Title

Data and Analysis Preservation, Recasting, and Reinterpretation

Permalink

https://escholarship.org/uc/item/6gv3q31n

Authors

Bailey, Stephen
Bierlich, Christian
Buckley, Andy
et al.

Publication Date

2022-03-18

March 21, 2022

Submitted to the Proceedings of the US Community Study
on the Future of Particle Physics (Snowmass 2021)

Data and Analysis Preservation, Recasting, and Reinterpretation

TF07 (Collider Phenomenology in the Theory Frontier)
COMPF7 (Reinterpretation and long-term preservation of data and code)

Stephen Bailey[1], Christian Bierlich[2], Andy Buckley[3], Jon Butterworth[4], Kyle Cranmer[5], Matthew Feickert[6*], Lukas Heinrich[7], Axel Huebl[1], Sabine Kraml[8‡], Anders Kvellestad[9], Clemens Lange[10], Andre Lessa[11], Kati Lassila-Perini[12], Christine Nattrass[13], Mark S. Neubauer[6], Sezen Sekmen[14], Giordon Stark[15], Graeme Watt[16]

1 Lawrence Berkeley National Laboratory, USA **2** Lund University, Lund, Sweden **3** University of Glasgow, UK **4** University College London, UK **5** New York University, USA **6** University of Illinois at Urbana-Champaign, USA **7** Technische Universität München, Germany **8** Univ. Grenoble Alpes, CNRS, Grenoble INP, LPSC-IN2P3, Grenoble, France **9** University of Oslo, Norway **10** Paul Scherrer Institute, Villigen, Switzerland **11** Universidade Federal do ABC, Brazil **12** Helsinki Institute of Physics, Finland **13** University of Tennessee, Knoxville, USA **14** Kyungpook National University, Korea **15** SCIPP, UC Santa Cruz, CA, USA **16** IPPP, Durham University, UK

Corresponding authors:
* matthew.feickert@cern.ch, ‡ sabine.kraml@lpsc.in2p3.fr

Abstract

We make the case for the systematic, reliable preservation of event-wise data, derived data products, and executable analysis code. This preservation enables the analyses' long-term future reuse, in order to maximise the scientific impact of publicly funded particle-physics experiments. We cover the needs of both the experimental and theoretical particle physics communities, and outline the goals and benefits that are uniquely enabled by analysis recasting and reinterpretation. We also discuss technical challenges and infrastructure needs, as well as sociological challenges and changes, and give summary recommendations to the particle-physics community.

Contents

Executive summary. To achieve their full scientific impact, HEP experiments need to integrate extensive data and analysis preservation efforts into their publication processes, alongside the communication of results in reusable form and preservation of data products, and making event-level data publicly available. Without this, the influence of the hundreds of published analyses from the LHC, HL-LHC, EIC, and other future experiments will be limited mainly to the physics ideas in vogue at the time the collaboration collected their data. The public investment in experimental programs underscores the importance of going beyond the original paper publication and ensuring that analyses continue providing scientific value in perpetuity.

1 Introduction

The scientific value of the output from analyses performed at experiments in High Energy Physics (HEP) is immense. To maximise the scientific return of the data obtained and the analyses performed at these unique machines, it is imperative that there are strategic community plans in practice for the reuse and reinterpretation of the analyses and all of their associated data products. Enabling reuse and the ability to explore not-yet-thought-of-theories in the long-term future (i.e., on the 10 to 50-year time scale) needs to become a community priority. The benefits extend beyond improving and extending searches for new physics, and also allow models, including new precision Standard Model calculations, to be tested in new physics regimes. This will require reuse and preservation to become parts of the planning process for future analyses, and additionally be incorporated into the operations and facilities planning for future experiments, as well as for the phenomenology community, who will need to be major shareholders in this work.

These changes, and the technical infrastructure necessary for long-term storage and improved preservation standards, are not inconsequential or easy [1, 2]. They will require development of standards, software, cyberinfrastructure, stewardship roles, and additional support from funding agencies. Changes at fundamental levels to the community's approach to analysis will necessarily require additional sociological changes. The particle physics community is not homogeneous in its experiences and practices with preservation and reinterpretation, so it is to be expected that different subfields will have different adoption times of recommendations and experience different levels of sociological challenges. There are steps that can be taken to smooth this process in the field. Current procedures in national laboratories and universities for publishing data and software — e.g., by asserting copyright — are non-uniform and often require significant overhead of individual scientists. Reducing such administrative entry burdens to contribute to repositories by using fast and streamlined procedures, could significantly increase contributions to data and software repositories.

In this paper, we distinguish between preserving collision and simulated event data used as input to physics analyses, described in Section 2, and preserving workflows and data products connected to specific analyses, covered in Sections 3 and 4, respectively. Section 5 addresses the usage of the available material for reinterpretation in terms of new theoretical ideas and global analyses. To help with clarity of discussion, we summarise key terms used in the paper in Figure 1. This paper focuses primarily on the Large Hadron Collider (LHC), but its recommendations apply to particle physics experiments in general, and beyond physics analyses.

Glossary of terms

1.1: **Data**: In the context of this paper, this refers explicitly to data recorded from experiments that have been passed through the experiment's event reconstruction.

1.2: **Derived data**: Data that have passed through a size reduction step that prunes information, but which might also add additional calculated quantities to the data files.

1.3: **Data products**: All files containing selections of derived data and synthesised information from the various stages of an analysis. These might include summary plots and tables, histograms of kinematic distributions, fiducial cross sections, cross section limits, simplified model results, correlation information, analysis statistical workspace binaries or full statistical models, etc. [2]

1.4: **Data preservation**: The procedures, practices, and standards of ensuring the long-term (i.e., decades beyond the end of an experiment) preservation, accessibility, and usability of data and derived data from experiments.

1.5: **Analysis preservation**: The procedures, practices, and standards of ensuring the long-term preservation, accessibility, and usability of information necessary to repeat an experimental analysis (starting from its associated preserved data) and generate all associated data products. As discussed in Section 3, there are multiple levels of analysis preservation fidelity with distinct advantages and trade offs.

1.6: **Reinterpretation**: Any type of new, alternative or updated interpretation of an experimental analysis or result, including the combination in, e.g., global fits or global averages.

1.7: **Recasting**: Reproducing the analysis logic in a simulation, considering a different physical process with a different phase space distribution, which might have different efficiencies and acceptances than the originally hypothesised model.

Figure 1: Definitions of a few key terms used in this paper.

