Reinterpretation material in CMS searches

Reinterpretation workshop @ Imperial

April 2nd, 2019

Andreas Albert on behalf of CMS

III. Physikalisches Institut A

RWTH AACHEN UNIVERSITY

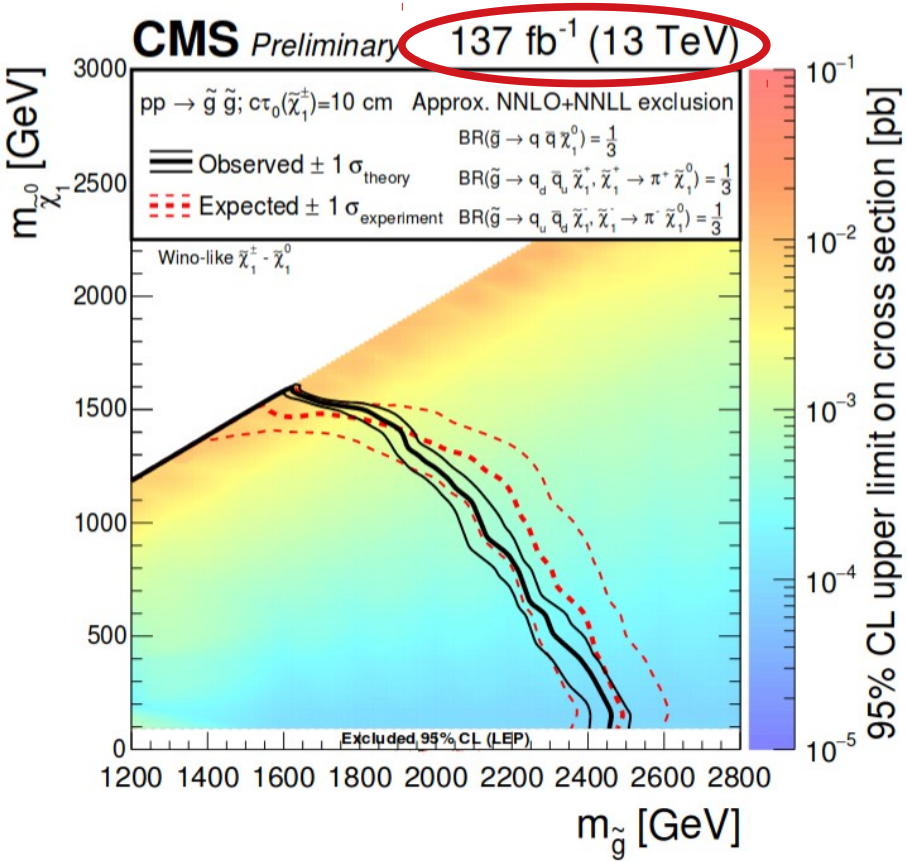
SPONSORED BY THE

Federal Ministry of Education and Research

Run-II legacy

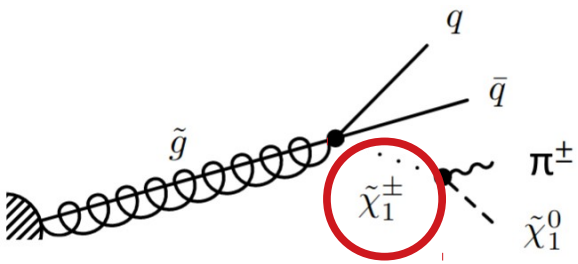
CMS-PAS-SUS-19-005

First full Run-II results out ≈ now



Absence of obvious BSM until now

→ Ensure experimental coverage



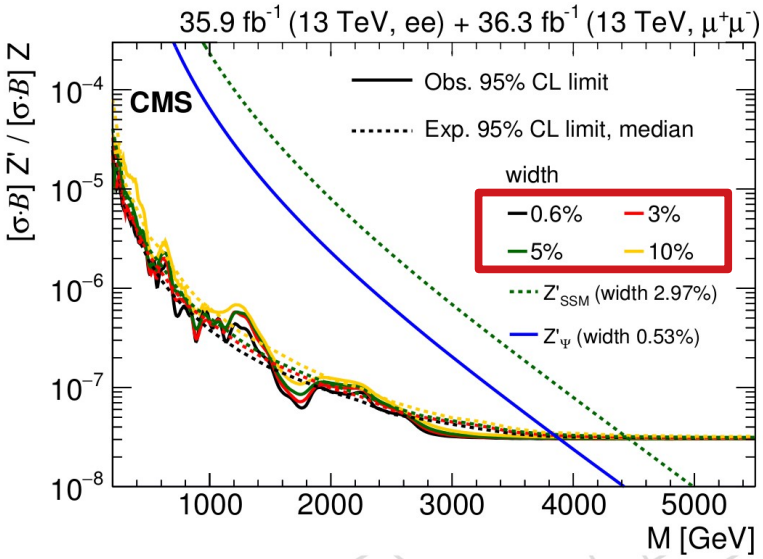
long-lived

→ Clear need to ensure reinterpretability

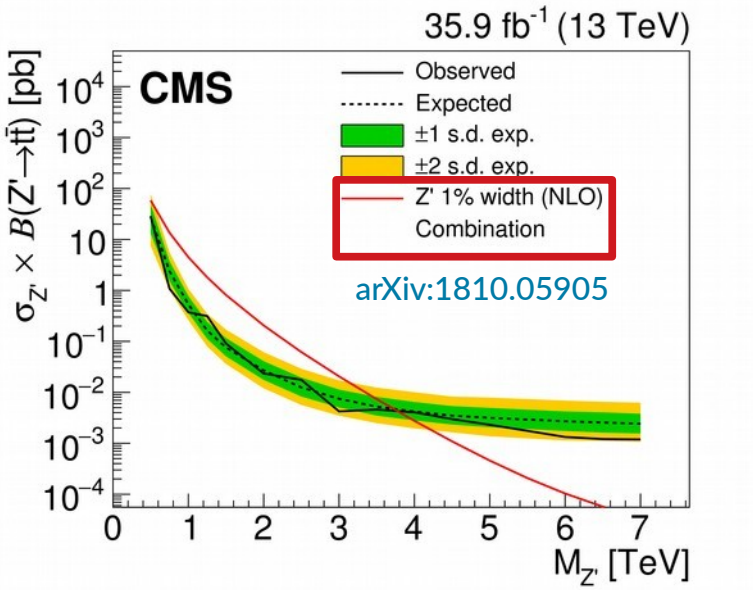
Sometimes, reinterpretation is easy

Best case: Results can be formulated generically, e.g. as cross-section limits

Z' → dilepton



Z' → tt



10.1007/JHEP06(2018)120

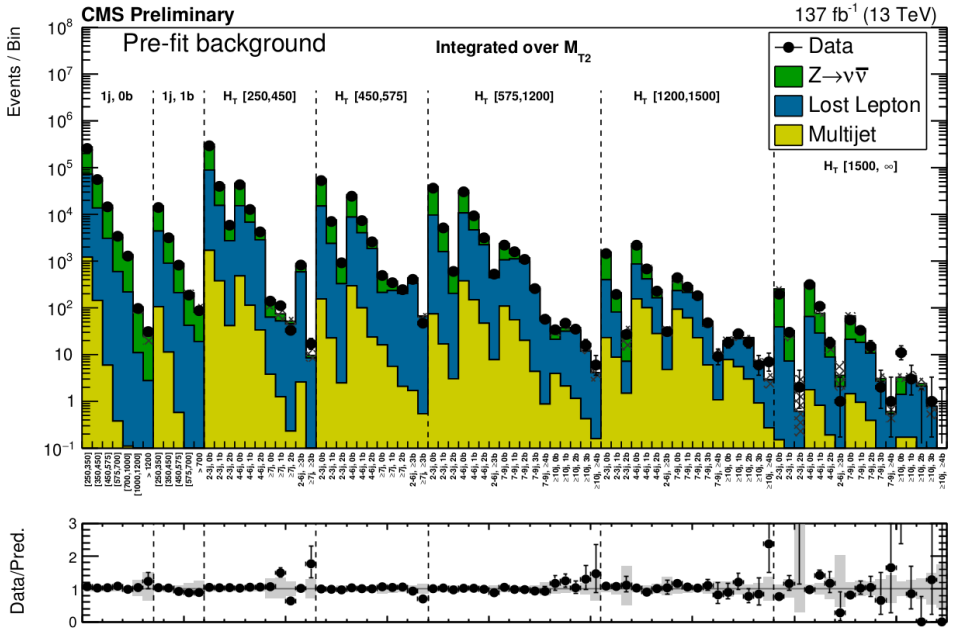
Parametrize in few free parameters: Resonance mass, width

