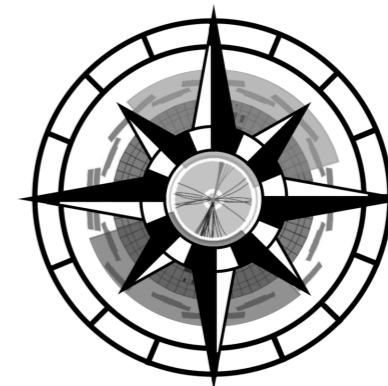


# HEPData and OpenMAPP



OpenMAPP

**Graeme Watt** (IPPP Durham)

Reinterpretation & OpenMAPP mini-workshop, Grenoble

Monday 17<sup>th</sup> June 2024

<https://hepdata.net>

**Email:** [info@hepdata.net](mailto:info@hepdata.net)

**Forum:** [hepdata-forum.cern.ch](https://hepdata-forum.cern.ch)

 Follow @HEPData

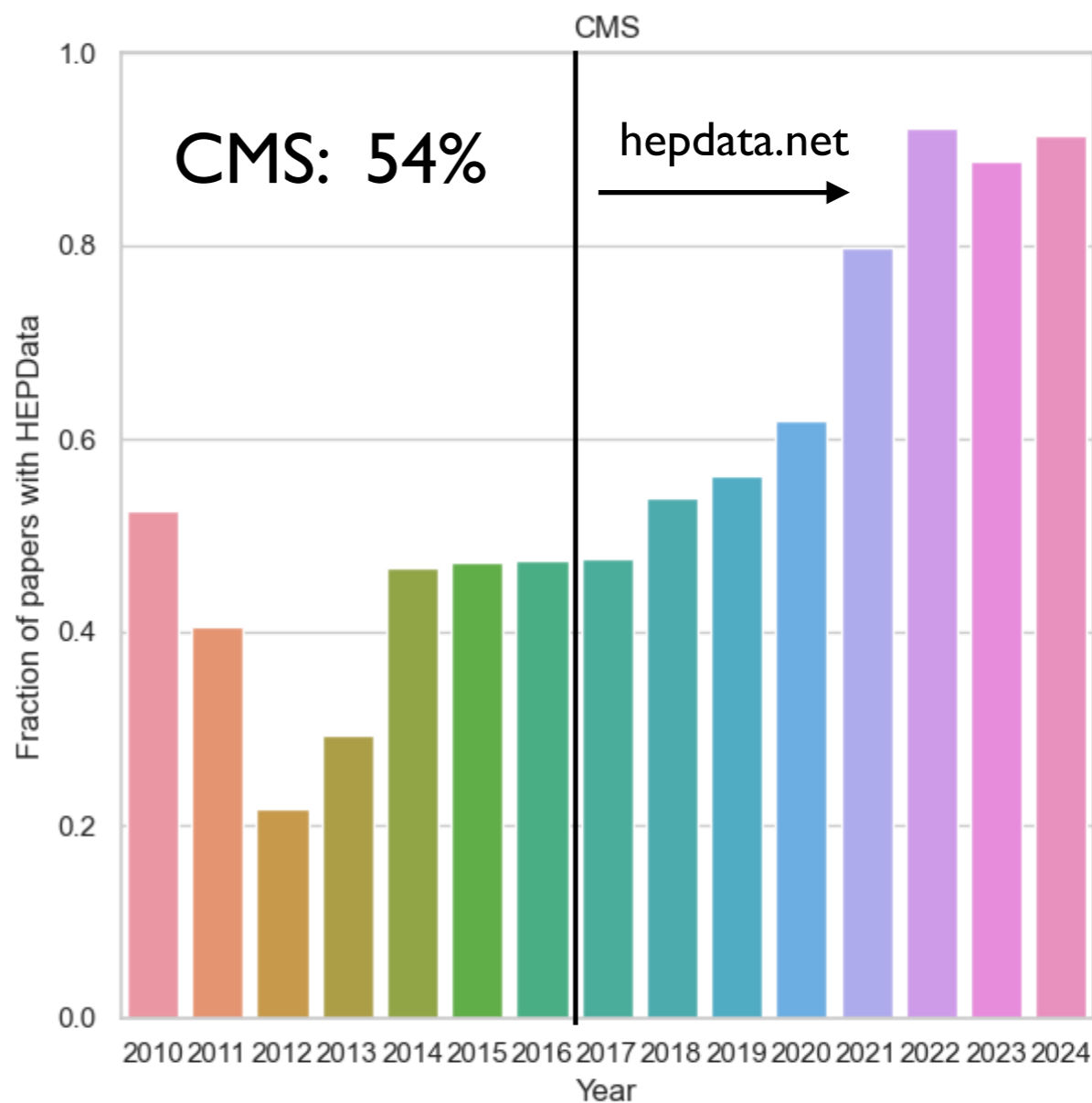
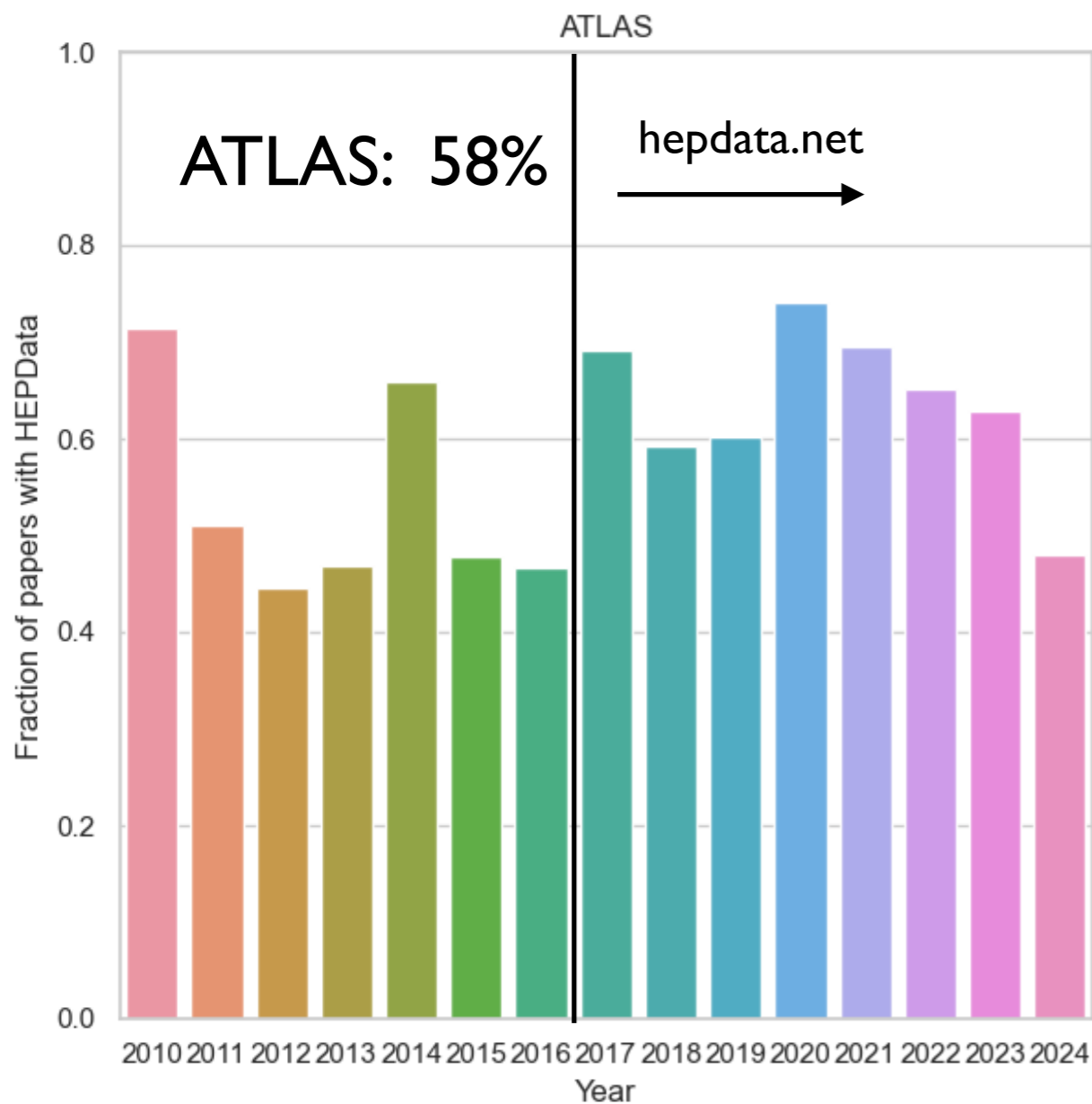
Code: <https://github.com/HEPData>

