University of Colorado Boulder

Reproducing a CMS higgsino search from public data

Bill Ford

December 13, 2022

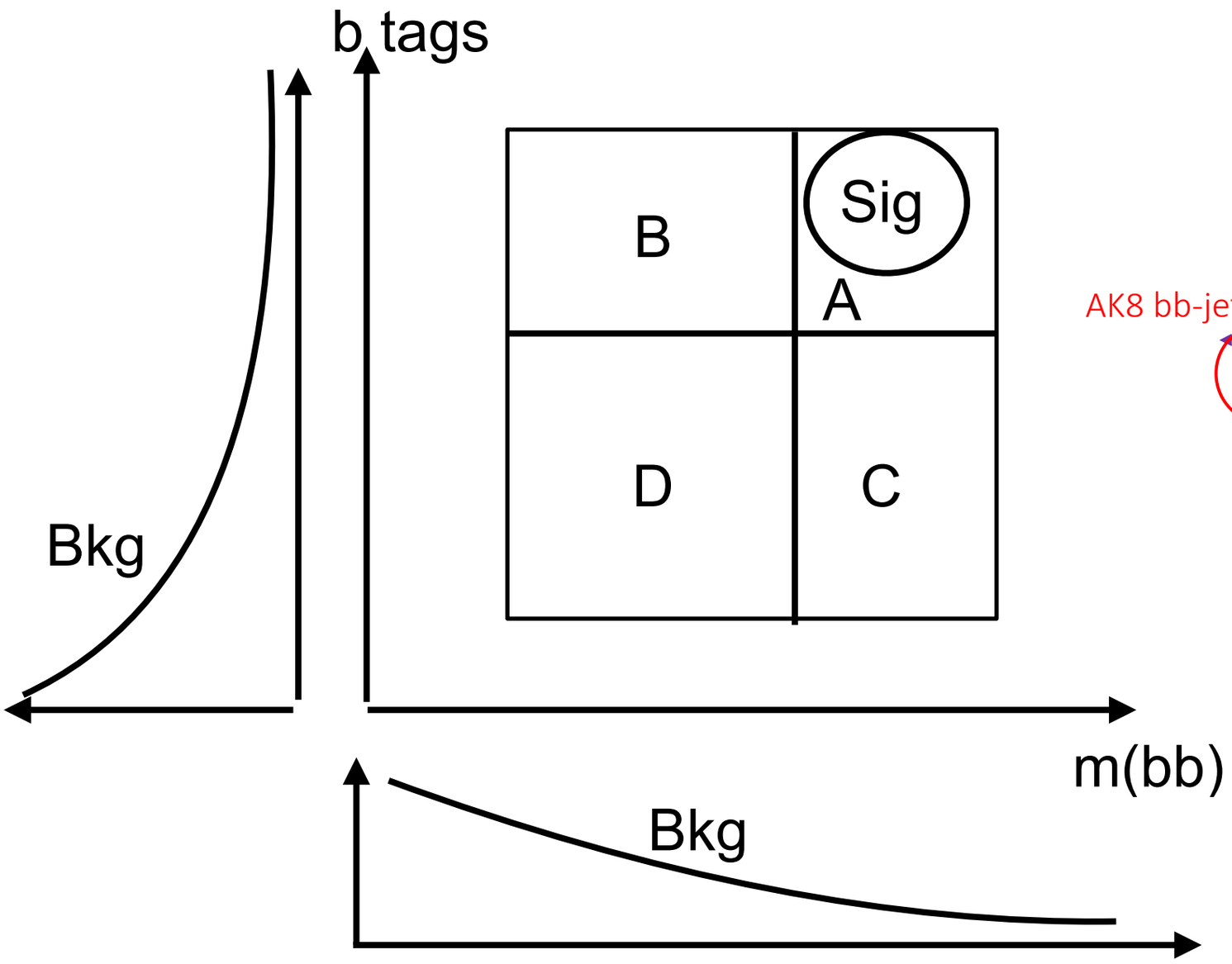
Approaches to extraction of limits from the data in search experiments

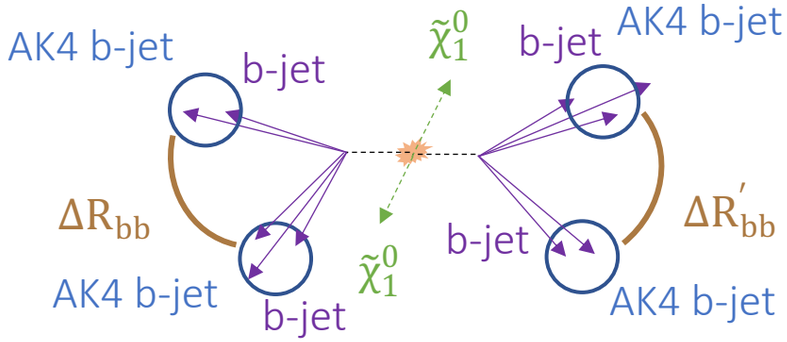
- Focus here on CMS-SUS-20-004 [1]: Higgsino decaying to LSP+H(bb)
  - The likelihood is built and analyzed with the CMS likelihood builder
- Question: how well can one reproduce these results from the information published in HEPData?
- Simplified Likelihood approaches
- Results, comparisons
- Application to alternate models

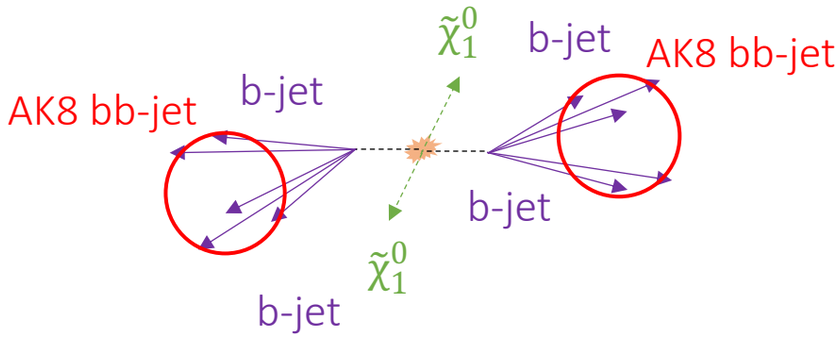
[1]: JHEP 05 (2022) 014

CMS-SUS-20-004: $pp \to \tilde{\chi}_3^0\tilde{\chi}_2^0 \to \mathrm{H(bb)H(bb)}\tilde{\chi}_1^0\tilde{\chi}_1^0$

