FAIROS-HEP + AGC

Analysis Preservation & Reinterpretation

Kyle Stuart Cranmer (Wisconsin)
Mike Hildreth (Notre Dame)
Peter Elmer (Princeton)

FAIR and Open Science in High Energy Physics
OAC-2226378, OAC-2226379 and OAC-2226380

FAIROS-HEP

Recently, the US National Science Foundation funded a new Research Coordination Network project titled “FAIROS-HEP”.

Findable

Accessible

Interoperable

Reusable

Open **S**cience

The NSF’s FAIROS Research Coordination Networks

Findable, Accessible, Interoperable, Reusable, Open Science Research Coordination Networks (FAIROS RCN) program represent a pooled investment of over $12.5 million in open science from all directorates comprising NSF.

FAIROS RCN supports groups of investigators to communicate, innovate, coordinate, and standardize research practices, training, and educational activities across disciplinary, organizational, geographic, and international boundaries to achieve the goals of FAIR and other open-science guiding principles.

Research coordination networks are a form of awards that NSF makes to advance scientific practices and standards broadly across multiple research fields. These RCN awards will be for **three-year projects**.

FAIRQS-HEP Continues a Legacy of Contributions

DASPOS (2012-2016)

- https://daspos.crc.nd.edu/
- Contributions to RECAST led to REANA as a spinoff project now led by CERN
- Supported REANA Common Workflow Language

DIANA-HEP (2015-2021)

- https://diana-hep.org/
- Contributions to REANA, RECAST, launched pyhf likelihood publishing, Active Learning for reinterpretation
- Supported GitHub -> Zenodo DOI minting

IRIS-HEP (2018-?)

- https://iris-hep.org/
- Major contributions to likelihood publishing, HEPData integration,

SCAILFIN (2018-2021)

- https://scailfin.github.io/
- Contributions to REANA (Slurm and HPC backends, applications built on top of REANA, etc.), Active Learning for reinterpretation

FAIROS-HEP (2022-2025)

- https://fairos-hep.org/ (under construction)
- **Continue the legacy of contributions, help coordinate the ecosystem**

What is FAIROS-HEP?

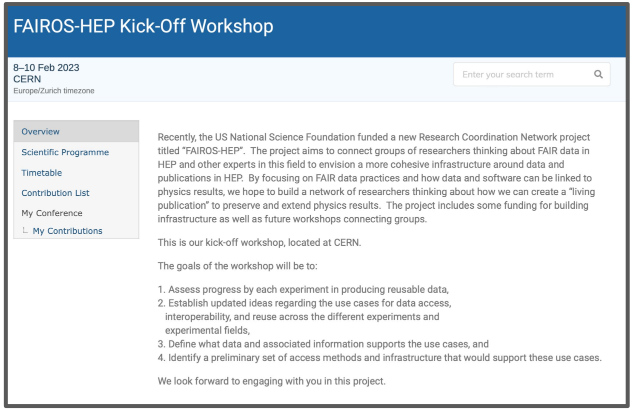
The FAIROS-HEP project aims to connect groups of researchers thinking about FAIR data in HEP and other experts in this field to envision a more cohesive infrastructure around data and publications in HEP.

- By focusing on FAIR data practices and **how data and software can be linked to physics results**, we hope to **build a network of researchers** thinking about how we can create a "**living publication**" to preserve and extend physics results.
- The project includes **some funding for building infrastructure** as well as **future workshops connecting groups**.

Kickoff workshop

The goals of the workshop were to:

1. Assess progress by each experiment in producing reusable data,
2. Establish updated ideas regarding the use cases for data access, interoperability, and reuse across the different experiments and experimental fields,
3. Define what data and associated information supports the use cases, and
4. Identify a preliminary set of access methods and infrastructure that would support these use cases.

