Using an agentic harness to write the AGC

Introduction and Motivation

Analyses in high energy physics require a number of software libraries integrated into the workflow in order to transform raw detector data into final physical results. While numerous software tools exist such as ServiceX, FuncADL, coffea, and cabinetry, physicists must nevertheless spend a lot of time learning how they work as well as writing and debugging the code for their analysis. This project proposes to use Claude Code's Sonnet LLM model to do a large part of this heavy lifting, allowing the physicist to translate the high-level analysis decisions into a software implementation with less effort. This would also improve reproducibility and standardization of workflows as manual implementation would be reduced.

This project proposes development of a number of Claude Code skills that can implement specific steps of a standard analysis workflow, from the datasets of detector data and simulations all the way to the final result. These skills will not replace the physicist's role in designing an analysis, but rather allow them to translate the high-level information such as what background processes to consider, which variables to plot, and which systematic uncertainties to include, into a software implementation. Each skill would have explanations of the relevant concepts, and generalized instructions for how to use the software so that the LLM understands enough about the software to implement the code and variations of the code for different contexts. Furthermore, the skills will include numerous validations to scrutinize the generated code and the output it generates, allowing Claude to correct errors while ensuring agreement with established results.

Scope of the Project

The project will focus on automating implementation of an existing analysis where the physicist has already decided the high-level structure of the analysis, including:

- The physics process or measurement of interest.
- The signal and background processes to be modeled.
- The discriminating variable(s) and event categories to be used.
- The desired event selections and analysis strategy.

The objective is to provide an efficient and reproducible method for translating these decisions into a functioning analysis pipeline using existing software tools. The project will focus on the already established ATLAS $H \rightarrow ZZ$ Analysis Grand Challenge (AGC) workflow, using this analysis to test the Claude skills initially but to have the skills be generalized to other analyses. Testing on other analyses will be done towards the end.

Software Deliverables and Timeline

The primary deliverable will be a set of Claude Code skills, each corresponding to a major stage of the AGC analysis pipeline. Each skill will include general explanations of the concepts, instructions, examples, and ways to validate the written code.

The project will run for 8 weeks from **Monday June 22, 2026** through **Friday August 14, 2026**.

This plan involves producing one new skill each week. As the coffea implementation of weighting, cuts, event selection, and systematic uncertainties is expected to be more involved due to more concepts needing to be understood and implemented, I set aside approximately two weeks. I will begin with the last stage of the analysis pipeline, the statistical analysis using cabinetry, and work backwards. This allows me to use established inputs for each stage of the analysis before trying to recreate them.

A tentative schedule is:

- **Week 1:** Statistical analysis (cabinetry/pyhf)
- **Week 2:** Plotting and saving histograms of relevant variables (uproot)
- **Weeks 3–4:** Event selections, weights, cuts, and systematic uncertainties (coffea)
- **Week 5:** Dataset retrieval and import (ServiceX/FuncADL)
- **Week 6:** Combining skills, debugging, and validation across skills (if needed)
- **Weeks 7-8:** Testing skills on other established analysis workflows, and investigations on eliminating redundant input information

Following development of the individual skills, the final two weeks will be spent mainly testing the skills for other established analyses to improve generalization, such as other precision measurements as well as searches for new physics. Numerous validations and tests will be implemented at each stage, ensuring a final result consistent with expectation and prior analysis. I will also investigate the minimization of necessary information for the user to supply, and allow Claude to figure out some of this information from earlier parts of the workflow. For example, the statistical analysis skill will need the user to specify samples, regions, and systematic uncertainties; however, this information would need to be already specified earlier in the analysis, and can be inferred by what datasets are chosen. The final outcome will be a prototype framework of skills demonstrating how LLM-guided software generation can streamline implementation of analyses in high energy physics while preserving the user's control over analysis design and interpretation.