# What is HEPData?

- Unique *open-access* repository for tabular high-level **data** from more than 10k **HEP** publications (130k data tables).
- **FAIR** data: **F**indable, **A**ccessible, **I**nteroperable, **R**eusable.
- *Complementary* to other HEP information providers, e.g. INSPIRE-HEP (literature), PDG (particle properties), CERN Open Data (event-level data), Zenodo (files).
- Historically based at Durham University (UK) from 1970s.
- Transition in 2017 to hepdata.net site, hosted at CERN. Partnership with CERN Scientific Information Service.  
J. Phys.: Conf. Ser. 898 102006 [arXiv:1704.05473]
- *Staff*: **G.W.** (Manager, 2013-), **Jordan Byers** (RSE, 2022-).

# Coverage of ATLAS/CMS publications

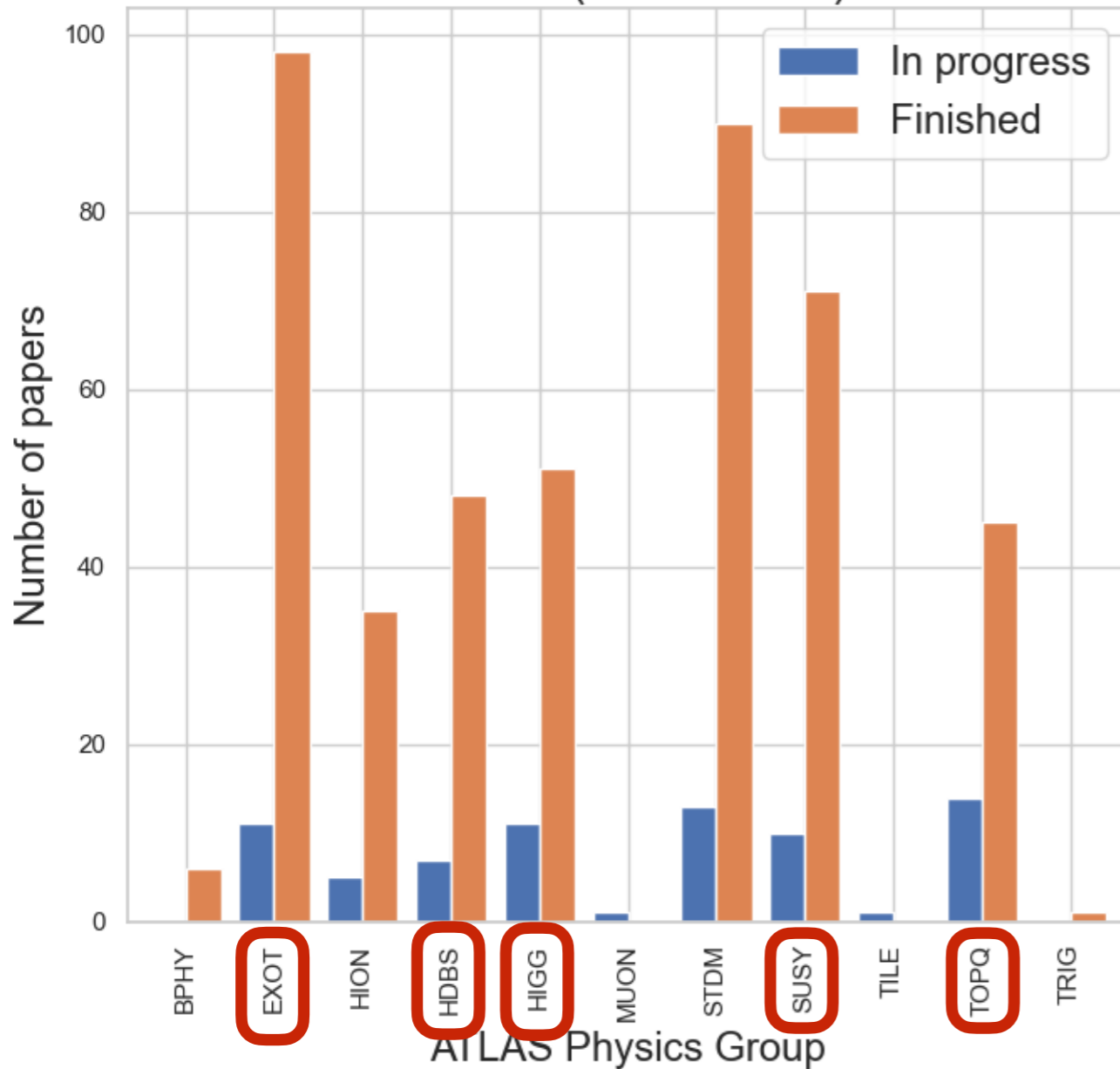
LHC publications with HEPData records (2024-06-12)