Complex production may pose problems through additional objects

Sometimes, it is more complicated

CMS-PAS-SUS-19-005

Complex categorization: Kinematic properties important



Goal: Empower you to do what we do

- Derive your own signal prediction
- Use our background prediction
- Perform statistical analysis

→ This talk

Predicting the signal

Making signal predictions: Object efficiencies

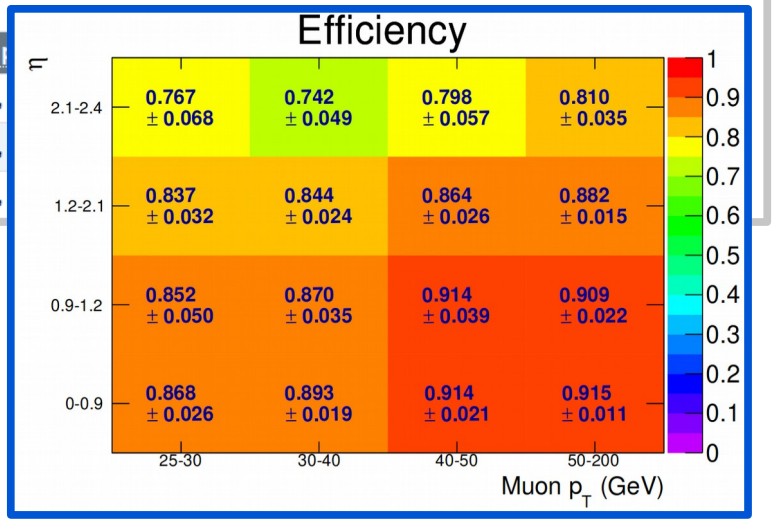
Standard objects systematically provided in SUSY searches on public twiki page

Light Leptons Selection Efficiency

Representative Muon and Electron efficiencies for the WPs of the identification techniques used in SUS-16 analyses:

- only the analyses with at least one light lepton in the final state are considered and only the efficiency for the signal-lepton selections is reported (no veto selections);
- the efficiency refers to the reconstruction + identification + isolation + vertexing requirements for generator-level leptons from W decay in a simulated sample of ttbar events;
- the efficiency is corrected with the corresponding data/simulation scale factors extracted from 2016 data

CADI	Analysis	Muon pT and eta	Muon Selection Eff. vs (pT, eta)	.root file	El
SUS-16-042	1L + Jets with ΔΦ	25, 2.4	eff	root	25,
SUS-16-037	1L + Jets with MJ	20, 2.4	eff	root	20,
SUS-16-040	1L RPV	20, 2.4	eff	root	20,



E.g. signal leptons: Efficiency from $t\bar{t}$ MC, corrected to match data

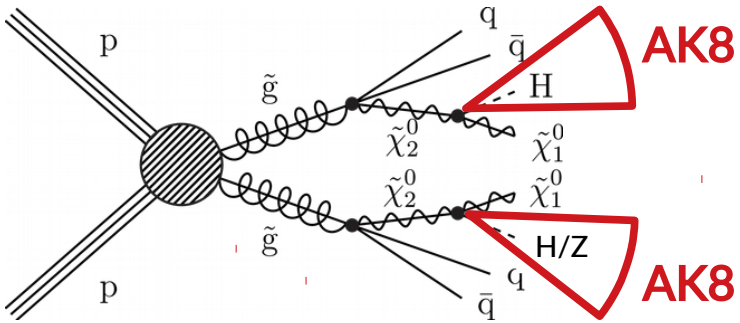
+ τ_{had}, b tag, photons

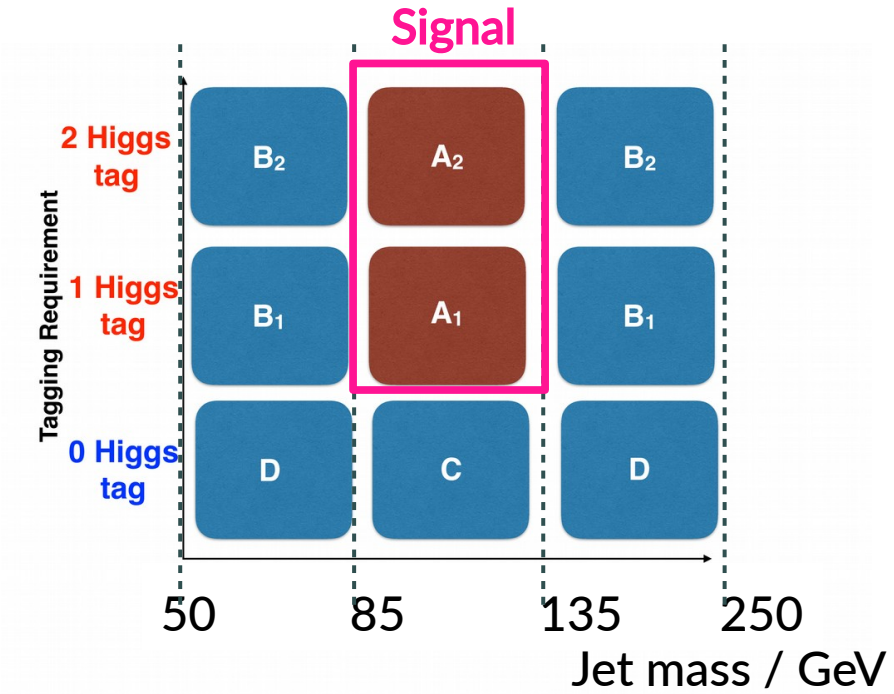
Cross-check / correction to Delphes

More complex objects: SUSY with H(bb)

10.1103/PhysRevLett.120.241801
Public twiki for this result

Example: double b tagged fat jets to look for BSM with H(bb)





Main selection:

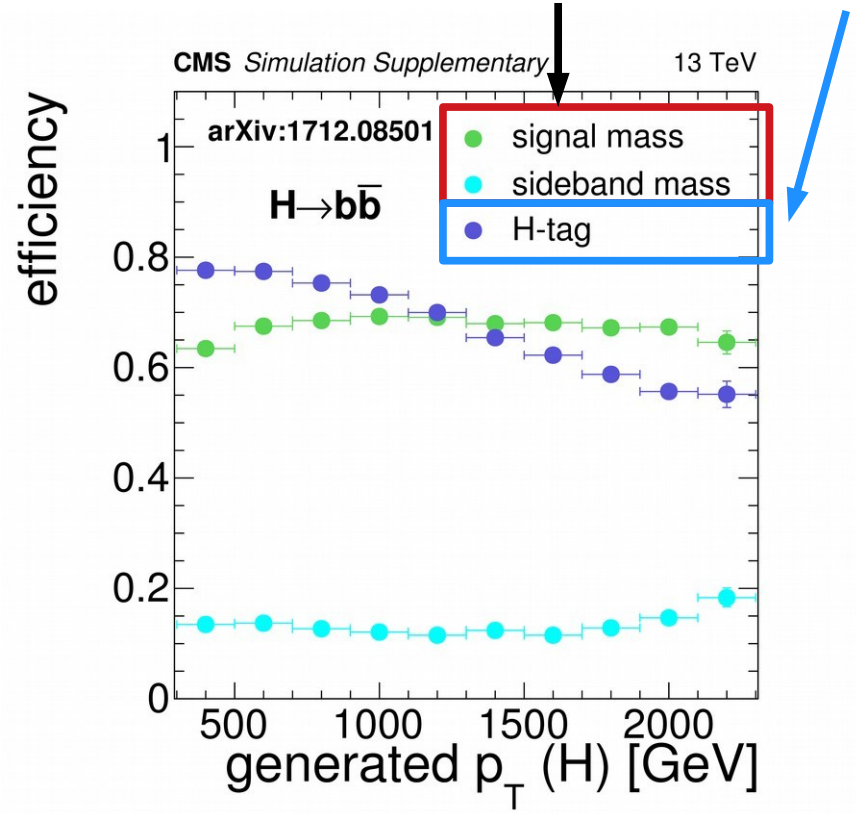
- Two R = 0.8 anti-kt jets
- p_T^{miss} > 300 GeV
- H_T^{miss} > 200 GeV
- H_T > 600 GeV

