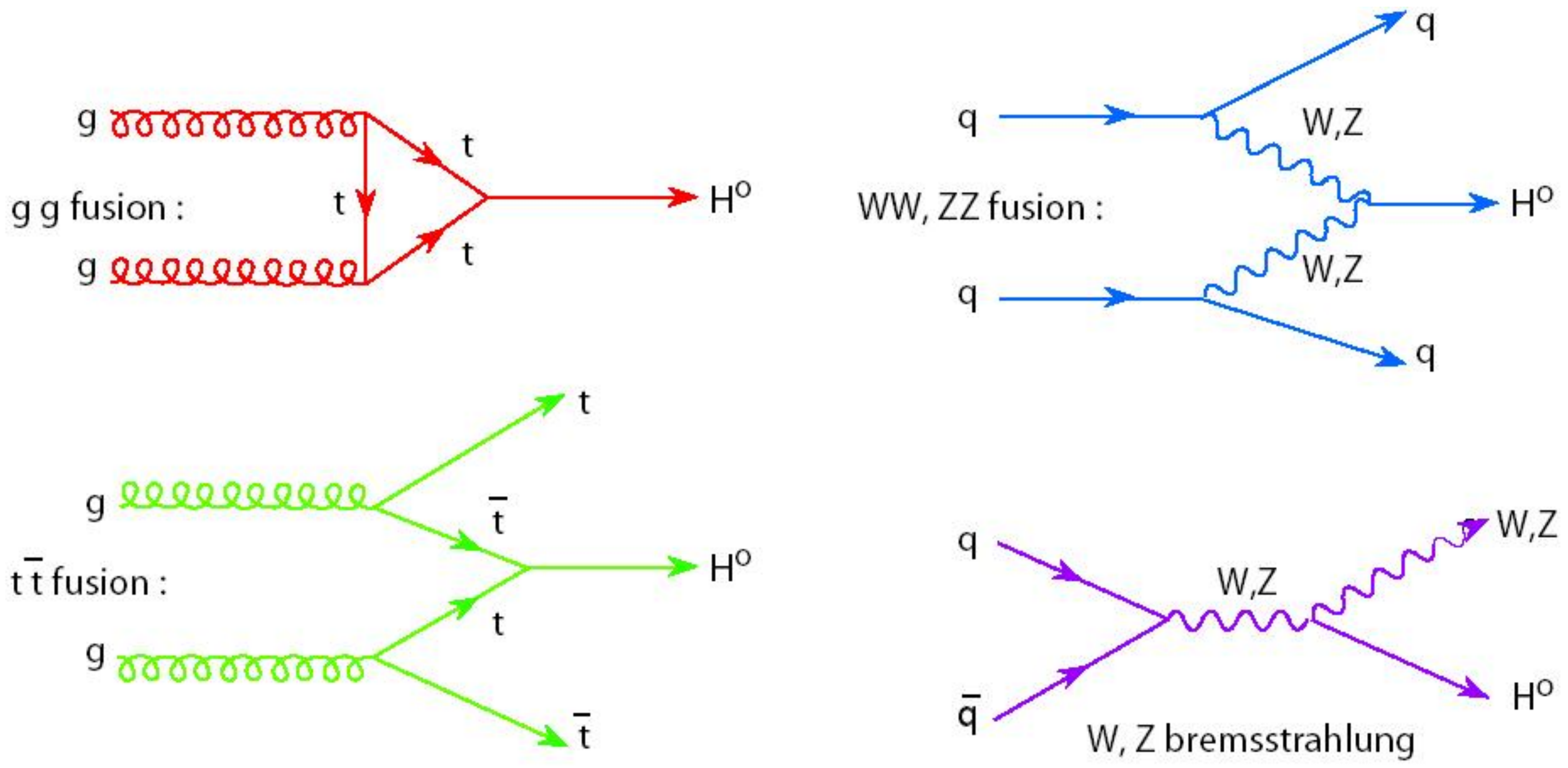
数据分析A-Z

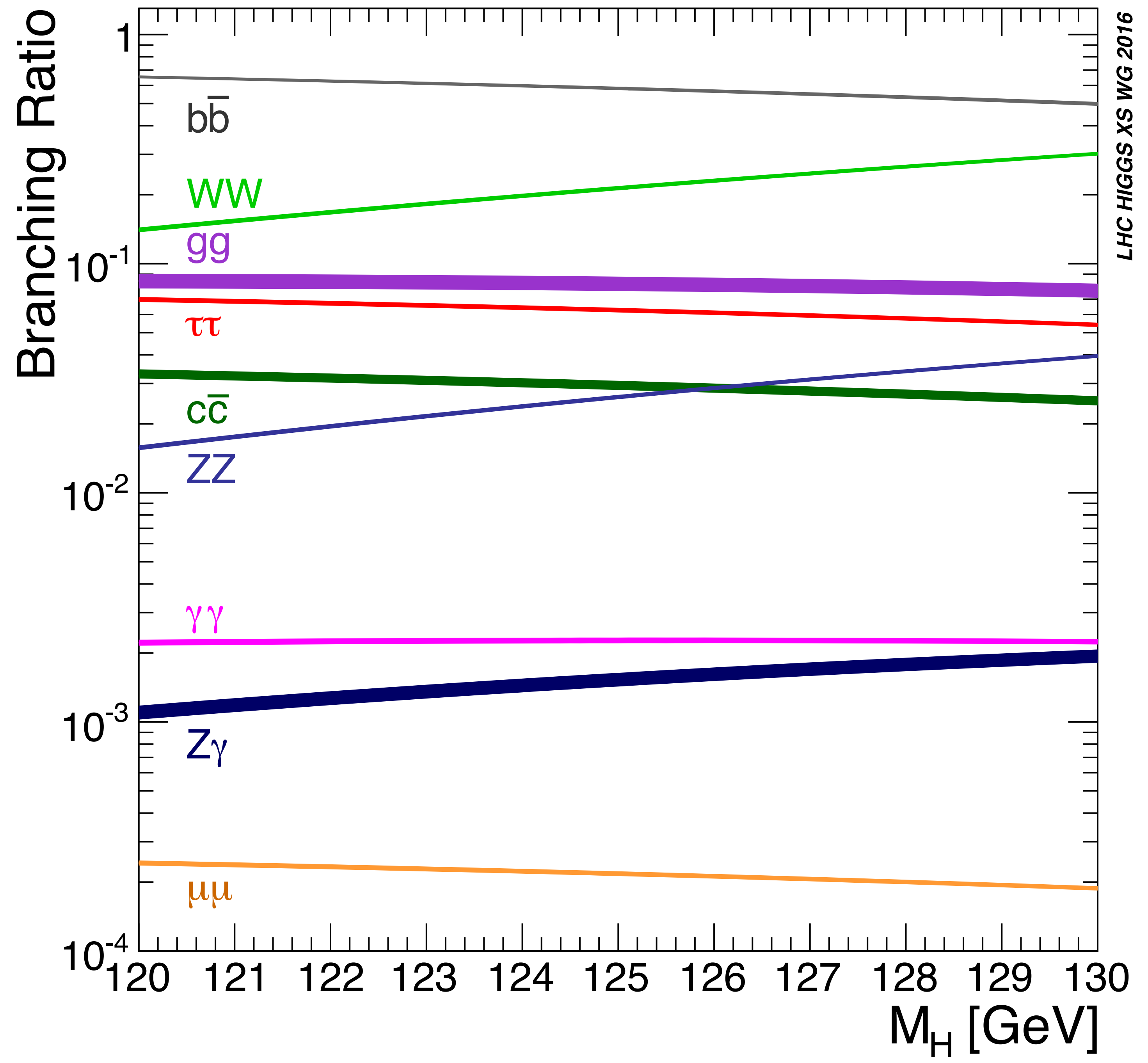
肖朦 （浙江大学）

中国CMS冬令营，2021.12.18

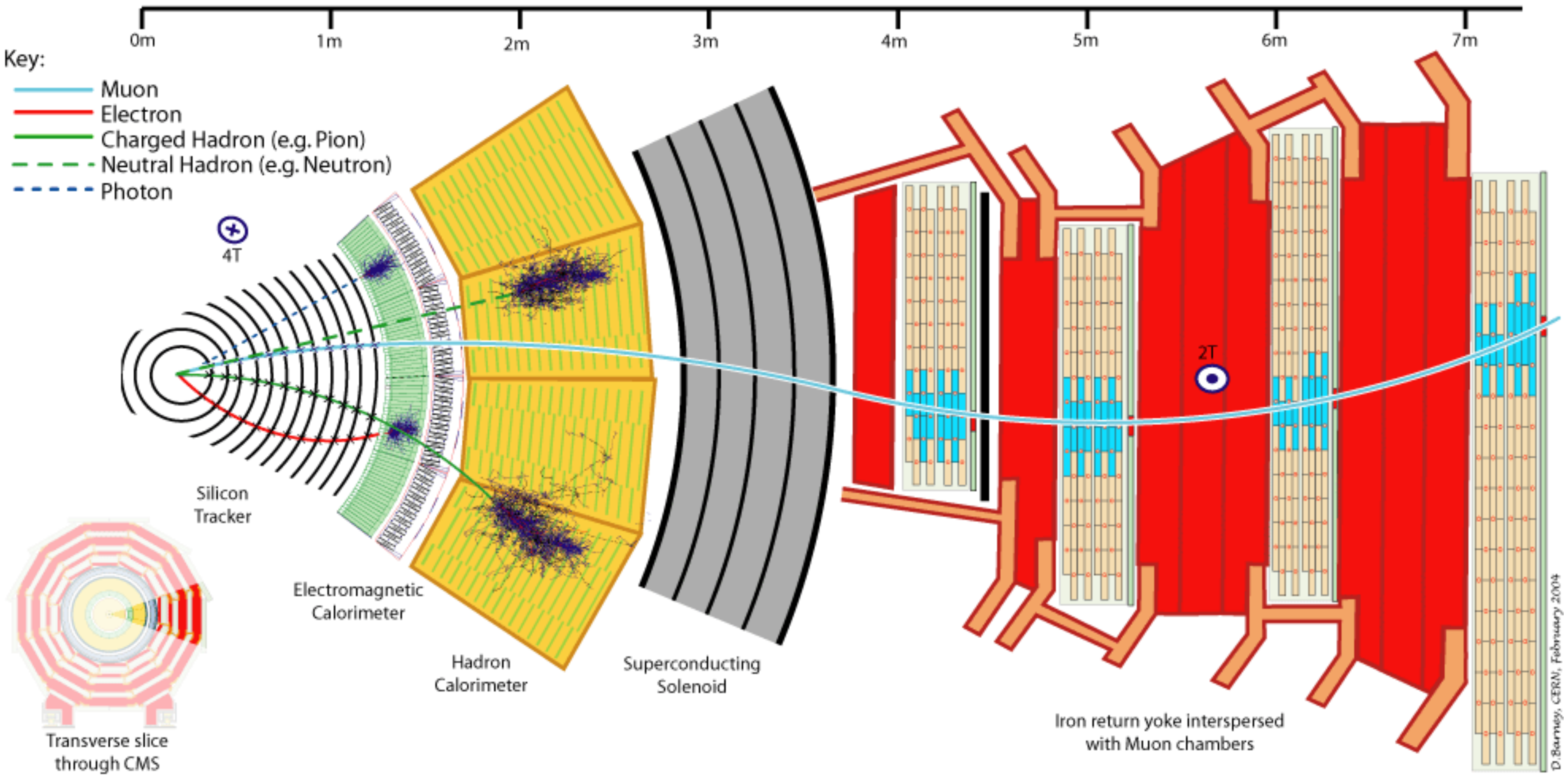
Physics analysis

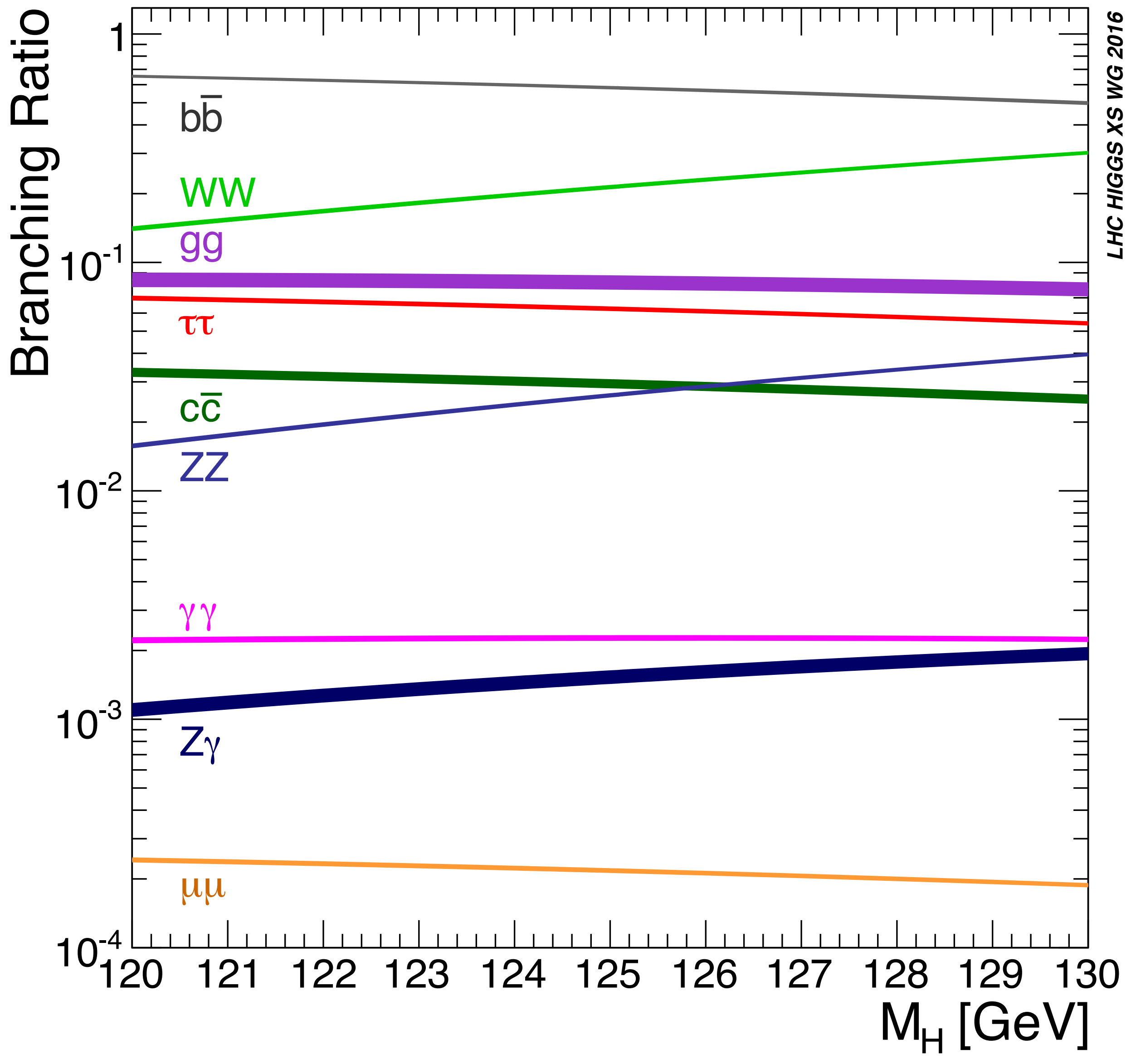


- Why: physics motivation
- Where: optimal channel
- How: optimal strategy

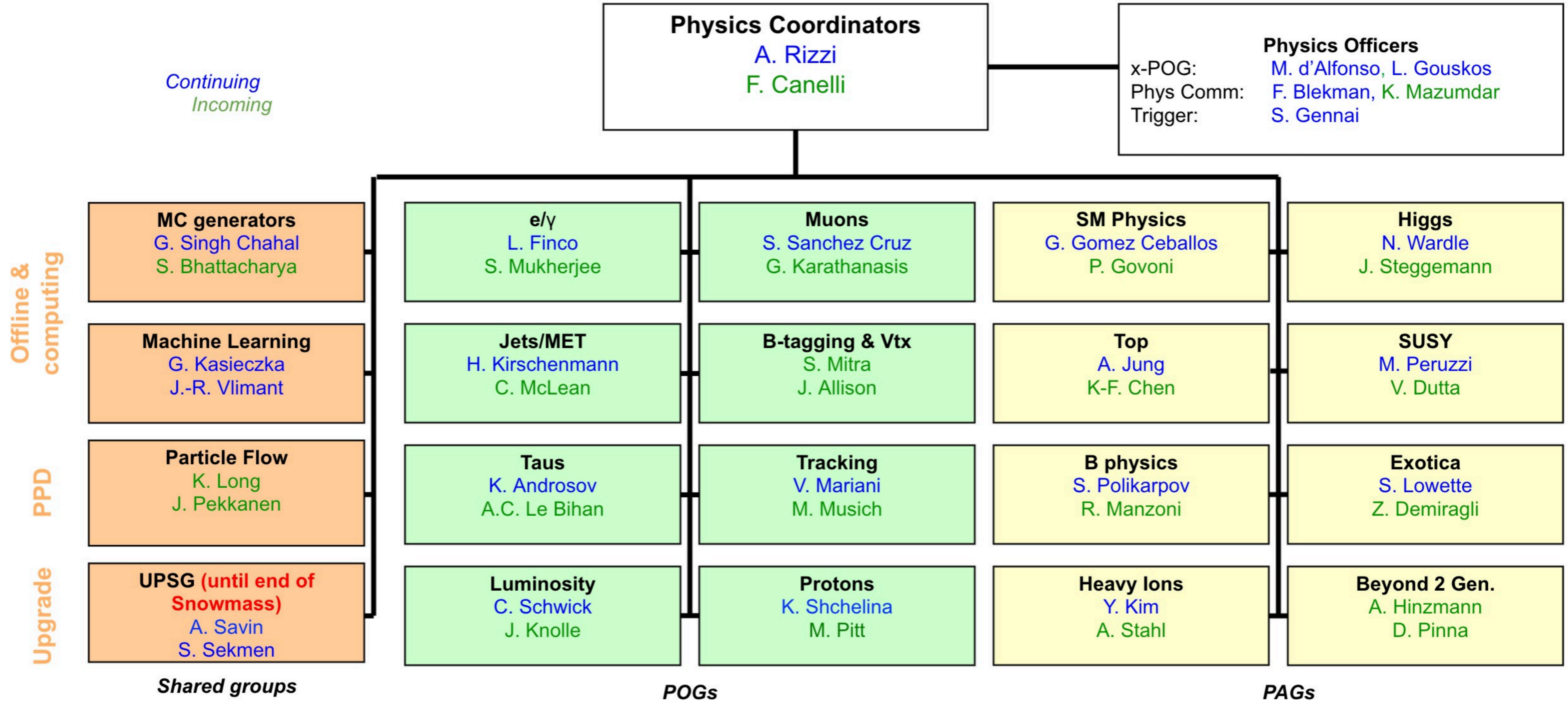