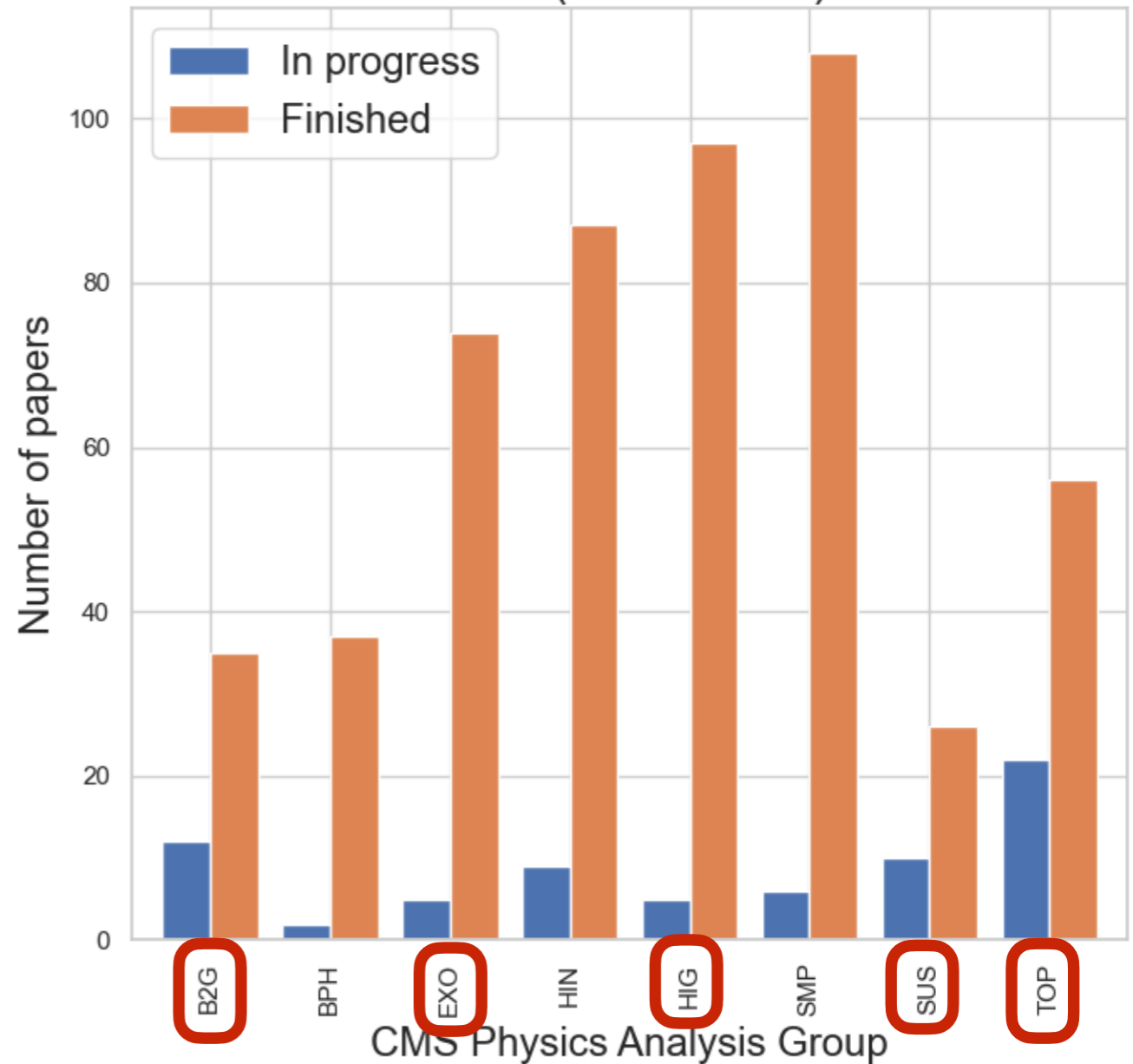
- Search INSPIRE for publications with HEPData (GitHub/Binder).

# Submissions by ATLAS/CMS groups (2017-)

ATLAS (2024-06-12)