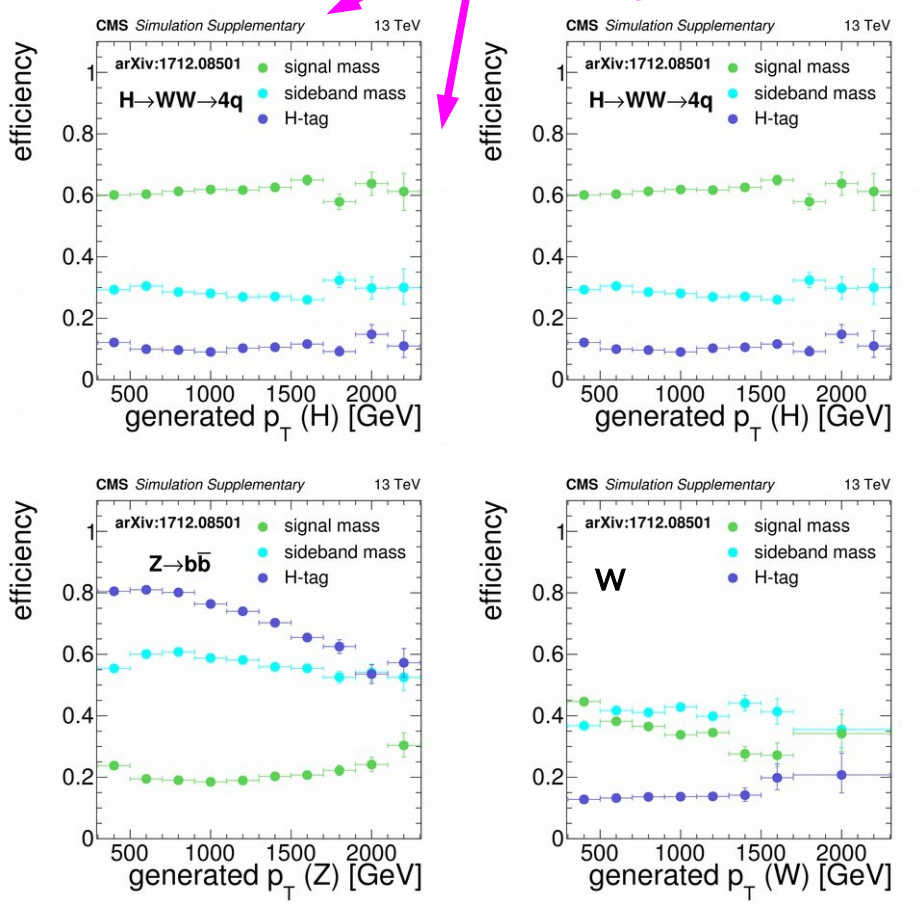
Use jet mass and H(bb) tags to define signal, control regions

SUSY with H(bb): Efficiencies

10.1103/PhysRevLett.120.241801
Public twiki for this result

Provide mass region and tagging efficiencies for different jet topologies





SUSY with H(bb): Full reinterpretation workflow

10.1103/PhysRevLett.120.241801
Public twiki for this result

1. Apply / verify object efficiencies
2. Implement signal selection, validate with cutflow
3. (If signal contaminates control regions:
   → Rederive ABCD estimate)
4. Construct likelihood , set limits

Cutflow

	Model	HH	HH	ZH	ZH
	Gluino mass (GeV)	1300	2200	1300	2200
	Total yield at 35.9 fb^{-1}	1660	12.8	1651	12.9
+	Lepton & isolated track vetoes	1253	10.1	1277	10.3
+	$\Delta\phi_{1,2,3,4}$	1037	8.3	1049	8.4
+	HT > 600 & p_T^{miss} > 300 GeV	886	7.8	834	7.6
+	≥2 AK8 jets with p_T > 300 GeV	602	6.5	523	5.4
+	Jet 1,2 mass ϵ [50, 250 GeV]	374	4.1	313	3.3
+	Jet 1,2 mass ϵ [85, 135 GeV]	209	2.5	101	1.0
+	One or two H-tags	168	2.0	67	0.6

Yields and correction factors for ABCD

$A_{predicted} = \kappa * B * C / D$

N_H	p_T^{miss} (GeV)	κ	$A_{predicted}$	A	B	C	D
1	300 – 500	0.98 ± 0.11	17.7 ± 3.8	15	112	44	273
1	500 – 700	0.86 ± 0.16	3.4 ± 1.5	2	20	12	60
1	>700	0.86 ± 0.17	0.61 ± 0.45	1	5	4	28
2	300 – 500	0.73 ± 0.14	1.52 ± 0.57	1	13	44	273
2	500 – 700	0.43 ± 0.12	0.09 ± 0.08	0	1	12	60
2	>700	0.62 ± 0.30	$0.09^{+0.11}_{-0.09}$	0	1	4	28

Selection rather general → widely applicable

Bin correlations irrelevant → easy limits

10.1103/PhysRevD.98.092011, HEPData

Displaced vertices

Main variable: two-vertex distance d_{VV}

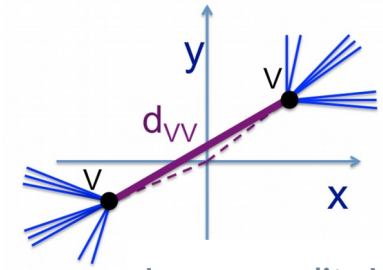
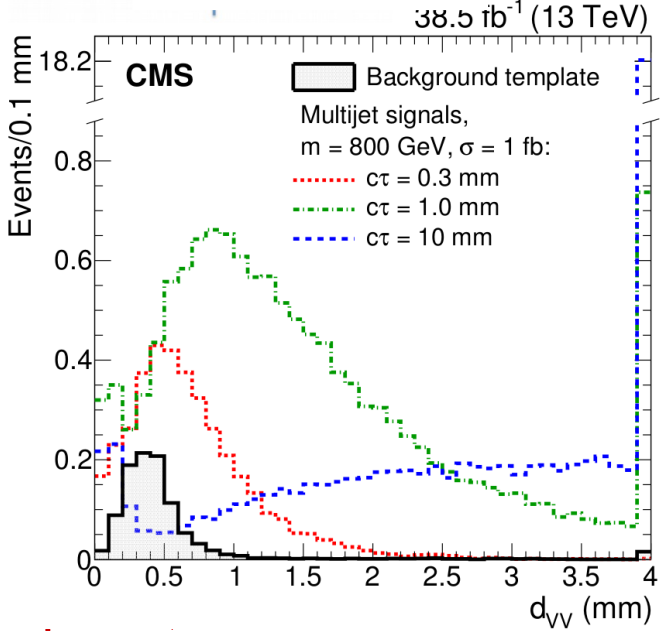


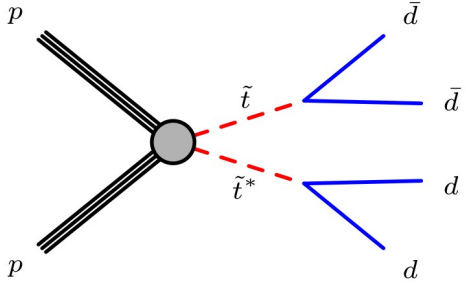
Image credit: J. Chu

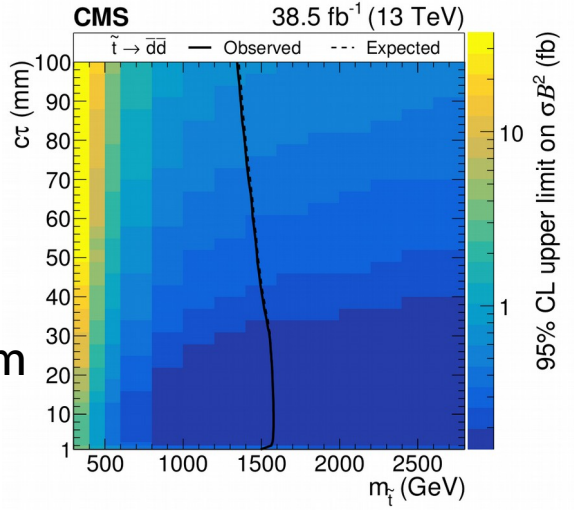


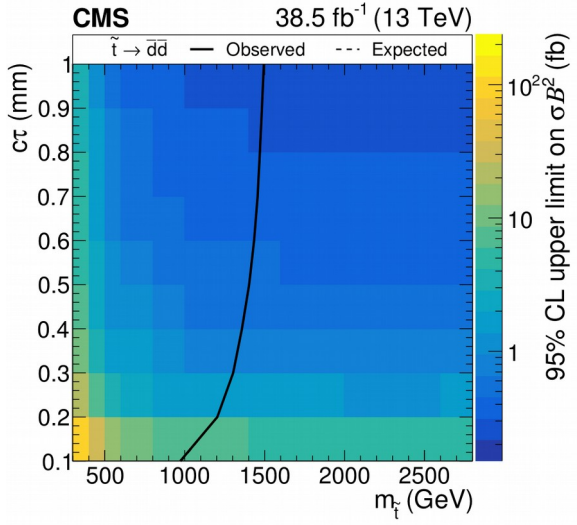
Search strategy is generic

Main sensitivity: cτ = 0.1 – 100 mm

Example interpretation: RPV stops







10.1103/PhysRevD.98.092011, HEPData

Displaced vertices: How to reinterpret?