What does CMS see



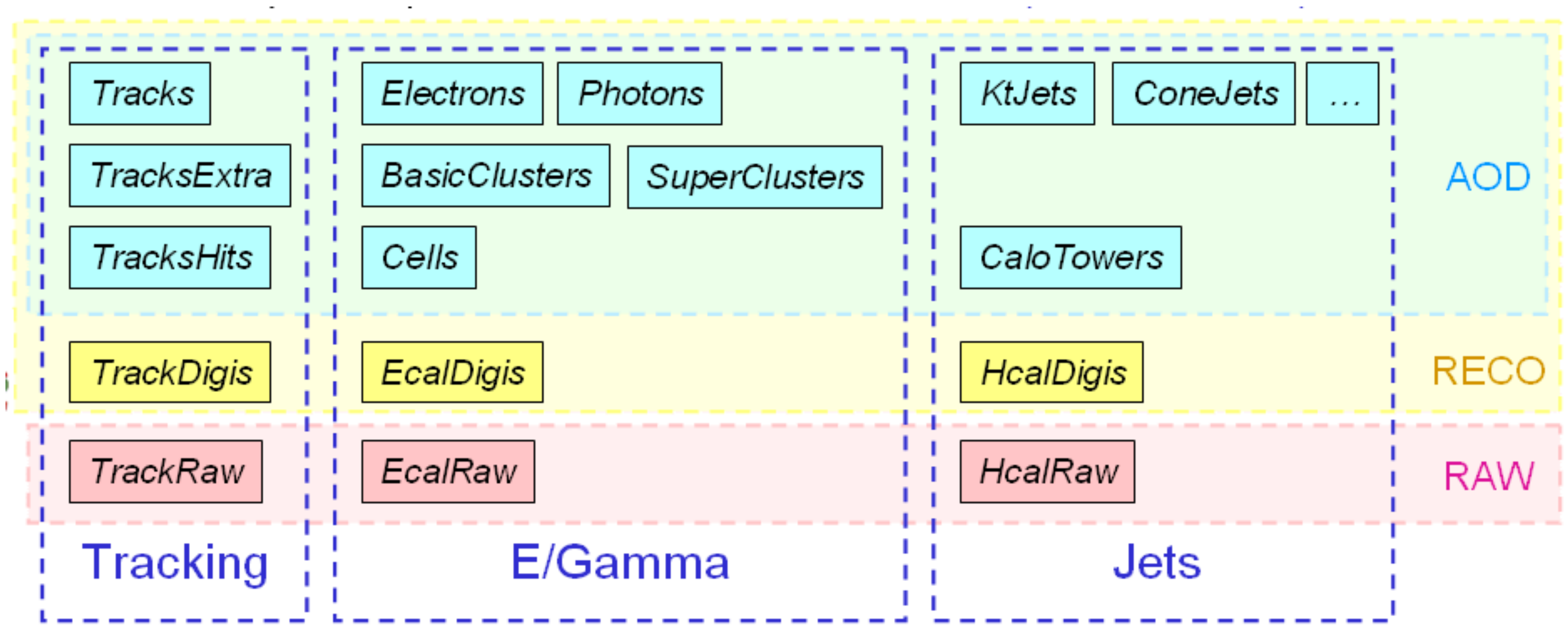


What does CMS see



POG: CMS Physics Object Groups   PAG: CMS Physics Analysis Groups

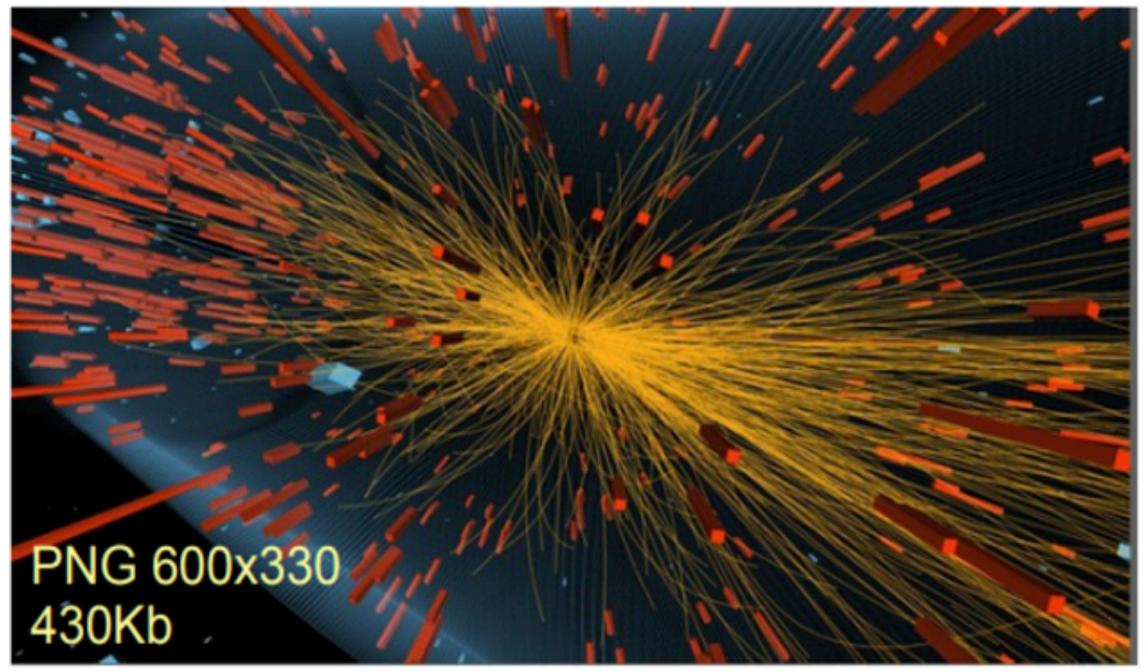
Data/MC format



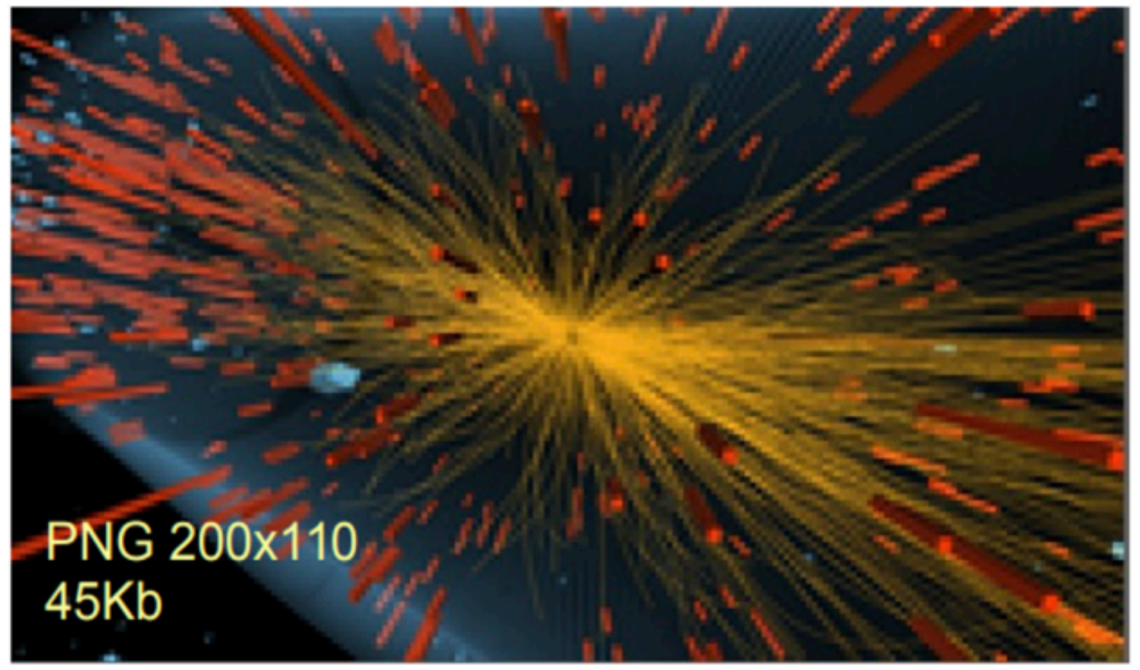
- For analyses, mostly miniAOD and nanoAOD, objects are all reconstructed