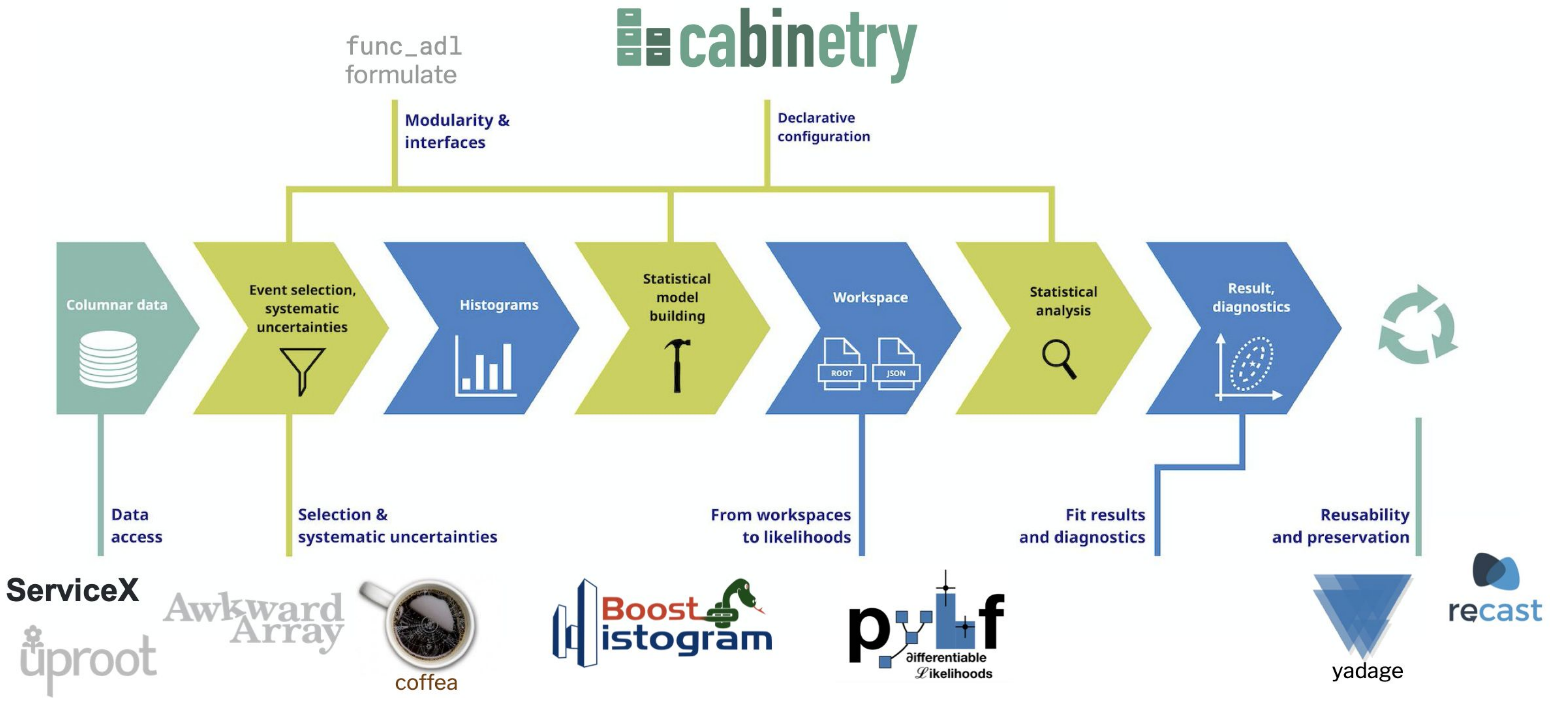


https://indico.cern.ch/event/1234612/

Technical Recommendations: Provide initial direction for which elements of the cyber ecosystem will be most relevant for the first round of technical improvements. Initiate investigations.

IRIS-HEP

The IRIS-HEP Analysis Systems focus area extends to reuse and preservation, but it has not been a major area of activity. FAIROS-HEP will coordinate closely.



Some Recent Developments in HEP

Open Science & Beyond

The field is at a tipping point. CERN has publicly embraced Open Science and the experiments are adopting new policies.

But we also realize Open Data is not the end of the story.
Reuse is key!