Challenge: Custom vertex reconstruction, can't rely on Delphes

→ Map **reco-level** selection onto purely **generator-level** selection

Analysis selection	Generator-level selection
at least four jets	at least four generated jets
HLT_PFHT800 or HLT_PFHT900 trigger	
H_T (jets with p_T > 40 GeV) > 1000 GeV	H_T (generated jets with p_T > 40 GeV) > 1000 GeV
two vertices	
for each vertex:	for each long-lived particle:
d_{BV} > 0.1 mm	xy position of decay > 0.1 mm
xy distance from detector origin < 20 mm	xy position of decay < 20 mm
at least five tracks	
uncertainty in d_{BV} < 25 μm	Σp_T of daughter particles > 350 GeV
d_{VV} > 0.4 mm	generated d_{VV} > 0.4 mm

generated jets: p_T > 20 GeV, |η| < 2.5, electron energy fraction < 0.9, muon energy fraction < 0.8
daughter particles: u,d,s,c,b,e,μ,τ with p_T > 20 GeV, |η| < 2.5, |dxy| > 0.1 mm
in calculating Σp_T of daughter particles, multiply the p_T of b quarks by a factor of 0.65

Approximates reco result within 20%

Main signal uncertainty: vertex reco eff.

→ Model independent

Simple signal region

→ easy reinterp.

d_{VV} range	Fitted background yield	Observed
0–0.4 mm	0.51 ± 0.01 (stat) ± 0.13 (syst)	1
0.4–0.7 mm	0.37 ± 0.02 (stat) ± 0.09 (syst)	0
0.7–40 mm	0.12 ± 0.02 (stat) ± 0.08 (syst)	0

Slide credit: J. Chu

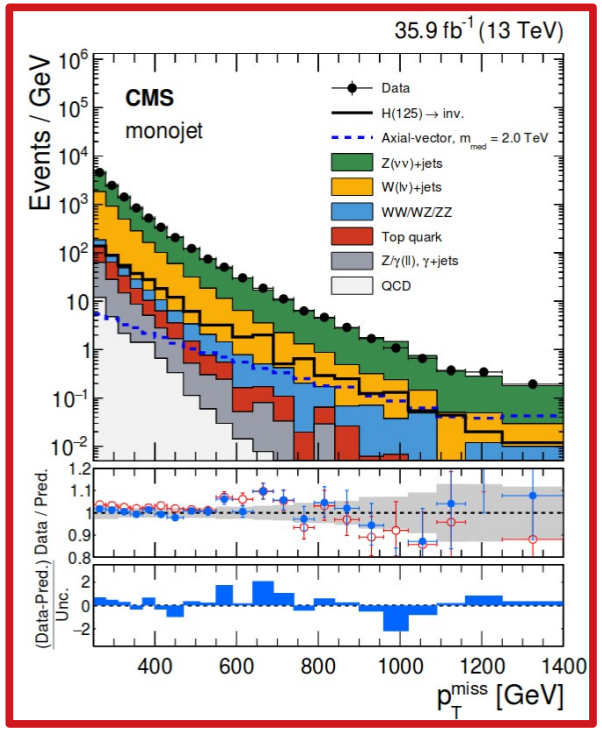
Background modelling for reinterpretation

Complex background modelling

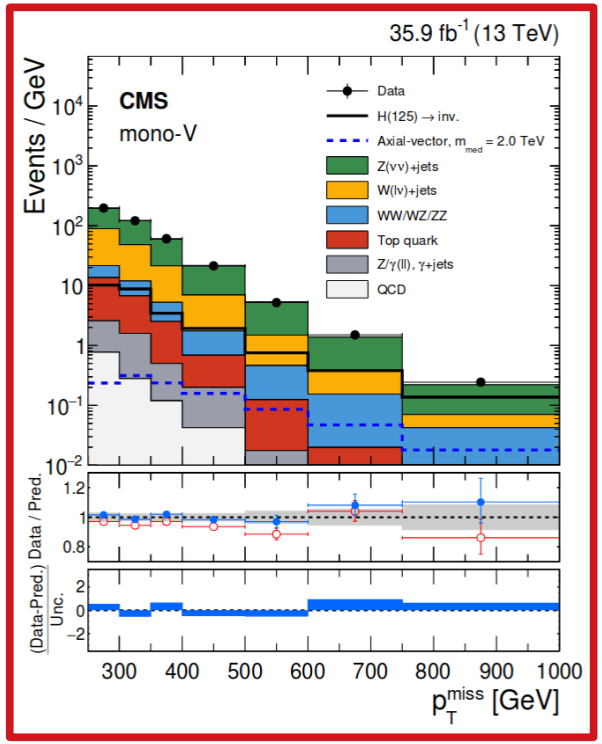
10.1103/PhysRevD.97.092005
Auxilliary material

Example: Monojet / V, uncertainties determined in combined ML fit

Two categories, 29 signal bins total

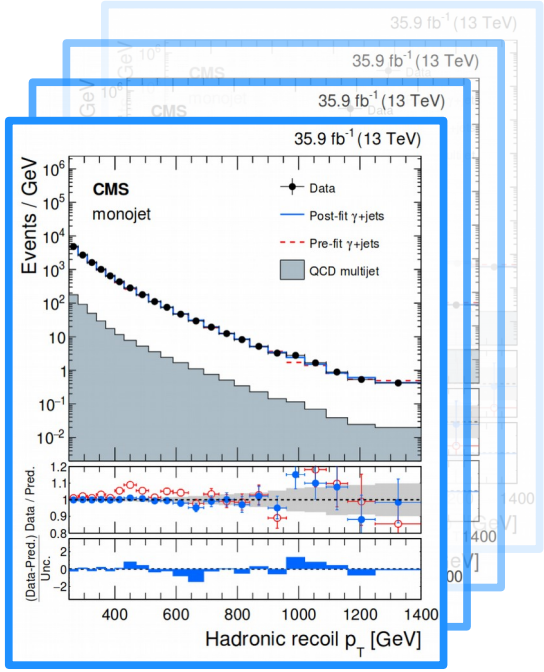


+



+

5 × 2 control regions



Elemental issue: bin yields are correlated through common systematics

Simplified likelihood

Simplified likelihood documentation

Statistical interpretation typically uses likelihood approach

Free parameters θ = independent physical uncertainty sources

$$\mathcal{L}(\mu, \boldsymbol{\theta}) = \mathcal{P}(\text{data}|\mu \cdot s(\boldsymbol{\theta}) + b(\boldsymbol{\theta})) \cdot p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$$

Simplified implementation

Free parameters θ = deviation from central value in each bin

$$\mathcal{L}_S(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu \cdot s_i + b_i + \theta_i)^{n_i} e^{-(\mu \cdot s_i + b_i + \theta_i)}}{n_i!} \cdot \underbrace{\exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \mathbf{V}^{-1} \boldsymbol{\theta}\right)}$$

(bins)

Assumption: Gaussian distribution for bin contents, covariance info encoded in V

Simplified likelihood: How do I use it?