2 Data preservation (open data)

Following the positive experience and feedback from the open releases of recorded and simulated event-level datasets by the CMS experiment since 2014, European Organization for Nuclear Research (CERN) formulated an open data policy for the LHC experiments [3] in 2020. As the host laboratory, CERN maintains and develops the infrastructure needed for the portal, provides storage resources, and takes the custodial responsibility of released open data in the long term. All LHC experiments are committed to releasing research-quality data through the CERN Open Data portal [4]. The amount of data and release timeline varies from one experiment to another [3, 5–9].

CMS is the only experiment having released research-quality data at the time of writing. The CMS Open Data releases contain complete reprocessing of collision data from each data-taking period and the simulated data corresponding to these data. They are made available in the format and with the exact data quality requirements used by analyses of the CMS collaboration. The volume of data, both actual and Monte Carlo simulation, amounts currently to 2.8 PB. CMS also provides a compatible version of the CMSSW and additional information necessary to perform a research-level physics analysis on the public data. The documentation comes with example code and specific guide pages to instruct new users.

Data released through the CERN Open Data portal satisfy Findable, Accessible, Interoperable, and Reusable (FAIR) principles for scientific data management [10] for data and metadata to a large extent. Due to the complexity of experimental particle physics data, the FAIR principles alone do not guarantee the reusability of these data, and additional effort is needed to pass on the knowledge needed to use and interpret them correctly. In large experiments, collaboration members benefit from the existing knowledge infrastructure (meetings, discussion forums, mentoring, specific groups responsible for providing data assets needed for physics analysis, and more) that is unavailable to external users or long-term.

Early testing is of utmost importance to ensure that the open data are usable for physics research and that all necessary additional information is captured, stored, and provided. Therefore, CMS emphasises releasing data relatively early after the data-taking. External users' feedback and questions indicate the eventual missing data assets and lack of information. The release can then be complemented while the expertise on these data is still present in the collaboration. To encourage the use of open data and to get direct feedback, the CMS Open Data group is organizing regular workshops with hands-on tutorials. It is essential to reach out to a diverse user community, from newcomers to seniors with different computing backgrounds and physics skills. While the CMS Open Data group is looking forward to organizing the next workshop in presence, the value of remote, virtual events with many participants who would not have been able to travel is acknowledged.

Setting up example analysis workflows is one of the priorities for making open data reusable. They should demonstrate all necessary steps from data access and selection to any eventual corrections and address the particularities of experimental data, such as estimating efficiencies and uncertainties. Activities for preserving analysis workflows internally in the collaboration will facilitate building such workflows. In use with open data, the software container technology is suitable for automating workflows. The experience gained with them can be used for current analyses within the collaboration.

There have been successful data-preservation efforts in the LEP and HERA experiments. However, no public access to these data exists in terms defined in the CERN data policy for the LHC experiments. Access to preserved data is made possible through different modalities, such as joining or working with collaboration members. The BABAR collaboration has decided to make its data available for future analyses through the CERN Open Data portal. Lack of person-power is often a bottleneck, and while valid for experiments in the data-taking phase, it is even more evident for experiments past data-taking. Open access initiatives have to start at an early stage; see also the reports at the recent DPHEP workshop [11].

Building a community of users for these data is paramount to making open data initiatives successful. The theory community is a critical player in this, and promoting open data despite the substantial work required in their analysis is a strong demonstration of their value and an invaluable validation effort for their usability for future generations.

Data Preservation Recommendations

2.1: Agree on data preservation and public event-data releases as a means to maximise scientific outcomes, and allocate resources and responsibilities to achieve this goal in the experiments' organization.

2.2: Give long-term custodial responsibility of public data to the host laboratory or an organization that persists beyond the experiment's lifetime and uses common distribution platforms with other experiments.

2.3: Incorporate preparing data for public releases and invest in preserving the knowledge needed for their use in the data processing and analysis operations and facilities.

2.4: Encourage and promote the use of open data to explore and improve usability and to ensure that all necessary information for research-level use is available.

Figure 2: Recommendations for data preservation.

3 Analysis preservation

Preservation of analysis logic and workflows in any specific form enables the reuse of the original analysis process and associated data products. Such reuse may, for instance, be the reinterpretation in terms of physics models not considered in the original analysis by the experimental collaboration; this may range from Beyond-the-Standard Model (BSM) theories, to new ideas and implementations of non-perturbative Quantum Chromodynamics (QCD), and everything in between.

As such, an experiment needs to integrate analysis preservation into its publication processes alongside open data and data-product preservation to achieve its full scientific impact. Without this, the influence of the hundreds of published analyses from the LHC, High-Luminosity LHC (HL-LHC), Electron-Ion Collider (EIC), and other modern collider experiments will be limited mainly to the physics ideas in vogue at the time the collaboration collected collider data. The public investment in experimental programs underscores the importance of going beyond the first publication and ensuring that analyses continue providing scientific value in perpetuity.

Different levels of event-wise analysis preservation are possible. They already exist, from full preservation of the experimental analysis workflow, to "fast" or "lightweight" emulations of the entire analysis that emphasise speed (and often greater public availability) to the expense of some precision. Both the accurate-but-expensive and fast-but-approximate approaches (and points in between) have value for physical applications. Given an ample model space to explore, combining many preserved event analyses to obtain a more statistically significant result, the emphasis lies more on speed and efficiency than total accuracy, so a rough measure of which parameter regions are viable. However, the emphasis naturally shifts to precision with a smaller range of models to explore, perhaps after such fast-interpretation triage. The expense of re-running full experimental software stacks becomes justifiable and tractable.

In the following sections, we review first the full and then the lightweight preservation paradigms, and finally the current status and issues of both, in both the high-energy *pp* and heavy-ion programs.

3.1 Full preservation

The most faithful approach to analysis preservation is to store, in a reproducible form, the exact software chain used to perform the analysis in the experiment. As each experiment has its internal data formats and software frameworks, injecting new models for analysis typically requires running the full post-generation software stack from detector simulation and reconstruction to the physics analysis code.

The obvious consequence is that simulation and reconstruction are computationally expensive, so is event reinterpretation via such analysis preservations. A secondary implication is for intellectual property: the ability to perform detailed simulation of an experiment's detector response is strategically sensitive information and not made explicitly usable outside the collaborations. If the code is hosted on publicly accessible platforms, this does not necessarily allow public use. Without explicit support and computing infrastructure, the complexity and resource requirements of these frameworks makes their use tractable only within the experimental collaborations.

Preserving full analysis chains is complicated by the diversity of analysis software frameworks, even within a given experiment. Versioned central production systems within one framework typically perform event generation, simulation, reconstruction, and data reduction ("derivation"). By contrast, most physics analyses within an experiment start from various tiers of derived data, and the lack of constraints has led to a proliferation of alternative analysis frameworks. These often lack extensive documentation, may be version controlled in different locations, and without coordination and standardization, naturally, evolve as many incompatible technical interfaces as there are packages.