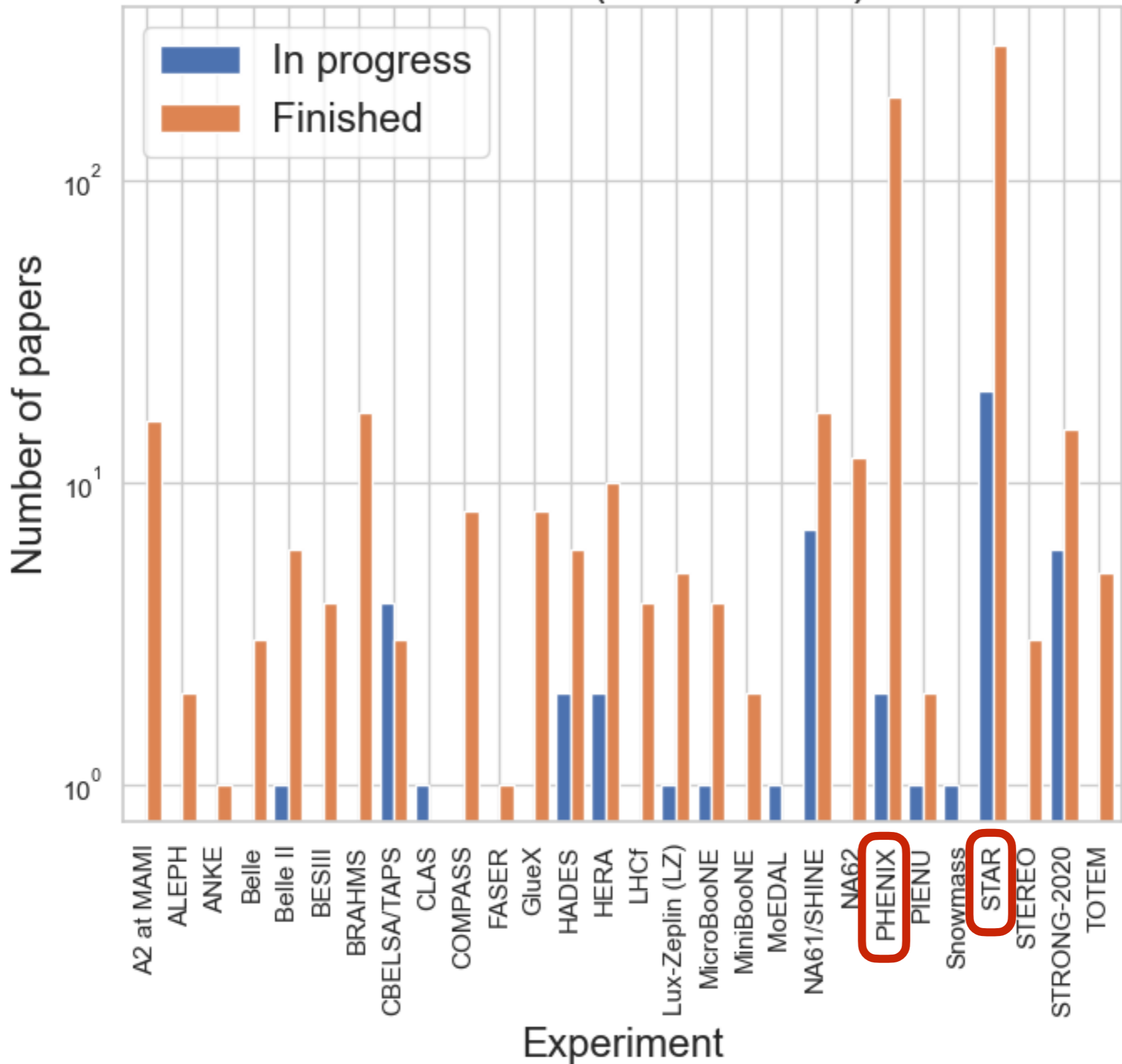


CMS (2024-06-12)



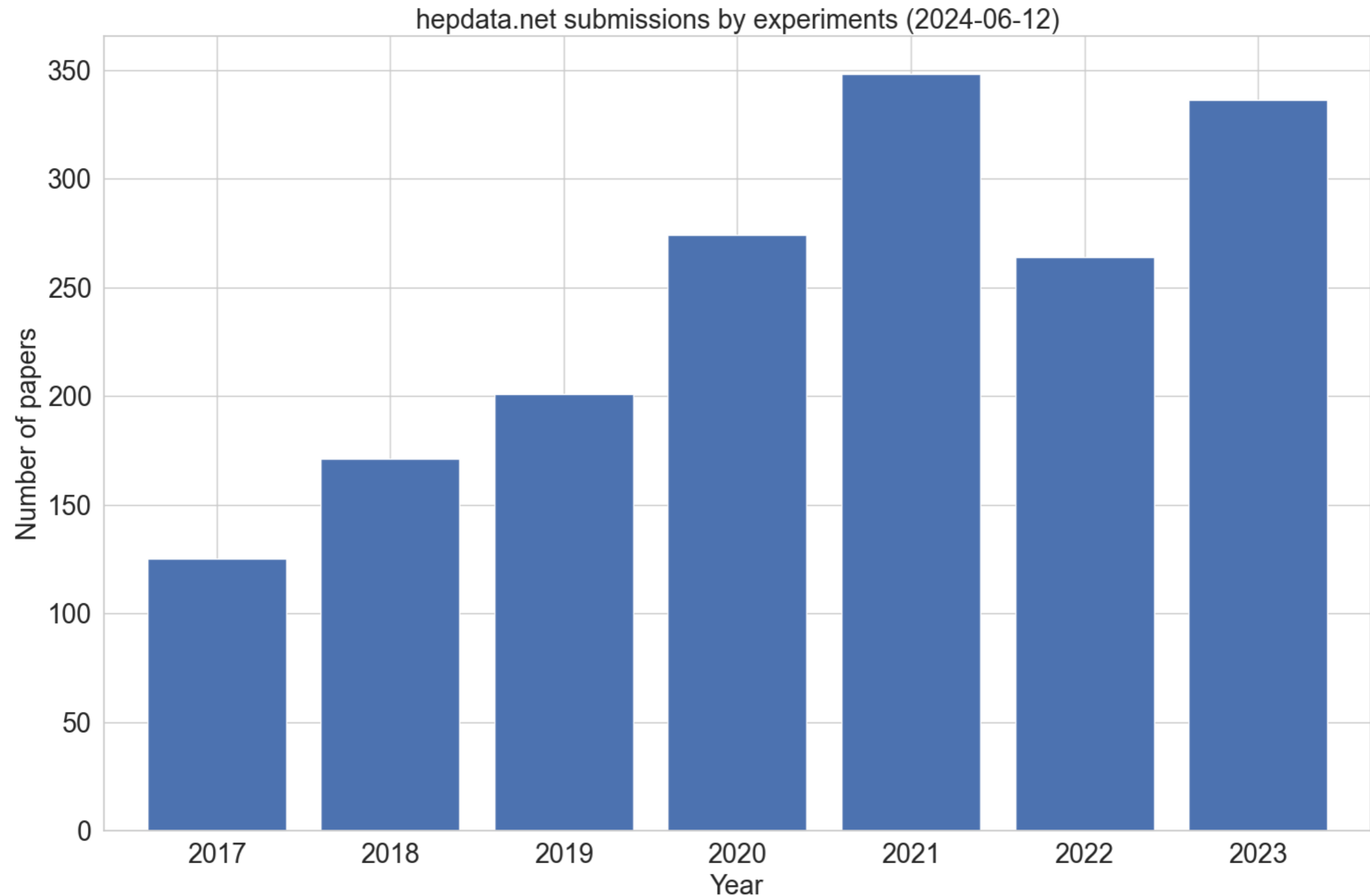
- Some ATLAS/CMS physics groups more active than others.

