ATLAS approach to releasing likelihoods for reinterpretations

ReINPS2021

Eric Schanet

on behalf of the ***ATLAS*** collaboration

February 15, 2021

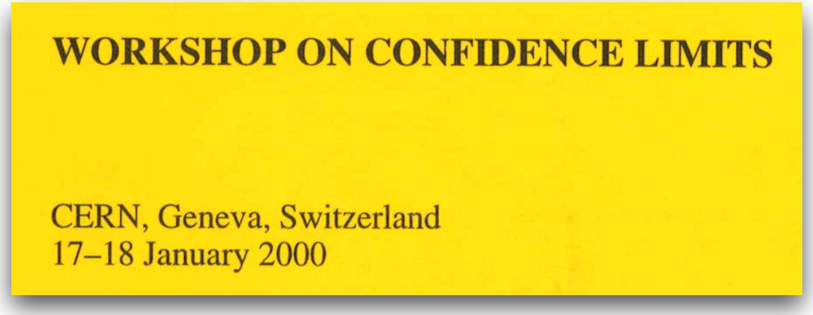
Luxembourg National Research Fund

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ATLAS EXPERIMENT

Publishing likelihoods

- **General agreement over importance of publishing likelihoods (LHs)**
  - Already back in 2000!


WORKSHOP ON CONFIDENCE LIMITS

CERN, Geneva, Switzerland
17–18 January 2000

- **Why is it important?**
  - Likelihood one of the most important data products of HEP analyses.
  - Nearly everything in an analysis affects the LH (trigger, detector, systematic uncertainties, event selection, …)
  - Most of the analysis products we publish on HEPdata are lossy projections of our LHs.
    - Theorists still have to do “guesses” to build realistic LH.

Massimo Corradi

It seems to me that there is a general consensus that what is really meaningful for an experiment is *likelihood*, and almost everybody would agree on the prescription that experiments should give their likelihood function for these kinds of results. Does everybody agree on this statement, to publish likelihoods?

Louis Lyons

Any disagreement ? Carried unanimously. That’s actually quite an achievement for this Workshop.

https://cds.cern.ch/record/452080

- **Actually publishing/preserving is tricky though …**
  - What do we want to preserve exactly? And how? In what format?
  - Do not really have a software-independent format of the LH to put on HEPdata …

➡ **Start with a single more tractable model first:** HistFactory

HistFactory

CERN-OPEN-2012-016

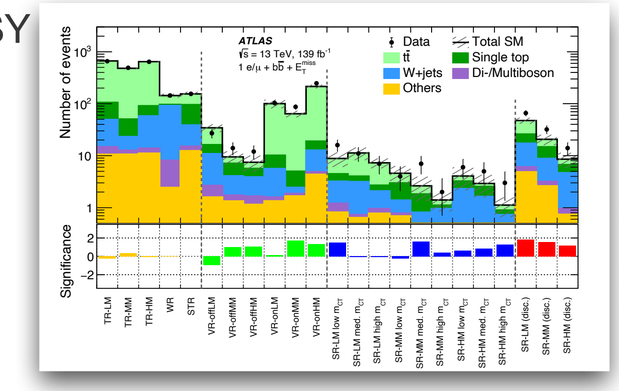
- **Flexible pdf template for building statistical models from binned distributions**
  - Widely used in HEP community for SM measurements as well as BSM searches.

- **Uses surprisingly simple formula to model full LH**

$$f(\boldsymbol{n}, \boldsymbol{a} \mid \boldsymbol{\eta}, \boldsymbol{\chi}) = \underbrace{\prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}\left(n_{cb} \mid \nu_{cb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\text{Simultaneous measurement of multiple channels}} \underbrace{\prod_{\chi \in \boldsymbol{\chi}} c_\chi(a_\chi \mid \chi)}_{\text{constraint terms for “auxiliary measurements”}} ,$$

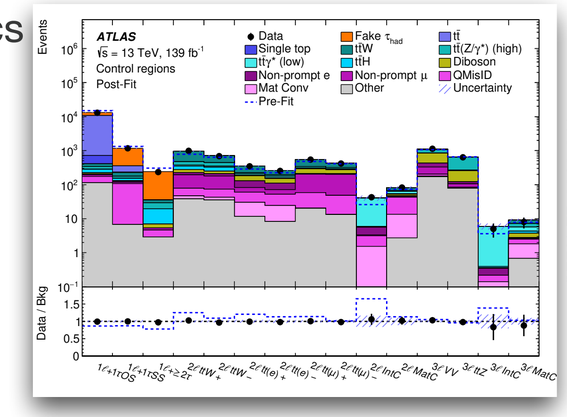
  - No software specification defined by the above formula!
  - Historically: HistFactory only implemented in RooStats/RooFit
  - Some downsides to this:
    - Requires knowledge in ROOT.
    - No straightforward interface to modern tools for minimisation, computation of pdf.
    - Data for LH only needs arrays of floats, not entire histograms in binary ROOT format

SUSY



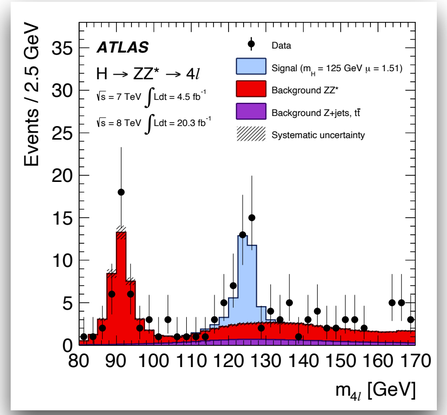
Eur. Phys. J. C 80 (2020) 691

Exotics



CERN-EP-2020-241

SM



Phys. Rev. D 91, 012006 (2015)

Enter **pyhf**

- **Pure-python implementation of HistFactory pdf templates**
  - Developed by Lukas Heinrich, Matthew Feickert, Giordon Stark and Kyle Cranmer.
  - Part of Scikit-HEP, source code available on Github.
  - Already being used in various publications (also outside the collaboration).

