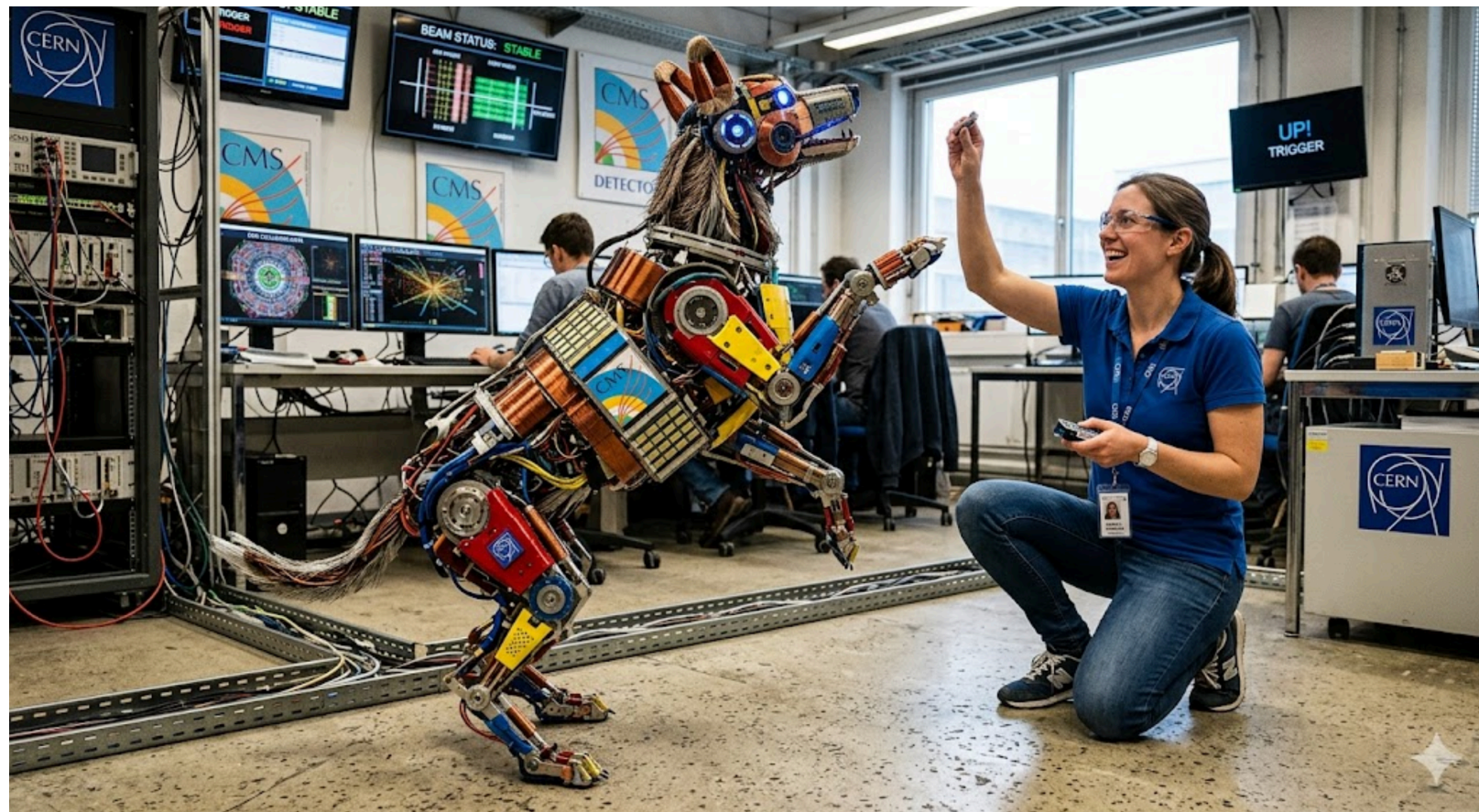


From Prompts to Publications: Agentic AI for Collider Analysis



ML4Jets Vienna 2026

Eric A. Moreno, Andrzej Novak, Sam Bright-Thonney, Dolores Garcia, Yiyang Zhao, Phil Harris

How many of you are using AI to
write papers / solve problems /
conduct research / think for you?

100%

Hopefully

How long until you're
replaced?

How do you make yourself
useful as to not be replaced?

Agentic science is the new hot topic

Automating High Energy Physics Data Analysis with LLM-Powered Agents

Haichen Wang
UC Berkeley
haichenwang@berkeley.edu

Joshua Ho
UC Berkeley
ho22joshua@berkeley.edu

Luc Tomas Le Pottier
UC Berkeley
luclepot@berkeley.edu

Dongwon Kim
UC Berkeley
dwkim@berkeley.edu

Chengxi Yang
UC Berkeley
cxyang@berkeley.edu

Eli Gendreau-Distler
UC Berkeley
egendreaudistler@berkeley.edu

Abstract

We present a proof-of-principle study demonstrating the use of large language model (LLM) agents to automate a representative high energy physics (HEP) analysis. Using the Higgs boson diphoton cross-section measurement as a case study with ATLAS Open Data, we design a hybrid system that combines an LLM-based supervisor-coder agent with the Snakemake workflow manager. In this architecture, the workflow manager enforces reproducibility and determinism, while the agent autonomously generates, executes, and iteratively corrects analysis code in response to user instructions. We define quantitative evaluation metrics – success rate and error distribution – to assess agent performance across multi-stage workflows. To characterize variability across architectures, we benchmark a representative selection of state-of-the-art LLMs, spanning the *Gemini* and *GPT-5* series, the *Claude* family, and leading open-weight models. While the workflow manager ensures deterministic execution of all analysis steps, the final outputs still show stochastic variation. Although we set the temperature to zero, other sampling parameters (e.g., top-p, top-k) remained at their defaults, and some reasoning-oriented models internally adjust these settings. Consequently, the models do not produce fully deterministic results. This study establishes the first LLM-agent-driven automated data-analysis framework in HEP, enabling systematic benchmarking of model capabilities, stability, and limitations in real-world scientific computing environments. The baseline code used in this work is available at <https://huggingface.co/HWresearch/LLM4HEP/tree/main>.

AI Agents Can Already Autonomously Perform Experimental High Energy Physics

Eric A. Moreno^{*†1,2}, Samuel Bright-Thonney^{*†1,2}, Andrzej Novak^{*§1,2}, Dolores Garcia^{#3}, and Philip Harris^{†1,2}

¹Department of Physics, Massachusetts Institute of Technology
²NSF AI Institute for Artificial Intelligence and Fundamental Interactions
³CERN

Article | [Open access](#) | Published: 25 March 2026

Towards end-to-end automation of AI research

[Chris Lu](#), [Cong Lu](#), [Robert Tjarko Lange](#), [Yutaro Yamada](#) [✉](#), [Shengran Hu](#), [Jakob Foerster](#)

[Jeff Clune](#) [✉](#)

[Nature](#) **651**, 914–919 (2026) | [Cite this article](#)

183k Accesses | **3** Citations | **625** Altmetric | [Metrics](#)

An End-to-end Architecture for Collider Physics and Beyond

Shi Qiu,¹ Zeyu Cai,¹ Jiashen Wei,¹ Zeyu Li,^{1,2} Yixuan Yin,¹ Qing-Hong Cao,^{1,3,2} Chang Liu,^{1,4} Ming-xing Luo,⁵ Xing-Bo Yuan,⁶ and Hua Xing Zhu^{1,2}

¹*School of Physics, Peking University, Beijing 100871, China*

²*Center for High Energy Physics, Peking University, Beijing 100871, China*

³*School of Physics, Zhengzhou University, Zhengzhou 450001, China*

⁴*State Key Laboratory of Nuclear Physics and Technology, Peking University, Beijing 100871, China*

⁵*Beijing Computational Science Research Center, Beijing 100193, China*

⁶*Institute of Particle Physics and Key Laboratory of Quark and Lepton Physics (MOE), Central China Normal University, Wuhan, Hubei 430079, China*

We present, to our knowledge, the first language-driven agent system capable of executing end-to-end collider phenomenology tasks, instantiated within a decoupled, domain-agnostic architecture for autonomous High-Energy Physics phenomenology. Guided only by natural-language prompts supplemented with standard physics notation, COLLIDERAGENT carries out workflows from a theoretical Lagrangian to final phenomenological outputs without relying on package-specific code. In this framework, a hierarchical multi-agent reasoning layer is coupled to MAGNUS, a unified execution backend for phenomenological calculations and simulation toolchains. We validate the system on representative literature reproductions spanning leptoquark and axion-like-particle scenarios, higher-dimensional effective operators, parton-level and detector-level analyses, and large-scale parameter scans leading to exclusion limits. These results point to a route toward more automated, scalable, and reproducible research in collider physics, cosmology, and physics more broadly.

Agentic AI – Physicist Collaboration in Experimental Particle Physics: A Proof-of-Concept Measurement with LEP Open Data

Anthony Badea,^a Yi Chen,^b Marcello Maggi,^c Yen-Jie Lee,^d Electron-Positron Alliance

^a*University of Chicago, Enrico Fermi Institute, 60637 IL, USA*

^b*Vanderbilt University, Nashville, Tennessee, USA*

^c*Istituto Nazionale di Fisica Nucleare, Bari Division, BA 70126, Italy*

^d*Laboratory for Nuclear Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

Autonomous Agents for Scientific Discovery: Orchestrating Scientists, Language, Code, and Physics

Lianhao Zhou¹ Hongyi Ling¹ Cong Fu¹ Yepeng Huang²
Michael Sun³ Wendi Yu¹ Xiaoxuan Wang⁴ Xiner Li¹
Xingyu Su¹ Junkai Zhang⁴ Xiusi Chen⁵ Chenxing Liang¹
Xiaofeng Qian^{6,7,8} Heng Ji⁵ Wei Wang⁴ Marinka Zitnik² Shuiwang Ji^{1,6,9*}

¹Department of Computer Science and Engineering, Texas A&M University

²Department of Biomedical Informatics, Harvard Medical School

³Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

⁴Department of Computer Science, University of California, Los Angeles

⁵Siebel School of Computing and Data Science, University of Illinois Urbana Champaign

⁶Department of Materials Science and Engineering, Texas A&M University

⁷Department of Electrical and Computer Engineering, Texas A&M University

⁸Department of Physics and Astronomy, Texas A&M University

⁹J. Mike Walker ’66 Department of Mechanical Engineering, Texas A&M University

CPTNP-2026-012

Agentic Diagrammatica: Towards Autonomous Symbolic Computation in High Energy Physics

Tony Menzo^{1,2*}, Alexander Roman¹, George T. Fleming², Sergei Gleyzer¹, Konstantin T. Matchev¹, Stephen Mrenna²

¹*Department of Physics and Astronomy, University of Alabama, Tuscaloosa, AL 35487, USA*

²*Fermi National Accelerator Laboratory, Batavia, IL 60510, USA*

March 31, 2026

Abstract

We present *Diagrammatica*, a symbolic computation extension to the HEP-TAPOD agentic framework, which enables LLM agents to plan and execute multi-step theoretical calculations. Symbolic computation poses a distinctive reliability challenge for LLM agents, as correctness is governed by implicit mathematical conventions that are not encoded in a form that can be easily checked in the computational backend. We identify two complementary remedies, tool-constrained computation and targeted knowledge grounding, and pursue the first as the primary architecture. Concretely, we concentrate the agent’s action distribution onto tool calls with convention-fixing semantics, in which the agent specifies a compact, human-auditable diagram specification and a trusted backend performs the symbolic or numerical manipulations exactly. The toolkit provides two complementary calculation paths consuming a shared diagram specification: Naive Dimensional Analysis (NDA) for order-of-magnitude rate estimates and Exact Diagrammatic Analysis (EDA) for tree-level symbolic calculations via automatic FeynCalc code generation,

Dr.Sai: An agentic AI for real-world physics analysis at BESIII

Mingfeng He^{1,2}, Fayu Jiang¹, Junkun Jiao³, Mingrun Li^{1,2}, Ke Li^{1,2*}, Yipu Liao^{1,2*}, Beiji Jiang Liu^{1,2}, Tong Liu^{1,2}, Fazhi Qi^{1,2}, Zijie Shang⁴, Weimin Song³, Yue Sun^{1,2}, Xiongfei Wang⁴, Hong Wang^{1,2}, Dongbo Xiong^{1,2}, Changzheng Yuan^{1,2}, Bolun Zhang^{1,2*}, Zhengde Zhang^{1,2*}, Xuliang Zhu^{5,6,7}

^{1*}Institute of High Energy Physics, CAS, Beijing, 100049, China.

²University of Chinese Academy of Sciences, Beijing, 100049, China.

³Jilin University, Changchun, Jilin, 130012, China.

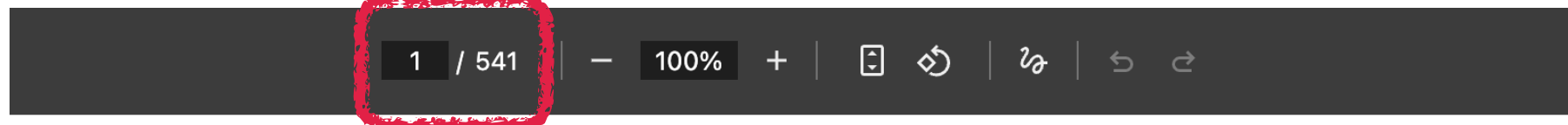
⁴Lanzhou University, Lanzhou, Gansu, 730000, China.

⁵Tsung-Dao Lee Institute, Shanghai Jiao Tong University, Shanghai, 201210, China.

⁶Institute of Nuclear and Particle Physics, School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai, 200240, China.

⁷Key Laboratory for Particle Astrophysics and Cosmology (MOE), Shanghai Key Laboratory for Particle Physics and Cosmology (SKLPPC), Shanghai, 200240, China.

On the pursuit of physics automation



A lot of you have seen this paper...

AI Agents Can Already Autonomously Perform
Experimental High Energy Physics

Eric A. Moreno^{*†1,2}, Samuel Bright-Thonney^{*†1,2}, Andrzej Novak^{*§1,2}, Dolores Garcia^{#3},
and Philip Harris^{b1,2}

¹Department of Physics, Massachusetts Institute of Technology

²NSF AI Institute for Artificial Intelligence and Fundamental Interactions

³CERN

April 2, 2026

Abstract

Large language model-based AI agents are now able to autonomously execute substantial portions of a high energy physics (HEP) analysis pipeline with minimal expert-curated input. Given access to a HEP dataset, an execution framework, and a corpus of prior experimental literature, we find that Claude Code succeeds in automating all stages of a typical analysis: event selection, background estimation, uncertainty quantification, statistical inference, and paper drafting. We argue that the experimental HEP community is underestimating the current capabilities of these systems, and that most proposed agentic workflows are too narrowly scoped or scaffolded to specific analysis structures. We present a proof-of-concept framework, *Just Furnish Context* (JFC), that integrates autonomous analysis agents with literature-based knowledge retrieval and multi-agent review, and show that this is sufficient to plan, execute, and document a credible high energy physics analysis. We demonstrate this by conducting analyses on open data from ALEPH, DELPHI, and CMS to perform electroweak, QCD, and Higgs boson measurements. Rather than replacing physicists, these tools promise to offload the repetitive technical burden of analysis code development, freeing researchers to focus on physics insight, truly novel method development, and rigorous validation. Given these developments, we advocate for new strategies for how the community trains students, organizes analysis efforts, and allocates human expertise.

03.20179v2 [hep-ex] 31 Mar 2026

Talk is based directly off of it: <https://arxiv.org/abs/2603.20179>

<https://jfc-mit.github.io/>

AI Agents Can Already Autonomously Perform Experimental High Energy Physics



Eric A. Moreno^{1,2,*}
RESEARCHER WITH LHC
PHYSICS PHD @ MIT



Samuel Bright-Thonney^{1,2,*}
IAIFI FELLOW @ MIT



Andrzej Novak^{1,2,*}
POSTDOC @ MIT
HEP-EX & ML



Dolores Garcia³
FELLOW @ CERN



Philip Harris^{1,2}
ASSOCIATE PROFESSOR
OF PHYSICS @ MIT

¹ Department of Physics, Massachusetts Institute of Technology

² NSF AI Institute for Artificial Intelligence and Fundamental Interactions

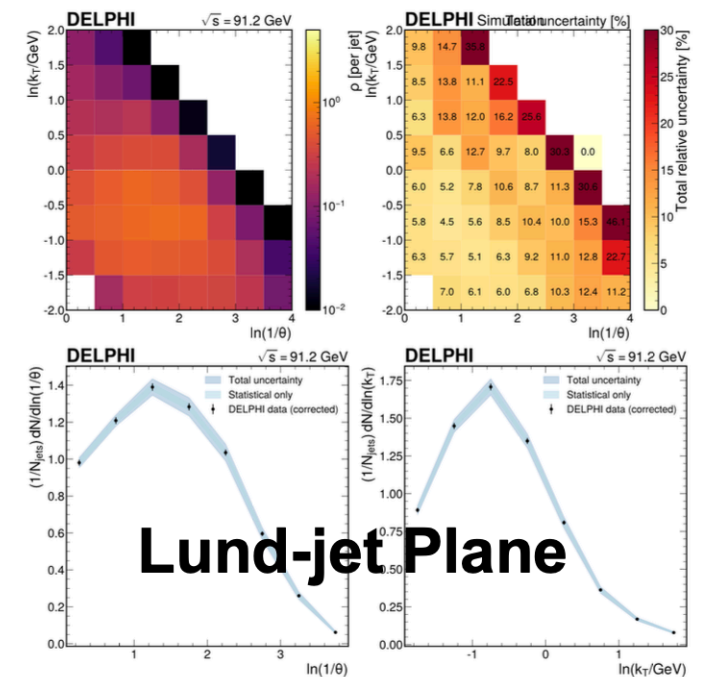
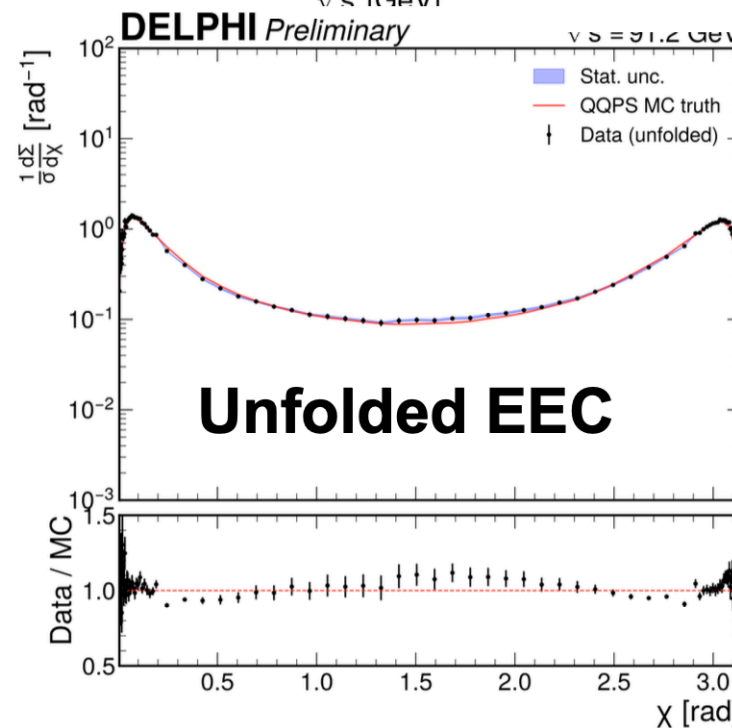
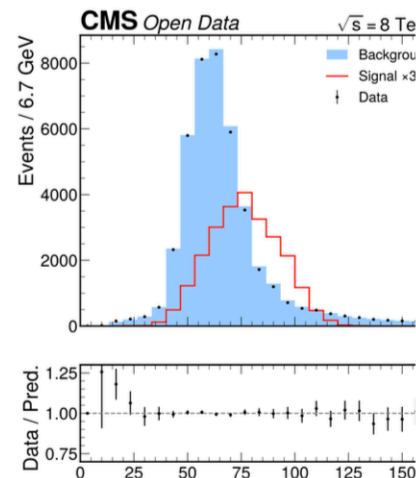
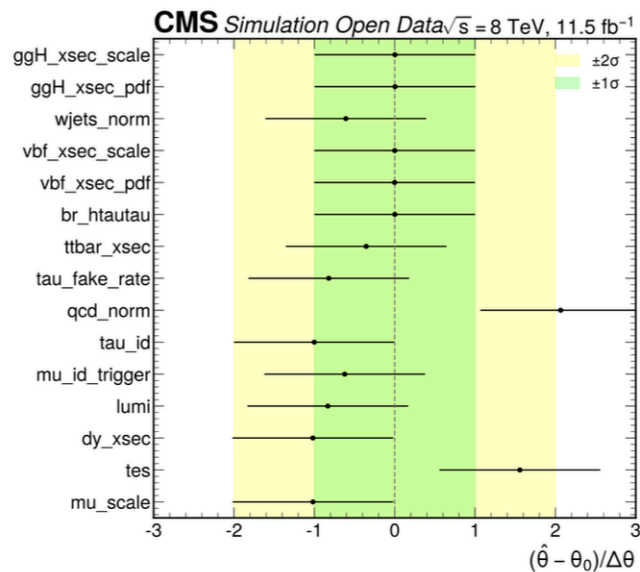
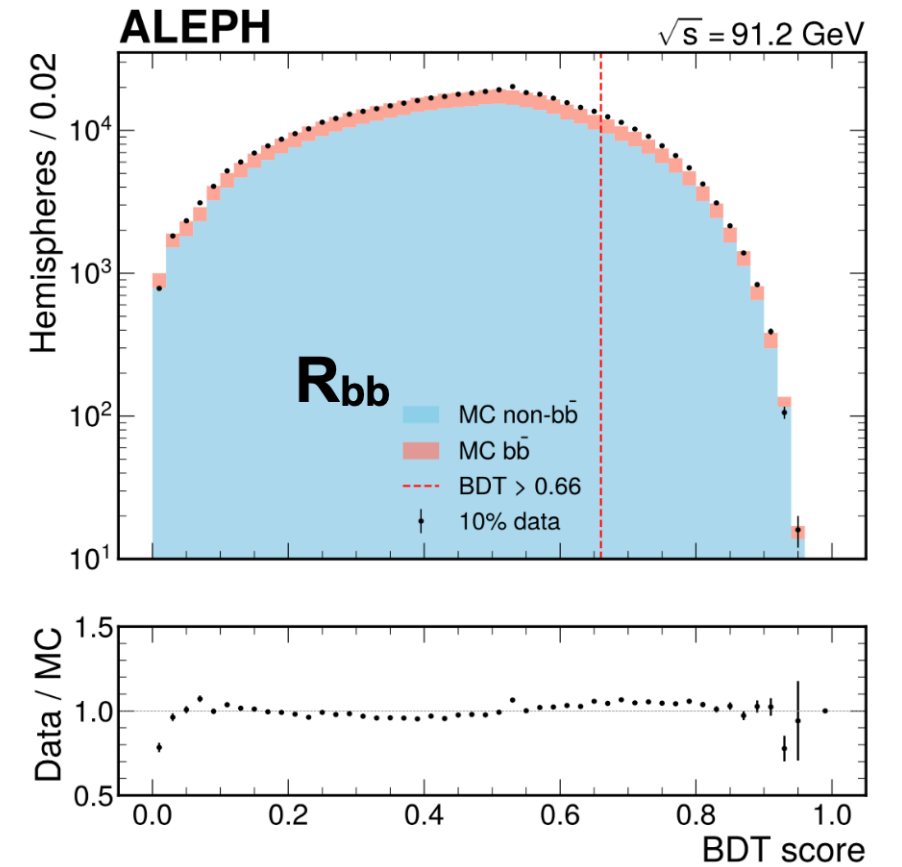
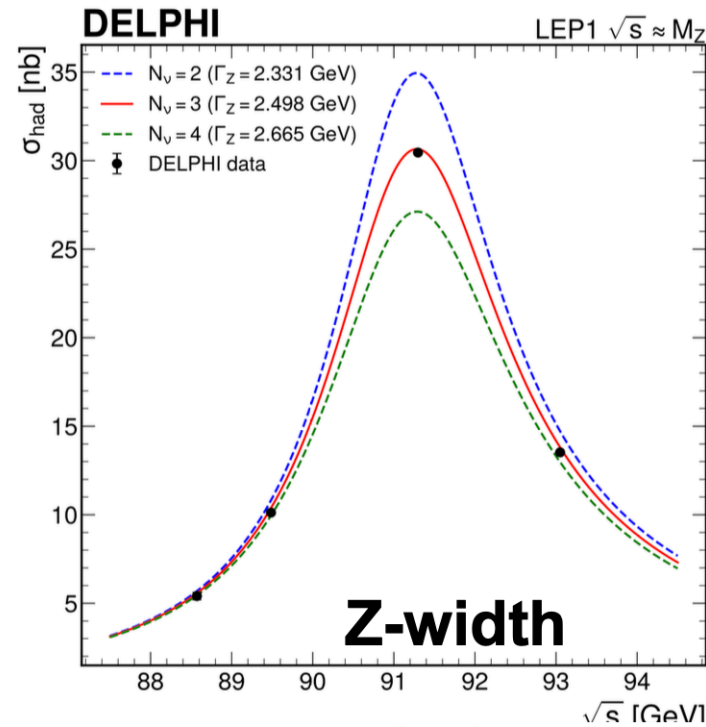
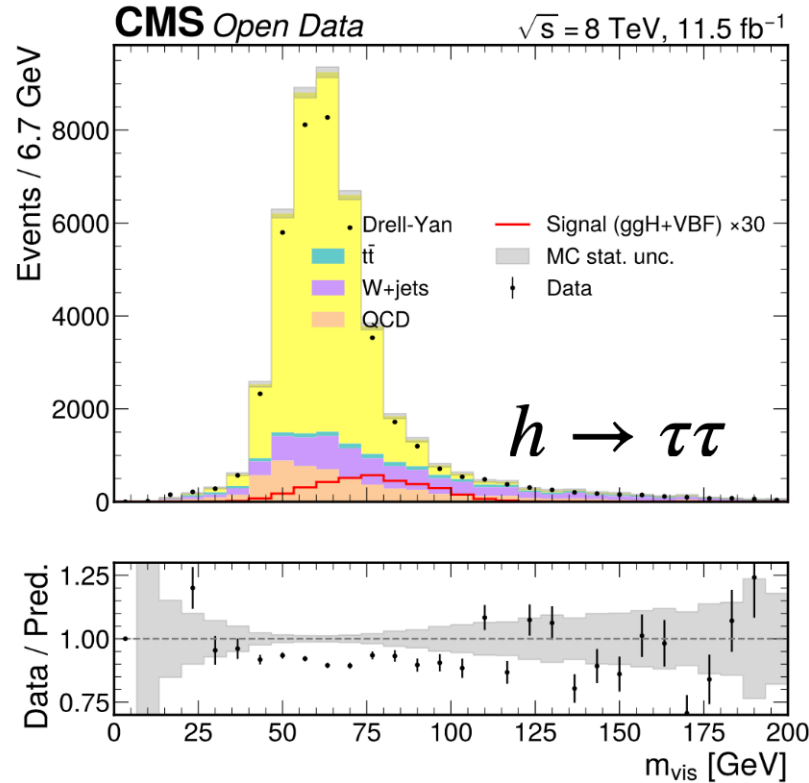
³ CERN

** These authors contributed equally.*

Take away message: This is not hard, we don't have some secret sauce

How good is it?