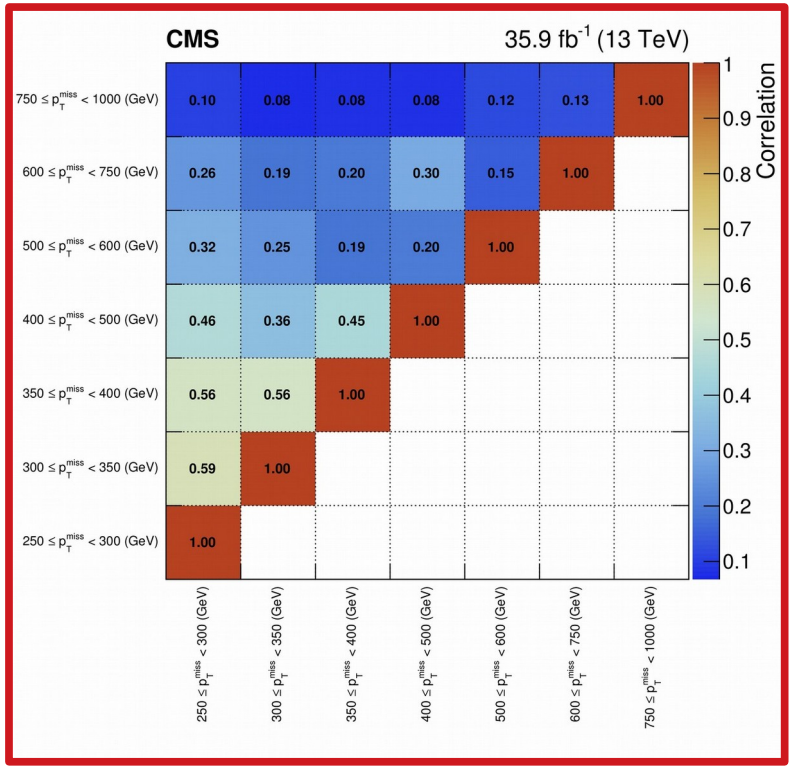
10.1103/PhysRevD.97.092005
Auxilliary material

Inputs for fully defined likelihood:

- You calculate signal yields s_i in each bin
- We provide:
  - BG yields b_i + uncertainties
  - data yields n_i
  - Covariance matrix V

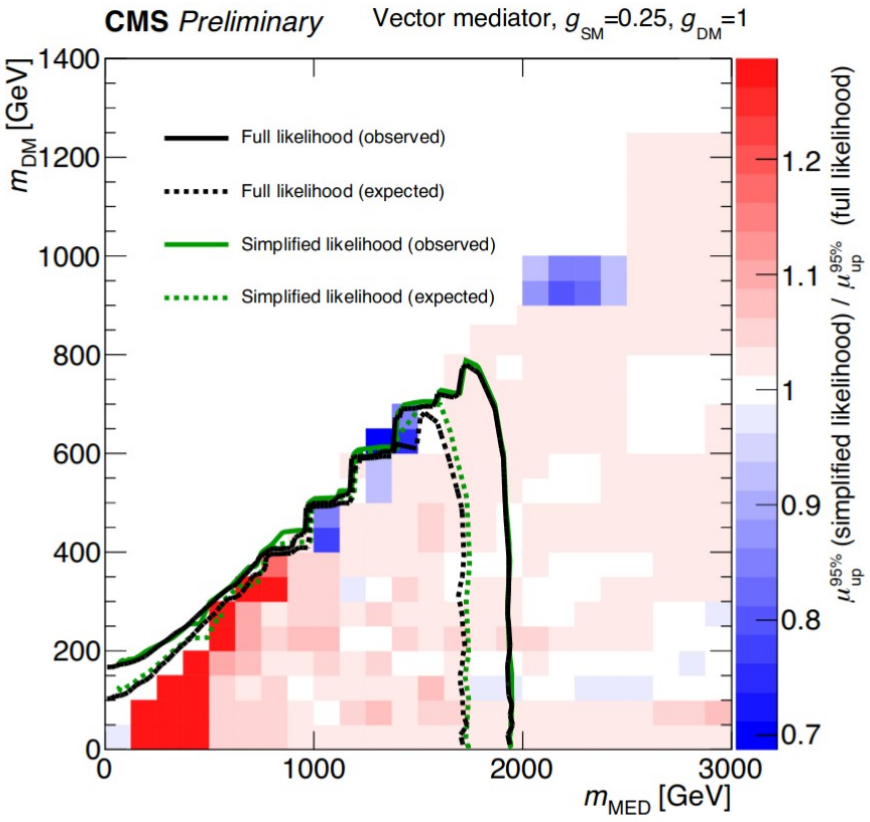
Plug everything into formula from two slides ago

$$\mathcal{L}_S(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu \cdot s_i + b_i + \theta_i)^{n_i} e^{-(\mu \cdot s_i + b_i + \theta_i)}}{n_i!} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \mathbf{V}^{-1} \boldsymbol{\theta}\right)$$



Simplified likelihood: Validation in previous monojet/V result

Compare simplified and full likelihood results



Good agreement observed
Significantly better than without correlation info

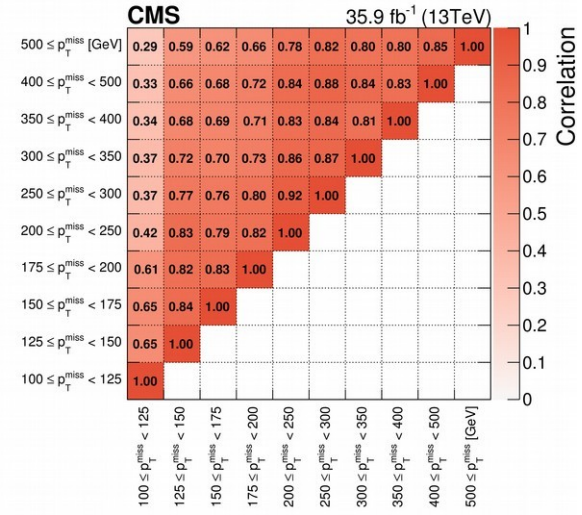
Simplified likelihood documentation

Simplified likelihood: Availability

Make simplified likelihood information available wherever possible

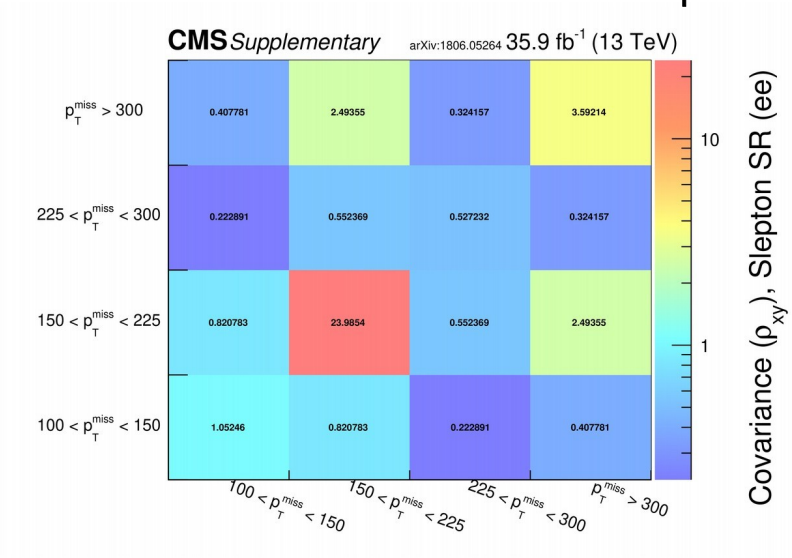
Some example cases:

Mono-Z



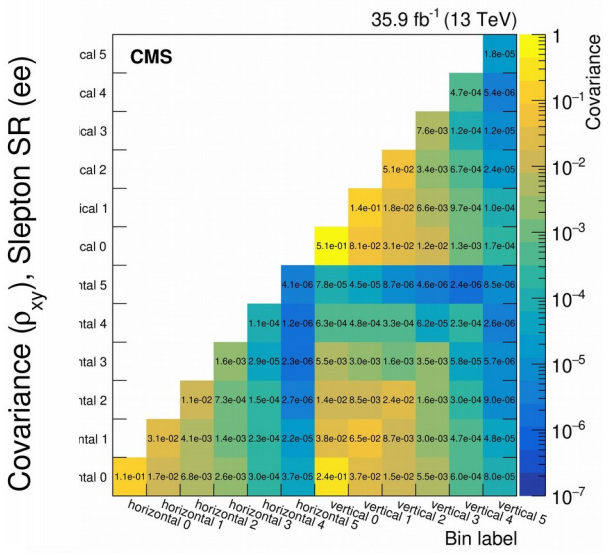
10.1140/epjc/s10052-018-5740-1

Slepton pairs → leptons + p_T^{miss}



Phys. Lett. B 790 (2019) 005

Monophoton



10.1007/JHEP02(2019)074

+ Many more

Distribution of material

Distribution of material: CMS public pages

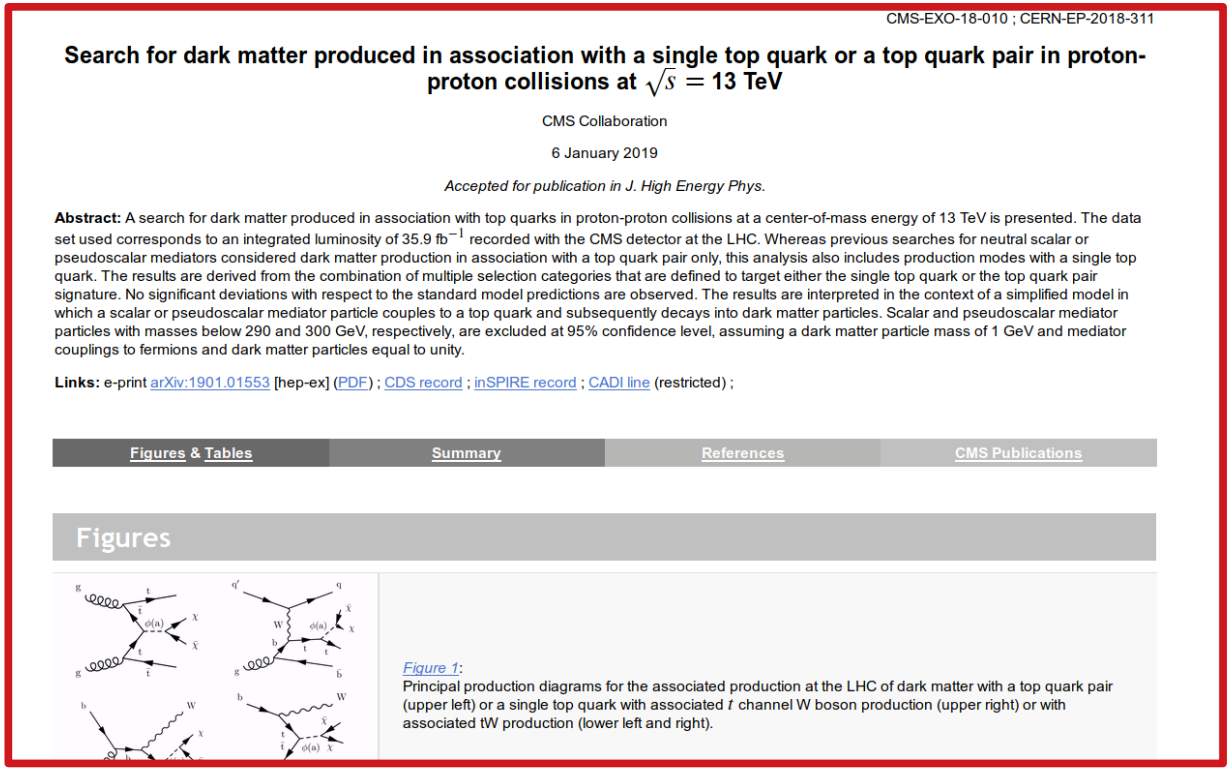
Every publication has a public web page

Collect material from paper

+ auxilliary items

Accessible from central CMS page:

http://cms-results.web.cern.ch/cms-results/public-results/publications/



CMS-EXO-18-010 ; CERN-EP-2018-311

Search for dark matter produced in association with a single top quark or a top quark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV

CMS Collaboration

6 January 2019

Accepted for publication in J. High Energy Phys.

Abstract: A search for dark matter produced in association with top quarks in proton-proton collisions at a center-of-mass energy of 13 TeV is presented. The data set used corresponds to an integrated luminosity of 35.9 fb^{-1} recorded with the CMS detector at the LHC. Whereas previous searches for neutral scalar or pseudoscalar mediators considered dark matter production in association with a top quark pair only, this analysis also includes production modes with a single top quark. The results are derived from the combination of multiple selection categories that are defined to target either the single top quark or the top quark pair signature. No significant deviations with respect to the standard model predictions are observed. The results are interpreted in the context of a simplified model in which a scalar or pseudoscalar mediator particle couples to a top quark and subsequently decays into dark matter particles. Scalar and pseudoscalar mediator particles with masses below 290 and 300 GeV, respectively, are excluded at 95% confidence level, assuming a dark matter particle mass of 1 GeV and mediator couplings to fermions and dark matter particles equal to unity.

Links: e-print arXiv:1901.01553 [hep-ex] (PDF) ; CDS record ; inSPIRE record ; CADI line (restricted) ;

Figures & Tables | Summary | References | CMS Publications

Figures

Figure 1:
Principal production diagrams for the associated production at the LHC of dark matter with a top quark pair (upper left) or a single top quark with associated t channel W boson production (upper right) or with associated tW production (lower left and right).

Distribution of material: HEPData

Goal: Provide entry for every analysis

Clear advantages in accessibility

Making progress

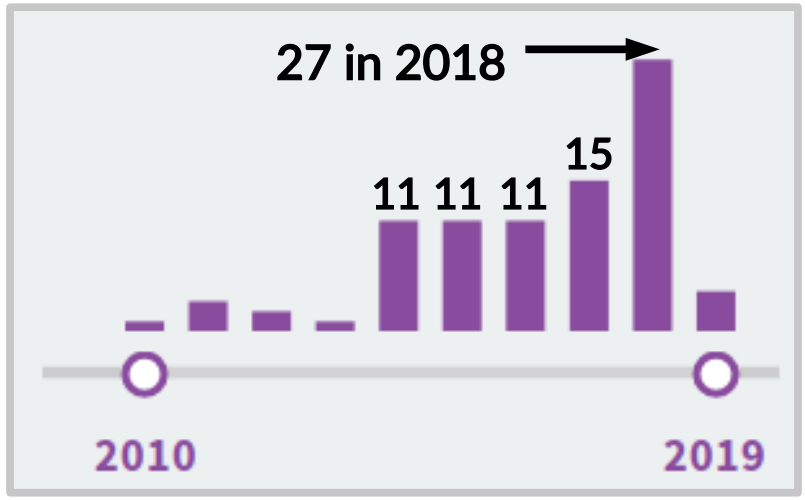
Two main ways:

1. Top down: institutionalized strategy

   = encourage centrally

2. Bottom up: Technical support for adoption

   = remove hurdles



Link to HepData web

HepData input format

YAML-based data format:

one data file per table

+ one meta data file per submission

```yaml
dependent_variables:
- header: {name: '$m_{\tilde{\chi}^0_1}$', units: GeV}
  values:
  - {value: 10.8}
    (...)
  - {value: 112.5}

independent_variables:
- header: {name: '$m_{\tilde{t}}$', units: GeV}
  values:
  - {value: 737.5}
    (...)
  - {value: 337.5}
```

```yaml
(...)
data_file: Fig11b_observed.yaml
description: Figure 11 (lower). Observed exclusion region at 95% CL assuming 100%
  branching fraction.
keywords:
- name: observables
  values: [regions]
- name: cmenergies
  values: [13000.0]
- name: phrases
  values: [Top squark, Dark Matter, T2bW, dilepton, SUSY]
location: Data from Figure 11 (lower) of publication
name: Figure 11b observed contour
table_doi: 10.17182/hepdata.79809.v1/t24
(...)
```

Easy to read, harder to write

Nested dictionary structure

What goes where?

