HEPData and pyhf

Graeme Watt (IPPP Durham)

pyhf Users and Developers Workshop 2023

CERN, Tuesday 5th December 2023

https://hepdata.net

Email: info@hepdata.net

Forum: 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 10,000 **H**igh-**E**nergy **P**hysics publications.
- **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. (2013-), A. Clarke (2019-22), J. Byers (2022-).

Modes of data entry

1. Manually harvested from data given in publications. HEPData staff extracted tables from `.tex` source.

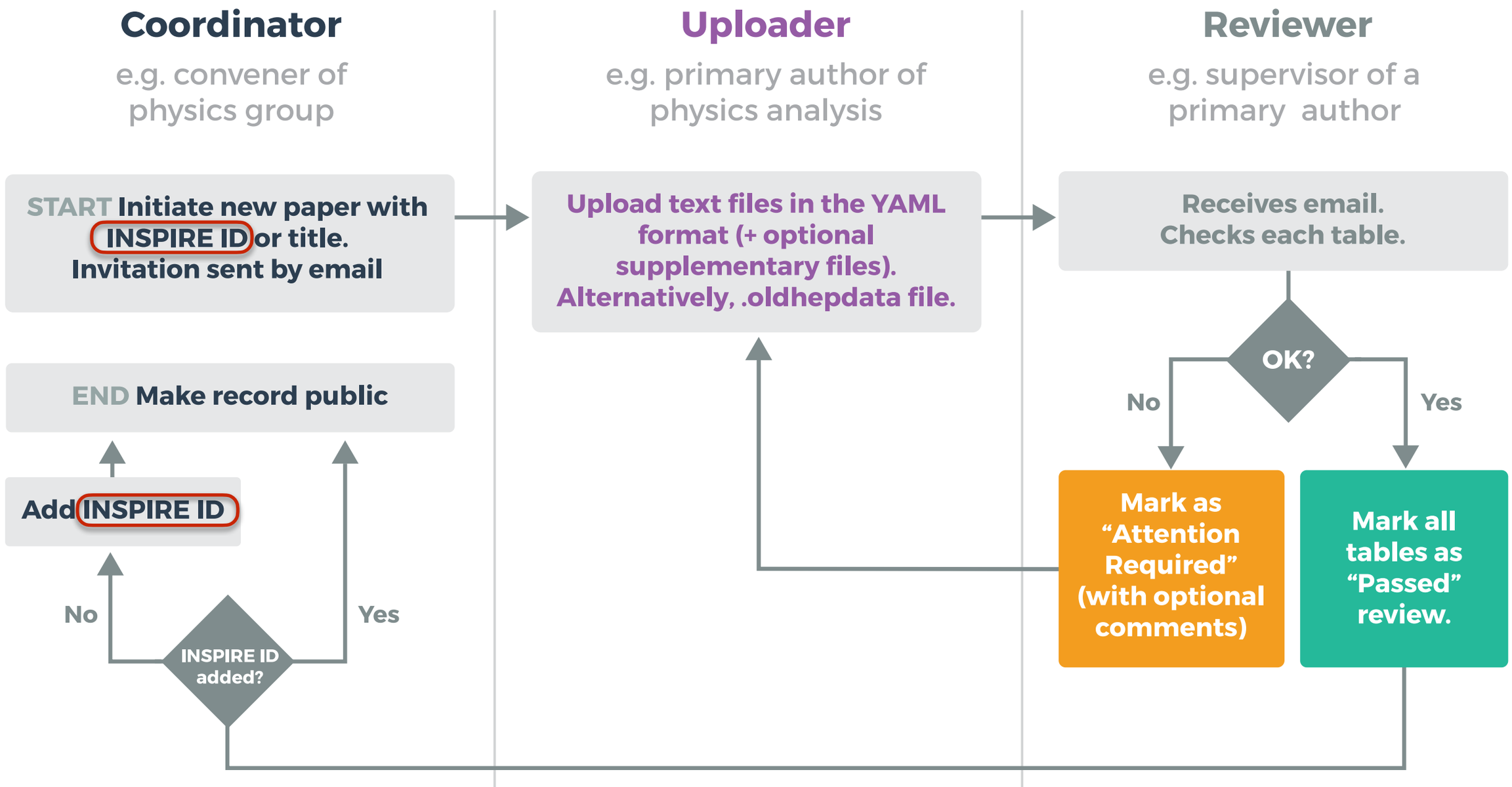
2. Data points directly submitted by experiments.

   - Pre-2014: no guidelines on preferred format.
   - Early 2014: encourage standard "input" format.
   - Late 2014: introduce online submission system.
   - Early 2017: allow submissions from hepdata.net.

Mode 1. now phased out in favour of mode 2.

