

# Likelihood Publication and Preservation

Matthew Feickert

(University of Illinois at Urbana-Champaign)



[matthew.feickert@cern.ch](mailto:matthew.feickert@cern.ch)

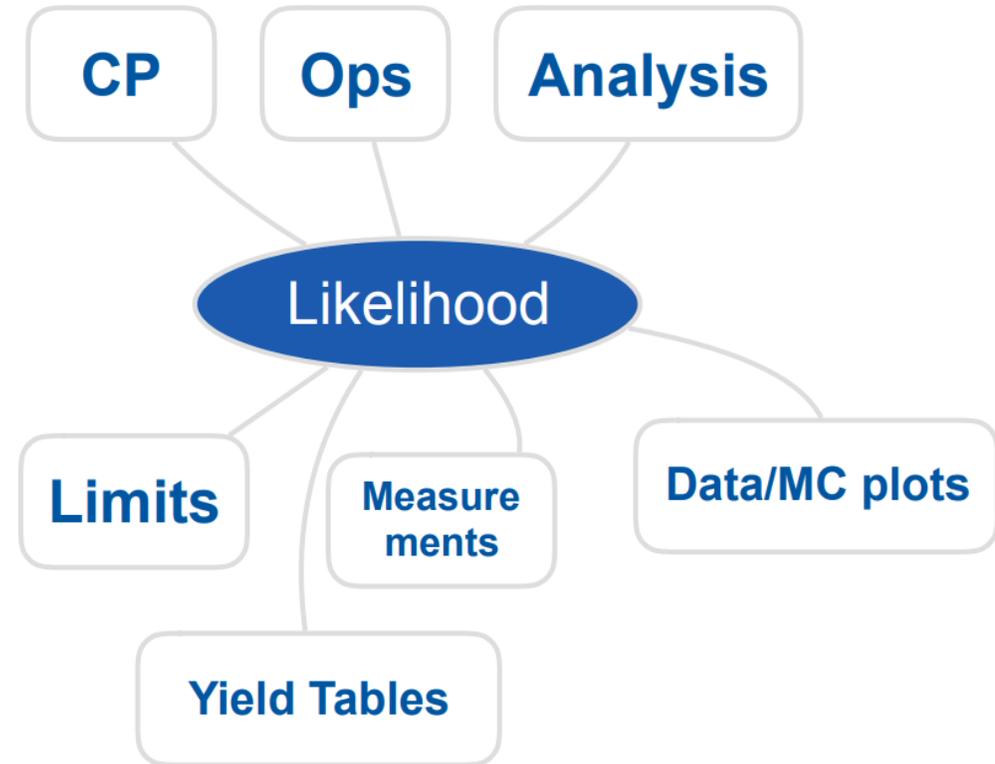
Snowmass 2021 Computational Frontier Workshop

August 10th, 2020



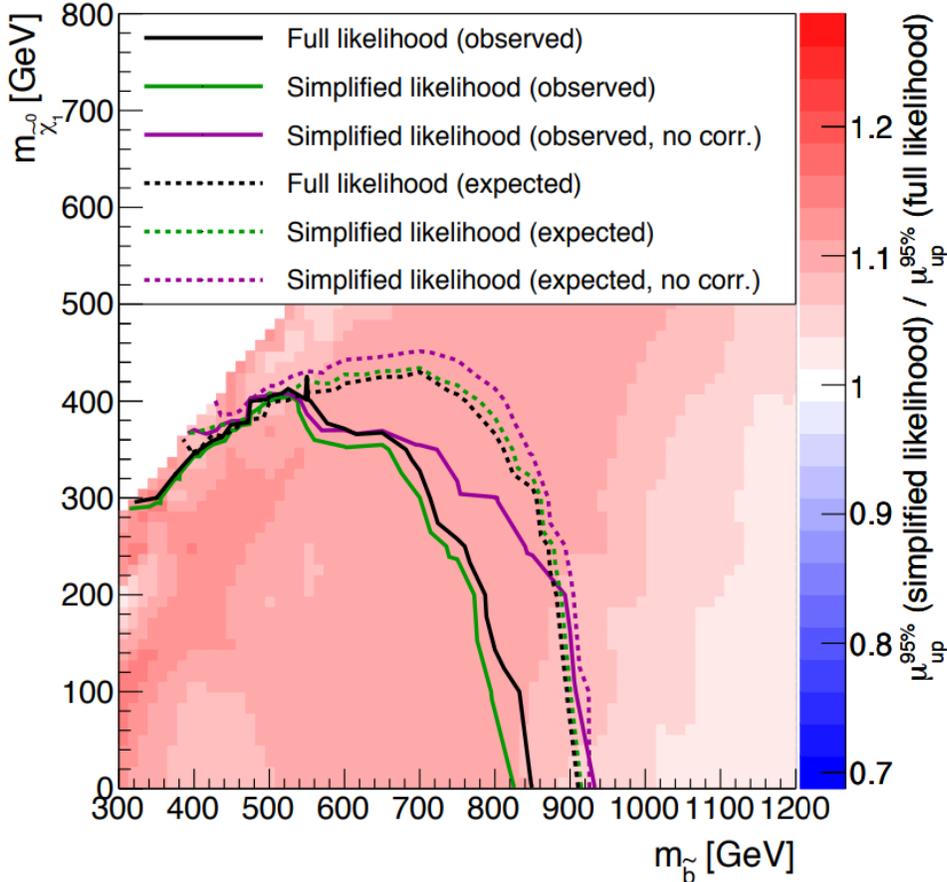
# Why is the likelihood important?