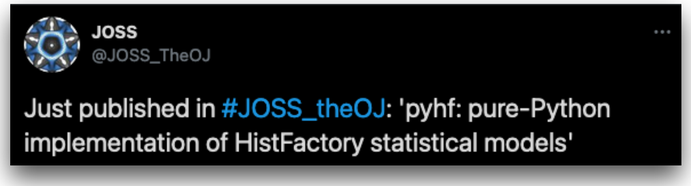
pyhf
∂ifferentiable
ℒikelihoods

DOI 10.5281/zenodo.4484948

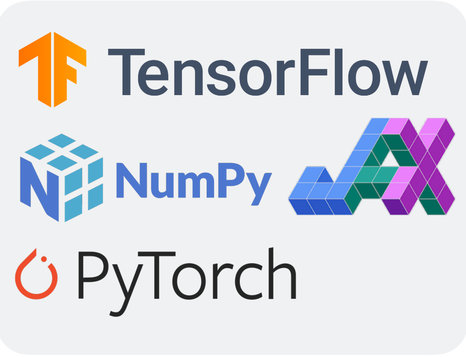
JOSS 10.21105/joss.02823

- **Some nice design features**
  - Numeric operations implemented through thin layer of n-D array operations powered by various tensor algebra backends
  - Supports modern computational graph libraries like PyTorch, TensorFlow, JAX
    - Auto-differentiation of full gradient of likelihood, hardware acceleration, …

The Journal of Open Source Software



Computational backends



- **Comes with JSON specification fully describing HistFactory template**
  - ATLAS is starting to publish these!
  - ATLAS PubNote on JSON schema: ATL-PHYS-PUB-2019-029

Simple likelihood example in **JSON format**

JSON defining two-bin single-channel counting experiment with systematics

```
{
    "channels": [ // List of regions
        { "name": "singlechannel",
          "samples": [ // List of samples in region
            { "name": "signal",
              "data": [5.0, 10.0],
             // List of rate factors and systematic uncertainties
              "modifiers": [ { "name": "mu", "type": "normfactor", "data": null} ]
            },
            { "name": "background",
              "data": [50.0, 60.0],
              "modifiers": [ {"name": "uncorr_bkguncrt", "type": "shapesys", "data": [5.0, 12.0]} ]
            }
          ]
        }
    ],
    "observations": [ // Observed data
        { "name": "singlechannel", "data": [50.0, 60.0] }
    ],
    "measurements": [ // Parameter of Interest and additional configuration
        { "name": "Measurement", "config": {"poi": "mu", "parameters": []} }
    ],
    "version": "1.0.0" // Version of JSON spec
}
```

- **Industry standard**
  - JSON is not going away anytime soon!
- **Human and machine readable**
  - Highly portable, no lock in, not implementation-dependent.
- **Perfect for preserving analyses**
  - Highly compressible
  - Can be put under version control.
  - JSON supported by HEPData.
    - Ongoing effort to natively support JSON likelihoods (#163, #164)

JSON likelihoods and **reinterpretations**

- **Natively built-in due to JSON patches (RFC 6902)**
  - Test new theory with a simple patch on top of original analysis LH

Original analysis (model A) Reinterpreted analysis (model B)

- **Very simple to do using pyhf and JSON patches**

Simple JSON patch for signal model B

```
$ curl -sL https://git.io/JtEl0
[{
    "op": "replace",
    "path": "/channels/0/samples/0/data",
    "value": [10.0,6.0]
}]
```

Interpreting analysis in different signal models

```
# Original likelihood
$ curl -sL https://git.io/JtElR | pyhf cls | jq .CLs_obs
0.05251497423736956

# Patching original likelihood with new model patch
$ curl -sL https://git.io/JtElR | pyhf cls --patch <(curl -sL https://git.io/JtEl0) | jq .CLs_obs
0.15582915780714504
```

  - How to get an estimation of the signal yields?
    - E.g. through Rivet, ATLAS SimpleAnalysis (truth), or ATLAS RECAST (full analysis chain)

Published likelihoods

- **Published on HEPdata**
  - Background-only model as JSON + signal patches
- **Official collaboration policy?**
  - No, but a lot of support/feedback from hep-ph and hep-th.
  - **SUSY WG:** Analyses are actively encouraged to publish, but it is not a strict publication requirement.

Discover likelihoods on ATLAS public results page

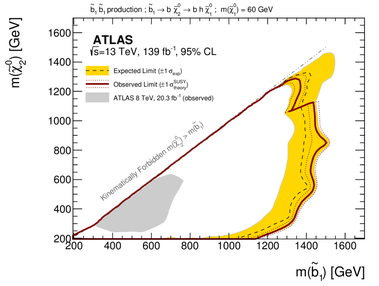
Applied filters:

Keywords: Likelihood available

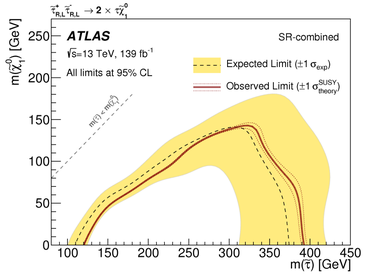
Short Title	Group	Journal Reference	Date	√s (TeV)	L	Links
Search for displaced leptons	SUSY	Submitted to PRL	13-NOV-20	13	139 fb^{-1}	Documents | 2011.07812 | Inspire HepData | Briefing | Internal
Chargino-neutralino pair; 3 leptons, weak-scale mass splittings	SUSY	Phys. Rev. D 101 (2020) 072001	18-DEC-19	13	139 fb^{-1}	Documents | 1912.08479 | Inspire HepData | Internal
Staus; taus	SUSY	Phys. Rev. D 101 (2020) 032009	15-NOV-19	13	139 fb^{-1}	Documents | 1911.06660 | Inspire HepData | Briefing | Internal
Chargino-neutralino pair; Higgs boson in final state, 2 b-jets and 1 lepton	SUSY	Eur. Phys. J. C 80 (2020) 691	19-SEP-19	13	139 fb^{-1}	Documents | 1909.09226 | Inspire HepData | Internal
Stop pair, sbottom pair, gluino pair; two same-sign leptons or three leptons	SUSY	JHEP 06 (2020) 46	18-SEP-19	13	139 fb^{-1}	Documents | 1909.08457 | Inspire HepData | Internal
Sbottom; b-jets	SUSY	JHEP 12 (2019) 060	08-AUG-19	13	139 fb^{-1}	Documents | 1908.03122 | Inspire HepData | Briefing | Internal