# Non-LHC (2024-06-12)



- Big efforts by STAR and PHENIX at RHIC (BNL News).

# Submissions per year from 2017 to 2023

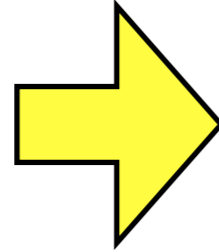


- Increase for first few years, then around 300 per year from 2020.

# Data output formats

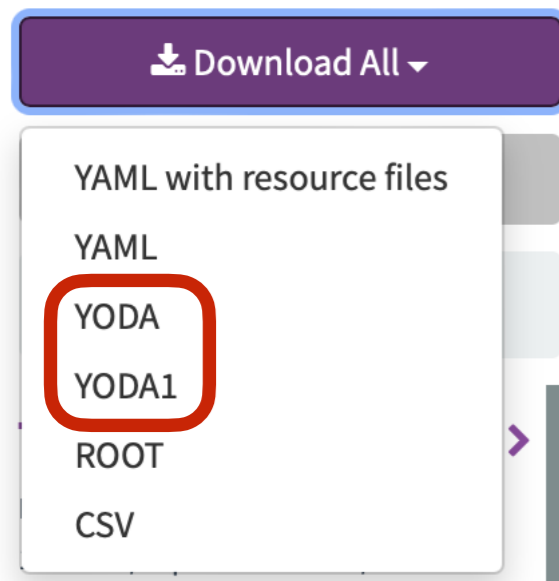
[hepdata.net/formats](https://hepdata.net/formats)

YAML: native  
HEPData format.



submission.yaml  
+ YAML data files for each table  
+ optional resource files

- JSON: JavaScript Object Notation.
- CSV: comma-separated values.
- ROOT: binary .root file.
- YODA: for inclusion in a Rivet analysis.



from 17<sup>th</sup>  
Nov 2023

- **NEW** “YODA” now gives new YODA2 format, for use with Rivet version 4 (released Feb 2024).
- Legacy YODA format still available as YODA1.
- Thanks to [Chris Gütschow \(UCL\)](#) for work on implementing the YAML → YODA2 conversion.

# Links to Rivet analysis code


<http://rivet.hepforge.org/analyses.json>

- JSON file maps INSPIRE IDs to Rivet analysis names:

```
{ "100016" : [ "GAMMAGAMMA_1975_I100016" ], ...,  
  "954993" : [ "ATLAS_2011_I954993" ] }
```

- Badge appears in search results and link on record:

 Rivet Analysis Measurement of the  $t\bar{t}$  production cross-section as a function of jet multiplicity and jet transverse momentum in 7 TeV proton-proton collisions with the ATLAS detector

 View Analyses ▾

 Rivet

- Extendable to other analysis frameworks containing publication-specific code.



NEW

from 5<sup>th</sup>  
Oct 2023

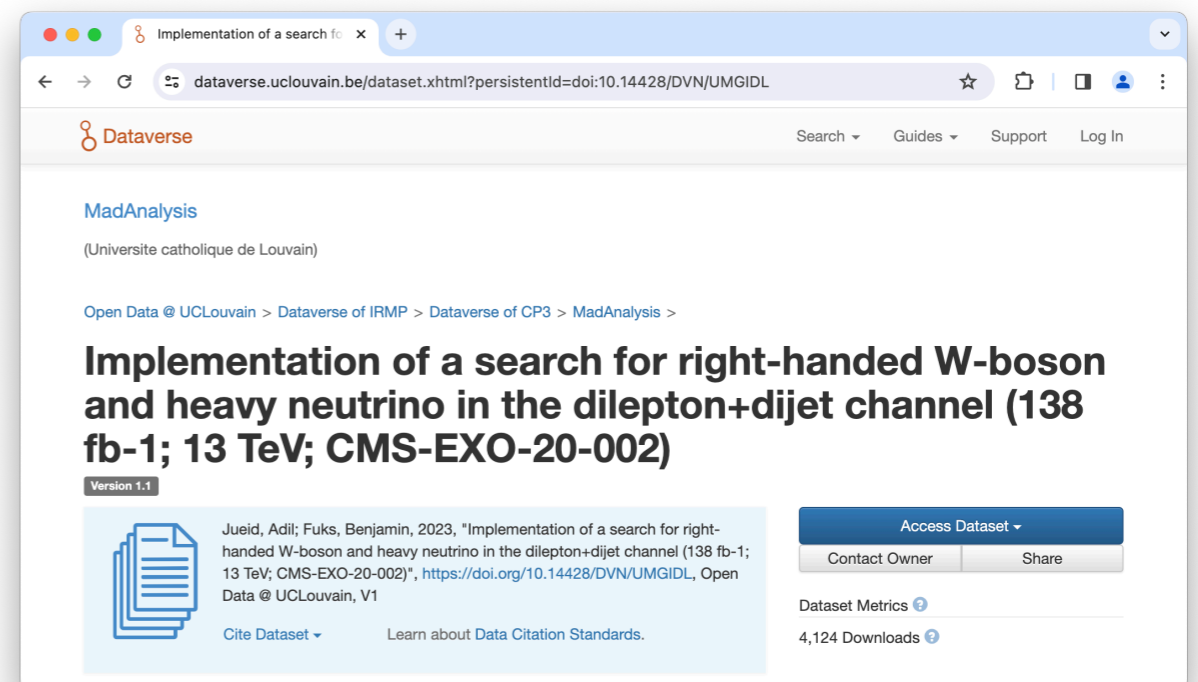
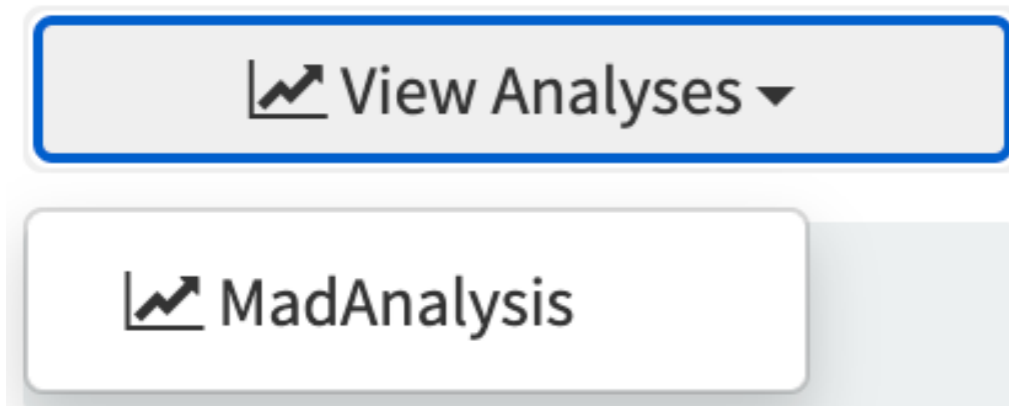
# Links to MadAnalysis 5