- Resolved (4 b jets) & boosted (2 fat bb jets) signatures
- Data-driven background prediction, “ABCD”





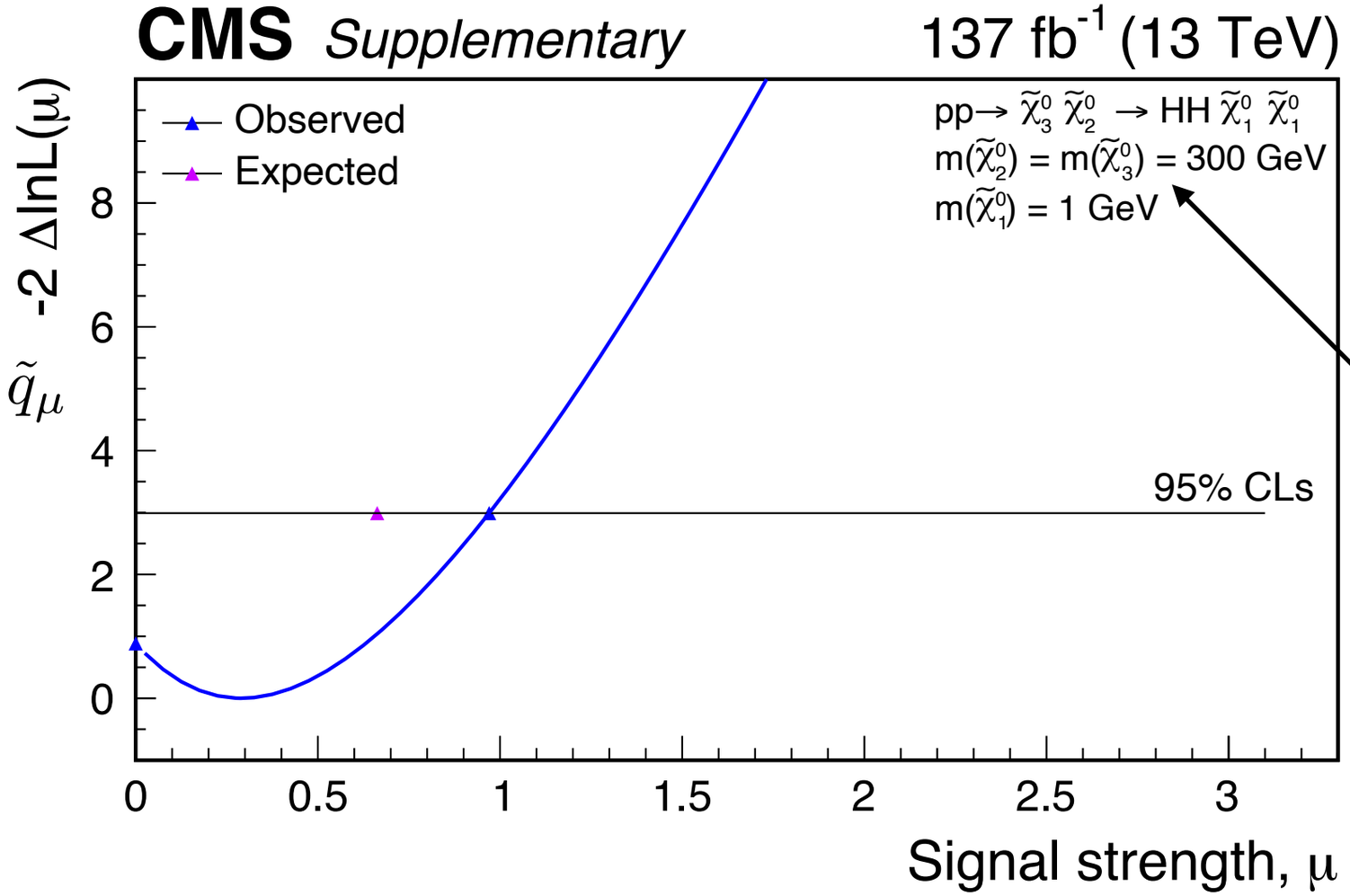


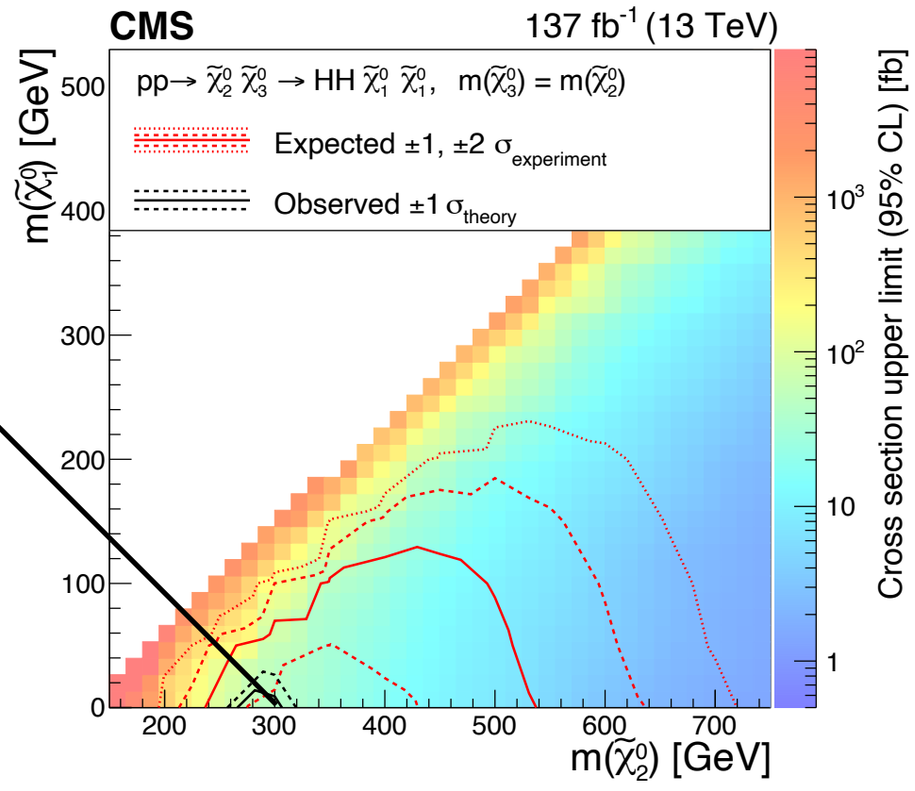
A, C: H mass peak in m(bb)
B, D: sidebands in m(bb)