Submission system on hepdata.net

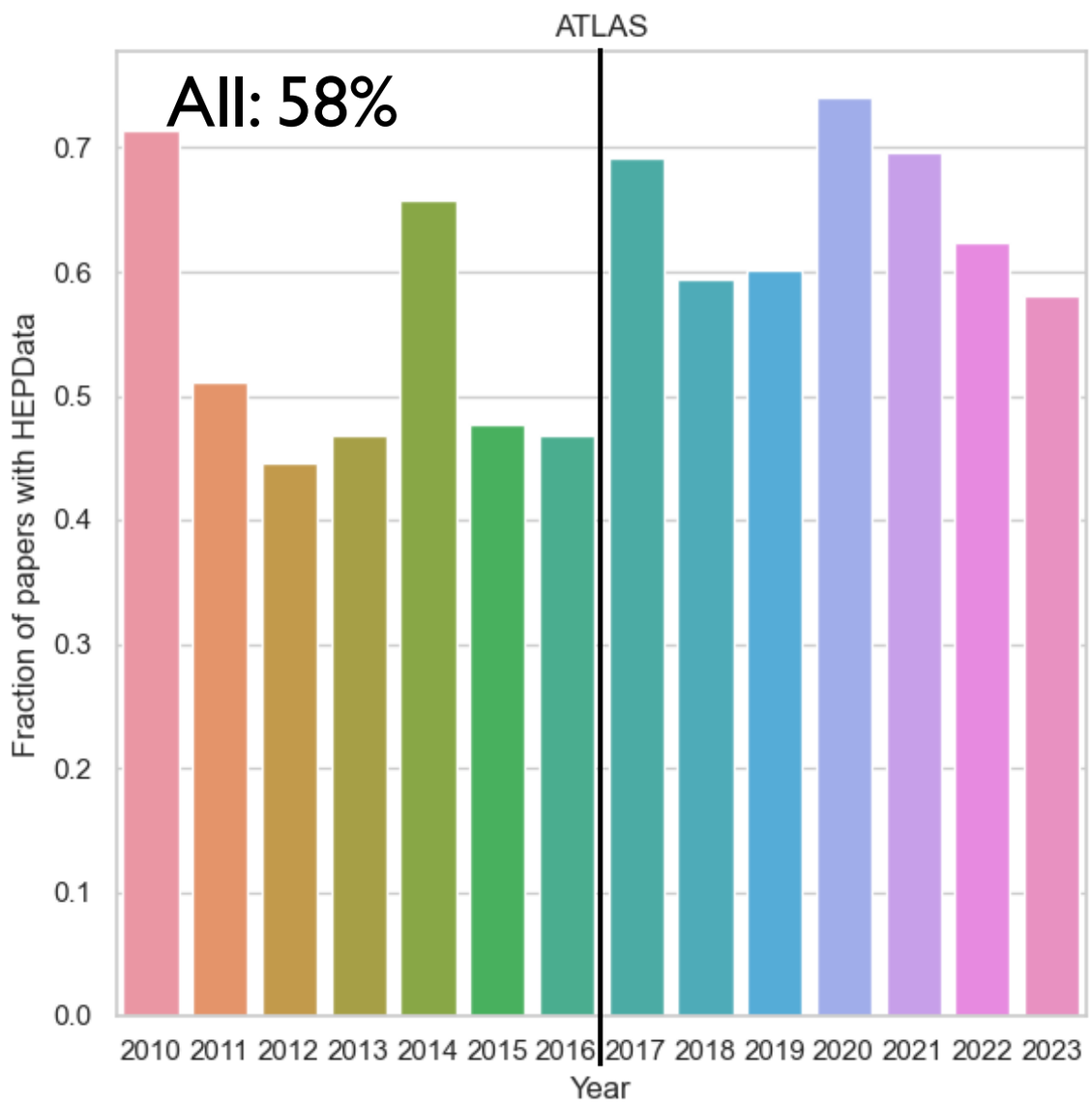
https://hepdata.net/submission

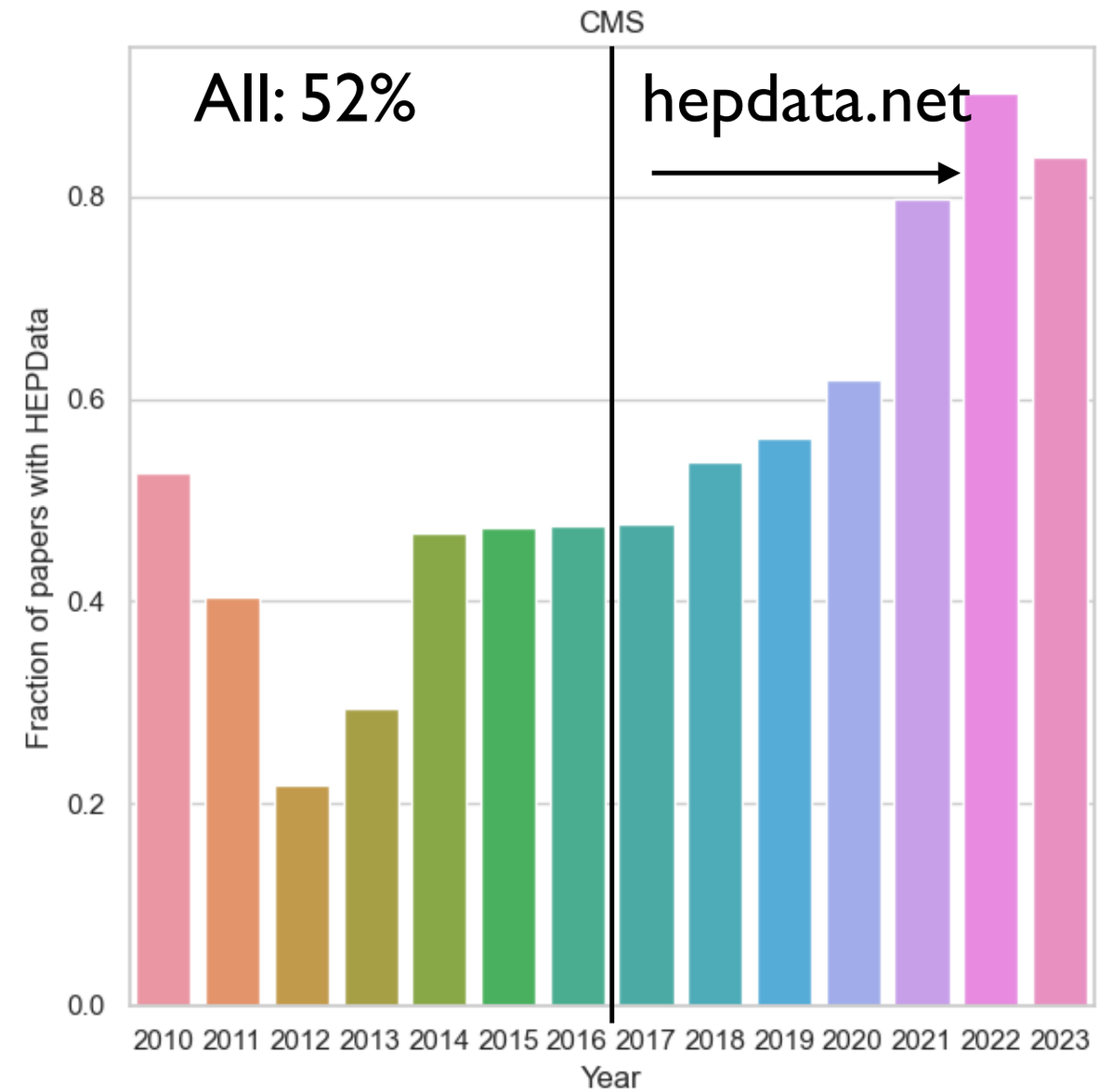


- Submissions managed by Coordinators within each experiment/group.