- analyses.json file for MadAnalysis 5 analyses:

```
{ "1458270": [ "10.14428/DVN/MHPXX4" ], ...,  
  "1750186": [ "10.14428/DVN/OFAE1G" ] }
```

- Search analysis:MadAnalysis gives 26 records.

 **MadAnalysis** Search for a right-handed W boson and a heavy neutrino in proton-proton collisions at  $\sqrt{s} = 13$  TeV



- Thanks to Jack Araz.

# hepdata lib

- hepdata lib package by *Clemens Lange* (and *Andreas Albert*).  
Library to read in text/ROOT and write HEPData YAML.  
[https://github.com/HEPData/hepdata\\_lib](https://github.com/HEPData/hepdata_lib)
- **Recent improvements** by new contributors (releases):
  - Convert from scikit-hep/hist histograms by *Yi-Mu Chen*.
  - Revamp of CI workflow by *Matthew Feickert* (and *G.W.*).
  - Update docs for add\_additional\_resource by *G.W.*
  - Fix broken code documentation on *Read the Docs* by *G.W.*
  - Add option to suppress warning if zero uncertainties by *G.W.*
  - Allow some functionality without **ROOT** installation by *G.W.*
  - Remove pin on **Pylint** to support **Python 3.11** by *G.W.*
  - Add functions for related records and tables by *Jordan Byers*.
  - Add functions to provide license information by *Jordan Byers*.
- **Future:** add ability to read in YODA files (issue)?

from 31<sup>st</sup>  
May 2024

# Licenses

- HEPData Terms of Use:

*Unless specified otherwise for selected datasets, all metadata and datasets in the HEPData service are made available under the terms of CC0.*

CC0 (aka CC Zero) is a public dedication tool, which enables creators to give up their copyright and put their works into the worldwide public domain. CC0 enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, with no conditions.



- Specify a license other than CC0 using fields `data_license` (for data tables) or `license` (for resources) in `submission.yaml` file.
- Recent changes (motivated by feedback from **CMS & Lukas Heinrich**):
  - Render license (including default CC0) on HEPData web pages.
  - Functions to add license information included in `hepdata_lib`.
- CMS Higgs boson observation statistical model uses CC BY 4.0.

NEW

from 22<sup>nd</sup>  
Aug 2023

# Record-specific `<title>` tags

- Suggestion by *Andy Buckley* in August 2023 for easier management of a user's browser history.

Recently Closed

- History
- HEPData | CMS | 2024 | Search for long-lived heavy neutral lept...uon detectors in proton-proton collisions at  $\sqrt{s} = 13$  TeV
- HEPData | ALICE | 2023 | Common femtoscopic hadron-emission source in pp collisions at the LHC
- HEPData | CMS | 2024 | Search for a standard model-like Higgs...oton final state in proton-proton collisions at  $\sqrt{s} = 13$  TeV
- HEPData | CMS | 2024 | Study of WH production through vector...e Higgs boson in proton-proton collisions at  $\sqrt{s} = 13$  TeV
- HEPData | CMS | 2024 | Search for heavy neutral leptons in final...ing tau leptons in proton-proton collisions at  $\sqrt{s} = 13$  TeV
- HEPData | LZ | 2024 | Constraints On Covariant WIMP-Nucleon...ctions from the First Science Run of the LUX-ZEPLIN Experiment
- HEPData | STAR | 2023 | Collision-energy Dependence of Deute...ts and Proton-deuteron Correlations in Au+Au collisions at RHIC
- HEPData | ARGUS | 1990 | Measurement of  $\chi(c)$  production in  $e^+ e^-$  annihilation at 10.5-GeV center-of-mass energy
- HEPData | PHENIX | 2018 | Measurements of multiparticle correl...Au collisions at 200 GeV and implications for collective behavior

Recently Visited

- HEPData | ATLAS | 2024 | Search for non-resonant Higgs boson...ATLAS detector | Table 3 (95% CL upper limits on cross-section)
- HEPData | ATLAS | 2024 | Search for non-resonant Higgs boson... $\sqrt{s}$  with the ATLAS detector | Table 2 (post-fit yields)
- HEPData | ATLAS | 2024 | Search for non-resonant Higgs boson... $\sqrt{s}$  with the ATLAS detector | Table 1 (pre-fit yields)
- HEPData | CMS | 2024 | Observation of the  $\Xi^- \rightarrow \Lambda^0 p$  baryon in proton-proton collisions at  $\sqrt{s} = 13$  TeV
- HEPData | ATLAS | 2024 | Underlying-event studies with strang...in  $pp$  collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector
- HEPData | CMS | 2024 | Search for  $CP$  violation in  $D^0 \rightarrow S$  decays in proton-proton collisions at  $\sqrt{s} = 13$  TeV
- HEPData | CMS | 2024 | Girth and groomed radius of jets recoili...d proton-proton collisions at  $\sqrt{s} = 5.02$  TeV
- HEPData Search
- HEPData Search
- HEPData Homepage
- HEPData | CMS | 2024 | Search for long-lived heavy neutral lept...uon detectors in proton-proton collisions at  $\sqrt{s} = 13$  TeV
- HEPData | ALICE | 2023 | Common femtoscopic hadron-emission source in pp collisions at the LHC
- HEPData | CMS | 2024 | Search for a standard model-like Higgs...oton final state in proton-proton collisions at  $\sqrt{s} = 13$  TeV
- HEPData | CMS | 2024 | Study of WH production through vector...e Higgs boson in proton-proton collisions at  $\sqrt{s} = 13$  TeV
- HEPData | CMS | 2024 | Search for heavy neutral leptons in final...ing tau leptons in proton-proton collisions at  $\sqrt{s} = 13$  TeV

NEW

from 7<sup>th</sup>  
Dec 2023

# Deferred loading of large files

- Validator restricts YAML data files to be smaller than **10 MB**.
- YAML data files **1-10 MB** can be slow to load in browser.
- Introduce deferred table loading if YAML data file **> 1 MB**:
- On resource file landing page if text file **> 1 MB**, only display “Download” button without trying to display in browser:

**This table is too large to load automatically.**  
The table size is 9.25 MB, which is greater than our threshold of 1.00 MB.

