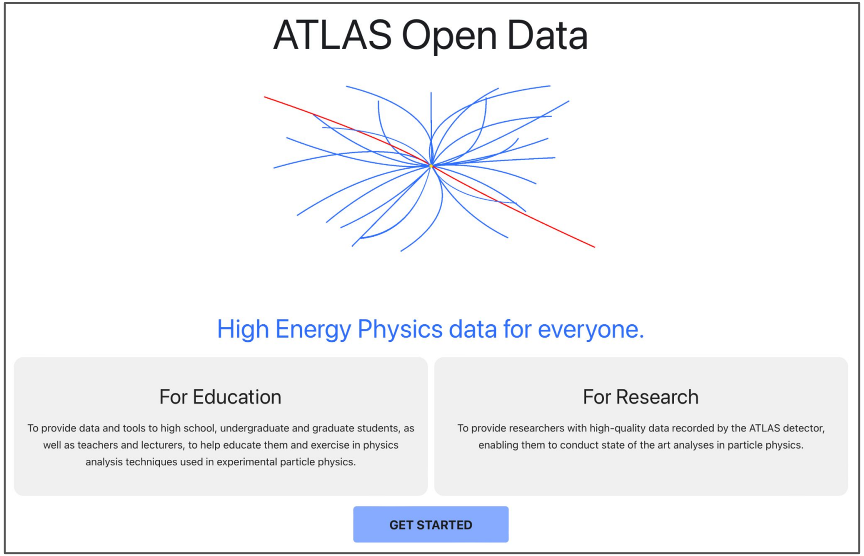


ATLAS open and preserved data: status and plans

4th DPHEP Collaboration Workshop

3 October 2024

Zach Marshall (LBNL) on behalf of the ATLAS Collaboration

ATLAS Open and Preserved Data

- RAW data and software will be preserved; RAW data will not be made public; software is all public (nothing new at Level 4)
- Good advances on other areas of open data and data / analysis preservation
- Managed to maintain a high level of commitment to open data despite the **huge** number of publications in the last year

Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

Data / Analysis Preservation (External)

- For ~all analyses we provide Plot records and HepData
  - In many cases the HepData are sufficient to create a simplified analysis externally
  - For long-lived particle searches (and some more complex ‘standard’ searches) we provide instructions for use of reinterpretation material, since this is not trivial

Papers and publications of ATLAS physics and performance results 207 documents (Published: 204 - Accepted: 2 - Submitted: 1)
(Full list of ATLAS papers, List/RSS from CDS)

Hide table

Showing 21 to 30 of 207 entries

Overall search (incl. keywords - click here for full list):