nature physics

PERSPECTIVE

https://doi.org/10.1038/s41567-018-0342-2

Corrected: Publisher Correction

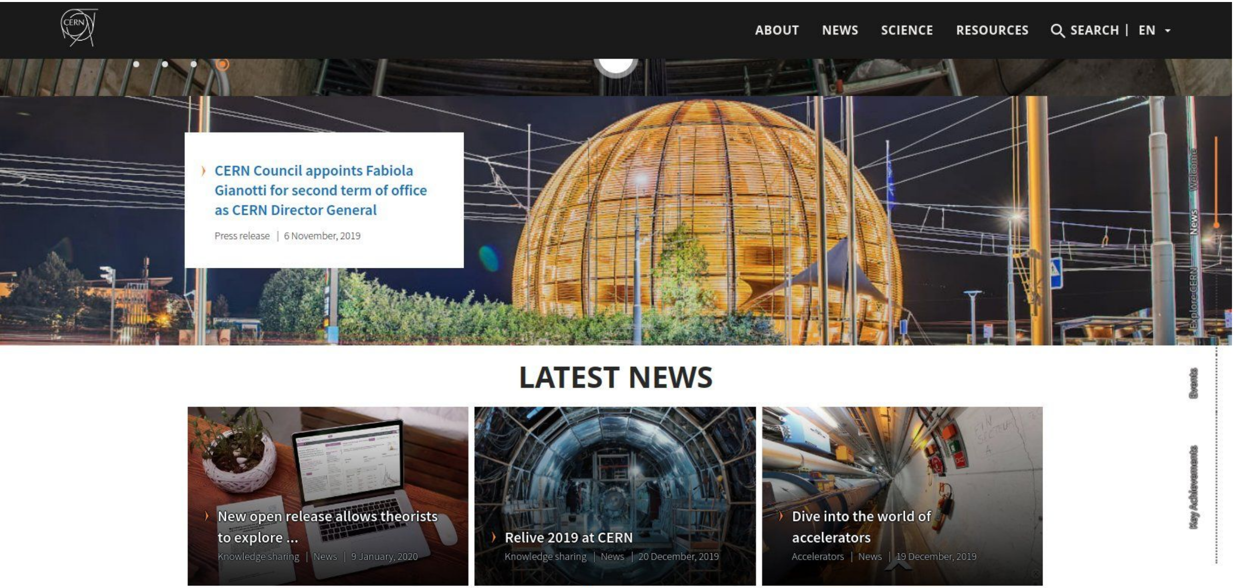
OPEN

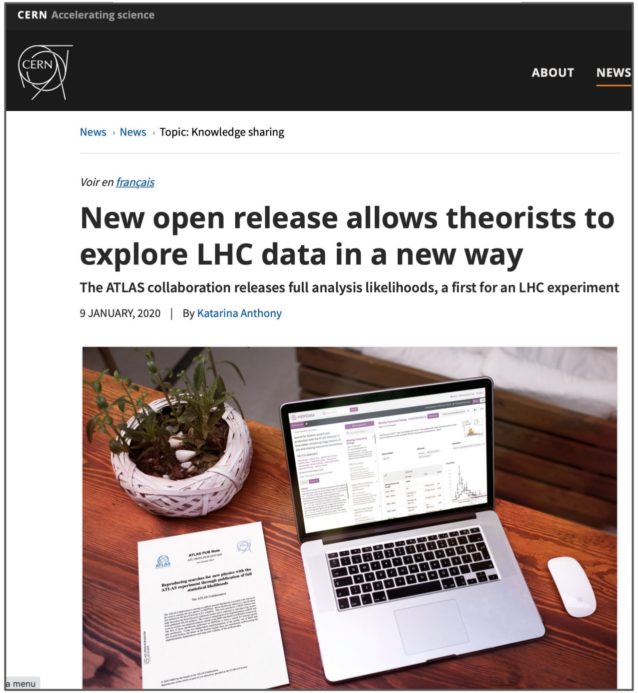
Open is not enough

Xiaoli Chen[1,2], Sünje Dallmeier-Tiessen[1*], Robin Dasler[1,11], Sebastian Feger[1,3], Pamfilos Fokianos[1], Jose Benito Gonzalez[1], Harri Hirvonsalo[1,4,12], Dinos Kousidis[1], Artemis Lavasa[1], Salvatore Mele[1], Diego Rodriguez Rodriguez[1], Tibor Šimko[1*], Tim Smith[1], Ana Trisovic[1,5*], Anna Trzcinska[1], Ioannis Tsanaktsidis[1], Markus Zimmermann[1], Kyle Cranmer[6], Lukas Heinrich[6], Gordon Watts[7], Michael Hildreth[8], Lara Lloret Iglesias[9], Kati Lassila-Perini[4] and Sebastian Neubert[10]

The solutions adopted by the high-energy physics community to foster reproducible research are examples of best practices that could be embraced more widely. This first experience suggests that reproducibility requires going beyond openness.

