Experimental Methods and Physics at the LHC - II

Sezen Sekmen
Kyungpook National University / CMS

26th Vietnam School of Physics: Particles and Dark Matter
29 Nov - 11 Dec 2020, Quy Nhon & virtual

Background estimation

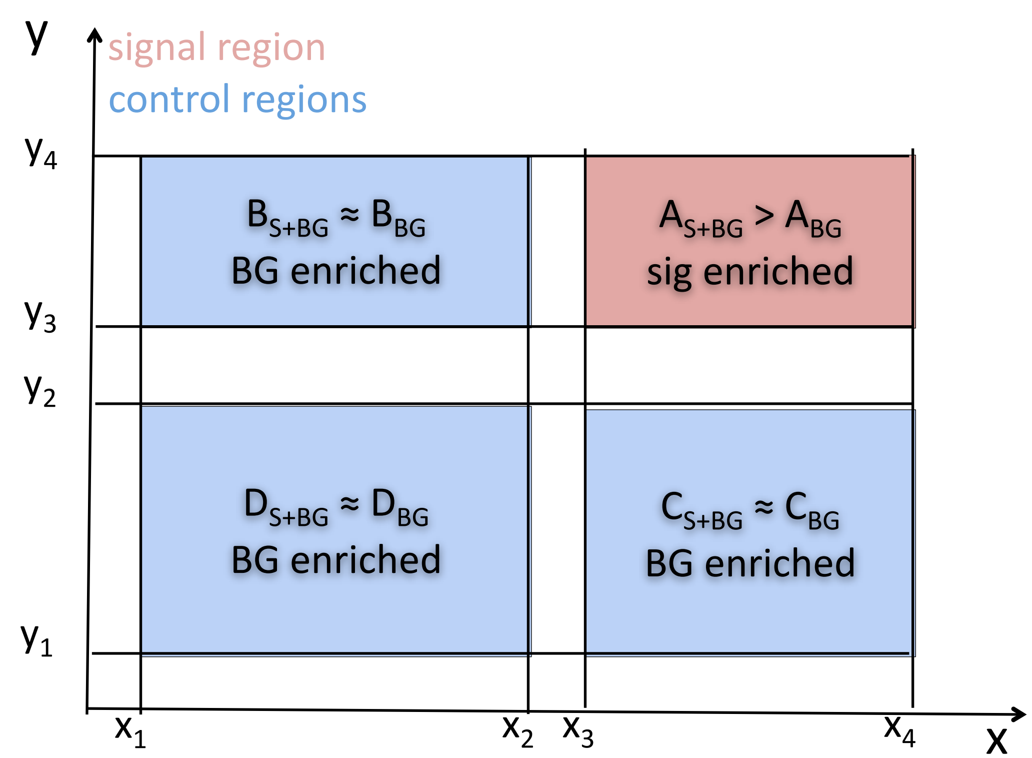
Matrix – or ABCD - method

When there exist two variables x and y for which the BG is uncorrelated, i.e. factorizable:

$$f^{BG}(x,y) = f^{BG}(x) \cdot f^{BG}(y)$$

- Apply all cuts except those on x and y on data
- Divide the x-y plane into 4-regions:
- When there is no signal, we have

$$\frac{N_A^{BG}}{N_B^{BG}} = \frac{N_C^{BG}}{N_D^{BG}}, \qquad \frac{N_A^{BG}}{N_C^{BG}} = \frac{N_B^{BG}}{N_D^{BG}}$$



- In the presence of signal, A will be contaminated by the signal. But we can estimate the number of BG events in A from

$$N_A^{BG} = \frac{N_C^{BG} N_B^{BG}}{N_D^{BG}}$$

Note: Always beware the signal contamination in the control regions. Add it as a systematic.

Matrix – or ABCD - method

When there exist two variables x and y for which the BG is uncorrelated, i.e. factorizable:

$$f^{BG}(x,y) = f^{BG}(x) \cdot f^{BG}(y)$$

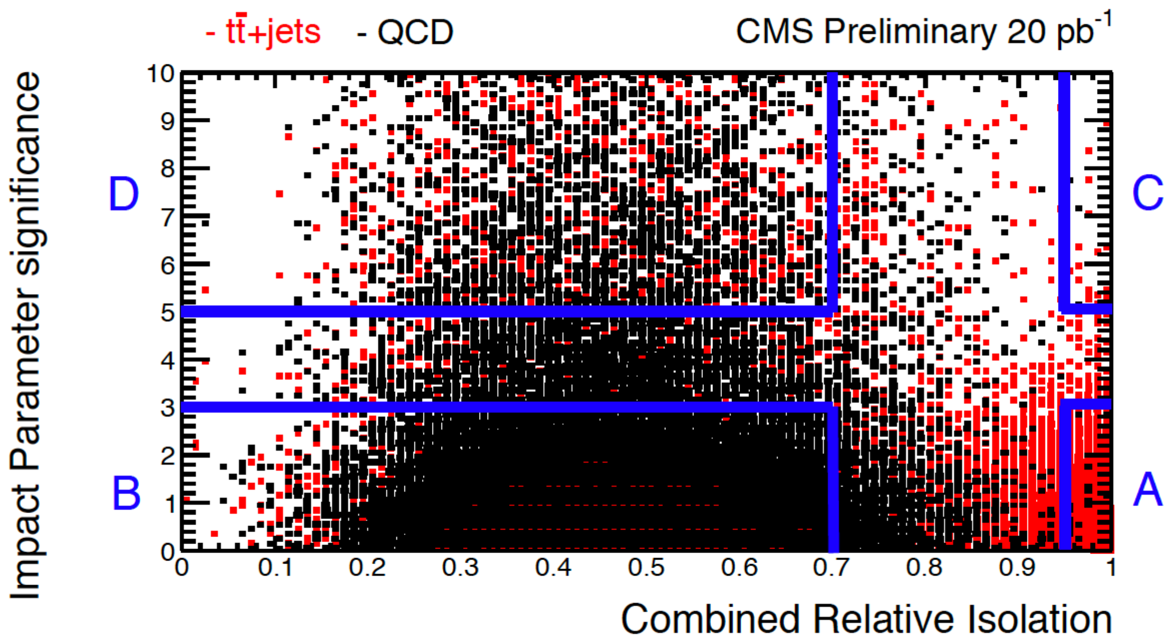
- Apply all cuts except those on x and y on data
- Divide the x-y plane into 4-regions:
- When there is no signal, we have

$$\frac{N_A^{BG}}{N_B^{BG}} = \frac{N_C^{BG}}{N_D^{BG}}, \quad \frac{N_A^{BG}}{N_C^{BG}} = \frac{N_B^{BG}}{N_D^{BG}}$$

- In the presence of signal, A will be contaminated by the signal. But we can estimate the number of BG events in A from

$$N_A^{BG} = \frac{N_C^{BG} N_B^{BG}}{N_D^{BG}}$$

CMS tt+jets cross section measurement in the muon+jets channel.



Note: Always beware the signal contamination in the control regions. Add it as a systematic.

Background estimation
Fake rates method

The ratios of objects found by a tight identification over objects found by a loose identification is widely used as a BG estimation tool.

Suppose we would like to estimate QCD BG in a signal region that has leptons with tight ID. Signal contribution comes from real leptons and QCD contribution comes from fake leptons (jets faking leptons). We define a control region with looser lepton ID. CR and SR can be decomposed as:

CR: $N_{loose} = N_{loose}^{real} + N_{loose}^{fake}$

SR: $N_{tight} = N_{tight}^{real} + N_{tight}^{fake}$

$$\epsilon^k \equiv N_{tight}^k / N_{loose}^k \rightarrow \quad = \epsilon^{real} N_{loose}^{real} + \epsilon^{fake} N_{loose}^{fake}$$

Background estimation

Fake rates method

The ratios of objects found by a tight identification over objects found by a loose identification is widely used as a BG estimation tool.

Suppose we would like to estimate QCD BG in a signal region that has leptons with tight ID. Signal contribution comes from real leptons and QCD contribution comes from fake leptons (jets faking leptons). We define a control region with looser lepton ID. CR and SR can be decomposed as:

Get these counts from data

CR: $$\boxed{N_{loose}} = N^{real}_{loose} + N^{fake}_{loose}$$

SR: $$\boxed{N_{tight}} = N^{real}_{tight} + N^{fake}_{tight}$$

$$\epsilon^k \equiv N^k_{tight}/N^k_{loose} \rightarrow \quad = \epsilon^{real} N^{real}_{loose} + \epsilon^{fake} N^{fake}_{loose}$$

Fake rates method

The ratios of objects found by a tight identification over objects found by a loose identification is widely used as a BG estimation tool.

Suppose we would like to estimate QCD BG in a signal region that has leptons with tight ID. Signal contribution comes from real leptons and QCD contribution comes from fake leptons (jets faking leptons). We define a control region with looser lepton ID. CR and SR can be decomposed as:

Get these counts from data

CR: $\boxed{N_{loose}} = N_{loose}^{real} + N_{loose}^{fake}$

SR: $\boxed{N_{tight}} = N_{tight}^{real} + N_{tight}^{fake}$

$$\epsilon^k \equiv N_{tight}^k / N_{loose}^k \rightarrow \quad = \boxed{\epsilon^{real}} N_{loose}^{real} + \boxed{\epsilon^{fake}} N_{loose}^{fake}$$

Find the ID efficiency, using e.g. tag and probe method.

Find this efficiency, i.e. the fake rate using e.g. a QCD control sample (e.g. low MET)

Fake rates method

The ratios of objects found by a tight identification over objects found by a loose identification is widely used as a BG estimation tool.

Suppose we would like to estimate QCD BG in a signal region that has leptons with tight ID. Signal contribution comes from real leptons and QCD contribution comes from fake leptons (jets faking leptons). We define a control region with looser lepton ID. CR and SR can be decomposed as:

Get these counts from data

CR: $\boxed{N_{loose}} = \boxed{N^{real}_{loose}} + \boxed{N^{fake}_{loose}}$

SR: $\boxed{N_{tight}} = N^{real}_{tight} + N^{fake}_{tight}$

$\epsilon^k \equiv N^k_{tight}/N^k_{loose} \rightarrow \; = \boxed{\epsilon^{real}}\boxed{N^{real}_{loose}} + \boxed{\epsilon^{fake}}\boxed{N^{fake}_{loose}}$

Solve the two equations simultaneously to get these numbers.

Find the ID efficiency, using e.g. tag and probe method.

Find this efficiency, i.e. the fake rate using e.g. a QCD control sample (e.g. low MET)

Fake rates method

The ratios of objects found by a tight identification over objects found by a loose identification is widely used as a BG estimation tool.

Suppose we would like to estimate QCD BG in a signal region that has leptons with tight ID. Signal contribution comes from real leptons and QCD contribution comes from fake leptons (jets faking leptons). We define a control region with looser lepton ID. CR and SR can be decomposed as:

Get these counts from data

CR: $\boxed{N_{loose}} = \boxed{N^{real}_{loose}} + \boxed{N^{fake}_{loose}}$

SR: $\boxed{N_{tight}} = N^{real}_{tight} + N^{fake}_{tight}$

$\epsilon^k \equiv N^k_{tight}/N^k_{loose} \rightarrow \quad = \boxed{\epsilon^{real}}\boxed{N^{real}_{loose}} + \boxed{\epsilon^{fake}}\boxed{N^{fake}_{loose}}$

Solve the two equations simultaneously to get these numbers.

Find the ID efficiency, using e.g. tag and probe method.

Find this efficiency, i.e. the fake rate using e.g. a QCD control sample (e.g. low MET)

Finally obtain the number of BG events from $\epsilon^{fake} N^{fake}_{loose} = N^{fake}_{tight} = N_{BG}$

Tag-and probe method

- Tag and probe (TP) is a data-driven method used for measuring particle efficiencies. It is used for obtaining trigger, reconstruction, identification efficiencies. Mainly used for leptons.
- For TP, we need a mass resonance decaying to the object whose efficiency we want to measure (e.g. J/psi, upsilon, Z)
- We select two objects, a tag object and a probe object.
  - Tag object : Tight selection/ID criteria- we assume this is a real object.
  - Probe object: Very loose selection/ID criteria.
- We compute the diobject invariant mass of the tag object + probe object.
  - If the invariant mass is close to the resonance mass value, we assume that the probe object was a real object. Otherwise it should be a fake object.
- We take the real leptons inside the resonance mass window and apply on them the criteria of the selection, whose efficiency we want to measure
- Selection efficiency is computed as

$$\epsilon_{\text{selection}} = \mathrm{N}_{\text{selection}}^{\text{in mass window}} / \mathrm{N}_{\text{total}}^{\text{in mass window}}$$

Validating the estimates

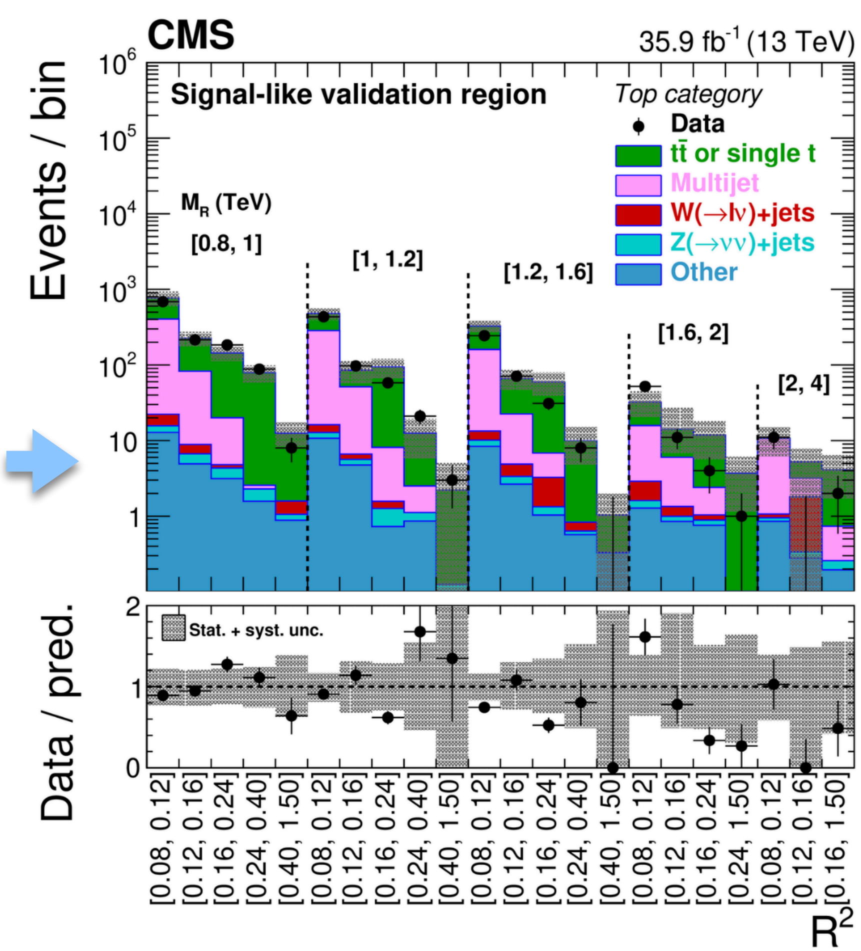
BG estimation methods must always be validated with closure tests or independent validation regions, or alternative methods.