2 days of running



Reactions

#- Also sent to the channel
Irrespective of the code, The choice of words in the "Title" are so generalized as if this represents the whole of HEP analysis problem.. probably written by the AI itself ? (edited)

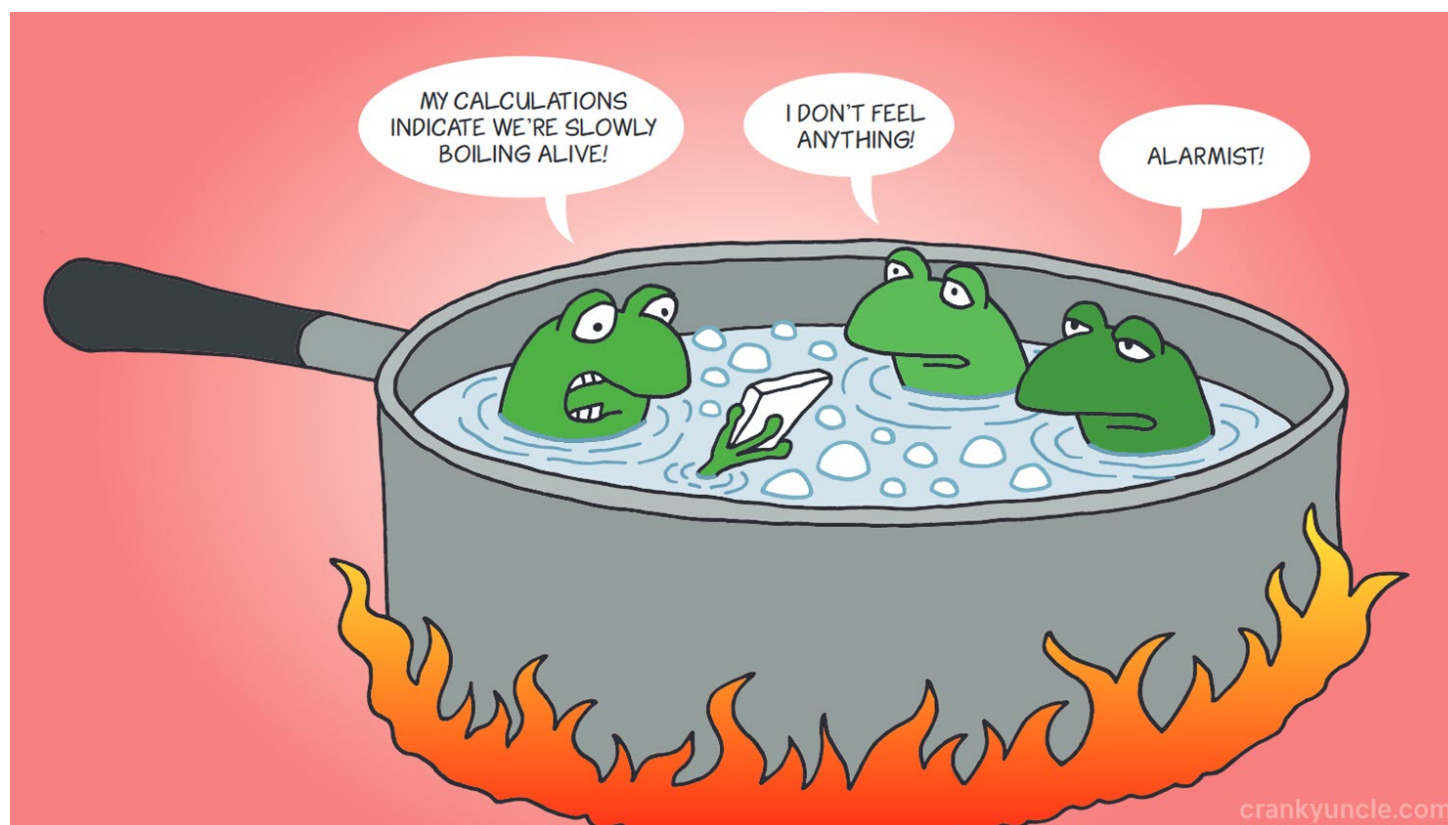
welcome to the era of SLOP

What is this AI slop I just found from you: "with data and MC agreeing within 2%"?!?

300 pages dumped by an llm

the CMS example seems to be ~3k lines of physics code and `phase4/scripts` in <https://github.com/jfc-mit/higgs>

This paper is unreviewable
I don't want to see it



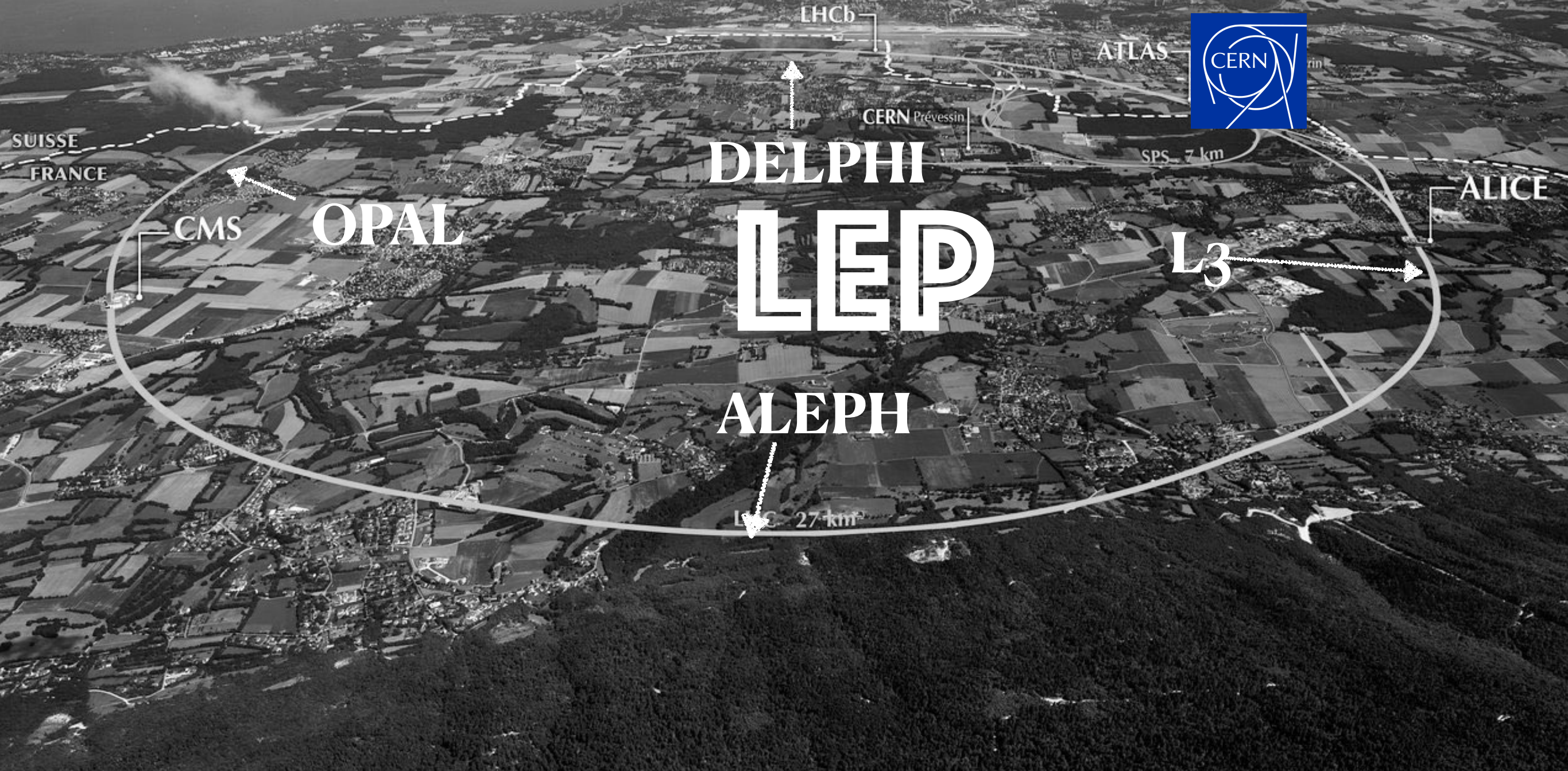
**The pot is warming up:
people who don't adapt will
be left behind**

US HEP PI's already feeling
the effects:

Genesis Mission is a national initiative to build the world's most powerful scientific platform to accelerate discovery science, strengthen national security, and drive energy innovation.

Genesis Mission will develop an integrated platform that connects the world's best supercomputers, experimental facilities, AI systems, and unique datasets across every major scientific domain to double the productivity and impact of American research and innovation within a decade.

THE SUSPECTS



DELPHI
LEP

ALEPH

LHC 27 km



SUISSE
FRANCE

CMS

OPAL

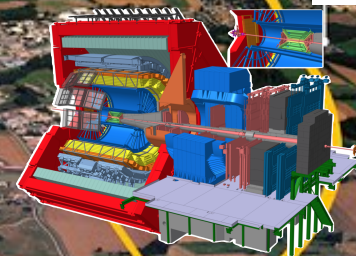
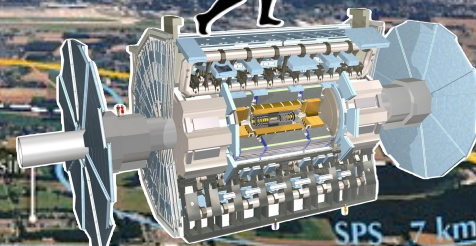
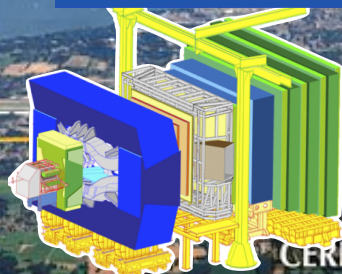
ATLAS

SPS 7 km

ALICE

L3

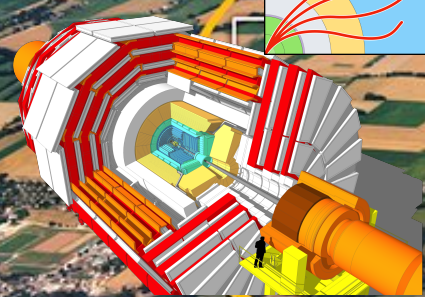
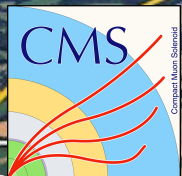
THE SUSPECTS



LHC

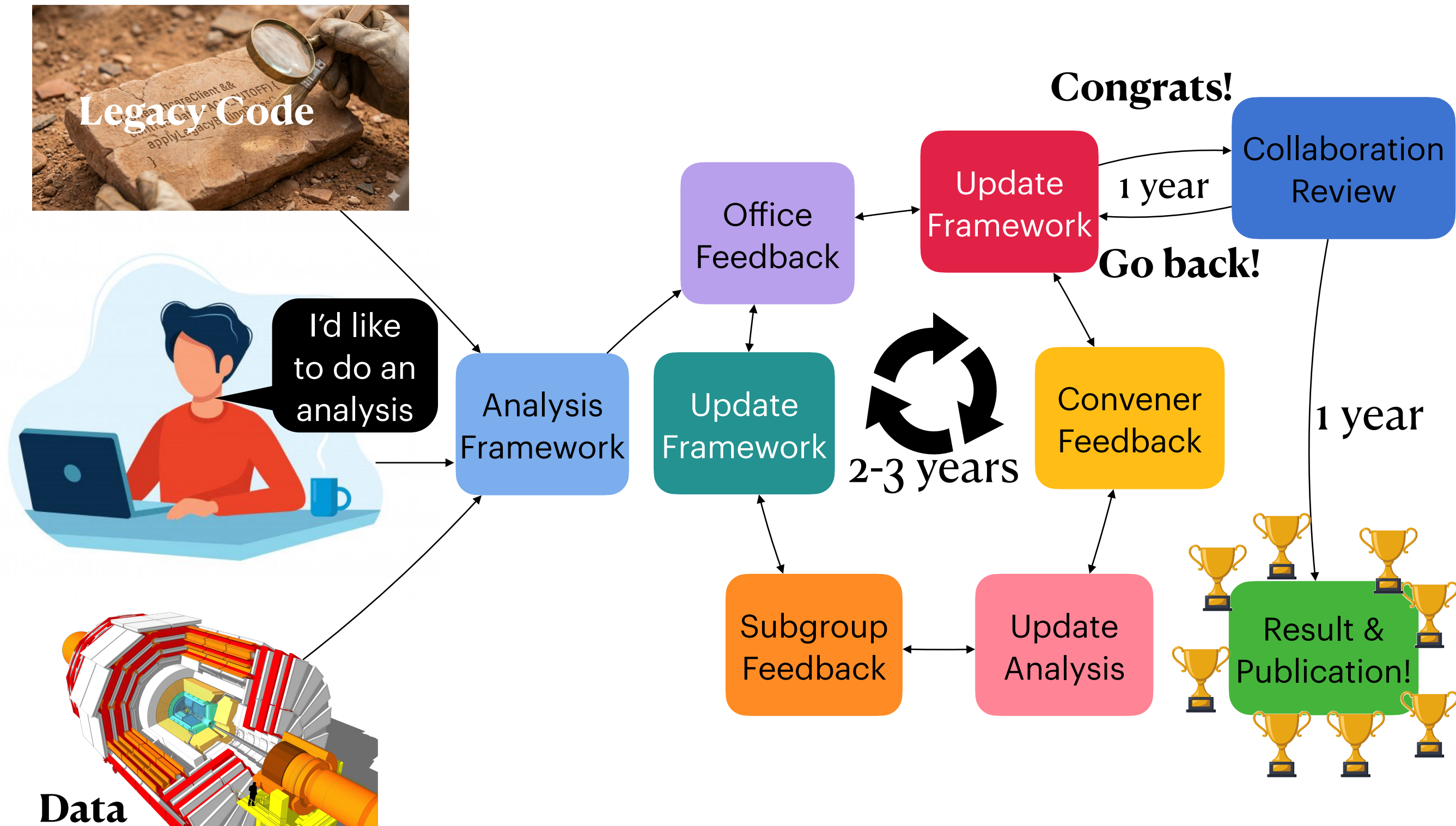


LHC 27 km



SUISSE
FRANCE

How a HEP Analysis Goes

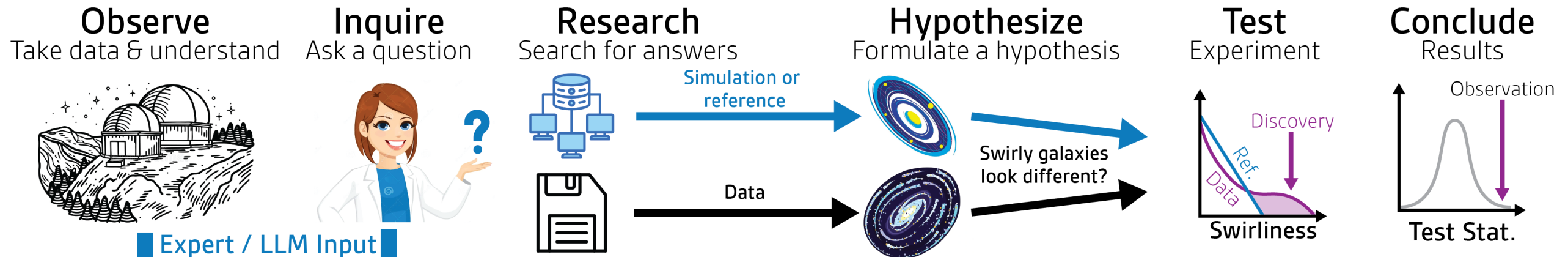


AutoSciDACT

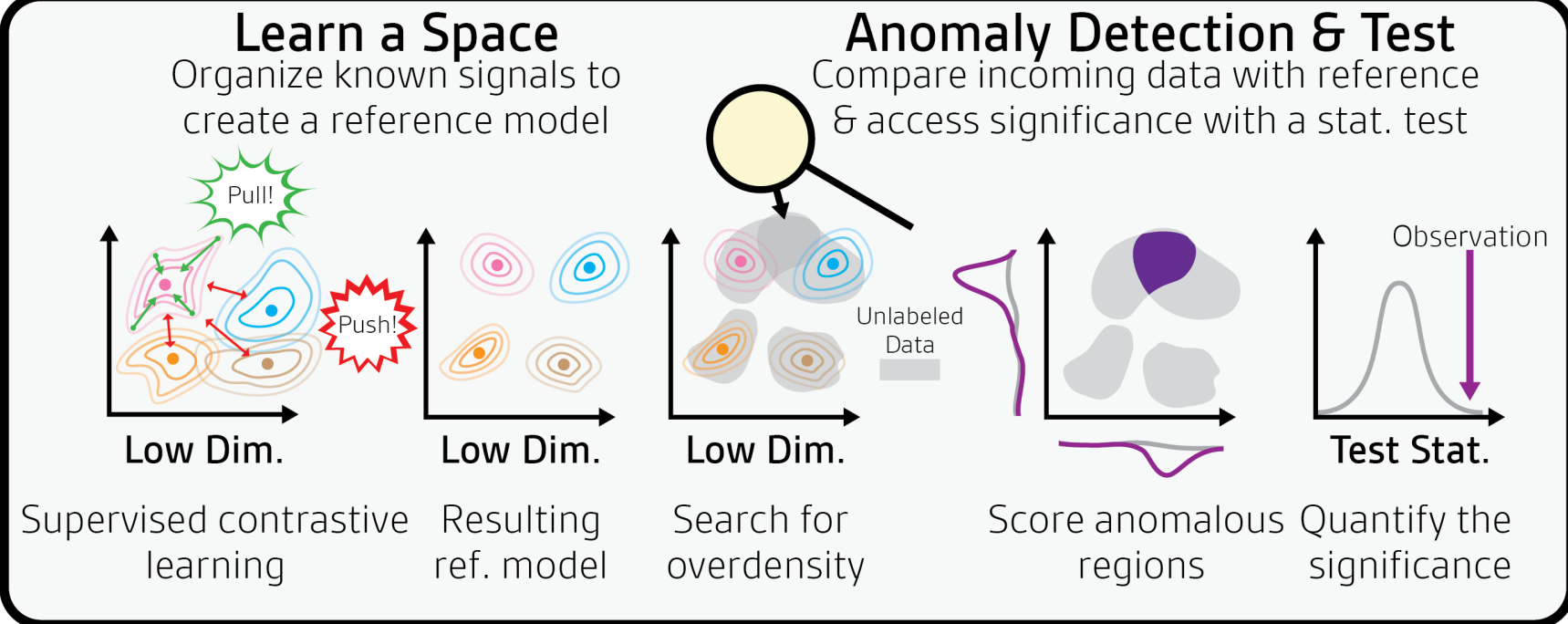
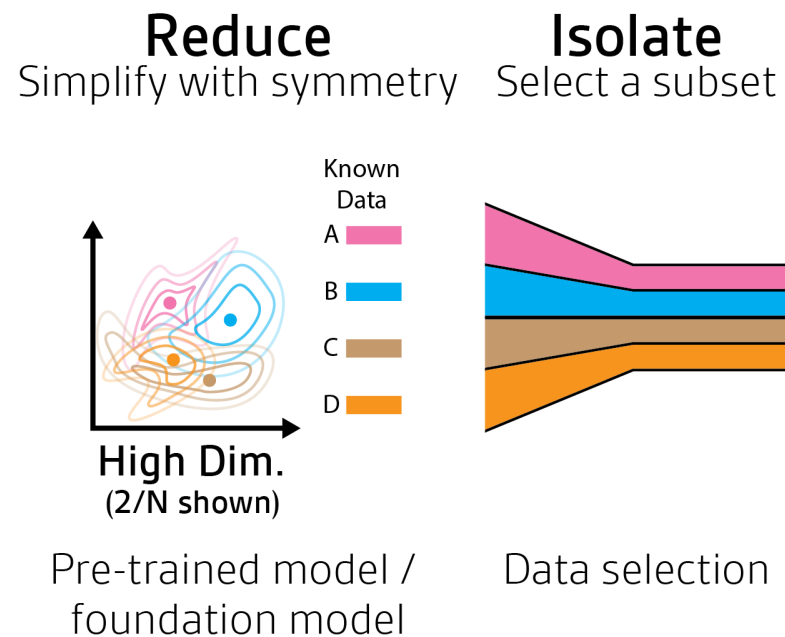
How do we automate scientific discovery?

<https://arxiv.org/abs/2510.21935>

Samuel Bright-Thonney et al.



AutoSciDACT



Split-N-Fit

<https://neurips.cc/virtual/2025/loc/san-diego/122964>

https://ml4physicalsciences.github.io/2025/files/NeurIPS_ML4PS_2025_347.pdf

Andrzej Novak et al.