A, B: >2 b tagged (of 4) jets,
or >0 bb tagged (of 2) fat jets
C, D: lower b-tag requirements

- Predicted N_{bkg} in A = N_B (N_C / N_D)
- All N’s are event counts (some small), so Poisson distributed

Full profile likelihood vs μ





- blue triangles: significance, 95% CLs limit
  - μ < 1 @ 95% CLs ⇒ this (300, 1) point is (barely) excluded
- purple triangle: expected limit

From HEPData

Resolved signature

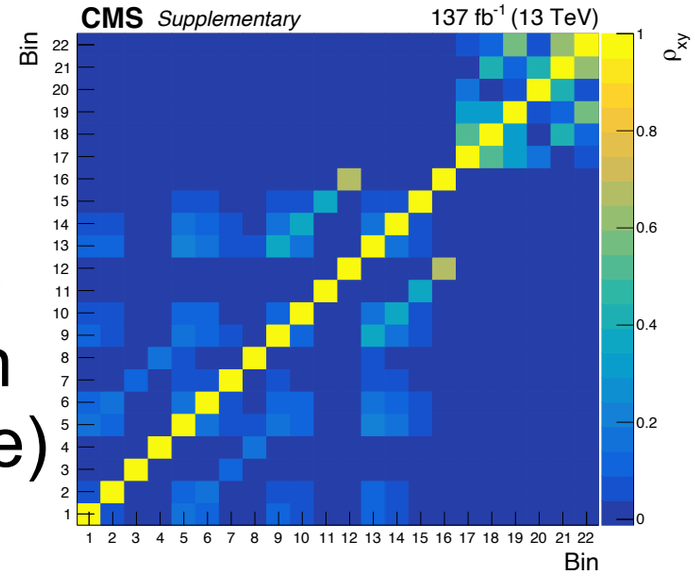
Bin	ΔR_{max}	N_{b}	$p_{\text{T}}^{\text{miss}}$ [GeV]	κ	$N_{\text{SR}}^{\text{pred}}$	$N_{\text{SR}}^{\text{fit}}$	$N_{\text{SR}}^{\text{obs}}$
1	1.1–2.2	3	150–200	$1.09 \pm 0.04 \pm 0.02$	161^{+14}_{-13}	$149.7^{+8.9}_{-8.5}$	138
2			200–300	$0.92 \pm 0.04 \pm 0.02$	$90.4^{+9.7}_{-9.0}$	$91.5^{+6.9}_{-6.5}$	91
3			300–400	$0.94 \pm 0.09 \pm 0.01$	$11.5^{+3.4}_{-2.7}$	$12.8^{+2.6}_{-2.2}$	14
4			>400	$0.98^{+0.19}_{-0.16} \pm 0.02$	$2.8^{+2.3}_{-1.4}$	$2.8^{+1.4}_{-1.0}$	3
5		4	150–200	$1.13 \pm 0.09 \pm 0.08$	$53.5^{+8.8}_{-7.8}$	$54.1^{+5.6}_{-5.2}$	54
6			200–300	$0.96 \pm 0.07 \pm 0.07$	$28.3^{+5.6}_{-4.8}$	$33.2^{+4.2}_{-3.9}$	38
7			300–400	$0.89^{+0.16}_{-0.15} \pm 0.05$	$2.6^{+1.5}_{-1.1}$	$3.2^{+1.3}_{-1.0}$	4
8			>400	$0.92^{+0.27}_{-0.22} \pm 0.07$	$2.6^{+2.4}_{-1.4}$	$1.27^{+0.98}_{-0.63}$	0
9	<1.1	3	150–200	$1.05^{+0.18}_{-0.15} \pm 0.12$	$5.1^{+1.6}_{-1.3}$	$5.9^{+1.4}_{-1.2}$	8
10			200–300	$1.04^{+0.14}_{-0.13} \pm 0.11$	$2.17^{+0.79}_{-0.60}$	$2.31^{+0.73}_{-0.57}$	2
11			300–400	$0.72^{+0.33}_{-0.22} \pm 0.08$	$0.06^{+0.11}_{-0.04}$	$0.72^{+0.53}_{-0.33}$	4
12			>400	$1.24^{+0.67}_{-0.45} \pm 0.10$	$0.89^{+1.42}_{-0.60}$	$0.52^{+0.65}_{-0.35}$	0
13		4	150–200	$1.26^{+0.21}_{-0.20} \pm 0.23$	$2.68^{+1.06}_{-0.79}$	$2.58^{+0.85}_{-0.67}$	1
14			200–300	$1.21^{+0.22}_{-0.21} \pm 0.22$	$1.26^{+0.62}_{-0.44}$	$1.62^{+0.65}_{-0.48}$	3
15			300–400	$2.35^{+0.88}_{-0.72} \pm 0.34$	$0.42^{+0.61}_{-0.27}$	$1.16^{+0.87}_{-0.55}$	1
16			>400	$0.94^{+0.53}_{-0.36} \pm 0.13$	$0.67^{+1.10}_{-0.46}$	$0.78^{+0.76}_{-0.43}$	1

Boosted sig.