Therefore, for reproducibility, experimental collaborations and host institutions must arrange to store analysis data in well-defined repositories and structures, with a plan for long-term archiving. The host institutions should monitor the correctness and continuing validity of the software by using continuous integration testing on a range of suitable data for each analysis. As a part of these tests, container images (e.g. Docker [12], Podman [13], or Singularity [14]) should be built and preserved on accessible container registries to capture the software environment needed for data processing.

3.2 Lightweight preservation

The approach described in Section 3.1 represents a faithful version of preservation, with implications for the original analysis software chain and typically high computing power requirements on any large-scale reuse of the preserved software stack. Even for organizations with the computational resources of active HEP experiments — where at present all such preservation has been performed — physicists cannot use this form of preservation to explore large model-parameter spaces, e.g. $\gg 2$ dimensions, via adaptive samplers.

These limitations motivate a more lightweight form of preservation, with more modest computing demands. It is simplest to make this alternative form independent of the original experiment software chain, with the desirable side-effect that it is also simpler to release publicly. The cost of being lightweight is that the approximations involved reduce the precision of the preservation; this makes lightweight preservation suitable for identifying model regions of interest rather than for making definitive statements of discovery. Public availability is essential for theorists' studies outside the experimental collaborations, such as testing new theoretical ideas against the data from several analyses and experiments in a global approach.

Several lightweight frameworks have arisen, ensuring not only preservation and reproducibility of experimental analyses but enabling reinterpretation studies for the whole HEP community [1]. The Rivet tool [15] is the clear choice for (differential) measurements, particularly where detector effects have been unfolded to a fiducial phase-space by the original experiment (and the statistical consequences fully reported, cf. Section 4). The LHC experiments have established Rivet-based measurement-preservation programs, and similar efforts are gaining traction at the Relativistic Heavy Ion Collider (RHIC), e.g. STAR and PHENIX programmes based around Rivet's heavy-ion features [16], and are being planned at the EIC. Such initiatives need to be built into the data publication and exploitation plans of ongoing and future experiments, to maximise the scientific impact of the analyses.

Analysis preservation at the reconstruction-level is more ambiguous, as some form of detector response needs to be encoded, and this will necessarily be less accurate than the original detector simulation+reconstruction code. Rivet incorporates a transfer-function approach for experiment-provided efficiencies and kinematic distortions to be applied to generator-level events but currently contains few reconstruction-level (usually BSM searches) analyses.

In parallel, driven by the need of open access to BSM recasting, the theory community has been developing various simulation-based reinterpretation frameworks for reconstruction-level analyses, in particular CheckMATE [17, 18], MadAnalysis5 [19, 20] and GAMBIT's ColliderBit [21]. They either also use a transfer-function approach, or rely on Delphes [22] for the emulation of detector effects; in some cases also generator-level events are used. A detailed overview of approaches and public frameworks is given in [1]. Experiments can also publish analyses in lightweight, executable form such as ATLAS' semi-public SimpleAnalysis system [23] (describing in particular, the analysis logic but not necessarily reconstruction efficiencies, and more.)

As in the cases of full analysis preservation and Rivet-based lightweight measurement preservation, active policy and support efforts will be required in experiments as part of the publication process to ensure the required level of preservation coverage. An example of good practice in this respect is the CMS jets+missing energy analysis [24], which provided a MadAnalysis5 recast code [25, 26] together with the paper, and in ATLAS and CMS' established integration of Rivet-analysis release into the publication process of suitable data-analyses.

All current tools support standard MC-generator tools such as the HepMC event record. However, essential features such as propagation of systematic uncertainties via weight vectors are not wholly or consistently implemented, and MC generators also have not yet standardised weight-coding conventions. Placing emphasis again on standard interfaces will help speed the convergence of the reconstruction-level BSM search tools in particular.

While public, lightweight preservation of cut-and-count analyses is increasingly becoming standard, the preservation of analyses that employ machine learning (ML) is still in its infancy. Indeed there are very few examples of analyses where communicating machine-learned models has been tried. In particular, the CMS all-hadronic search for supersymmetry [27] published a simplified version of their top-quark tagger [28], which is based on a Random Forest decision tree, and the ATLAS 0-lepton gluino/squark search [29, 30] published the boosted decision tree (BDT) weights of their event selection as `XML` files on HEPData [31].

Standards exist for a degree of ML interoperability, including the exchange of neural networks and BDTs, e.g. the Open Neural Network Exchange (ONNX) [32] format or direct preservation of decision trees as framework-independent code, but their long-term stability is unclear; the inclusion of implementation-specific behaviours mean that for long-term preservation either exact versions of frameworks (often Python-specific) need to be re-used, or functionality limited to a minimal subset. Trained networks have bee published in ONNX format by the ATLAS search for R-parity-violating supersymmetry [33, 34]; this is also included in the ATLAS SimpleAnalysis framework [23]. However, detailed documentation of, e.g. the input variables is missing, and it is unclear how to verify that physics objects from any given fast-simulation package will produce the intended ML responses within acceptable uncertainties.

Further development along these lines is highly needed. Successful sharing of a ML model requires not only sharing the model itself (including architectures, weights, and complete specification of software dependencies) but also the detailed specification of the input data. See also the corresponding discussion in Ref. [35] in this context.

Finally, a better meta-data stewardship for preserved analysis codes is in order, to make them searchable and findable (see also Section 5 in this context).

3.3 Analysis description languages

The full and lightweight preservation frameworks described above constitute a crucial step in analysis preservation, with the significant benefit of being executable. However, an equally important step is to preserve analysis logic in a more accessible and easily communicable, yet unambiguous format.[1] The thousands of analyses designed by experimental collaborations and phenomenologists have accumulated a tremendous source of physics content. This diverse content can inspire and inform new analysis ideas or train the new generation of particle physicists. Therefore, the third facet of analysis preservation is to encode an abstract description of the analysis logic in a *human- and machine-readable* declarative language.

Physics analyses typically consist of multiple, separately executed steps (e.g. event, selection and signal extraction). It is also important to capture the workflows for running and combining them over multiple event samples. There is currently no *de facto* standard for this workflow language across the experimental community: for example, Common Workflow Language [36], Snakemake [37], Yadage [38, 39], and Argo Workflows [40] are in use by different collaborations. Standardization and evolution around a smaller (or unique) subset of such languages would improve interoperability and knowledge transfer. In making such a choice, we stress the importance of a *declarative* rather than *imperative* model [41], as the former enables researchers to concentrate on the physics task with minimal need to consider technical details such as scalable job orchestration. Developers can significantly reduce the complexity of workflow descriptions by establishing standard interfaces for tools implementing processing steps so that they can be more easily composed.