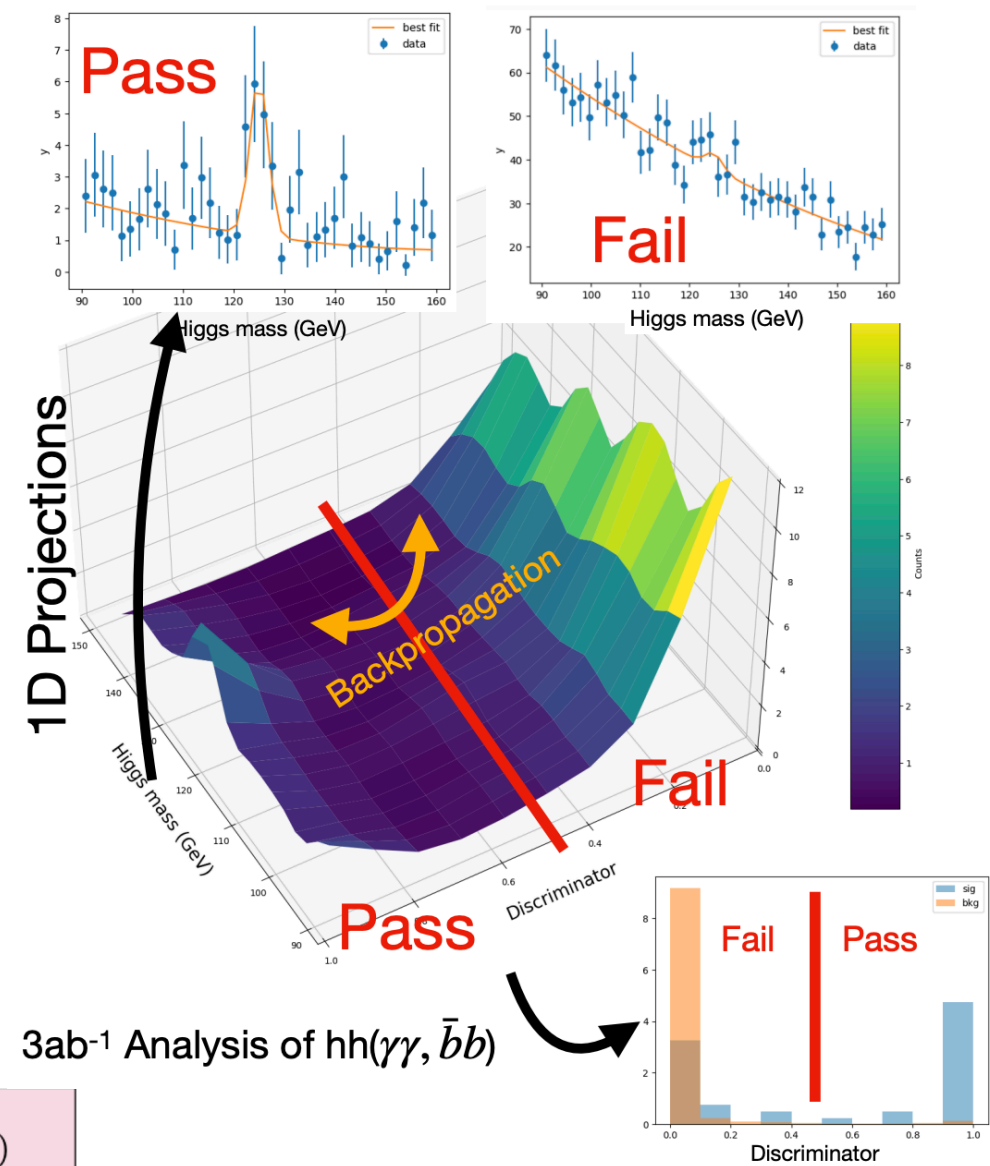
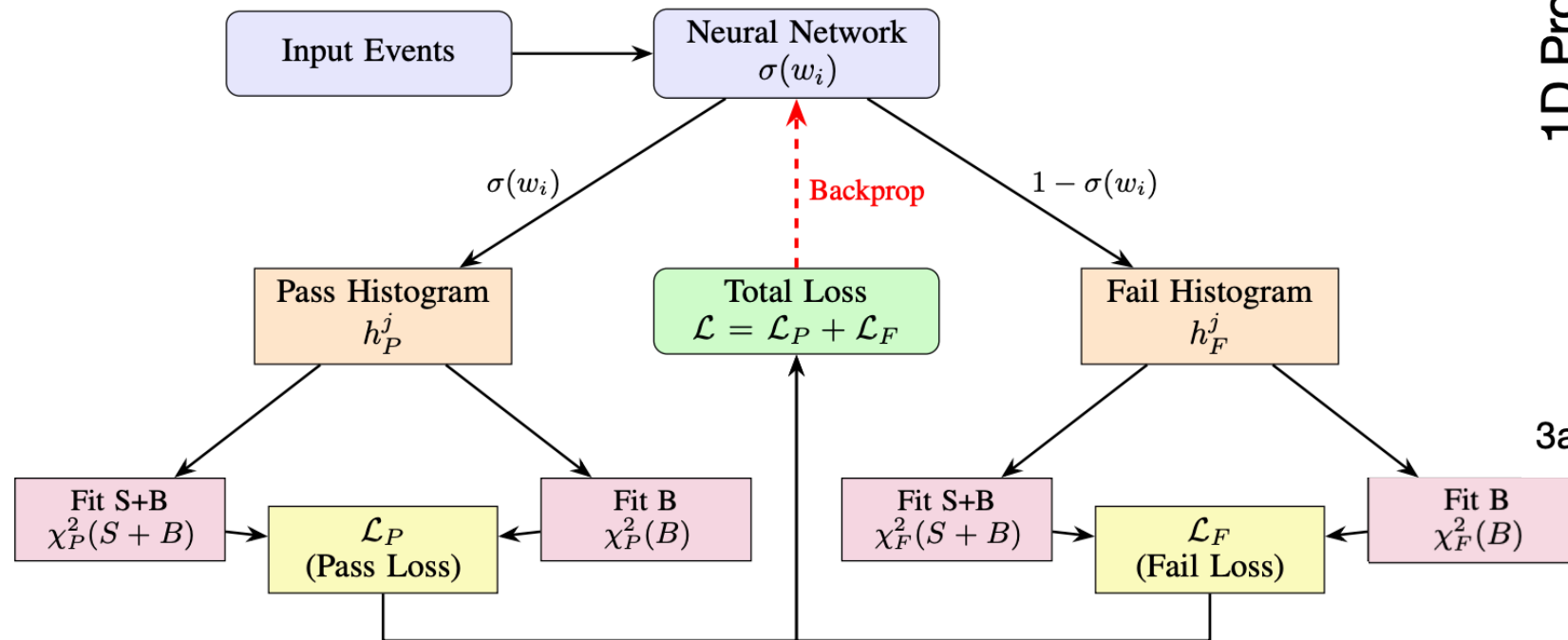


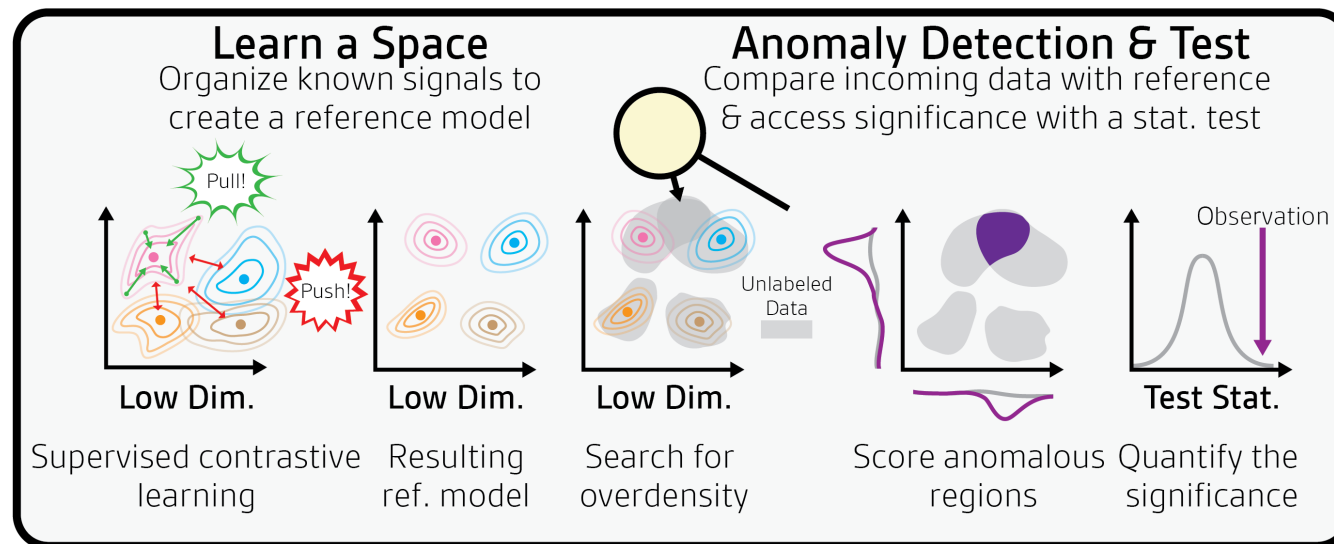
Figure 1: Schematic overview of the Split-N-Fit method. Events are split into pass and fail histograms. Both S+B and B functions are fit to each histogram, yielding χ^2 values that define individual loss terms $\mathcal{L}_P$ and $\mathcal{L}_F$, which combine to form the total differentiable loss.

- Automate event selection using differentiable NN
- Instead of training a classifier and only later doing the physics fit, they merge the two into one end-to-end optimization

The Automation Landscape

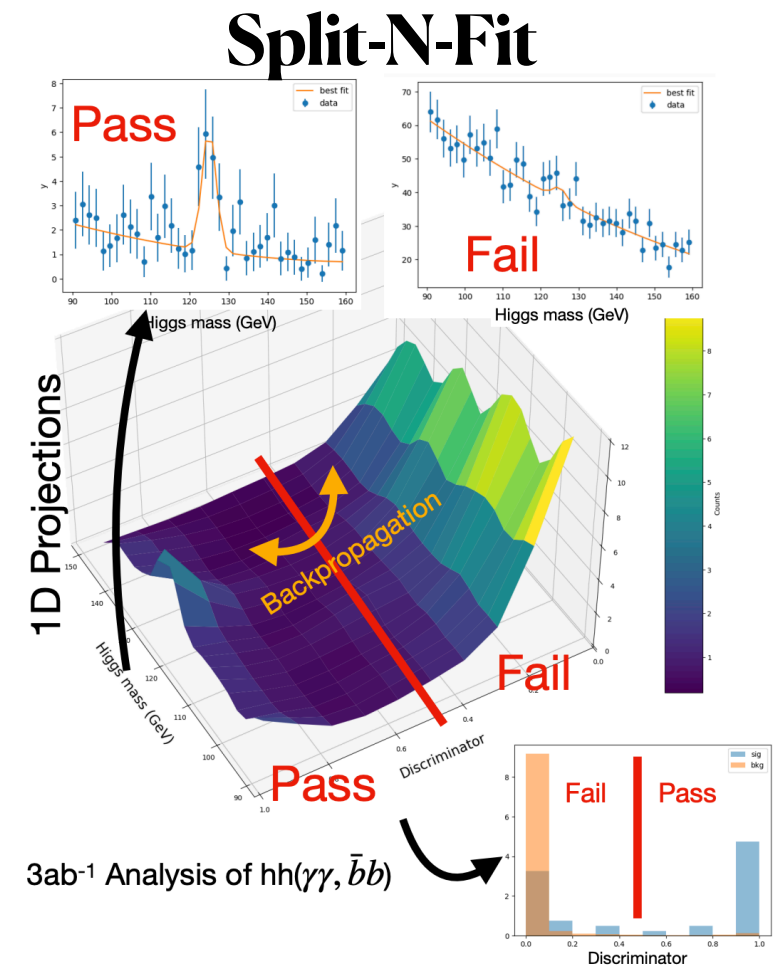
You can do something clever:

AutoSciDACT



<https://arxiv.org/abs/2510.21935>

Or you can do something slop:



<https://neurips.cc/virtual/2025/loc/san-diego/122964>

 Claude

 ChatGPT

 Gemini

Is slop really the answer?

Are We Really That Different?

Attention asks: what should I focus on right now?

Memory asks: what from the past should I bring forward right now?

LLM

$$a_t = f(x_t, \text{Attn}(x_t, m_t))$$

- a_t = action at time t
- x_t = current input or situation
- m_t = memory/state
- $\text{Attn}(x_t, m_t)$ = the part of memory your attention brings into focus
- f = decision process

$$x_{t+1} = x_t + a_t$$

Humanity

$$a_t = f(x_t, \text{Attn}(x_t, m_t), \mathbf{g}_t)$$

- a_t = action at time t
- x_t = current input or situation
- m_t = memory/state
- $\text{Attn}(x_t, m_t)$ = the part of memory your attention brings into focus
- $\mathbf{g}_t$ = **goals, drives, or mood**
- f = decision process

What you do now is a function of what is happening now, conditioned on your past memories, how your attention retrieves them, and what you currently want*

LLMs Can Approximate Us (Maybe)

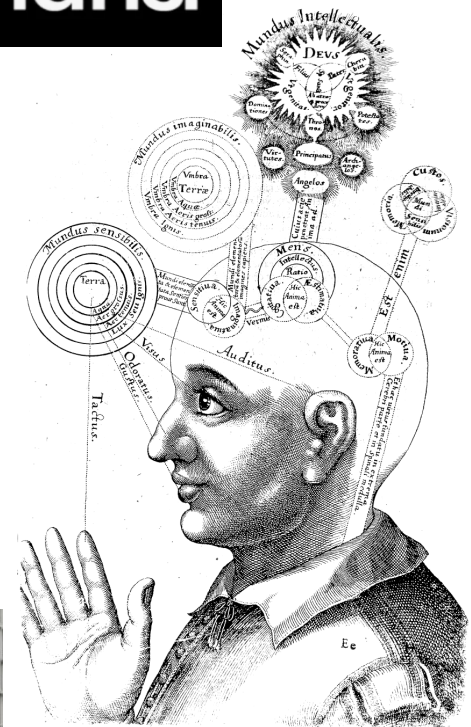
 Claude

 ChatGPT

≈

 Gemini

For humans,
by humans.



First Attempt: Hadronic Z-Line Shape



I want to use claude to create an automated analysis pipeline for high energy physics data. Analysis in HEP (Im thinking LEP, CMS, ATLAS) is one of the most complex undertakings of data analysis anywhere in the world and requires serious reasoning, coding skills, statistical knowledge, and physics knowledge to complete.

I want to structure the analysis bot like a team of humans. There will be one agent looking over the project (call him boss bot), gathering updates from all the subagents, looking at the project hollistically, and deciding where to devote effort or what to improve. Next, there should be a variety of subagents relaying information to boss bot, one for reading relevant literature on the subject forming an idea about the analysis strategy (call them strategy bot) and forming ideas on how to improve on previous analyses. some for processing the data from the "Raw" data format into a usable format (processing bot) that could be a friendly histogram format (for example I use coffea for my analyses this can be decided later). somebody making physics informed cuts on the data and or designing state of the art ML models to filter through the data (filter bot) optimizing the sensitivity of the analysis. somebody designing a statistical fit for the data assigning nuisances working closely with the rest of the bots to import theory nuisances or assign nuisances to different selections and uncertainties and making the fit results look nice like goodness of fit, Data/MC, impacts (stat bot). finally a bot that is looking over the work and preparing slides and a latex paper on the various updates by the bots as they work (paper bot).

Have Claude set a project structure up...

```
# HEP Multi-Agent Analysis

This is a high energy physics analysis managed by a team of AI agents.

## Project Structure
- `.claude/agents/` - Agent definitions (boss-bot, strategy-bot, etc.)
- `config/` - Shared configuration (SAMPLES.yaml, CUTS.yaml, SYSTEMATICS.yaml)
- `strategy/` - Strategy Bot's workspace
- `processing/` - Processing Bot's workspace
- `selections/` - Filter Bot's workspace
- `statistics/` - Stat Bot's workspace
- `paper/` - Paper Bot's workspace
- `messages/` - Inter-agent communication
- `proposals/` - Change proposals (PENDING → APPROVED/REJECTED)
- `lib/` - Shared Python utilities (messaging.py)

## Key Rules
1. Each agent owns specific files - see agent definitions in .claude/agents/
2. Major changes require proposals in proposals/PENDING/
3. All communication happens via messages/ directory using lib/messaging.py
4. ANALYSIS_OVERVIEW.md is the single source of truth
5. Every agent rewrites its STATE.md each cycle (never append)
```

```
name: boss-bot
description: >
  Analysis convener / orchestrator. Invoke for project coordination,
  priority setting, proposal review, and holistic analysis oversight.
  Use when multiple agents need coordination or when strategic decisions
  are needed about the analysis direction.
model: opus
tools: Read, Write, Glob, Grep, Bash
```

```
name: stat-bot
description: >
  Statistical modeling and fitting agent. Invoke when you need to build
  HistFactory workspaces, assign systematics, run fits, compute limits
  or significance, make pull/impact plots, or diagnose fit issues.
model: opus
tools: Read, Write, Glob, Grep, Bash
```

```
# ALEPH_HadronicZLineshape -- Analysis Overview

> **Experiment:** ALEPH (LEP)

## Signal Process
- Primary signal: e+e- → Z/gamma* → q qbar (hadronic final states)
- Observable(s): hadronic cross section sigma_had(sqrt(s)) vs sqrt(s); fit to extract M_Z, Gamma_Z, and sigma^0_had
- Typical signature: high-multiplicity hadronic events (many charged tracks, large visible energy)
- Phase-space: s'/s > 0.10 (accept events with up to ~90% ISR energy loss)

## Target Dataset

### Real Data (3,050,610 total events across 6 files)
- Path: /eos/home-e/emoreno/AnalysisBot/ALEPH_Archive_Data
- File format: ROOT TTrees, primary tree 't' (particle-level, one entry per event)
- sqrt(s) stored per-event in 'Energy' branch (Float_t, GeV)

File | Events | Energy (GeV) |
-----|-----|-----|
LEP1Data1992_recons_aftercut-MERGED.root | 551,474 | 91.28 (single point) |
LEP1Data1993_recons_aftercut-MERGED.root | 538,601 | 89.4 / 91.2 / 91.3 / 93.0 (scan) |
LEP1Data1994P1_recons_aftercut-MERGED.root | 433,947 | ~91.20 |
LEP1Data1994P2_recons_aftercut-MERGED.root | 447,844 | ~91.19 |
LEP1Data1994P3_recons_aftercut-MERGED.root | 483,649 | ~91.20 |
LEP1Data1995_recons_aftercut-MERGED.root | 595,095 | 89.4 / 91.3 / 91.7 / 93.0 (scan) |
```

- ANALYSIS_ROADMAP.md
- background_values.json
- BTAG_STRATEGY.md
- EFFICIENCY_CORRECTION.md
- efficiency_per_period.json
- efficiency_values.json
- ENERGY_BINNING.md
- FINAL_ASSESSMENT.md
- GAMMA_Z_ASSESSMENT.md
- ISR_ALPHA2_FORMULA.md
- ISR_ASSESSMENT.md
- JES_STRATEGY.md
- LEP_ENERGY_ENVIRONMENT.md
- LITERATURE_REVIEW.md
- LUMINOSITY_NOTES.md
- REPORT.md
- SIGNAL_DEFINITION.md
- STATE.md

Does an analysis?

Measurement of the Hadronic Z Lineshape with ALEPH LEP1 Data

ALEPH HEP Analysis Bot Collaboration
March 5, 2026

Abstract

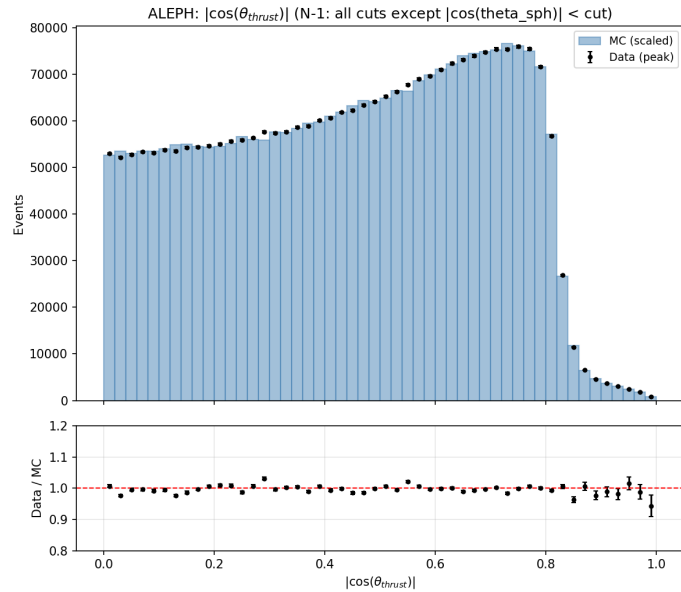
We present a measurement of the hadronic cross-section of the Z boson as a function of centre-of-mass energy using data collected by the ALEPH detector at LEP during the 1992–1995 running periods. A total of approximately 3.05 million hadronic Z decay candidates are selected using a charged-track-based selection. The hadronic cross-section is measured at eight centre-of-mass energy points spanning 89.4–93.0 GeV, split by year and energy, with per-period efficiency corrections derived from published ALEPH cross-sections. Background cross-sections from two-photon and τ -pair processes are subtracted, and systematic uncertainties from luminosity, beam energy calibration, and selection efficiency are incorporated as nuisance parameters in a Poisson negative-log-likelihood fit. A Breit–Wigner $\otimes O(\alpha^2)$ additive ISR lineshape fit with γ -Z interference yields $M_Z = 91.200 \pm 0.005$ GeV, $\Gamma_Z = 2.485 \pm 0.007$ GeV, and $\sigma_{\text{had}}^0 = 41.42 \pm 0.06$ nb, with deviance/ndf = 7.39/5 ($p = 0.19$). The total Z width agrees with the published ALEPH value within 1.5σ , and the hadronic pole cross-section within 2.5σ , demonstrating the critical importance of the additive ISR radiator with hard photon shape corrections $h_1(x)$ and $h_2(x)$. This analysis validates the feasibility of extracting precision electroweak measurements from archived LEP data with modern tools.

Contents

1	Introduction	3
2	Data and Monte Carlo Samples	3
2.1	Collision Data	3
2.2	Luminosity	4
2.3	Monte Carlo Simulation	5
3	Event Selection	5
3.1	Good Track Definition	5
3.2	Event-Level Selection	6
3.3	Validation Against Pre-computed Flags	6
3.4	Cut-Flow	6
3.5	Selected Events per Energy Cluster	7
4	Cross-Section Measurement	7
4.1	Method	7
4.2	Efficiency	8

Not Half Bad

Data/MC Post Selection



Detailed Background Estimation

6.2 Tau-Pair Background

Tau leptons produced in $e^+e^- \rightarrow Z \rightarrow \tau^+\tau^-$ decay into hadrons with a combined branching ratio of approximately 65% (for both taus decaying hadronically) or 35% (one hadronic, one leptonic). In rare cases, the hadronic decay products of both taus, combined with photons from $\pi^0 \rightarrow \gamma\gamma$ and initial-state radiation, can produce events that satisfy the hadronic selection criteria, particularly the charged multiplicity requirement $N_{trk} \geq 5$.

The tau-pair contamination is estimated as:

$$f_\tau = \frac{\sigma_\tau}{\sigma_{had}} \times \frac{\varepsilon_\tau}{\varepsilon_{had}} \quad (6)$$

where:

- $\sigma_\tau/\sigma_{had} = \Gamma_\tau/\Gamma_{had} = 84.08/1744.4 = 0.0482$ is the ratio of partial widths (and hence of Born cross-sections at the Z pole);
- $\varepsilon_\tau = 0.35\% \pm 0.10\%$ is the acceptance of the hadronic selection for tau-pair events, taken from [ALEPH, Eur. Phys. J. C 14 (2000) 1];
- $\varepsilon_{had} = 75.07\%$ is the hadronic selection efficiency (Section 5).

This gives:

$$f_\tau = 0.0482 \times \frac{0.0035}{0.7507} = 2.2 \times 10^{-4} = 0.022\% \quad (7)$$

At the peak, this corresponds to approximately $2,573,249 \times 2.2 \times 10^{-4} \approx 566$ tau-pair events in the selected hadronic sample.

Statistical Method - Fit Model

8.2 Lineshape Model

The theoretical hadronic cross-section as a function of $\sqrt{s}$ is modelled in three stages:

8.2.1 Born-level Cross-Section

The Born-level cross-section for $e^+e^- \rightarrow Z/\gamma^* \rightarrow \text{hadrons}$ follows the standard LEP parameterisation [LEP EWWG, Phys. Rept. 427 (2006) 257] with an s -dependent width:

$$\sigma_{had}^{Born}(s) = \sigma_{had}^0 \cdot \frac{s\Gamma_Z^2}{(s - M_Z^2)^2 + s^2\Gamma_Z^2/M_Z^2} + \sigma_{int}(s) + \sigma_\gamma(s) \quad (13)$$

where:

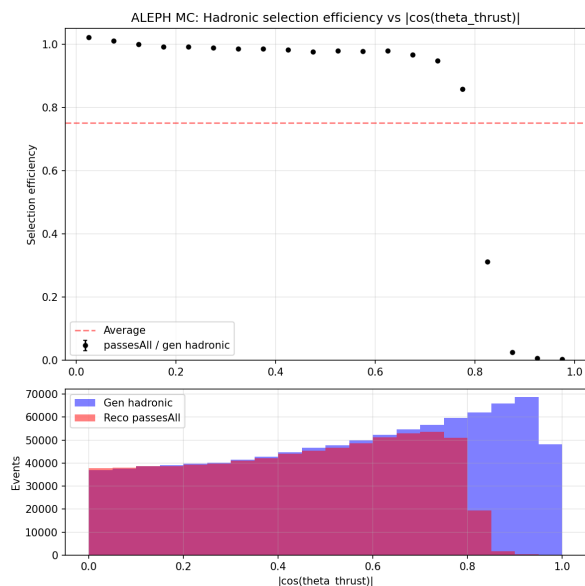
- The first term is the Z-exchange Breit-Wigner with running width $\Gamma(s) = \Gamma_Z \cdot s/M_Z^2$;
- $\sigma_{int}(s)$ is the γ -Z interference term, proportional to $(s - M_Z^2)/(\text{denominator})$, involving the product of electromagnetic and weak vector couplings;
- $\sigma_\gamma(s) = (4\pi\alpha^2)/(3s) \cdot R_\gamma$ is the pure photon-exchange contribution, with $R_\gamma = N_c \sum_f Q_f^2 = 11/3$ for five quark flavours.

The interference and pure-gamma terms are sub-percent corrections at the Z pole but become relevant at the off-peak scan points where $|s - M_Z^2|$ is large.