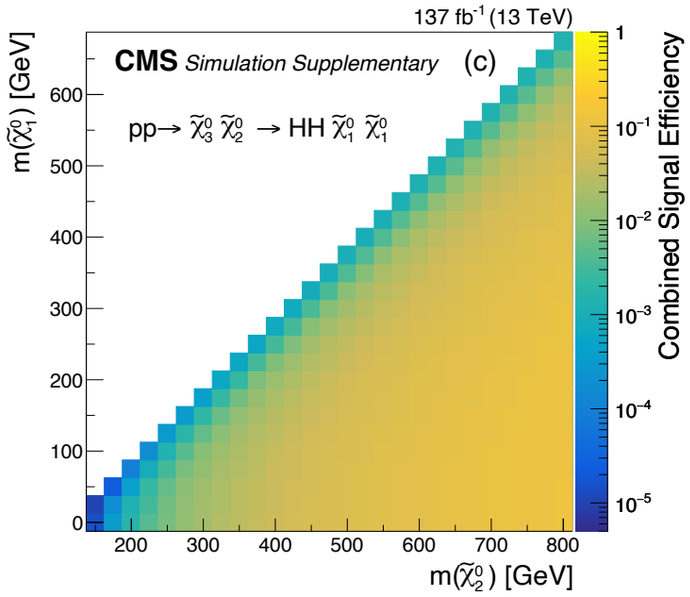
Bin	N_{H}	$p_{\text{T}}^{\text{miss}}$ [GeV]	$N_{\text{SR, tot}}^{\text{pred}}$	f_{pTmiss}	$N_{\text{SR}}^{\text{pred}}$	$N_{\text{SR}}^{\text{fit}}$	$N_{\text{SR}}^{\text{obs}}$
17	1	300–500	42.6 ± 4.2	0.789 ± 0.030	$33.6^{+6.1}_{-5.2}$	$37.0^{+4.2}_{-4.0}$	42
18		500–700		0.172 ± 0.028	$7.3^{+2.0}_{-1.6}$	$7.2^{+1.5}_{-1.3}$	6
19		>700		0.039 ± 0.014	$1.65^{+1.04}_{-0.66}$	$1.50^{+0.75}_{-0.53}$	1
20	2	300–500	5.1 ± 1.0	0.789 ± 0.030	$4.0^{+1.5}_{-1.1}$	$4.0^{+1.2}_{-1.0}$	4
21		500–700		0.172 ± 0.028	$0.88^{+0.40}_{-0.28}$	$0.74^{+0.29}_{-0.21}$	0
22		>700		0.039 ± 0.014	$0.20^{+0.21}_{-0.10}$	$0.14^{+0.13}_{-0.07}$	0

https://www.hepdata.net/record/ins2009652

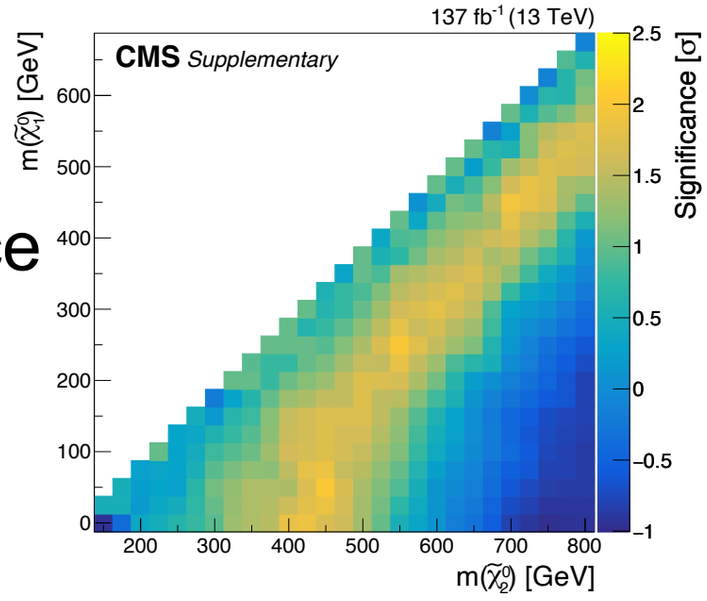
covariance (correlation shown here)



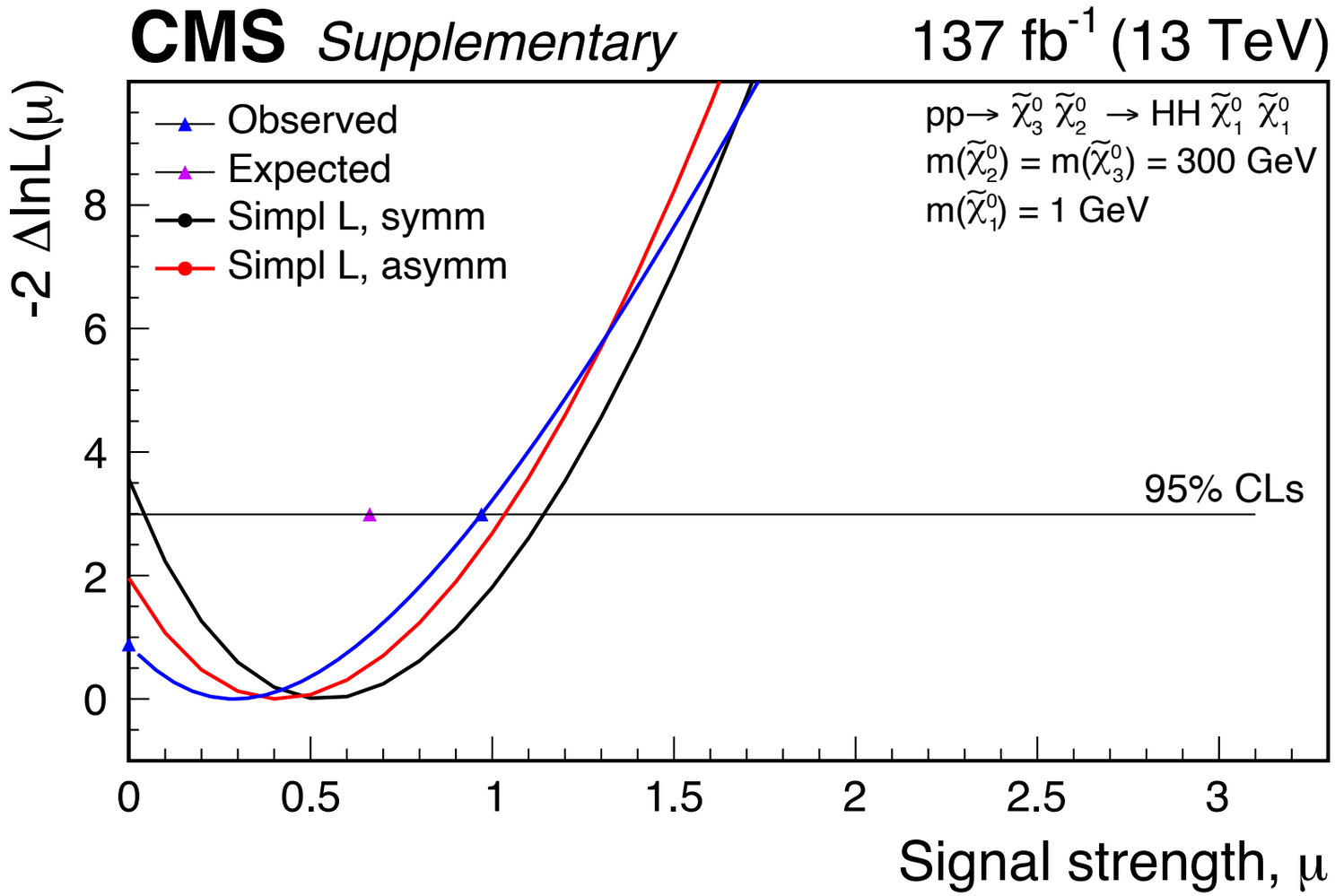
efficiency (by bin available)