Highlight: 2020 CERN homepage





Featured on CERN homepage

In the press

CERN Accelerating science

ABOUT NEWS

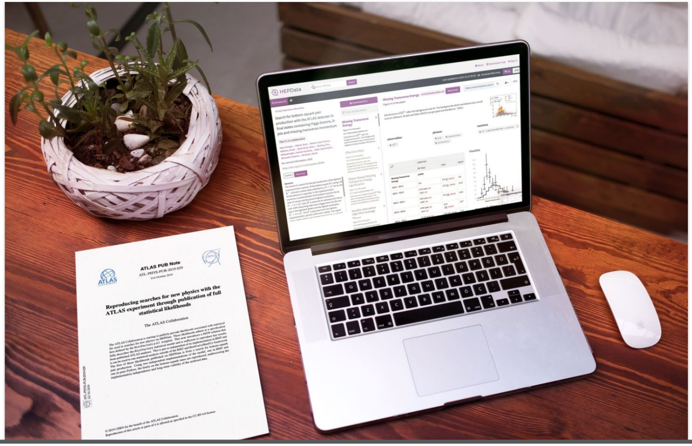
News › News › Topic: Knowledge sharing

Voir en français

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

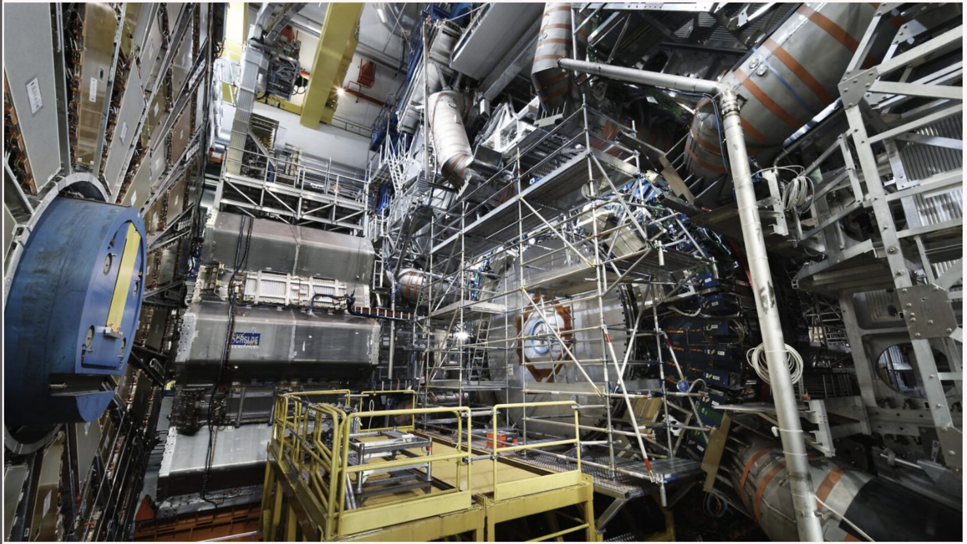
9 JANUARY, 2020 | By Katarina Anthony



a menu

https://www.symmetrymagazine.org/article/atlas-releases-full-orchestra-of-analysis-instruments

symmetry topics follow +



Courtesy of CERN

ATLAS releases 'full orchestra' of analysis instruments

01/14/21 | By Stephanie Melchor

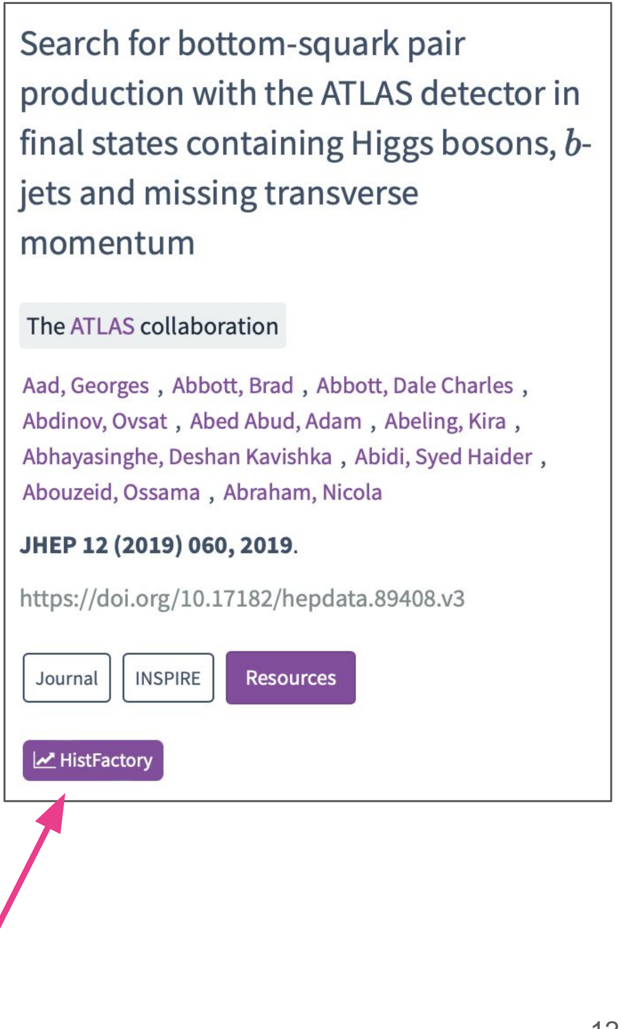
The ATLAS collaboration has begun to publish likelihood functions, information that will allow researchers to better understand and use their experiment's data in future analyses.