Coverage of ATLAS/CMS publications

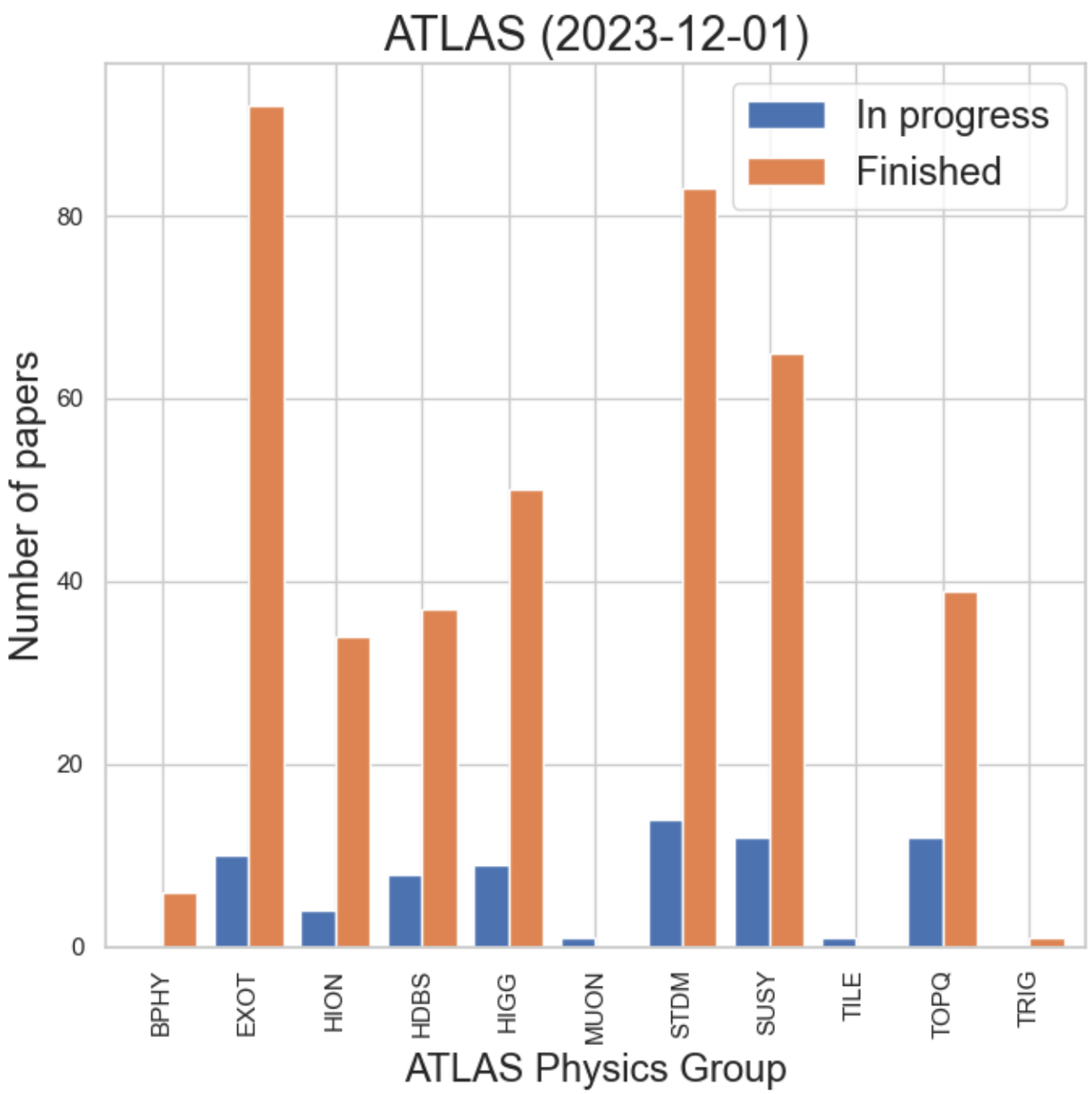
LHC publications with HEPData records (2023-12-01)