- High information-density summary of analysis
- Almost everything we do in the analysis ultimately affects the likelihood and is encapsulated in it
  - Trigger
  - Detector
  - Combined Performance / Physics Object Groups
  - Systematic Uncertainties
  - Event Selection
- Unique representation of the analysis to reuse and preserve

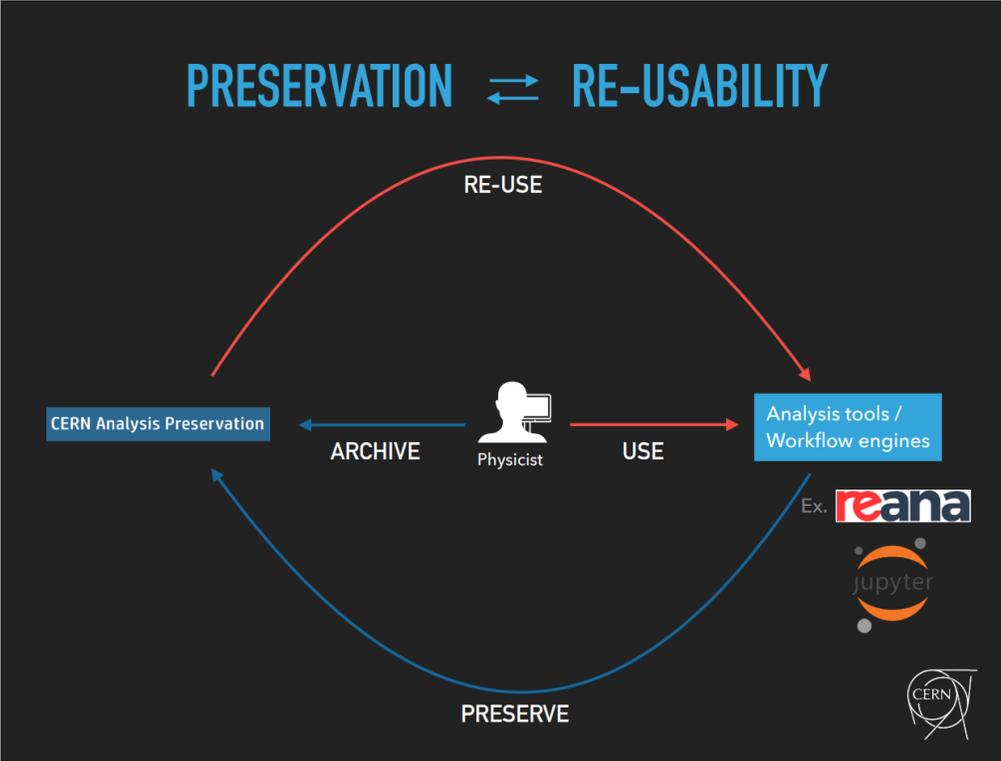


# Partial likelihoods have been published/preserved

**CMS Preliminary**



(CMS, 2017)



CERN Analysis Preservation implements FAIR data

(CERN, CHEP 2019)

# Full likelihood serialization...

...making good on [19 year old agreement to publish likelihoods](#)

## 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.](#)

[\(1st Workshop on Confidence Limits, CERN, 2000\)](#)

## This hadn't been done in HEP until 2019

- In an "open world" of statistics this is a difficult problem to solve
- What to preserve and how? All of ROOT?
- Idea: Focus on a single more tractable binned model first

# Enter HistFactory and pyhf

$$f(\text{data}|\text{parameters}) = f(\vec{n}, \vec{a}|\vec{\eta}, \vec{\chi}) = \prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}(n_{cb}|\nu_{cb}(\vec{\eta}, \vec{\chi})) \prod_{x \in \vec{\chi}} c_x(a_x|\chi)$$