As with declarative workflow description formats, a recent trend is to decouple analysis logic from the software frameworks used to execute it. Analysis frameworks internally used in ATLAS and CMS are moving towards accumulating all information related to object and event processing in a single location, e.g. in a configuration file, at least for part of the processing chain. An approach that takes this one step further is to use domain specific

[1] As should be the case in the paper publication, but rarely is given to the level of detail required for full reproduction.

languages (DSLs) dedicated to expressing the physics content of HEP analyses. DSLs can either rely on the syntax of an existing computer language (embedded DSL) or have a custom syntax (external DSL) more tailored to the semantics of the HEP-analysis context. Here also, a declarative approach has many benefits due to abstracting practical details into framework implementations. Several DSL approaches are studied, and various working prototypes exist.

On the embedded DSL side, the `F.A.S.T.` framework [42], used in CMS, LZ, and DUNE analyses, incorporates a `YAML` DSL. More recent Python based developments include `NAIL` [43], used in the CMS analysis that observed the first evidence for Higgs to two muon decays [44], and bamboo [45] which has been used in studies for Snowmass. Additionally, FuncADL [46] constructs hierarchical data queries using SQL-like concepts in Python and has been used to explore applications of functional programming.

On the external DSL side, the most advanced example is the so-called analysis description language (ADL) [47,48], see also [1]. It organizes the analysis description in multiple blocks separating object, variable and event selection definitions, and has keywords specifying analysis concepts and operations. ADL can be executed by any framework capable of parsing its syntax, notably the runtime interpreter CutLang [49]. ADL together with CutLang thus provide a functioning example of lightweight preservation in the sense of Section 3.2. The LHC analyses implemented by ADL are currently preserved in a GitHub repository [50]. ADL principles and prospects are discussed further in a dedicated Snowmass White Paper contribution [51].

One caveat to note is that HEP analyses are not always fully domain-specific and often need to execute custom logic in addition to predefined behaviors. This could be remedied either by extending the DSL to incorporate new known behaviors via an “escape” mechanism to inject general-purpose code (compromising its framework independence), or combining them with pre-and post-processing code to cover gaps in the DSL capabilities. ADL follows the former approach, and relies on an external database of injectable functions for each target framework.

The applications of ADL and ADL-type preservation are numerous. It may be used for analysis design, be directly included in publications, serve for pedagogical purposes and knowledge transfer, act as a configuration file that transpiles to code in lightweight analysis preservation frameworks, or as a source to transpile between full and lightweight preservation frameworks.

4 Preservation of data products

There is widespread consensus in the community that experiments should systematically provide all relevant derived data and data products in an open-access numerical form for future reuse [1]. Such data products are also called publication-related or “Level 1” data in the DPHEP categories. They include observed and expected event counts (including

Analysis Preservation Recommendations

3.1: Ensure use of interoperable systems to maximise the preservability and reusablility of experiment simulation and analysis software chains. This includes the use of version control, archival systems, containerisation, common software interfaces and data formats, and commitments from experimental collaborations and their host laboratories to maintain documentation and provide long-term support.

3.2: Ensure that all operational and in-preparation experiments have a planned and resourced programme for capture and long-term reproduction of their complete computational processing chain, including validation regression-tests.

3.3: Ensure that release of analysis preservation logic via public frameworks for the community to use is integrated with experiment publication and data-release processes, to maximise analysis impact. This also includes providing clear documentation and making all dependent frameworks available and documented for community consumption.

3.4: Support continuing development and uptake of new technologies for increasingly framework-independent analysis specifications, such as via declarative domain-specific analysis description languages.

Figure 3: Recommendations for analysis preservation.

error sources), efficiency functions, bin-to-bin correlations, profile likelihoods, full statistical models, signal efficiencies, simplified model results, and much more, as discussed extensively in Refs. [1,2].

With regards to publication infrastructure, HEPData [52] is the primary open-access repository for data products from particle physics experiments, with a long history going back to the 1970s. Funding is provided by the UK Science and Technology Facilities Council (STFC) to Durham University (UK) for staff to maintain the operation of the hepdata.net site, provide user support, and develop the open source software (available on the HEPData GitHub organisation) underlying the web application.

In the past, HEPData staff at Durham University handled data preparation in a standard format and uploaded it to the repository. However, now these tasks are delegated to the experimental collaborations. Data submitted to HEPData (as `YAML`) is primarily in a tabular form that can be interactively plotted in the web application and automatically converted to other standard formats (i.e. `CSV`, `JSON`, `ROOT`, `YODA`). The interactive nature of HEPData means that data tables must be kept sufficiently small (∼MB or less) that they can render in a web browser. In practice, tables with more than $\sim 10,000$ rows (for example, a covariance matrix for a measurement with ~ 100 bins) cannot efficiently render in a web browser. However, moderately large tables or non-tabular data files can be attached to a HEPData record as additional resources (in any format). The original files are downloadable, but the interactive nature is lost. HEPData imposes an overall size limit of 50 MB on the uploaded archive file to avoid problems caused by the attempted upload. Data products that are not suitable for HEPData, due to either being too large or predominantly in a non-tabular format, might be submitted to another data repository like Zenodo. Zenodo currently plugs a gap to host data (and software) that do not fit into other repositories.[2] A new HEP-specific instance of Zenodo could perhaps better serve the particle physics community in the future.

Reporting of results on HEPData has become a standard procedure in the LHC community, with ATLAS, CMS, ALICE, and LHCb all providing the results and data products from publications. This is also increasingly the case in the heavy-ion community with the STAR, PHENIX, and NA61/SHINE collaborations publishing their data products to HEPData as well in recent years. However, as use of HEPData is not a particle physics wide community norm yet, coverage remains incomplete[3] and will continue to until there are cultural shifts, as the heavy ion community has recently experienced.

Often, instead of being provided on HEPData, digitised results/plots are available only on collaboration web pages, without appropriate documentation, versioning, or other data stewardship. Sometimes, the linked ROOT files are wrong (not corresponding to the associated plot) or otherwise corrupted. In some cases, the digital material is missing altogether. Experience shows that missing or wrong resources can rarely be retrieved or corrected, as often the analysis team has disbanded with analyzers leaving the field, or the relevant files become lost. Part of the problem is missing time and community recognition to providing material on HEPData. The RAMP ("Reinterpretation: Auxiliary Material Presentation") seminar series aims at providing more visibility and recognition for such efforts, but more is needed for a sustainable change of culture.