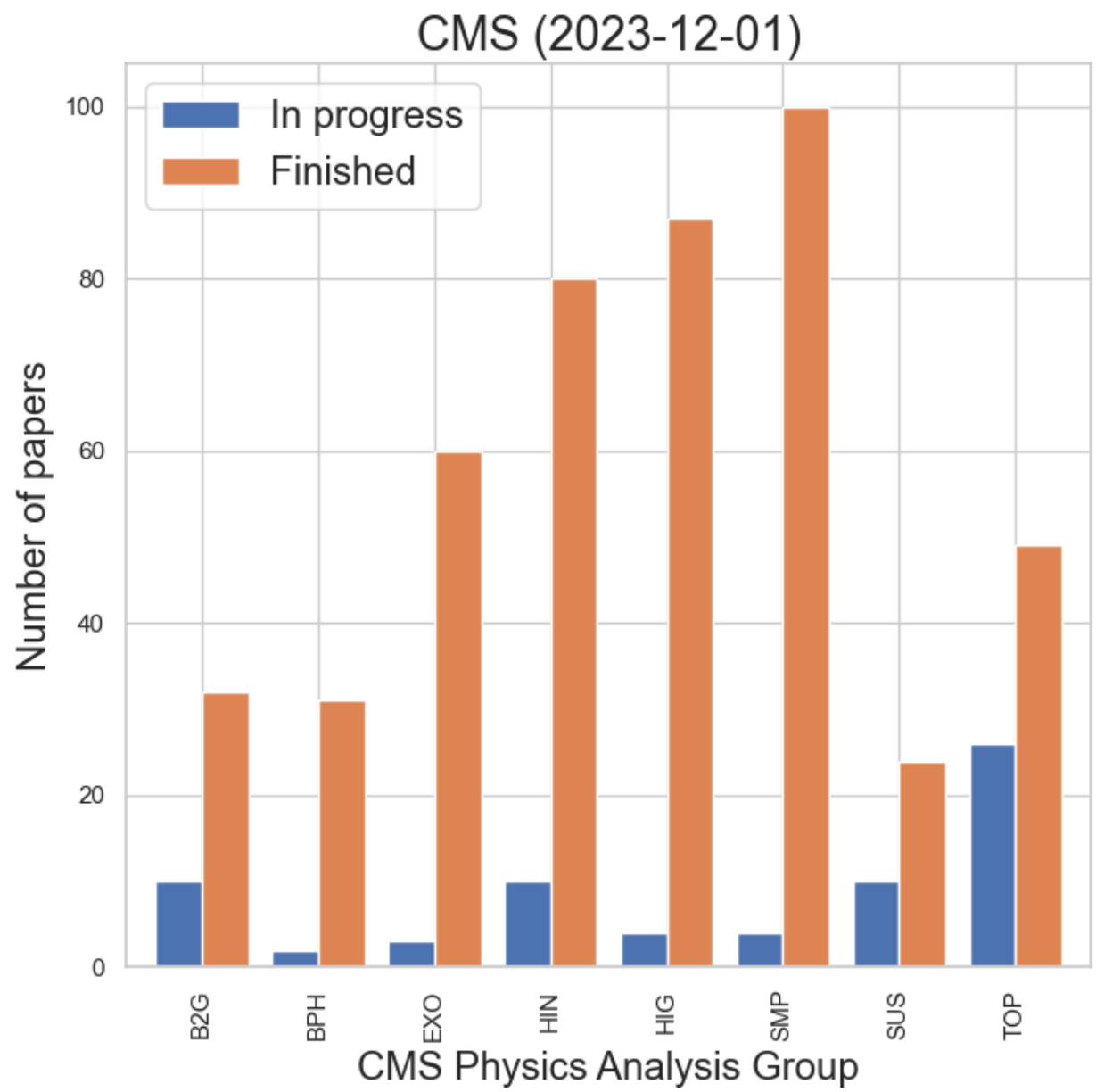




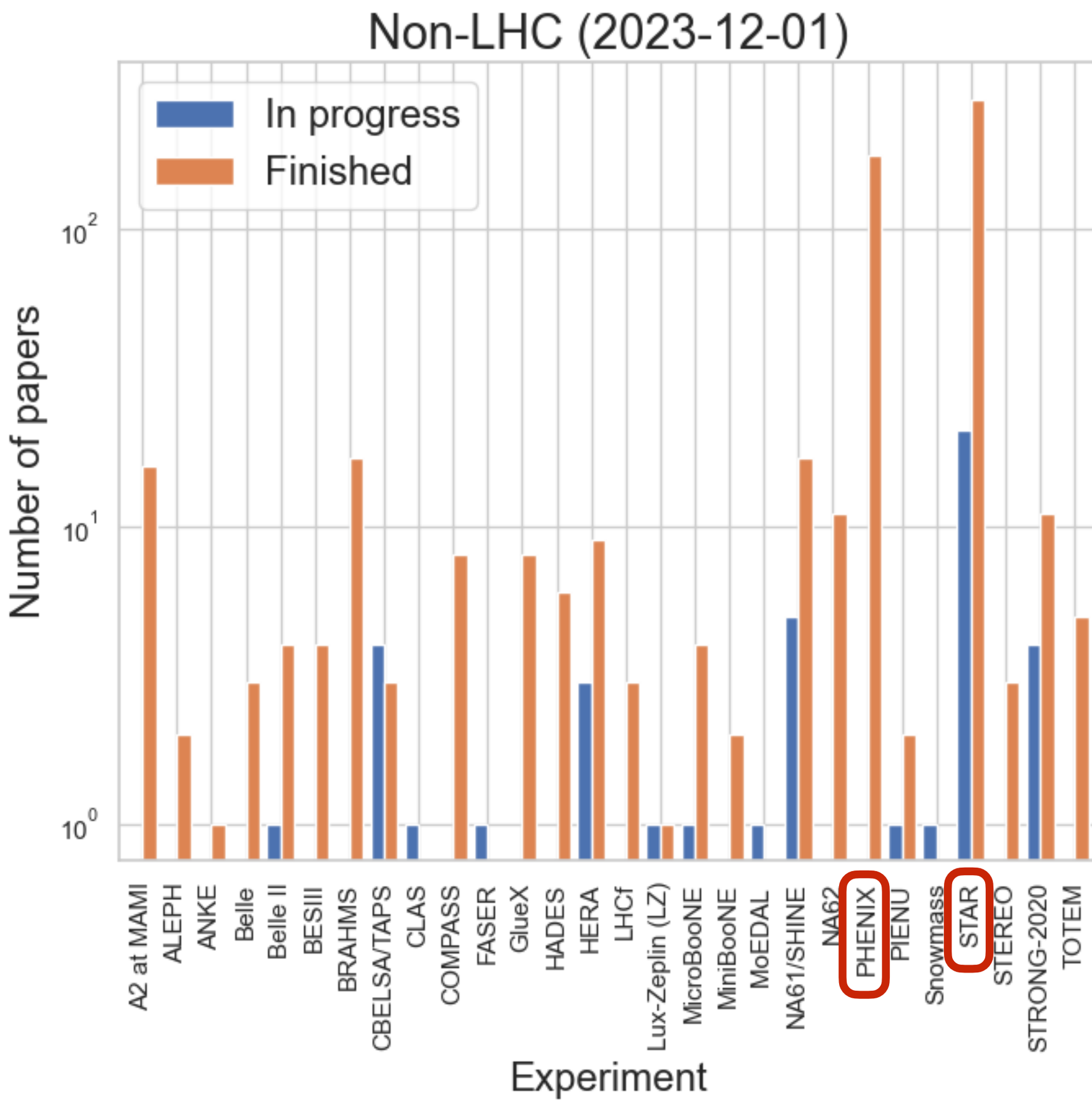
- Search INSPIRE for publications with HEPData (GitHub/Binder).

Submissions by ATLAS/CMS groups (2017-)





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



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

Record from hepdata.net

HEPData

hepdata.net/record/ins1860766

HEPData | Search HEPData | Search | About | Submission Help | File Formats | Sign in

Browse all | Abudinén, F. et al. | Last updated on 2022-08-29 13:50 | Accessed 278 times | Cite | JSON

Hide Publication Information

Search for $B^+ \to K^+ \nu\bar{\nu}$ Decays Using an Inclusive Tagging Method at Belle II

The Belle-II collaboration

Abudinén, F. , Adachi, I. , Adamczyk, K. , Ahlburg, P. , Aihara, H. , Akopov, N. , Aloisio, A. , Ky, N. Anh , Asner, D.M. , Atmacan, H.

Phys.Rev.Lett. 127 (2021) 181802, 2021.

https://doi.org/10.17182/hepdata.130199

Journal | INSPIRE | Resources

Abstract (data abstract)

SuperKEKB Belle II. Measurement of the branching fraction of $B^+ \to K^+ \nu\bar{\nu}$ at the Belle II experiment at the SuperKEKB. The analysed data sample corresponds to an integrated luminosity of 63 fb^{-1} collected at the $\Upsilon(4S)$ resonance and a sample of 9 fb^{-1} collected at an energy 60 MeV below the resonance between 2019-2021. Since no significant signal was observed, limit of 4.1×10^{-5} was set using CL$_s$ method.

Download All

- YAML with resource files
- YAML
- YODA
- ROOT
- CSV

Postfit yields off-resonance

Figure 1 in https://journals.aps.org/prl/supplemental/1

10.17182/hepdata.130199.v1/t3

Yields in off-resonance data and as predicted by the simultaneous fit to the on- and off-resonance data, corresponding to an...

Expected and observed limit

Figure 2 in https://journals.aps.org/prl/supplemental/1

10.17182/hepdata.130199.v1/t4

CL$_s$ value as a function of the branching fraction of $B^+ \to K^+ \nu\bar{\nu}$ for expected and observed signal...

Selection efficiency

Figure 3 in https://journals.aps.org/prl/supplemental/1

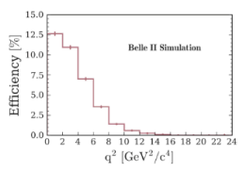
10.17182/hepdata.130199.v1/t5

Signal efficiency as a function of the dineutrino invariant mass squared ...