significance



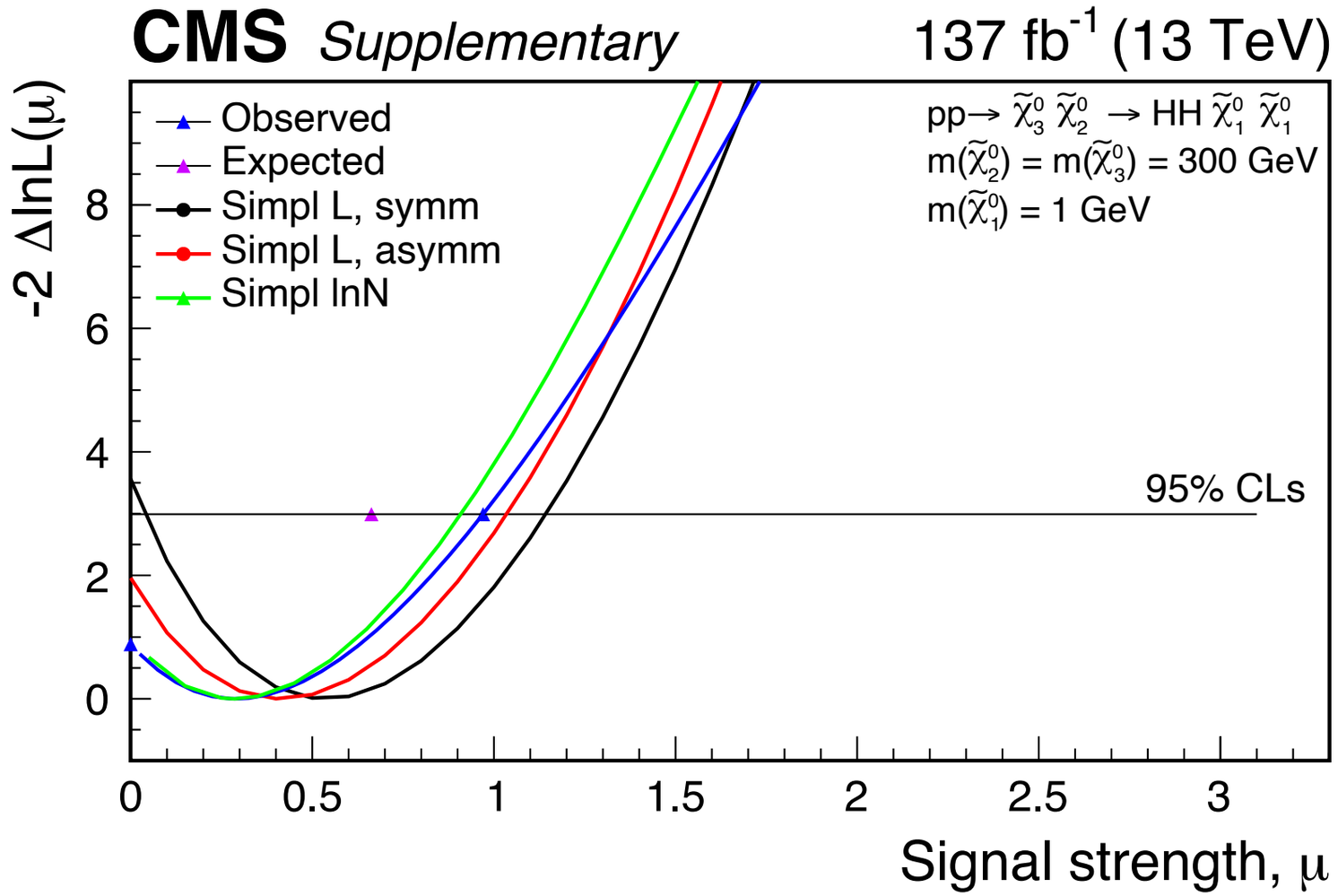
Compare simplified (SL) with full likelihood



- SL treats N^{obs} as Poisson, N^{bkg} as (asymmetric) Gaussian
  - Asymmetry term computed from bifurcated Gaussian bkg pdf.
  - Doesn't fully account for Poisson fluctuations of low-stats CR yields
- Including the asymmetry improves the agreement.

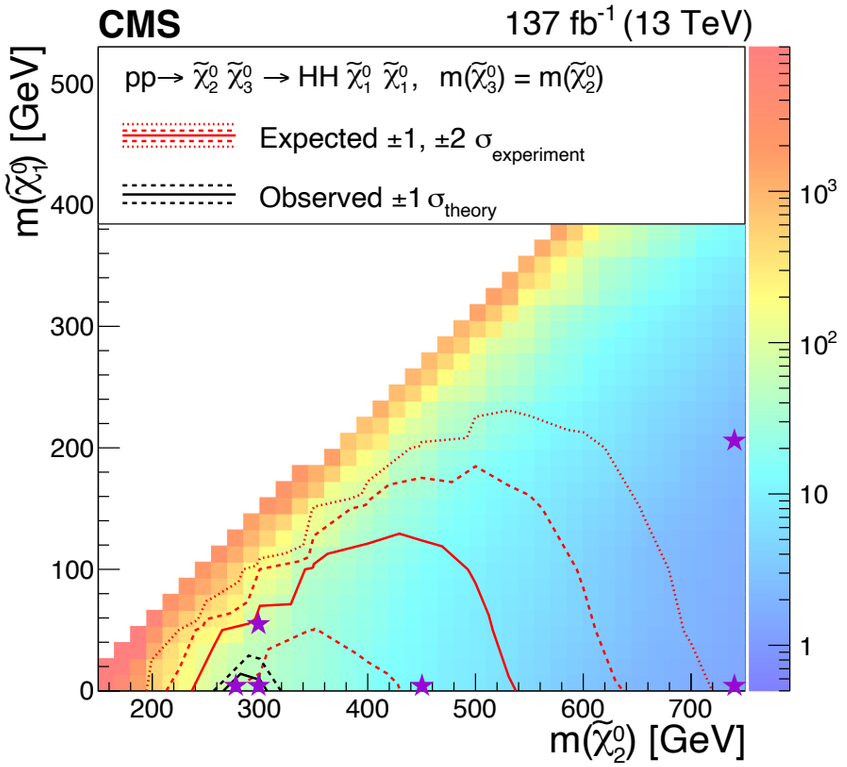
Alternate SL: bkg uncertainties as log-normal