Published Probability Models

Published Probability Models

Updated list of HEPData entries for publications using `HistFactory` JSON statistical models:

- Search for charginos and neutralinos in final states with two boosted hadronically decaying bosons and missing transverse momentum in pp collisions at s√=13 TeV with the ATLAS detector. 2021. doi:10.17182/hepdata.104458
- Measurement of the tttt production cross section in pp collisions at s√=13 TeV with the ATLAS detector. 2021. doi:10.17182/hepdata.105039
- Search for R-parity violating supersymmetry in a final state containing leptons and many jets with the ATLAS experiment using s√=13 TeV proton-proton collision data. 2021. doi:10.17182/hepdata.104860
- Search for chargino−neutralino pair production in final states with three leptons and missing transverse momentum in s√=13 TeV pp collisions with the ATLAS detector. 2021. doi:10.17182/hepdata.95751
- Measurements of the inclusive and differential production cross sections of a top-quark-antiquark pair in association with a Z boson at s√=13 TeV with the ATLAS detector. 2021. doi:10.17182/hepdata.100351
- Search for pair production of third-generation scalar leptoquarks decaying into a top quark and a τ-lepton in pp collisions at s√=13 TeV with the ATLAS detector. 2021. doi:10.17182/hepdata.100174.
- Search for squarks and gluinos in final states with one isolated lepton, jets, and missing transverse momentum at s√=13 TeV with the ATLAS detector. 2021. doi:10.17182/hepdata.97041
- Search for trilepton resonances from chargino and neutralino pair production in s√=13 TeV pp collisions with the ATLAS detector. 2020. doi:10.17182/hepdata.99806.
- Search for displaced leptons in s√=13 TeV pp collisions with the ATLAS detector. 2020. doi:10.17182/hepdata.98796.
- Search for squarks and gluinos in final states with jets and missing transverse momentum using 139 fb−1 of s√=13 TeV pp collision data with the ATLAS detector. 2020. doi:10.17182/hepdata.95664.
- Measurement of the tt production cross-section in the lepton+jets channel at s√=13 TeV with the ATLAS experiment. 2020. doi:10.17182/hepdata.95748.
- Search for long-lived, massive particles in events with a displaced vertex and a muon with large impact parameter in pp collisions at s√=13 TeV with the ATLAS detector. 2020. doi:10.17182/hepdata.91760
- Search for chargino-neutralino production with mass splittings near the electroweak scale in three-lepton final states in s√ = 13 TeV pp collisions with the ATLAS detector. 2019. doi:10.17182/hepdata.91127.
- Searches for electroweak production of supersymmetric particles with compressed mass spectra in s√=13 TeV pp collisions with the ATLAS detector. 2019. doi:10.17182/hepdata.91374
- Search for direct stau production in events with two hadronic τ-leptons in s√=13 TeV pp collisions with the ATLAS detector. 2019. doi:10.17182/hepdata.92006.
- 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 s√=13 TeV with the ATLAS detector. 2019. doi:10.17182/hepdata.90607.
- Search for squarks and gluinos in final states with same-sign leptons and jets using 139 fb−1 of data collected with the ATLAS detector. 2019. doi:10.17182/hepdata.91214.
- Search for bottom-squark pair production with the ATLAS detector in final states containing Higgs bosons, b-jets and missing transverse momentum. 2019. doi:10.17182/hepdata.89408.



Reuse in <30 lines of code

```python
import json
import cabinetry
import pyhf
from cabinetry.model_utils import prediction
from pyhf.contrib.utils import download

# download the ATLAS bottom-squarks analysis probability models from HEPData
download("https://www.hepdata.net/record/resource/1935437?view=true", "bottom-squarks")

# construct a workspace from a background-only model and a signal hypothesis
bkg_only_workspace = pyhf.Workspace(json.load(open("bottom-squarks/RegionC/BkgOnly.json")))
patchset = pyhf.PatchSet(json.load(open("bottom-squarks/RegionC/patchset.json")))
workspace = patchset.apply(bkg_only_workspace, "sbottom_600_280_150")

# construct the probability model and observations
model, data = cabinetry.model_utils.model_and_data(workspace)

# produce visualizations of the pre-fit model and observed data
prefit_model = prediction(model)
cabinetry.visualize.data_mc(prefit_model, data)

# fit the model to the observed data
fit_results = cabinetry.fit.fit(model, data)

# produce visualizations of the post-fit model and observed data
postfit_model = prediction(model, fit_results=fit_results)
cabinetry.visualize.data_mc(postfit_model, data)
```

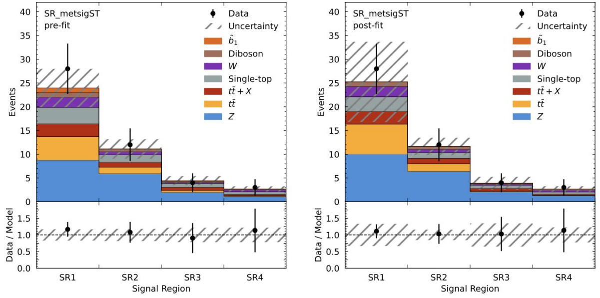
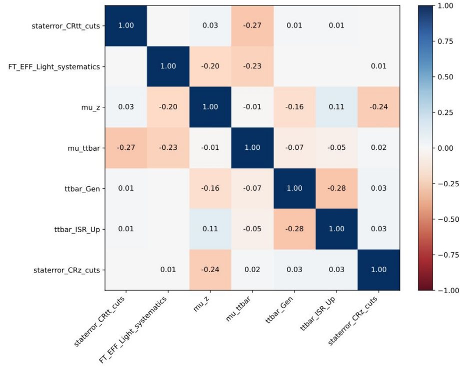