Selection efficiency 10.17182/hepdata.130199.v1/t5 | Resources | https://www.hepdata.ne | JSON

Figure 3 in https://journals.aps.org/prl/supplemental/10.1103/PhysRevLett.127.181802/suppl_mat.pdf

Signal efficiency as a function of the dineutrino invariant mass squared q^2 for events in the signal region (SR) ($\mathrm{BDT}_1 > 0.9$ and $\mathrm{BDT}_2 > 0.95$). The error bars indicate the statistical uncertainty.



phrases

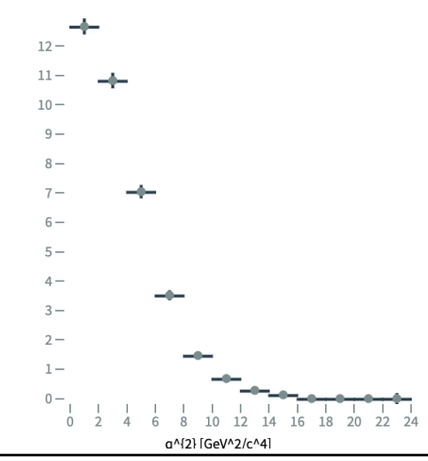
FCNC | b --> s l l transition | electroweak penguin decay | missing energy

reactions

$B^+ \to K^+ \nu\bar{\nu}$

Luminosity	63+9 fb^{-1}
q^2 [GeV$^2/c^4$]	Efficiency
0.0 - 2.0	12.66745696 ±0.27207295
2.0 - 4.0	10.82571692 ±0.26463688
4.0 - 6.0	7.04488885 ±0.2278063
6.0 - 8.0	3.51769225 ±0.1711566
8.0 - 10.0	1.46683133 ±0.11813559
10.0 - 12.0	0.68175914 ±0.08670158
12.0 - 14.0	0.27954132 ±0.06074916

Visualize



- DOIs minted for whole (versioned) record and each table.

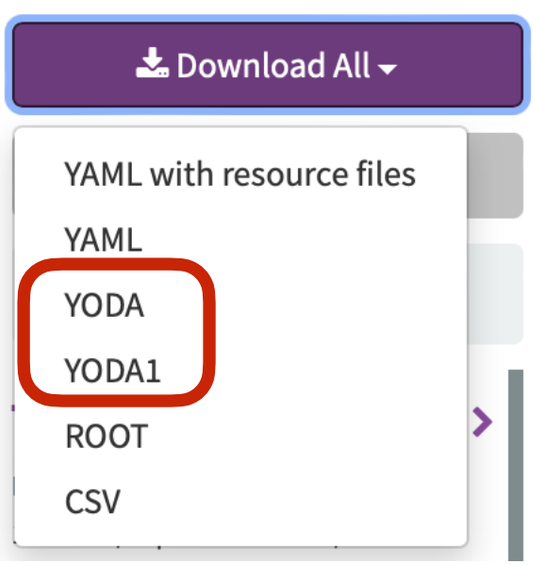
Data output formats

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.
- YODA: for inclusion in a Rivet analysis.
- ROOT: binary `.root` file.



- NEW "YODA" now gives new YODA2 format.
- Legacy YODA format still available as YODA1.
- Thanks to Chris Gütschow for work on implementing the YAML → YODA2 conversion.

Machine-readable metadata

- Schema.org JSON-LD embedded in web pages (→ Google):
  `curl https://www.hepdata.net/record/ins1419244`
- Search and record pages each have a JSON equivalent:
  `https://www.hepdata.net/search/?q=collaborations:ATLAS&format=json`
  `https://www.hepdata.net/record/ins1419244?format=json`
- Example: `access_count` of ATLAS records (GitHub/Binder).
- `format={original,yaml,json,csv,yoda,root}`
- Additional (optional) arguments for `version` and `table`.
- Option `light=true` to remove data tables from the JSON.
- See `hepdata.net/formats` for more information.

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`
- Examples of usage:

```
hepdata-cli find 'collaborations:"Belle-II"' -i inspire
```

```
hepdata-cli fetch-names 1860766 -i inspire
```

```
hepdata-cli download 1860766 -f csv -i inspire
```

```
hepdata-cli upload /path/to/TestHEPSubmission.tar.gz -e
my@email.com -p $PASSWORD –r 123456 -i $INVITATION_COOKIE -s False
```

Code: `https://github.com/HEPData/hepdata-cli`

submission.yaml file

- Separate YAML "document" for each data table, e.g.

```yaml
---
# This is Table 3.
name: "Table 3"
location: Data from Figure 8A
description: Normalized ZZ fiducial cross section (multiplied by 10^6 for
keywords: # used for searching, possibly multiple values for each keyword
- {name: reactions, values: [P P --> Z0 Z0 X]}
- {name: observables, values: [DSIG/DPT]}
- {name: cmenergies, values: [7000.0]}
- {name: phrases, values: [Inclusive, Single Differential Cross Section,
data_file: data3.yaml
additional_resources:
- {description: Image file, location: figFigure8A.png}
- {description: Thumbnail image file, location: thumb_figFigure8A.png}
```

- Validated against `submission_schema.json`

YAML data file

- "Independent" and "dependent" variables:

```yaml
independent_variables:
- header: {name: Leading dilepton PT, units: GEV}
  values:
  - {low: 0, high: 60}
  - {low: 60, high: 100}
  - {low: 100, high: 200}
  - {low: 200, high: 600}
dependent_variables:
- header: {name: 10**6 * 1/SIG(fiducial) * D(SIG(fiducial))/DPT, units: GEV**-1}
  qualifiers:
  - {name: RE, value: P P --> Z0 < LEPTON+ LEPTON- > Z0 < LEPTON+ LEPTON- > X}
  - {name: SQRT(S), units: GEV, value: 7000}
  values:
  - value: 7000
    errors:
    - {symerror: 1100, label: stat}
    - {symerror: 79, label: 'sys,detector'}
    - {symerror: 15, label: 'sys,background'}
  - value: 9800
    errors:
    - {symerror: 1600, label: stat}
    - {symerror: 75, label: 'sys,detector'}
    - {symerror: 15, label: 'sys,background'}
  - value: 1600
    errors:
    - {symerror: 490, label: stat}
    - {symerror: 41, label: 'sys,detector'}
    - {symerror: 2, label: 'sys,background'}
  - value: 80
    errors:
    - {symerror: 60, label: stat}
    - {symerror: 2, label: 'sys,detector'}
    - {symerror: 0, label: 'sys,background'}
```