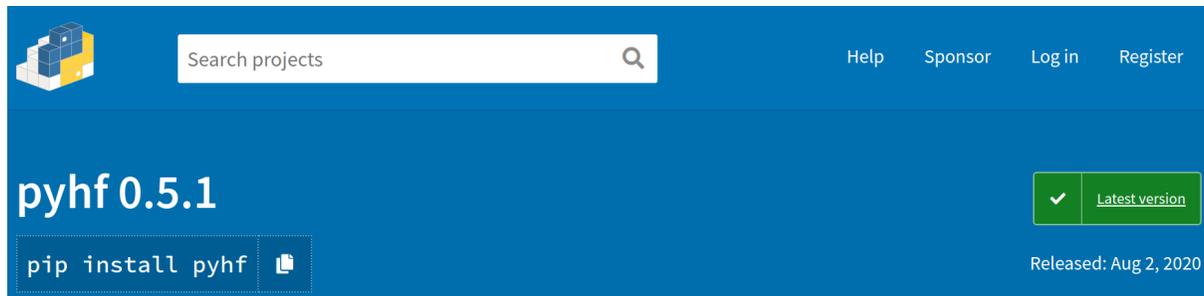
**Use:** Multiple disjoint **channels** (or regions) of binned distributions with multiple **samples** contributing to each with additional (possibly shared) systematics between sample estimates

HistFactory is used ubiquitously in binned analyses

Focus on this flexible p.d.f. template rather than "open world" of models

**This is a mathematical representation!** Nowhere is any software spec defined

**pyhf: HistFactory in pure Python hardware accelerated with autodiff**



Lukas  
Heinrich

CERN



Matthew  
Feickert

Illinois



Giordon  
Stark

UCSC SCIPP

# JSON spec fully describes the HistFactory model

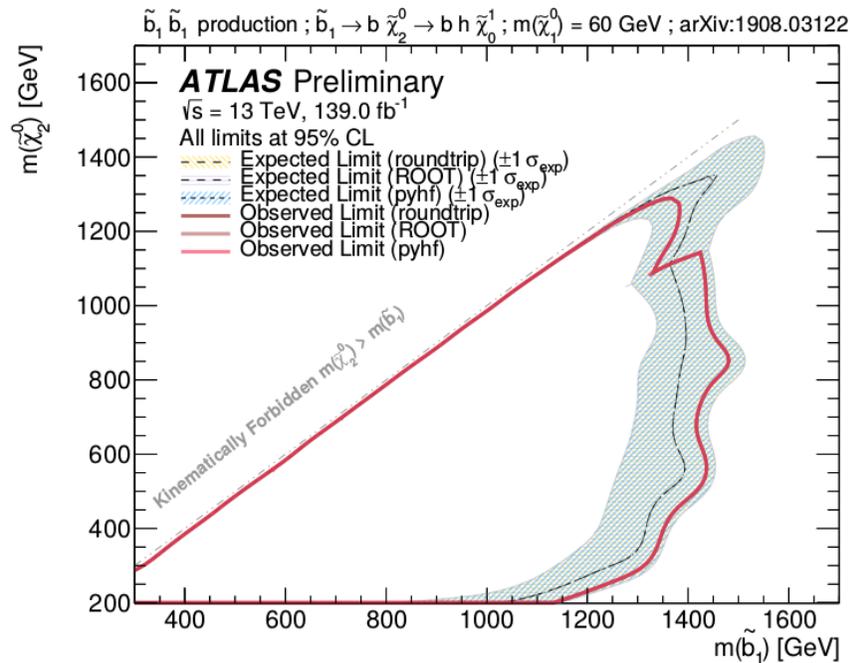
- Human & machine readable **declarative** statistical models
- Industry standard
  - Will be with us forever
- Parsable by every language
  - Highly portable
  - Bidirectional translation with ROOT
- Versionable and easily preserved
  - JSON Schema [describing HistFactory specification](#)
  - Attractive for analysis preservation
  - Highly compressible

```
{
  "channels": [ # List of regions
    { "name": "singlechannel",
      "samples": [ # List of samples in region
        { "name": "signal",
          "data": [20.0, 10.0],
          # List of rate factors and/or systematic uncertainties
          "modifiers": [ { "name": "mu", "type": "normfactor", "data": null} ]
        },
        { "name": "background",
          "data": [50.0, 63.0],
          "modifiers": [ {"name": "uncorr_bkguncrt", "type": "shapesys", "data": [5.0, 12.0]} ]
        }
      ]
    }
  ],
  "observations": [ # Observed data
    { "name": "singlechannel", "data": [55.0, 62.0] }
  ],
  "measurements": [ # Parameter of interest
    { "name": "Measurement", "config": {"poi": "mu", "parameters": []} }
  ],
  "version": "1.0.0" # Version of spec standard
}
```

JSON defining a single channel, two bin counting experiment with systematics

# ATLAS validation and publication of likelihoods

ATLAS Note	
Report number	ATL-PHYS-PUB-2019-029
Title	Reproducing searches for new physics with the ATLAS experiment through publication of full statistical likelihoods
Corporate Author(s)	The ATLAS collaboration

