Overview

An overhaul of established approaches to analyses at the LHC is needed to meet the challenge of handling an order of magnitude more data expected from the High-Luminosity LHC.

Existing software in use by the LHC to perform binned template fits is typically in the form of monolithic frameworks, and often not available for use outside the experiments.

IRIS-HEP approaches the challenge with a modular workflow focused on well-defined interfaces. It is containerized for analysis preservation and reusability. The modular nature allows for natural integration of tools developed within IRIS-HEP and beyond.

Modularity & interfaces

IRIS-HEP investigates the performance and usability of existing software for various parts of the workflow ($\S2$). The studies compare the use of an established monolithic framework (TRExFitter, $\S2$) to approaches that make use of pyhf ($\S3$) and tools developed in the FAST-HEP ($\S4$) project.

To accommodate novel analysis methods, the full workflow is envisioned to be end-to-end differentiable.

Declarative configuration

Different stages of the workflow require similar information. The user specifies this information in a configuration file. The declarative format makes the configuration highly readable and intuitive to use. The format can be serialized for example as JSON and YAML and easily parsed for the relevant steps in the workflow. A possible design of such a configuration file is being investigated in $\S5$.

Columns data

The typical dataset size at this stage is multiple terrabyte, and can scale up to petabyte depending on how much filtering was already applied. ServiceX from the DOMA focus area provides relevant parts of the dataset on demand. The scalability and intelligent caching are crucial for fast turnaround times.

The connection between existing analysis frameworks and ServiceX is being explored in $\S6$.

Selection & systematic uncertainties

Event selection, columnar operations and the various kinds of processing performed in this step make it the most compute-intensive part of the workflow. A wide range of packages with IRIS-HEP involvement, including coffea ($\S7$), enter at this stage. To achieve modularity, a common ground for a selection language is needed.

The SSL focus area allows IRIS-HEP to benchmark realistic analysis examples at scale.

From models to likelihoods

The so-called workspace serializes all information needed to build the likelihood function for subsequent inference. While they have traditionally been ROOT-based, pyhf ($\S3$) now provides a python-based alternative for workspaces in the HistFactory ($\S8$) scheme, easily serializable as JSON.

Extensions to the definition of the workspace are planned in order to accommodate novel analysis methods.

Fit results and diagnostics

Fit results are typically presented with a small set of common visualizations used for diagnostics. We have developed user stories for these visualizations to facilitate the development of a common declarative format to specify them. This promotes modularity and allows for easier preservation of the results. For details, see $\S9$.

Reusability and preservation

Analysis reusability and preservation are guiding design principles. The modular containerized workflow is of great use to achieve these goals. Well-defined schemas and interfaces further help preserve analyses in a common format.