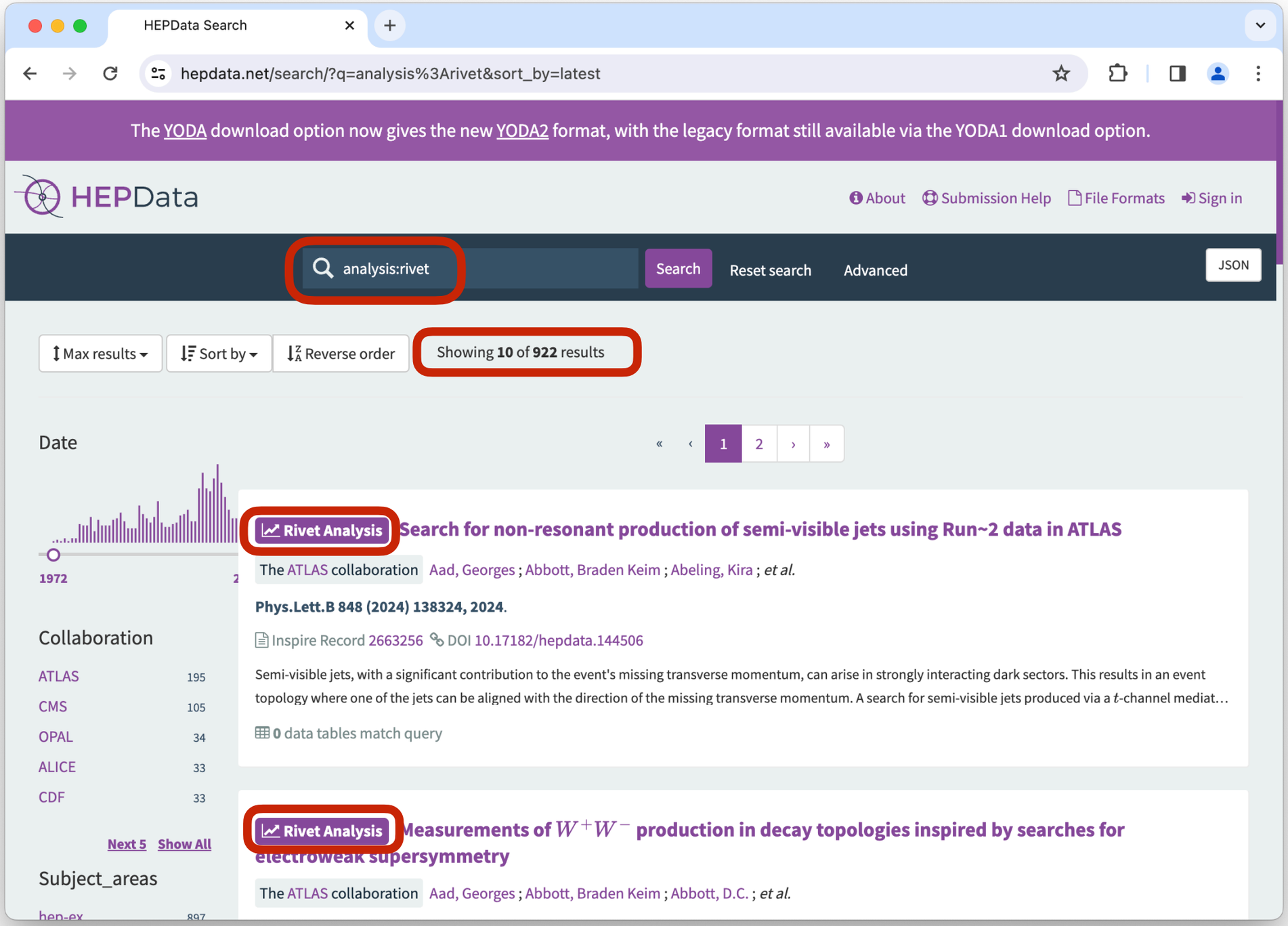
RE	P P --> Z0 < LEPTON+ LEPTON- > Z0 < LEPTON+ LEPTON- > X
SQRT(S)	7000.0 GeV
Leading dilepton PT [GEV]	**10**6 * 1/SIG(fiducial) * D(SIG(fiducial))/DPT [GEV**-1]**
0.0 - 60.0	7000.0 ±1100.0 stat ±79.0 sys,detector ±15.0 sys,background
60.0 - 100.0	9800.0 ±1600.0 stat ±75.0 sys,detector ±15.0 sys,background
100.0 - 200.0	1600.0 ±490.0 stat ±41.0 sys,detector ±2.0 sys,background
200.0 - 600.0	80.0 ±60.0 stat ±2.0 sys,detector

- Validated against `data_schema.json`

Submission documentation

- Documentation at hepdata-submission.readthedocs.io. Includes example Python scripts (simple, complicated).
- HEPData YAML files checked against JSON schema by validation code during submission. `hepdata-validate`
- 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
  🆕 Latest v0.14.0 can convert scikit-hep/hist histograms.
- Similar Python package by *Christian Holm Christensen*. https://gitlab.com/cholmcc/hepdata
- Experiments often develop internal HEPData tools/docs.

Connection to Rivet analyses



- HEPData exports data in YODA format for Rivet.

Links to analysis code

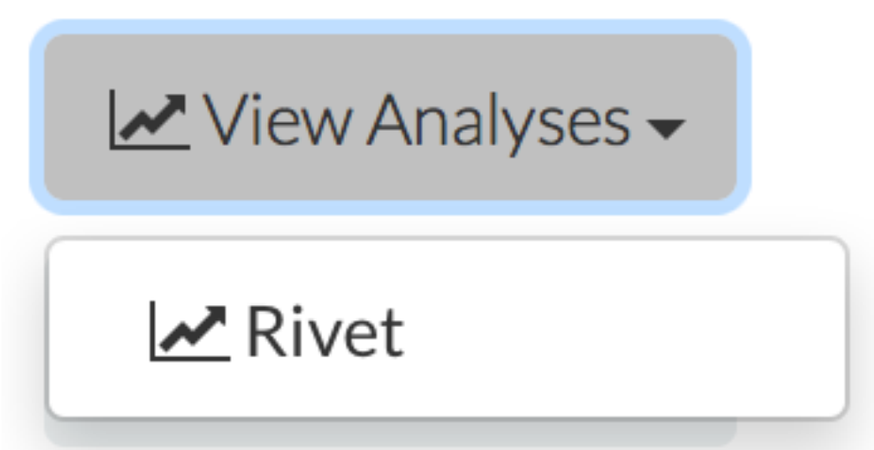
http://rivet.hepforge.org/analyses.json

- JSON file maps INSPIRE IDs to Rivet analysis names:

```
{"422172": [“SLD_1996_S3398250”], …,
 “1304688”: [“ATLAS_2014_I1304688”]}
```

- 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



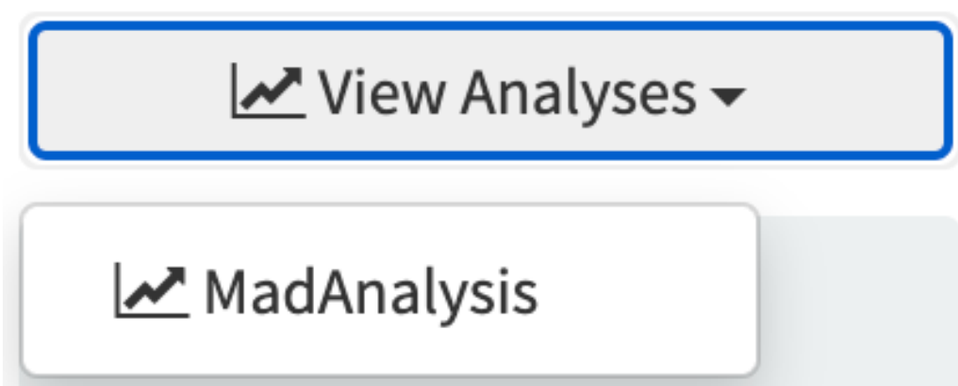
- Extendable to other analysis frameworks containing publication-specific code.