The three free parameters of the lineshape fit are:

- M_Z : the Z boson mass [GeV];
- Γ_Z : the Z boson total width [GeV];
- σ_{had}^0 : the hadronic pole cross-section [nb], defined as $\sigma_{had}^0 = (12\pi/M_Z^2) \cdot \Gamma_e\Gamma_{had}/\Gamma_Z^2$.

Selection Efficiency



Systematic Uncertainties

7.7 ISR Modelling

The lineshape model includes initial-state radiation through a convolution with the $\mathcal{O}(\alpha^2)$ radiator function (see Section 8.2). The uncertainty from higher-order ISR corrections ($\mathcal{O}(\alpha^3)$ and beyond) is taken from the published ALEPH analysis, which quotes $\delta\Gamma_Z = 0.2$ MeV and $\delta\sigma_{had}^0 = 0.010$ nb from ISR modelling.

Method: The published ALEPH values for the ISR modelling uncertainty are adopted directly, as the same radiator function formalism is used.

Result: $\delta M_Z = 0.00$ MeV, $\delta\Gamma_Z = 0.20$ MeV, $\delta\sigma_{had}^0 = 0.010$ nb.

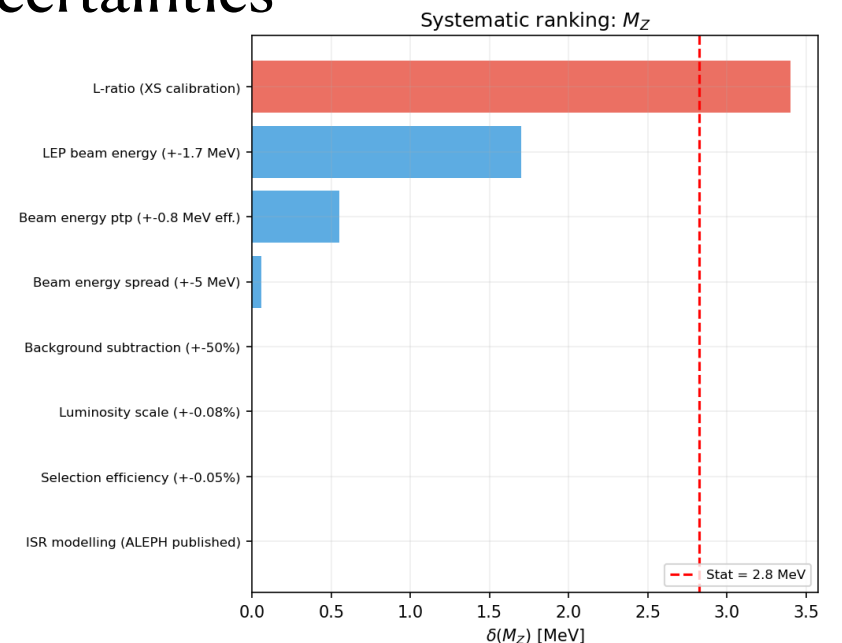
The ISR uncertainty has no effect on M_Z because the ISR convolution preserves the symmetry of the lineshape around the pole. It affects Γ_Z and σ_{had}^0 through the ISR correction to the peak cross-section.

7.8 Luminosity Ratios (Cross-Section Calibration)

This is the dominant systematic uncertainty specific to this reanalysis. As described in Section 8.3, the integrated luminosity ratios between energy points are determined using published ALEPH hadronic cross-sections as external calibration. The published cross-sections were measured using independent Bhabha-based luminosity from the SiCAL detector, making this procedure non-circular.

The luminosity ratio uncertainties are:

Ratio	Value	Uncertainty
$\mathcal{L}(P-2)/\mathcal{L}(\text{Peak})$	0.15511	± 0.00053 (0.34%)
$\mathcal{L}(\text{Intermediate})/\mathcal{L}(\text{Peak})$	0.00099	± 0.00001 (1.0%)
$\mathcal{L}(P+2)/\mathcal{L}(\text{Peak})$	0.15345	± 0.00042 (0.28%)



Not Half Bad

Fitting Plan

8.4 Chi-Square Fit

The Z resonance parameters are extracted by minimising the chi-square:

$$\chi^2(M_Z, \Gamma_Z, \sigma_{\text{had}}^0) = \sum_{i=1}^4 \left(\frac{\sigma_i^{\text{obs}} - \sigma_i^{\text{model}}(M_Z, \Gamma_Z, \sigma_{\text{had}}^0)}{\delta\sigma_i} \right)^2 \quad (18)$$

where:

- σ_i^{obs} is the measured cross-section at energy point i ;
- σ_i^{model} is the theoretical prediction from the full lineshape model (Born + ISR + BES);
- $\delta\sigma_i$ is the statistical uncertainty on σ_i^{obs} .

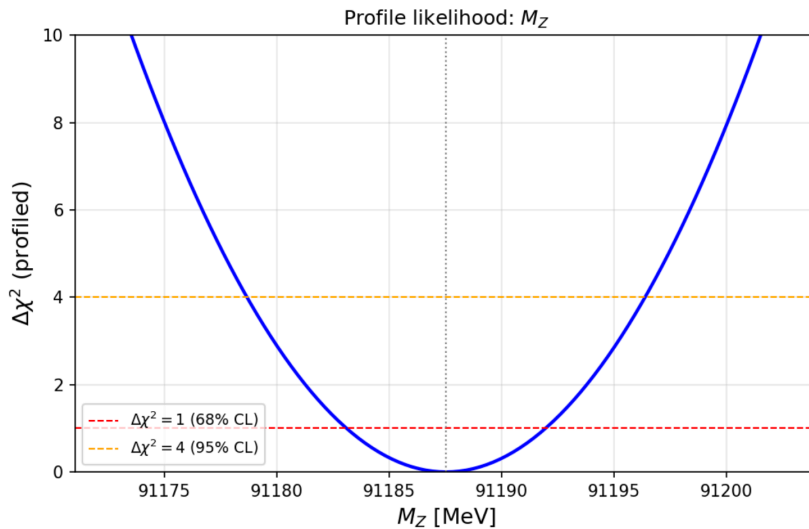
The minimisation is performed using the Nelder-Mead simplex algorithm (`scipy.optimize.minimize`) with convergence tolerances of 10^{-8} on the parameter values and 10^{-14} on the chi-square.

Statistical uncertainties on the fitted parameters are obtained from the inverse of the Hessian matrix of the chi-square at the minimum:

$$\text{Cov}(\hat{\theta}) = 2 \cdot [H(\chi^2)]^{-1} \quad (19)$$

where $H_{ij} = \partial^2 \chi^2 / \partial \theta_i \partial \theta_j$ is the Hessian, evaluated numerically using finite differences with step sizes $\Delta M_Z = \Delta \Gamma_Z = 10^{-5}$ GeV and $\Delta \sigma_{\text{had}}^0 = 5 \times 10^{-3}$ nb.

Profile Scans



Blinding Procedure

8.5 Blinding Procedure

This analysis follows a blinding protocol to prevent unconscious bias in the selection optimisation and systematic evaluation. The blinding procedure is as follows:

1. **Selection development** (Waves 1–2): The hadronic selection is based entirely on the pre-existing `passesAll` flag in the archived ntuples. No optimisation on data is performed. Cut efficiencies are compared between data and MC, but no lineshape fit to data is performed.
2. **Background estimation and validation** (Wave 3): Background fractions are estimated from published cross-sections and MC acceptances. Data-MC comparisons are performed for kinematic distributions, but no cross-sections are computed from data.
3. **Systematic evaluation** (Wave 4): All systematic uncertainties are evaluated using an **Asimov dataset** — a pseudo-dataset constructed from the Standard Model predictions at each energy point with the PDG values of M_Z , Γ_Z , and σ_{had}^0 . The Asimov event counts are:

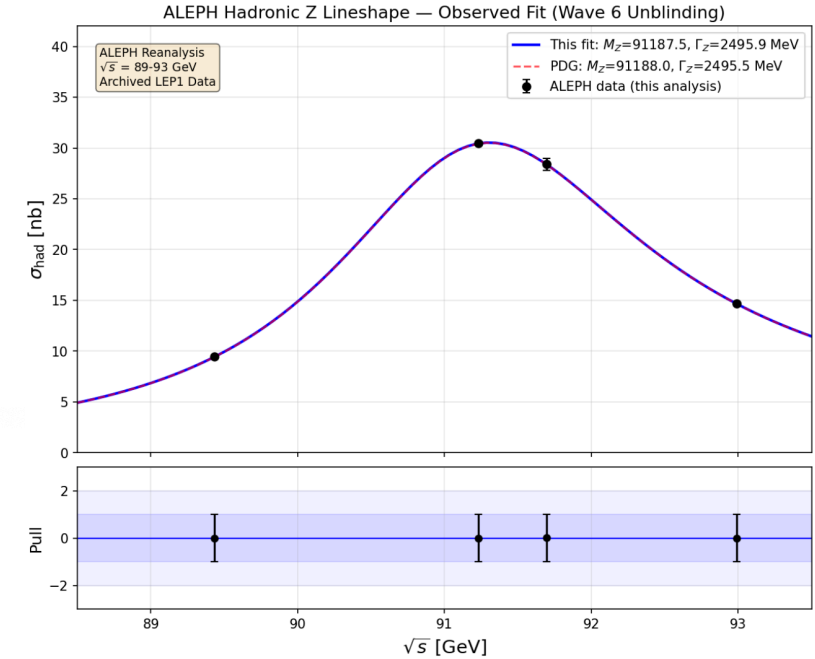
$$N_i^{\text{Asimov}} = \varepsilon_i \times \mathcal{L}_i \times \sigma_{\text{had}}^{\text{SM}}(\sqrt{s}_i) \quad (20)$$

1. **Pre-unblinding review** (Wave 5): The analysis note is drafted using Asimov results. The full analysis chain is frozen: selection, backgrounds, efficiencies, systematics, statistical model.
2. **Unblinding** (Wave 6): The lineshape fit is performed on the actual data cross-sections. This constitutes the unblinding. No changes to the analysis are permitted after unblinding (see below).

Expected Sensitivity

Parameter	This analysis (expected)	ALEPH published	LEP combined
δM_Z (stat)	2.8	2.5	—
δM_Z (syst)	3.8	1.7	—
δM_Z (total)	4.8	3.1	2.1
$\delta \Gamma_Z$ (stat)	4.3	4.3	—
$\delta \Gamma_Z$ (syst)	5.1	1.3	—
$\delta \Gamma_Z$ (total)	6.6	4.5	2.3
$\delta \sigma_{\text{had}}^0$ (total)	0.055	0.057	0.037

Fit to Data

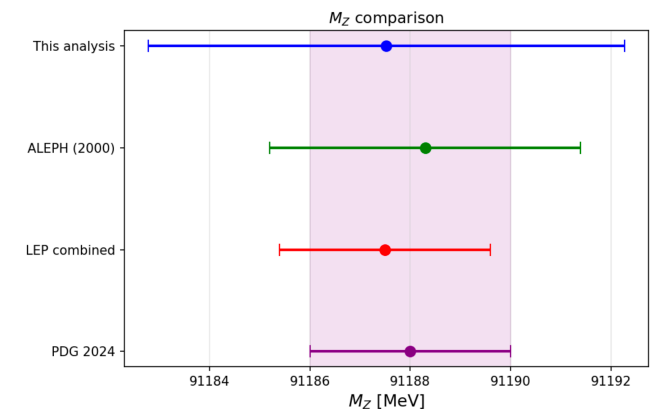


Fit Results & Comparison

$$M_Z = 91187.52 \pm 2.90 \text{ (stat)} \pm 3.84 \text{ (syst)} \text{ MeV} = 91187.52 \pm 4.81 \text{ (total)} \text{ MeV}$$

$$\Gamma_Z = 2495.94 \pm 4.17 \text{ (stat)} \pm 5.05 \text{ (syst)} \text{ MeV} = 2495.94 \pm 6.55 \text{ (total)} \text{ MeV}$$

$$\sigma_{\text{had}}^0 = 41.507 \pm 0.029 \text{ nb}$$



Not Half Bad

Fitting Plan

8.4 Chi-Square Fit

The Z resonance parameters are extracted by minimising the chi-square:

$$\chi^2(M_Z, \Gamma_Z, \sigma_{\text{had}}^0) = \sum_{i=1}^4 \left(\frac{\sigma_i^{\text{obs}} - \sigma_i^{\text{model}}(M_Z, \Gamma_Z, \sigma_{\text{had}}^0)}{\delta\sigma_i} \right)^2 \quad (18)$$

where:

- σ_i^{obs} is the measured cross-section at energy point i ;
- σ_i^{model} is the theoretical prediction from the full lineshape model (Born + ISR + BES);
- $\delta\sigma_i$ is the statistical uncertainty on σ_i^{obs} .

The minimisation is performed using the Nelder-Mead simplex algorithm (`scipy.optimize.minimize`) with convergence tolerances of 10^{-8} on the parameter values and 10^{-14} on the chi-square.

Statistical uncertainties on the fitted parameters are obtained from the inverse of the Hessian matrix of the chi-square at the minimum:

$$\text{Cov}(\hat{\theta}) = 2 \cdot [H(\chi^2)]^{-1} \quad (19)$$

where $H_{ij} = \partial^2 \chi^2 / \partial \theta_i \partial \theta_j$ is the Hessian, evaluated numerically using finite differences with step sizes $\Delta M_Z = \Delta \Gamma_Z = 10^{-5}$ GeV and $\Delta \sigma_{\text{had}}^0 = 5 \times 10^{-3}$ nb.

Blinding Procedure

8.5 Blinding Procedure

This analysis follows a blinding protocol to prevent unconscious bias in the selection optimisation and systematic evaluation. The blinding procedure is as follows:

1. **Selection development** (Waves 1–2): The hadronic selection is based entirely on the pre-existing `passesAll` flag in the archived ntuples. No optimisation on data is performed. Cut efficiencies are compared between data and MC, but no lineshape fit to data is performed.

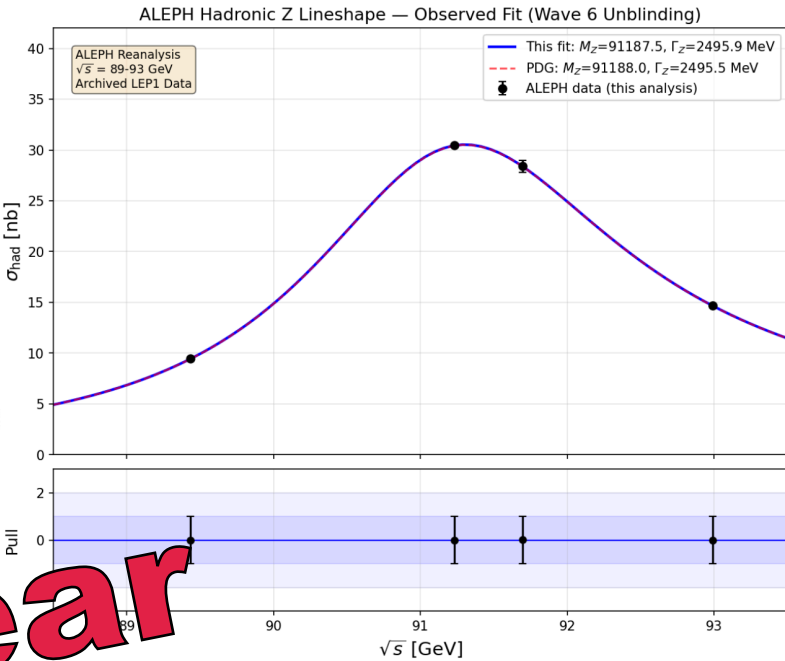
2. **Background estimation and validation** (Wave 3): Backgrounds are estimated from published cross-sections and Monte Carlo (MC) simulations. No optimisations are performed for kinematic distributions, but no cross-sections are optimised.

3. **Systematic evaluation** (Wave 4): All systematic uncertainties are evaluated using an Asimov dataset — a pre-generated dataset constructed from the Standard Model predictions for the Z resonance, with the energy point $\sqrt{s}$ fixed to the PDG values of M_Z , Γ_Z , and σ_{had}^0 . The Asimov event count is $N_{\text{Asimov}} = \mathcal{L}_{\text{int}} \times \sigma_{\text{had}}^{\text{SM}}(\sqrt{s})$.

$$N_{\text{Asimov}} = \mathcal{L}_{\text{int}} \times \sigma_{\text{had}}^{\text{SM}}(\sqrt{s}) \quad (20)$$

1. **Pre-unblinding** (Wave 5): The analysis note is drafted using Asimov results. This constitutes the pre-unblinding. No changes to the analysis are permitted after pre-unblinding.
2. **Unblinding** (Wave 6): The lineshape fit is performed on the actual data cross-sections. This constitutes the unblinding. No changes to the analysis are permitted after unblinding (see below).

Fit to Data



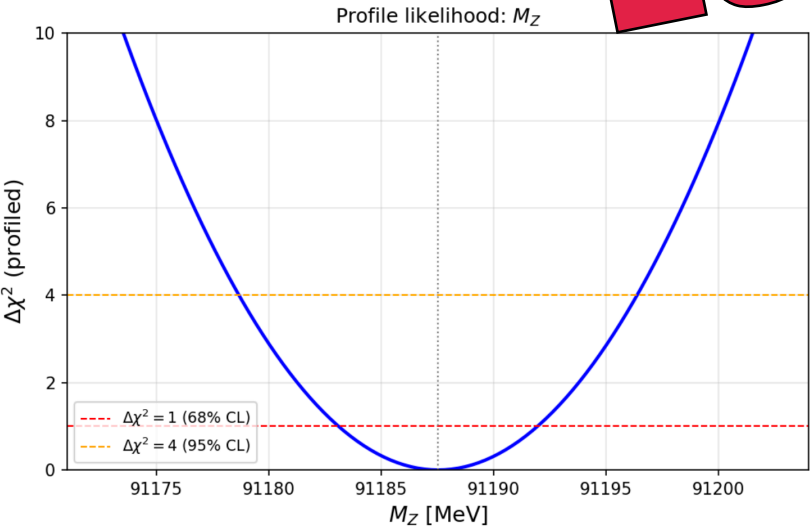
Fit Results & Comparison

$$M_Z = 91187.52 \pm 2.90 \text{ (stat)} \pm 3.84 \text{ (syst)} \text{ MeV} = 91187.52 \pm 4.81 \text{ (total)} \text{ MeV}$$

$$\Gamma_Z = 2495.94 \pm 4.17 \text{ (stat)} \pm 5.05 \text{ (syst)} \text{ MeV} = 2495.94 \pm 6.55 \text{ (total)} \text{ MeV}$$

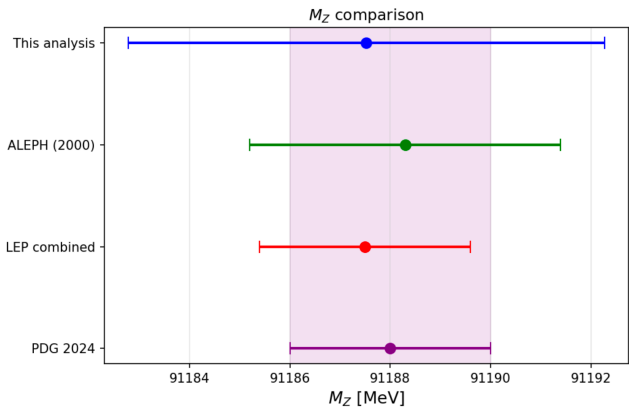
$$\sigma_{\text{had}}^0 = 41.507 \pm 0.029 \text{ nb}$$

Profile Scans



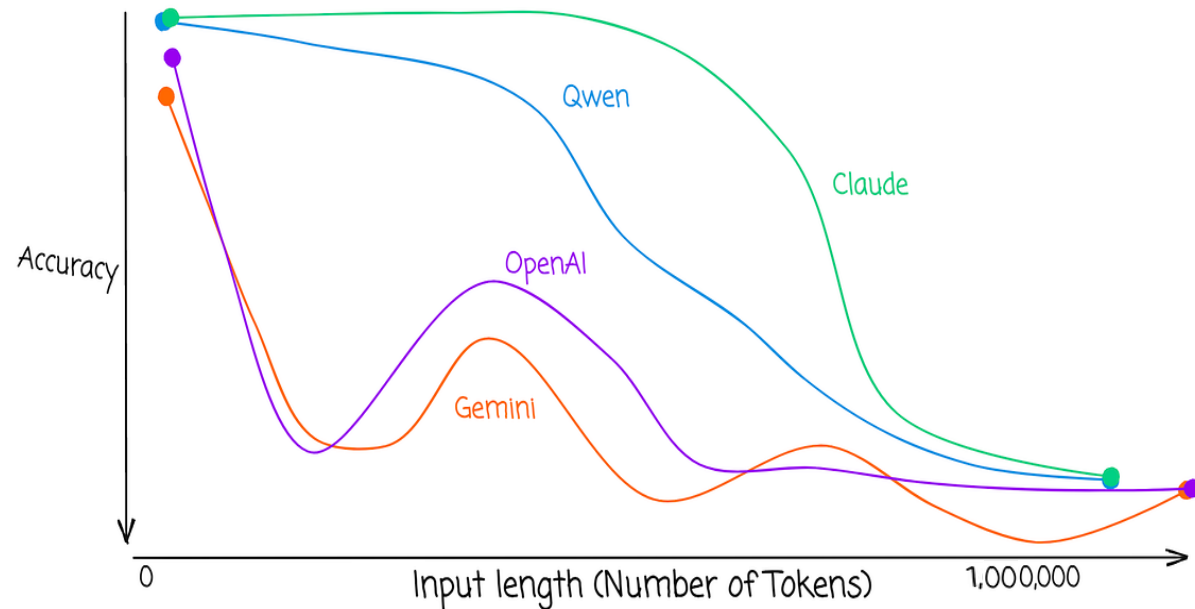
Expected Sensitivity

Parameter	This analysis (expected)	ALEPH published	LEP combined
δM_Z (stat)	2.8	2.5	—
δM_Z (syst)	3.8	1.7	—
δM_Z (total)	4.8	3.1	2.1
$\delta \Gamma_Z$ (stat)	4.3	4.3	—
$\delta \Gamma_Z$ (syst)	5.1	1.3	—
$\delta \Gamma_Z$ (total)	6.6	4.5	2.3
$\delta \sigma_{\text{had}}^0$ (total)	0.055	0.057	0.037



Problems to Solve

Context Rot

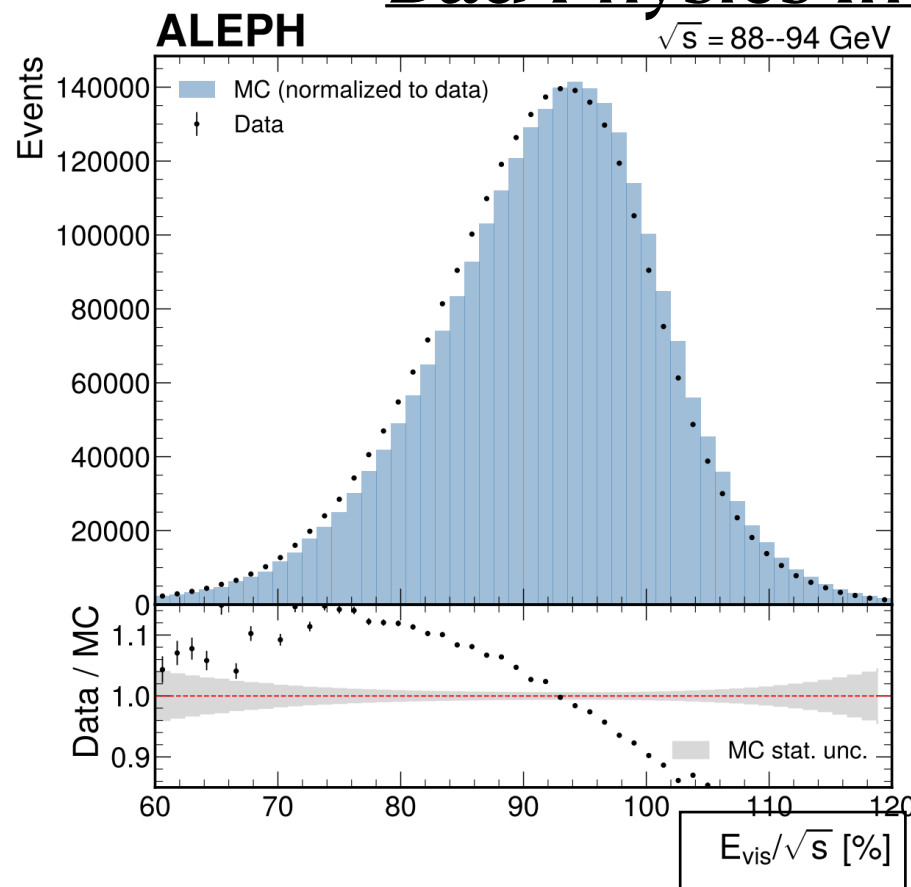


Data Leakage

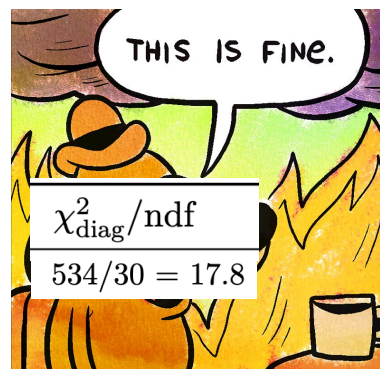
Measurement of the Z Resonance Parameters at LEP The ALEPH Collaboration

The properties of the Z resonance are measured from the analysis of 4.5 million Z decays into fermion pairs collected with the ALEPH detector at LEP. The data are consistent with lepton universality. The resonance parameters are measured to be $M_Z = (91.1885 \pm 0.0031) \text{ GeV}/c^2$, $\Gamma_Z = (2.4951 \pm 0.0043) \text{ GeV}$, $\sigma_{\text{had}}^0 = (41.559 \pm 0.058) \text{ nb}$ and, combining the three lepton flavours, $R_\ell = 20.725 \pm 0.039$. The corresponding number of light neutrino species is $N_\nu = 2.983 \pm 0.013$ and the strong coupling constant is $\alpha_s(M_Z) = 0.114 \pm 0.004 \pm 0.002_{\text{QCD}} + 0.005 \log_{10} \left[\frac{M_H}{150 \text{ GeV}/c^2} \right]$. The lepton pair forward-backward asymmetry is measured to be $A_{\text{FB}}^{0,\ell} = 0.0173 \pm 0.0016$ from which the effective weak mixing angle is derived: $\sin^2 \theta_{\text{eff}}^{\text{lept}} = 0.23089 \pm 0.00089$. The measurement of the leptonic width $\Gamma_{\ell\ell} = 84.02 \pm 0.15 \text{ MeV}$ leads to a determination of the effective ρ parameter $\rho_{\text{eff}}^{\text{lept}} = 1.0064 \pm 0.0018$. The data support the Standard Model and favour a light Higgs.