- CMSSW and root based

AOD 450kb/ev



MiniAOD 50kb/ev

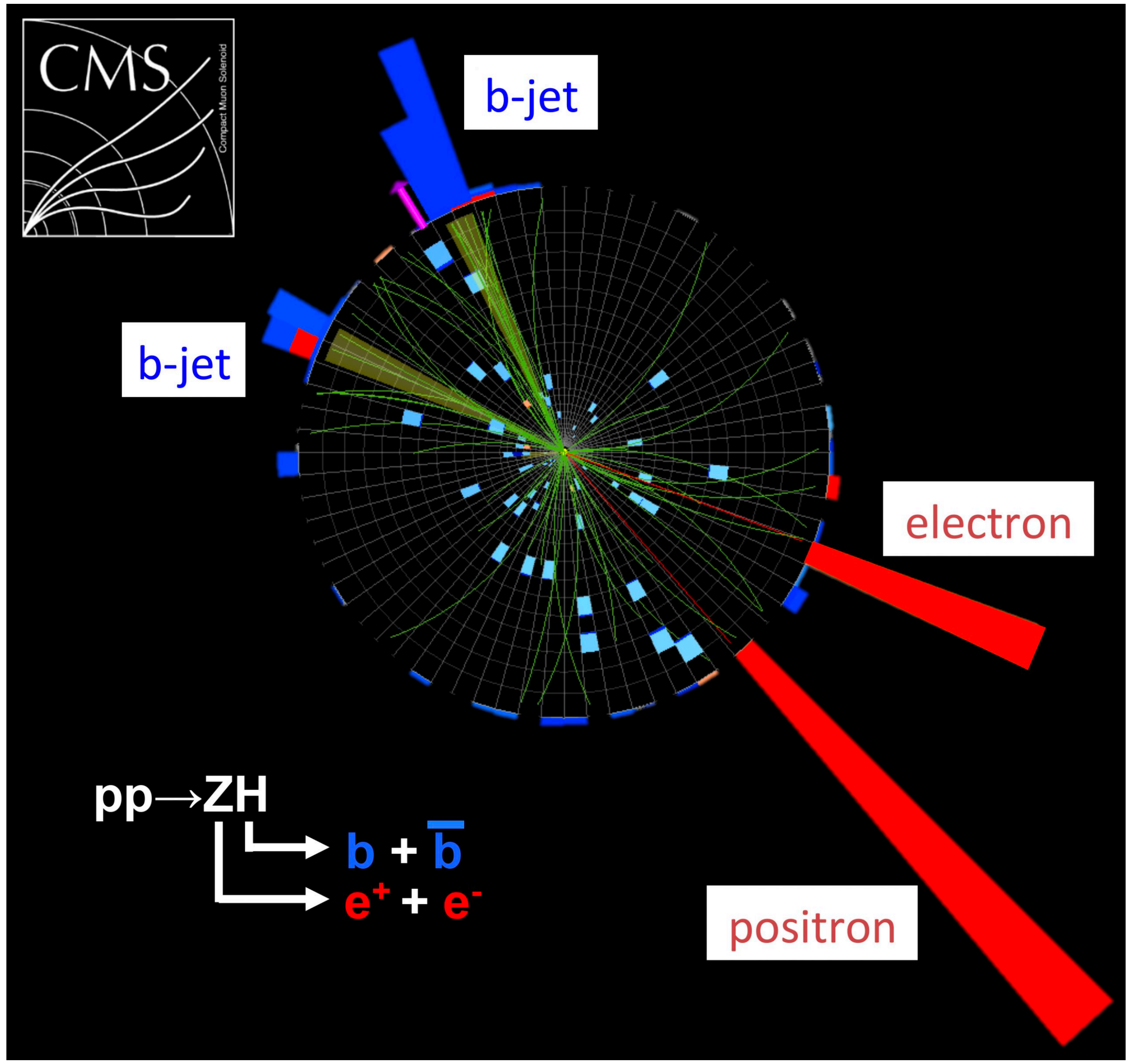


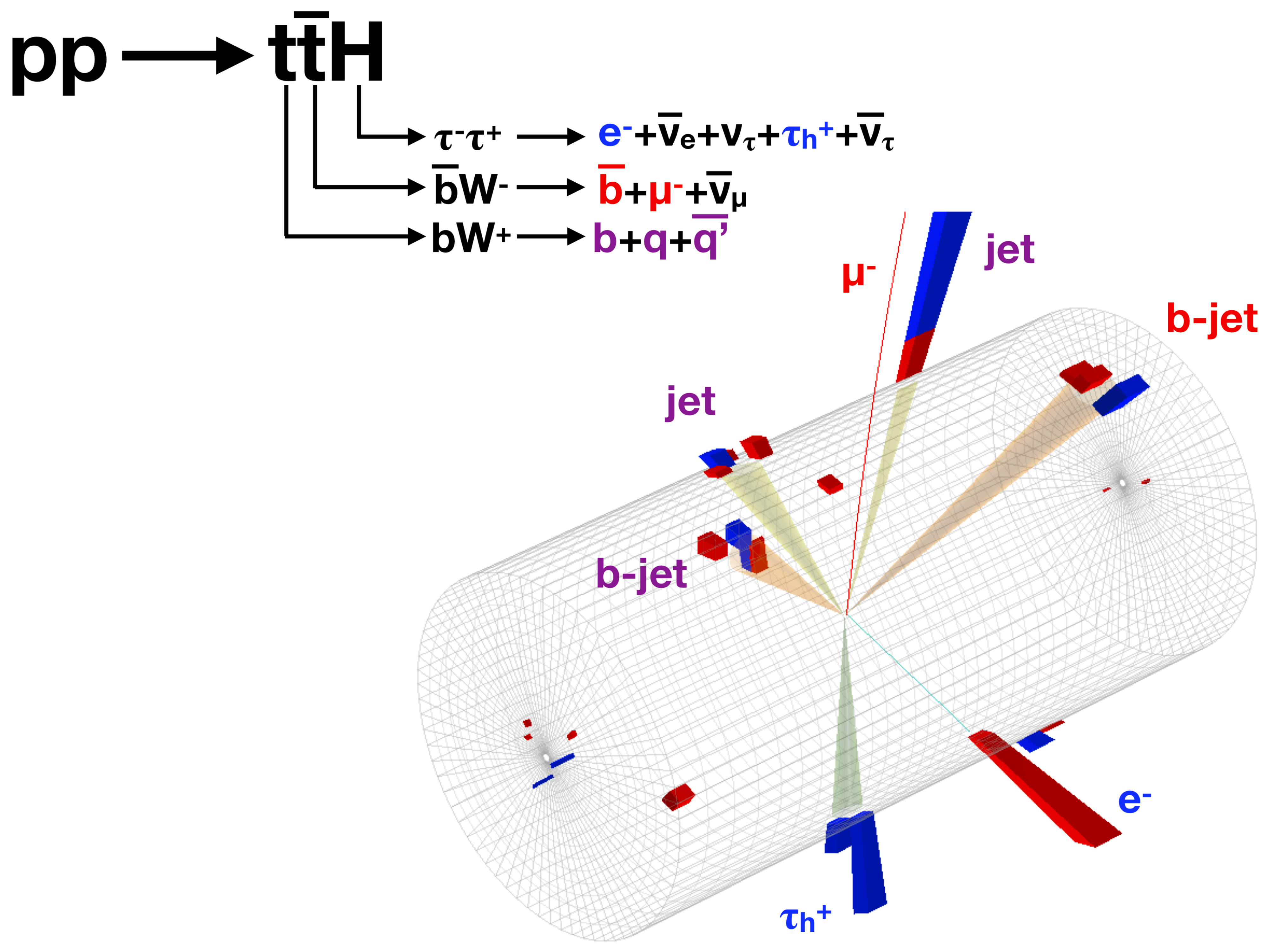
NanoAOD 1-2kb/ev

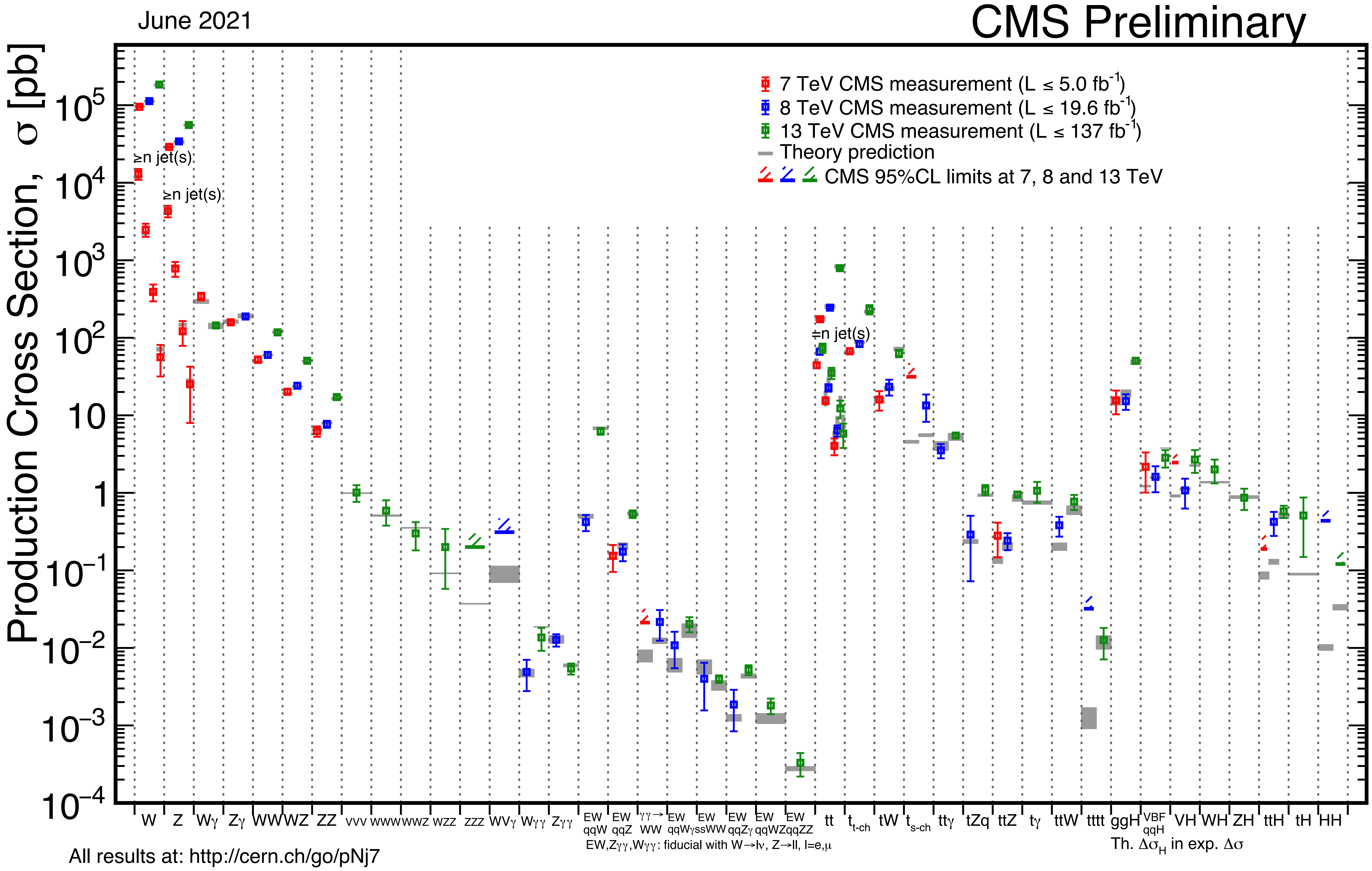
Taken from induction course here

Data/MC format

- What’s the difference between data and MC?