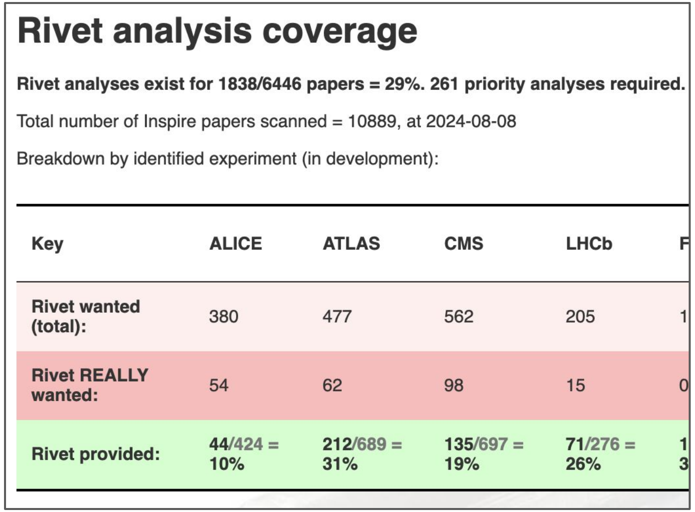
Short Title	Group	Journal Reference	Date	√s ( )	L	Links
search...	search...	search...	search...	search...		search...
H(125)→invisible combination	HIGG	Phys. Lett. B 842 (2023) 137963	2023-01-25	13	139 fb^{-1}	Documents | 2301.10731 | Inspire Internal
Search for gluinos in multi-b final states	SUSY	Eur. Phys. J. C 83 (2023) 561	2022-11-15	13	139 fb^{-1}	Documents | 2211.08028 | Inspire HepData |
WIMP DM with heavy flavours combination	SUSY	Eur. Phys. J. C 83 (2023) 503	2022-11-10	13	139 fb^{-1}	Documents | 2211.05426 | Inspire HepData |
Search for electroweak SUSY in events with 2 leptons and no jets	SUSY	JHEP 06 (2023) 031	2022-09-28	13	139 fb^{-1}	Documents | 2209.13935 | Inspire HepData |
Non-pointing photons	SUSY	Phys. Rev. D 108 (2023) 032016	2022-09-02	13	139 fb^{-1}	Documents | 2209.01029 | Inspire Internal
Unfolding of WW from 2L0J SUSY EWK search	SUSY	Eur. Phys. J. C 83 (2023) 718	2022-06-30	13	139 fb^{-1}	Documents | 2206.15231 | Inspire HepData |
Search for new phenomena in final states with photons, jets and missing transverse momentum	SUSY	JHEP 07 (2023) 021	2022-06-13	13	139 fb^{-1}	Documents | 2206.06012 | Inspire HepData |
Search for charged long-lived particles in final states with high pixel ionization loss	SUSY	JHEP 06 (2023) 158	2022-05-12	13	139 fb^{-1}	Documents | 2205.06013 | Inspire HepData |
Search for SUSY in events with 2 leptons, jets and MET	SUSY	Eur. Phys. J. C 83 (2023) 515	2022-04-27	13	139 fb^{-1}	Documents | 2204.13072 | Inspire HepData |
Disappearing track	SUSY	Eur. Phys. J. C 82 (2022) 606	2022-01-07	13	139 fb^{-1}	Documents | 2201.02472 | Inspire Internal

Show 10 entries

Previous 1 2 3 4 5 … 21 Next

Data / Analysis Preservation (External)

- For ~all analyses we provide Plot records and HepData
  - In many cases the HepData are sufficient to create a simplified analysis externally
  - For long-lived particle searches (and some more complex ‘standard’ searches) we provide instructions for use of reinterpretation material, since this is not trivial
- For Standard Model measurements we also provide Rivet routines
- For many searches we provide ‘Simple Analysis’ routines
  - Easily converted to your favorite standard



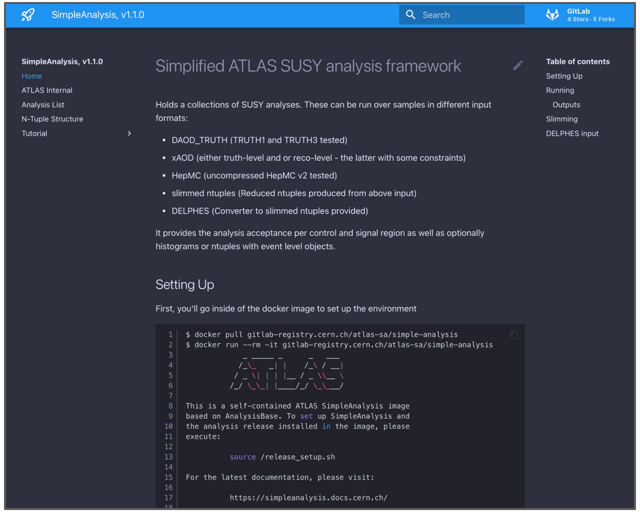
Rivet analysis coverage

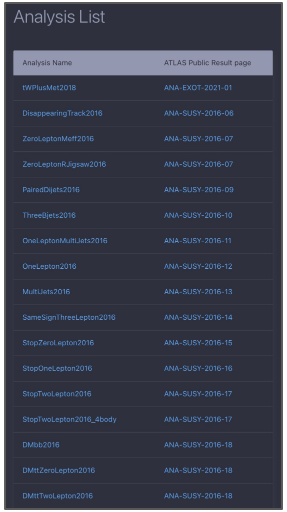
Rivet analyses exist for 1838/6446 papers = 29%. 261 priority analyses required.