note: some of these actually have more than one published contour

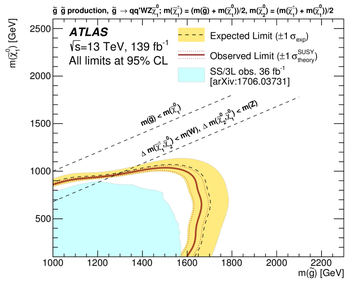
DOI 10.17182/hepdata.89408.v2



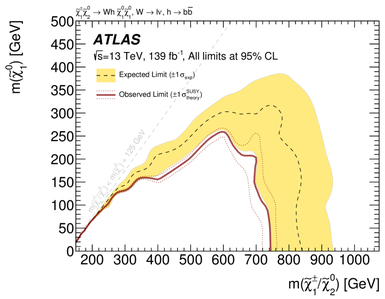
DOI 10.17182/hepdata.92006.v2



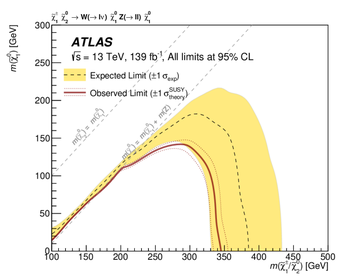
DOI 10.17182/hepdata.91214.v3



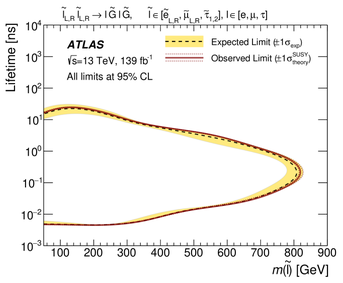
DOI 10.17182/hepdata.90607.v3



DOI 10.17182/hepdata.91127.v2



DOI 10.17182/hepdata.98796.v2

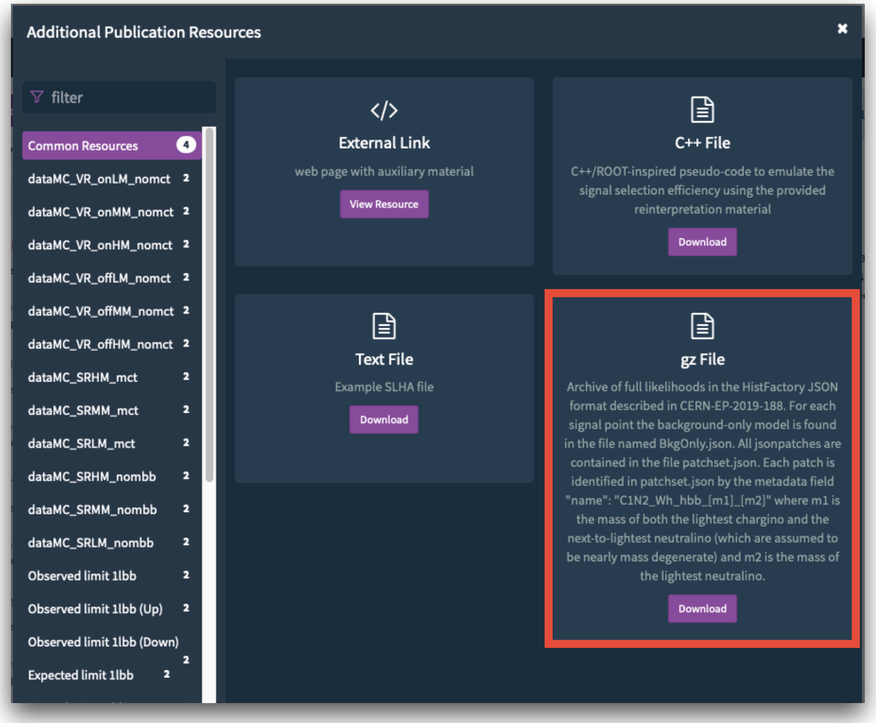


Published likelihoods

- **Published on HEPData**
  - Contains:
    - background-only model in JSON format (BkgOnly.json)
    - patchset file containing the $\mathcal{O}(10^2)$ original signal models

You can get the likelihoods e.g. using their HEPData DOI

```
$ pyhf contrib download https://doi.org/10.17182/hepdata.90607.v3/r3 1Lbb-likelihoods
$ cd 1Lbb-likelihoods && ls
BkgOnly.json   README.md      patchset.json
$ pyhf cls BkgOnly.json \
> --backend pytorch \
> --patch <(pyhf patchset extract patchset.json --name "C1N2_Wh_hbb_700_50") \
> | jq .CLs_obs
0.020306234596335727
```