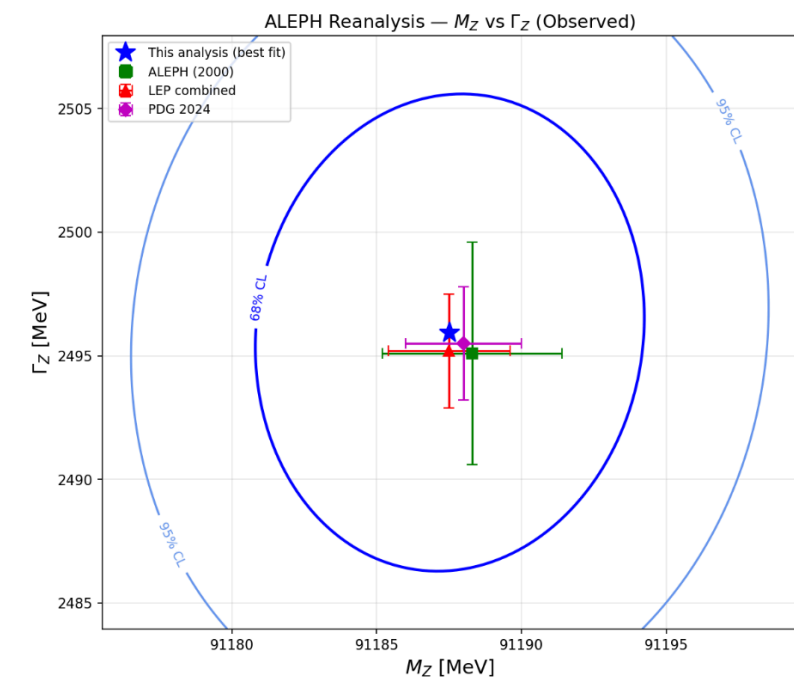
Bad Physics Intuition?



“The distribution demonstrates excellent calorimetric performance of the ALEPH detector, with data and MC agreeing within 2%”

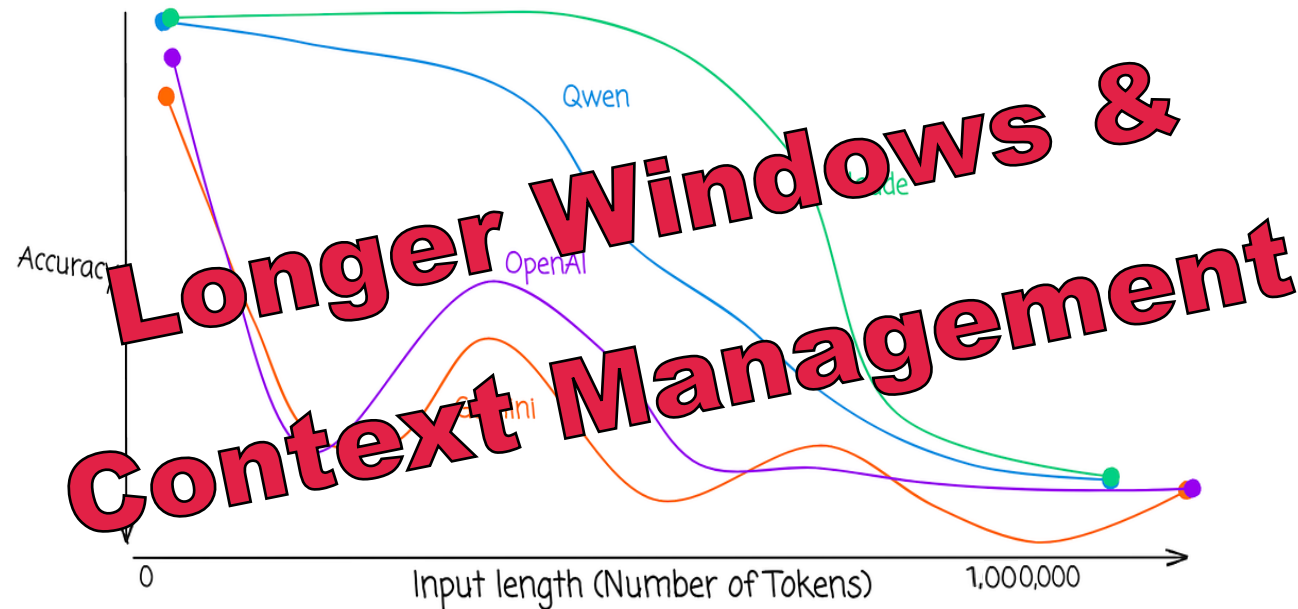


Worse Performance / Laziness



Problems to Solve

Context Rot



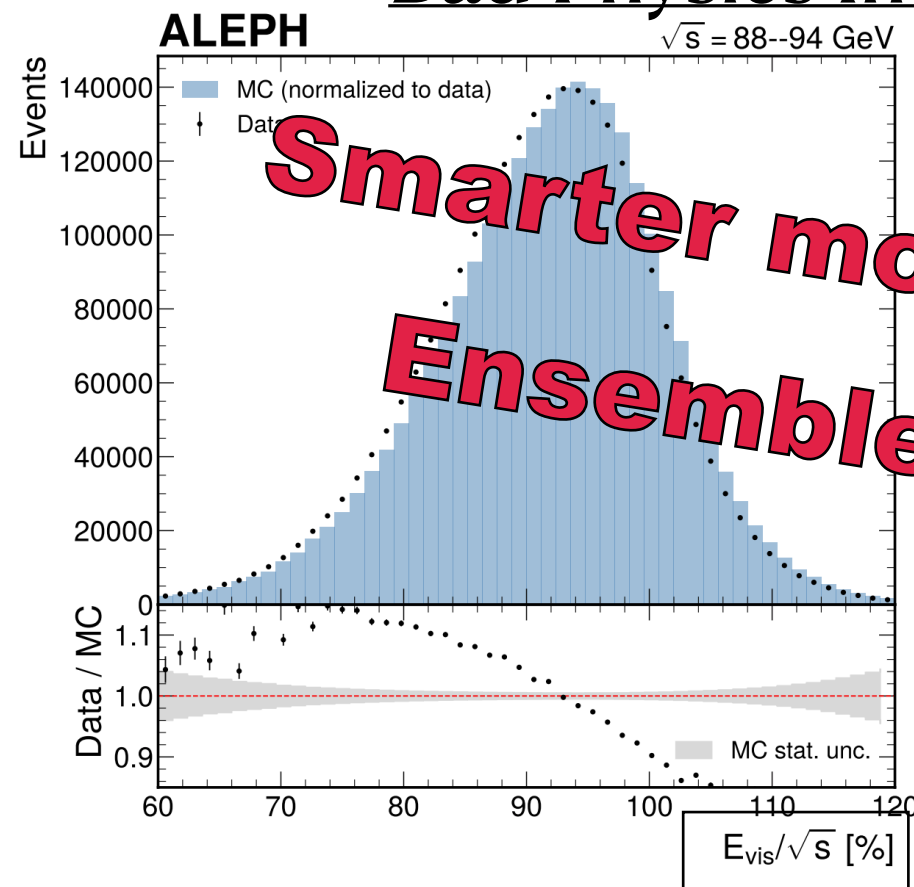
Data Leakage

Measurement of the Z Resonance Parameters at LEP

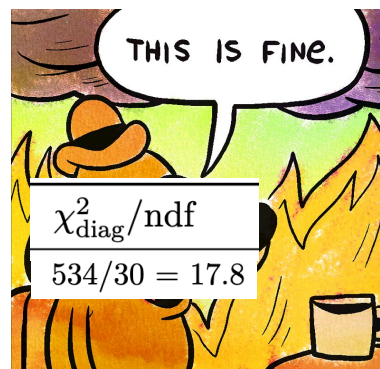
The ALEPH Collaboration

The properties of the Z resonance are measured from the analysis of 4.5 million Z decays into fermion pairs collected with the ALEPH detector at LEP. The data are consistent with lepton universality. The resonance parameters are measured to be $M_Z = (91.1885 \pm 0.0031) \text{ GeV}/c^2$, $\Gamma_Z = (2.4951 \pm 0.0043) \text{ GeV}$, $\sigma_{\text{had}}^0 = (41559 \pm 2158) \text{ nb}$ and, combining the three lepton flavours, $R_\ell = 20.725 \pm 0.039$. The corresponding number of light neutrino species is $N_\nu = 2.983 \pm 0.013$ and the strong coupling constant is $\alpha_s(M_Z) = 0.118 \pm 0.004 \pm 0.002_{\text{QCD}} + 0.005 \log_{10} \left[\frac{M_H}{150 \text{ GeV}/c^2} \right]$. The lepton pair forward-backward asymmetry is measured to be $A_{\text{FB}}^{0,\ell} = 0.0173 \pm 0.0016$ from which the effective weak mixing angle is derived: $\sin^2 \theta_{\text{eff}}^{\text{lept}} = 0.23089 \pm 0.00089$. The measurement of the leptonic width $\Gamma_{\ell\ell} = 84.02 \pm 0.15 \text{ MeV}$ leads to a determination of the effective ρ parameter $\rho_{\text{eff}}^{\text{lept}} = 1.0064 \pm 0.0018$. The data support the Standard Model and favour a light Higgs.

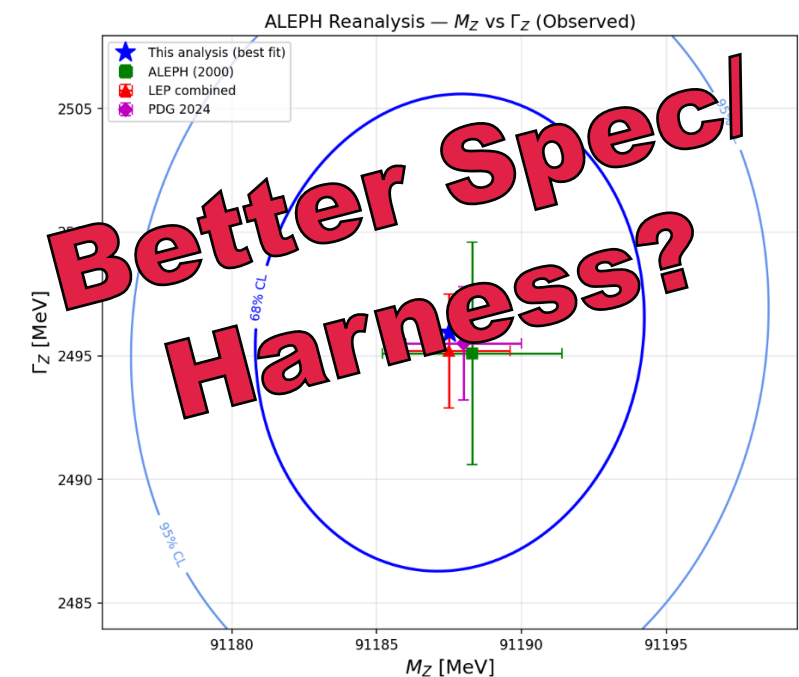
Bad Physics Intuition?



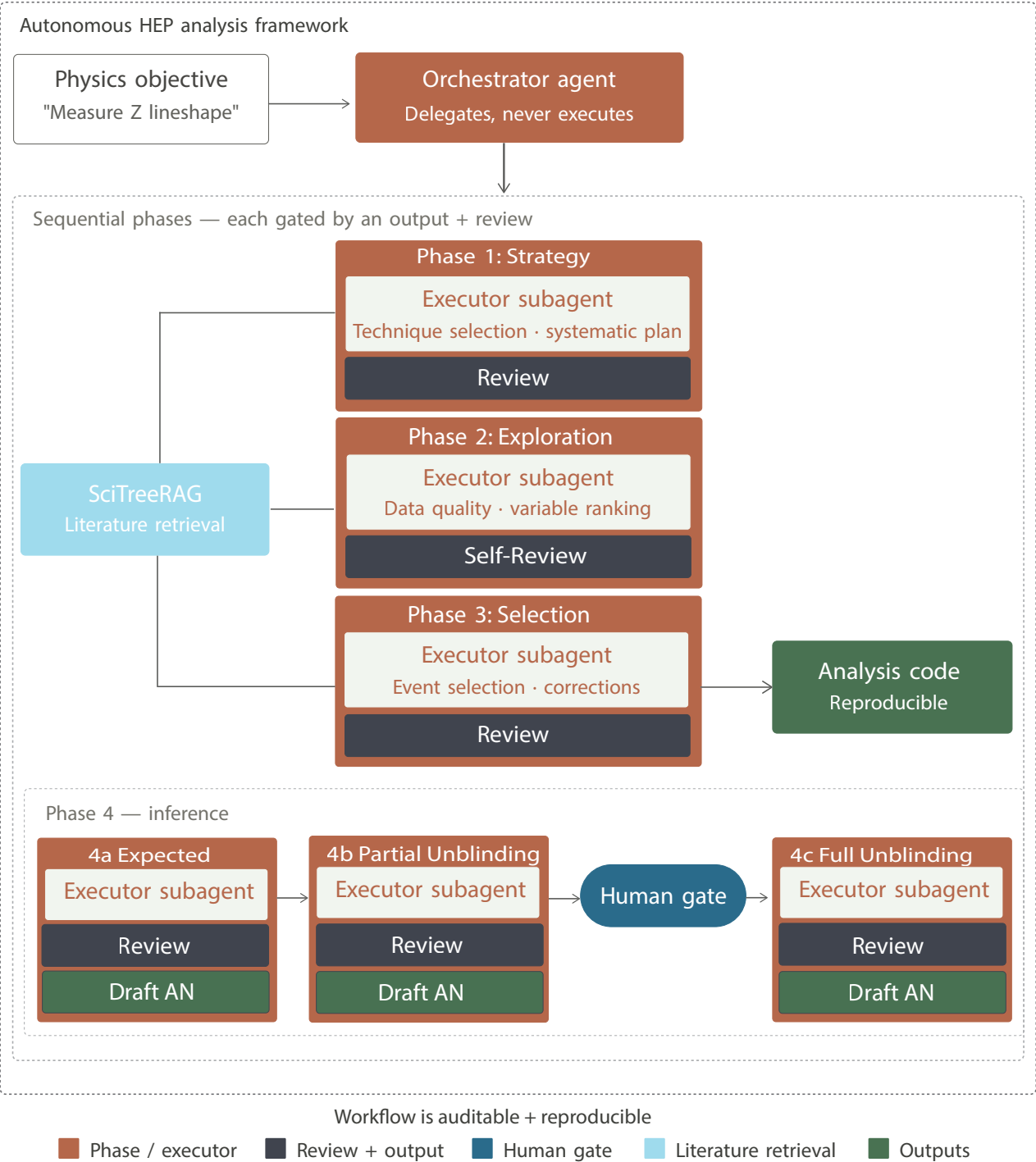
"The distribution demonstrates excellent calorimetric performance of the ALEPH detector, with data and MC agreeing within 2%"



Worse Performance / Laziness



JFC: Just Furnish Context



Analysis Agents

lead-analyst, data-explorer, detector-specialist, theory-scout, signal-lead, background-estimator,

Literature Retrieval

SciTreeRAG¹: Indexes a corpus of published LEP (ALEPH, DELPHI) papers, to build contextual retrievals.

¹<https://arxiv.org/abs/2509.06855>

arXiv MCP: Query published arXiv papers, with answers to prompted questions.

INSPIRE-HEP/HEPData: queries for recent measurements and

Reviewer agents

physics-reviewer, critical-reviewer, constructive-reviewer, plot-validator, bibtex-validator, arbiter and investigator

Phases (# Reviews)

Phase 1 — Strategy (4)
Phase 2 — Exploration (1)
Phase 3 — Selection and Background Modeling
Phase 4a — Expected Results (4)
Phase 4b — Partial Unblinding (4)

Blinding Protocol

1. Background model validated
2. Systematic uncertainties evaluated and fit stable.
3. Expected results physically sensible
4. Signal injection tests confirm fit recovers injected signals,
5. 10% partial unblinding shows no unexpected pathologies
6. All agent review cycles resolved (arbiter PASS)
7. Human gate: APPROVE, ITERATE, HALT
8. Full Unblinding: full statistical fit on the complete dataset

Spec & Prompting

- “Spec” is created by repeatedly injecting physics analysis expert knowledge
- Models are smart enough to figure out most things, just need structure.

Accumulated domain knowledge

conventions

- extraction.md
- lint_plots.py
- postprocess_tex.py
- preamble.tex
- README.md
- search.md
- TEMPLATE.md
- unfolding.md

Role definitions

agents

- arbiter.md
- bibtex_validator.md
- constructive_reviewer.md
- critical_reviewer.md
- executor.md
- fixer.md
- investigator.md
- note_writer.md
- physics_reviewer.md
- plot_validator.md

Analysis Rulebook

methodology

- 01-principles.md
- 02-inputs.md
- 03a-orchestration.md
- 03-phases.md
- 04-blinding.md
- 05-artifacts.md
- 06-review.md
- 07-tools.md
- 09-multichannel.md
- 11-coding.md
- 12-downscoping.md
- analysis-note.md
- appendix-automation.md
- appendix-checklist.md
- appendix-dependencies.md
- appendix-heuristics.md
- appendix-integration.md
- appendix-plotting.md
- appendix-prompts.md
- appendix-sessions.md
- README.md

Scaffolding blueprints

templates

- doc4a_claude.md
- doc4b_claude.md
- doc4c_claude.md
- phase1_claude.md
- phase2_claude.md
- phase3_claude.md
- phase4a_claude.md
- phase4b_claude.md
- phase4c_claude.md
- pixi.toml
- root_claude.md

Provided Toolkits

Task	Required tool
ROOT file I/O	uproot [21] (≥5.0)
Array operations	awkward-array [22], numpy [23]
Histogramming	hist [24], boost-histogram [25]
Plotting	matplotlib [26] + mplhep [27]
Statistical modeling	pyhf [28] (binned), zfit [29] (unbinned)
4-vectors	vector [30], particle [31]
Document preparation	LaTeX + tectonic
Environment management	pixi [32]

Spec & Prompting

- “Spec” is created by repeatedly injecting physics analysis expert knowledge
- Models are smart enough to figure out most things, just need structure.

Bad Prompt (Leave it all to LLM)

```
# Analysis Title: H→ττ Analysis with CMS Open Data (μτ_h final state)

# You are performing a Higgs boson search in the τ+τ- decay channel using CMS Open Data from 2012 at √s = 8 TeV. The final state is one muon and one hadronically decaying tau lepton (μτ_h). Your goal is to produce distributions of key observables – particularly the visible di-tau mass – showing the Higgs signal contribution on top of Standard Model backgrounds.
```

- Found that expert input at initiation is crucial - **LLMs struggle with the “higher-level” tasks**
- You become the higher-level thinker rather than the code monkey!

Good Prompt (Inject Physics Knowledge)

```
# Analysis Title: H→ττ Analysis with CMS Open Data (μτ_h final state)

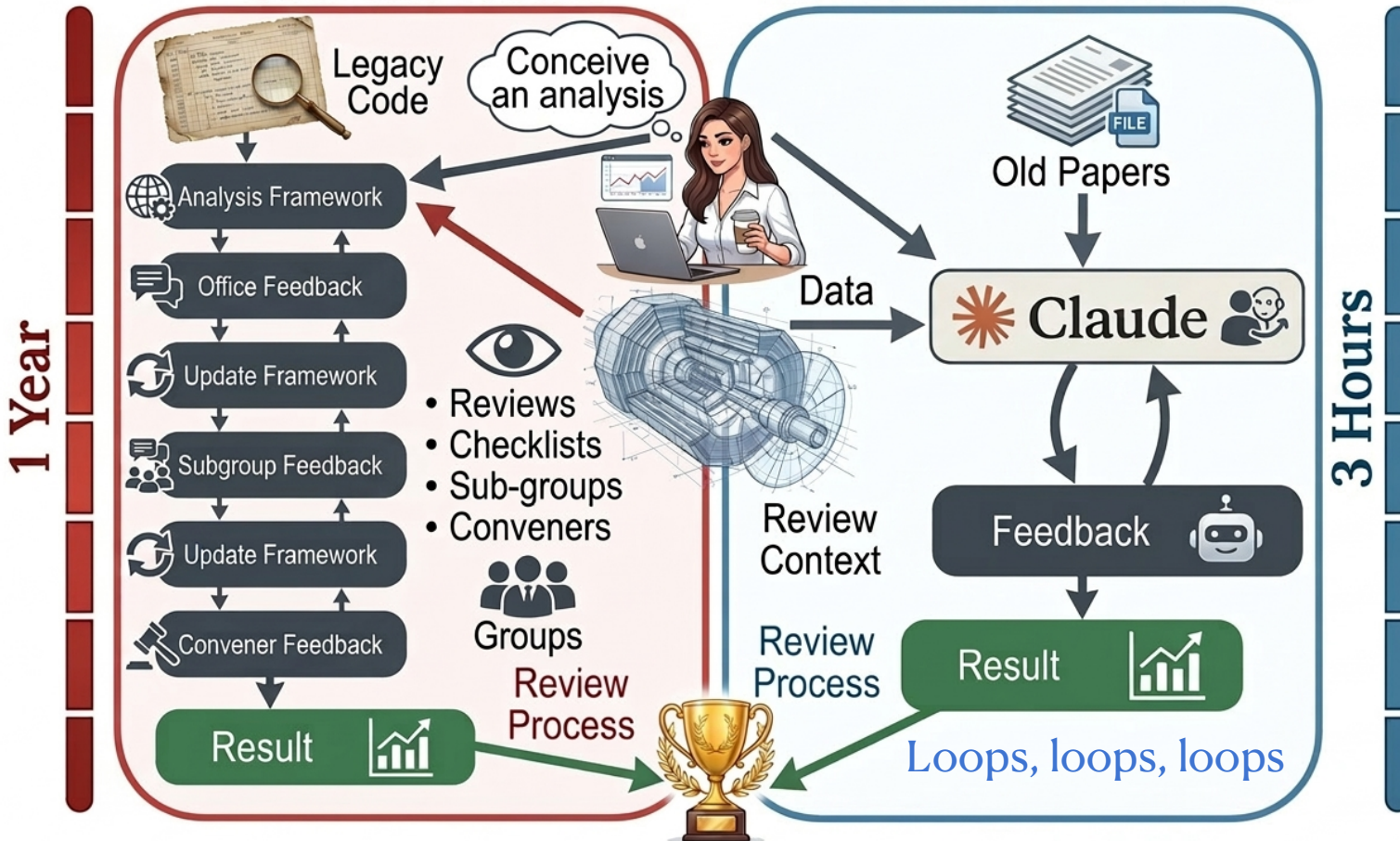
# You are performing a Higgs boson search in the τ+τ- decay channel using CMS Open Data from 2012 at √s = 8 TeV. The final state is one muon and one hadronically decaying tau lepton (μτ_h). Your goal is to produce distributions of key observables – particularly the visible di-tau mass – showing the Higgs signal contribution on top of Standard Model backgrounds. This loosely follows the official CMS publication (Phys. Lett. B 779 (2018) 283 and JHEP 05 (2014) 104). To optimize the higgs signal selection, do some categorization, in particular add a VBF category, be sure to fit all categories simultaneously. In addition our dataset is missing the full trigger and tau efficiency scale factors, please loosen the tau efficiency selection to 10 to 15% to allow for a good good agreement of the drell-yan peak. In addition to a baseline analysis, we would like 3 more approaches that focus on fitting a different final observable these include a. perform the analysis fitting an fit NN discriminator to find the higgs propagating the systematic uncertainties in the fit b. train a NN to regress the direction of the genMET and genMET phi in the final state and use this to make a combined mass with the visible objects that you use as the final analysis, and finally c. fit the mass adding the missing energy to the mass distribution. Also, make sure to do a tight anti muon veto, and to normalize the W+jets sample from data using the high mT region. Finally, make sure you put a tight anti-muon veto on the hadronic tau id. Also make sure you Z normalization uncertainty is larger than the MC predictions, at a level of 10–15% to reflect missing trigger turn on scale factors and a larger tau efficiency scale factor
```