Load Table

**workspace\_tHu.json** [10.17182/hepdata.150998.v1/r1](https://doi.org/10.17182/hepdata.150998.v1/r1)

License: [CC0](#)

Full likelihood of the tHu fit in the HistFactory JSON format described in ATL-PHYS-PUB-2019-029

**This file (2.21 MB) is larger than our loading threshold (1.00 MB), and is only available for download below.**

Download via DOI: `curl -OJLH "Accept: application/json" https://doi.org/10.17182/hepdata.150998.v1/r1`

Download

NEW

from 22<sup>nd</sup>  
Aug 2023

# Bidirectional linking

- Suggestion by Jon Butterworth in December 2022. Technical implementation by Jordan Byers (Durham).
- Enable bidirectional links *between* HEPData **tables** possibly in different records in `submission.yaml`:

```
related_to_table_dois:
```

```
- 10.17182/hepdata.12345.v1/t2  
- 10.17182/hepdata.67890.v3/t4
```

or use hepdata\_lib

- Similar bidirectional links *between* HEPData **records**:

```
related_to_hepdata_records:
```

```
- 12345  
- 67890
```

or use hepdata\_lib

NEW

from 22<sup>nd</sup>  
Aug 2023

# Bidirectional linking

Measurement and interpretation of same-sign  $W$  boson pair production in association with two jets in  $pp$  collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector

**Table 1** 10.17182/hepdata.141650.v1/t1  
License: CC0  
Data from Fig. 5(a)  
Fiducial differential cross section of the electroweak  $W^+W^+jj$  production as a function of  $m_{\ell\ell}$ . The correlation of uncertainties...

This table is related to:  
• Table 11

This table is referred to by:  
• Table 11

**Table 11** 10.17182/hepdata.141650.v1/t11  
License: CC0  
Data from Fig. 37(a)  
Observed correlations between the bins of the LH-unfolded cross section of the electroweak  $W^+W^+jj$  production as a function...

**Table 12** 10.17182/hepdata.141650.v1/t12  
Data from Fig. 37(b)  
Observed correlations between the bins of the LH-unfolded cross section of the electroweak  $W^+W^+jj$  production as a function...

**Table 13** 10.17182/hepdata.141650.v1/t13  
Data from Fig. 37(c)  
Observed correlations between the bins of the LH-unfolded cross section of the electroweak  $W^+W^+jj$  production as a function...

**Table 14** 10.17182/hepdata.141650.v1/t14  
Data from Fig. 37(d)  
Observed correlations between the bins of the LH-unfolded cross section of the electroweak  $W^+W^+jj$  production as a function...

**cmenergies** 13000.0

**observables** SIG

**phrases** Proton-Proton Scattering, Vector Boson Scattering, Single Differential Cross Section

**reactions** P P → W+ W+ JET JET X, P P → W- W- JET JET X

**Visualize**

SQRT(S)	13 TeV
LUMINOSITY	139 fb <sup>-1</sup>

Measurement and interpretation of same-sign  $W$  boson pair production in association with two jets in  $pp$  collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector

**Table 11** 10.17182/hepdata.141650.v1/t11  
License: CC0  
Data from Fig. 37(a)  
Observed correlations between the bins of the LH-unfolded cross section of the electroweak  $W^+W^+jj$  production as a function...

This table is related to:  
• Table 1

This table is referred to by:  
• Table 1

**Table 1** 10.17182/hepdata.141650.v1/t1  
License: CC0  
Data from Fig. 5(a)  
Fiducial differential cross section of the electroweak  $W^+W^+jj$  production as a function of  $m_{\ell\ell}$ . The correlation of uncertainties...

**Table 12** 10.17182/hepdata.141650.v1/t12  
Data from Fig. 37(b)  
Observed correlations between the bins of the LH-unfolded cross section of the electroweak  $W^+W^+jj$  production as a function...

**Table 13** 10.17182/hepdata.141650.v1/t13  
Data from Fig. 37(c)  
Observed correlations between the bins of the LH-unfolded cross section of the electroweak  $W^+W^+jj$  production as a function...

**Table 14** 10.17182/hepdata.141650.v1/t14  
Data from Fig. 37(d)  
Observed correlations between the bins of the LH-unfolded cross section of the electroweak  $W^+W^+jj$  production as a function...

**cmenergies** 13000.0

**observables** CORR

**phrases** Proton-Proton Scattering, Vector Boson Scattering, Single Differential Cross Section

**reactions** P P → W+ W+ JET JET X, P P → W- W- JET JET X

**Visualize**

First Bin	Second Bin	Correlation coefficient, total	Correlation coefficient, stat. only
#bin1	#bin1	1.0	1.0
#bin1	#bin2	0.052768	-0.022928
#bin1	#bin3	0.072057	0.0072439

related\_to\_table\_dois:

[10.17182/hepdata.141650.v1/t11]

~~related\_to\_table\_dois:~~

~~[10.17182/hepdata.141650.v1/t1]~~

- First record: <https://www.hepdata.net/record/ins2729396>
- Links are automatically *bidirectional*: only specify one direction.
- Issue with previewing bidirectional links before record finalised.

NEW

from 3<sup>rd</sup>  
May 2024

# Searching for resources

## Searching resources by field

Text-based description searching:

`resources:"Created with hepdata_lib"`

Quotes force a full match.

Resource-type searching:

`resources.type:png`

Examples: png, html, github, zenodo etc.