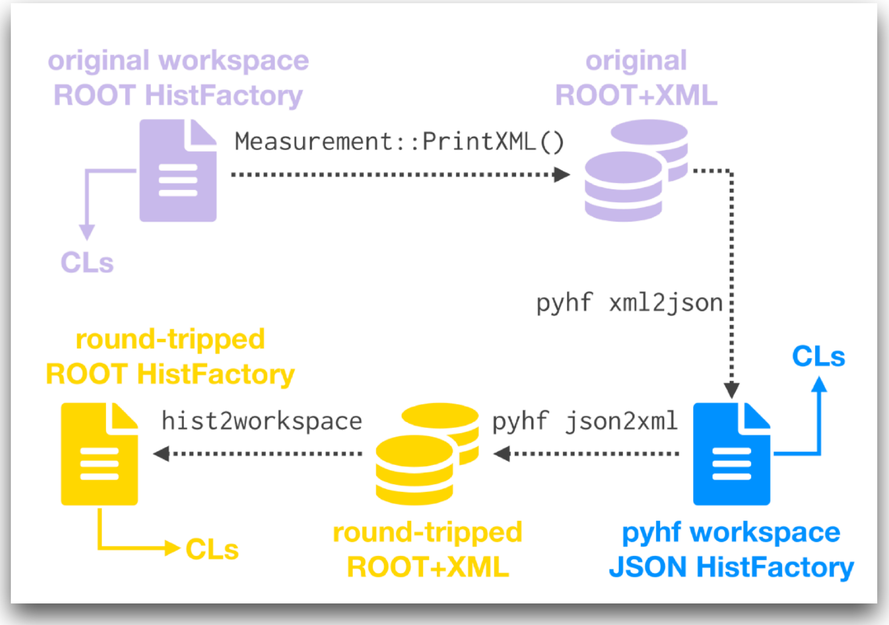


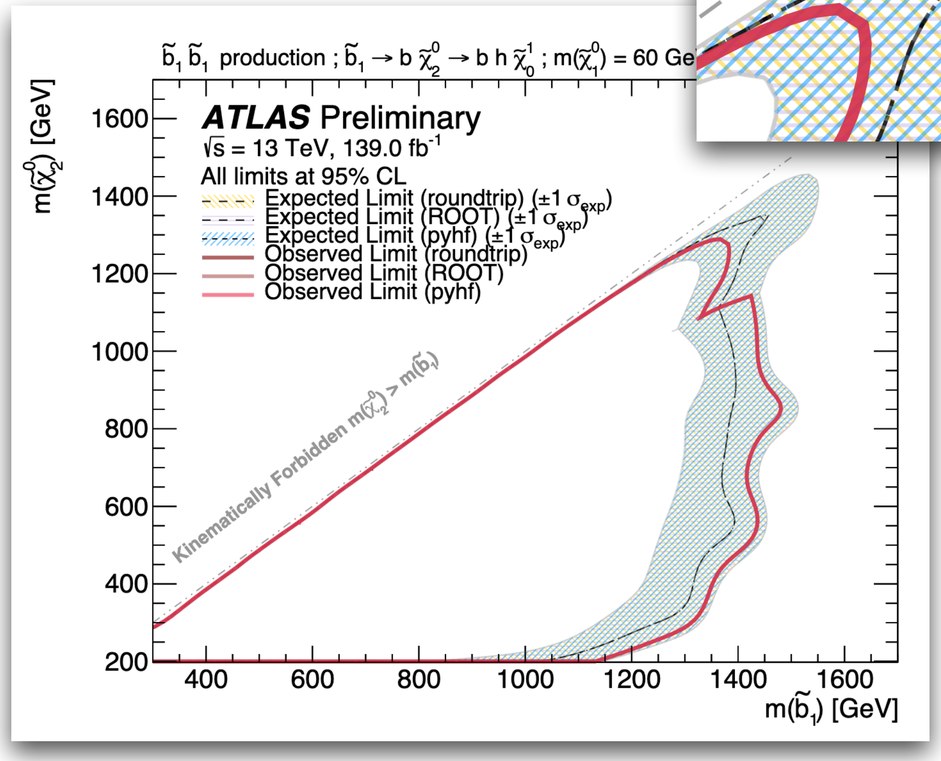
Cross-checking **ROOT** and **pyhf**

ATL-PHYS-PUB-2019-029

DOI 10.17182/hepdata.89408.v2

- **Do JSON+python and XML+ROOT implementations of HistFactory give same results?**
  - Short answer: yes!
  - pyhf has routines to go back and forth between JSON LHs and ROOT workspaces.
  - Significant speed-up using pyhf
    - **ROOT:** 10+ hours (fitting + hypothesis tests)
    - **pyhf:** 30min





Simplified likelihoods

- **Sometimes full likelihoods are just too complex**
  - Either because full statistical precision not needed, or too CPU expensive.

  ➡ Full LHs contain all the information needed for creating simplified LHs!

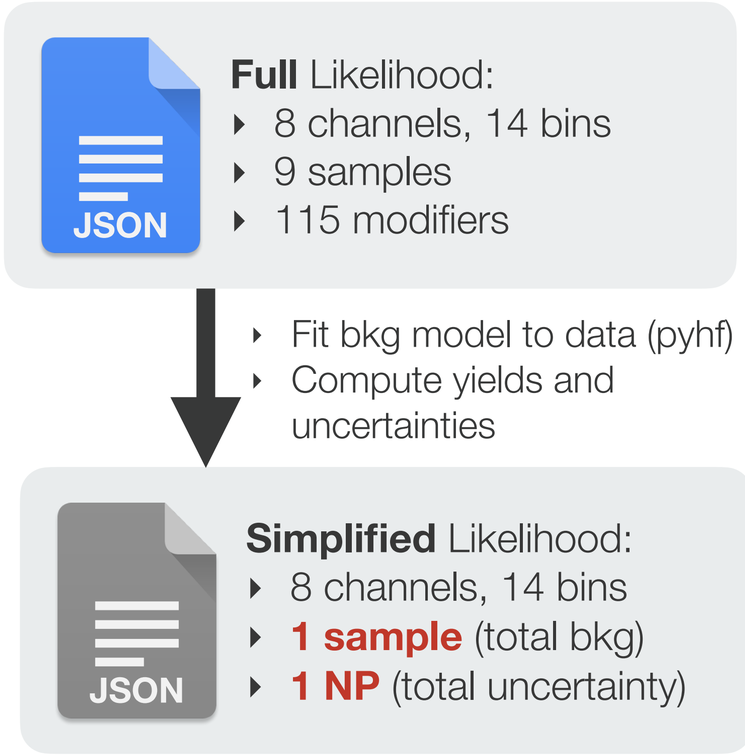
Python tool: simplify

ATLAS-SUSY-2019-08

Example: DOI 10.17182/hepdata.90607.v3

- **Proof-of-concept python tool producing simplified JSON LHs available**

```
$> pip install simplify-hep
$> curl http://foo/likelihood.json | simplify convert
```

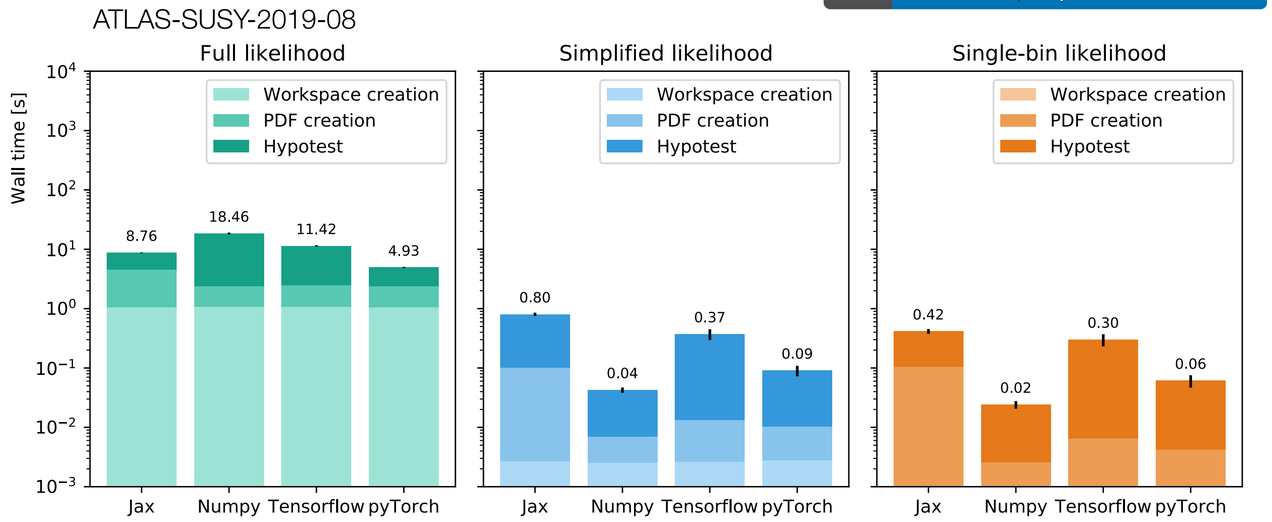


  - Currently implements one version of simplified LHs:
    - 1 background sample (total fitted background), 1 constrained parameter (bin-wise total uncertainty), obtained from fit in full likelihood.
    - Not all likelihoods can be simplified using this naive approach.
  - Can be extended to support other forms of simplified likelihoods.