Total number of Inspire papers scanned = 10889, at 2024-08-08

Breakdown by identified experiment (in development):

Key	ALICE	ATLAS	CMS	LHCb	F
Rivet wanted (total):	380	477	562	205	1
Rivet REALLY wanted:	54	62	98	15	0
Rivet provided:	44/424 = 10%	212/689 = 31%	135/697 = 19%	71/276 = 26%	1 3



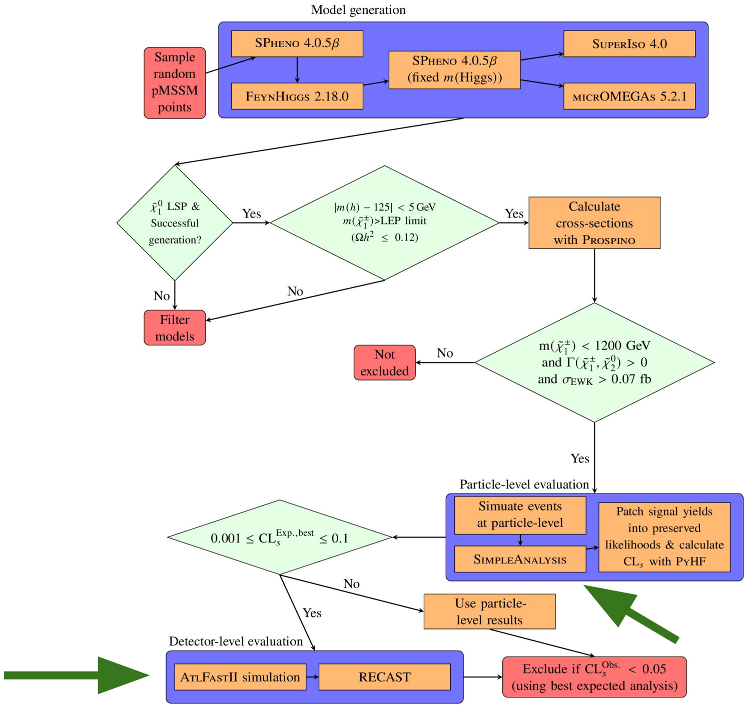


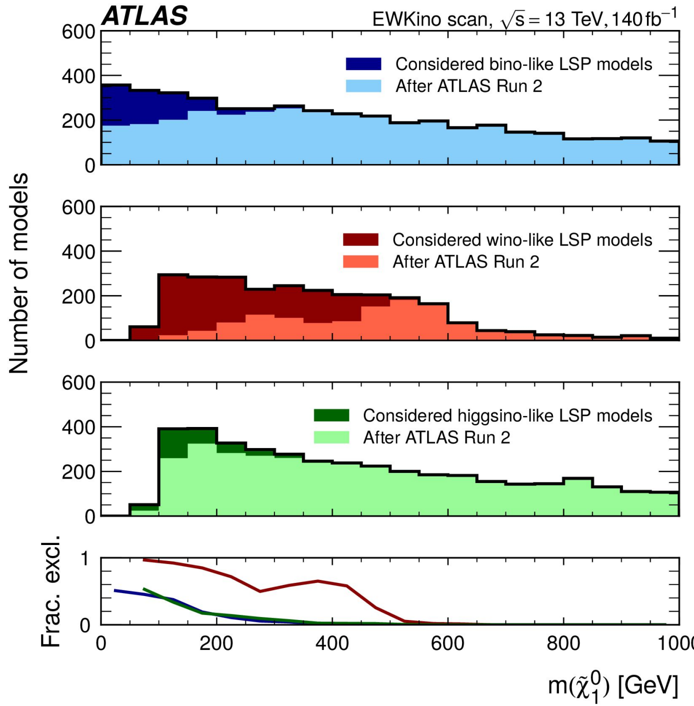
Analysis List

Analysis Name	ATLAS Public Result page
tWPlusMet2018	ANA-EXOT-2021-01
DisappearingTrack2016	ANA-SUSY-2016-06
ZeroLeptonMeff2016	ANA-SUSY-2016-07
ZeroLeptonRJigsaw2016	ANA-SUSY-2016-07
PairedDijets2016	ANA-SUSY-2016-09
ThreeBjets2016	ANA-SUSY-2016-10