JFC: Just Furnish Context

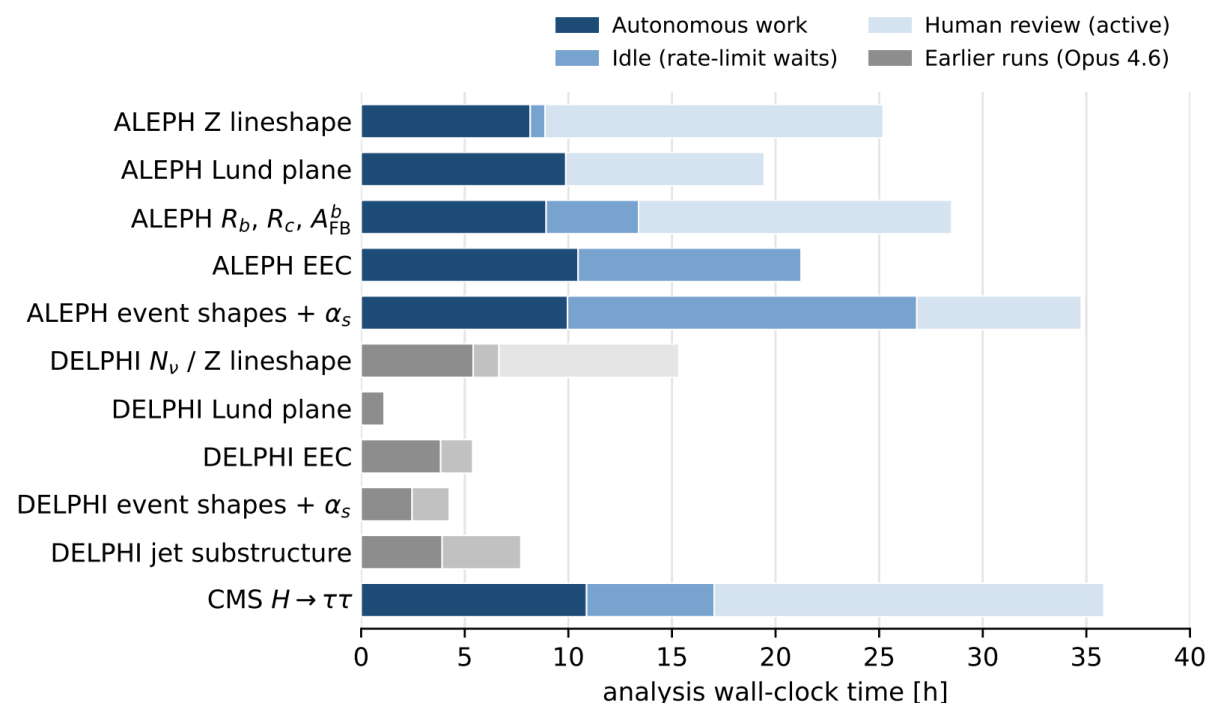
High Level Scheme

Current World

AI-Assisted Workflow



Analysis	Initial prompt (abridged)	Data	Time	Ref.
$H \rightarrow \tau\tau$	Measure the Higgs boson signal strength in the $\mu\tau_h$ final state at $\sqrt{s} = 8$ TeV	CMS	6h 30m	A/Github
Z lineshape + α_s	Measure Z lineshape parameters (M_Z , Γ_Z , σ_{had}^0), N_ν from invisible width, and extract $\alpha_s(M_Z)$ from R_ℓ	ALEPH	4h 58m	B/Github
Lund plane	Measure the primary Lund jet plane density in hadronic Z decays	ALEPH	7h 14m	C/Github
R_b , R_c , and A_{FB}^b	Measure R_b , R_c , and A_{FB}^b using signed impact parameter b-tagging and jet charge	ALEPH	4h 06m	D/Github
Energy-energy correlators	Measure the two-point EEC, its asymmetry (AEEC), and mean charged multiplicity in hadronic Z decays	ALEPH	3h 19m	E/Github
N_ν from Γ_{inv}	Measure the number of light neutrino generations from the Z invisible width	DELPHI	5h 15m	F/Github
Lund jet plane	Measure the primary Lund jet plane density in hadronic Z decays	DELPHI	6h 59m	G/Github
Energy-energy correlators	Measure the two-point energy-energy correlator in hadronic Z decays	DELPHI	5h 23m	H/Github
Jet substructure	Measure groomed and ungroomed jet substructure observables in hadronic Z decays	DELPHI	7h 42m	Github
Event shapes + α_s	Measure six event shape distributions and extract $\alpha_s(M_Z)$ from NLO QCD fits	DELPHI	4h 15m	Github



All in 1 week! X5 iterations
~\$300 worth of licenses

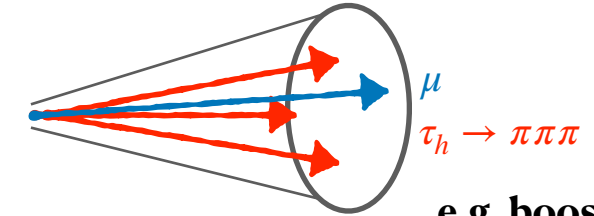
Better models think longer
Still <24 hours per analysis

E.g. $H \rightarrow \tau\tau$ Open Data Analysis

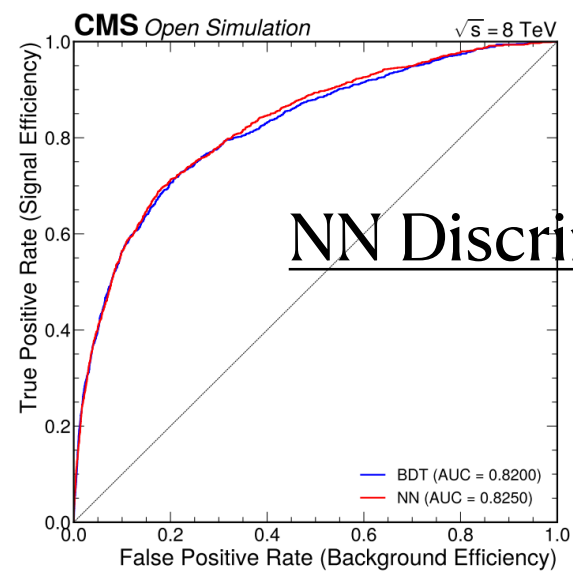
Topology-Driven Categorization: 1 jet, 3 jet, VBF event topologies*

3 different methods: m_{vis} , m_{col} , NN Score

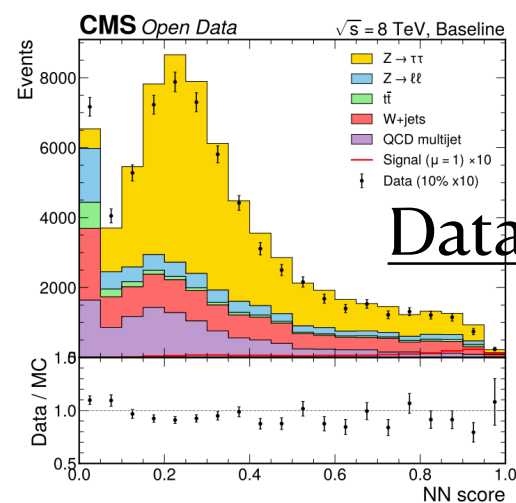
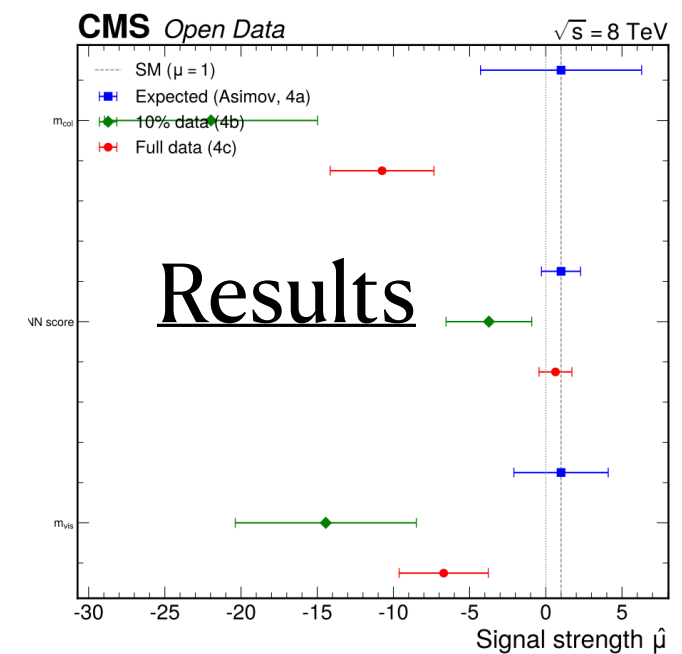
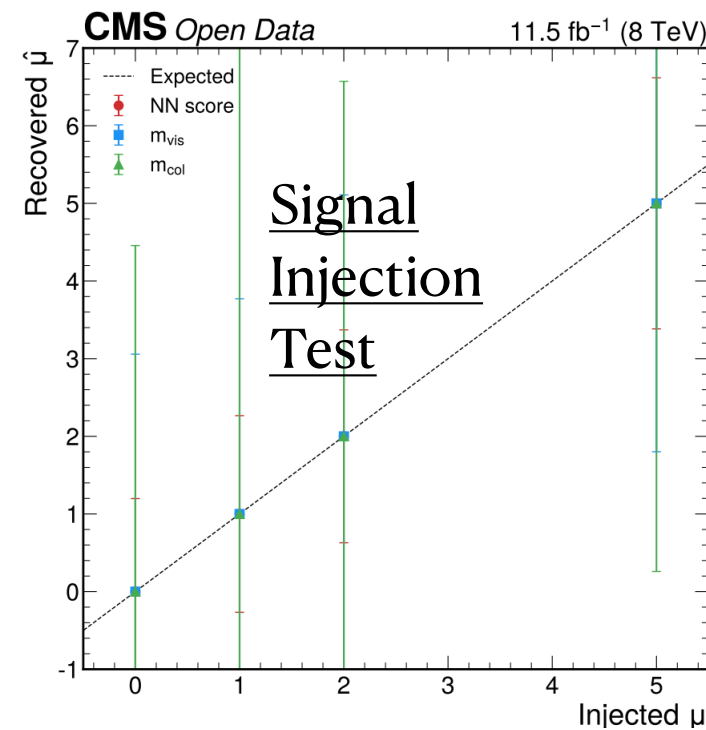
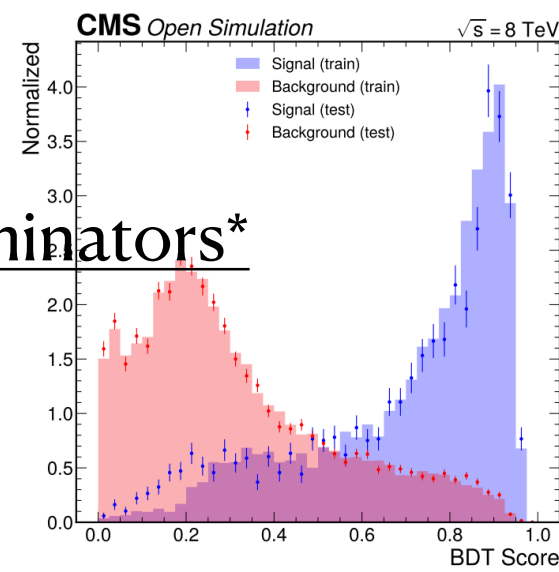
*Similar strategy to ATLAS 7/8TeV!



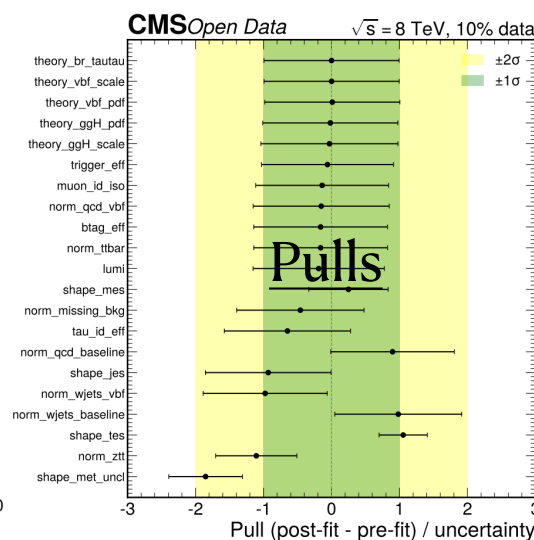
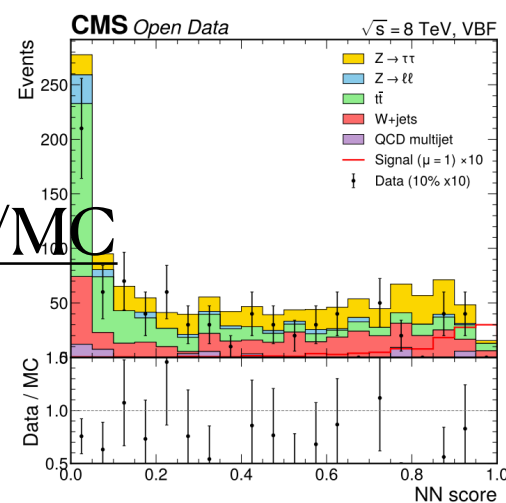
e.g. boosted $\mu\tau_h$



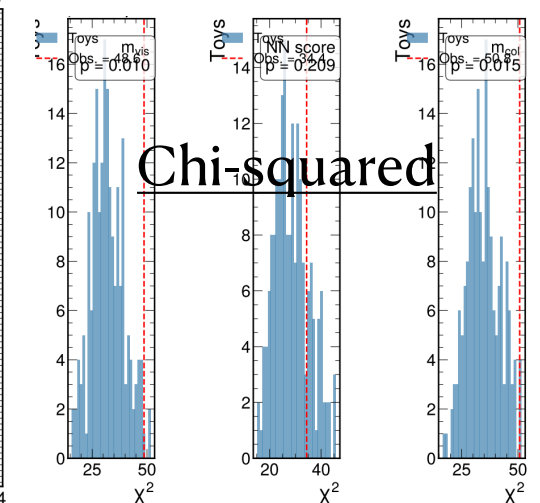
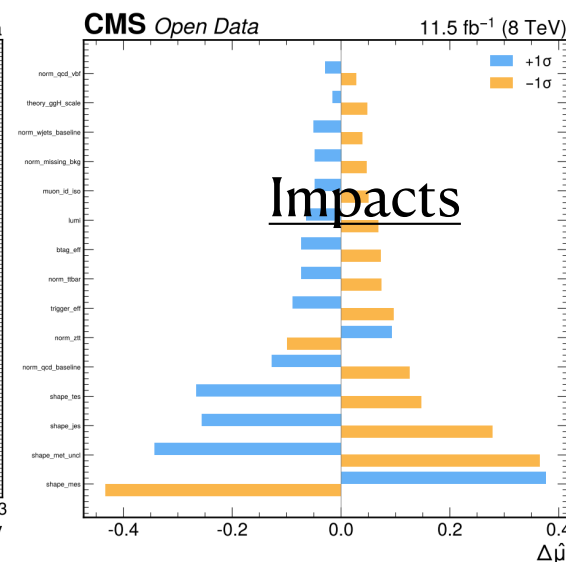
NN Discriminators*



Data/MC



Pulls

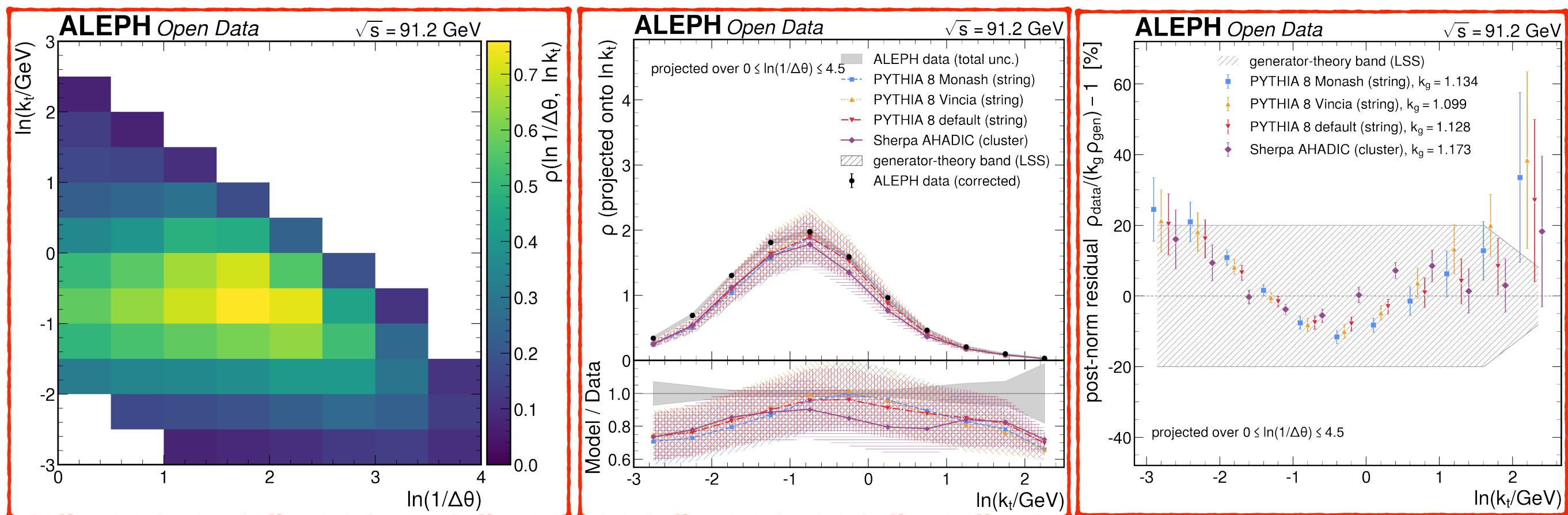


E.g. Lund Jet Plane Measurement @ ALEPH

First-ever e^+e^- Lund jet plane: 1.3M ALEPH $Z \rightarrow q\bar{q}$ events, thrust hemispheres, C/A declustering, 2D Bayesian unfolding*

Clean pure-quark (C_F) reference: no UE / MPI / pileup $\rightarrow$ soft wide-angle corner is pure QCD + hadronization, traces $\alpha_s(k_t)$ running

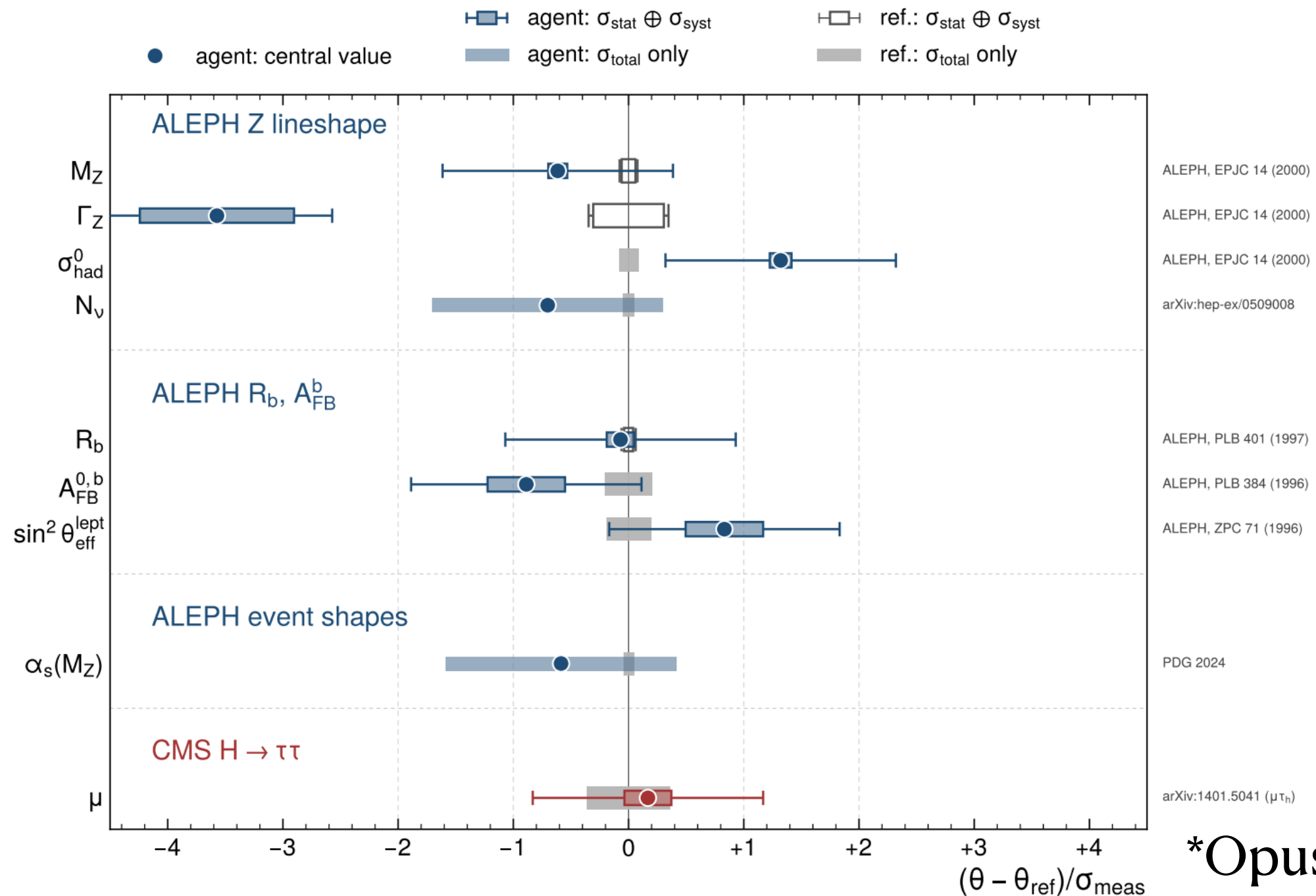
New generator benchmark: PYTHIA 8 & Sherpa all 9–18% low; mostly a normalization ($k_g \approx 1.1$ – 1.2), Sherpa cluster model worst



How Good is it*?