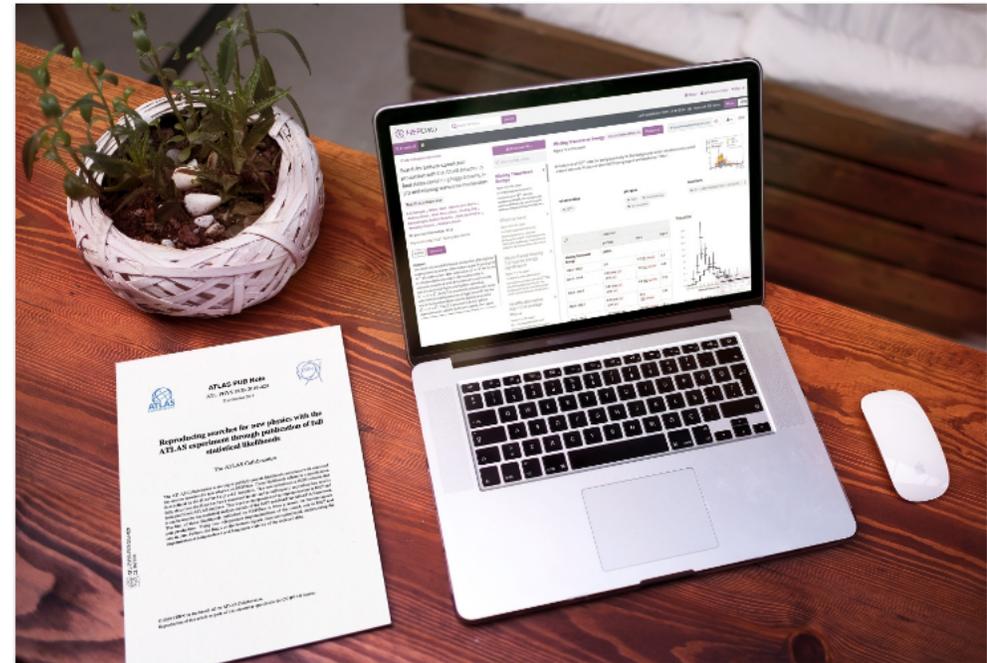


(ATLAS, 2019)

## New open release allows theorists to explore LHC data in a new way

The ATLAS collaboration releases full analysis likelihoods, a first for an LHC experiment

9 JANUARY, 2020 | By Katarina Anthony



Explore ATLAS open likelihoods on the HEPData platform (Image: CERN)

(CERN, 2020)

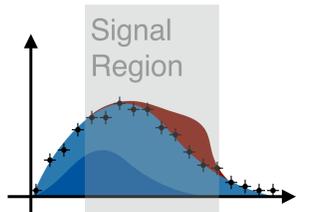
# JSON Patch for signal model (reinterpretation)

JSON Patch gives ability to **easily mutate model**

Think: test a **new theory** with a **new patch!**

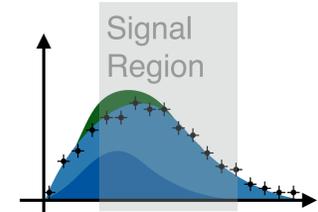
(c.f. [Lukas's RECAST talk](#))

Combined with RECAST gives powerful tool for **reinterpretation studies**



Signal model A

```
● ● ●  
  
# Using CLI  
$ pyhf cls example.json | jq .CLs_obs  
0.053994246621274014  
  
$ cat new_signal.json  
[  
  {  
    "op": "replace",  
    "path": "/channels/0/samples/0/data",  
    "value": [10.0, 6.0]  
  }  
]  
  
$ pyhf cls example.json --patch new_signal.json | jq .CLs_obs  
0.3536906623262466
```



Signal model B

# Likelihoods preserved on HEPData

- Background-only model JSON stored
- Hundreds of signal model JSON Patches stored together as a "patch set" file
- Together are able to publish and fully preserve the full likelihood (with own DOI! DOI [10.17182/hepdata.90607.v2/r2](https://doi.org/10.17182/hepdata.90607.v2/r2) )

HEPData Search HEPData

Search

Q Browse all

Hide Publication Information

Search for direct production of electroweakinos in final states with one lepton, missing transverse momentum and a Higgs boson decaying into two  $b$ -jets in (pp) collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector

The ATLAS collaboration

Aad, Georges , Abbott, Brad , Abbott, Dale Charles , Abed Abud, Adam , Abeling, Kira , Abhayasinghe, Deshan Kavishka , Abidi, Syed Haider , Abouzeid, Ossama , Abraham, Nicola , Abramowicz, Halina

PhD Thesis, 2020.