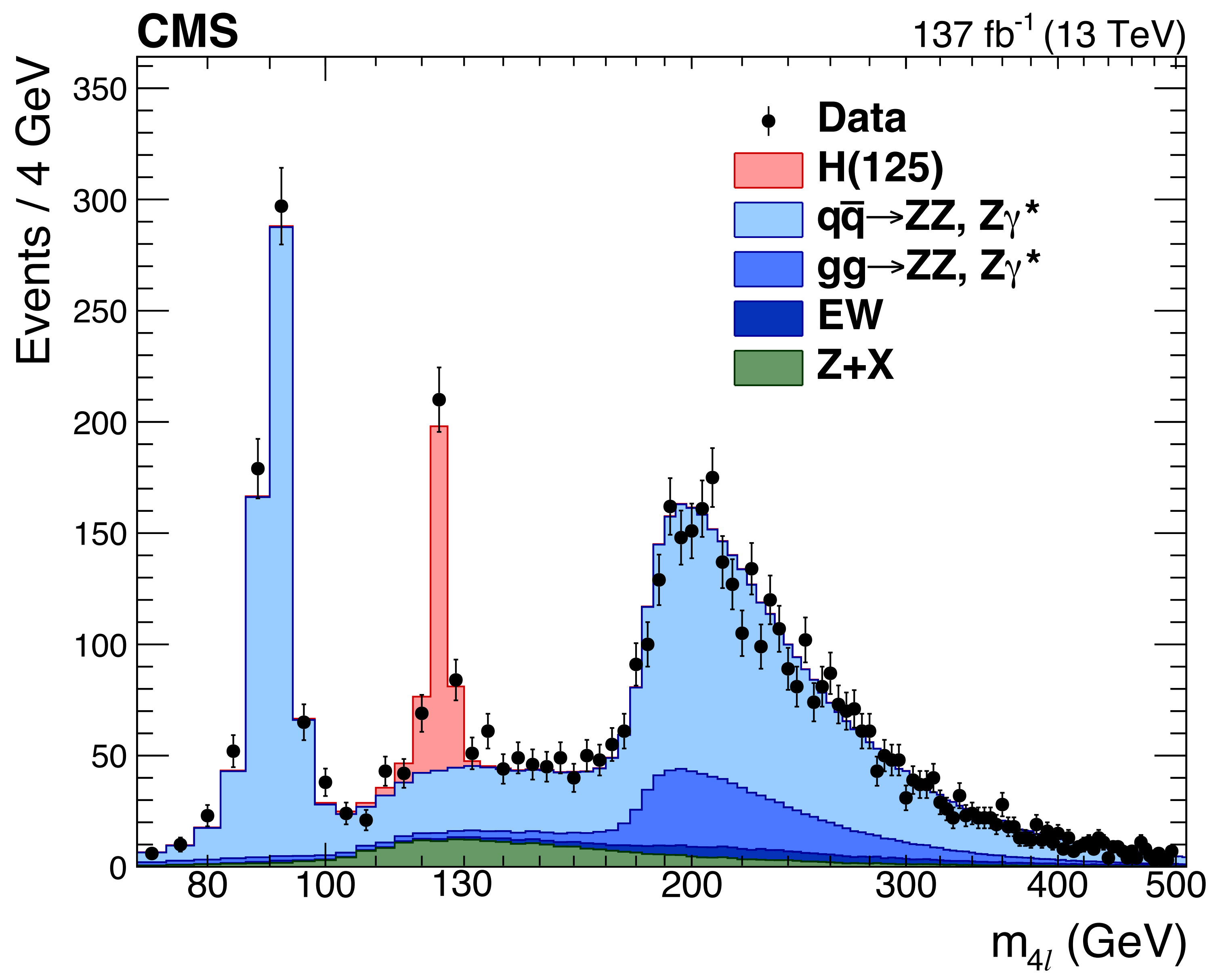





June 2021
CMS Preliminary
Production Cross Section, σ [pb]
7 TeV CMS measurement (L ≤ 5.0 fb-1)
8 TeV CMS measurement (L ≤ 19.6 fb-1)
13 TeV CMS measurement (L ≤ 137 fb-1)
Theory prediction
CMS 95%CL limits at 7, 8 and 13 TeV
≥n jet(s)
≥n jet(s)
=n jet(s)
W Z Wγ Zγ WW WZ ZZ VVV WWW WWZ WZZ ZZZ WVγ Wγγ Zγγ
EW qqW EW qqZ γγ→WW EW qqWγ EW ssWW EW qqZγ EW qqWZ EW qqZZ
tt t t-ch tW t s-ch ttγ tZq ttZ tγ ttW tttt ggH VBF qqH VH WH ZH ttH tH HH
EW,Zγγ,Wγγ: fiducial with W→lν, Z→ll, l=e,μ
Th. ΔσH in exp. Δσ
All results at: http://cern.ch/go/pNj7

MC

- Tools to simulate real data, usually for one particular process, e.g H→ 4l, qq→ 4l
  - Hard scattering: theoretical ME calculation at different orders, POWHEG, Madgraph (MC@NLO)…
  - Parton shower and hadronization: Pythia, Herwig, Sherpa…
  - Detector response: Geant4, Delphes…



Data and MC

- Where to look? CMS data aggregation system (DAS)
  - Website: https://cmsweb.cern.ch/das/
  - Terminal: dasgoclient -- query "dataset=/*/*/*"
- Naming conventions
  - Data: /DoubleEG/Run2016H-17Jul2018-v1/MINIAOD
    / Primary datasets / DataPeriod / Format
    Primary dataset: combination of a certain collections of HLT paths
  - MC: /QCD_HT1000to1500_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM

Trigger

- Data rates very high, 100 kHz, most of them not interesting
- HLT: high level trigger, 0.5 kHz, only events passing HLT stored
- Primary Datasets: grouped events with similar physical contents
  - Non-exclusive, one event could pass more than 1 HLT, thus 1 PD
- MC: apply the same HLT trigger and estimate scale factors

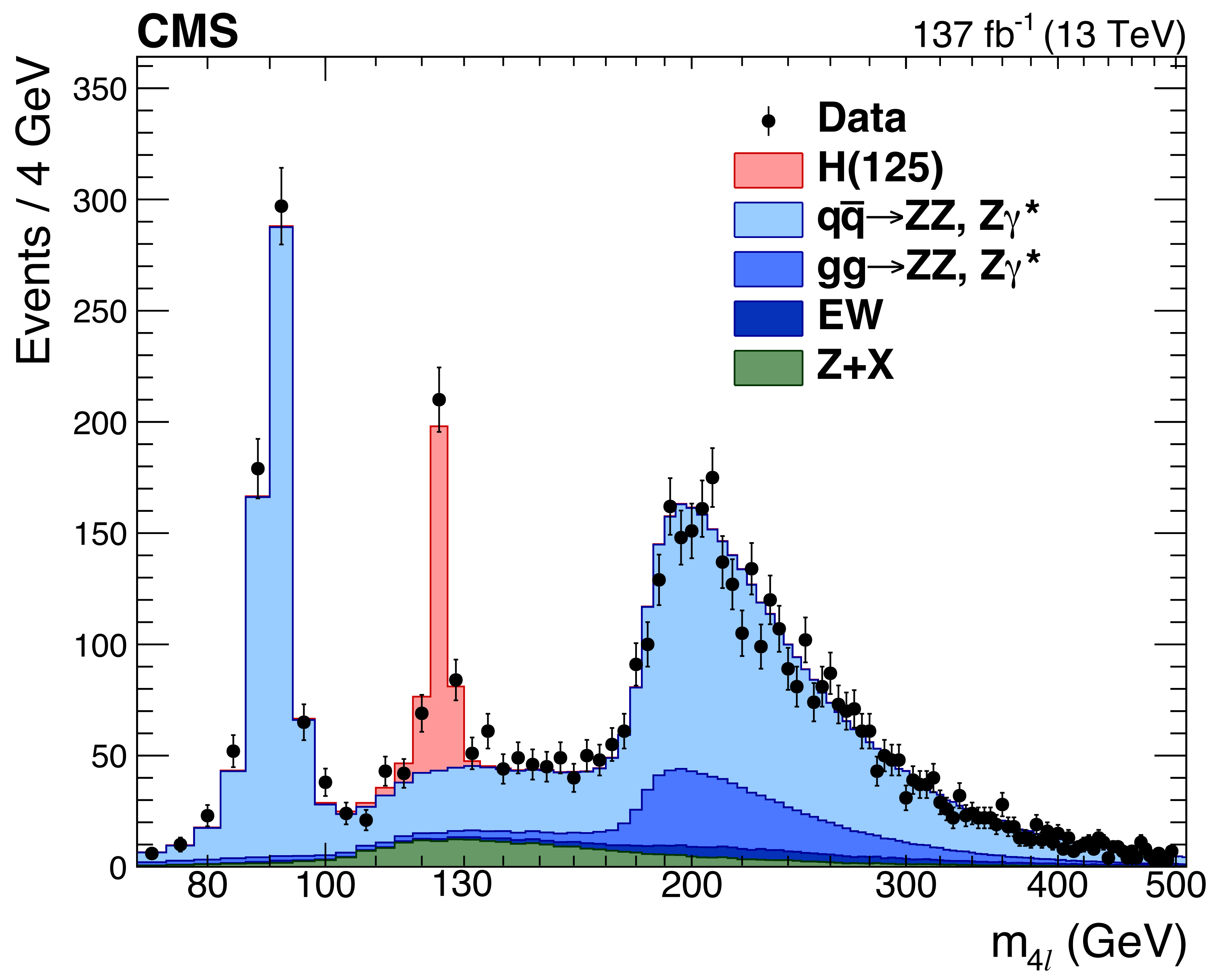
Primary Dataset	rate [Hz]
SingleElectron	68
DoubleElectron	15
ElectronHad	14
SinglePhoton	16
SinglePhotonParked	69
DoublePhoton	33
DoublePhotonHighPt	9
PhotonHad	11
MuEG	19
SingleMu	63
MuHad	14
DoubleMu	23
DoubleMuParked	26
MuOnia	38
MuOniaParked	94
Multijet	23
Multijet1Parked	174
JetHT	17
HTMHT	16
JetMon	6
VBF1Parked	201
MET	16
METParked	43
Tau	21
TauParked	40
TauPLusX	36
BTag	3
BTagPlusX	3

Numbers from here

HLT path	prescale	primary dataset
HLT_Ele17_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	1	DoubleEG
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	1	DoubleEG
HLT_DoubleEle33_CaloIdL_GsfTrkIdVL	1	DoubleEG
HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL	1	DoubleEG

Event selection

- Based on the final signatures, design selections and strategies (blind!)
- H→ ZZ→ 4l: final state 4 leptons, find them and calculate the invariant mass
- You will always see backgrounds
    - How to determine what the major bkg are?
    - Look for the same signature with large xsec