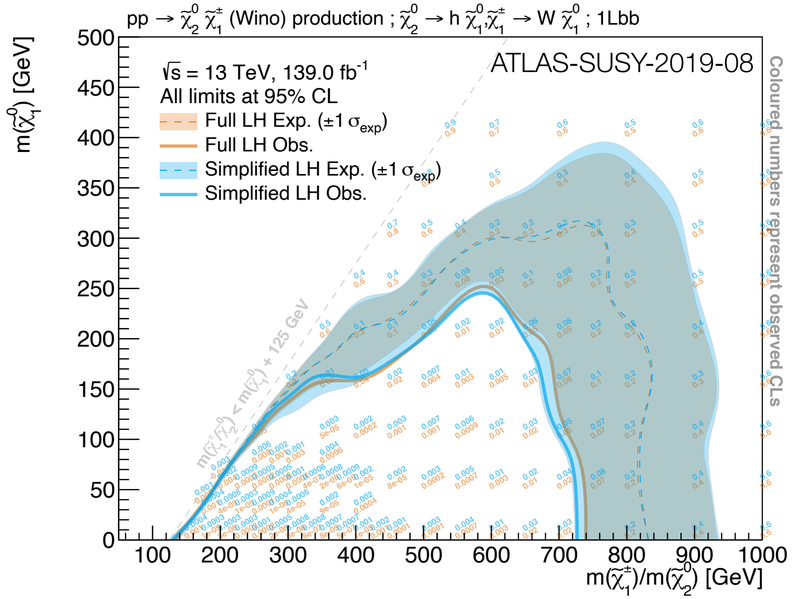
Performance of simplified likelihoods

Analysis paper DOI 10.1140/epjc/s10052-020-8050-3
Likelihood DOI 10.17182/hepdata.90607.v3

- **Computational performance**
  - Simplified LH offers factor $\mathcal{O}(10^2)$ in speed-up (benchmarked on CPU)
    - PyTorch fastest for the full LH,
    - Simplified LH speed in same order of magnitude than model-independent limits (single-bin).



- **Physics performance**
  - First application on ATLAS SUSY analyses shows accurate reproduction of full analysis results.
  - DIY: This is only using public information
  - Caveat: Not every workspace can be simplified in this straightforward way



Summary

- **Plain-text JSON schema for full likelihoods!** ATL-PHYS-PUB-2019-029
  - Implementation-independent.
  - Versionable, highly compressible → optimal for long-term archival on e.g. HEPdata.
  - Native support for reinterpretations through JSON patches.

- **Towards fully reproducible analyses**
  - Significant interest/support from both experimental and theory communities.
  - Considerable increase in scientific impact of our analyses!
  - Together with tools like, ATLAS RECAST / SimpleAnalysis, provides powerful setup for reinterpretations.

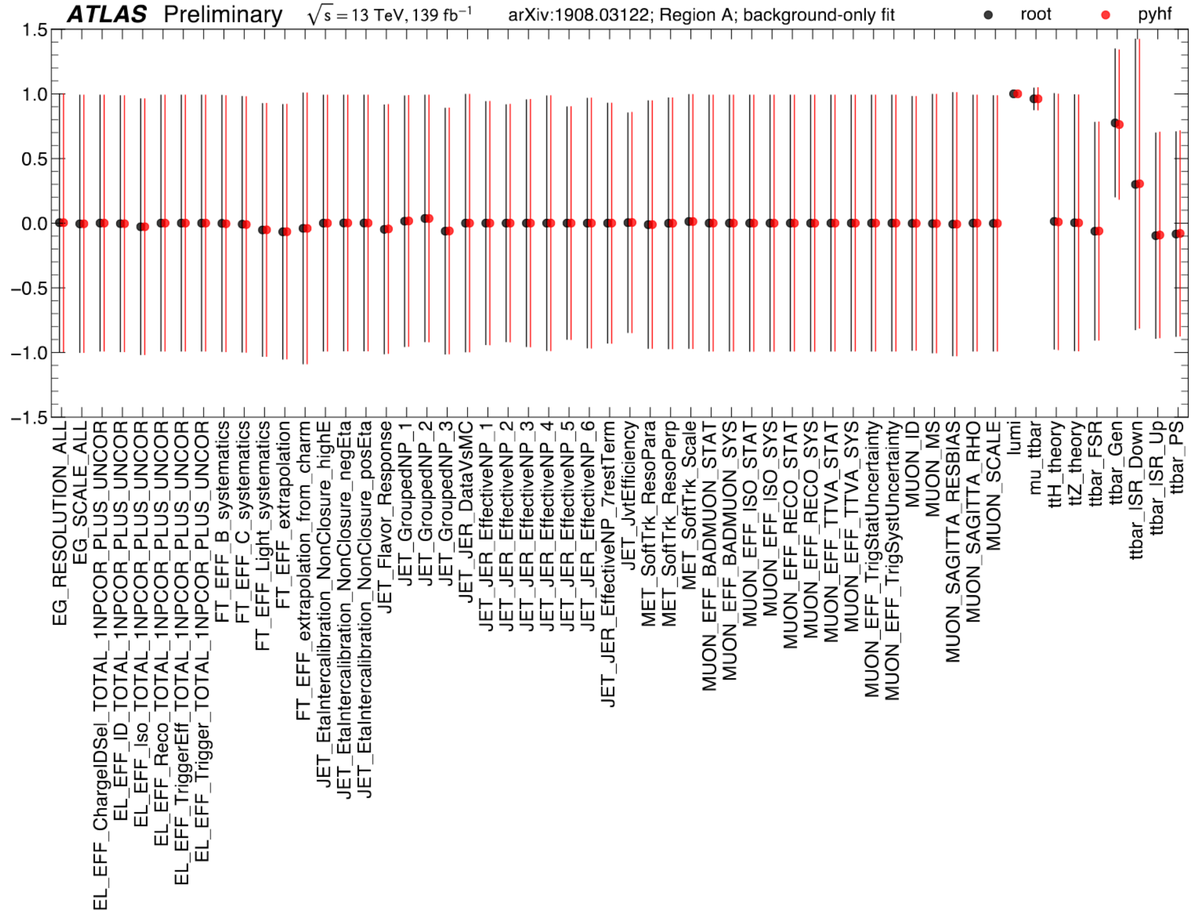
- **Simplified likelihoods**
  - No more “guessing” based on lossy projections of the LH analysis usually publish on HEPdata.
  - ATLAS is already using simplified LHs internally; but tool to generate them from full LH is public.