<https://doi.org/10.17182/hepdata.90607.v2>

INSPIRE Resources

**Abstract**

The results of a search for electroweakino pair production  $pp \rightarrow \tilde{\chi}_1^0 \tilde{\chi}_2^0$  in which the chargino ( $\tilde{\chi}_1^+$ ) decays into a  $W$  boson and the lightest neutralino ( $\tilde{\chi}_1^0$ ), while the heavier neutralino ( $\tilde{\chi}_2^0$ ) decays into the Standard Model 125 GeV Higgs boson and a second  $\tilde{\chi}_1^0$  are presented. The signal selection requires a pair of  $b$ -tagged jets consistent with those from a Higgs boson decay, and either an electron or a muon from the  $W$  boson decay, together with missing transverse momentum from the corresponding neutrino and the stable neutralinos. The analysis is based on data corresponding to  $139 \text{ fb}^{-1}$  of  $\sqrt{s} = 13$  TeV pp collisions provided by the Large Hadron Collider and recorded by the ATLAS detector. No statistically significant evidence of an excess of events above the Standard Model expectation is found. Limits are set on the direct production of the electroweakinos in simplified models, assuming pure wino cross-sections. Masses of  $\tilde{\chi}_1^+ / \tilde{\chi}_2^0$  up to 740 GeV are excluded at 95% confidence level for a massless  $\tilde{\chi}_1^0$ .

**Additional Publication Resources**

filter

**Common Resources**

- dataMC\_VR\_onLM\_nomct 2
- dataMC\_VR\_onMM\_nomct 2
- dataMC\_VR\_onHM\_nomct 2
- dataMC\_VR\_offLM\_nomct 2
- dataMC\_VR\_offMM\_nomct 2
- dataMC\_VR\_offHM\_nomct 2
- dataMC\_SRHM\_mct 2
- dataMC\_SRRM\_mct 2
- dataMC\_SRLM\_mct 2
- dataMC\_SRHM\_nombb 2
- dataMC\_SRRM\_nombb 2
- dataMC\_SRLM\_nombb 2
- Observed limit 1lbb 2
- Observed limit 1lbb (Up) 2
- Observed limit 1lbb (Down) 2
- Expected limit 1lbb 2
- Upper limits 1Lbb 2

**External Link**  
web page with auxiliary material  
View Resource

**C++ File**  
C++/ROOT-inspired pseudo-code to emulate the signal selection efficiency using the provided reinterpretation material  
Download

**Text File**  
Example SLHA file  
Download

**gz File**  
Archive of full likelihoods in the HistFactory JSON format described in CERN-EP-2019-188. For each signal point the background-only model is found in the file named BkgOnly.json. All jsonpatches are identified in patchset.json by the metadata field "name": "C1N2\_Wh\_hbb\_[m1]\_[m2]" where m1 is the mass of both the lightest chargino and the next-to-lightest neutralino (which are assumed to be nearly mass degenerate) and m2 is the mass of the lightest neutralino.  
Download

```
$ tree archive-likelihoods-hepdata
archive-likelihoods-hepdata
├── BkgOnly.json
├── patchset.json
└── README.md

0 directories, 3 file
```

# ...can be used from HEPData

- Background-only model JSON stored
- Hundreds of signal model JSON Patches stored together as a "patch set" file
- Together are able to publish and fully preserve the full likelihood (with own DOI! DOI [10.17182/hepdata.90607.v2/r2](https://doi.org/10.17182/hepdata.90607.v2/r2) )

```
# signal patchset for the SUSY EWK 1Lbb analysis
$ curl -sL https://www.hepdata.net/record/resource/1267798?view=true | tar -xzv
$ cd 1Lbb-likelihoods-hepdata

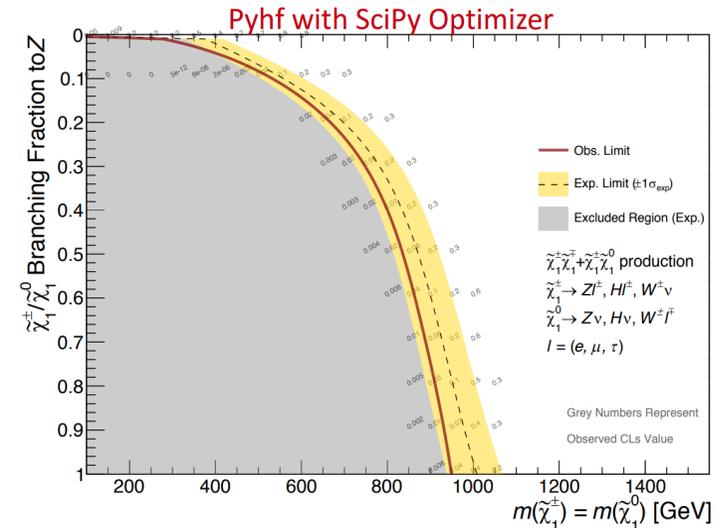
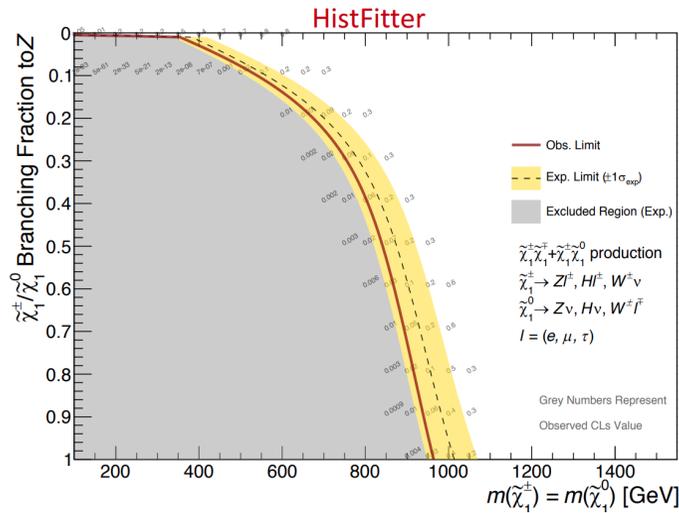
# verify patchset is valid
$ pyhf patchset verify BkgOnly.json patchset.json
All good.

# signal model: m1 = 900, m2 = 300 (chain CLI API output)
$ cat BkgOnly.json | \
  pyhf cls --patch <(pyhf patchset extract --name C1N2_Wh_hbb_900_300 patchset.json) | \
  jq .CLs_obs
0.5004154596172573

# new signal model: m1 = 900, m2 = 400 (use serialized CLI API output)
$ pyhf patchset extract --name C1N2_Wh_hbb_900_400 --output-file C1N2_Wh_hbb_900_400_patch.json patchset.json
$ pyhf cls --patch C1N2_Wh_hbb_900_400_patch.json BkgOnly.json | jq .CLs_obs
0.5735044179801038
```