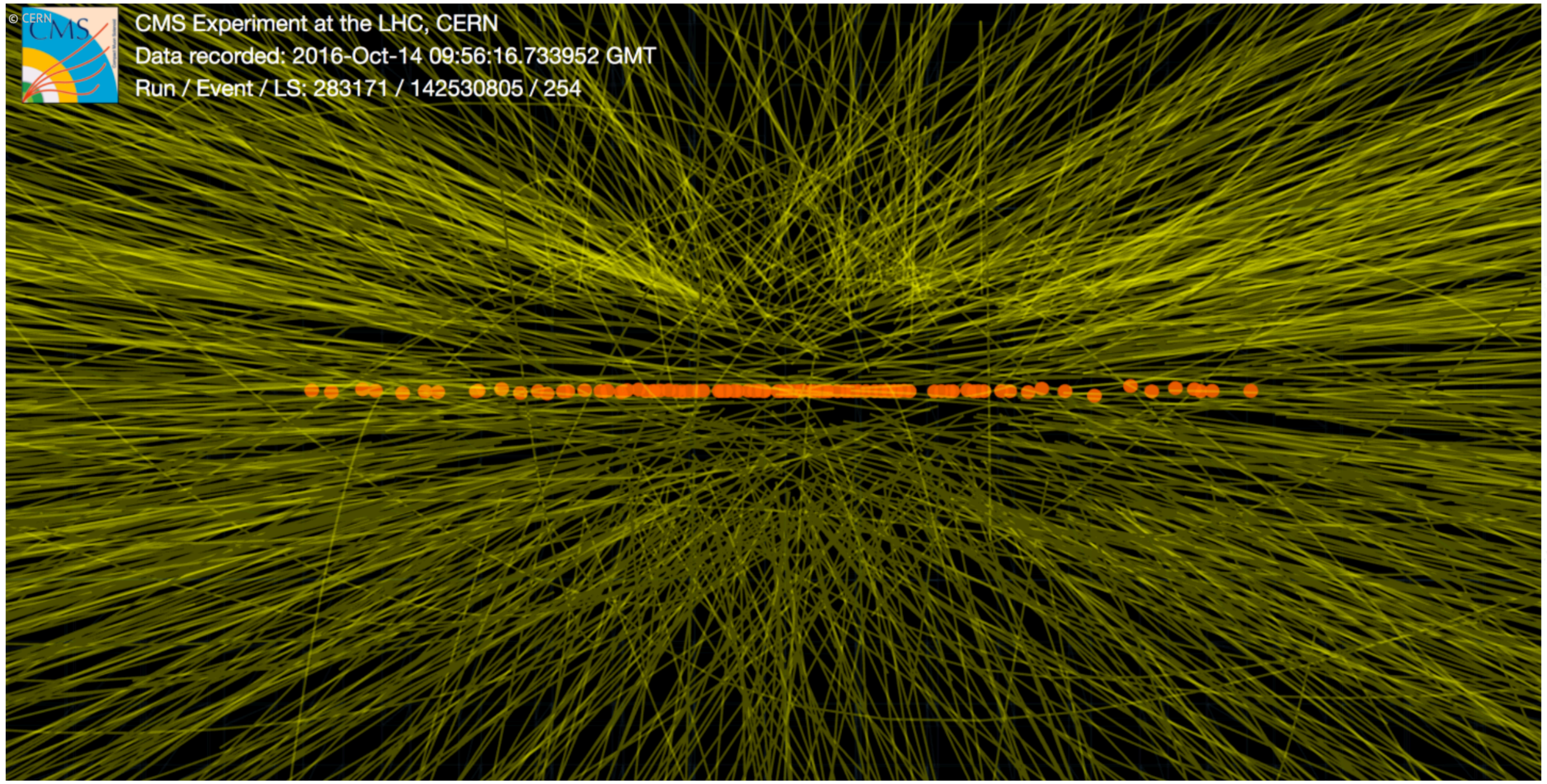


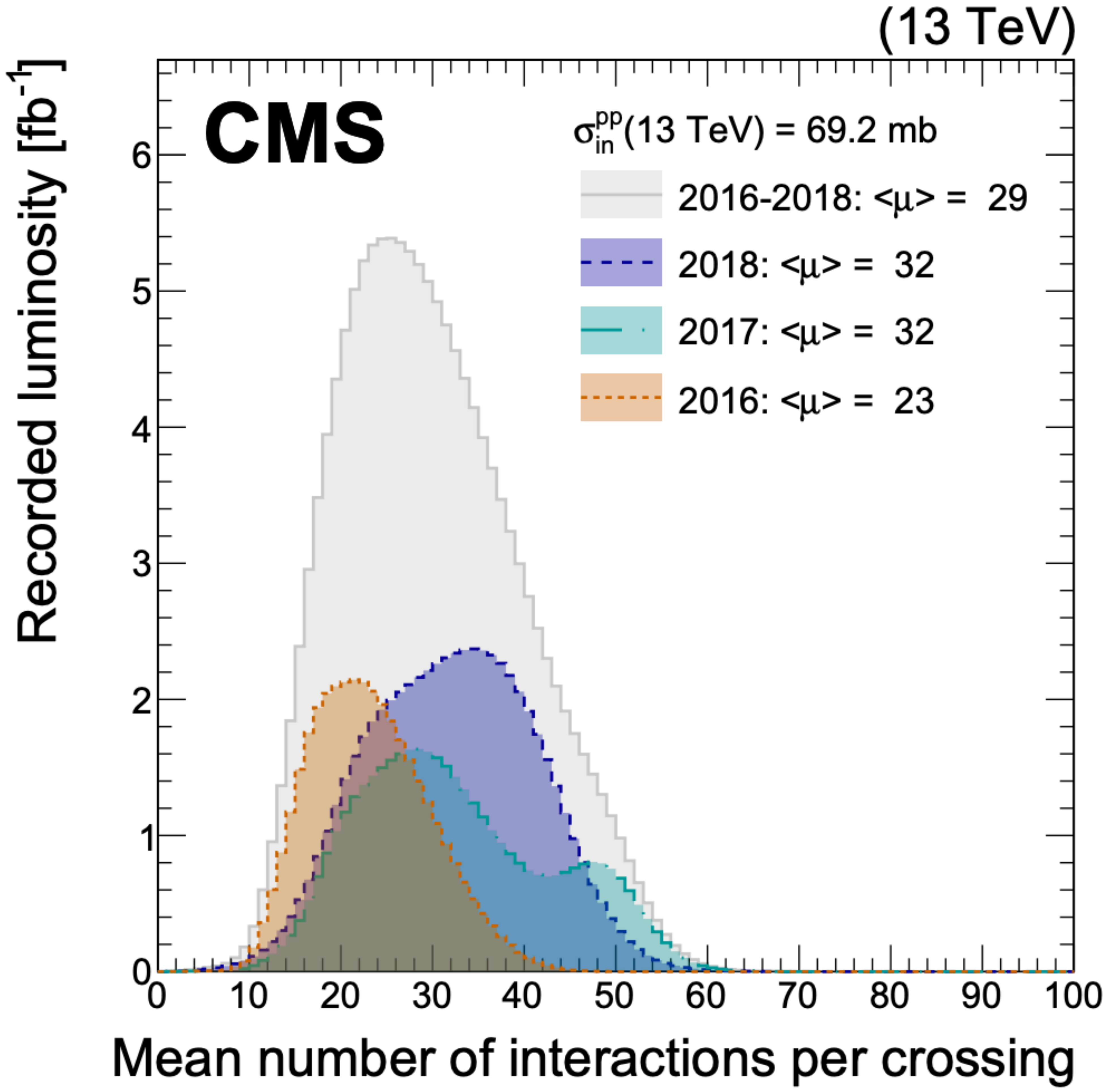
Event selection

- Trigger
- Pileup reweighting: match MC to data
- Object selection: good quality objects
- Analysis specific selections: enhance signal, reduce bkg
- Observable & category building: differentiate signal and bkg

Pileup

Multiple protons collide, concentrate on one, others become pileup in the event





Object selection

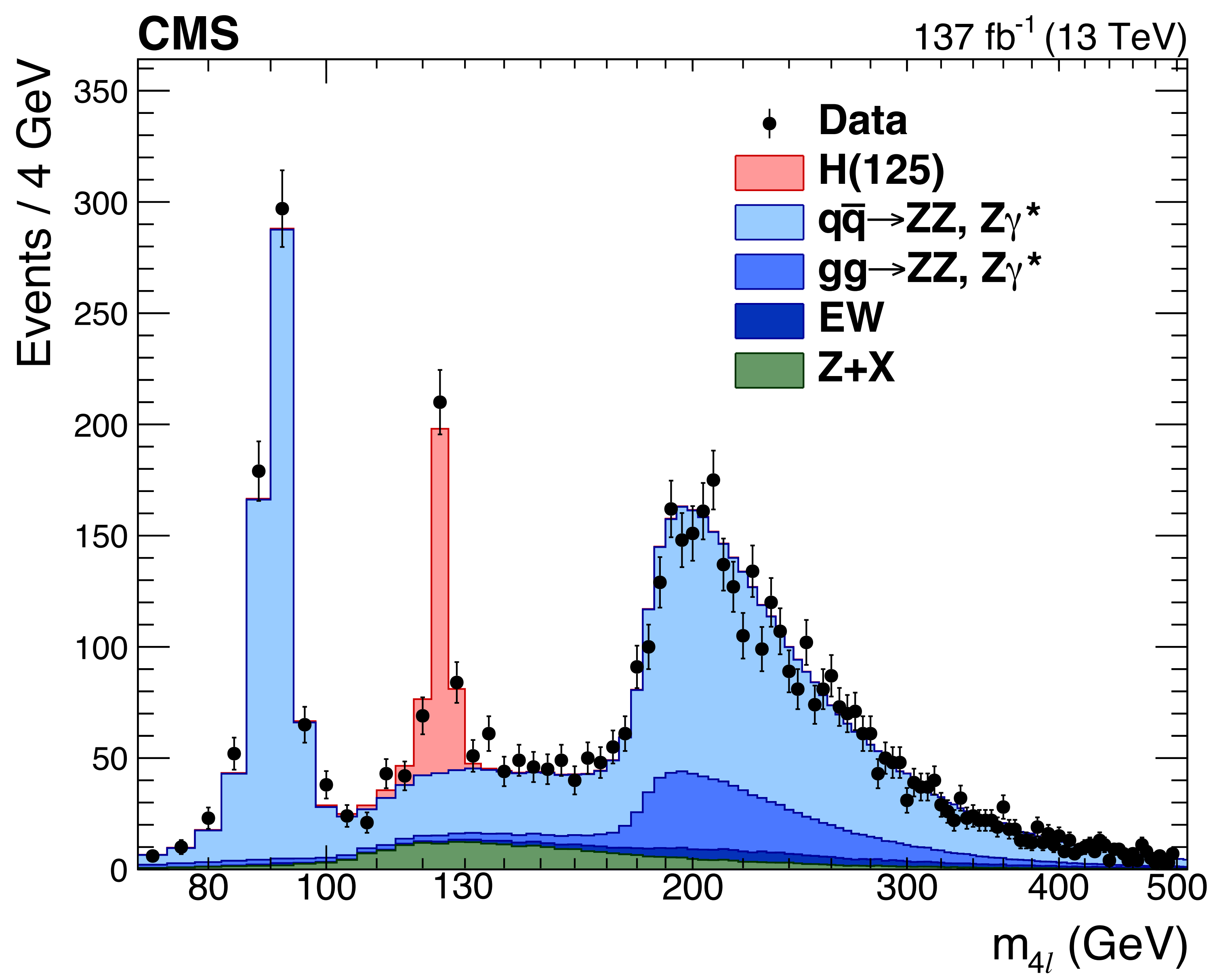
- Detector has certain acceptance/resolution, follow POG recommendations
- Kinematic selection: pT > xx, | η | < xx
- Object identification (ID): cut-based or machine learning techniques to identify objects from fakes
- Isolation & Significance of impact parameter (SIP)
- Energy/momentum calibration
- Scale factors of reconstruction/ID/Isolation/SIP: different selection efficiency in data/MC, use tag & probe
- Many analyses require development of object ID!

Good p_T measurement	$\frac{p_T}{\sigma_{p_T}} < 0.3$
Vertex compatibility $(x - y)$	$d_{xy} < 2$ mm
Vertex compatibility (z)	$d_z < 5$ mm
Pixel hits	At least one pixel hit
Tracker hits	Hits in at least six tracker layers