Figure 3: Pre-fit (left) and post-fit (right) visualizations of a selected signal hypothesis for four signal regions of the ATLAS search [41] of a bottom-squark of mass 600 GeV with a second-lightest neutralino of mass 280 GeV and lightest supersymmetric particle of mass 150 GeV generated from the full statistical models published in Ref. [20] using code from Ref. [40].



A SModelS interface for pyhf likelihoods #6

Gaël Alguero (LPSC, Grenoble), Sabine Kraml (LPSC, Grenoble), Wolfgang Waltenberger (Vienna, OAW and Vienna U.) (Sep 3, 2020)

Published in: *Comput.Phys.Commun.* 264 (2021) 107909 • e-Print: 2009.01809 [hep-ph]

pdf DOI cite claim reference search 20 citations

Community Contributions

Reinterpretation and Long-Term Preservation of Data and Code
#1

Stephen Bailey (LBL, Berkeley), K.S. Cranmer (Wisconsin U., Madison), Matthew Feickert (Wisconsin U., Madison), Rob Fine (Los Alamos), Sabine Kraml (LPSC, Grenoble) et al. (Sep 16, 2022)

Contribution to: 2022 Snowmass Summer Study • e-Print: 2209.08054 [physics.comp-ph]

pdf | cite | claim | reference search | 2 citations

Data and Analysis Preservation, Recasting, and Reinterpretation
#2

Stephen Bailey (LBL, Berkeley), Christian Bierlich (Lund U. (main)), Andy Buckley (Glasgow U.), Jon Butterworth (University Coll. London), Kyle Cranmer (New York U.) et al. (Mar 18, 2022)

Contribution to: 2022 Snowmass Summer Study • e-Print: 2203.10057 [hep-ph]

pdf | cite | claim | reference search | 4 citations

Signal region combination with full and simplified likelihoods in MadAnalysis 5
#2

Gaël Alguero (LPSC, Grenoble and Annecy, LAPTH), Jack Y. Araz (Durham U., IPPP), Benjamin Fuks (Paris, LPTHE), Sabine Kraml (LPSC, Grenoble) (Jun 29, 2022)

Published in: *SciPost Phys.* 14 (2023) 009 • e-Print: 2206.14870 [hep-ph]

pdf | DOI | cite | claim | reference search | 1 citation

Publishing statistical models: Getting the most out of particle physics experiments
#4

Kyle Cranmer (New York U.), Sabine Kraml (LPSC, Grenoble), Harrison B. Prosper (Florida State U.), Philip Bechtle (Bonn U.), Florian U. Bernlochner (Bonn U.) et al. (Sep 10, 2021)

Published in: *SciPost Phys.* 12 (2022) 1, 037, *SciPost Phys.* 12 (2022) 037 • e-Print: 2109.04981 [hep-ph]

pdf | DOI | cite | claim | reference search | 23 citations

Reinterpretation of LHC Results for New Physics: Status and Recommendations after Run 2
#8

LHC Reinterpretation Forum Collaboration • Waleed Abdallah (Harish-Chandra Res. Inst. and Cairo U.) et al. (Mar 19, 2020)

Published in: *SciPost Phys.* 9 (2020) 2, 022 • e-Print: 2003.07868 [hep-ph]

pdf | links | DOI | cite | claim | reference search | 57 citations

SciPost Physics — Submission

Publishing statistical models: Getting the most out of particle physics experiments

Kyle Cranmer[1*], Sabine Kraml[2‡], Harrison B. Prosper[3§] (editors), Philip Bechtle[4], Florian U. Bernlochner[4], Itay M. Bloch[5], Enzo Canonero[6], Marcin Chrzaszcz[7], Andrea Coccaro[8], Jan Conrad[9], Glen Cowan[10], Matthew Feickert[11], Nahuel Ferreiro Iachellini[12,13], Andrew Fowlie[14], Lukas Heinrich[15], Alexander Held[1], Thomas Kuhr[13,16], Anders Kvellestad[17], Maeve Madigan[18], Farvah Mahmoudi[15,19], Knut Dundas Morå[20], Mark S. Neubauer[11], Maurizio Pierini[15], Juan Rojo[8], Sezen Sekmen[22], Luca Silvestrini[23], Veronica Sanz[24,25], Giordon Stark[26], Riccardo Torre[8], Robert Thorne[27], Wolfgang Waltenberger[28], Nicholas Wardle[29], Jonas Wittbrodt[30]