# Rapid adoption in ATLAS...

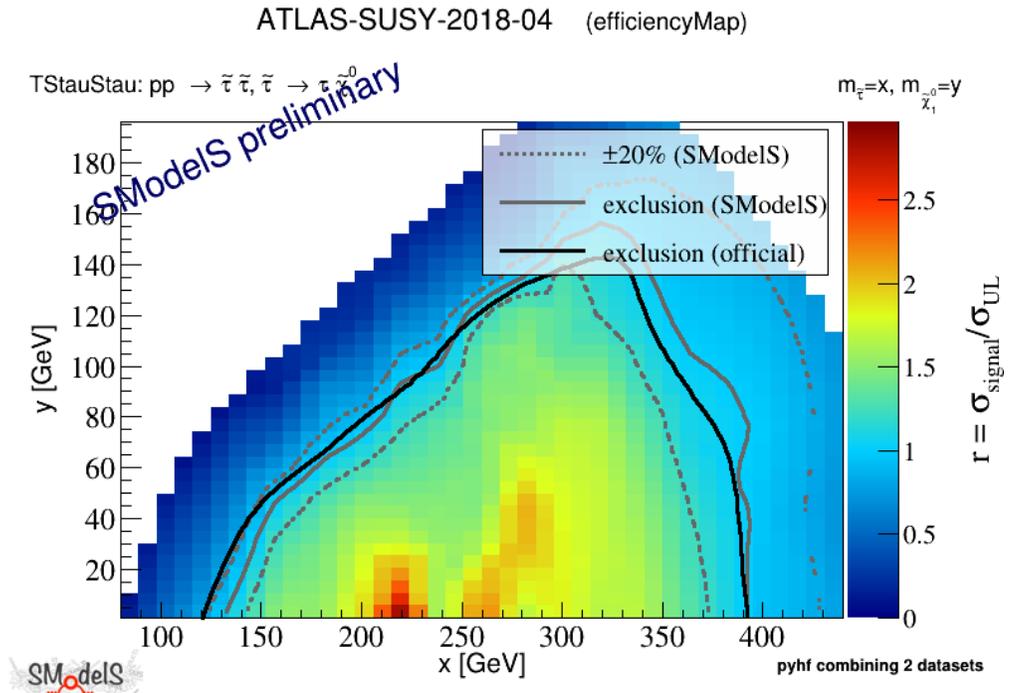
- Four ATLAS analyses with full likelihoods published to HEPData
- ATLAS SUSY will be continuing to publish full Run 2 likelihoods
- direct staus, [doi:10.17182/hepdata.89408](https://doi.org/10.17182/hepdata.89408) (2019)
- sbottom multi-b, [doi:10.17182/hepdata.91127](https://doi.org/10.17182/hepdata.91127) (2019)
- 1Lbb, [doi:10.17182/hepdata.92006](https://doi.org/10.17182/hepdata.92006) (2019)
- 3L eRJR, [doi:10.17182/hepdata.90607.v2](https://doi.org/10.17182/hepdata.90607.v2) (2020)



SUSY EWK 3L RPV analysis ([ATLAS-CONF-2020-009](https://arxiv.org/abs/ATLAS-CONF-2020-009)): Exclusion curves as a function of mass and branching fraction to Z bosons