BACKUP

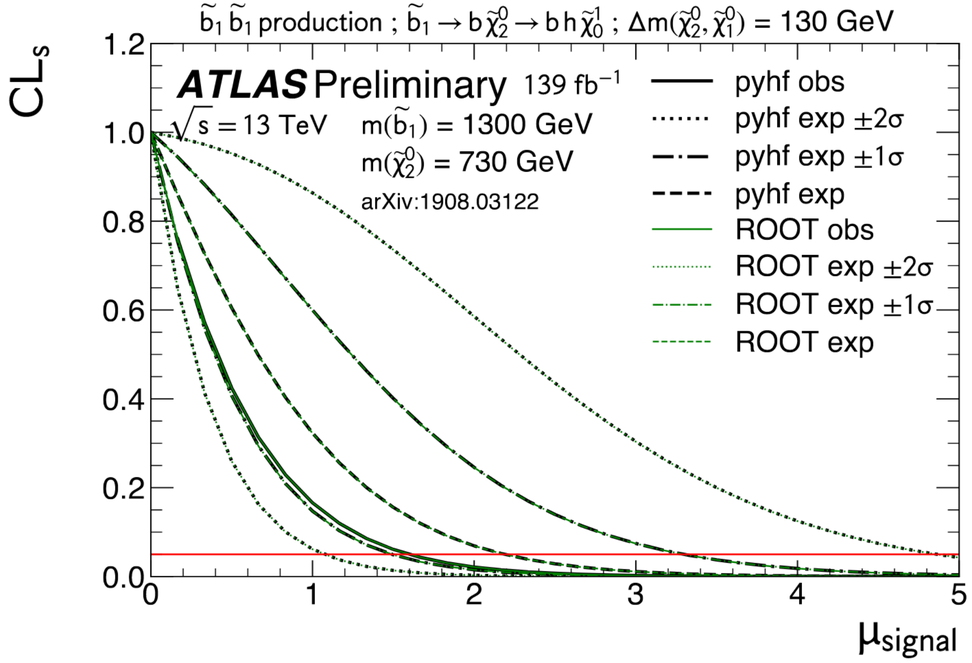
Cross-checking **ROOT** and **pyhf**

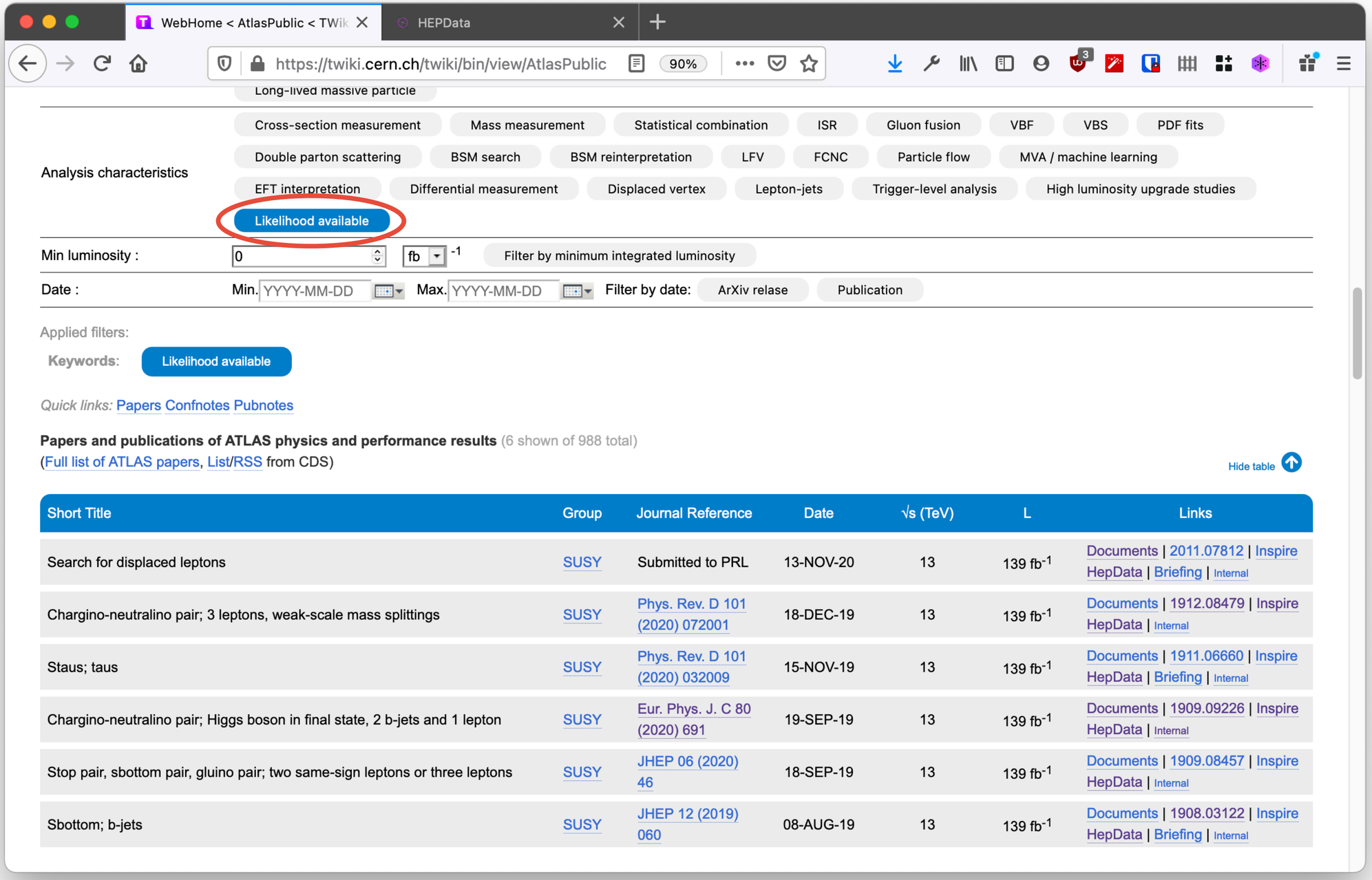
ATL-PHYS-PUB-2019-029

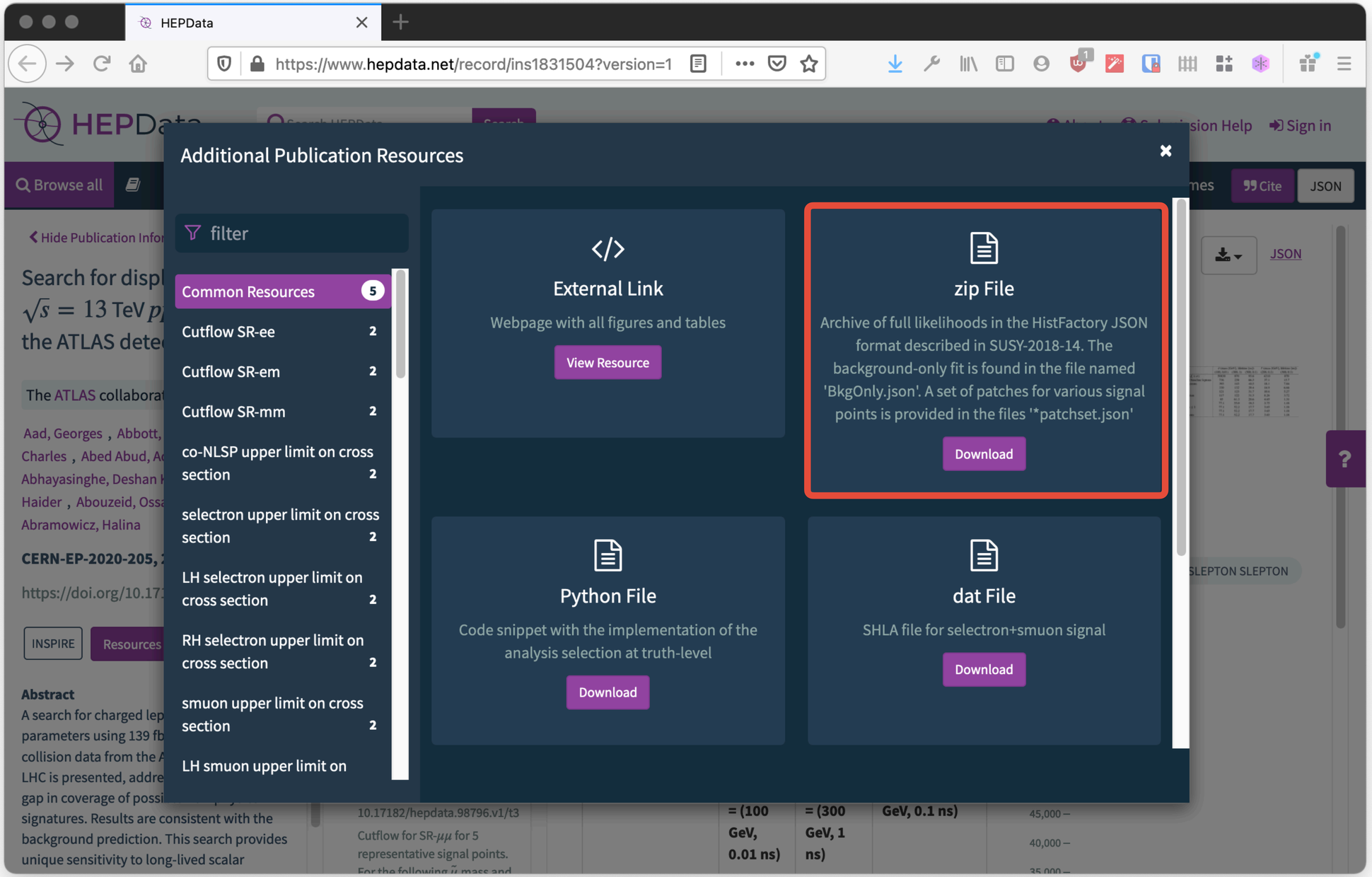
- Systematic pulls
  - No large differences between XML+ROOT and JSON+pyhf



- Signal strength scan
  - Again no difference




WebHome < AtlasPublic < TWik
HEPData
https://twiki.cern.ch/twiki/bin/view/AtlasPublic
90%
Long-lived massive particle
Analysis characteristics
Cross-section measurement
Mass measurement
Statistical combination
ISR
Gluon fusion
VBF
VBS
PDF fits
Double parton scattering
BSM search
BSM reinterpretation
LFV
FCNC
Particle flow
MVA / machine learning
EFT interpretation
Differential measurement
Displaced vertex
Lepton-jets
Trigger-level analysis
High luminosity upgrade studies
Likelihood available
Min luminosity :
0
fb -1
Filter by minimum integrated luminosity
Date :
Min. YYYY-MM-DD
Max. YYYY-MM-DD
Filter by date:
ArXiv relase
Publication
Applied filters:
Keywords:
Likelihood available
Quick links: Papers Confnotes Pubnotes
Papers and publications of ATLAS physics and performance results (6 shown of 988 total)