As with analysis preservation, current heavy-ion experiments have less comprehensively established procedures for data-product preservation in HEPData than their LHC counterparts, but such procedures are now integrated into the publication processes for STAR and PHENIX, and decoupled from publication in sPHENIX. Coverage and process for BRAHMS and Phobos is less developed. STAR and PHENIX have both also instituted programmes of transcription of previously published data to HEPData, including as a form of experimental shift in 2020–22. Ensuring systematic and sufficiently detailed preservation of analysis products from all experiments in a central subject database such as HEPData is a central component of maximising scientific impact and data re-use, and should be designed into new experiment investment and deliverables from the start.

Regarding HEPData itself, material beyond digitised plots (like MC run cards, input files for benchmark points, or, most importantly, statistical models) are "additional resources", often lumped together in compressed archives without any standard structure. The types of data products being preserved has become much richer and diverse than flat tables. To

[2] This concerns also model files, Monte Carlo simulation, and data products from phenomenological studies.

[3] Even in ATLAS and CMS.

be able to provide the necessary infrastructure for all of these data products will require additional funding. There is room, and a clear need, for FAIR-ification of these precious material.

Data Product Preservation Recommendations

4.1: Make the provisioning of all data products associated with an experimental analysis a mandatory step for publication. Establishing appropriate person power, time, and community recognition is essential to that end.

4.2: Assure appropriate resources and funding for further development of the cyber-infrastructure, such as HEPData and other repositories like Zenodo, to preserve the data products and metadata, and extend the current data structure to include more rich data products and information beyond paper plots and flat tables, e.g., statistical models, in an individually searchable and citeable form.

Figure 4: Recommendations for data product preservation.

5 Reinterpretation and recasting

The physics impact of an experimental analysis can be increased well beyond its original purpose through reinterpretation. The various kinds of reinterpretation include:

updates of existing analyses using e.g. more precise theoretical calculations, improved experimental calibrations, or a different probability model;

parametric reinterpretation reparametrizing the likelihood through rescaling, without altering the efficiencies and acceptances that might modify the distributions — this is the approach taken for example when reusing simplified model results from the LHC, or in in the context of Higgs signal strengths;

kinematic reinterpretation considering a different physical process with a different phase space distribution, which might have different efficiencies and acceptances — this is what we generally refer to as recasting; a concrete example from these proceedings is the reuse of multi-boson and top-quark measurements to constrain new scalar states in composite Higgs models in [53];

combinations of analyses or datasets in model surveys,[4] global fits or global averages; this includes global Effective Field Theory (EFT) and BSM analyses as well as the reuse of datasets for the determination of parton distribution functions (discussed in more detail in [2]).

[4] Reinterpretation of several analyses within a given (usually BSM) scenario is relevant for displaying the complementary of distinct searches as well as identifying possible gaps in coverage. Such gaps can then be used as motivation for designing new experimental searches.

For this to be possible, the preservation of data products and analyses (usually in a lightweight format) are essential, as discussed in Sections 3 and 4. It is relevant to note here, that for many purposes of recasting,[5] analyses need only be preserved to the extent that the new signal yields can be determined and the subsequent statistical analysis using them can be performed. In this case, details of e.g. the derivation of background estimates need not be captured as they should already be preserved in the data products described in Section 4.

Reinterpretations are most often done by physicists (from both, the theory and the experiment sides) outside the experimental collaborations, but sometimes also collaboration-internally. In ATLAS, analysis preservation for reinterpretation has been more seriously pursued in recent years. A large number of analyses is preserved using Docker images and the yadage workflow language [38, 54]. This has led to the first successful uses of the RECAST [55] paradigm, in which existing analyses are reinterpreted at full fidelity within the collaboration [56–58]. Currently the selection of candidate reinterpretations is done mostly within the experiment. In the future, a public portal in which the wider HEP community has access to the catalog of RECAST-able analyses and can provide input for future full-fidelity reinterpretations will be desirable. Reinterpretations like such provide additional scientific value and should be published in peer-reviewed venues in the future. As such, they are also producers of data products in their own right. Products such as yields of the newly studied signal, statistical model fragments ("patches"), etc., should be submitted to archives such as HEPData or Zenodo as well.

Within CMS, reinterpretation or recasting is largely performed within ongoing analyses or via statistically combining analyses that explore complementary final states. While the analysis code is not systematically archived, the data products for signal extraction are usually preserved so that uncertainties across different analyses can be treated consistently in a statistical combination. We refer to Section 4 of [2] for more discussion of collaboration practices and use cases, including inter-experiment combinations, EFT fits, measurements in the flavour sector, etc..

Outside of the collaborations several software tools are being developed (CheckMATE [17, 18], MadAnalysis5 [19, 20], Contur/Rivet [59], SModelS [60, 61], and others as detailed in Section III of Ref. [1]) and are publicly available for the task of reinterpretation and/or recasting. They typically supply a database of implemented analyses, but also allow the user to implement new ones. In addition, they provide functionality for making statistical statements about the results, either by implementing the statistical models supplied by the collaboration (if available) or by taking some simplifying assumptions. We note that maintaining these tools and implementing new analyses requires considerable person power and funding.

The public tools provide distinct analyses coverage and sometimes different implementations of the same analysis. The proliferation of analyses and tools, and the lack of interoperability between the tools, can make the complete coverage of a physics case in reinterpretation studies seriously difficult. A unified format for analysis implementation, which could be

[5] Except, for instance, when signal-background interference effects play a role.

used interchangeably by the different tools, would significantly improve the reinterpretation potential of the phenomenology community. Even though this proposal has been discussed in the past [62], not much progress has been made in this direction. A possible interface format to make recast codes interchangeable between frameworks might be based on ADL [47,48] if parsers for the most common public frameworks (best including automatised validation) are developed.

Meanwhile, a few steps could be taken by the community with immediate benefits. One is a centralised (meta)database where the analyses available in the specific tools and the corresponding validation material can easily be found.[6] A searchable database covering all tools, including the major reinterpretation frameworks, and all analysis types, would clearly be of great benefit. In addition, whenever possible, it would be helpful to adopt a few validation guidelines to allow for a proper estimate of the recasting uncertainties introduced by each analysis implementation. As a second step, adopting basic standards for the input and output formats would also help the user to efficiently use distinct tools.

Another vital aspect of reinterpretation is the statistical treatment of the results. In recent years, the amount of information provided by experimental collaborations has increased significantly, allowing for a more robust statistical interpretation in phenomenological studies. However, to take full advantage of these new developments, it would be desirable to coordinate and unify the statistical output format and treatment within the specific recasting tools to have a common ground for comparison. Another critical aspect to be taken into account is the possibility for a global analysis of the results. Such global approaches are attempted by, e.g., the GAMBIT collaboration [64, 65] and the "protomodelling" project in [66]. They could, in principle, also combine results from different recasting tools. To help in this direction, the statistical analysis should be factorised as much as possible within each tool, allowing again for interoperability. In addition, standards and guidelines should be established for presenting the results and providing the required output for the statistical interpretation.