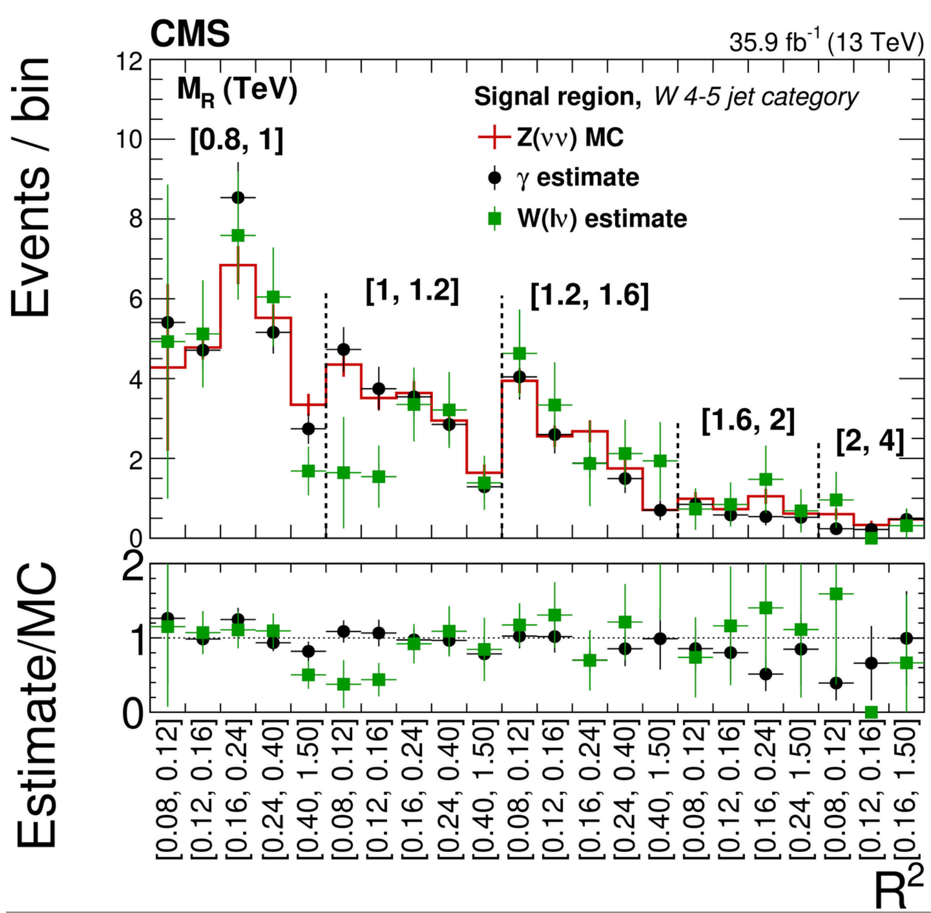
Closure tests : Validate the internal consistency of the method, e.g. validate the method using purely MC events.

Validation regions : Validate the method in independent dedicated regions. These can have a composition similar to the signal regions but be dominated by BG. Estimate should be equal to data.

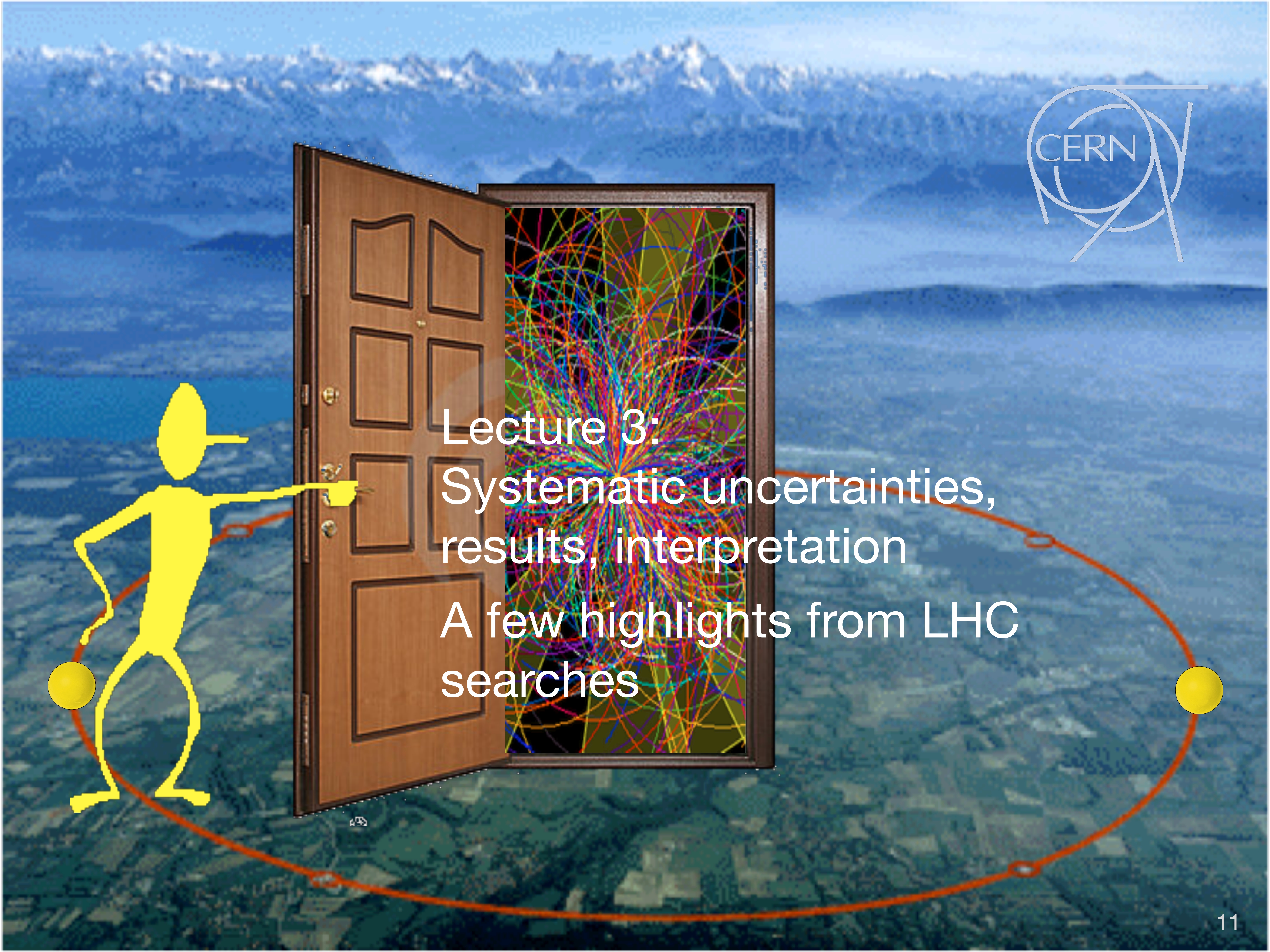
Alternative methods : Estimate the BG with multiple methods and compare the results.

Signal-like validation region from CMS SUSY razor.



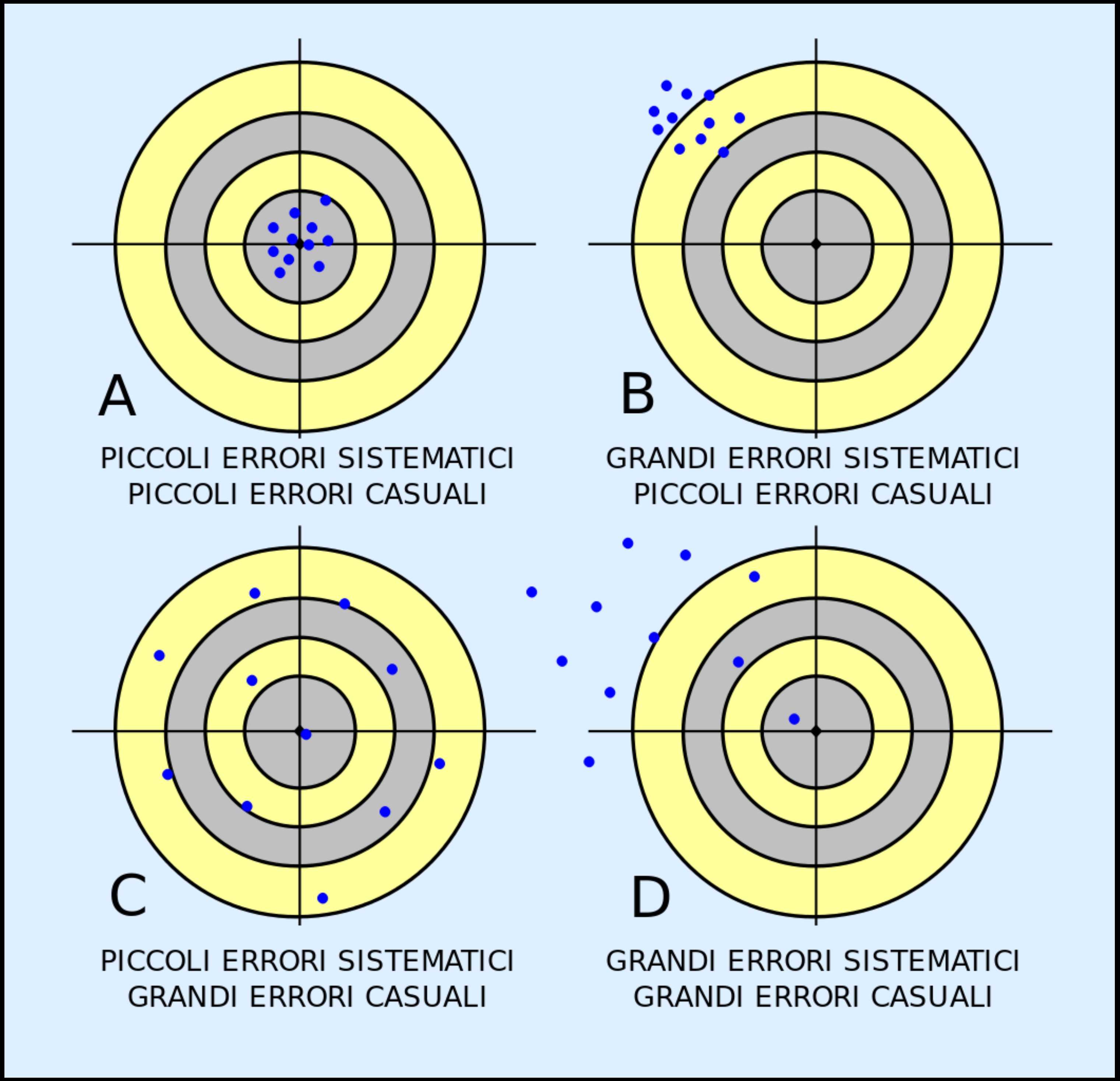


Estimating Z(vv)+jets in 3 different ways from CMS SUSY razor.



Lecture 3:
Systematic uncertainties, results, interpretation

A few highlights from LHC searches



(6) Systematic uncertainties

Statistical vs. systematic

- Statistical uncertainties are the result of stochastic fluctuations arising from the fact that a measurement is based on a finite set of observations.
    - Repeated measurements of the same phenomenon will therefore result in a set of observations that will differ. Statistical uncertainty is a measure of the range of this variation.
- Systematic uncertainties or "systematics" arise from uncertainties associated with the nature of the experiment, assumptions made by the experimenter, or the theoretical model used to make inferences based on the observed data.
    - They cause a shift in the mean of a measurement from the true value due to effects from the experimental setup / theoretical calculations. They are different from statistical uncertainties.
    - They can effect either the normalization or the shape of a given distribution.

Effects of both uncertainties on the measured quantity is included in an experimental result.

Types of systematics

- Experimental uncertainties (coming from the accelerator/detector setup) on:
  - Luminosity measurements
  - Trigger efficiency measurements
  - Jet energy scale, jet energy resolution measurements
  - Lepton, photon, b-jet, W-jet, top-jet, etc. efficiency measurements
  - Uncertainties derived in background estimations, e.g. from closure tests, validation regions or comparison of alternative estimation methods
  - …
- Theoretical uncertainties (due to insufficient accuracy of calculations) on:
  - Cross section and branching ratio calculations
  - Parton distribution functions
  - Initial state / final state radiation, renormalization / factorization scale effects
  - …

When using MC in an analysis, systematics must be applied on the MC events to reflect the uncertainties in the predictions.

Types of systematics

Source	Signal	Lost lepton	1ℓ (not from t)	$Z \to \nu\bar{\nu}$
Data statistical uncertainty	—	5–50%	4–30%	—
Simulation statistical uncertainty	6–36%	3–68%	5–70%	4–41%
$t\bar{t}$ p_T^{miss} modeling	—	3–50%	—	—
Signal p_T^{miss} modeling	1–25%	—	—	—
QCD scales	1–5%	0–3%	2–5%	1–40%
Parton distribution	—	0–4%	1–8%	1–12%
Pileup	1–5%	1–8%	0–5%	0–7%
Luminosity	2.3–2.5%	—	—	2.3–2.5%
$W + b(\bar{b})$ cross section	—	—	20–40%	—
$t\bar{t}Z$ cross section	—	—	—	5–10%
System recoil (ISR)	1–13%	0–3%	—	—
Jet energy scale	2–24%	1–16%	1–34%	1–28%
p_T^{miss} resolution	—	1–10%	1–5%	—
Trigger	2–3%	1–3%	—	2–3%
Lepton efficiency	3–4%	2–12%	—	1–2%
Merged t tagging efficiency	3–6%	—	—	5–10%
Resolved t tagging efficiency	5–6%	—	—	3–5%
b tagging efficiency	0–2%	0–1%	1–7%	1–10%
Soft b tagging efficiency	2–3%	0–1%	0–1%	0–5%

Example set of systematic uncertainties and their effects on expected signal yields and predicted background yields from a CMS SUSY analysis.