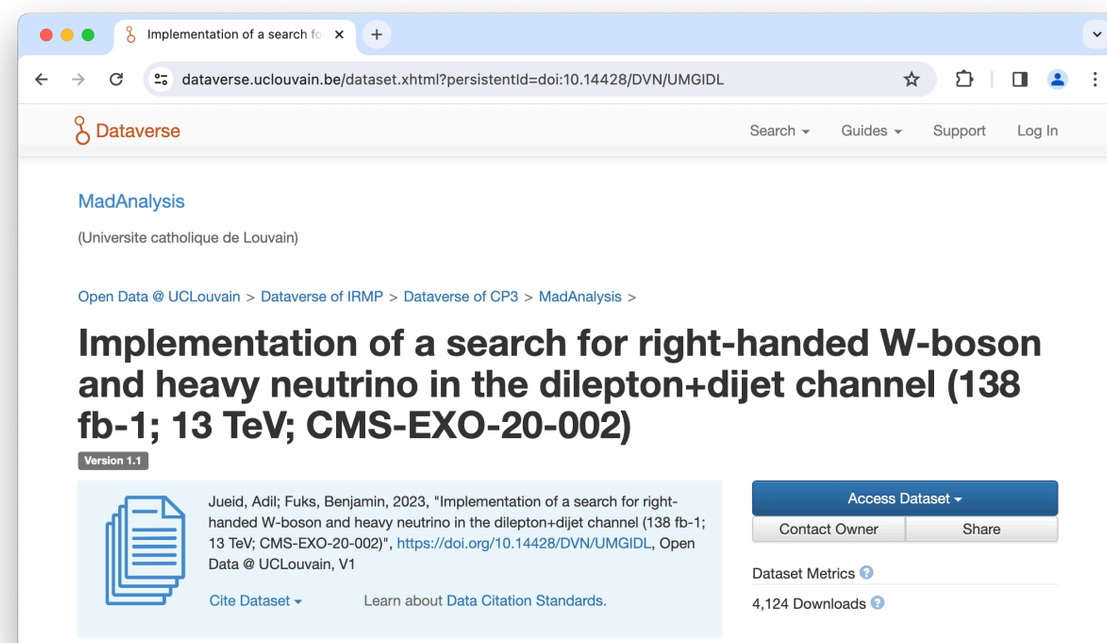
Links to MadAnalysis 5

- `analyses.json` file for MadAnalysis 5 analyses:

```
{"1458270": [“10.14428/DVN/MHPXX4”], …,
  “1750186”: [“10.14428/DVN/OFAE1G"]}
```

- Search `analysis:MadAnalysis` gives 26 records.





- Thanks to Jack Araz.

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`:

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

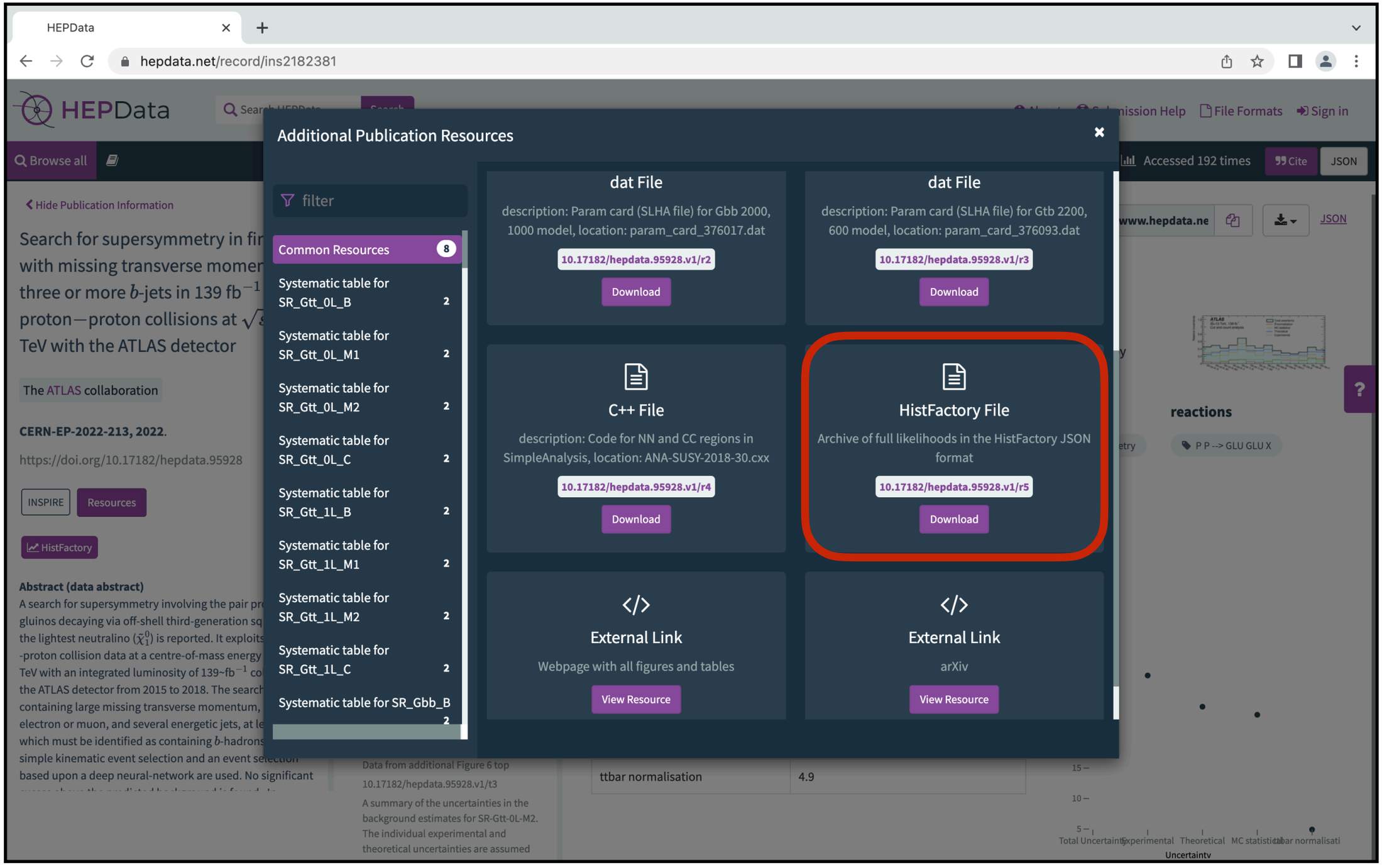
or use hepdata_lib

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

```yaml
related_to_hepdata_records:
- 12345
- 67890
```