A further motivation for facilitating collider reinterpretations in global analyses is that it enables large-scale and adaptive exploration of the complementarities between collider results and other experimental results. Such global fits have the potential of both uncovering gaps in the experimental coverage, and identify which uncovered BSM scenarios are most plausible in light of other experiments, and thus constitute well-motivated targets for future analyses/experiments. Realising this potential, however, depends critically on computational efficiency, since a global fit faces the additional computational cost of reinterpreting all relevant non-collider results and explore a typically many-dimensional theory parameter space. As such, a focus on code efficiency, stability and parallelisability in reinterpretation tools, and on development of fast approximations for expensive computations (e.g. computation of higher-order cross-sections), is important to enable proper utilisation of experimental results.

Last but not least, it is important to ensure reproducibility and preservation of the results

[6] In this regard, the long-lived particle (LLP) recasting community created a centralised location for stand-alone LLP analysis preservation and validation material in form of a GitHub repository [63].

obtained. The same platforms used for data product preservation (see Section 4) can also be used by the theory community to preserve the reinterpretation results and data. In particular, Zenodo has already been used by a few groups to publish auxiliary material from phenomenological studies, see e.g. [67,68]. On the code side, is imperative for the purpose of reproducibility that the tools used for obtaining the results are properly documented and *versioned*. This policy should be largely encouraged within the theory community.

Reinterpretation and Recasting Recommendations

5.1: Encourage that reinterpretability and reuse be kept in mind early on in the analysis design. This concerns, for instance, the choice of input parameters in ML models, the full specification of the fiducial phase space of a measurement in terms of the final state, including any vetos applied, and generally the choice of non-overlapping regions and standard naming of shared nuisances to facilitate the combination of analyses.

5.2: Design the format and nature of the public and internally preserved data products, such as statistical models, with reinterpretation use-cases in mind.

5.3: Improve the coordination among the different public reinterpretation frameworks with the goal of a centralised database of recast codes, common input/output formats, and a unified statistical treatment.

5.4: Encourage the FAIR-ification of codes and data products from (theory) reinterpretation studies outside the experimental collaborations at the same level of sophistication as asked for experimental analyses and results. Suitable repositories are, e.g., GitHub and Zenodo; appropriate versioning is essential.

Figure 5: Recommendations for reinterpretation and recasting.

6 Conclusions

The recommendations we put forth in this paper are designed to give the particle physics community actionable steps in ensuring robust preservation of data, analysis logic, and tools that will allow us to get the most scientific value possible out of the rich data and data products from the experiments. One primary goal is to reduce the amount of overhead by our colleagues in HEP, either in re-reproducing existing analysis logic, or having to re-derive or re-produce data products that should have already been made more accessible from the start. For example, by looking towards harmonisation of analysis logic, experimental collaborations could begin automatising the data products necessary for reinterpreation and reusability, and reduce the burden on their colleagues.

Our recommendations in Figures 2 to 5 will ideally be enacted early on in the design and planning phases for experiments and analyses, and provisioned with appropriate resources and funding. If adopted rigorously, our guidelines will allow for new scientific results, including testing of new models, for decades to come through reinterpretation of published experimental analyses and results, much beyond the lifetime of the experiment.

Acknowledgments

Andy Buckley, Jon Butterworth, and Graeme Watt are supported by the UK STFC Consolidated Grant programme. Kyle Cranmer, Matthew Feickert, and Mark Neubauer are supported in part by the National Science Foundation under Cooperative Agreement OAC-1836650. Lukas Heinrich is supported by the Excellence Cluster ORIGINS, which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC-2094-390783311. Axel Huebl acknowledges support by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. Sabine Kraml acknowledges support by the IN2P3 master project "Théorie – BSMGA" and the joint ANR-FWF project PRCI SLDNP grant no. ANR-21-CE31-0023. Andre Lessa is supported by Sao Paulo Research Foundation (FAPESP) grants no. 2018/25225-9 and 2021/01089-1. Christine Nattrass acknowledges support from the Division of Nuclear Physics of the U.S. Department of Energy under Grant No. DE-FG02-96ER40982 and from the National Science Foundation under Grant No. OAC-1550300. Sezen Sekmen is supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under contracts NRF-2021R1I1A3048138, NRF-2018R1A6A1A06024970 and NRF-2008-00460.

References

[1] LHC Reinterpretation Forum, *Reinterpretation of LHC Results for New Physics: Status and Recommendations after Run 2*, *SciPost Phys.* **9** (2020) 022 [2003.07868].

[2] K. Cranmer et al., *Publishing statistical models: Getting the most out of particle physics experiments*, *SciPost Phys.* **12** (2022) 037 [2109.04981].

[3] "CERN Open Data Policy for the LHC Experiments." DOI: 10.7483/OPENDATA.0XO6.HYY1, 2020.

[4] "CERN Open Data Portal." https://opendata.cern.ch/, 2021.

[5] "CERN Open Data Privacy Policy." http://opendata.cern.ch/docs/privacy-policy, 2021.

[6] CMS Collaboration, "2020 CMS data preservation, re-use and open access policy." DOI: 10.7483/OPENDATA.CMS.1BNU.8V1W, 2020.

[7] ATLAS Collaboration, “ATLAS Data Access Policy.”
DOI: 10.7483/OPENDATA.ATLAS.T9YR.Y7MZ, 2014.

[8] LHCb collaboration, “LHCb External Data Access Policy.”
DOI: 10.7483/OPENDATA.LHCb.HKJW.TWSZ, 2013.

[9] ALICE collaboration, “ALICE data preservation strategy.”
DOI: 10.7483/OPENDATA.ALICE.54NE.X2EA, 2013.
OPENDATA.ALICE.54NE.X2EA.

[10] M.D. Wilkinson et al., *The FAIR Guiding Principles for scientific data management and stewardship*, *Scientific Data* **3** (2016) .

[11] “3rd DPHEP Collaboration Workshop.”
`https://indico.cern.ch/event/1043155/`, 21–23 June, 2021.

[12] “Docker.” `https://www.docker.com/`, 2022.

[13] “Podman.” `https://podman.io/`, 2022.

[14] “SingularityCE.” `https://sylabs.io/singularity/`, 2022.

[15] C. Bierlich et al., *Robust Independent Validation of Experiment and Theory: Rivet version 3*, *SciPost Phys.* **8** (2020) 026 [1912.05451].