1 New York University, USA **2** LPSC Grenoble, France **3** Florida State University, USA **4** University of Bonn, Germany **5** School of Physics and Astronomy, Tel-Aviv University, Israel **6** University of Genova, Italy **7** Institute of Nuclear Physics, Polish Academy of Sciences, Krakow, Poland **8** INFN, Sezione di Genova, Italy **9** Oskar Klein Centre, Stockholm University, Sweden **10** Royal Holloway, University of London, UK **11** University of Illinois at Urbana-Champaign, USA **12** Max Planck Institute for Physics, Munich, Germany **13** Exzellenzcluster ORIGINS, Garching, Germany **14** Nanjing Normal University, Nanjing, PRC **15** CERN, Switzerland **16** Ludwig-Maximilians-Universität München, Germany **17** University of Oslo, Norway **18** DAMTP, University of Cambridge, UK **19** Lyon University, France **20** Columbia University 10027, USA **21** VU Amsterdam and Nikhef, The Netherlands **22** Kyungpook National University, Daegu, Korea **23** INFN, Sezione di Roma, Italy **24** University of Sussex, UK **25** IFIC, Universidad de Valencia-CSIC, Spain **26** SCIPP, UC Santa Cruz, CA, USA **27** University College London, UK **28** HEPHY and University of Vienna, Austria **29** Imperial College London, UK **30** Lund University, Sweden

* kyle.cranmer@nyu.edu, ‡ sabine.kraml@lpsc.in2p3.fr, § harry@hep.fsu.edu
(mailing list: open-likelihoods@cern.ch)

September 9, 2021

Abstract

The statistical models used to derive the results of experimental analyses are of incredible scientific value and are essential information for analysis preservation and reuse. In this paper, we make the scientific case for systematically publishing the full statistical models and discuss the technical developments that make this practical. By means of a variety of physics cases — including parton distribution functions, Higgs boson measurements, effective field theory interpretations, direct searches for new physics, heavy flavor physics, direct dark matter detection, world averages, and beyond the Standard Model global fits — we illustrate how detailed information on the statistical modelling can enhance the short- and long-term impact of experimental results.

1

Preservation & Reinterpretation

First results using the RECAST reinterpretation framework and publishing full statistical likelihoods (using pyhf)

recast

reana

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

The ATLAS Collaboration is starting to publicly provide likelihoods associated with statistical fits used in searches for new physics on HEPData. These likelihoods adhere to a specification first defined by the `HistFactory` p.d.f. template. This note introduces a JSON schema that fully describes the `HistFactory` statistical model and is sufficient to reproduce key results from published ATLAS analyses. This is per-se independent of its implementation in `ROOT` and it can be used to run statistical analysis outside of the `ROOT` and `RooStats/RooFit` framework. The first of these likelihoods published on HEPData is from a search for bottom-squark pair production. Using two independent implementations of the model, one in `ROOT` and one in pure Python, the limits on the bottom-squark mass are reproduced, underscoring the implementation independence and long-term viability of the archived data.

ATL-PHYS-PUB-2019-029
05 August 2019

ATLAS PUB Note

ATL-PHYS-PUB-2019-032

11th August 2019

RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two b-quarks

The ATLAS Collaboration

The reinterpretation of a search for dark matter produced in association with a Higgs boson decaying to b-quarks performed with RECAST, a software framework designed to facilitate the reinterpretation of existing searches for new physics, is presented. Reinterpretation using RECAST is enabled through the sustainable preservation of the original data analysis as re-executable declarative workflows using modern cloud technologies and integrated with the wider CERN Analysis Preservation efforts. The reinterpretation targets a model predicting dark matter production in association with a hypothetical dark Higgs boson decaying into b-quarks where the mass of the dark Higgs boson m_s is a free parameter, necessitating a faithful reinterpretation of the analysis. The dataset has an integrated luminosity of 79.8 fb^{-1} and was recorded with the ATLAS detector at the Large Hadron Collider at a centre-of-mass energy of $\sqrt{s} = 13$ TeV. Constraints on the parameter space of the dark Higgs model for a fixed choice of dark matter mass $m_\chi = 200$ GeV exclude model configurations with a mediator mass up to 3.2 TeV.