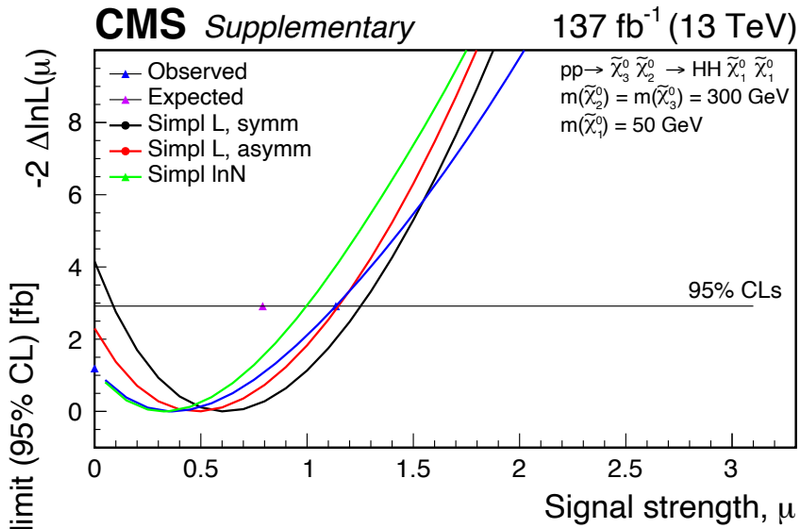
- Here implemented with the CMS likelihood builder.
- Published bkg central values, uncertainties as asymmetric log-normal nuisances.
- Multiply by correlation matrix for bin-bin correlations.

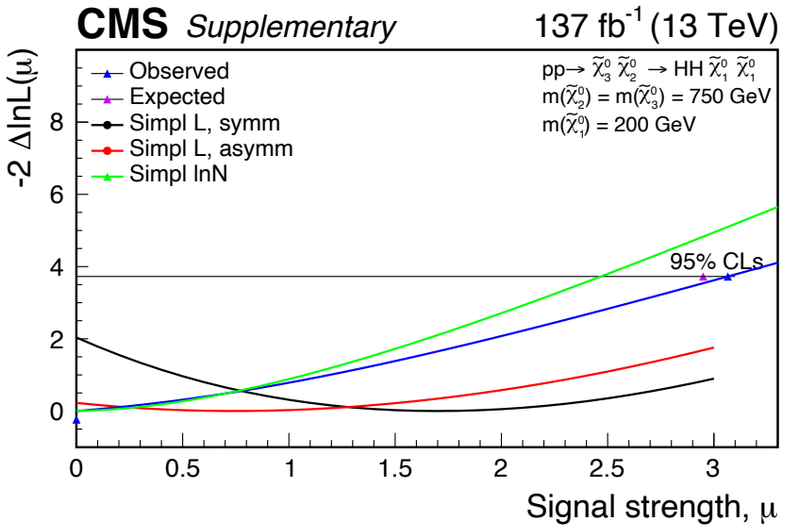


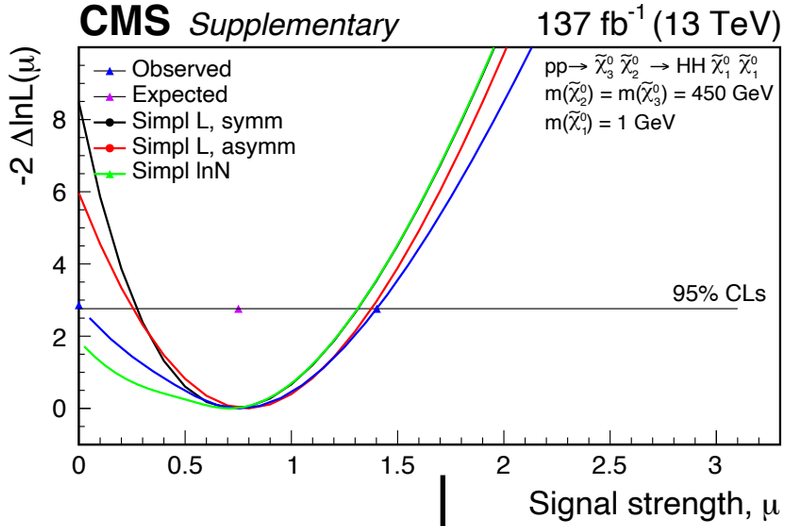
- Accurately fits the minimum and significance.
  - Again, doesn't fully account for Poisson fluctuations of low-stats CR yields.