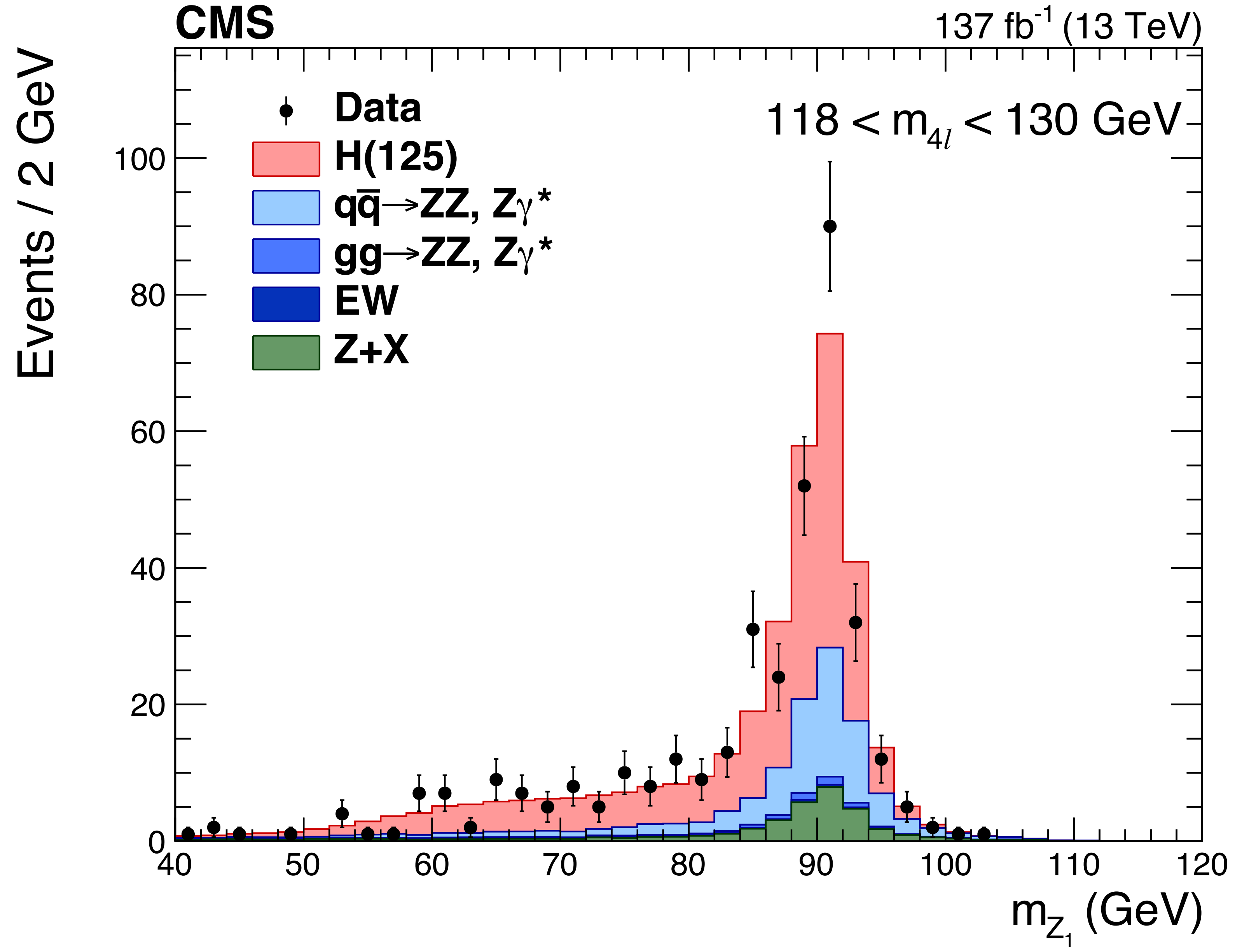
Muon ID

Analysis selection

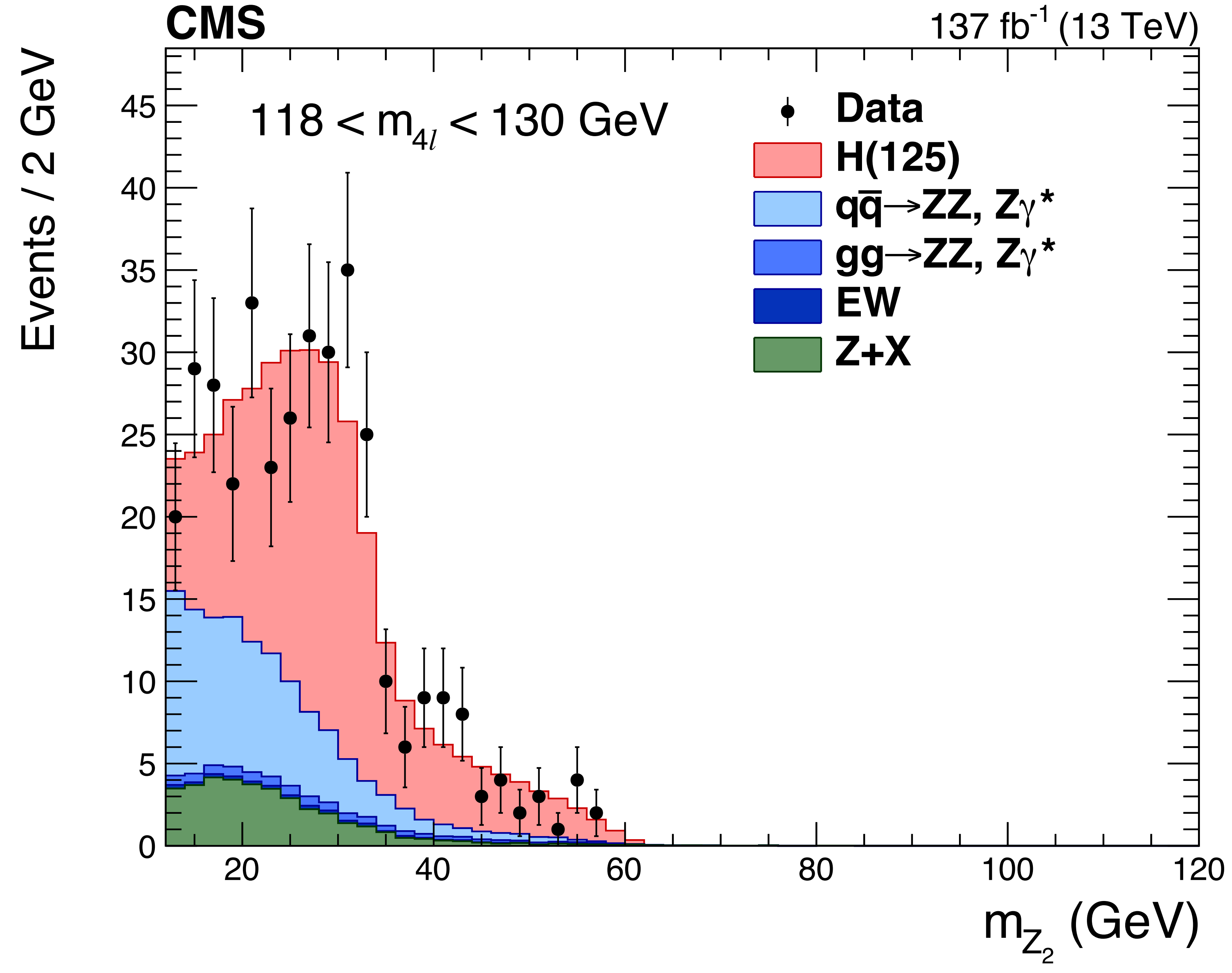
- Rely on careful MC based studies



M4l > 70 GeV

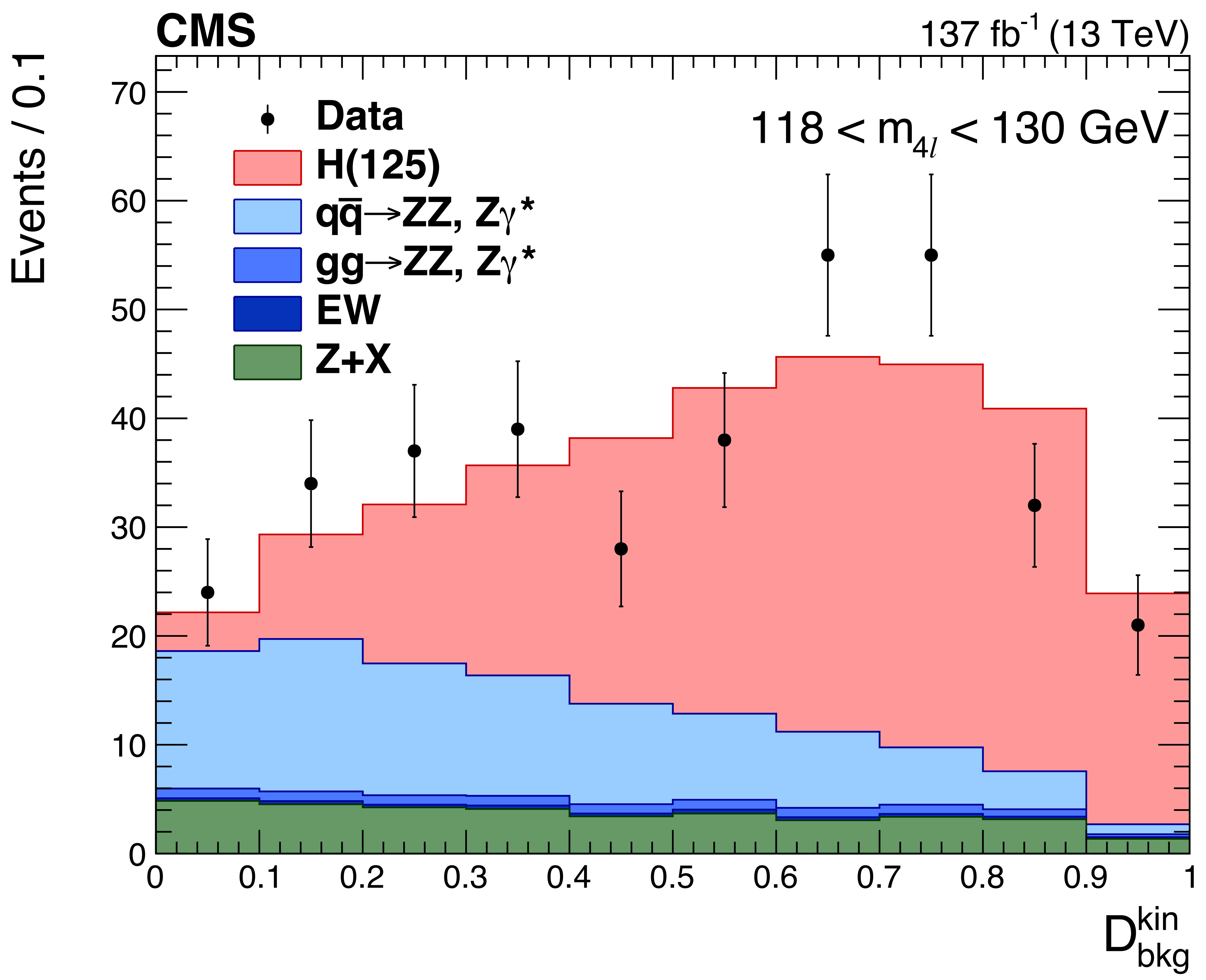


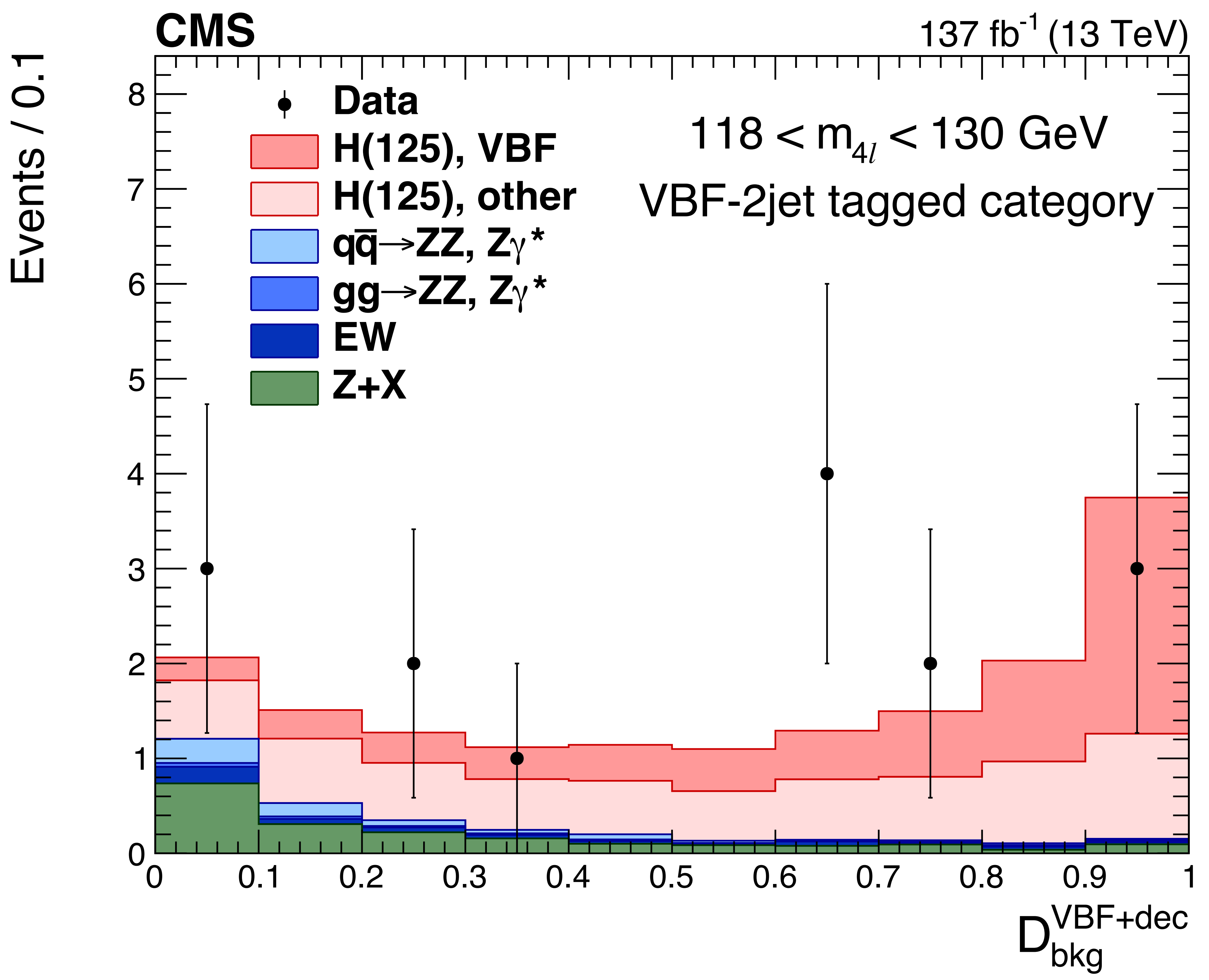
M12 > 40 GeV



M34 > 12 GeV

Observables & categories

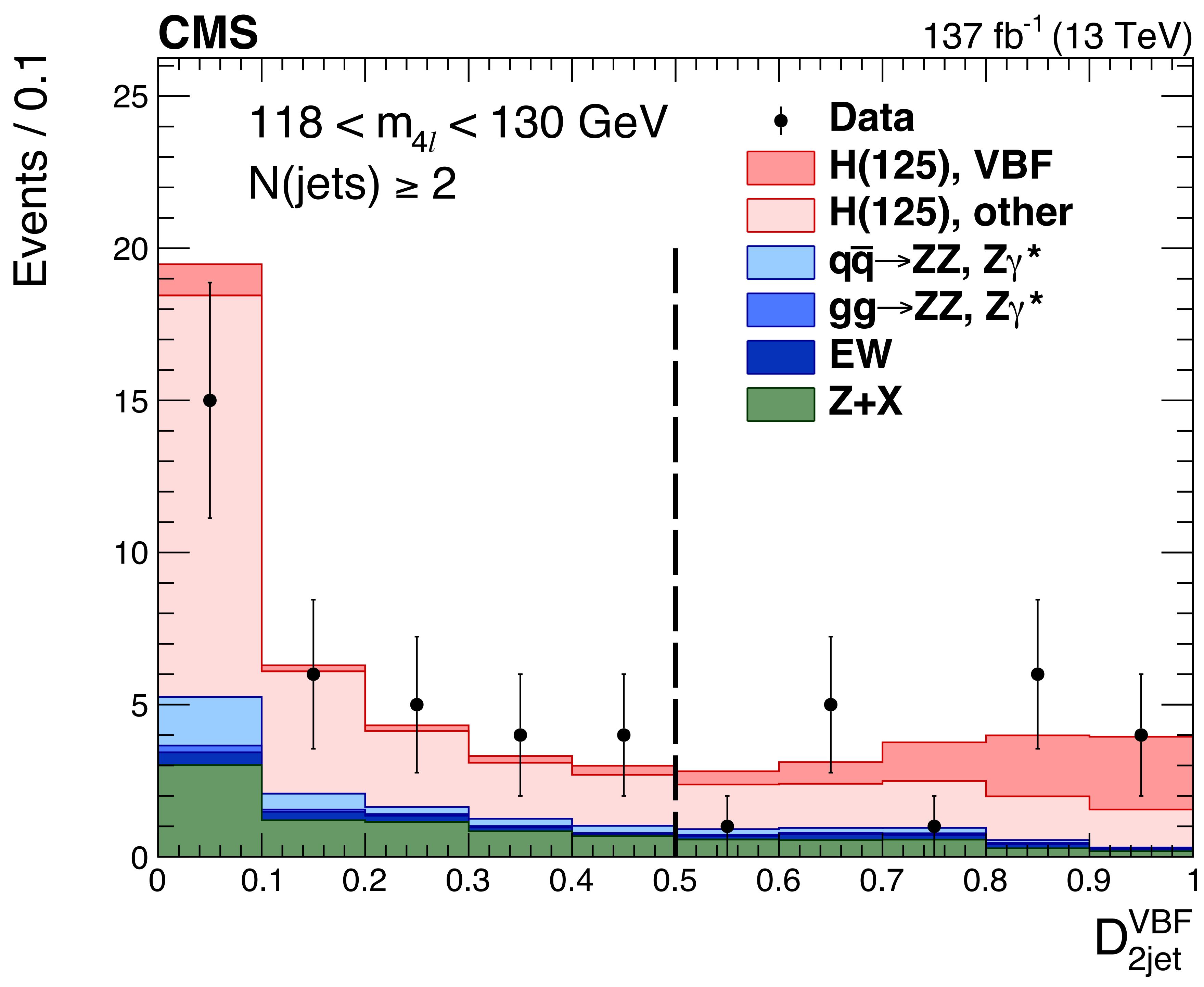


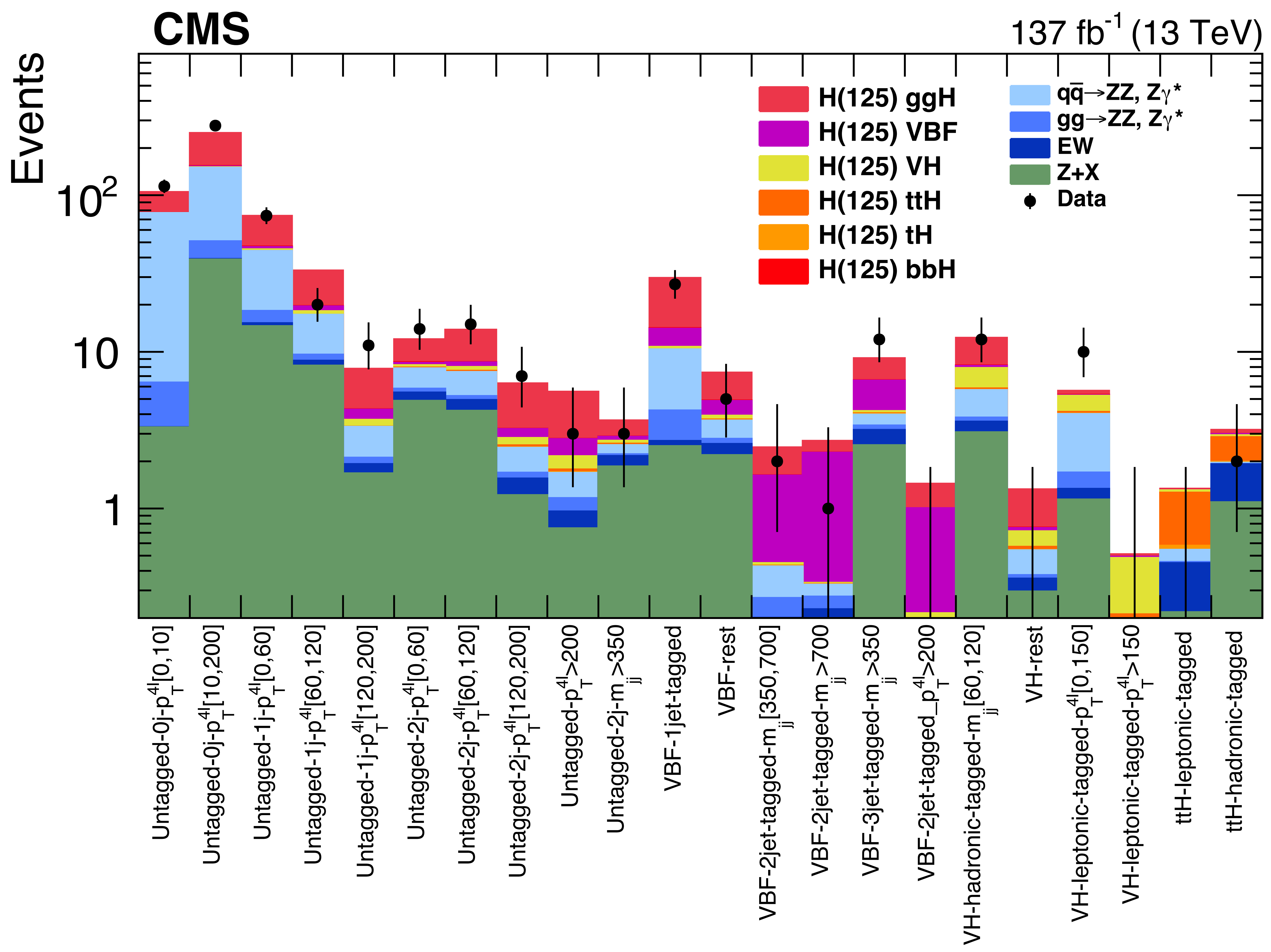


Cut based/Machine learning/ME based observables to differentiate signal from bkg

Observables & categories

- Design categories where S/B ratio is high





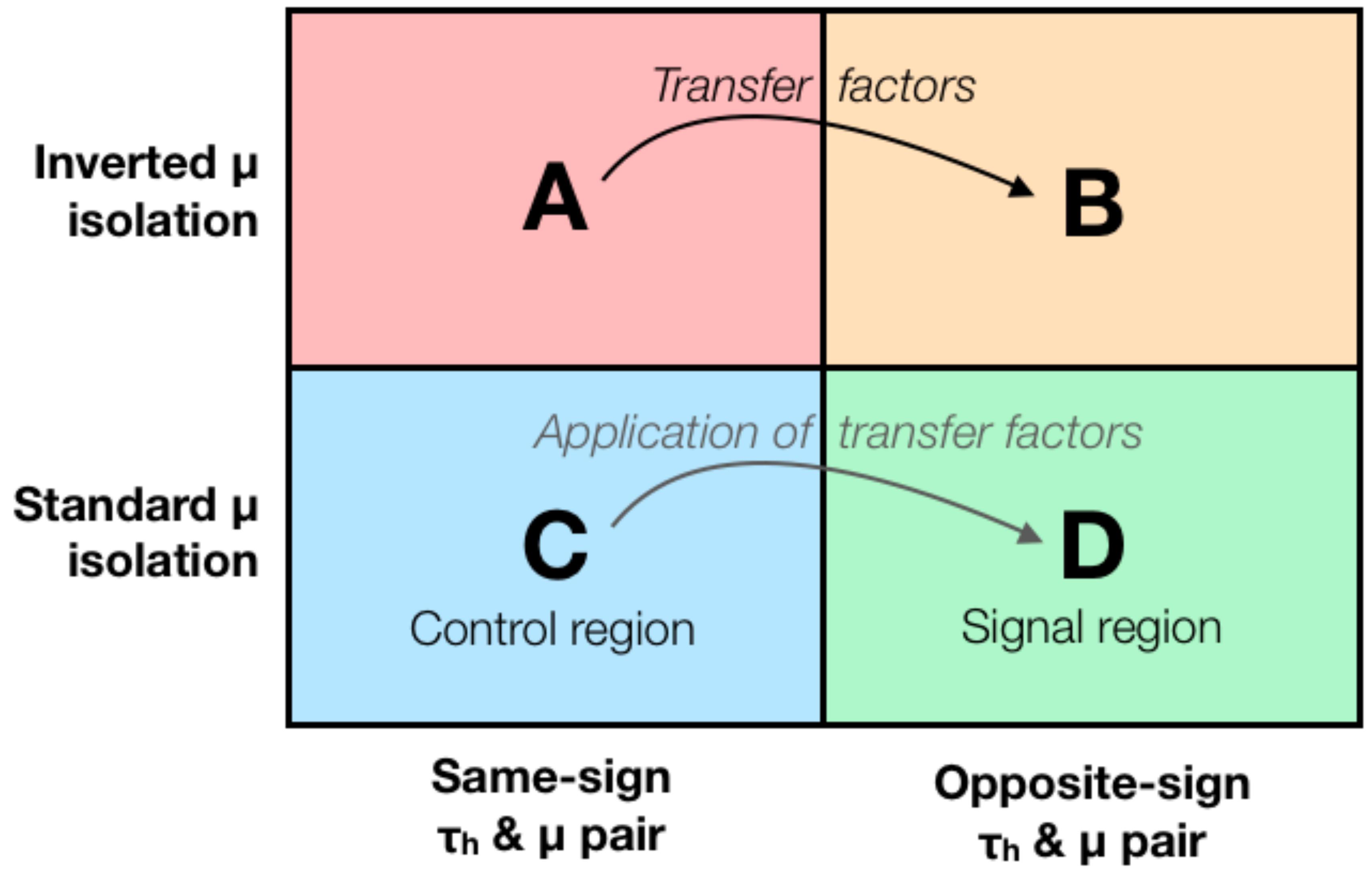
Expected #events

- MC could simulate as many as events
- What are the number of events to expect in reality after selection?
- Need to know
  - σ: xsec of the process
  - L: Luminosity of data taking
  - N_{MC}: #events generated
  - N_{sel}: #events after selection

$$N_{exp} = \frac{\sigma \mathscr{L}}{N_{MC}} N_{sel}$$

Reconstructed event	Signal							Background				Expected		Observed
category	ggH	VBF	WH	ZH	$t\bar{t}$H	$b\bar{b}$H	tH	$q\bar{q} \to$ ZZ	gg $\to$ ZZ	EW	Z+X	signal	total	
Untagged-0j-$p_T^{4\ell}$[0, 10]	27.7	0.09	0.03	0.03	0.00	0.15	0.00	71.5	3.06	0.01	3.21	27.9±0.1	106±0	114
Untagged-0j-$p_T^{4\ell}$[10, 200]	96.2	1.69	0.60	0.77	0.01	1.01	0.00	98.1	11.6	0.35	37.8	100±0	248±1	278
Untagged-1j-$p_T^{4\ell}$[0, 60]	26.8	1.51	0.56	0.48	0.01	0.45	0.01	25.3	3.02	0.64	14.2	29.8±0.1	72.9±0.4	74
Untagged-1j-$p_T^{4\ell}$[60, 120]	13.5	1.31	0.51	0.41	0.02	0.11	0.01	7.81	0.82	0.62	7.95	15.9±0.1	33.1±0.3	20
Untagged-1j-$p_T^{4\ell}$[120, 200]	3.51	0.60	0.17	0.17	0.01	0.02	0.00	1.15	0.19	0.25	1.63	4.48±0.05	7.69±0.16	11
Untagged-2j-$p_T^{4\ell}$[0, 60]	3.45	0.29	0.15	0.14	0.08	0.09	0.02	2.14	0.32	0.63	4.75	4.20±0.06	12.1±0.2	14
Untagged-2j-$p_T^{4\ell}$[60, 120]	5.26	0.56	0.24	0.19	0.12	0.04	0.03	2.19	0.30	0.72	4.14	6.43±0.06	13.8±0.2	15
Untagged-2j-$p_T^{4\ell}$[120, 200]	3.07	0.40	0.16	0.13	0.07	0.01	0.02	0.75	0.14	0.34	1.19	3.86±0.05	6.28±0.14	7