→ No need for everyone to learn this

File source

Alternative: Implement logic in simple python

HEPData / **hepdata_lib** ✓

Github

Object oriented code intuitively represents what you see on HepData web

```python
from hepdata_lib import Submission, Variable, Table

# Top-level object
sub = Submission()

# Table = HepData table
table = Table("Graviton limits")

# Define relations
table.add_variable(mass)
table.add_variable(limit)
sub.add_table(table)

# Write upload-ready submission tarball
sub.create_files("./output/")

# Done!
```

```python
# Variable = table column
mass = Variable("Graviton mass",
                is_independent=True,
                is_binned=False,
                units="GeV")
mass.values = [ 1, 2, 3 ]

limit = Variable("Cross-section limit",
                is_independent=False,
                is_binned=False,
                units="fb")
limit.values = [ 10, 5, 2 ]
```

- **Handles all HepData features**
- **Not CMS dependent**

Summary

CMS aims to provide reinterpration ingredients wherever possible

- Signal prediction
  - → Simple cases: Object efficiencies + cutflow enough
  - → Complex cases: Dedicated recipes, e.g. generator-level selection
- Background model
  - → Simplified likelihood

Improving our HEPData coverage

Recognize that ease of access is important

Always looking for feedback

Let us (me) know if ingredients are missing for a specific analysis or in general

Backup

hepdata_lib: Simple example (Reading)

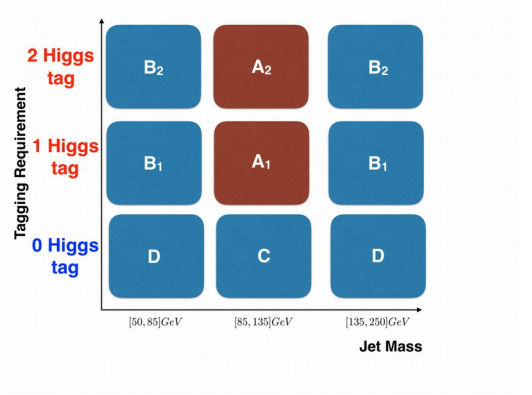
Minimize effort needed to translate your existing files into finished submission.
Example: Read 1D ROOT histogram from file

```
from hepdata_lib import RootFileReader
reader = RootFileReader("input/backgrounds.root")

# Read Histogram properties as dictionary
# Format ready to feed into Variable
QCD = reader.read_hist_1d("QCD")

# Access:
x_values = QCD["x"]
y_values = QCD["y"]
```

SUSY in H(bb): BG estimate

$A_{predicted} = \kappa * B * C / D$

κ is correction factor for residual non-closure of ABCD from MC



N_H	p_T^{miss} (GeV)	κ	$A_{predicted}$	A	B	C	D
1	300 – 500	0.98 ± 0.11	17.7 ± 3.8	15	112	44	273
1	500 – 700	0.86 ± 0.16	3.4 ± 1.5	2	20	12	60
1	>700	0.86 ± 0.17	0.61 ± 0.45	1	5	4	28
2	300 – 500	0.73 ± 0.14	1.52 ± 0.57	1	13	44	273
2	500 – 700	0.43 ± 0.12	0.09 ± 0.08	0	1	12	60
2	>700	0.62 ± 0.30	$0.09^{+0.11}_{-0.09}$	0	1	4	28

Signal contamination in sidebands → subtract → new BG estimate

SUSY in H(bb): Cutflow

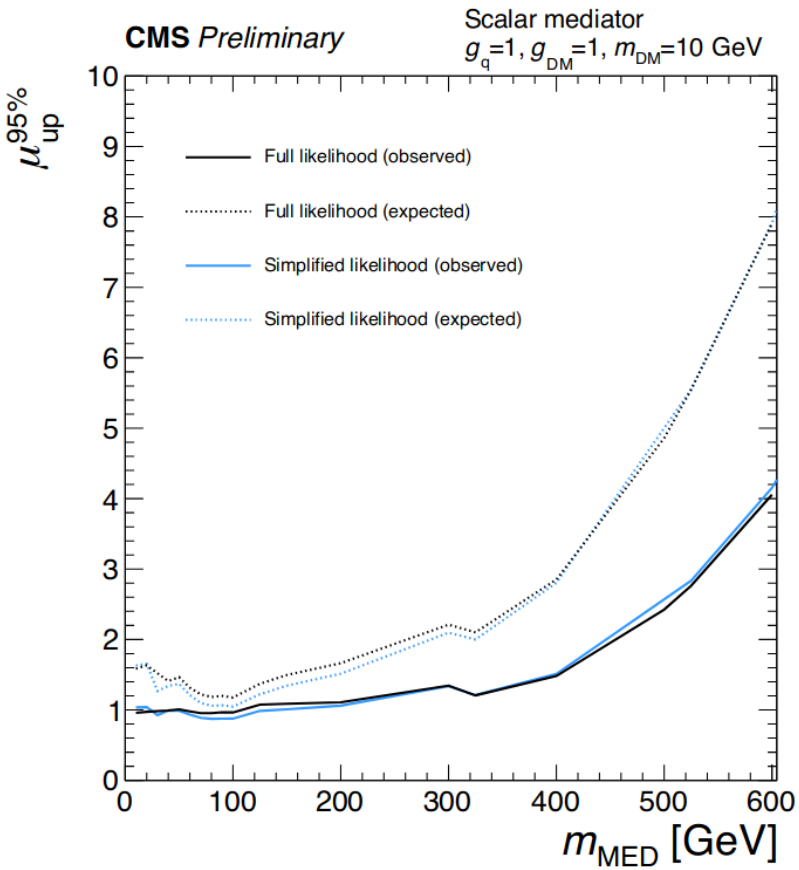
Cutflow for validation of selection

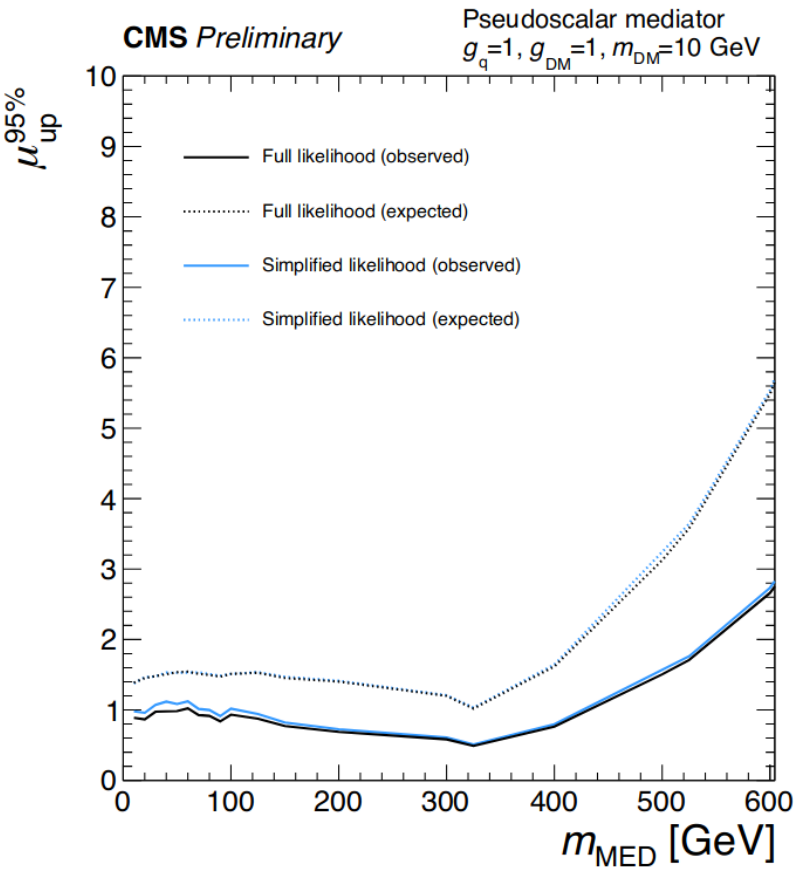
	Model	HH	HH	ZH	ZH
	Gluino mass (GeV)	1300	2200	1300	2200
	Total yield at 35.9 fb^{-1}	1660	12.8	1651	12.9
+	Lepton & isolated track vetoes	1253	10.1	1277	10.3
+	$\Delta\phi_{1,2,3,4}$	1037	8.3	1049	8.4
+	HT $>$ 600 & $p_T^{miss} > 300$ GeV	886	7.8	834	7.6
+	$\geq$2 AK8 jets with $p_T > 300$ GeV	602	6.5	523	5.4
+	Jet 1,2 mass ϵ [50, 250 GeV]	374	4.1	313	3.3
+	Jet 1,2 mass ϵ [85, 135 GeV]	209	2.5	101	1.0
+	One or two H-tags	168	2.0	67	0.6

Especially helpful to validate vetoes

($\Delta\varphi_i$ = angle between jet i and p_T^{miss})