Not as precise as original analyses. However, able to recover reference analysis with uncertainty a majority of the time

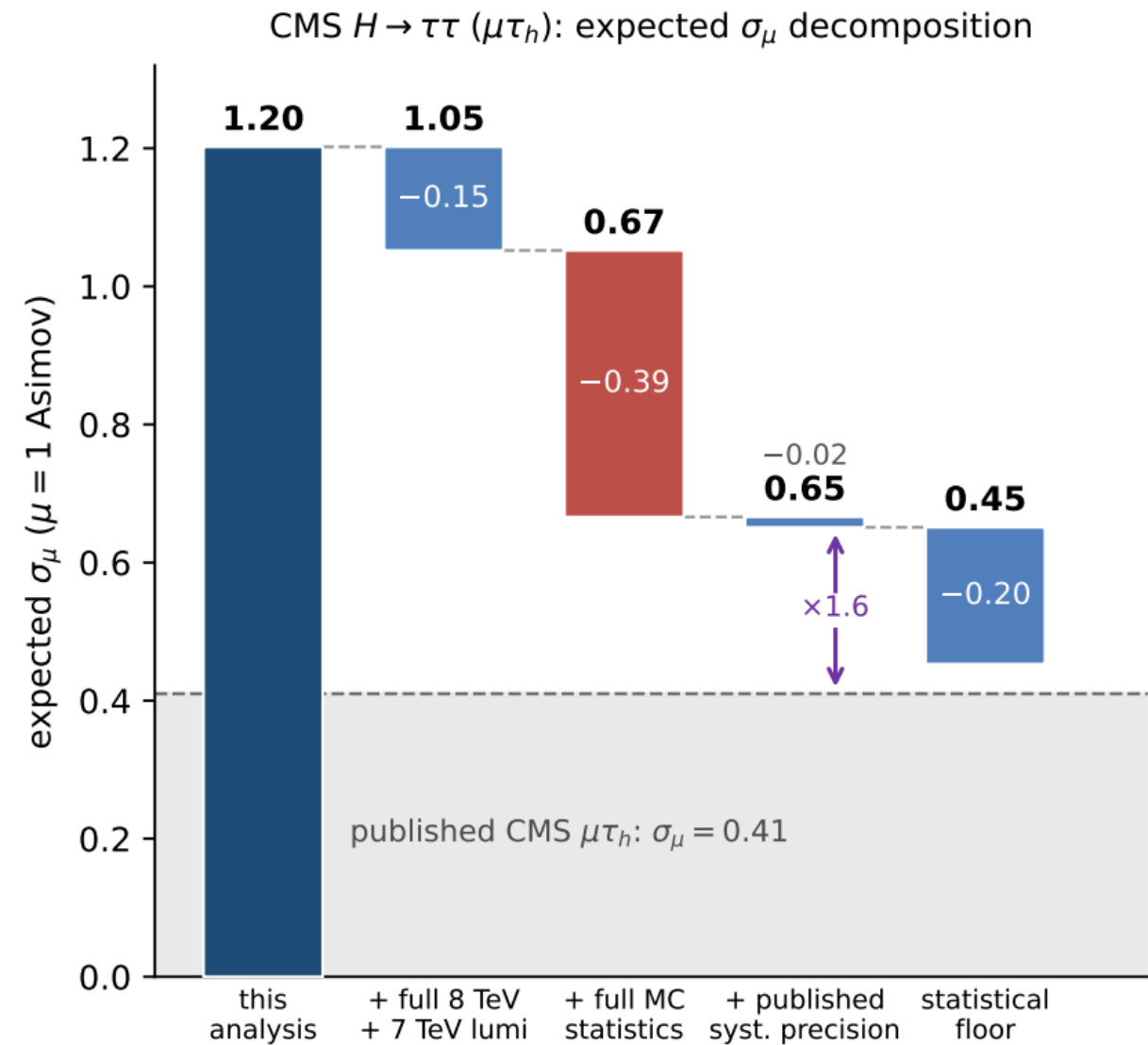
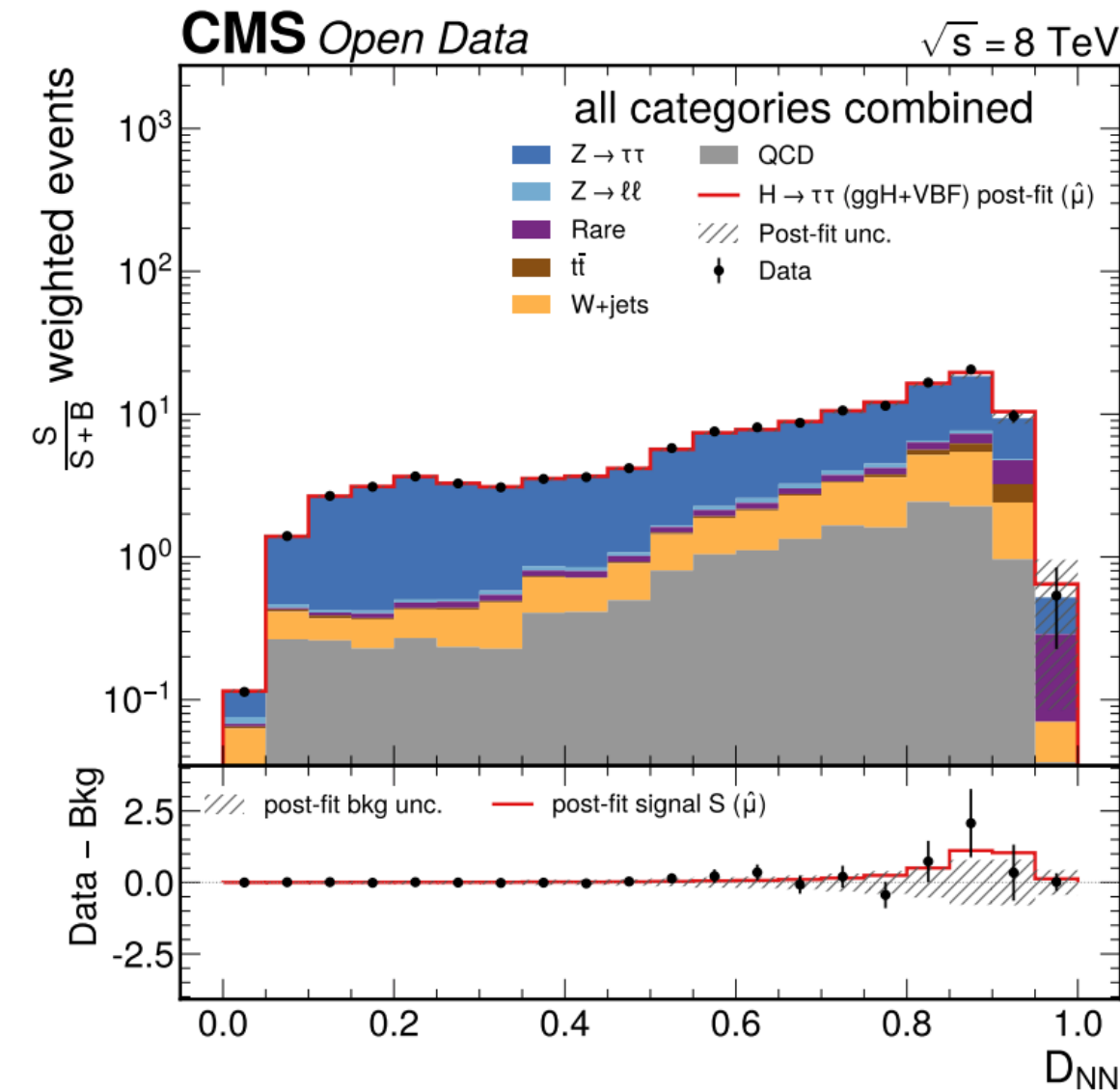
Working on: “benchmark” for more accurate measure for how to score a NN’s performance



*Opus 4.8

How Good is it?

How does it compare to a reference



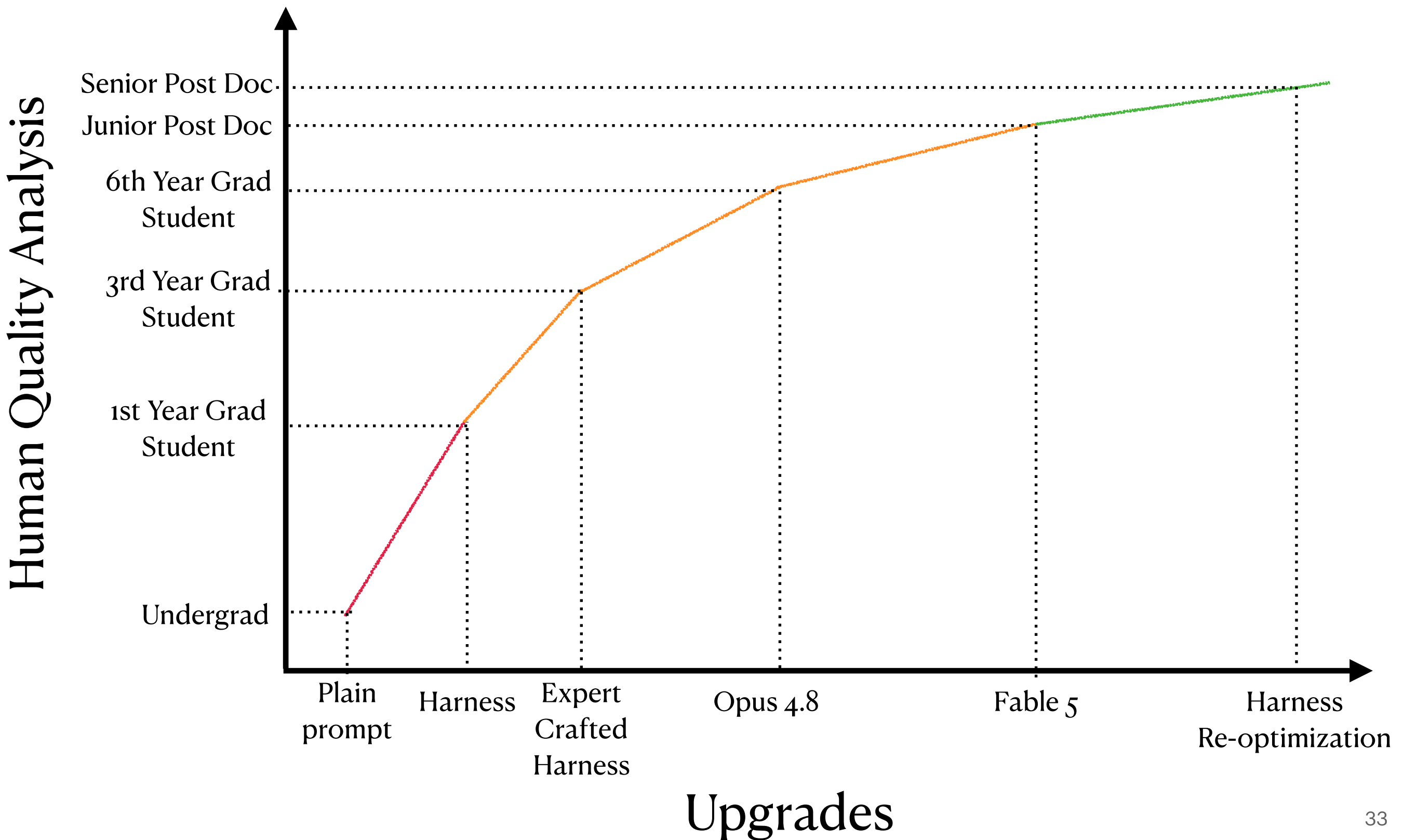
CMS*: 1.01 ± 0.59

ATLAS*: 1.1 ± 1.09

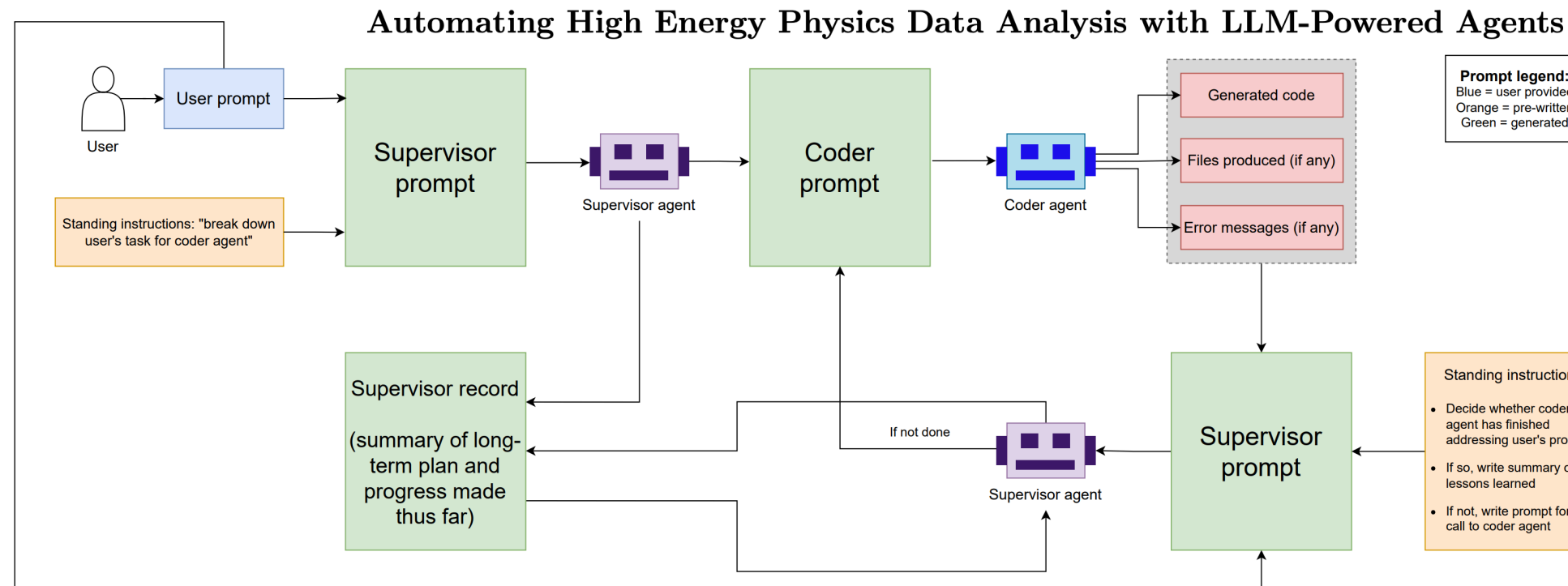
Same Uncertainty as the ATLAS experiment

***Very approximate scaling**

Heuristic Progression



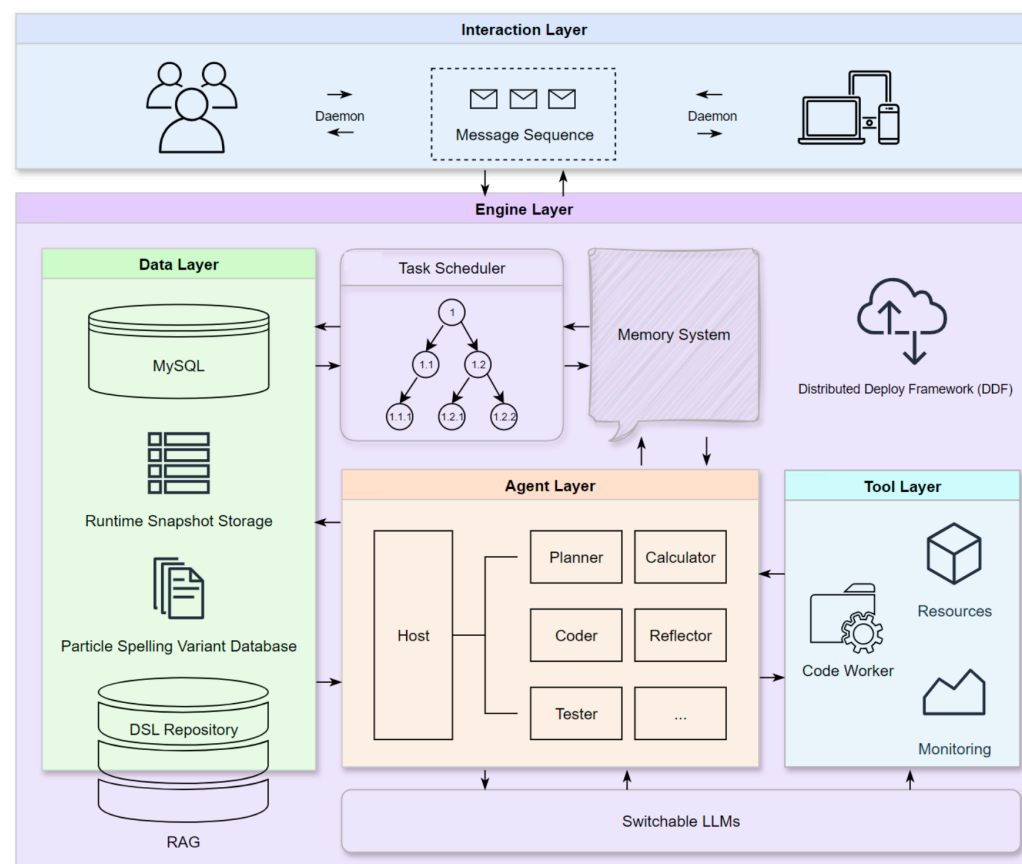
Other Experimental End-to-End Harnesses



[2512.07785](#)

Supervisor-coder agent + Snakemake on ATLAS Open Data.

First LLM-agent-driven automated analysis framework in HEP.



[2604.22541](#)

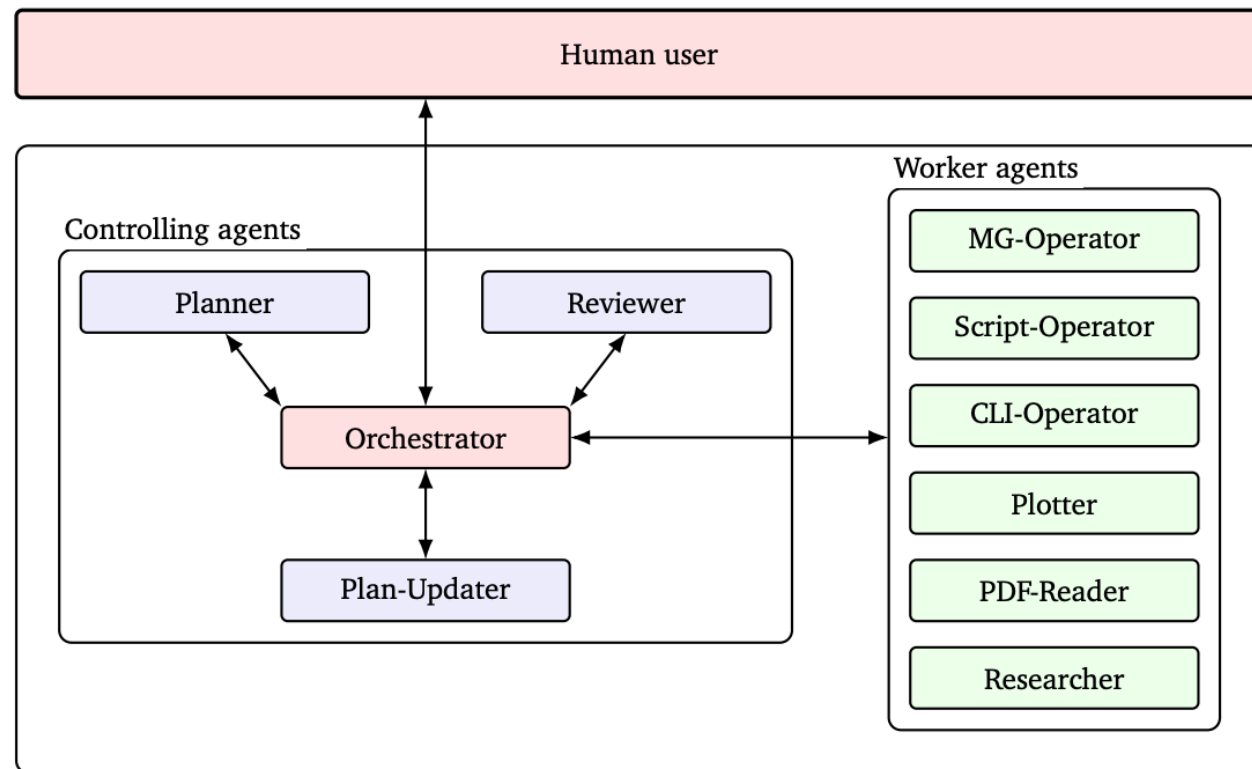
Dr.Sai: An agentic AI for real-world physics analysis at BESIII

Host/Planner/Coder/Tester/Calculator/Reflector.

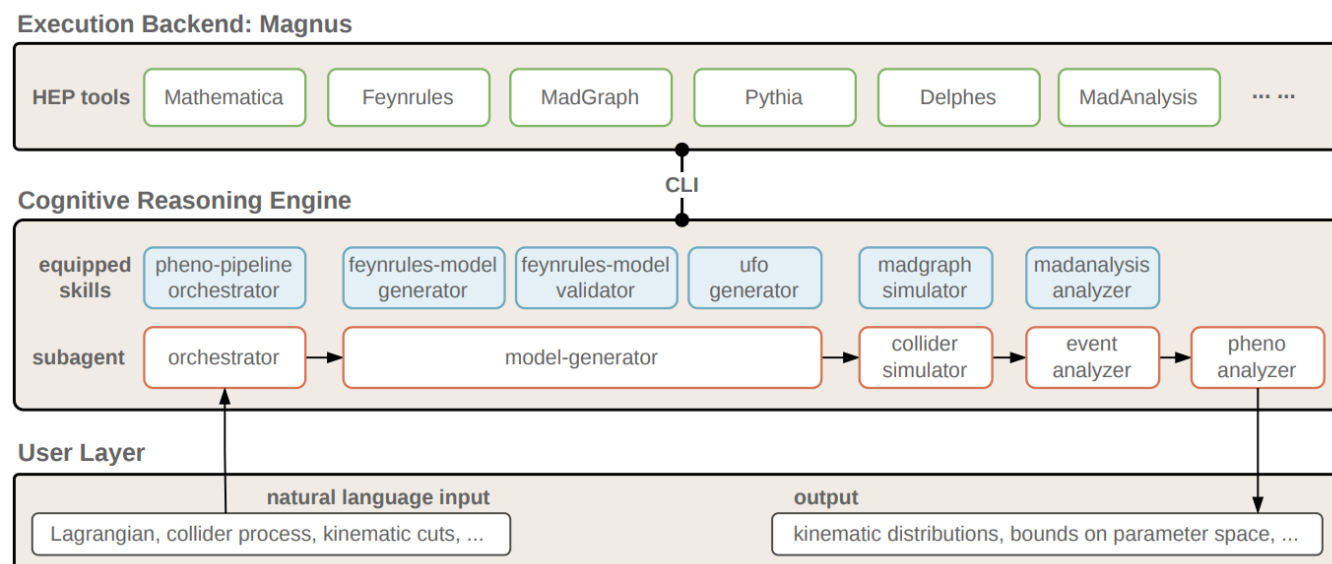
Conclusion: "the core challenge is no longer reasoning but precision of tool-calling."

Pheno End-to-End Agents

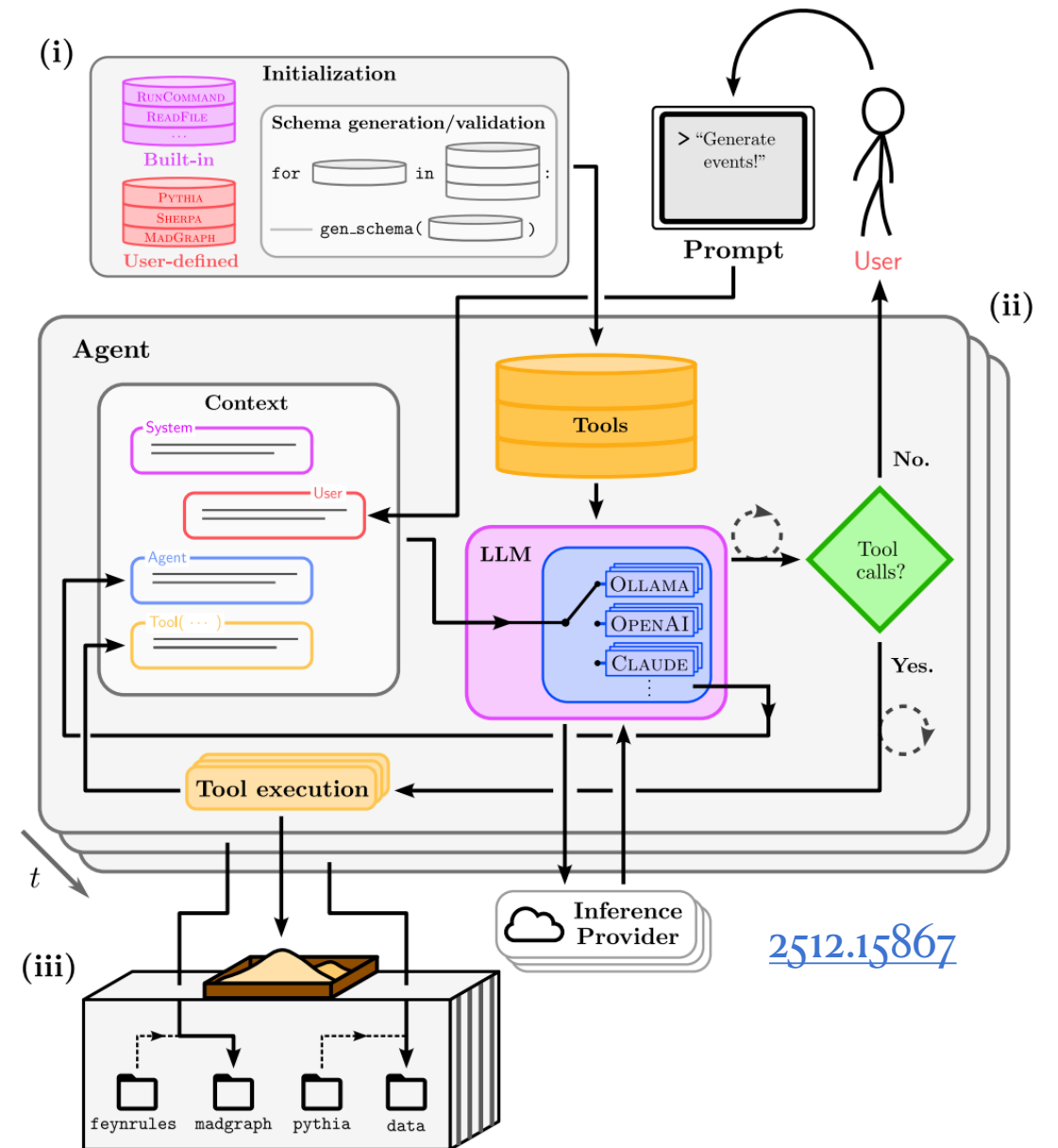
2601.21015



MadAgents: Agents drive MadGraph campaigns from a paper PDF



2603.14553

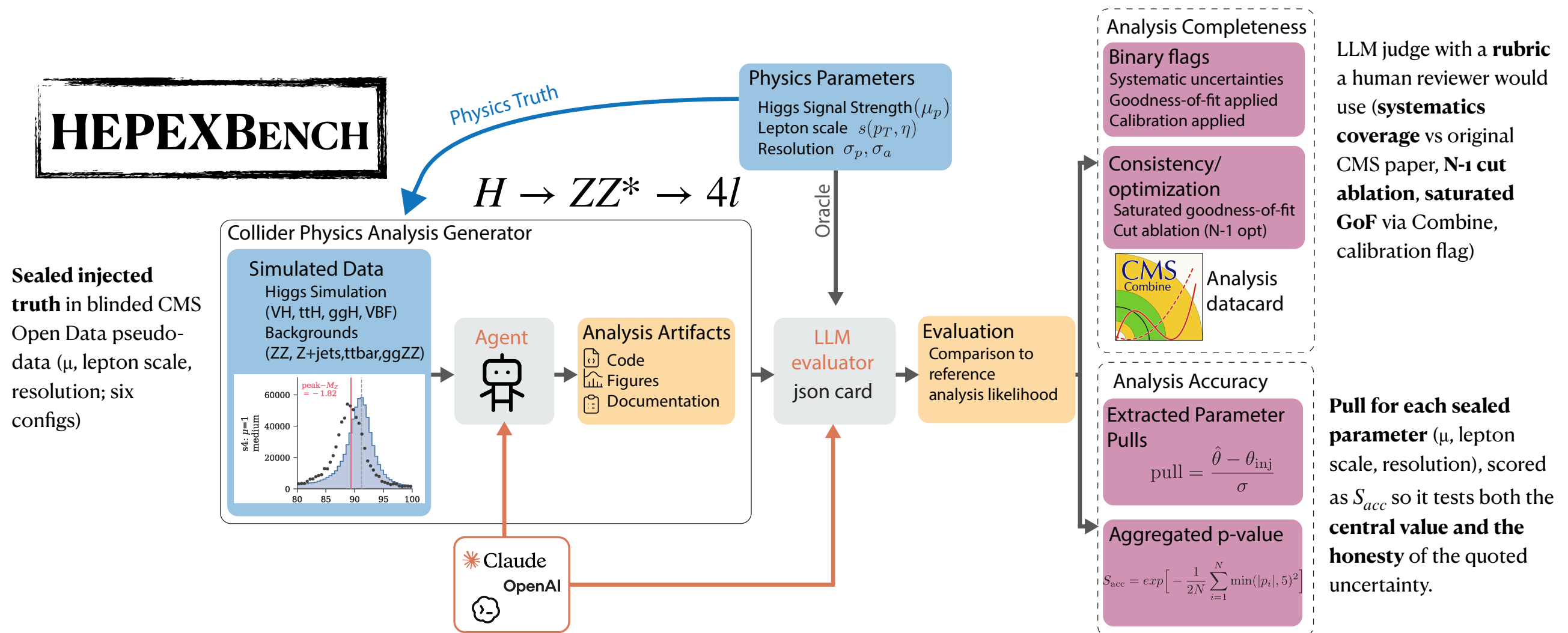


HEPTAPOD: Auditable orchestration layer for BSM MC pipelines, human-in-the-loop by design.