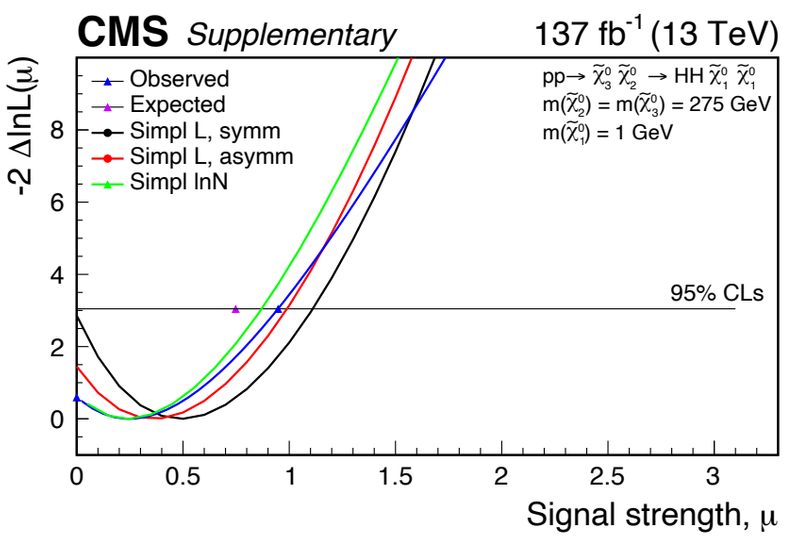
More sample scan points

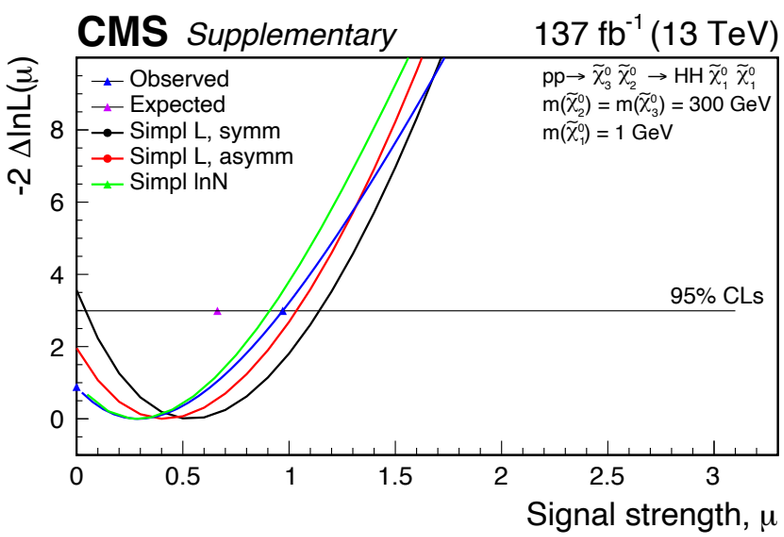


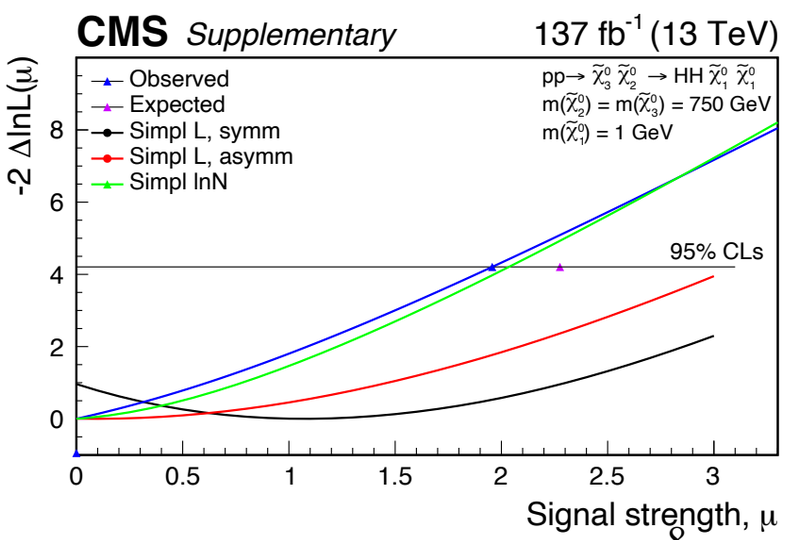












Thoughts on application to other models

- The b quark content would have to be the same, else sorting of the model into 3b, 4b, or 1bb, 2bb bins would be impossible
  - ⇒ (8 topology/kinematical bins for resolved + 3 for boosted)

    * 2 flavor bins
- From generator level information, sort the model events into
  - resolved/boosted, with cuts on ΔR between the H daughter quarks
  - p_T^{miss}, ΔR_{max} bins
- The bin efficiency is normalized to total cross section σ^0 of the reference model, so for a trial model m, need to scale the prediction by $S_i^m = \frac{\sigma_i^m/\sigma^m}{\sigma_i^0/\sigma^0}$
- Then the predicted signal yield for topology/kinematical bin i and flavor bin j of model m is

$$N_{i,j}^{\rm sig} = S_i^m \epsilon_{i,j} \mathcal{B}^2(H \to b\bar{b}) \sigma^m \mathcal{L}, \qquad i \subset 1-11,\ j \subset 1-2$$

Summary

- CMS search papers are typically accompanied by digitized results, with supplementary data, in a HEPData record.
- Here we exercised the use of HEPData tables from one of these searches to reproduce the results by approximate methods.
- The results agree reasonably well.
- We've sketched the steps to test other phenomenological models.

Additional material

Full likelihood

- Built from
  - Poisson pdfs for N^{obs}_i in all A, B, C, D regions
  - Constraints $N^{bkg} = A = \kappa\ B\ C\ /\ D$
  - Correction κ (~1) from MC with Gaussian uncertainty pdfs
  - Log-normal pdfs for other nuisances (calibration corrections)
- The expected yields N^{exp}_i in all ABCD regions are given by
  - $N^{exp}_i = N^{bkg}_i + \mu\ N^{sig}_i$, where μ is the signal strength
    - Accounts for signal contamination in control regions
- The criterion for 95% CL is that $CL_s = CL_{s+b} / CL_b = 0.05$
  - $CL_{s+b} = 1 - \Phi(\sqrt{\tilde{q}_\mu})$, where $\tilde{q}_\mu$ is the profile likelihood test statistic:

$$\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}(data|\mu, \hat{\theta}_\mu)}{\mathcal{L}(data|\hat{\mu}, \hat{\theta})}, \quad 0 \leq \hat{\mu} \leq \mu$$

    , and Φ is the normal cumulative density function.
  - CL_b measured with the Asimov data set (N^{obs} set to $N^{expected}$)
  - Details in CMS-NOTE-2011/005 (ATLAS/CMS)

Simplified Likelihood Framework (SL)