[16] C. Bierlich et al., *Confronting experimental data with heavy-ion models: RIVET for heavy ions*, *Eur. Phys. J. C* **80** (2020) 485 [2001.10737].

[17] M. Drees, H. Dreiner, D. Schmeier, J. Tattersall and J.S. Kim, *CheckMATE: Confronting your Favourite New Physics Model with LHC Data*, *Comput. Phys. Commun.* **187** (2015) 227 [1312.2591].

[18] D. Dercks, N. Desai, J.S. Kim, K. Rolbiecki, J. Tattersall and T. Weber, *CheckMATE 2: From the model to the limit*, *Comput. Phys. Commun.* **221** (2017) 383 [1611.09856].

[19] B. Dumont, B. Fuks, S. Kraml, S. Bein, G. Chalons, E. Conte et al., *Toward a public analysis database for LHC new physics searches using MADANALYSIS 5*, *Eur. Phys. J. C* **75** (2015) 56 [1407.3278].

[20] E. Conte and B. Fuks, *Confronting new physics theories to LHC data with MADANALYSIS 5*, *Int. J. Mod. Phys. A* **33** (2018) 1830027 [1808.00480].

[21] GAMBIT collaboration, *ColliderBit: a GAMBIT module for the calculation of high-energy collider observables and likelihoods*, *Eur. Phys. J. C* **77** (2017) 795 [1705.07919].

[22] DELPHES 3 collaboration, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [1307.6346].

[23] B. Petersen and G. Stark, "SimpleAnalysis, v1.1.0." https://gitlab.cern.ch/atlas-sa/simple-analysis. 10.5281/zenodo.6328569.

[24] CMS collaboration, *Search for new particles in events with energetic jets and large missing transverse momentum in proton-proton collisions at $\sqrt{s}$ = 13 TeV*, *JHEP* **11** (2021) 153 [2107.13021].

[25] A. Albert, "Material for the CMS monojet analysis for 101+36/fb (PAS-EXO-20-004). Reinterpretation: Auxiliary Material Presentation (RAMP) #3." https://cds.cern.ch/record/2774586, Jun, 2021.

[26] A. Albert, "Implementation of a search for new phenomena in events featuring energetic jets and missing transverse energy (137 fb-1; 13 TeV; CMS-EXO-20-004)." https://doi.org/10.14428/DVN/IRF7ZL, 2021. 10.14428/DVN/IRF7ZL.

[27] CMS collaboration, *Search for supersymmetry in proton-proton collisions at 13 TeV using identified top quarks*, *Phys. Rev. D* **97** (2018) 012007 [1710.11188].

[28] "Top tagger." https://github.com/susy2015/TopTagger, 2015.

[29] ATLAS collaboration, *Search for squarks and gluinos in final states with jets and missing transverse momentum using 139 fb^{-1} of $\sqrt{s}$ =13 TeV pp collision data with the ATLAS detector*, *JHEP* **02** (2021) 143 [2010.14293].

[30] K. Uno, "Material for the ATLAS 0-lepton gluino/squark search with 139/fb (incl. BDT weights!). Reinterpretation: Auxiliary Material Presentation (RAMP) kickoff meeting." https://cds.cern.ch/record/2763449, April 9, 2021.

[31] ATLAS Collaboration, ""ZeroLepton2018-SRBDT-weight.tar.gz" of "Search for squarks and gluinos in final states with jets and missing transverse momentum using 139 fb^{-1} of $\sqrt{s}$ =13 TeV *pp* collision data with the ATLAS detector" (Version 2)." DOI: 10.17182/hepdata.95664.v2/r8, 2021.

[32] ONNX Community, "Open standard for machine learning interoperability." https://github.com/onnx/onnx, September, 2017.

[33] ATLAS Collaboration, *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*, *Eur. Phys. J. C* **81** (2021) 1023 [2106.09609].

[34] ATLAS Collaboration, "'SUSY-2019-04-ONNX.tgz' of "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" (Version 1)." DOI: 10.17182/hepdata.104860.v1/r3, 2021.

[35] A. Butter, T. Plehn, S. Schumann et al., *Machine Learning and LHC Event Generation*, in *Proceedings of the US Community Study on the Future of Particle Physics (Snowmass 2021)*, 3, 2022 [2203.07460].

[36] P. Amstutz, M.R. Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer et al., "Common Workflow Language, v1.0." https://doi.org/10.6084/m9.figshare.3115156, Jul, 2016. 10.6084/m9.figshare.3115156.v2.

[37] F. Mölder, K.P. Jablonski, B. Letcher, M.B. Hall, C.H. Tomkins-Tinch, V. Sochat et al., *Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]*, *F1000Research* **10** (2021) 33.

[38] K. Cranmer and L. Heinrich, *Yadage and Packtivity - analysis preservation using parametrized workflows*, *J. Phys. Conf. Ser.* **898** (2017) 102019 [1706.01878].

[39] L. Heinrich and K. Cranmer, "yadage/yadage v0.20.2." https://doi.org/10.5281/zenodo.596276, Feb., 2022. 10.5281/zenodo.596276.

[40] "Argo Workflows – The workflow engine for Kubernetes." https://argoproj.github.io/argo-workflows/, 2022.

[41] T. Šimko, C. Lange, L.A. Heinrich, A.E. Lintuluoto, D.M. MacDonell, A. Mečionis et al., *Scalable declarative hep analysis workflows for containerised compute clouds*, *Frontiers in Big Data* **4** (2021) .

[42] B. Krikler, "FAST." https://fast-carpenter.readthedocs.io/en/latest/.

[43] A. Rizzi, "NAIL." https://indico.cern.ch/event/769263/timetable/#25-nail-a-prototype-analysis-l.

[44] CMS Collaboration, *Evidence for Higgs boson decay to a pair of muons*, *JHEP* **01** (2021) 148 [2009.04363].

[45] P. David, *Readable and efficient HEP data analysis with bamboo*, *EPJ Web Conf.* **251** (2021) 03052 [2103.01889].

[46] M. Proffitt and G. Watts, *FuncADL: Functional Analysis Description Language*, *EPJ Web Conf.* **251** (2021) 03068 [2103.02432].

[47] H. B. Prosper, S. Sekmen and G. Unel, "ADL Web Portal." https://cern.ch/adl.

[48] G. Unel, S. Sekmen, A.M. Toon, B. Gokturk, B. Orgen, A. Paul et al., *CutLang V2: towards a unified Analysis Description Language*, *Front. Big Data* **4** (2021) 659986 [2101.09031].

[49] S. Sekmen and G. Ünel, *CutLang: A Particle Physics Analysis Description Language and Runtime Interpreter*, *Comput. Phys. Commun.* **233** (2018) 215 [1801.05727].