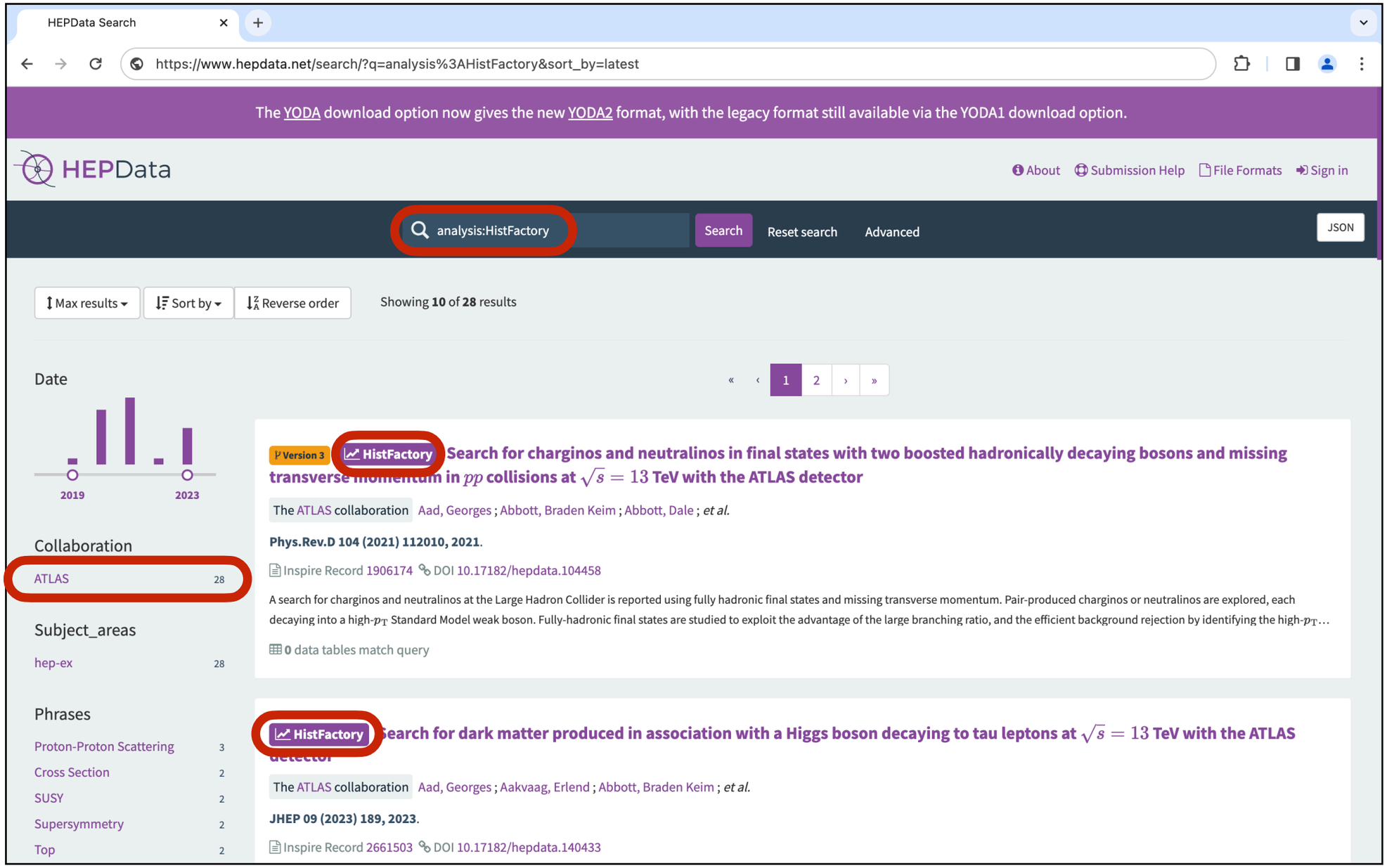
or use hepdata_lib

Additional resource files



- DOIs now minted via DataCite for additional resource files.

pyhf likelihoods



- Search `analysis:HistFactory`.

Native support for pyhf JSON?

- **Idea:** support submissions with pyhf JSON files replacing the usual HEPData YAML data tables. Provide a visualization of the pyhf JSON files within the HEPData record.
- First step completed by Sinclert Pérez (NYU) in 2020: modify `hepdata-validator` to use a remote JSON schema.
- Idea abandoned as too complicated and not well-defined.
- New idea (2022): standalone web application to visualize pyhf JSON data that can import resource files from HEPData:

1. hf Explorer: `http://www.hepexplorer.net/`
2. WorkspaceExplorer: `https://workspaceexplorer.app.cern.ch/`

Formats for pyhf resource files

- pyhf JSON files usually uploaded to HEPData in an archive (`.zip`, `.tar`, `.tar.gz`, `.tgz`, `.tar.xz`) or as `.json`.
- Visualization web apps should have ability to import archive.
- Identify as "HistFactory" if one of these file extensions and a (case-insensitive) trigger word {`histfactory`, `pyhf`, `likelihoods`, `workspaces`} in the `description`.
- Alternatively, submitter can add an explicit `type` field:

```
additional_resources:
- location: "Likelihoods.tar.gz"
  description: "Archive of full likelihoods in the HistFactory JSON format"
  type: "HistFactory" # (optional) currently supports 'HistFactory' type to allow HistFactory JSON (pyhf) file
```

- See submission docs and hepdata_lib docs for more details.

Follow @HEPData

Summary

Email: `info@hepdata.net`
Forum: hepdata-forum.cern.ch

- **HEPData** is *the* repository for publication-related HEP data.
- *Caveats*: design restricts size (~MB) and format (mostly tabular).
- `hepdata-cli` search/download/upload from command line.
- **New features:** YODA2, *bidirectional linking* of tables/records.
- Enhanced linking to Rivet analyses and MadAnalysis 5 analyses.
- pyhf JSON likelihood data hosted as additional resource files.
- Potential to add a "Visualize" button for "HistFactory" files that would open a visualization web app and load the pyhf JSON.

Code: `https://github.com/HEPData`