OneLeptonMultiJets2016	ANA-SUSY-2016-11
OneLepton2016	ANA-SUSY-2016-12
MultiJets2016	ANA-SUSY-2016-13
SameSignThreeLepton2016	ANA-SUSY-2016-14
StopZeroLepton2016	ANA-SUSY-2016-15
StopOneLepton2016	ANA-SUSY-2016-16
StopTwoLepton2016	ANA-SUSY-2016-17
StopTwoLepton2016_4body	ANA-SUSY-2016-17
DMbb2016	ANA-SUSY-2016-18
DMttZeroLepton2016	ANA-SUSY-2016-18
DMttTwoLepton2016	ANA-SUSY-2016-18

Data / Analysis Preservation (Internal)

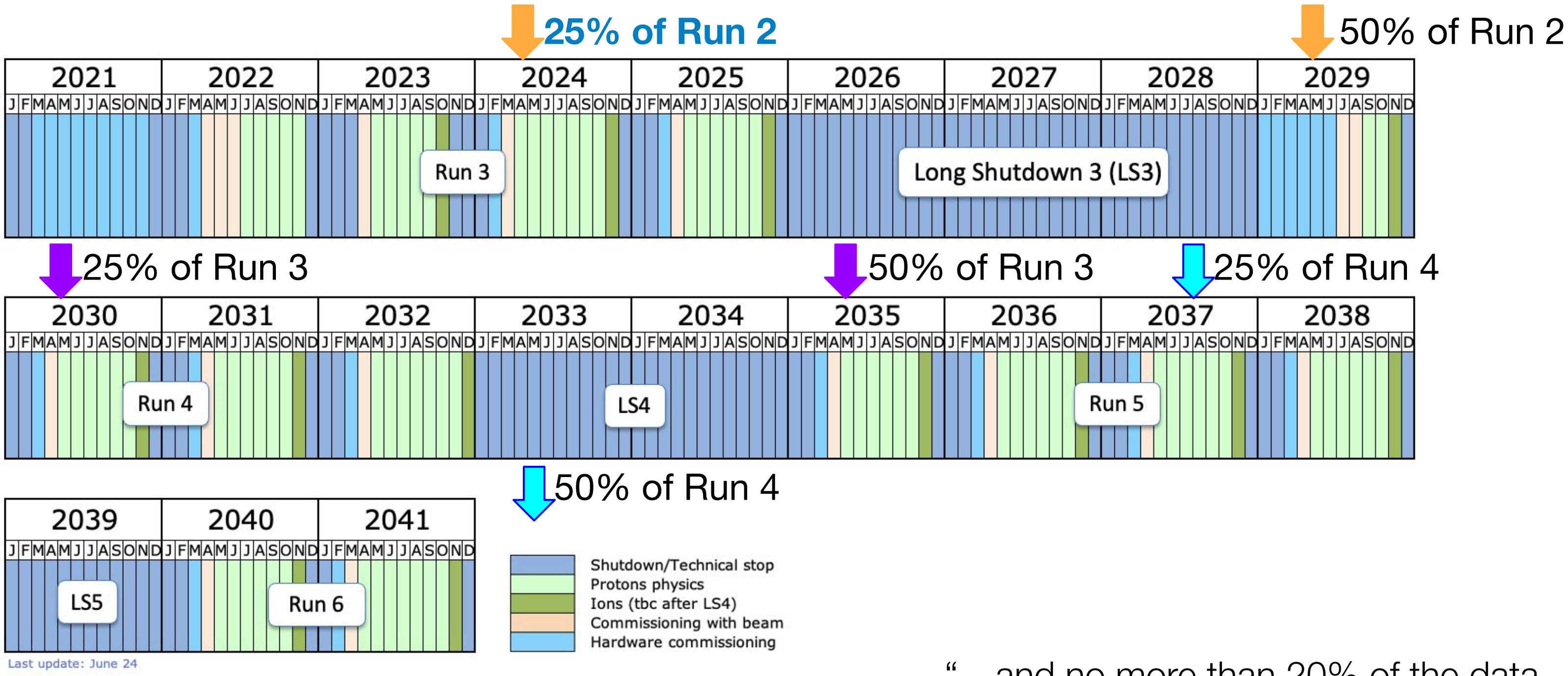
- Internally, we try to do a bit more for data preservation
- Saving searches using REANA and RECAST for later re-use in model scans
- Successful but **heavy** — often hard to finish this work with the paper