- The predicted yield in bin i is

$$N_i^{\text{pred}} \equiv N_i^{\text{bkg}} + \mu N_i^{\text{sig}},$$

$$N_i^{\text{bkg}} = a_i + b_i\theta_i + c_i\theta_i^2$$

  - a_i is the central value of the bkg prediction
  - θ_i is a nuisance parameter drawn from a unit Gaussian
  - b_i is the effective sigma of the bkg uncertainty, $\sqrt{V_{ii}}$ in the limit of symmetric uncertainties
  - c_i gives the asymmetry of the bkg uncertainty

- The simplified likelihood is

$$L_S(\mu,\theta) \propto \prod_i \text{Pois}(N_i^{\text{obs}}|N_i^{\text{pred}}(\mu,\theta)) \exp\left(-\tfrac{1}{2}\theta^{\mathsf{T}}\rho^{-1}\theta\right)$$

  - where $\rho \rightarrow$ correlation matrix for symmetric uncertainties

  - A. Buckley et al., CMS Note-2017/001
    A. Buckley et al., JHEP 2019, 64 (2019)
    gitLab

SL: asymmetric bkg uncertainties

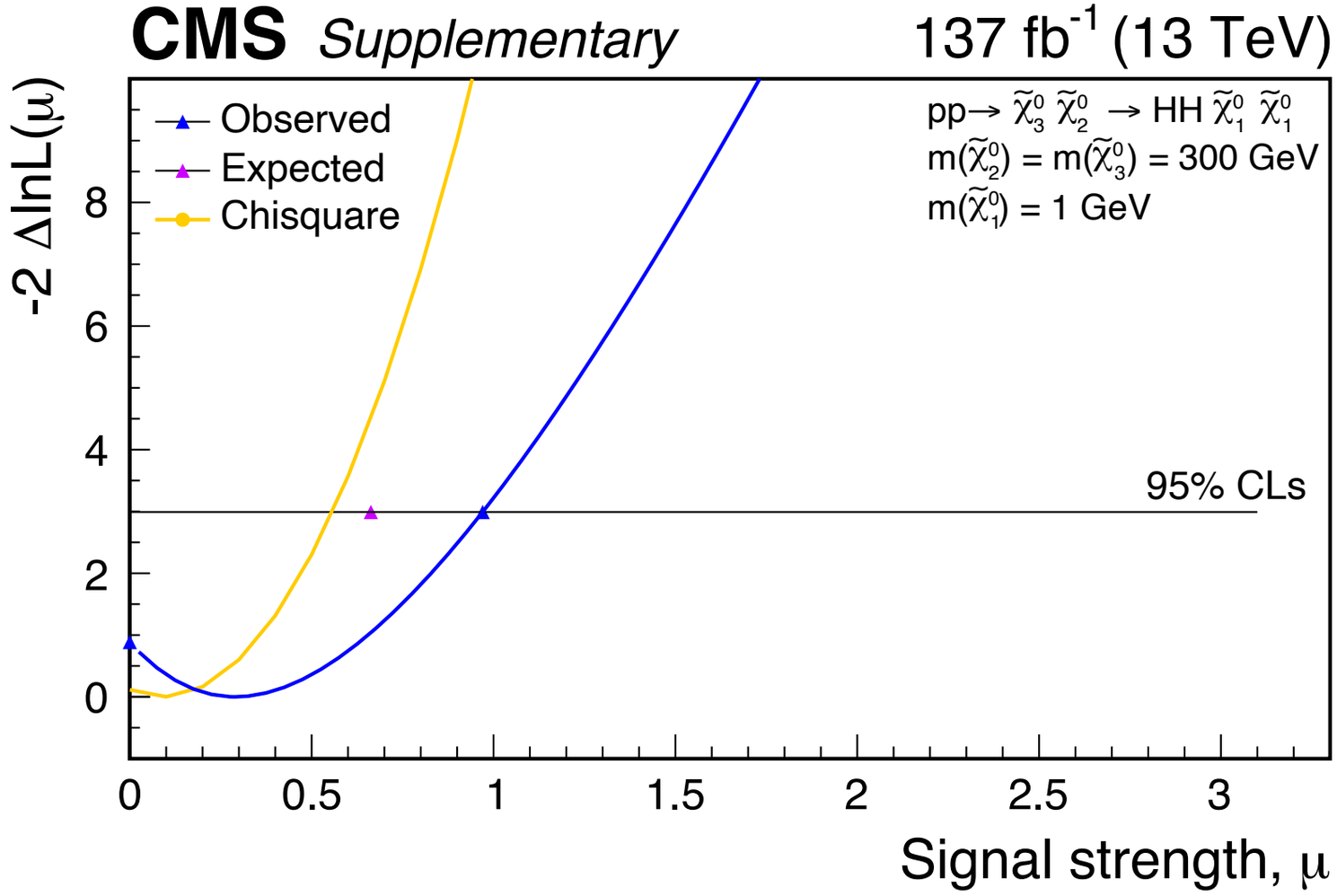
- The covariance matrix gives second moments, i.e., sigma^2, on the diagonal, and correlations, on off-diagonal elements
- To incorporate asymmetric uncertainties, SL uses the diagonal elements of the 3rd moment m_3 of the background nuisances.
- For CMS-SUS-20-004, we compute m_3 from a bifurcated Gaussian using the asymmetric uncertainties $\sigma_{1,2}$:

$$m_3 = \frac{2}{\sigma_1 + \sigma_2}\left[\sigma_1 \int_{-\infty}^{0} x^3 G(x;0,\sigma_1)dx + \sigma_2 \int_{0}^{+\infty} x^3 G(x;0,\sigma_2)dx\right]$$

Chisquare method

$$\chi^2 = \Delta_i V_{ij}^{-1} \Delta_j$$
$$\Delta_i \equiv N_i^{\rm obs} - N_i^{\rm pred},$$
$$N_i^{\rm pred} \equiv N_i^{\rm bkg} + \mu N_i^{\rm sig},$$
$$N_i^{\rm sig} = \epsilon_i \mathcal{B}^2(H \to b\bar{b}) \sigma \mathcal{L}$$
$$V = V^{\rm bkg} + {\rm diag}(N^{\rm obs})$$

underestimates μ_0 and high-side uncertainty



- Limitations
  - All errors Gaussian
  - Any tension between predicted bkg and observation is underestimated by artificial uncertainty on the observed yield.
    - E.g., the bin 11 contribution before squaring is (very nearly) (4 - 0)/ √4, which is 2 sigma, vs the detailed study giving 3.3 sigma local significance.