(Full list of ATLAS papers, List/RSS from CDS)
Hide table
Short Title | Group | Journal Reference | Date | √s (TeV) | L | Links
Search for displaced leptons | SUSY | Submitted to PRL | 13-NOV-20 | 13 | 139 fb-1 | Documents | 2011.07812 | Inspire HepData | Briefing | Internal
Chargino-neutralino pair; 3 leptons, weak-scale mass splittings | SUSY | Phys. Rev. D 101 (2020) 072001 | 18-DEC-19 | 13 | 139 fb-1 | Documents | 1912.08479 | Inspire HepData | Internal
Staus; taus | SUSY | Phys. Rev. D 101 (2020) 032009 | 15-NOV-19 | 13 | 139 fb-1 | Documents | 1911.06660 | Inspire HepData | Briefing | Internal
Chargino-neutralino pair; Higgs boson in final state, 2 b-jets and 1 lepton | SUSY | Eur. Phys. J. C 80 (2020) 691 | 19-SEP-19 | 13 | 139 fb-1 | Documents | 1909.09226 | Inspire HepData | Internal
Stop pair, sbottom pair, gluino pair; two same-sign leptons or three leptons | SUSY | JHEP 06 (2020) 46 | 18-SEP-19 | 13 | 139 fb-1 | Documents | 1909.08457 | Inspire HepData | Internal
Sbottom; b-jets | SUSY | JHEP 12 (2019) 060 | 08-AUG-19 | 13 | 139 fb-1 | Documents | 1908.03122 | Inspire HepData | Briefing | Internal