Untagged-$p_T^{4\ell}$ > 200	2.79	0.62	0.21	0.17	0.07	0.01	0.02	0.43	0.21	0.21	0.73	3.89±0.04	5.47±0.11	3
Untagged-2j-m_{jj} > 350	0.77	0.16	0.06	0.04	0.05	0.01	0.01	0.34	0.06	0.31	1.71	1.12±0.02	3.54±0.14	3
VBF-1jet-tagged	15.5	3.29	0.22	0.16	0.00	0.13	0.01	6.85	1.53	0.20	2.44	19.3±0.1	30.3±0.2	27
VBF-2jet-tagged-m_{jj}[350, 700]	0.83	1.19	0.01	0.01	0.00	0.01	0.00	0.19	0.07	0.11	0.14	2.05±0.03	2.55±0.05	2
VBF-2jet-tagged-m_{jj} > 700	0.43	1.96	0.00	0.00	0.00	0.00	0.00	0.07	0.05	0.12	0.03	2.40±0.02	2.67±0.03	1
VBF-3jet-tagged-m_{jj} > 350	2.52	2.35	0.06	0.06	0.03	0.03	0.05	0.62	0.21	0.64	2.43	5.11±0.05	9.01±0.17	12
VBF-2jet-tagged-$p_T^{4\ell}$ > 200	0.44	0.79	0.01	0.01	0.01	0.00	0.01	0.03	0.03	0.04	0.06	1.26±0.02	1.42±0.03	0
VBF-rest	2.48	0.94	0.13	0.09	0.04	0.04	0.01	0.98	0.20	0.39	2.18	3.74±0.05	7.49±0.17	5
VH-hadronic-tagged-m_{jj}[60, 120]	4.11	0.25	1.09	0.96	0.13	0.06	0.02	1.69	0.22	0.52	2.93	6.62±0.06	12.0±0.2	12
VH-rest	0.57	0.03	0.09	0.06	0.03	0.01	0.00	0.16	0.02	0.06	0.33	0.79±0.02	1.36±0.06	0
VH-leptonic-tagged-$p_T^{4\ell}$[0, 150]	0.33	0.04	0.85	0.26	0.10	0.03	0.03	2.16	0.36	0.19	1.11	1.64±0.02	5.47±0.13	10
VH-leptonic-tagged-$p_T^{4\ell}$ > 150	0.02	0.01	0.21	0.06	0.04	0.00	0.01	0.05	0.01	0.03	0.08	0.35±0.01	0.52±0.03	0
$t\bar{t}$H-leptonic-tagged	0.02	0.01	0.02	0.02	0.68	0.00	0.03	0.08	0.01	0.23	0.21	0.79±0.01	1.32±0.07	0
$t\bar{t}$H-hadronic-tagged	0.18	0.05	0.03	0.05	0.86	0.01	0.03	0.03	0.01	0.82	1.06	1.22±0.01	3.15±0.14	2

Background estimation



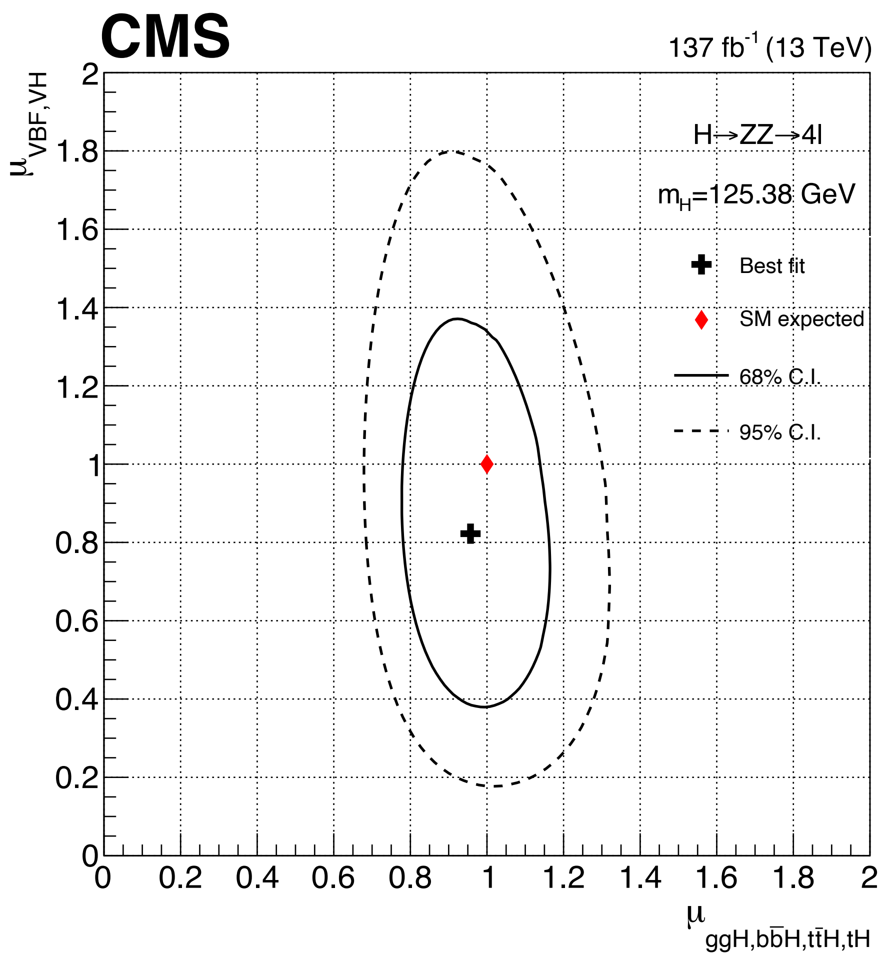
- In most cases, we rely on MC to model signals
- For bkg, if the MC prediction is reliable, use MC
  - Otherwise, e.g. QCD, DY+jets.. data driven methods
- Data-driven methods usually combines data and MC
  - Control region (CR): orthogonal to signal region, where bkg is dominant, trust data in CR
  - Transfer factor: from CR to signal region (SR), relies on MC or some other data (ABCD method)
  - Requires a lot of validations!