Open Data for Research

- New for us this year!
- ATLAS Schedule (to be updated when the HL-LHC shift is finalized…):



"… and no more than 20% of the data available to the collaboration…"

What did we Release?

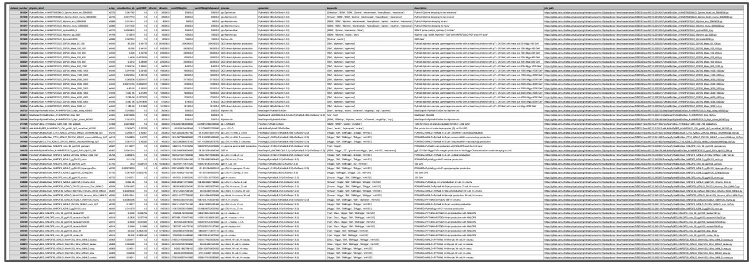
ATLAS DAOD_PHYSLITE format Run 2 2016 proton-proton collision data

ATLAS collaboration

Cite as: ATLAS collaboration (2024). ATLAS DAOD_PHYSLITE format Run 2 2016 proton-proton collision data. CERN Open Data Portal. DOI:10.7483/OPENDATA.ATLAS.4ZES.DJHA

Dataset characteristics

5383448881 events. **45571** files. **35.4 TiB** in total.



Documentation on PHYSLITE Variables for ATLAS Open Data

Page generated from sample: mc20_13TeV.410471.PhPy8EG_A14_ttbar_hdamp258p75_allhad.deriv.DAOD_PHYSLITE.e6337_s3681_r13167_p5631

List of Containers:

AnalysisElectrons | AnalysisJets | AnalysisLargeRJets | AnalysisMuons | AnalysisPhotons | AnalysisTauJets | AnalysisTrigMatch | AntiKt10TruthSoftDropBeta100Zcut10Jets | AntiKt4TruthDressedWZJets | BTagging_AntiKt4EMPFlow | CombinedMuonTrackParticles | egammaClusters | EventInfo | ExtrapolatedMuonTrackParticles | GSFConversionVertices | GSFTrackParticles | HardScatterParticles | HardScatterVertices | InDetTrackParticles | Kt4EMPFlowEventShape | MET_Core_AnalysisMET | MET_Truth | MuonSpectrometerTrackParticles | PrimaryVertices | TauTracks | TruthBoson | TruthBosonsWithDecayParticles | TruthBosonsWithDecayVertices | TruthBottom | TruthElectrons | TruthEvents | TruthForwardProtons | TruthMuons | TruthNeutrinos | TruthPhotons | TruthPrimaryVertices | TruthTaus | TruthTop

AnalysisElectrons[back to top]

Variable Name	Type	Description
ambiguityLink	vector<ElementLink<DataVector<xAOD::Egamma_v1> > >	Links Photon<-> Electron when ambiguous
ambiguityType	vector<unsigned char>	Ambiguity (almost surely electron 0 or photon 7/6 (>= rel22/rel21) or ambiguous 1-6/5, (>= rel22/rel21)
author	vector<unsigned short>	Electron, Photon, Ambiguous, Forward Electron
caloClusterLinks	vector<vector<ElementLink<DataVector<xAOD::CaloCluster_v1> > > >	Photon/electron -> Cluster
charge	vector<float>	Electron charge from best track match (+/- 1)
DFCommonElectronsECIDS	vector<char>	Charge selection (to reject wrong charge assignment)
DFCommonElectronsECIDSResult	vector<double>	BDT (Boosted Decision Tree) score for the charge selection
DFCommonElectronsLHLoose	vector<char>	Likelihood identification decision
DFCommonElectronsLHLooseBL	vector<char>	Likelihood identification decision