Searching for specific URLs:

`resources.url:atlas.web.cern.ch`

- Additional resource metadata now indexed for searching.



NEW

from 3<sup>rd</sup>  
May 2024

# Resource information in JSON

The screenshot shows the HEPData search results page for the query 'ONNX'. The search bar and the 'ONNX' query are circled in red. The results list includes:

- Search for R-parity violating supersymmetry in a final state containing leptons and many jets with the ATLAS experiment using  $\sqrt{s} = 13$  TeV proton-proton collision data**  
The ATLAS collaboration Aad, Georges; Abbott, Braden Keim; Abbott, Dale; *et al.*  
Eur.Phys.J.C 61 (2021) 1023, 2021.  
Inspire Record 1869040 DOI 10.17182/hepdata.104860
- Search for neutral long-lived particles in  $pp$  collisions at  $\sqrt{s} = 13$  TeV that decay into displaced hadronic jets in the ATLAS calorimeter**  
The ATLAS collaboration Aad, Georges; Abbott, Braden Keim; Abbott, Dale; *et al.*  
JHEP 06 (2022) 005, 2022.  
Inspire Record 2043503 DOI 10.17182/hepdata.115578

The screenshot shows the search results in JSON format. The 'resources' array contains the following entries:

- 0: description: "Created with hepdata\_lib 0.10.0", type: "zenodo", url: "https://zenodo.org/record/494627/"
- 1: description: "Webpage with all figures and tables", type: "html", url: "http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/EXOT-2019-23/"
- 2: description: "arXiv", type: "html", url: "http://arxiv.org/abs/arXiv:2203.01009"
- 3: description: "ONNX records of the NNs", type: "gz", url: "https://www.hepdata.net/record/resource/3170098?landing\_page=true"
- 4: description: "Pure-C++ and pure-python standalone executables of the BDTs", type: "gz", url: "https://www.hepdata.net/record/resource/3170099?landing\_page=true"
- 5: description: "Archive of full likelihoods in the HistFactory JSON format described in ATL-PHYS-PUB-2019-029", type: "HistFactory", url: "https://www.hepdata.net/record/resource/3170100?landing\_page=true"
- 6: description: "Example code to read and use the efficiency maps from Aux Mat Fig 11 and 12", type: "Python", url: "https://www.hepdata.net/record/resource/3170101?landing\_page=true"

```
import requests
query = 'ONNX'
url = f'https://www.hepdata.net/search/?q={query}&format=json'
request = requests.get(url).json()
for result in request['results']:
    for resource in result['resources']:
        if query in resource['description']:
            print(result['inspire_id'])
            print(resource['description'])
            print(resource['type'])
            print(resource['url'].replace('landing_page', 'view'))
            print()
```

1869040  
ONNX files for the neural networks for the EWK analysis  
tgz  
<https://www.hepdata.net/record/resource/2677521?view=true>


2043503  
ONNX records of the NNs  
gz  
<https://www.hepdata.net/record/resource/3170098?view=true>

● **Example:** get download links of ONNX files from Python.

# hepdata-cli

- CLI and Python API for HEPData search/download/upload.
- Summer project in 2020 by Giuseppe De Laurentis.
- Install (in venv) with: `pip install hepdata-cli`
- Reproduce previous example using `hepdata-cli`:

```
from hepdata_cli.api import Client
client = Client()
query = 'ONNX'
inspire_ids = client.find(query, ids='inspire').split()
results = client.find(query, keyword='resources')
for i, result in enumerate(results):
    print(inspire_ids[i])
    for resource in result['resources']:
        if query in resource['description']:
            print(resource['description'])
            print(resource['type'])
            print(resource['url'].replace('landing_page', 'view'))
            print()
```



```
1869040
ONNX files for the neural networks for the EWK analysis
tgz
https://www.hepdata.net/record/resource/2677521?view=true

2043503
ONNX records of the NNs
gz
https://www.hepdata.net/record/resource/3170098?view=true
```

**Note:** this slide added *after* my talk (on 22<sup>nd</sup> June 2024) when I realised that the last item of the next slide is already satisfied by `hepdata-cli`.

# OpenMAPP (03/2024 - 02/2026)

Tasks	
T1.1	<p><b>Extension of HEPData functionalities (M1–M6: responsible: 7; involved: 4,5,7)</b></p> <ul style="list-style-type: none"><li>- Add backend code to the HEPData system to store, retrieve, and be able to query data types beyond primary data-tables and uncertainties.</li><li>- Build a HEPData mechanism for linking between data objects via internal DOIs, e.g. for associating theory predictions to measurement tables, or reference simulated event samples to their validation datasets.</li><li>- Add a REST querying mechanism to HEPData, allowing interrogation of what resources of which types are associated with an analysis, and allowing each to be retrieved by a distinct URL.</li><li>- Provide a programmatic interface to the REST API via the <code>hepdata_cli</code> Python package.</li></ul> <p>The data types concerned by T1.1 include statistical models in JSON format, serialised machine-learning models (e.g., ONNX), binned and symbolic detector-response functions, combined sets of event-generator steering files for reference configurations and their corresponding reference outputs, etc.</p> <p>To be carried out by the UK partners, coordinated by partner 7 (Graeme Watt)</p>

- ✓ Resources indexed
- ✓ Bidirectional linking
- ✓ Resources in JSON
- ✓ `hepdata-cli`

- Most of Task T1.1 now implemented, but open to feedback.

# Summary

**Email:** [info@hepdata.net](mailto:info@hepdata.net)

**Forum:** [hepdata-forum.cern.ch](https://hepdata-forum.cern.ch)

- **HEPData** is *the* repository for publication-related HEP data.
- *Caveats:* design restricts size ( $\approx$ MB) and format (mostly tabular).
- Infrastructure via CERN, development and support via Durham.
- Open development process via <https://github.com/HEPData> .
- New features since LHC Reinterpretation Forum in Dec 2022:
  - Numerous improvements to `hepdata_lib` helper library.
  - Default CC0 license with option to specify a different license.
  - Record-specific `<title>` tags to manage browser history.
  - Deferred loading in browser of YAML and text files larger than 1 MB.
  - Bidirectional linking possible between HEPData tables or records.
  - Resource metadata indexed and available in JSON format.
- OpenMAPP Task T1.1 implemented (?), but feedback welcome.