HEPData
https://www.hepdata.net/record/ins1831504?version=1
Browse all
Cite
JSON
Hide Publication Infor
Search for displ
$\sqrt{s} = 13$ TeV p
the ATLAS dete
The ATLAS collaborat
Aad, Georges , Abbott,
Charles , Abed Abud, A
Abhayasinghe, Deshan
Haider , Abouzeid, Oss
Abramowicz, Halina
CERN-EP-2020-205,
https://doi.org/10.17
INSPIRE
Resources
Abstract
A search for charged lep
parameters using 139 fb
collision data from the A
LHC is presented, addre
gap in coverage of possi
signatures. Results are consistent with the
background prediction. This search provides
unique sensitivity to long-lived scalar
10.17182/hepdata.98796.v1/t3
Cutflow for SR-$\mu\mu$ for 5
representative signal points.
= (100 GeV, 0.01 ns)
= (300 GeV, 1 ns)
GeV, 0.1 ns)
45,000
40,000
Additional Publication Resources
filter
Common Resources 5
Cutflow SR-ee 2
Cutflow SR-em 2
Cutflow SR-mm 2
co-NLSP upper limit on cross section 2
selectron upper limit on cross section 2
LH selectron upper limit on cross section 2
RH selectron upper limit on cross section 2
smuon upper limit on cross section 2
LH smuon upper limit on
External Link
Webpage with all figures and tables
View Resource
zip File
Archive of full likelihoods in the HistFactory JSON format described in SUSY-2018-14. The background-only fit is found in the file named 'BkgOnly.json'. A set of patches for various signal points is provided in the files '*patchset.json'
Download
Python File
Code snippet with the implementation of the analysis selection at truth-level
Download
dat File
SHLA file for selectron+smuon signal
Download

The HistFactory template [1/2]

$$f(\boldsymbol{n}, \boldsymbol{a} \mid \boldsymbol{\eta}, \boldsymbol{\chi}) = \underbrace{\prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}\left(n_{cb} \mid \nu_{cb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\substack{\text{Simultaneous measurement} \\ \text{of multiple channels}}} \underbrace{\prod_{\chi \in \boldsymbol{\chi}} c_{\chi}(a_{\chi} \mid \chi)}_{\substack{\text{constraint terms} \\ \text{for “auxiliary measurements”}}} ,$$

- Analysis-specific model terms describing channels with observed events n_{cb} given expected events $\nu_{cb}(\boldsymbol{\eta},\boldsymbol{\chi})$.
- Analysis-independent constraint terms for constrained parameters $\boldsymbol{\chi}$.

$$\nu_{cb}\left(\boldsymbol{\phi}\right) = \sum_{s \in \text{samples}} \nu_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right) = \sum_{s \in \text{samples}} \underbrace{\left(\prod_{\kappa \in \boldsymbol{\kappa}} \kappa_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\text{multiplicative modifiers}} \left(\nu^{0}_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right) + \underbrace{\sum_{\Delta \in \boldsymbol{\Delta}} \Delta_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)}_{\text{additive modifiers}}\right)$$

- Sample rates ν_{scb} with nominal rate ν^{0}_{scb}
- Additive and multiplicative rate modifiers $\boldsymbol{\Delta}(\boldsymbol{\phi})$ and $\boldsymbol{\kappa}(\boldsymbol{\phi})$

The HistFactory template [2/2]

- Rate modifiers

ATL-PHYS-PUB-2019-029

	Description	Modification	Constraint Term c_χ	Input
constrained	Uncorrelated Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Pois}\left(r_b = \sigma_b^{-2} \middle\vert \rho_b = \sigma_b^{-2}\gamma_b\right)$	σ_b
	Correlated Shape	$\Delta_{scb}(\alpha) = f_p\left(\alpha \middle\vert \Delta_{scb,\alpha=-1}, \Delta_{scb,\alpha=1}\right)$	$\text{Gaus}\,(a = 0 \vert\, \alpha, \sigma = 1)$	$\Delta_{scb,\alpha=\pm1}$
	Normalisation Unc.	$\kappa_{scb}(\alpha) = g_p\left(\alpha \middle\vert \kappa_{scb,\alpha=-1}, \kappa_{scb,\alpha=1}\right)$	$\text{Gaus}\,(a = 0 \vert\, \alpha, \sigma = 1)$	$\kappa_{scb,\alpha=\pm1}$
	MC Stat. Uncertainty	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Gaus}\left(a_{\gamma_b} = 1 \middle\vert \gamma_b, \delta_b\right)$	$\delta_b^2 = \sum_s \delta_{sb}^2$
	Luminosity	$\kappa_{scb}(\lambda) = \lambda$	$\text{Gaus}\,(l = \lambda_0 \vert\, \lambda, \sigma_\lambda)$	$\lambda_0, \sigma_\lambda$
free	Normalisation	$\kappa_{scb}(\mu_b) = \mu_b$		
	Data-driven Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$		

- Summary of notation

	Symbol	Name
	$f(\boldsymbol{x}\vert\boldsymbol{\phi})$	model
	$\mathcal{L}(\boldsymbol{\phi})$	likelihood
data	$\boldsymbol{x} = \{\boldsymbol{n}, \boldsymbol{a}\}$	full dataset (including auxiliary data)
	$\boldsymbol{n}$	channel data (or event counts)
	$\boldsymbol{a}$	auxiliary data
	$\nu(\boldsymbol{\phi})$	calculated event rates
parameters	$\boldsymbol{\phi} = \{\boldsymbol{\eta}, \boldsymbol{\chi}\} = \{\boldsymbol{\psi}, \boldsymbol{\theta}\}$	all parameters
	$\boldsymbol{\eta}$	free parameters
	$\boldsymbol{\chi}$	constrained parameters
	$\boldsymbol{\psi}$	parameters of interest
	$\boldsymbol{\theta}$	nuisance parameters
rate modifiers	$\boldsymbol{\kappa}(\boldsymbol{\phi})$	multiplicative rate modifier
	$\boldsymbol{\Delta}(\boldsymbol{\phi})$	additive rate modifiers
	$c_\chi(a_\chi\vert\chi)$	constraint term for constrained parameter χ
	σ_χ	relative uncertainty in the constrained parameter