(8) Results and (9) Interpretation (including statistical model (7))

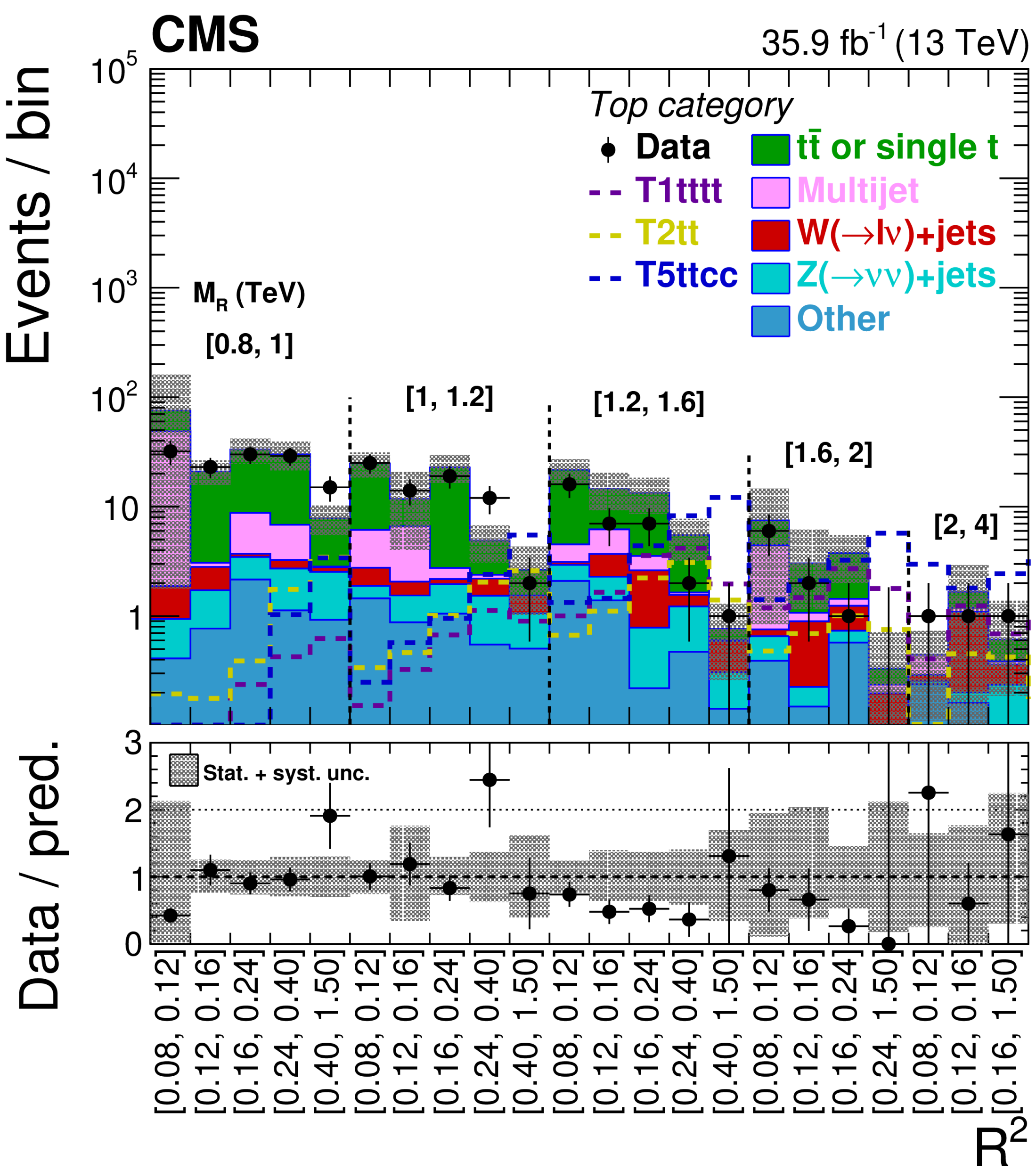
What do we measure?

An experimental result is the empirical outcome of the experiment, the measurement of some physical quantity, such as:

- Event counts <— mainly in new physics search analyses.
- Masses
- Cross sections/branching ratios
- Signal strengths
- Couplings
- Kinematic shapes, peaks, edges, endpoints
  (usually to derive masses or mass differences)
- Decay widths
- Charge asymmetry
- Spin correlations
- etc.

Experimental results

Event counts



The results of the SUSY razor search consist of event counts in 2-dimensional bins of razor variables M_R and R^2.

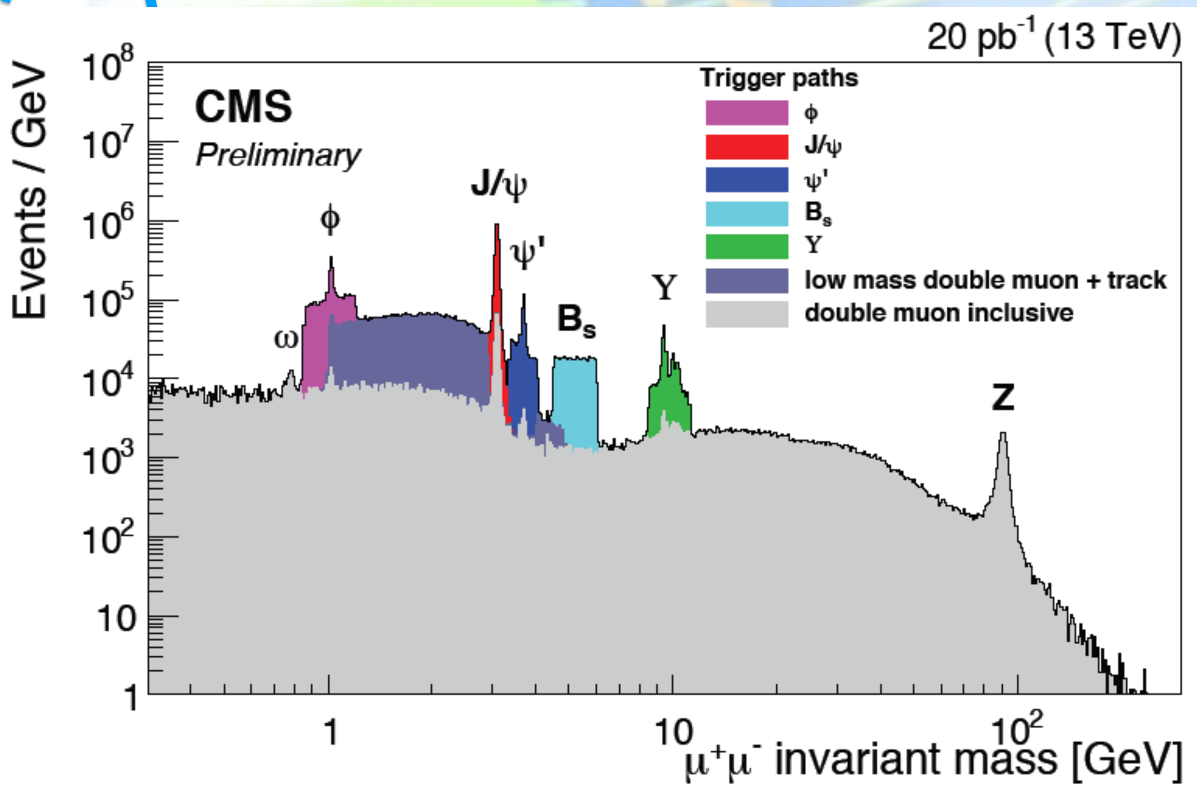
- Observed data counts
- Estimated yields from 4 different backgrounds.
- Systematic uncertainties on the estimated BG yields (gray shaded band).

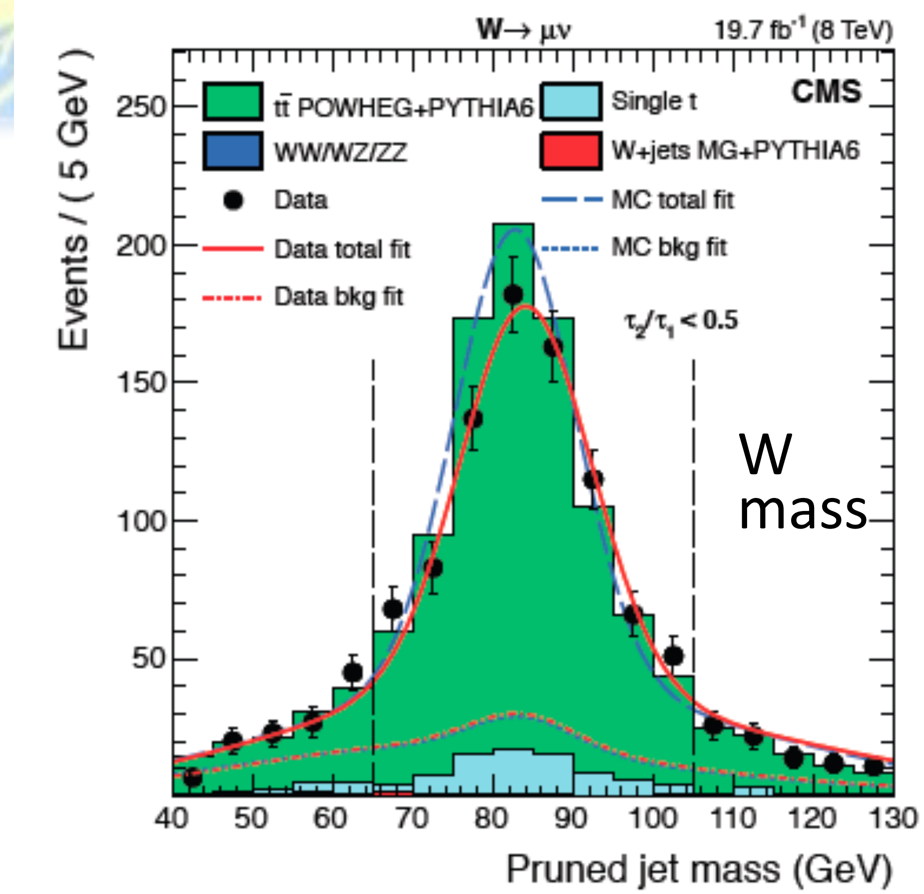
In a search analysis, we compare observed data and estimated BG yields to see if there is agreement or discrepancy.

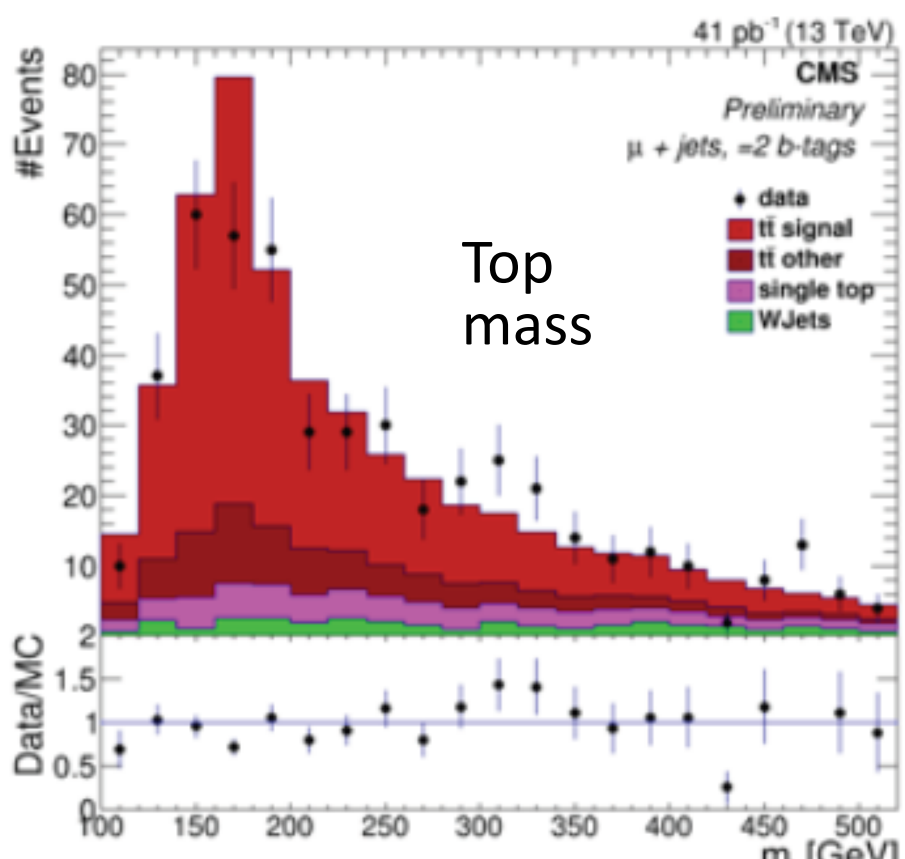
We quantify this with a statistical analysis.

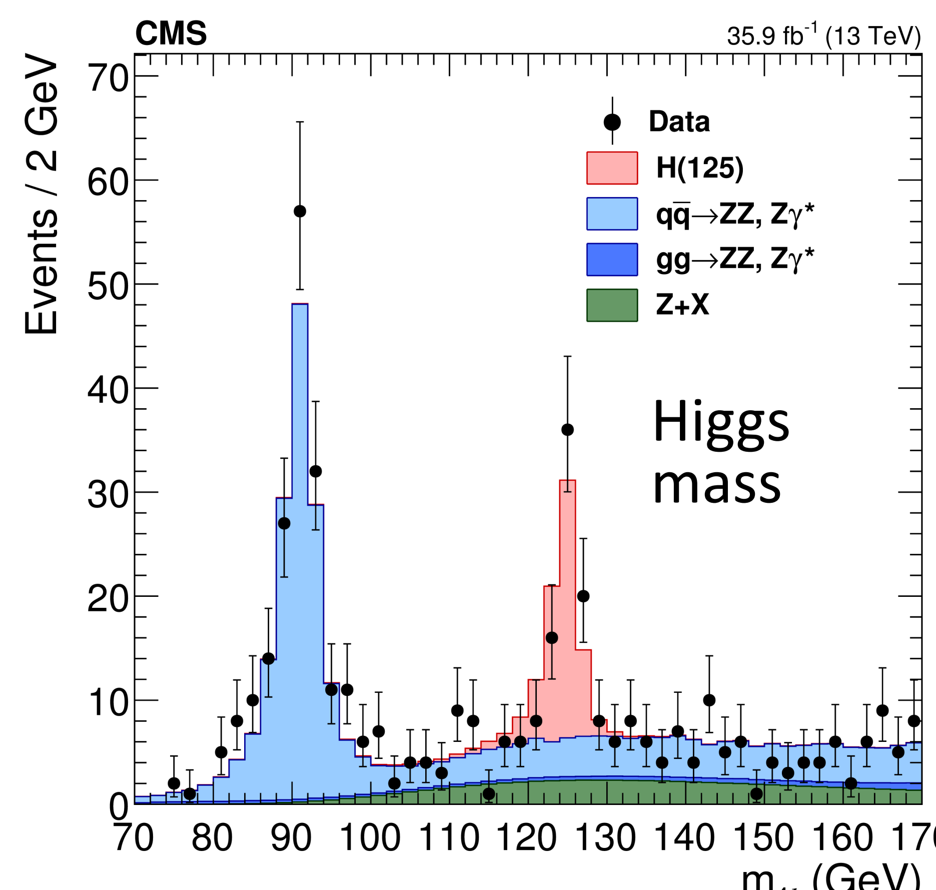
It is customary to show expected / MC distributions from a few signal points, but these are not a part of the results.