# ...and by theory

- `pyhf` likelihoods discussed in
  - [Les Houches 2019 Physics at TeV Colliders: New Physics Working Group Report](#)
  - [Higgs boson potential at colliders: status and perspectives](#)
- **SModelS** team has implemented a `SModelS/pyhf` interface
  - tool for interpreting simplified-model results from the LHC
  - designed to be used by theorists
- Have produced comparison for *Search for direct stau production in events with two hadronic tau leptons in  $\sqrt{s} = 13$  TeV pp collisions with the ATLAS detector (ATLAS-SUSY-2018-04) published likelihood*
  - Compare simplified likelihood (`SModelS`)
  - to full likelihood (`pyhf`)



So here is one of our first reasonable validation plots. It's preliminary, the black line is ATLAS-SUSY-2018-04 official exclusion curve. The grey line is `SModelS` using `pyhf`, running over the published data. — Wolfgang Waltenberger, CMS/`SModelS`

# Summary

- **JSON specification** of `pyhf` likelihoods
  - human/machine readable, versionable, HEPData friendly, orders of magnitude smaller, long term preservation
- **Bidirectional translation** of likelihood specifications
  - ROOT workspaces  $\leftrightarrow$  `pyhf` JSON
- Publication for the first time of the **full likelihood** of a search for new physics
  - Continued publications from **ATLAS SUSY full Run 2 results**
- **Open publication** on HEPData, reuse and **reinterpretation** with SModelS and RECAST



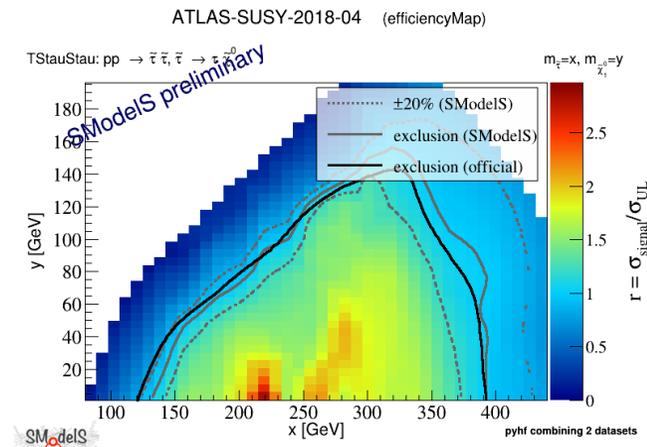
**ATLAS PUB Note**  
ATL-PHYS-PUB-2019-029  
5th August 2019



**Reproducing searches for new physics with the ATLAS experiment through publication of full statistical likelihoods**

The ATLAS Collaboration

(ATLAS, 2019)



(SModelS of ATLAS-SUSY-2018-04)



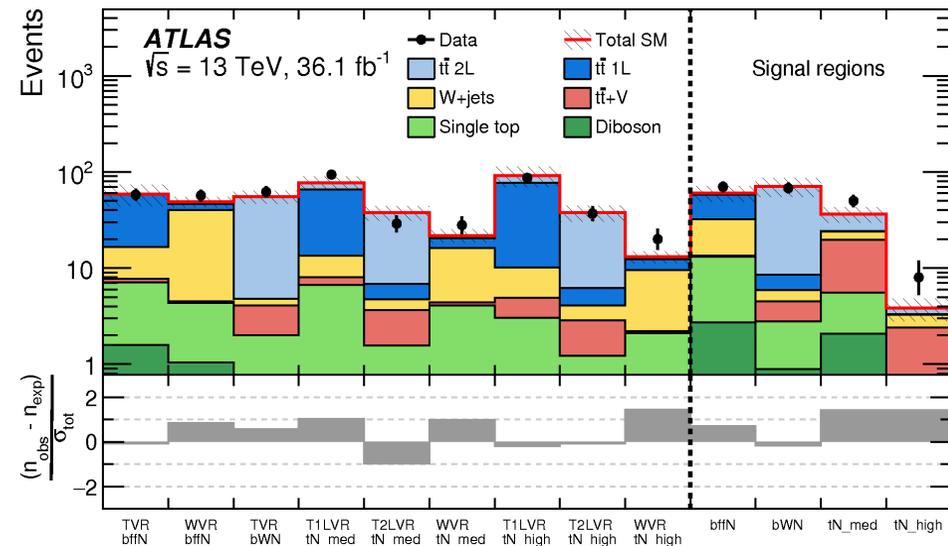
# HistFactory Template

$$f(\text{data}|\text{parameters}) = f(\vec{n}, \vec{a}|\vec{\eta}, \vec{\chi}) = \prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}(n_{cb}|\nu_{cb}(\vec{\eta}, \vec{\chi})) \prod_{\chi \in \vec{\chi}} c_{\chi}(a_{\chi}|\chi)$$

**Use:** Multiple disjoint **channels** (or regions) of binned distributions with multiple **samples** contributing to each with additional (possibly shared) systematics between sample estimates