- 2015+2016 Run 2 pp collision data
  - 45 TB of data, 6.3 kB/event, 7.1B events, 55k files in ~300 runs
  - 20 TB of MC, ~10 kB/event, 2B events, 16k files in ~300 MC datasets
- Explanation of our nomenclature
- Giant tables of **metadata**
  - Cross sections, k-factors, filters / efficiencies, processes, how to combine samples, configurations, …
- PHYSLITE (ROOT-based) format
  - Already columnar — Uproot friendly
  - Used for our own papers too
- Pre-calibrated (first for ATLAS)
  - Just draw a plot!
- Extensive effort to document usage
- **Super useful for us as well!**

Documentation!

- We've constructed "paths" through the Open Data for different kinds of users

🚀 Quick start

The quickest way to start learning with ATLAS Open Data.

🤿 Deep Dive

For extended use. Let's dive into what ATLAS has to offer!

🔬 Researchers Toolkit

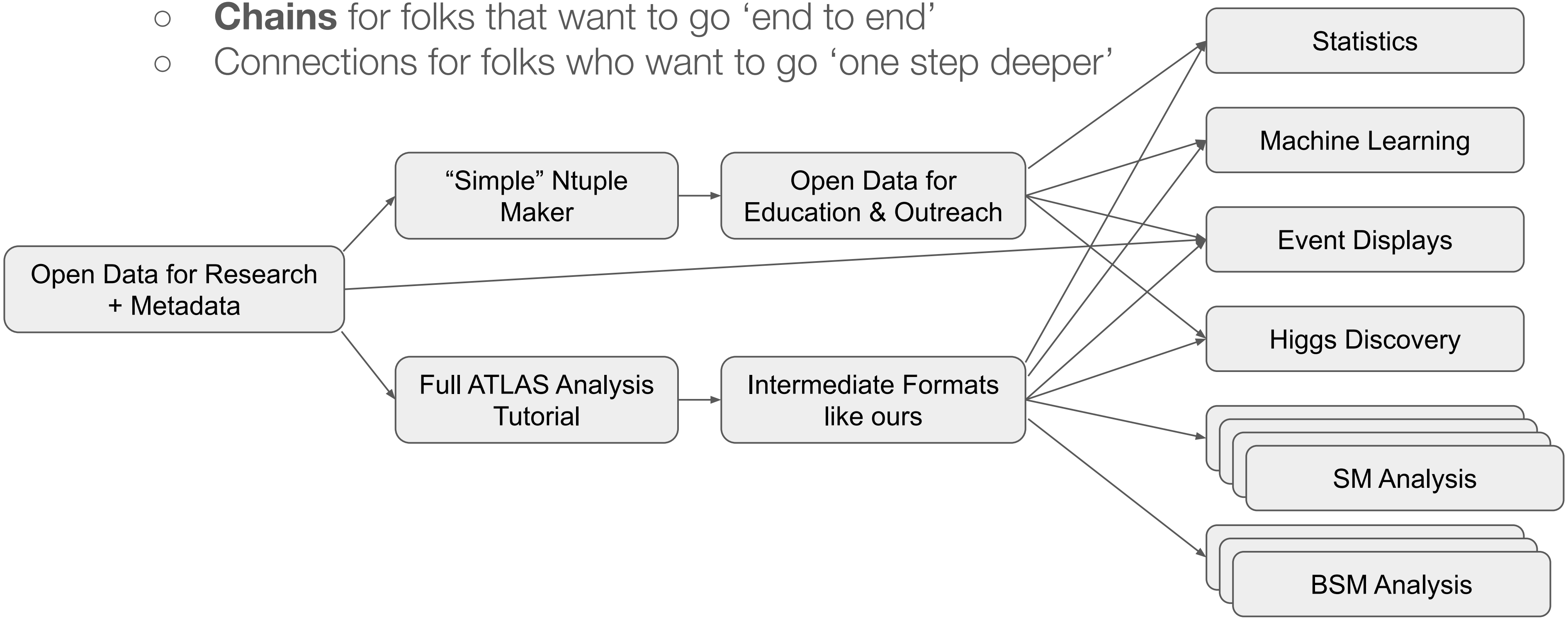
Detailed information and resources for researchers

📊 Online Data Analyser

Explore ATLAS Open Data at a glance!