[50] "ADL LHC analyses repository." https://github.com/ADL4HEP/ADLLHCanalyses.

[51] H.B. Prosper, S. Sekmen and G. Unel, *Analysis Description Language: A DSL for HEP Analysis*, in *Proceedings of the US Community Study on the Future of Particle Physics (Snowmass 2021)*, 2022.

[52] E. Maguire, L. Heinrich and G. Watt, *HEPData: a repository for high energy physics data*, *J. Phys. Conf. Ser.* **898** (2017) 102006 [1704.05473].

[53] A. Banerjee et al., *Phenomenological aspects of composite Higgs scenarios: exotic scalars and vector-like quarks*, in *Proceedings of the US Community Study on the Future of Particle Physics (Snowmass 2021)*, 3, 2022 [2203.07270].

[54] K. Cranmer, L. Heinrich, R. Jones and D.M. South, *Analysis preservation in ATLAS*, *Journal of Physics: Conference Series* **664** (2015) 032013.

[55] K. Cranmer and I. Yavin, *RECAST: Extending the Impact of Existing Analyses*, *JHEP* **04** (2011) 038 [1010.2506].

[56] ATLAS collaboration, *RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two b-quarks*, Tech. Rep. ATL-PHYS-PUB-2019-032, CERN, Geneva (Aug, 2019).

[57] ATLAS collaboration, *Reinterpretation of the ATLAS Search for Displaced Hadronic Jets with the RECAST Framework*, Tech. Rep. ATL-PHYS-PUB-2020-007, CERN, Geneva (Mar, 2020).

[58] ATLAS collaboration, *Constraining the Dark Sector with the monojet signature in the ATLAS experiment*, Tech. Rep. ATL-PHYS-PUB-2021-020, CERN, Geneva (Jun, 2021).

[59] A. Buckley et al., *Testing new physics models with global comparisons to collider measurements: the Contur toolkit*, *SciPost Phys. Core* **4** (2021) 013 [2102.04377].

[60] S. Kraml, S. Kulkarni, U. Laa, A. Lessa, W. Magerl, D. Proschofsky-Spindler et al., *SModelS: a tool for interpreting simplified-model results from the LHC and its application to supersymmetry*, *Eur. Phys. J. C* **74** (2014) 2868 [1312.4175].

[61] G. Alguero, J. Heisig, C. Khosa, S. Kraml, S. Kulkarni, A. Lessa et al., *Constraining new physics with SModelS version 2*, 2112.00769.

[62] L. Heinrich, "Towards a unified interface for Reinterpretation tools." `https://doi.org/10.5281/zenodo.6362700`, Oct., 2017.

[63] LLP Community, "LLP Recasting Repository." `https://github.com/llprecasting/recastingCodes`, 2019.

[64] GAMBIT collaboration, *Combined collider constraints on neutralinos and charginos*, *Eur. Phys. J. C* **79** (2019) 395 [1809.02097].

[65] A. Kvellestad, P. Scott and M. White, *GAMBIT and its Application in the Search for Physics Beyond the Standard Model*, 1912.04079.

[66] W. Waltenberger, A. Lessa and S. Kraml, *Artificial Proto-Modelling: Building Precursors of a Next Standard Model from Simplified Model Results*, *JHEP* **03** (2021) 207 [2012.12246].

[67] “(Re)interpretation of LHC results for BSM studies Zenodo community.” https://zenodo.org/communities/lhc-recasting/.

[68] “GAMBIT Zenodo community.” https://zenodo.org/communities/gambit-official/.

Glossary

ADL analysis description language. 11, 16

ALICE A detector specialised in heavy-ion physics at the LHC. 13

analysis preservation The procedures, practices, and standards of ensuring the long-term preservation, accessibility, and usability of information necessary to repeat an experimental analysis (starting from its associated preserved data) and generate all associated data products. 4, 7, 10, 12, 13, 15, 16

ATLAS A general-purpose detector at the LHC. 9, 10, 13, 15

BABAR A particle physics experiment at SLAC. 6

BDT boosted decision tree. 9

BSM Beyond-the-Standard Model. 6, 9, 14, 16

CERN European Organization for Nuclear Research. 5, 6

CMS A general-purpose detector at the LHC. 5, 9–11, 13, 15, 24

CMSSW Offline software for the CMS collaboration. 5

data Data recorded from experiments that have been passed through the experiment's event reconstruction. 4, 6, 8, 24

data preservation The procedures, practices, and standards of ensuring the long-term (i.e., decades beyond the end of an experiment) preservation, accessibility, and usability of data and derived data from experiments. 4, 6

data product All files containing selections of derived data and synthesised information from the various stages of an analysis. These might include summary plots and tables, histograms of kinematic distributions, fiducial cross sections, cross section limits, simplified model results, correlation information, analysis statistical workspace binaries or full statistical models, etc. 4, 6, 11, 14, 15, 17, 24

derived data Data that have passed through a size reduction step that prunes information, but which might also add additional calculated quantities to the data files.. 4

DESY German Electron Synchrotron. 25

DSL domain specific language. 10, 11

DUNE Deep Underground Neutrino Experiment. 11

EFT Effective Field Theory. 14, 15

EIC Electron-Ion Collider. 7, 8

FAIR Findable, Accessible, Interoperable, and Reusable. 5, 14, 17

HEP High Energy Physics. 3, 8, 11, 13, 15, 17

HERA Hadron-Electron Ring Accelerator at DESY. 6

HL-LHC High-Luminosity LHC. 7

LEP Large Electron-Positron Collider. 6

LHC Large Hadron Collider. 4–8, 11, 13, 24, 25

LHCb A detector specialised in *b*-physics at the LHC. 13

LZ LUX-ZEPLIN dark matter experiment. 11

ML machine learning. 9, 10, 17

ONNX Open Neural Network Exchange. 9

PHENIX Pioneering High Energy Nuclear Interaction eXperiment, was a detector at RHIC designed to investigate high energy collisions of heavy ions and protons. 8, 13, 25

QCD Quantum Chromodynamics. 7

recasting Reproducing the analysis logic in a simulation, considering a different physical process with a different phase space distribution, which might have different efficiencies and acceptances than the originally hypothesised model. 4, 14–16

reinterpretation Any type of new or updated interpretation of an experimental analysis or result, including the combination in, e.g., global fits or global averages.. 4, 6, 8, 14, 15

RHIC Relativistic Heavy Ion Collider. 8, 25

SLAC SLAC National Accelerator Laboratory. 24

sPHENIX A detector at RHIC studying quark-gluon plasma by combining PHENIX and STAR. 13

STAR One of four experiments at RHIC studying quark-gluon plasma. 8, 13, 25