Experimental results
Mass



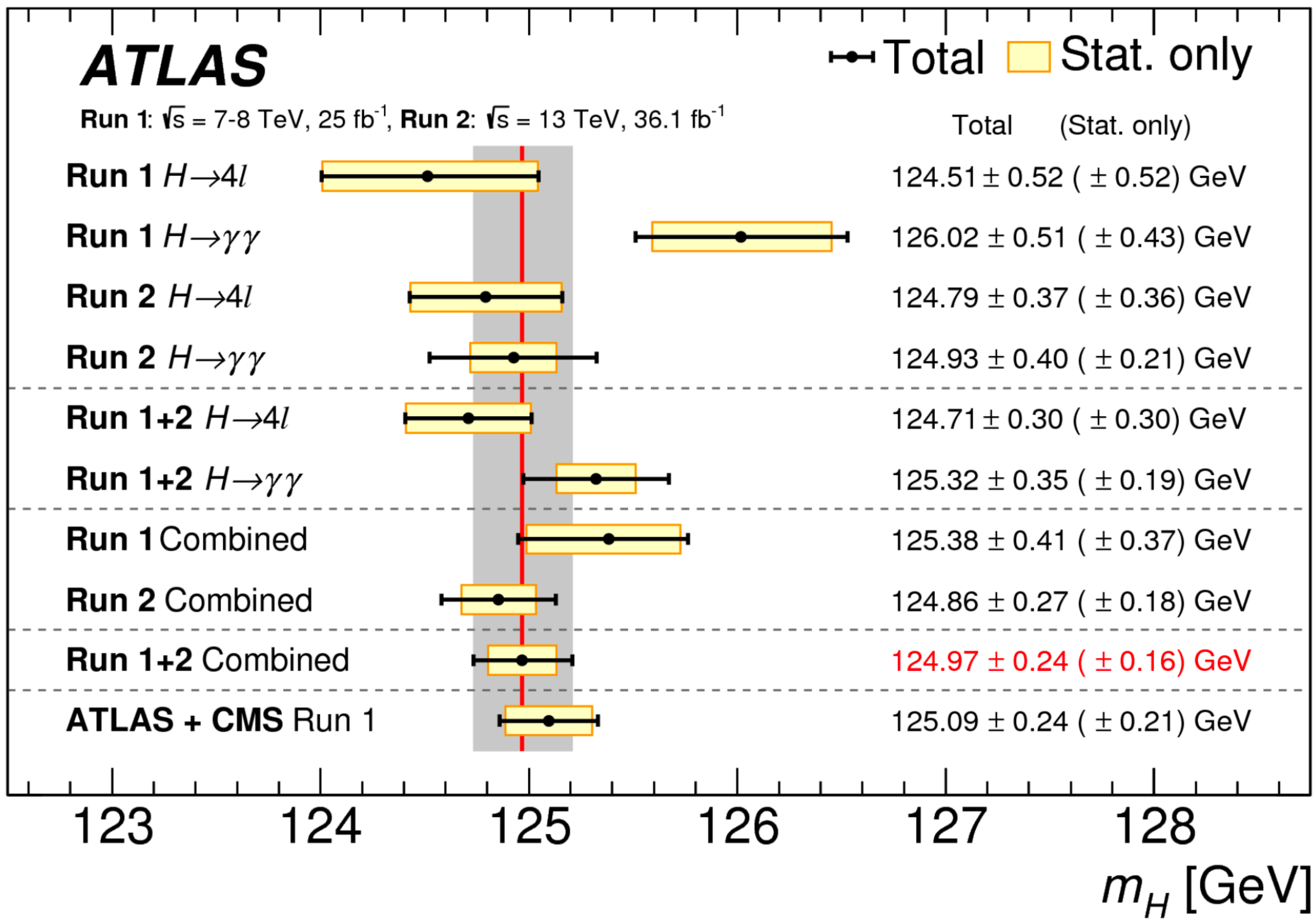






Experimental results

Higgs mass measurements



Cross section measurements

For a signal process, the number of expected signal events is given by

$$N_{signal} = \sigma_{signal} \cdot \int \mathcal{L}dt \cdot \epsilon_{signal} = N^{observed} - N_{BG}$$

where ε_{signal} is the product of all branching ratios, geometrical and kinematical acceptance, efficiencies for trigger, object reconstruction and identification.

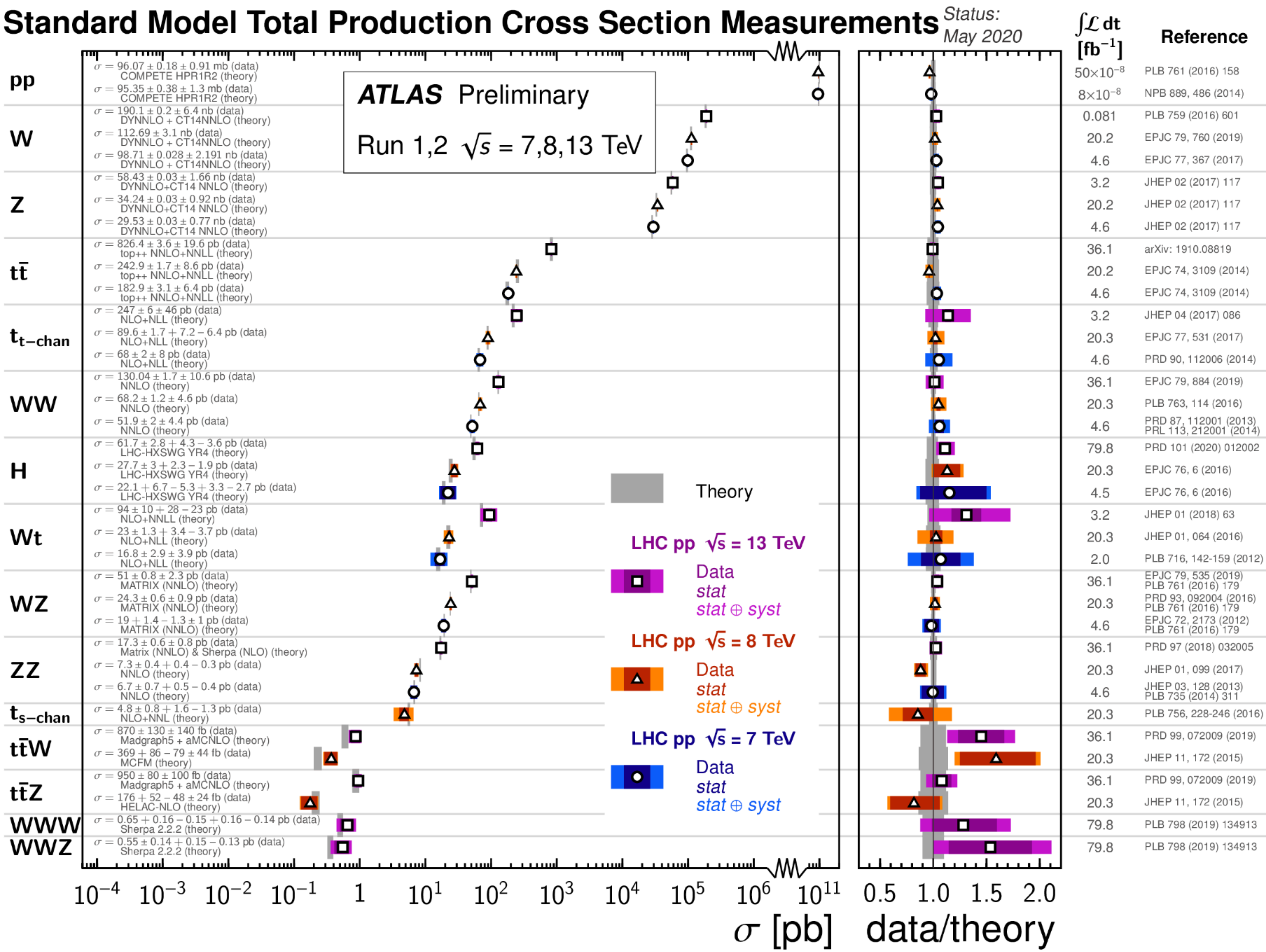
The cross section can be obtained by reverting the equation:

$$\sigma_{signal} = \frac{N^{observed} - N_{BG}}{\int \mathcal{L}dt \cdot \epsilon_{signal}}$$

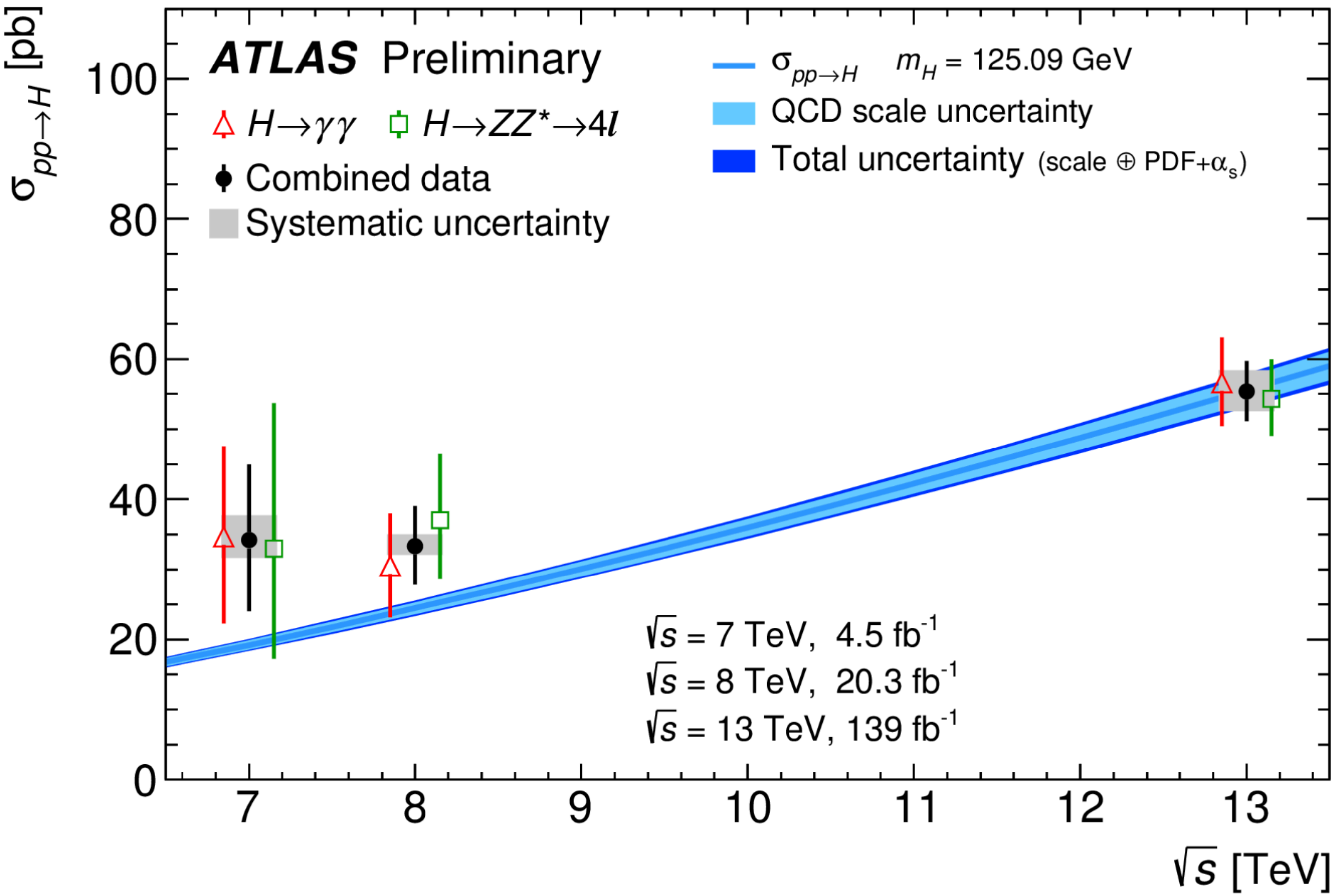
Cross sections for SM processes are continuously being refined at the LHC.

Experimental results

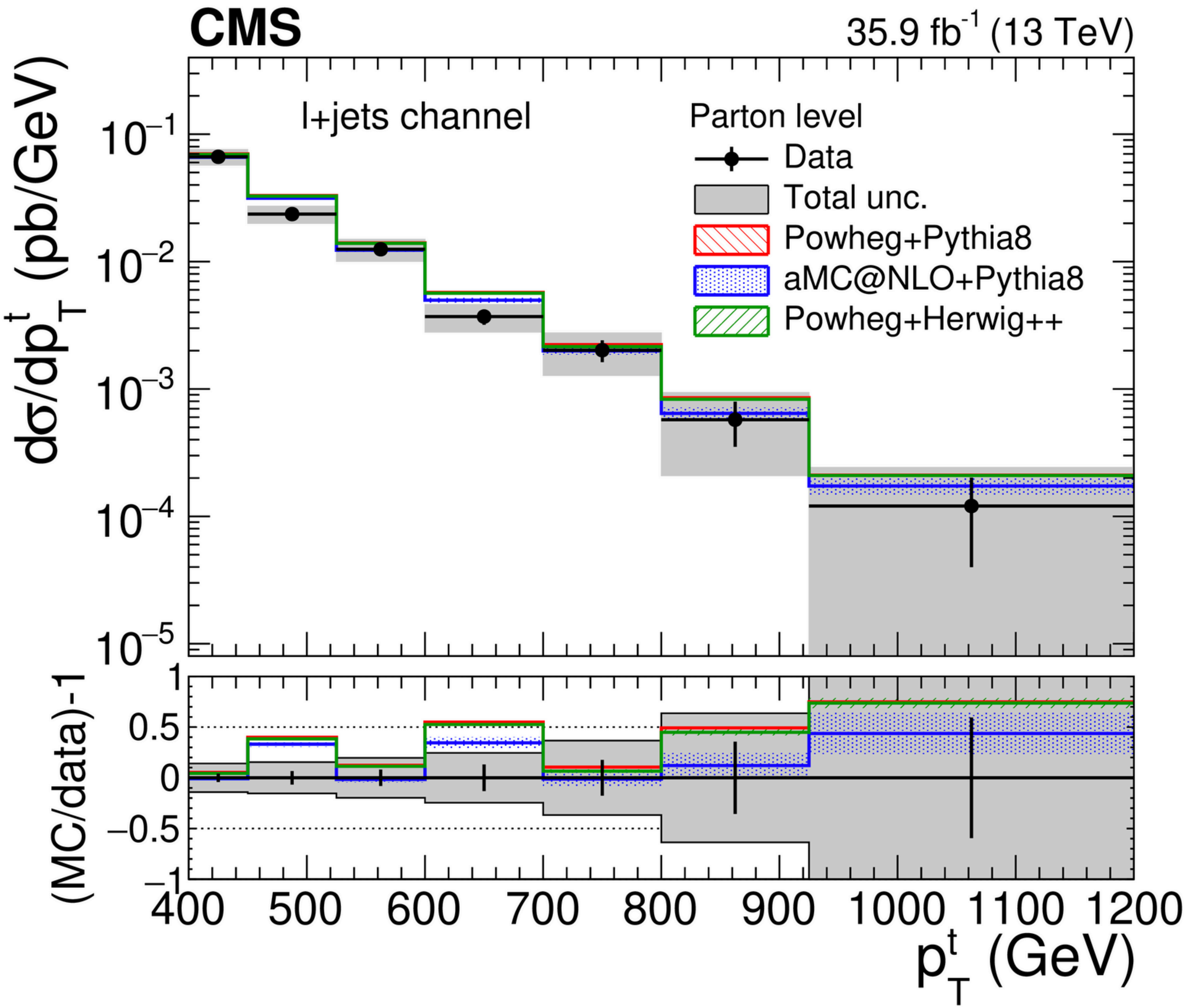
SM cross section measurements



Higgs cross section measurements



Differential ttbar cross section measurement



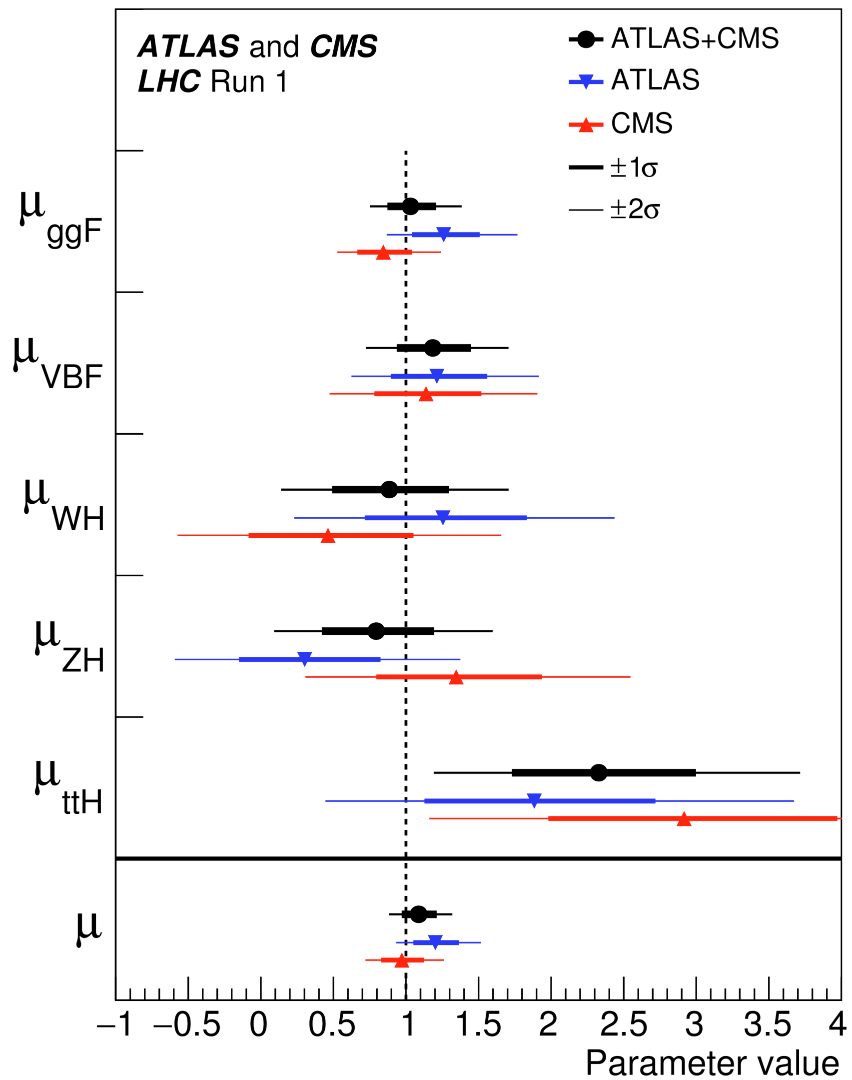
Experimental results
Signal strength

Signal strength in a search region:

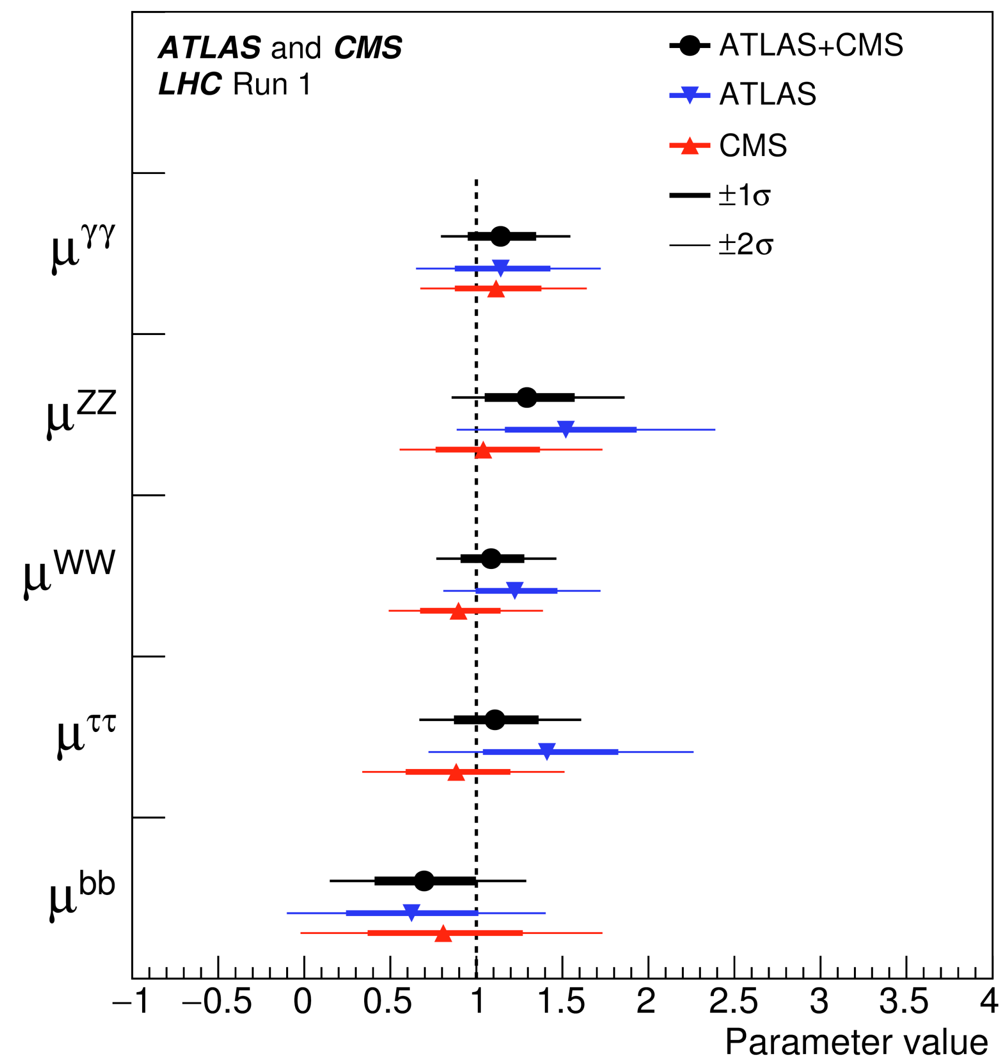
$$\mu = \frac{\text{Observed event yield for a process}}{\text{Expected SM event yield for the process}}$$

μ = 0: no signal
μ = 1: signal consistent with the SM
μ > 1: signal different from the SM —> BSM

Higgs cross sections



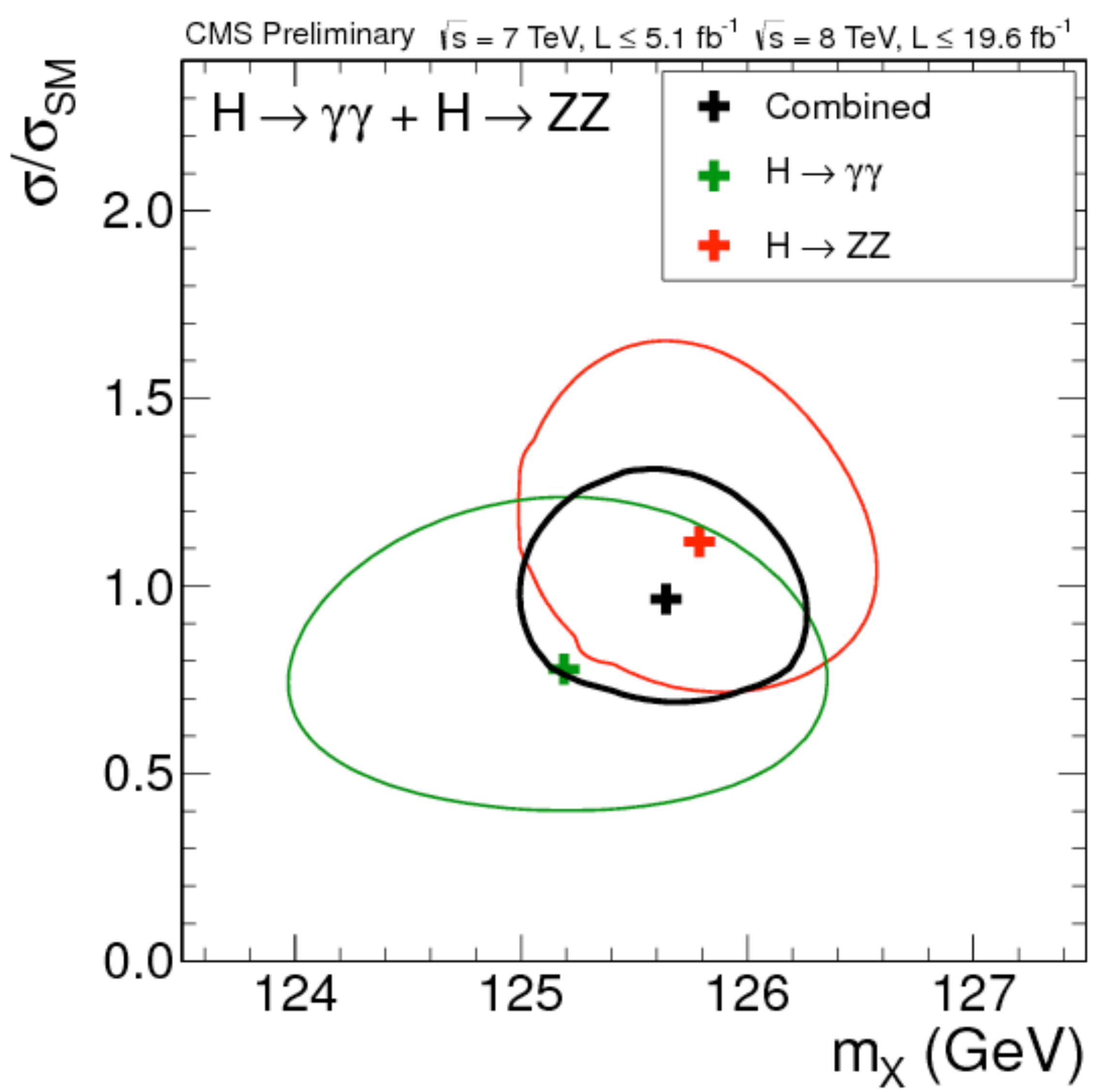
Higgs branching ratios



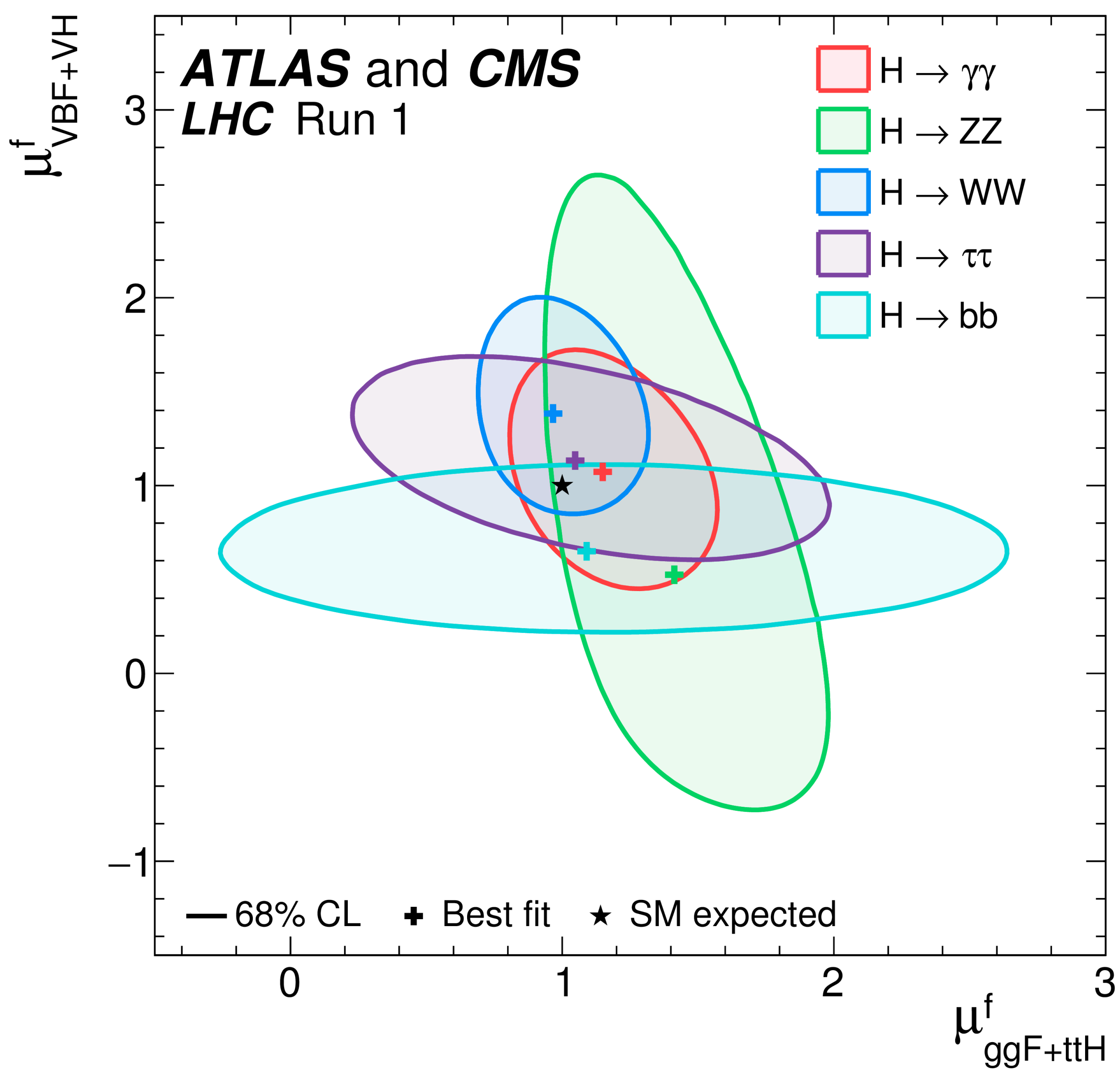
Experimental results

Signal strength - let's get fancier

Fitting the signal strength and the Higgs mass together to find their common best values.

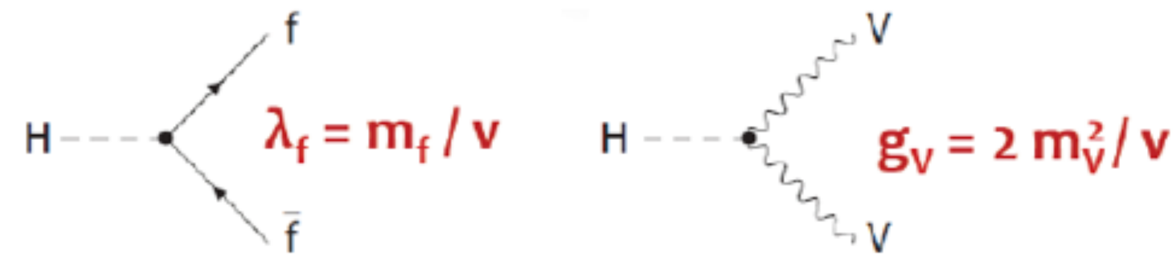


Comparing production signal strengths for different decay channels.