Documentation!

- We've constructed "paths" through the Open Data for different kinds of users
- The eventual goal is to have a **web** of paths that satisfy many users and interests
  - Independent **modules** for specific learning objectives (hours – months)
  - **Chains** for folks that want to go 'end to end'
  - Connections for folks who want to go 'one step deeper'



Documentation: The Really Hard Part

Setting Uncertainties

One of the most important parts of any data analysis is the inclusion of proper uncertainties. Uncertainties help quantify the reliability and precision of a conclusion obtained from data.

When comparing detector data to simulations, you can see a difference that might seem significant. However, whether that difference is interesting or important requires understanding uncertainties. Agreement within uncertainties implies that the observed and predicted values are consistent. If a number is measured to be 1000 and it was predicted to be 2000±1000, then the measurement and prediction agree. Despite the measurement appearing far from the prediction, the large uncertainty range indicates that the prediction is not very precise, allowing for agreement.

Similarly, it is important not misinterpret agreement that is better than the uncertainty suggests. If a number is measured to be 1000 and the prediction was 1000±500, that does not mean that the true value will be 1000. A more precise model might give a prediction of 600±100, which would be in consistent with the original prediction, but would no longer agree with the measurement.

A key part of scientific training is understanding when a difference between a prediction and an observation is meaningful and significant, and that comes down to understanding uncertainties.

Why Consider Uncertainties?

In ATLAS analyses we consider uncertainties for several reasons:

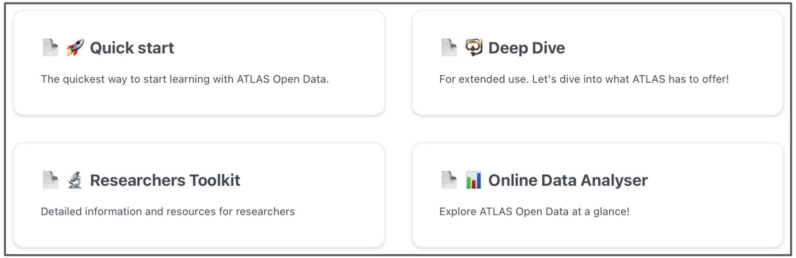
1. **Accurate Parameter Estimation:** To get reliable estimates of the parameters of interest (POIs), such as the Higgs boson couplings or the top quark mass, we need to account for all sources of uncertainty. Ignoring systematic uncertainties can lead to biased estimates and incorrect conclusions.
2. **Robust Hypothesis Testing:** In testing theoretical models against experimental data, systematic uncertainties ensure that discrepancies between the observed data and theoretical predictions are not mistakenly attributed to new physics or phenomena, instead that they are correctly identified as derived from known uncertainties in the experimental or theoretical setup.

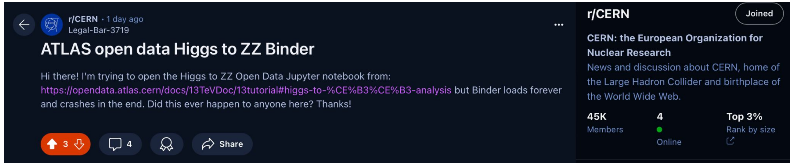
- The **hardest part** of an analysis is understanding and calculating systematic uncertainties
- Explaining how to do this in an approachable way is **extraordinarily difficult** and **important**