ATL-PHYS-PUB-2019-032
12 August 2019

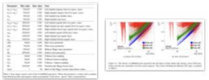
Scalable ATLAS pMSSM computational workflows using containerised REANA reusable analysis platform

M. Donadoni[1], M. Feickert[2], L. Heinrich[3], Y. Liu[4], A. Mečionis[1], V. Moisieienkov[1], G. Stark[5], T. Šimko[1], M. Vidal García[1]

[1] CERN, Geneva, Switzerland
[2] University of Madison, Wisconsin, United States
[3] Max-Planck-Institut für Physik, München, Germany
[4] Sun Yat-sen University, Guangzhou, China
[5] University of California, Santa Cruz, United States

ATLAS pMSSM searches

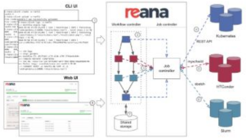
We have developed a streamlined framework for large-scale pMSSM reinterpretations of ATLAS analyses of LHC Run 2 using containerised computational workflows. The project is looking to assess the global coverage of BSM physics and requires running numerous computational workflows representing pMSSM model points.



ATLAS pMSSM studies from LHC Run 1. arXiv:1508.06608v2
K. Cranmer, I. Yavin, RECAST arxiv:1010.2506

REANA reusable analysis platform

The computational workflows were run at scale using the REANA reusable analysis platform. The workflows typically run on Kubernetes clusters.



The architecture of the REANA cloud platform. Users can use a command-line client and a web interface to submit containerised workflows that are then orchestrated on supported compute backends.

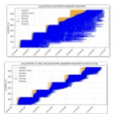
Selection of analyses

Following ATLAS analysis preservation policies, many ATLAS analyses have been preserved as containerised Yadage workflows. After validation they are added to a curated selection of analyses suitable for the pMSSM study.

ATLAS SUSY analyses preserved on GitLab.

Scaling out to O(5k) workflows

We have improved the REANA platform scheduling in order to maximise the scheduling throughput of incoming workflows.

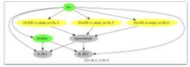


A scalability test submitting 200 workflows every 10 minutes. A cluster with 448 cores (top) cannot keep up with the load. A cluster with 1072 cores (bottom) can comfortably hold the incoming workload.

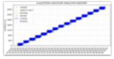
The benchmark tests were running in the CERN Computer Centre and on the Google Cloud Platform public cloud. The REANA scheduling parameters were optimised to maximise CPU utilisation for the pMSSM workloads taking into account the three time-consuming ntupling jobs per one workflow run.

Computational workflow

One typical containerised workflow consists of three time-consuming ntupling steps that run in parallel and the fitting steps that run afterwards. The workflow dependency graph is simple; the complexity relies in having to run O(5k) of these workflows in order to cover sufficient number of pMSSM model points.



A typical pMSSM workflow. The computational runtime is about 10 minutes without systematics (test payload) and about 10 hours with all systematics (real payload).



The workload burndown throughput rate is sustainable over a long period of time.

This project was carried out in a close collaboration between ATLAS and CERN IT.

Conclusions

- ATLAS searches for supersymmetry are effectively preserved with computational workflow recipes enabling their future reuse and reinterpretation.
- We have launched several ATLAS pMSSM workflows on REANA and studied the workflow burndown throughput rate as a function of increasing Kubernetes cluster size.
- REANA platform has been extended to support the workload of many concurrent workflows. The solution was benchmarked on medium to large clusters (from 500 to 5000 cores).
- It is essential to adapt cluster parameters to the type of workloads in order to ensure best throughput and cluster resource utilisation (CPU per node, RAM per node).
- The developed system is ready to run large-scale ATLAS pMSSM reinterpretations of LHC Run 2 analyses.

reana
www.reana.io

Docs
docs.reana.io

Email
info@reana.io

GitHub
@reanahub

Twitter
@reanahub

FAIROS-HEP Developed Infrastructure

http://hepexplorer.net

FAIROS-HEP: Future Workshop Trajectory

Workshop 2: Broader Community Engagement and Theory Reinterpretation

Attendees: Particle Physics Experimentalists and Theorists

Location: TBD

Workshop 3: Specific Reuse Case: Deriving EFT Results from Future LHC Data

Attendees: Participants from the LHC experiments and theorists working on EFT interpretations

Location: US University, TBD

Workshop 4: Broader Engagement: CNI and External Science Partners

Attendees: Core RCN members, CNI membership, External Science Partners

Location: Semi-Annual CNI Membership Meeting

Workshop 5: Reuse Case: Kinematic RECASTing for New Physics Discovery

Attendees: Participants from the LHC experiments and theorists working on new physics searches

Location: US University

Thank you!

Questions?

The NSF's FAIROS Research Coordination Networks

We are one of 10 funded projects:

DBER+ Commons
- FAIR in Education research

MaRCN
- Open science in materials

FAIR facilities and instruments
- PIDs for research instruments

FARR
- Best practices for ML/AI

Geospatial Big Data Infrastructure
- Environmental research

Paleobio/zooarchaeology databases
- Community-coordinated resources

SEEKCommons project
- Bridge social and environ. sciences

NoCTURN
- Non-clinical tomography

REPETO
- Reproducibility in CS Education