Simplified likelihood: Validation





Simplified likelihood: Test statistic

$$q(\mu) = -2\ln\frac{\mathcal{L}_S(\mu, \hat{\boldsymbol{\theta}}_\mu)}{\mathcal{L}_S(\hat{\mu}, \hat{\hat{\boldsymbol{\theta}}})}$$

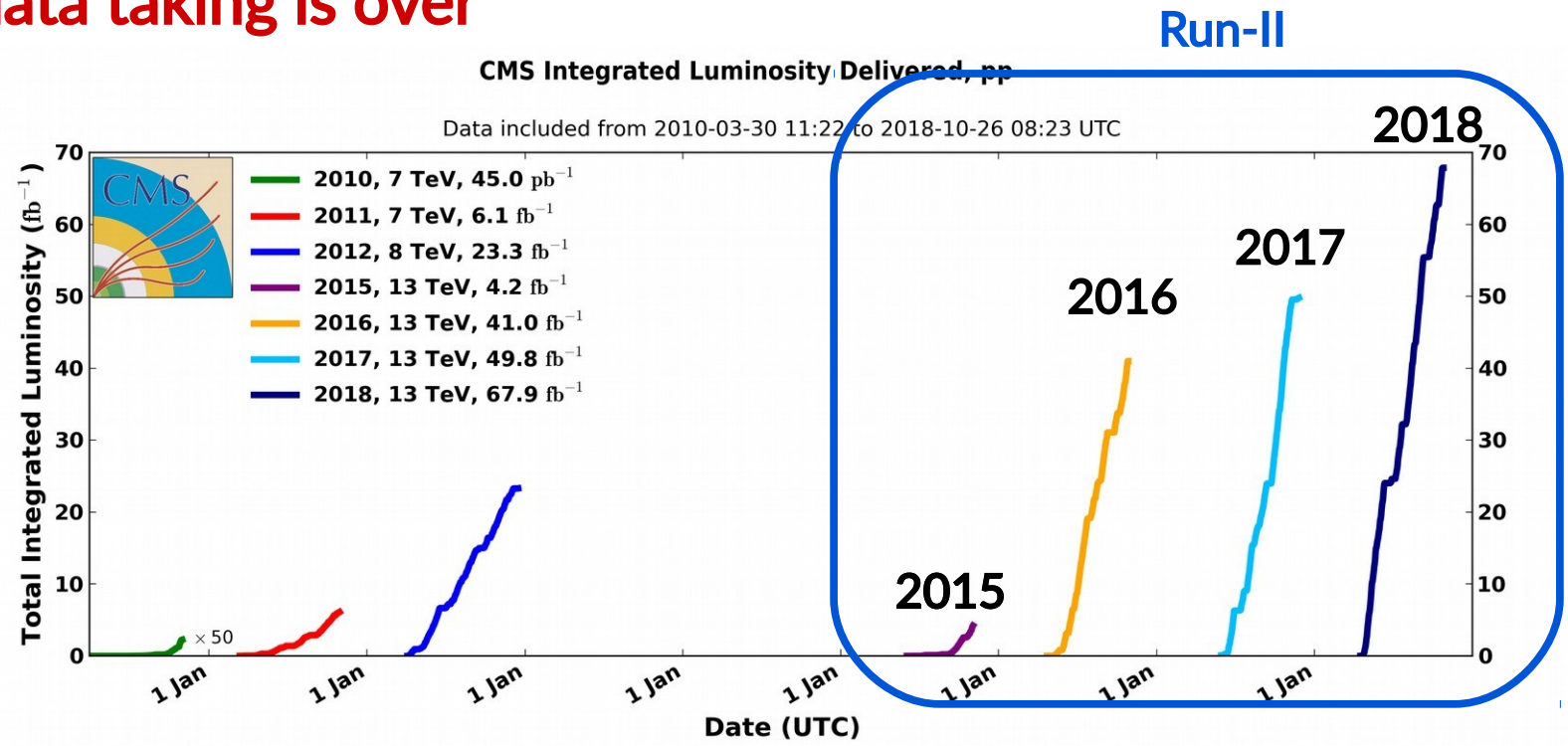
Numerator:

1) Fix μ
2) determine best fit θ
3) evaluate L

Denominator:

1) Simultaneously find best fit μ, θ
2) then evalue L

Run-II data taking is over



Beginning of Run-II:

Immediate results with 2015,

full set of analyses with 2016

Now:

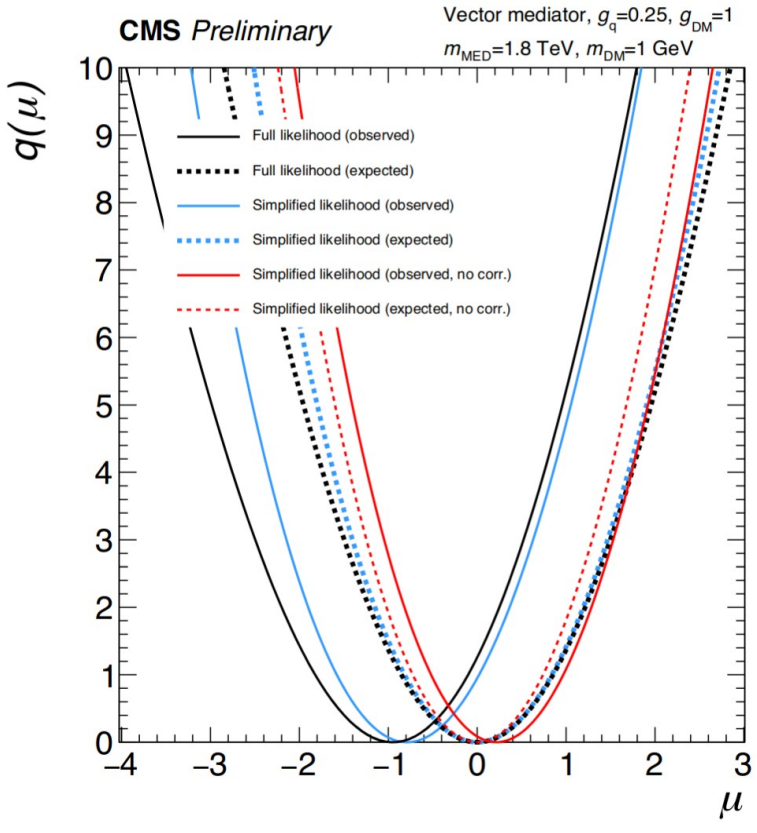
Incorporate lessons learnt

Full set of “legacy” analyses

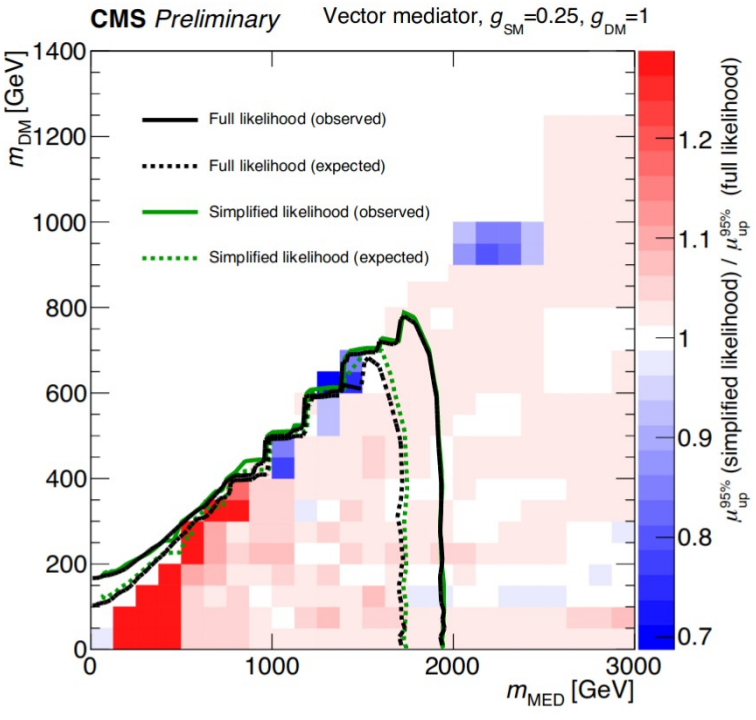
Simplified likelihood: Validation in previous monojet/V result

Compare simplified and full likelihood results

Likelihood ratio test statistic



Exclusion limits



Good closure
Significantly better than without correlation info

Simplified likelihood documentation