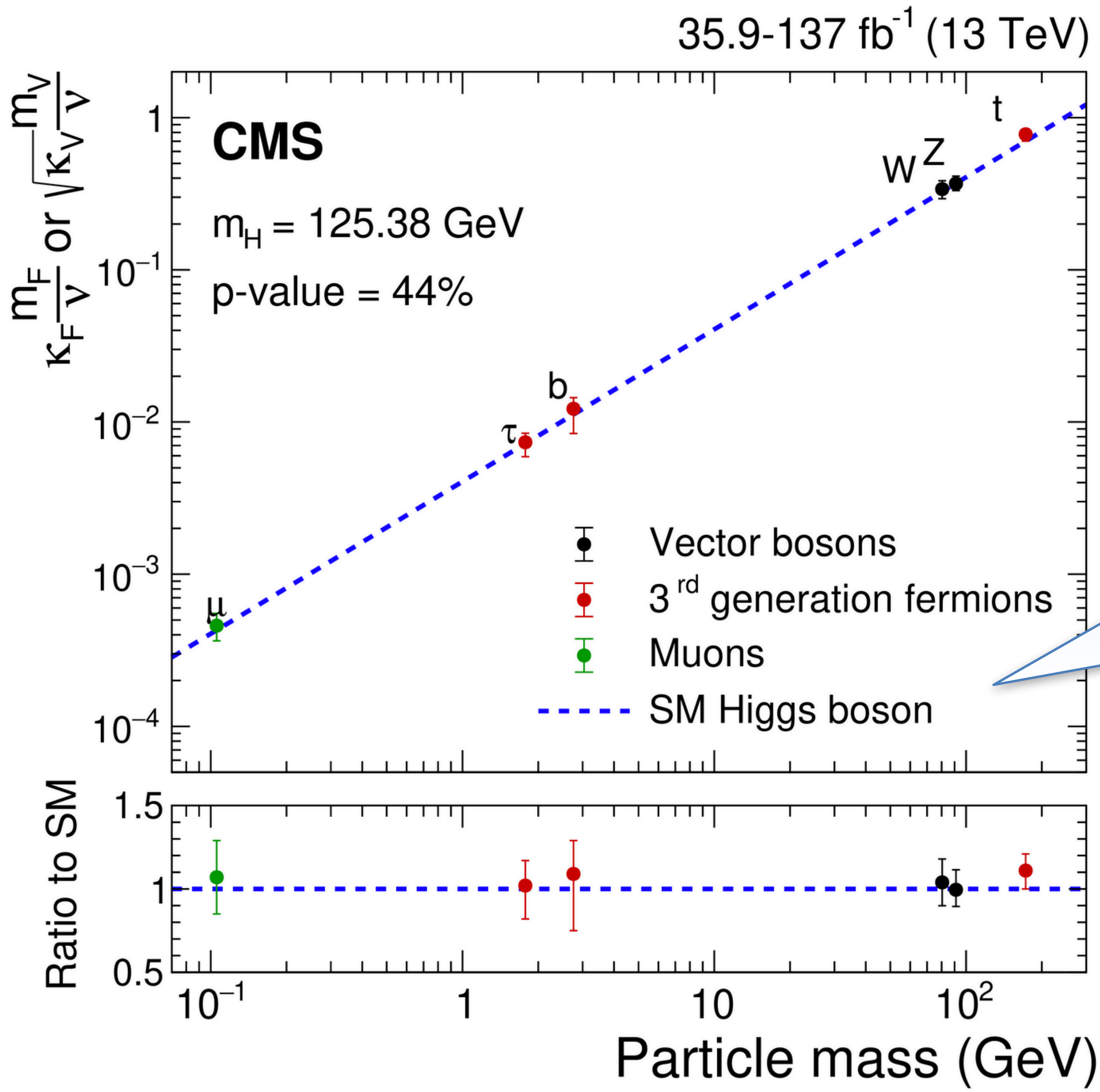


All these help to check the consistency of the discovered Higgs with the SM and and see if there are any hints for BSM

Experimental results
Couplings



v: Vacuum expectation value



Particles that interact more with the Higgs field are more massive.

Couplings of SM particles to the Higgs field versus the SM particles' masses.

A deviation from the SM expectation (blue line) would have been a sign of new physics.

Interpretation
The idea

- Interpretation is the comparison of experimental results with the expectations of a given a theoretical model.
- CAUTION! Interpretation is NOT the experimental result.
- We use a statistical model and likelihood to interpret the experimental result.
  - The statistical model of an analysis provides the complete mathematical description of that analysis.
  - It relates the observed quantities x to the theoretical model's parameters θ_i through the probability density $p(x|\theta_i)$. Parameters θ_i include signal model parameters θ_s and background model parameters θ_b.
  - The likelihood $L(\theta_i) = p(X_0|\theta_i)$ is the probability density $p(x|\theta_i)$ evaluated at the observed values X_0 of the observables x.
  - The likelihood is the starting point of any interpretation.
- We estimate parameters θ using statistical procedures, and test the validity of the model.
- An experimental result can be interpreted with multiple theoretical models.

Limits

Limits: Suppose we are testing a theory with a free parameter θ by an experimental analysis. For example, this parameter could be the mass m of a particle.

Suppose our analysis did not observe a sign of this particle. Data look consistent with the background hypothesis. What can we say about the theory and m?

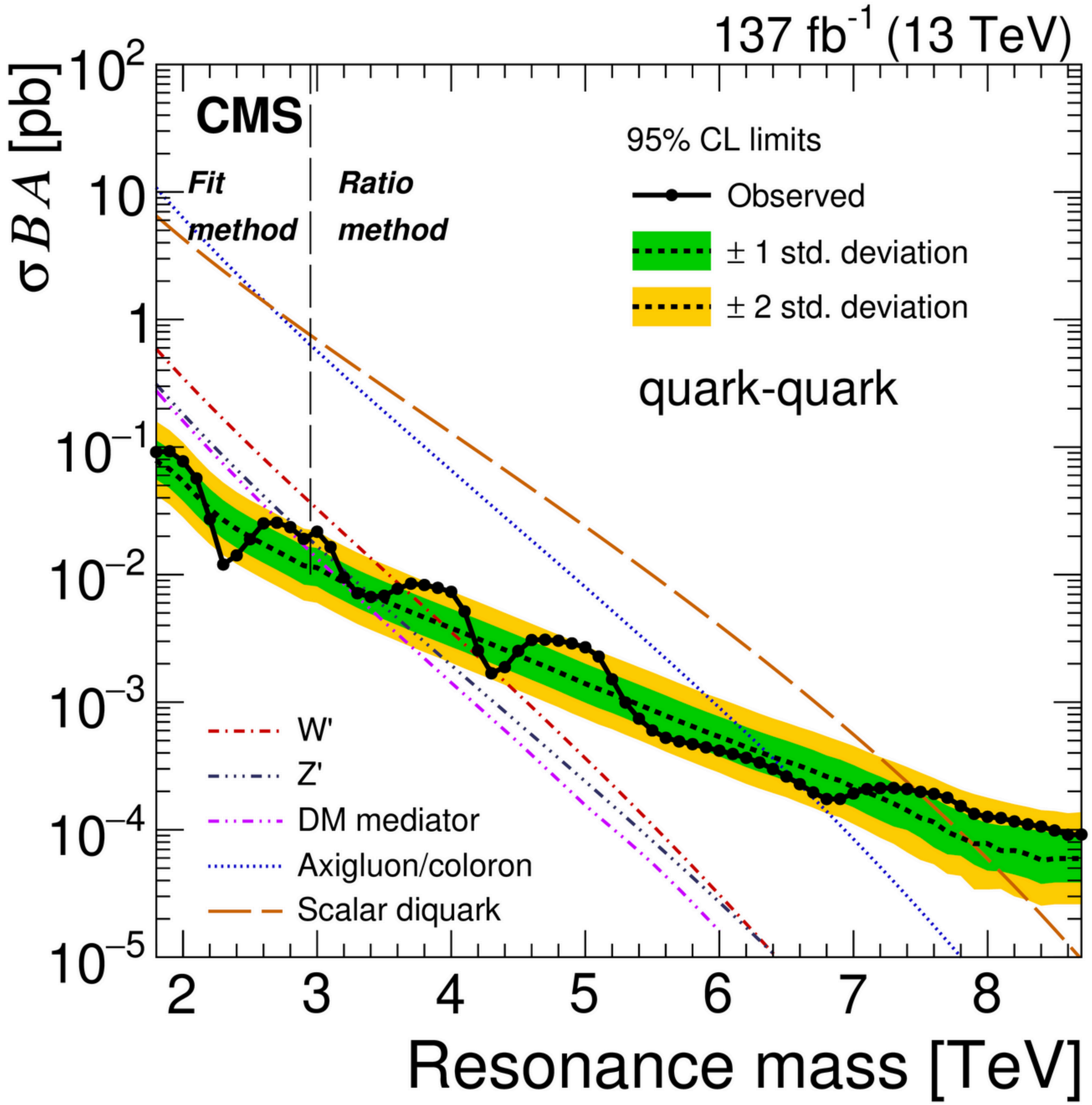
We can say up to which value of m our analysis could have observed m, if the theory were true. This is called a limit.

For a theory with free parameters θ_i, a limit establishes the boundaries defining the range where the experiment can make a statement about the theory.

Limits are obtained using elaborate statistical methods based on likelihoods.

- Observed limit: Obtained by comparing observed data with 1) signal MC + estimated BG and 2) with only estimated BG.
  - Checks the consistency of the observation with the signal + BG hypothesis and compares it to the BG-only hypothesis.
- Expected limit: Obtained by comparing estimated BG with signal + estimated BG.
  - Useful for predicting the analysis sensitivity.

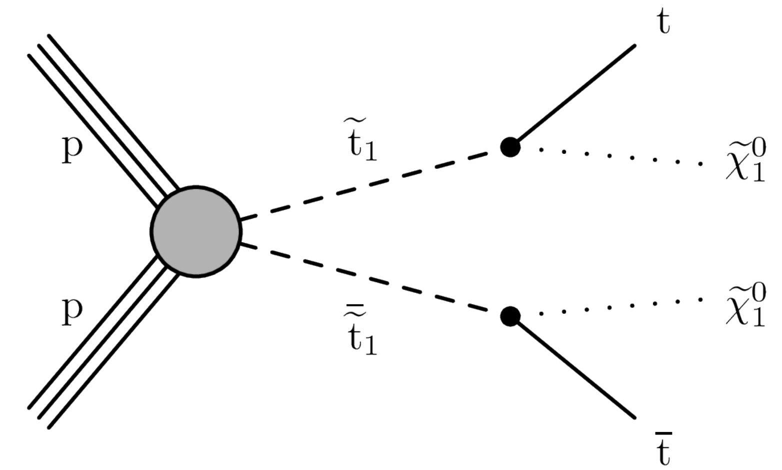
Limits for 1 free parameter

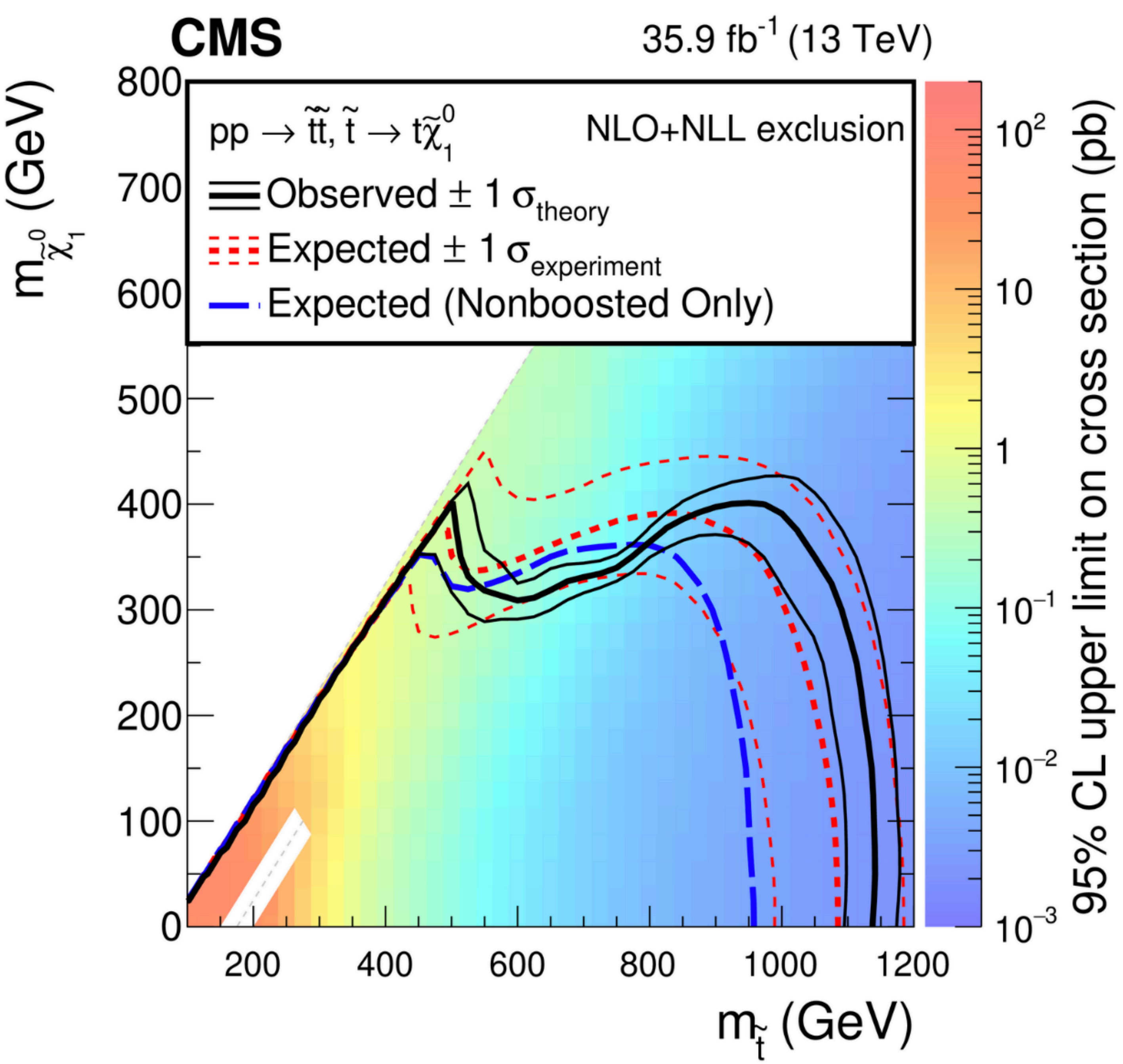


CMS search for high mass dijet resonances.

Limits for 2 free parameters

Interpreting SUSY razor analysis results in terms of a simplified SUSY model, with 2 particles with free masses and cross sections.