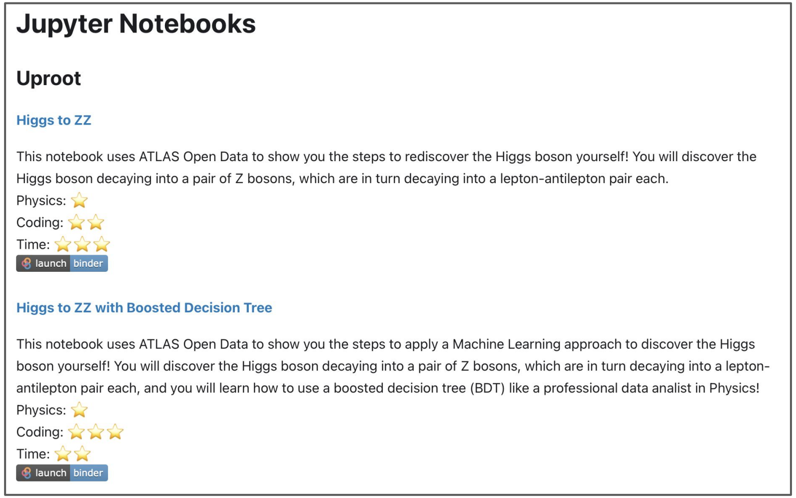
- **Evergreen documentation** of concepts
  - Useful for our students as well!
  - Can be integrated into our tutorials
- **Technical documentation** for code
  - **Momentarily** matches internal documentation until we move on (except CERN-specifics, Grid use…)
  - Needs to be **fool proof** to avoid science problems
  - Good examples are a **huge** help here
- Related documentation we haven't written yet: what you **cannot** do
  - Things our systematics don't cover
  - Things our datasets / simplification don't permit

Open Data Feedback

- July 1, Open Data for Research was released
  - ATLAS news article; EP news article
- People did find it, even outside the CERN Community
- Bulking up documentation (incl. available resources)
- Not a lot of feedback from large-scale users yet; good discussion of KPI monitoring in the CERN ODWG
  - We have citations and data access monitoring
  - Working on many other KPIs (see also the OSPF)
- Considering a workshop a la CMS in 2025
- Heavy Ion open data (next) in production / validation
  - Hope to get this release out before the end of the year
  - Also matching pp/HI Education and Outreach open data



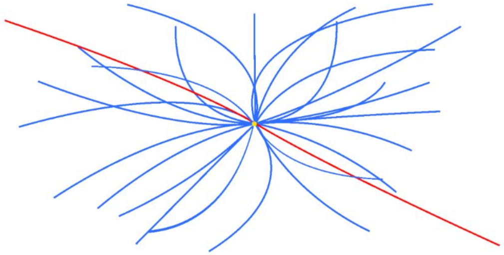




Other Open Data and Notes

- Several bespoke Open Data (mostly MC) sets in the last ~year
  - ATLAS Top Tagging Open Data Set with Systematic Uncertainties
    - Upgrade of the ATLAS Top Tagging Open Data Set
    - A publicly available data set for the development of Machine Learning (ML) based boosted top tagging algorithms, this time including systematic uncertainties
  - WH1lbb SUSY search datasets
    - BDT training datasets
  - Omnifold 24-dimensional Z+jets datasets
    - Unfolded measurement data, useable for reinterpretations
- These often accompany notes, papers, or other studies, and come when the teams are ready
  - Generally pretty good engagement experience, though not as much ‘true’ non-ATLAS usage as we might wish for
- Ongoing discussion re: Open Data as *the* support mechanism for legacy data
  - That is, do we try to continue to support simulation and reconstruction of our Run 1 RAW data forever? Or do we “just” keep the Open Data and that’s that?
- There’ll be a talk at CHEP describing some of this in more detail

ATLAS Open Data



High Energy Physics data for everyone.

NEW

For Education

To provide data and tools to high school, undergraduate and graduate students, as well as teachers and lecturers, to help educate them and exercise in physics analysis techniques used in experimental particle physics.

For Research

To provide researchers with high-quality data recorded by the ATLAS detector, enabling them to conduct state of the art analyses in particle physics.

GET STARTED

Thank you!

CERN

ATLAS EXPERIMENT