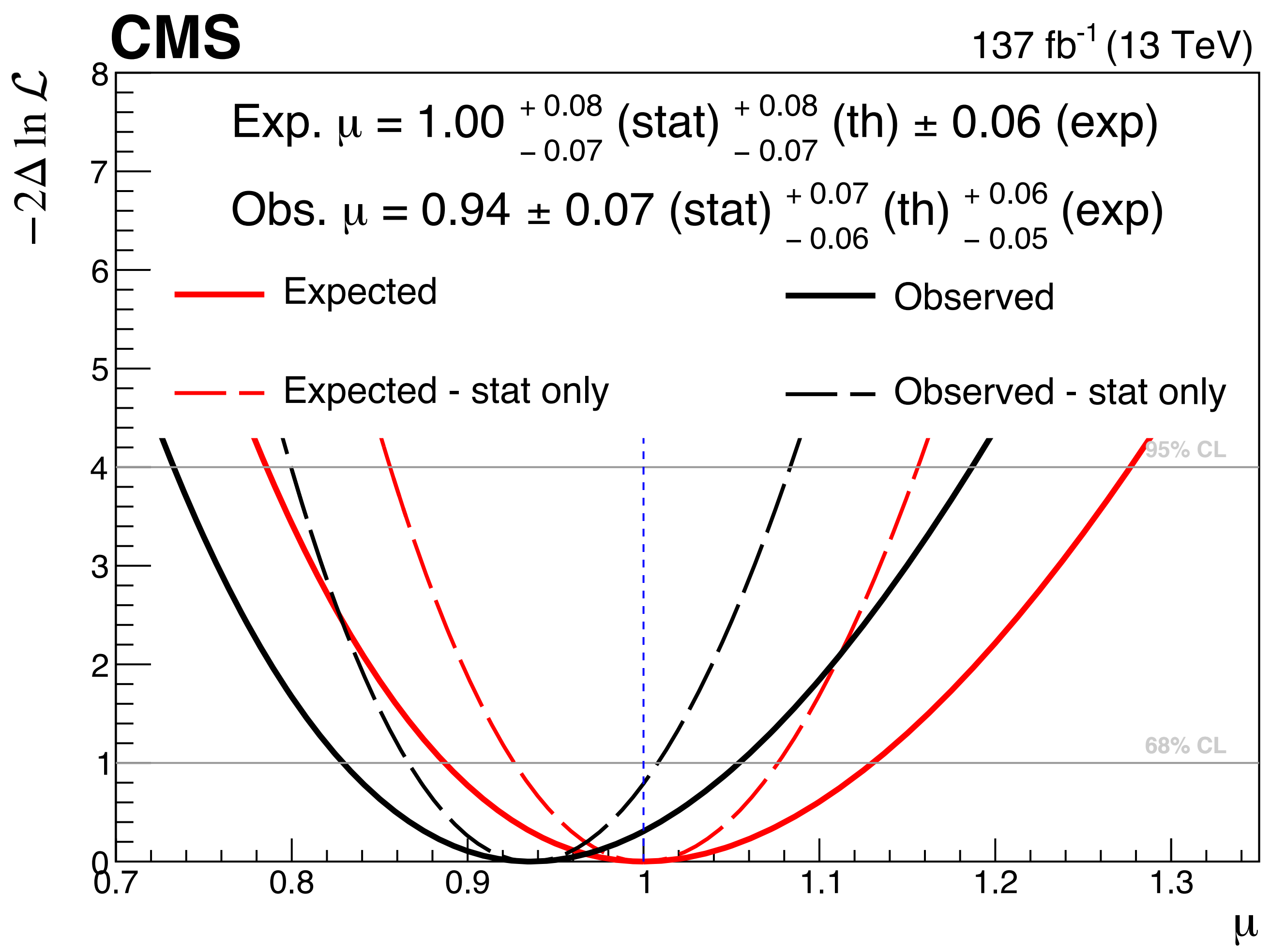
Systematics

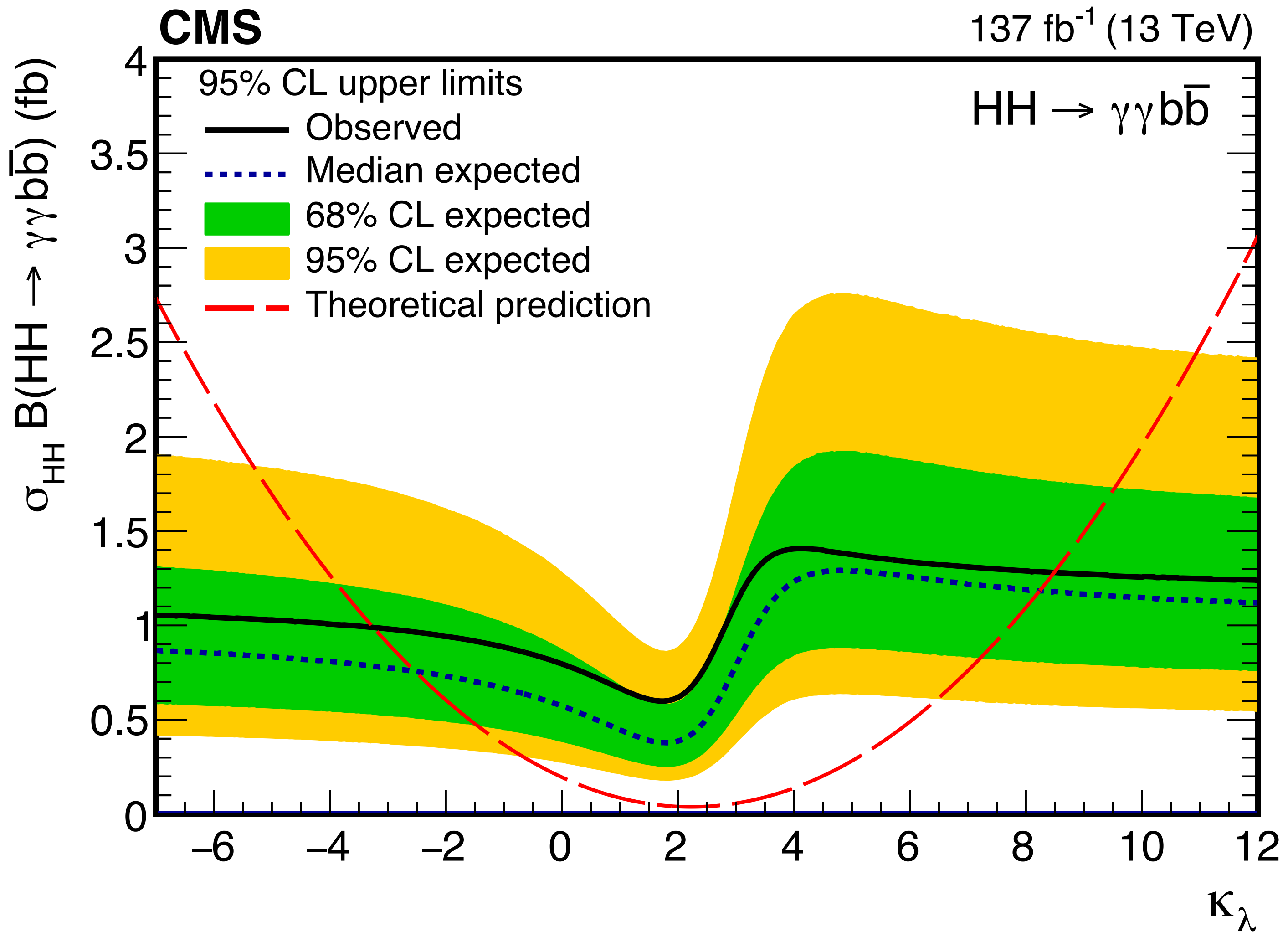
- MC is not reality
- Theoretical uncertainty
  - QCD scale, EW correction, PDF, parton shower, hadronization
- Experimental uncertainty
  - Luminosity
  - Object energy/momentum scale/resolution
  - Object reconstruction/identification efficiency
  - Data driven method
- Yield/Shape uncertainty, log-Normal nuisances

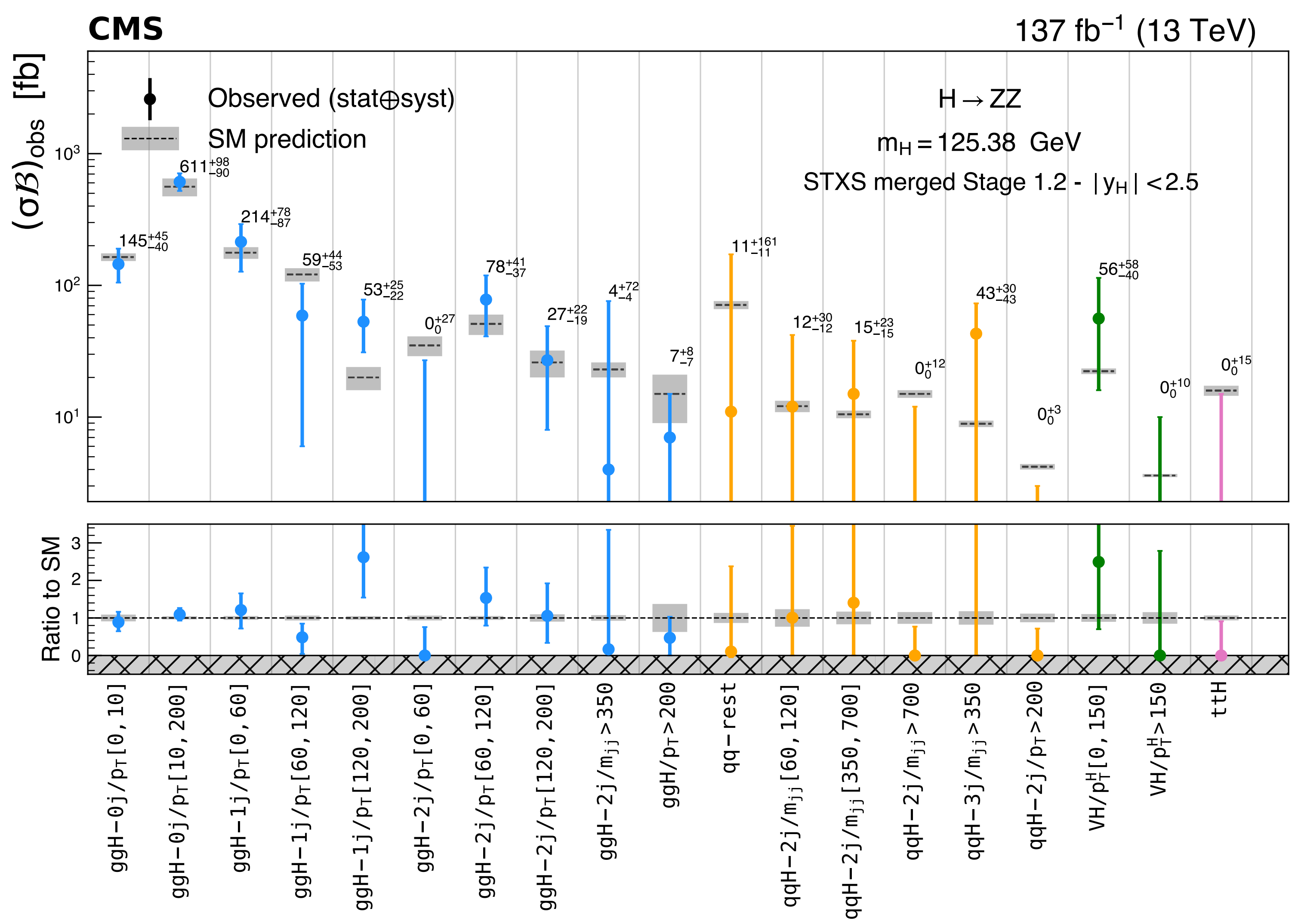
Statistical interpretation

- Precision measurement / Search
- Roofit/RooStats
- Tool in CMS: Higgs-CombinedLimit package
  - Parameter estimate, test statistics
  - Maximum log likelihood, chi2
  - 68% CL, 95% CL
  - Significance, p-value
  - Upper limit









To start in CMS

- To get familiar with CMS: CMS induction course, link
- To know what's being published: CMS public results, link
- To get a deeper understanding of an analysis: internal CADI line, link
- To get reminders of all internal communication on certain physics/object groups, subscribe to:
  - CMS hyper news, link
  - CMS Talk, link