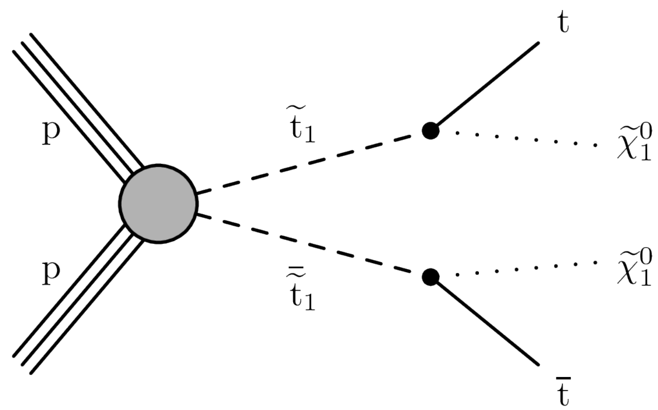


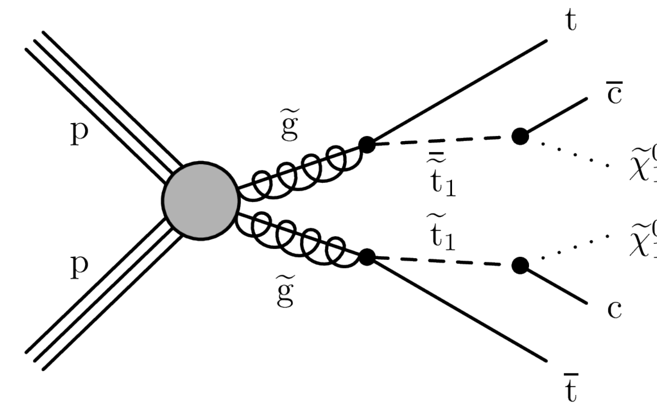


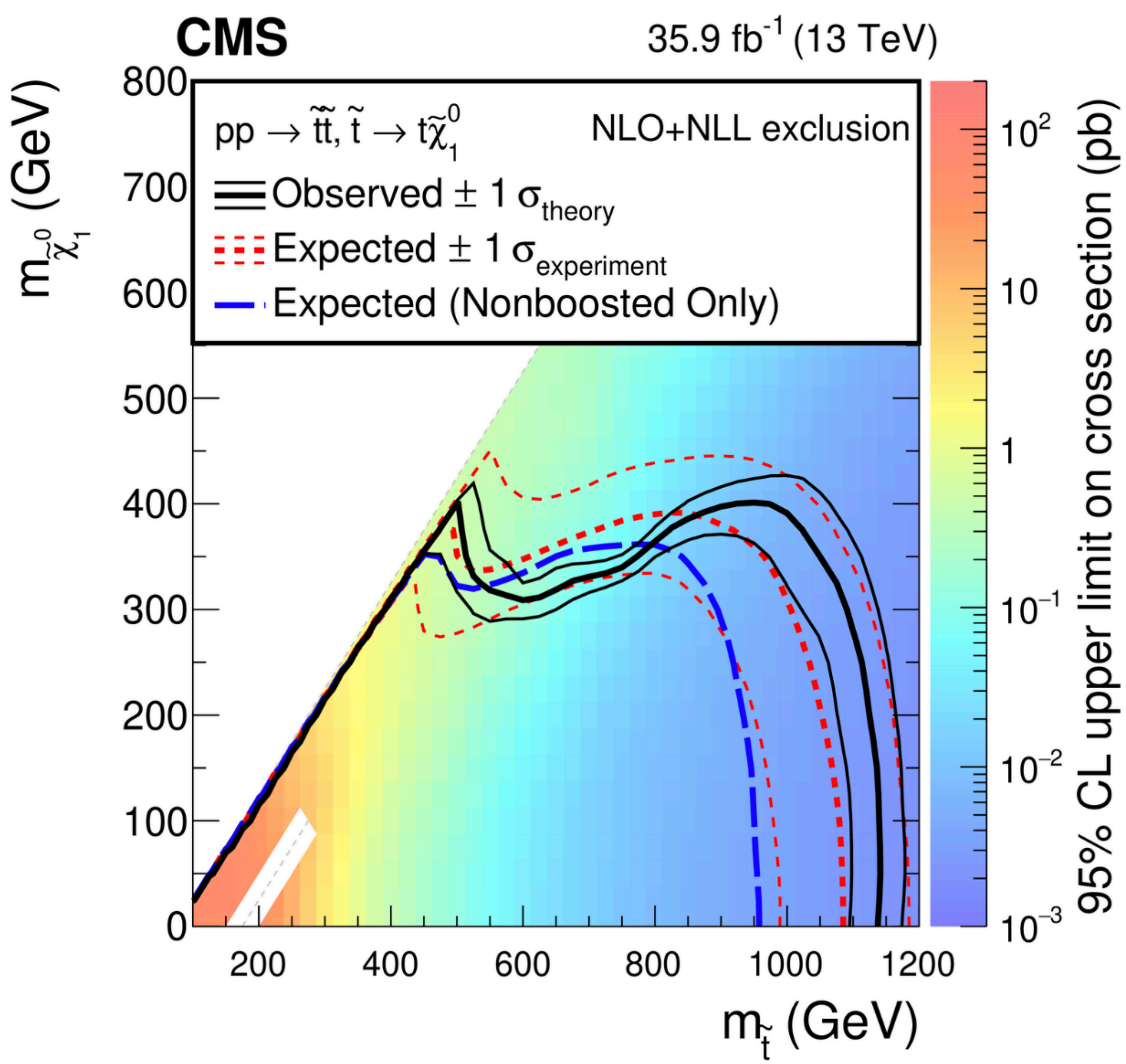
Exactly like the 1 free parameter case on the previous page, but for 2 free parameters!

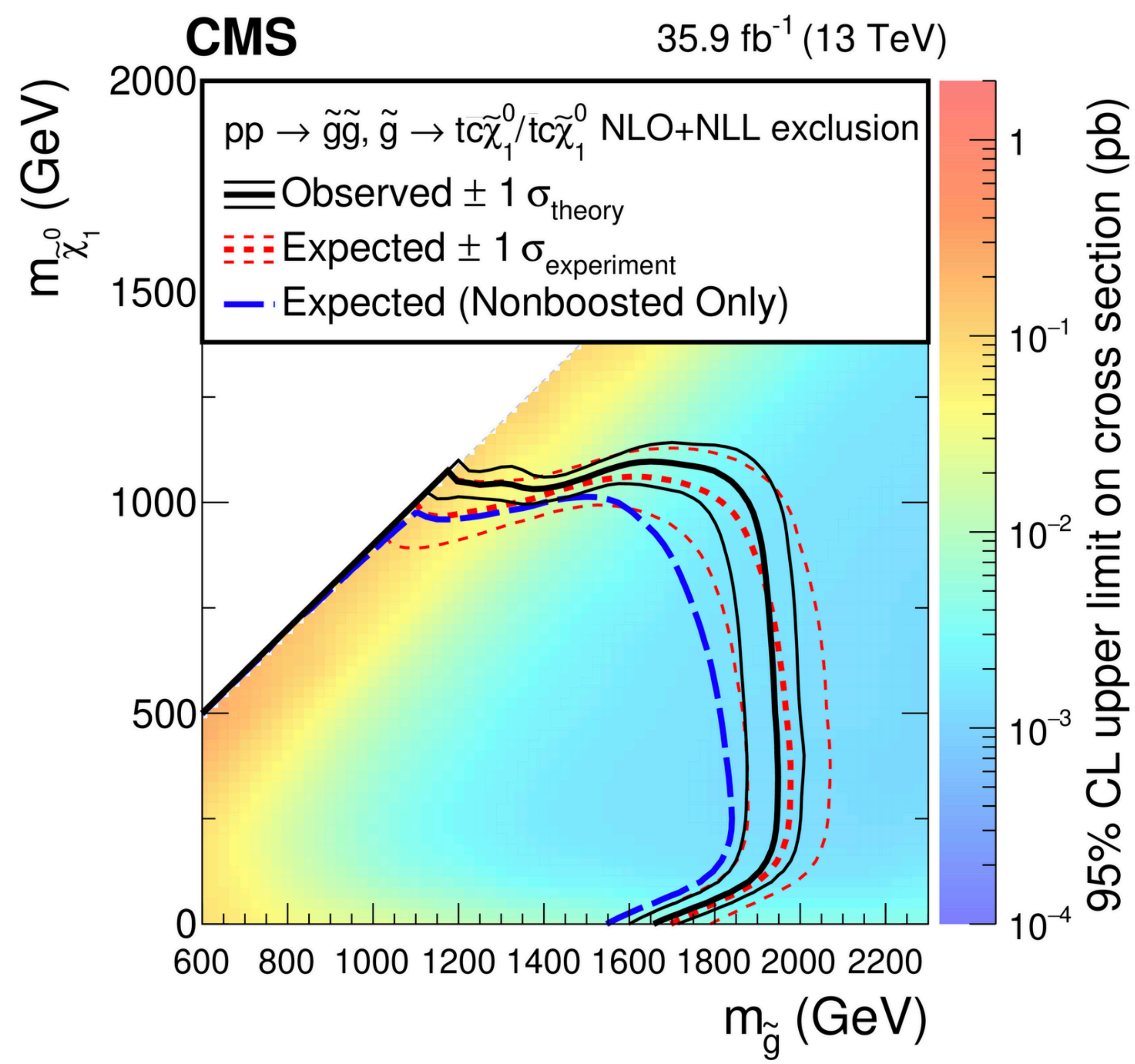
Interpreting with multiple models

SUSY razor analysis results interpreted for multiple theoretical models:





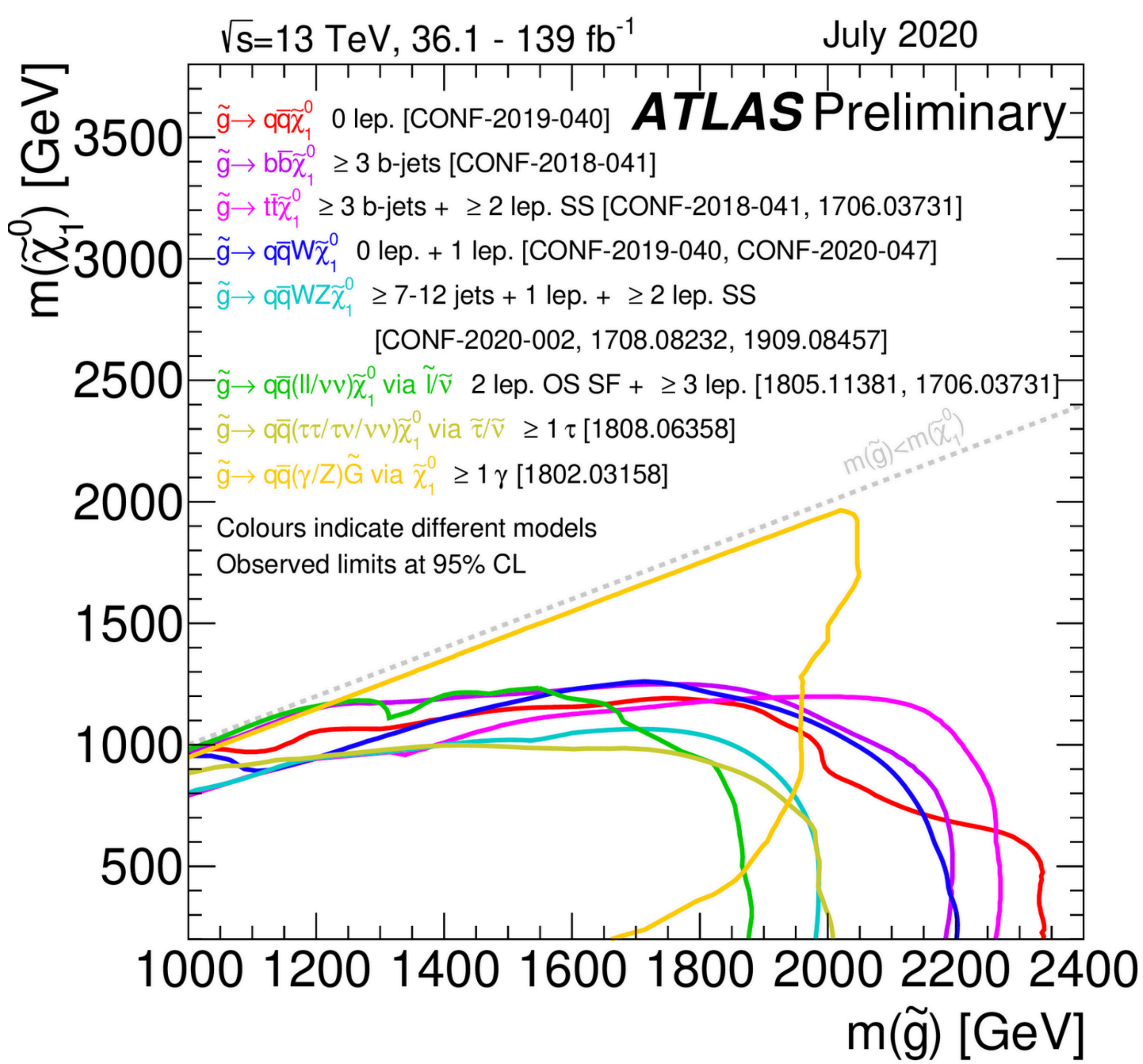


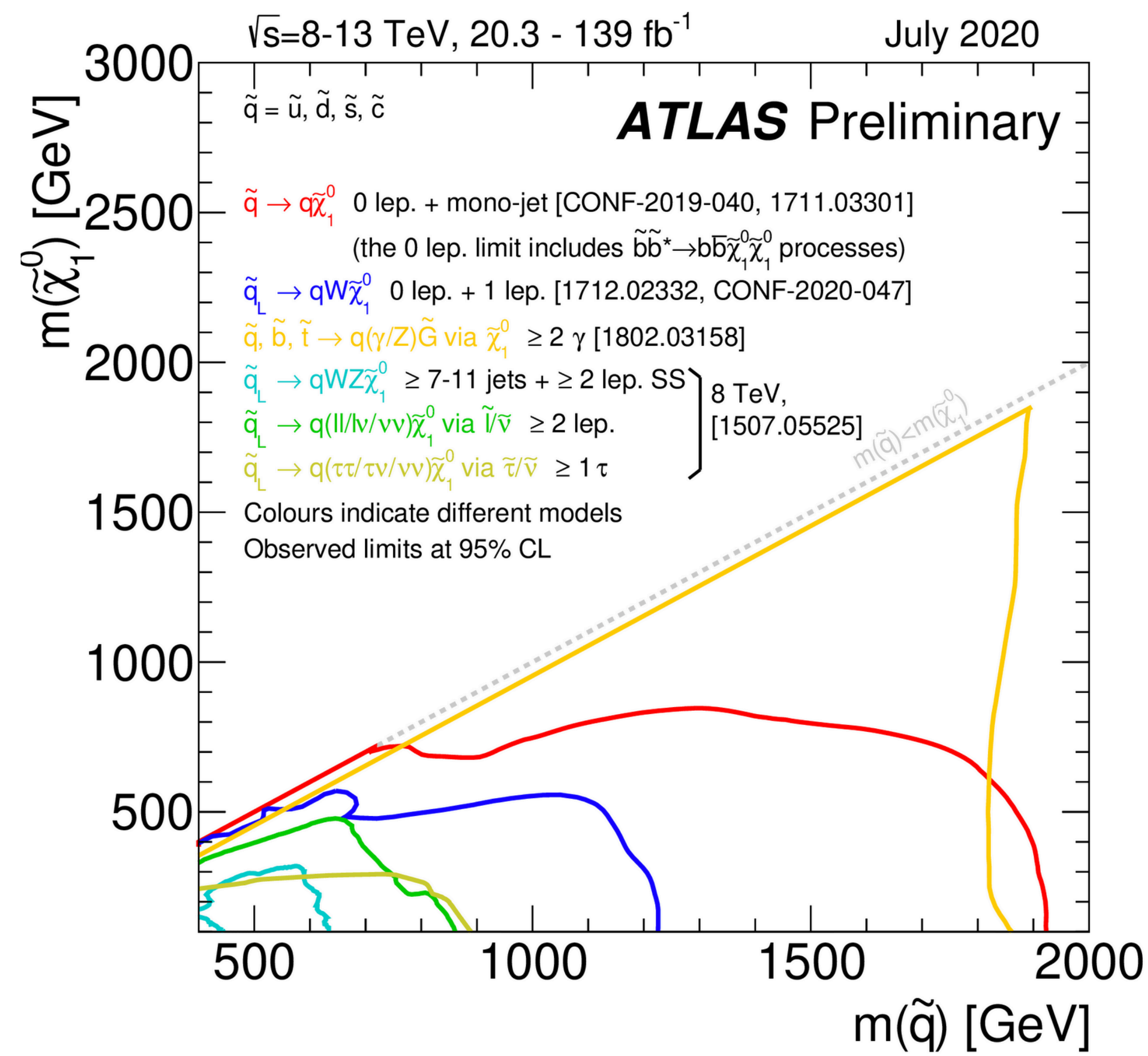


Interpretation
SUSY mass limits

Mass limits on SUSY particles gluinos, squarks in various decay channels to neutralinos. Interpretations were based on several analyses.