## Main pieces:

- Main Poisson p.d.f. for simultaneous measurement of multiple channels
- Event rates  $\nu_{cb}$  (nominal rate  $\nu_{scb}^0$  with rate modifiers)
- Constraint p.d.f. (+ data) for "auxiliary measurements"
  - encode systematic uncertainties (e.g. normalization, shape)
- $\vec{n}$ : events,  $\vec{a}$ : auxiliary data,  $\vec{\eta}$ : unconstrained pars,  $\vec{\chi}$ : constrained pars



Example: **Each bin** is separate (1-bin) **channel**, each **histogram** (color) is a **sample** and share a **normalization systematic** uncertainty

# HistFactory Template

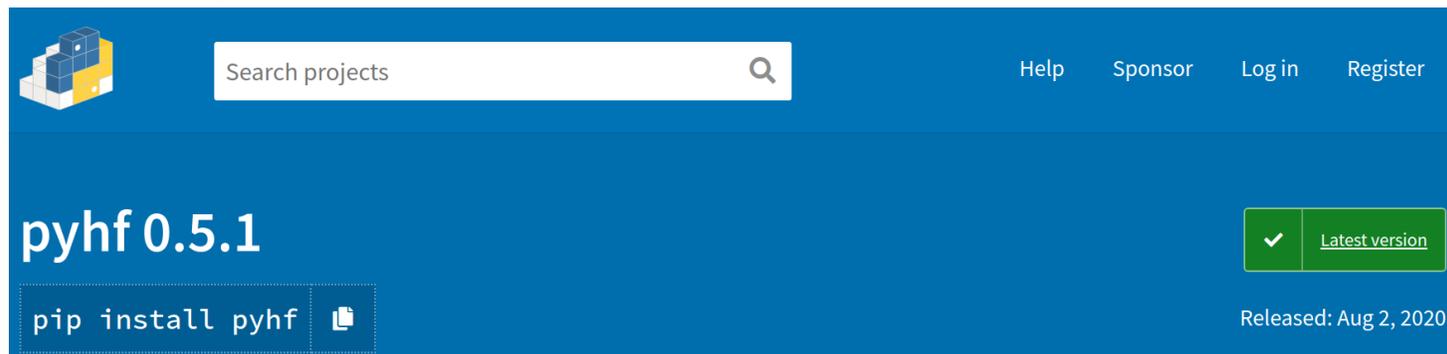
$$f(\vec{n}, \vec{a} | \vec{\eta}, \vec{\chi}) = \prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}(n_{cb} | \nu_{cb}(\vec{\eta}, \vec{\chi})) \prod_{\chi \in \vec{\chi}} c_{\chi}(a_{\chi} | \chi)$$

Mathematical grammar for a simultaneous fit with

- **multiple "channels"** (analysis regions, (stacks of) histograms)
- each region can have **multiple bins**
- coupled to a set of **constraint terms**

**This is a mathematical representation!** Nowhere is any software spec defined  
**Until now** (2018), the only implementation of HistFactory has been in **ROOT**

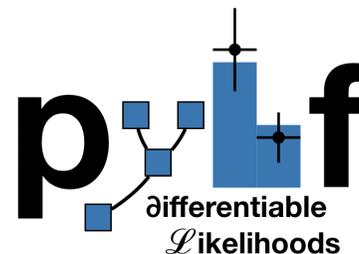
## pyhf: HistFactory in pure Python



The screenshot shows the PyPI page for the package 'pyhf'. At the top left is the package logo, a search bar with the text 'Search projects', and navigation links for 'Help', 'Sponsor', 'Log in', and 'Register'. The main content area features the package name 'pyhf 0.5.1' in large text. To the right of the name is a green checkmark icon and a link labeled 'Latest version'. Below the package name is a code block containing the command 'pip install pyhf' and a copy icon. At the bottom right, it says 'Released: Aug 2, 2020'.

# What is pyhf?

Please checkout the many resources we have starting with the [website](#) and the [SciPy 2020 talk!](#)



# References

1. F. James, Y. Perrin, L. Lyons, *Workshop on confidence limits: Proceedings*, 2000.
2. ROOT collaboration, K. Cranmer, G. Lewis, L. Moneta, A. Shibata and W. Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, 2012.
3. L. Heinrich, H. Schulz, J. Turner and Y. Zhou, *Constraining  $A_4$  Leptonic Flavour Model Parameters at Colliders and Beyond*, 2018.
4. A. Read, *Modified frequentist analysis of search results (the  $CL_s$  method)*, 2000.
5. K. Cranmer, *CERN Latin-American School of High-Energy Physics: Statistics for Particle Physicists*, 2013.
6. ATLAS collaboration, *Search for bottom-squark pair production with the ATLAS detector in final states containing Higgs bosons, b-jets and missing transverse momentum*, 2019
7. ATLAS collaboration, *Reproducing searches for new physics with the ATLAS experiment through publication of full statistical likelihoods*, 2019
8. ATLAS collaboration, *Search for bottom-squark pair production with the ATLAS detector in final states containing Higgs bosons, b-jets and missing transverse momentum: HEPData entry*, 2019