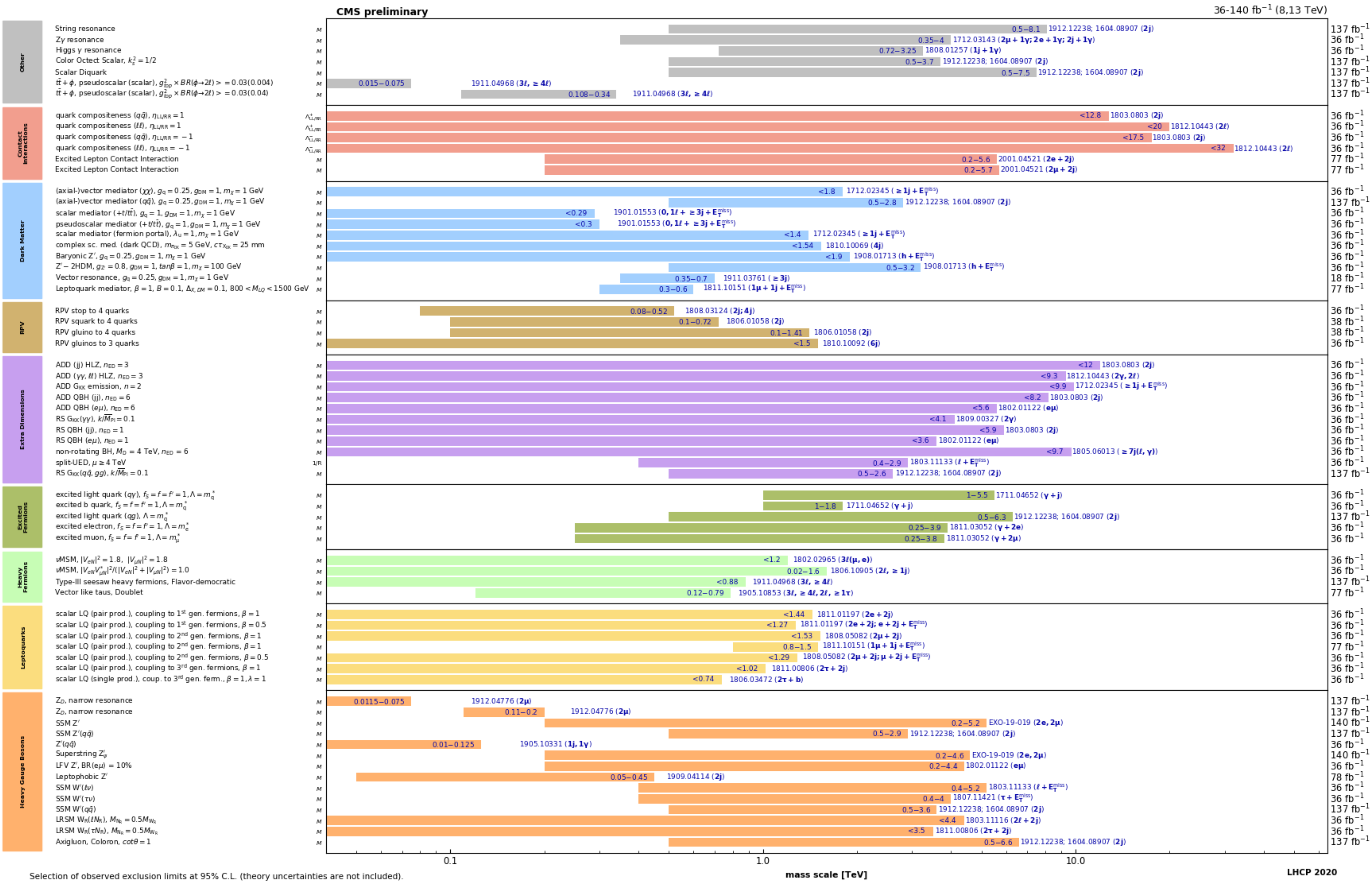
Interpreting the same model with multiple analyses provides complementarity.





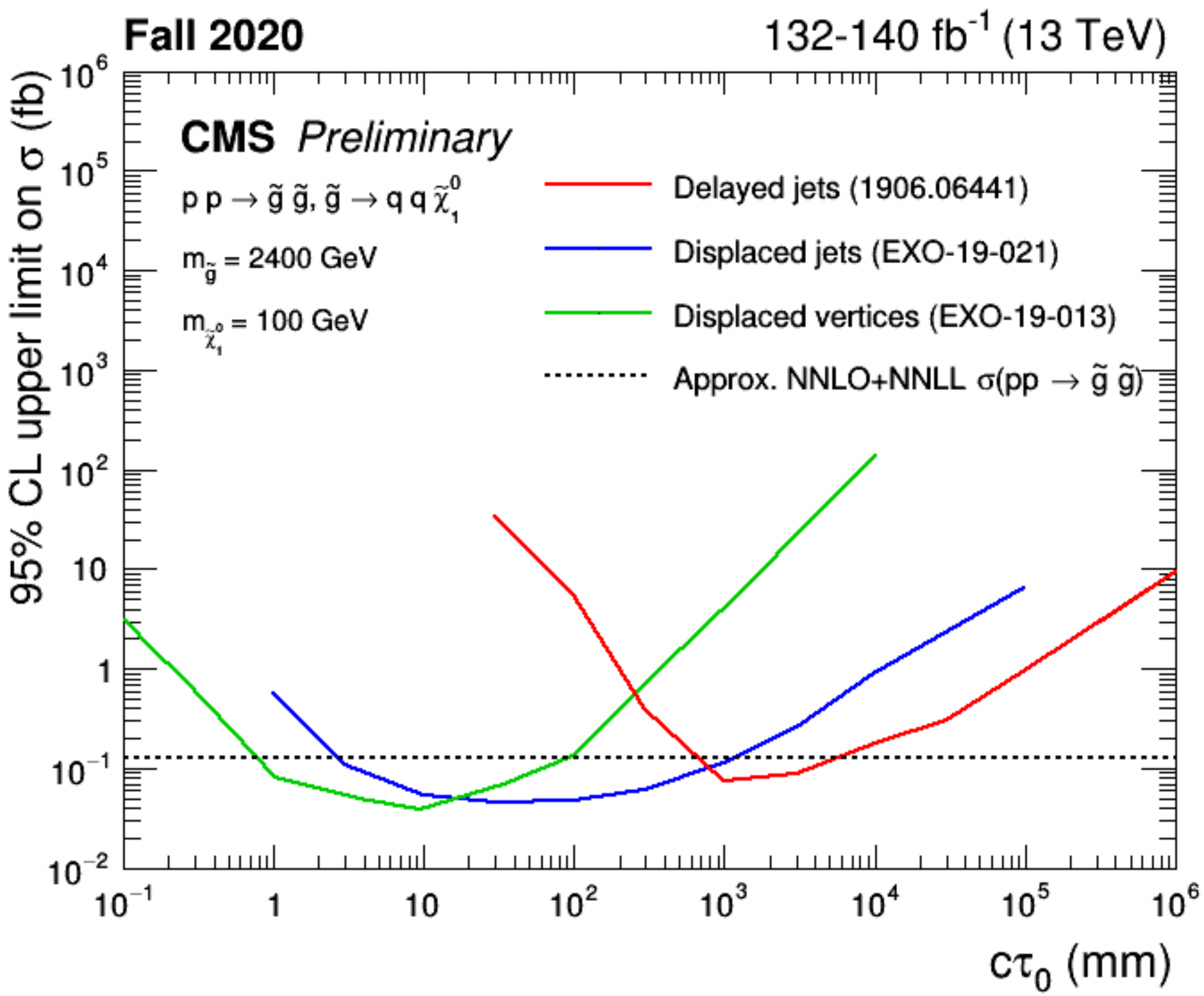
Interpretation
Limits on exotic models

Overview of CMS EXO results



Interpretation

Limits on long-lived particles

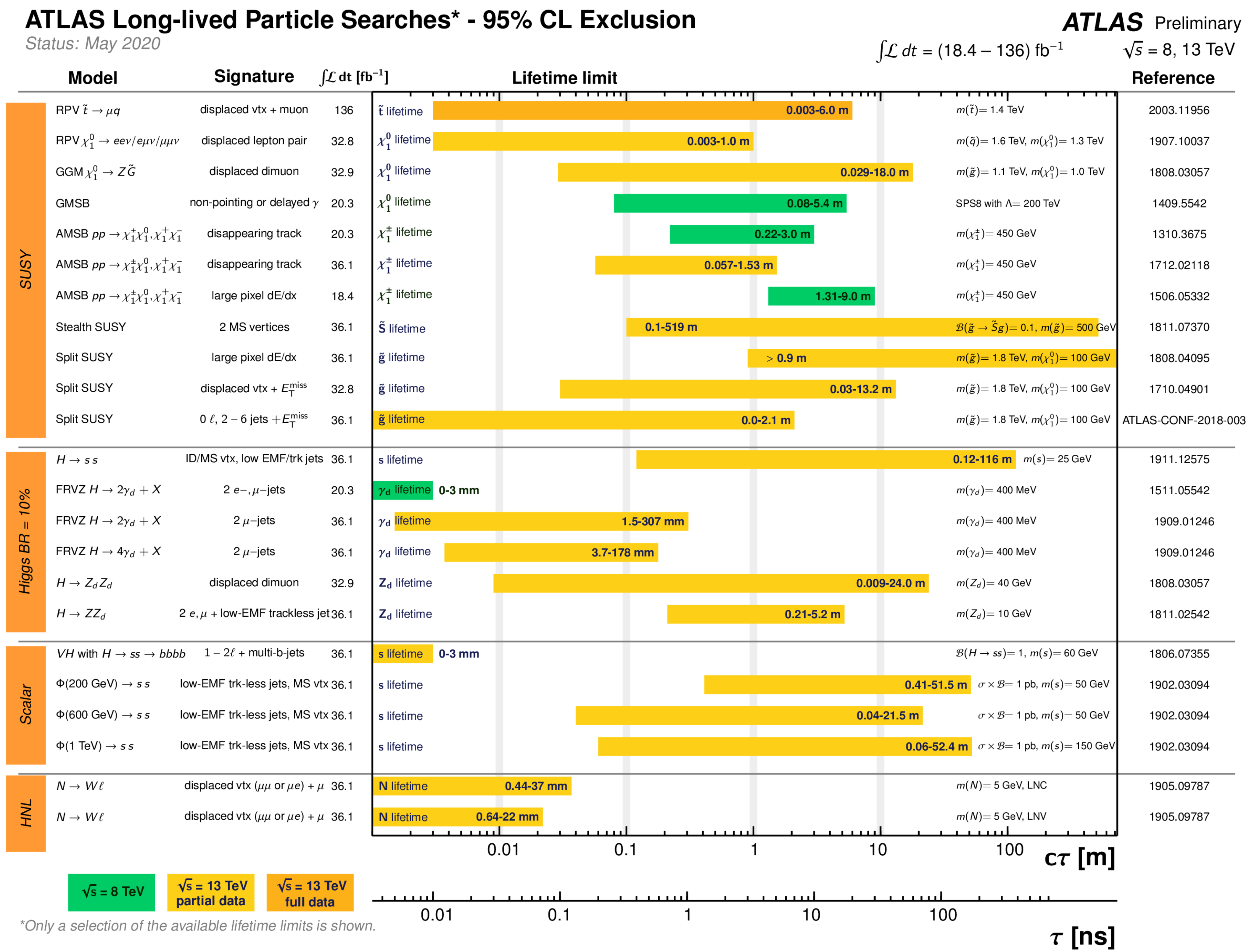


Now the free parameter is the long-lived particle's proper lifetime. Particle masses are fixed.

Different searches exclude different lifetimes, providing complementary information.

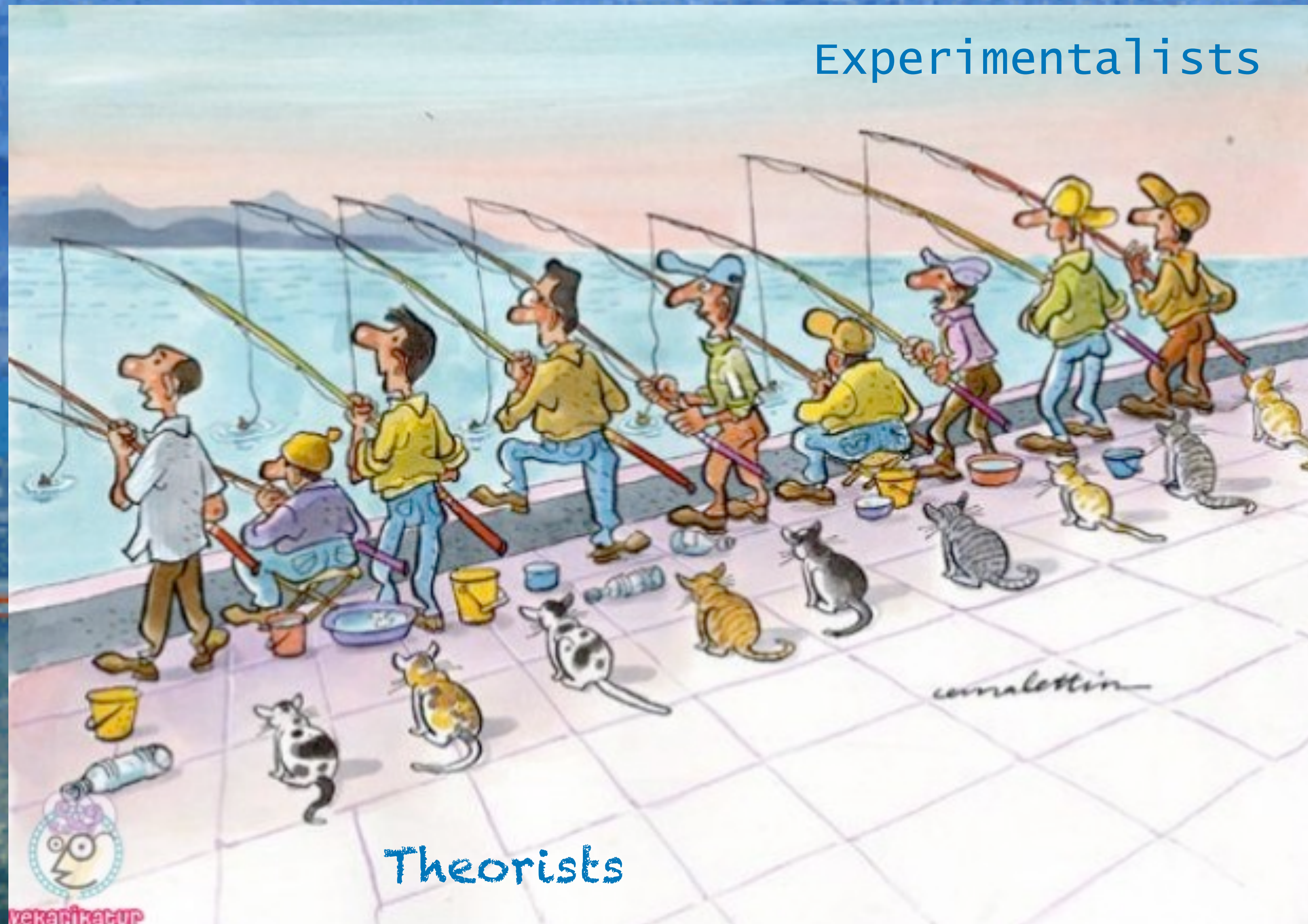
Interpretation

Limits on long-lived particles



"Data are coming! Data are coming!"

Actually, it is the other way around…



What we learn from the absence of discovery and excluding many models takes us closer to discovering the true nature of the Universe.

Hands-on exercises 1 and 2: Looking into signal and background events, plotting variable distributions, applying simple event selections.