COLLIDERAGENT: Multi-agent + Magnus backend; leptoquark/ALP/EFT scans to exclusion limits.

Up Next: HEP Agentic Benchmarks

- How do we know we can trust the Agentic results?
- How do you grade an analysis with no ground truth?

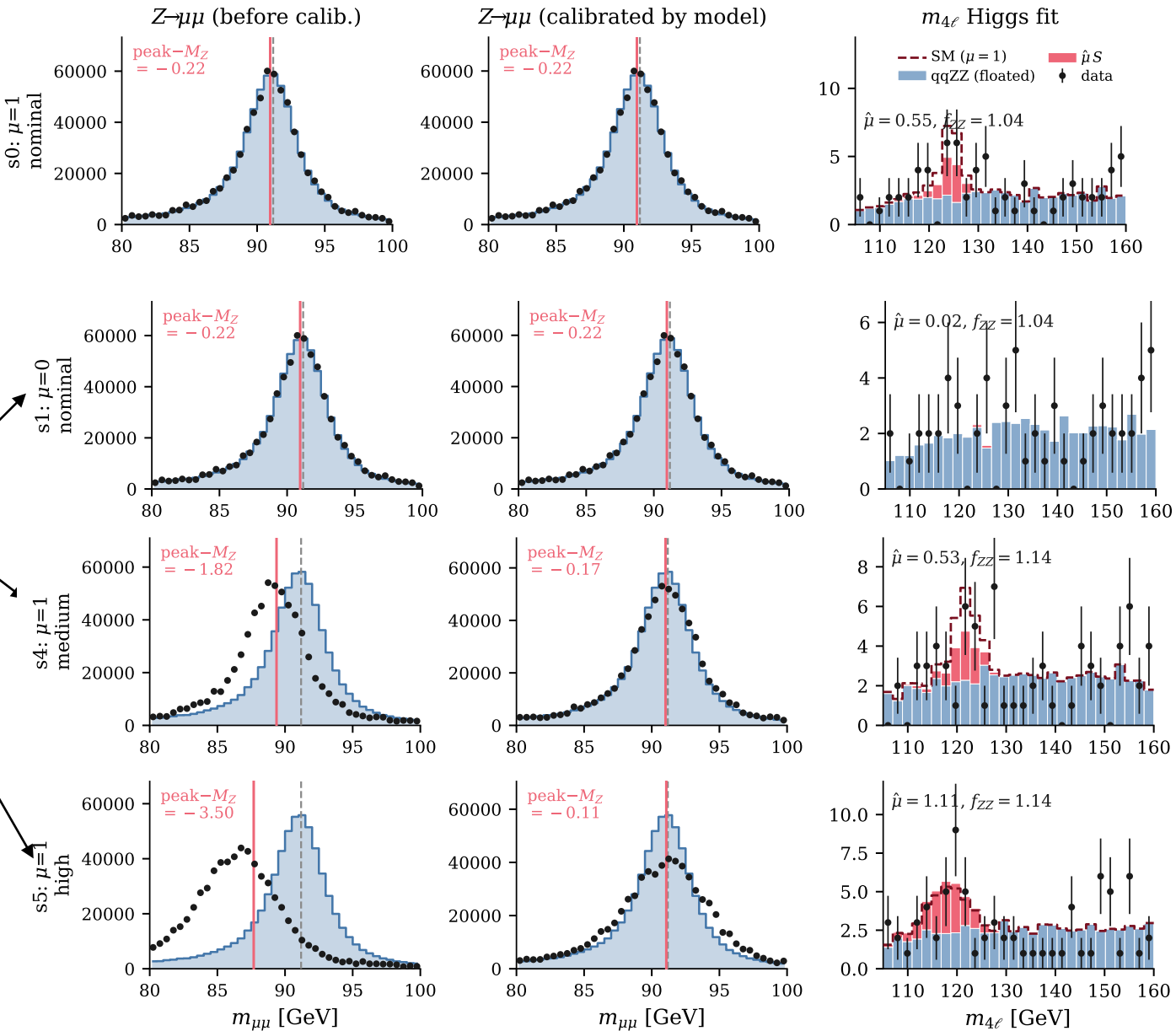


An agent asked to reproduce a published result can recall the answer instead of measuring it; sealed injected truth in blinded pseudo-data can score the physics.

HEPEXBENCH Results*

--- SM ($\mu = 1$) $\hat{\mu} S$
 ■ qqZZ (floated) • data

μ = Higgs
signal
strength



Higgs present, no distortion (s0)

$\hat{\mu} = 0.55$, pull -1.1σ . Cause: free-floated qqZZ normalisation absorbs signal under the peak. GoF healthy.

No Higgs, no distortion (s1: the null test)

$\hat{\mu} = 0.02$, unbiased. Null control passes: the μ deficit is not a calibration artifact.

Higgs present, 1.8 GeV lepton scale shift (s4)

$\hat{\mu} = 0.53$, same -1.1σ bias as nominal despite the 2% scale shift. Calibration and background error are independent.

Higgs present, 3.5 GeV lepton scale shift (s5)

$\hat{\mu} = 1.1$, right answer by accident: large calibration correction cancels the background bias. Saturated GoF $p \approx 0$ flags the residual mismodelling the pull alone would hide.

Distortion = hidden lepton momentum-scale shift and resolution smear injected into the pseudo-data, "low/medium/high" $\approx 0.5\%, 2\%, 5\%$ scale shift

configuration	Procedural					Scores	
	unc. inclusion %	cuts	GoF (p-value)	GoF test	calib. test	procedural score	accuracy score
$\mu=1$ nominal	75 ± 11	1.00 ± 0.00	$1.00 (0.74 \pm 0.02)$	✗	✓	0.92 ± 0.04	0.69 ± 0.13
$\mu=0$ nominal	72 ± 15	1.00 ± 0.00	$1.00 (0.46 \pm 0.02)$	✗	✓	0.91 ± 0.05	0.98 ± 0.02
$\mu=0$ medium	62 ± 10	1.00 ± 0.00	$1.00 (0.06 \pm 0.01)$	✗	✓	0.88 ± 0.03	0.90 ± 0.10
$\mu=1$ low	70 ± 10	1.00 ± 0.00	$0.50 (0.98 \pm 0.01)$	✗	✓	0.73 ± 0.03	0.60 ± 0.21
$\mu=1$ medium	68 ± 11	1.00 ± 0.00	$1.00 (0.24 \pm 0.02)$	✗	✓	0.89 ± 0.04	0.61 ± 0.17
$\mu=1$ high	79 ± 13	0.67 ± 0.00	$0.00 (< 0.002)$	✗	✓	0.49 ± 0.04	0.62 ± 0.44

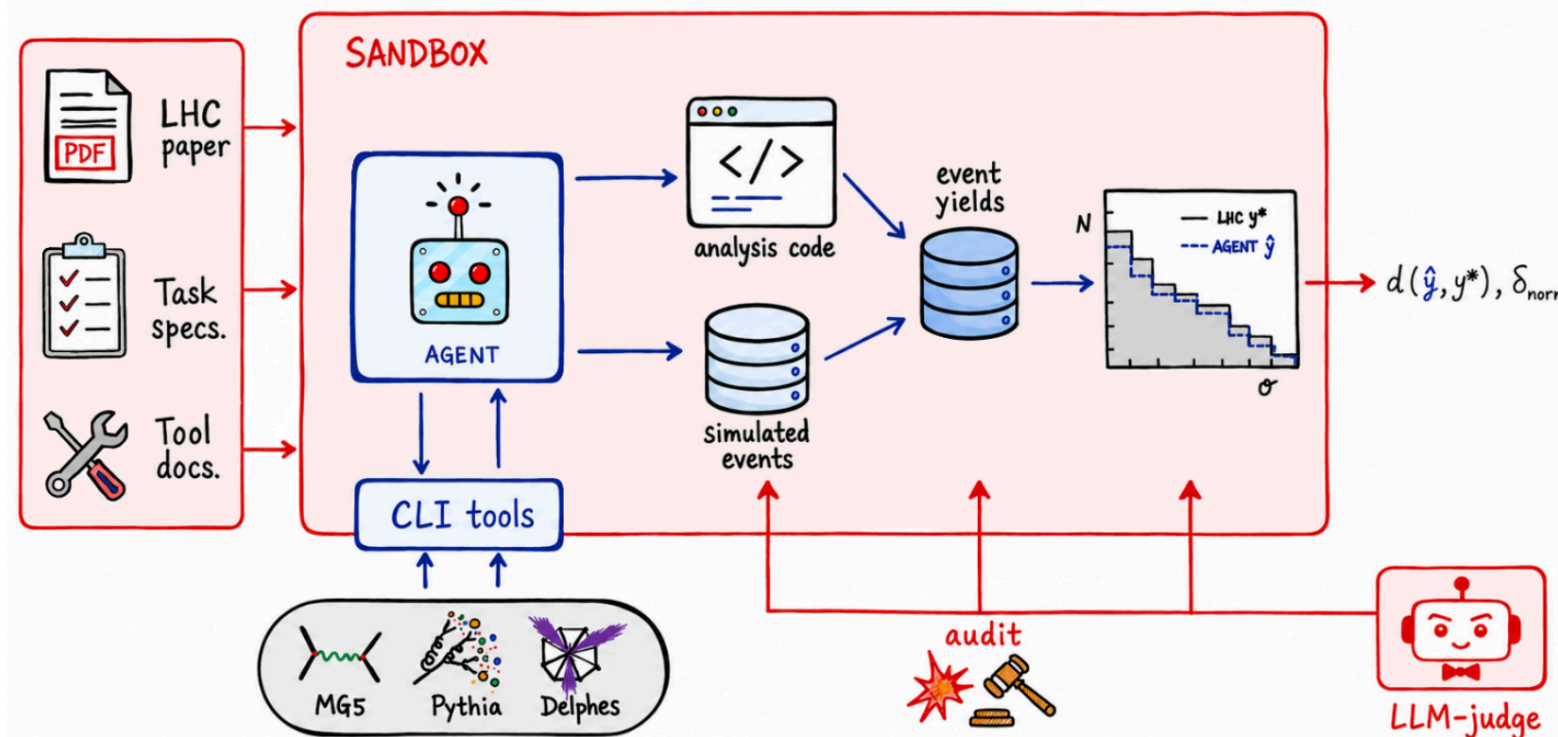
Method quality (procedural score) is high and nearly constant until the distortion is.

Answer accuracy is high when there is no Higgs and mediocre whenever there is one, independent of distortion

Another Agentic Benchmark: Collider-Bench

COLLIDER-BENCH

2605.13950

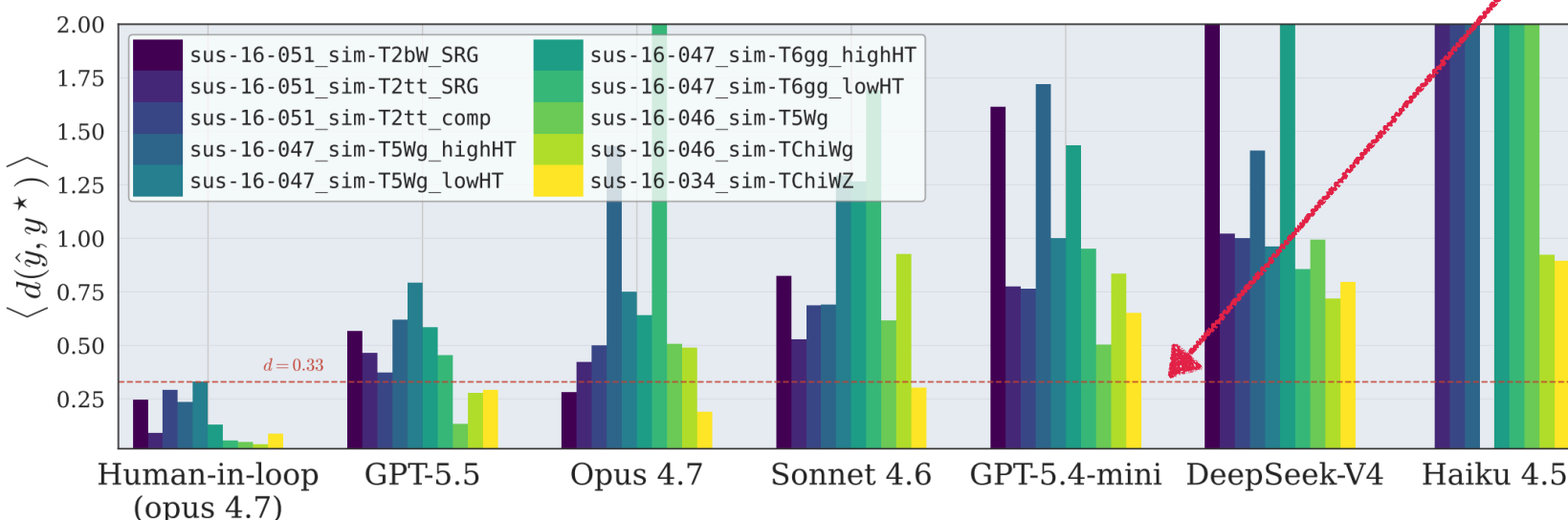


Analysis target	Signal s	obs. $\mathcal{O}$	LHC search	Reference	Diff.
leptons + jets	TChiWZ	E_T^{miss}	CMS-SUS-16-034	Sirunyan et al. [2018b]	★
photons	T5Wg	S_T^{γ}	CMS-SUS-16-046	Sirunyan et al. [2018a]	★
photons	TChiWg	S_T^{γ}	CMS-SUS-16-046	Sirunyan et al. [2018a]	★
photons	T5Wg, high- H_T	p_T^{miss}	CMS-SUS-16-047	Sirunyan et al. [2017b]	★★
photons	T5Wg, low- H_T	p_T^{miss}	CMS-SUS-16-047	Sirunyan et al. [2017b]	★★★★
photons	T6gg, high- H_T	p_T^{miss}	CMS-SUS-16-047	Sirunyan et al. [2017b]	★★
photons	T6gg, low- H_T	p_T^{miss}	CMS-SUS-16-047	Sirunyan et al. [2017b]	★
single lepton	T2tt	E_T^{miss}	CMS-SUS-16-051	Sirunyan et al. [2017a]	★
single lepton	T2tt, compressed	E_T^{miss}	CMS-SUS-16-051	Sirunyan et al. [2017a]	★★★★
single lepton	T2bW	E_T^{miss}	CMS-SUS-16-051	Sirunyan et al. [2017a]	★★

10 tasks from 4 CMS Run-2 SUSY searches

- **Agent reads the paper and builds MadGraph→Pythia→Delphes pipeline**, returning binned signal-region yields.
- Scored by relative L^2 against published yields; pass = beating the worst physicist-in-the-loop result ($\tau = 0.33$).
- **No agent reliably beats the physicist-in-the-loop baseline**: Shape usually right, normalization usually wrong .
- Provenance audit: 87% passed, 6% failed, 6% fabricated.
- **Reproduction benchmarks measure whether an agent can rebuild a known result**. The reference answer is on arXiv, so recall and fabrication are failure modes and need a provenance audit to catch.

Results

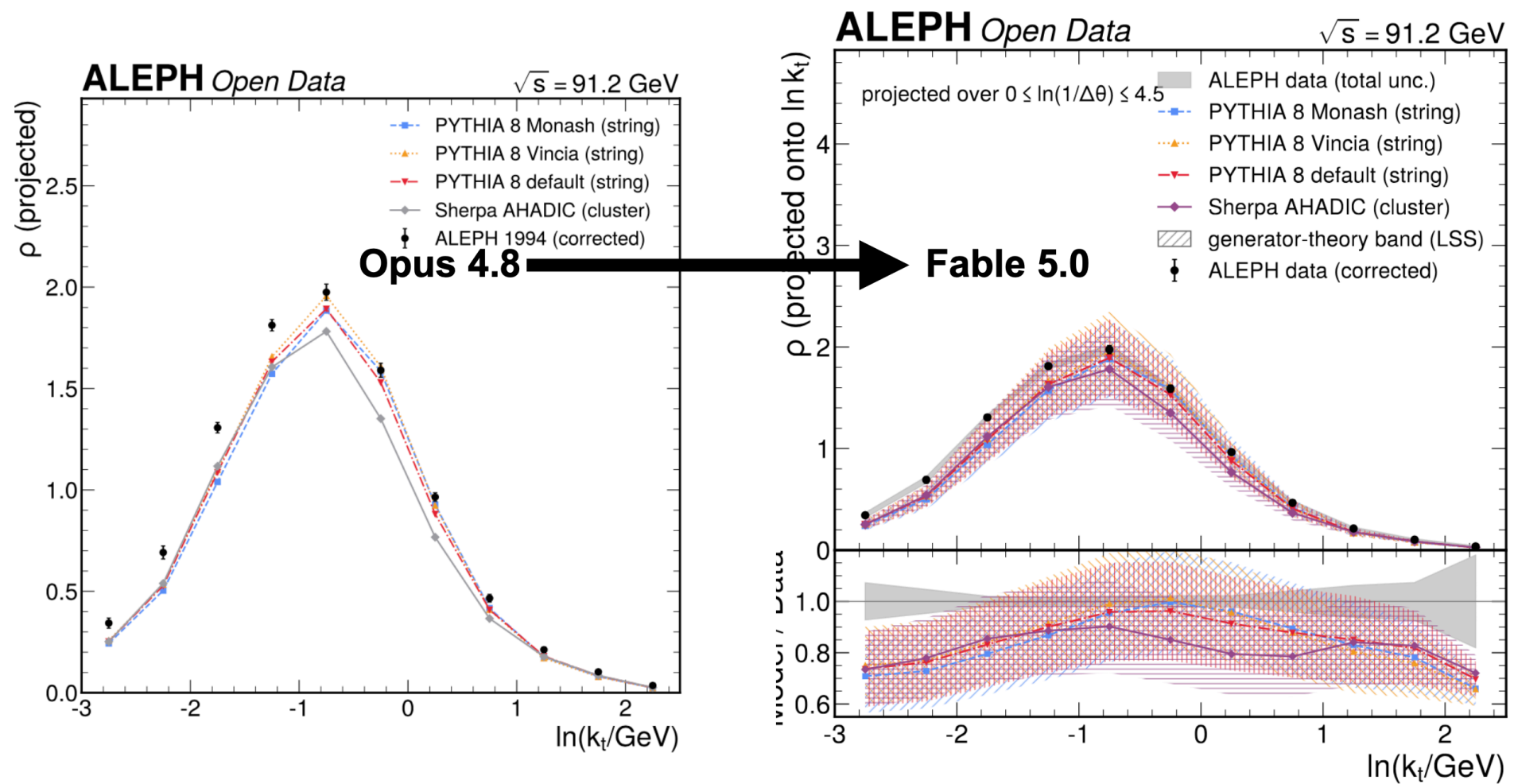


Conclusions

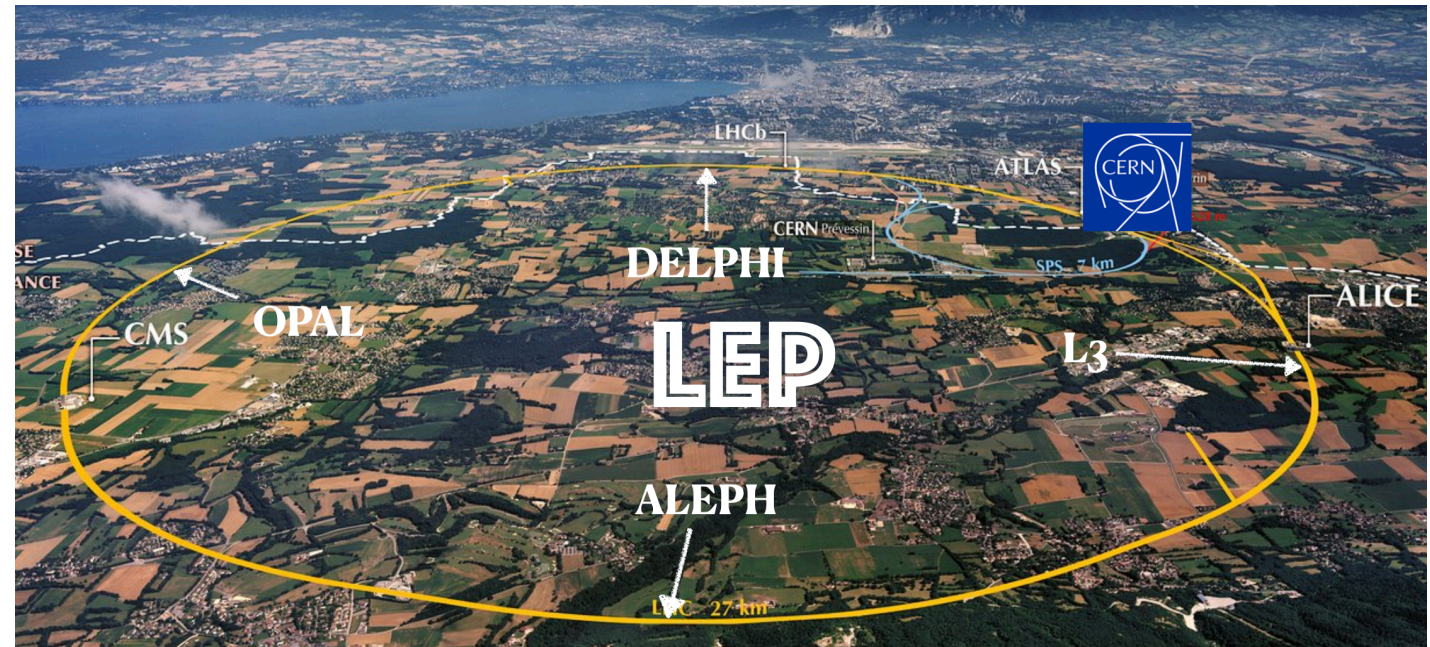
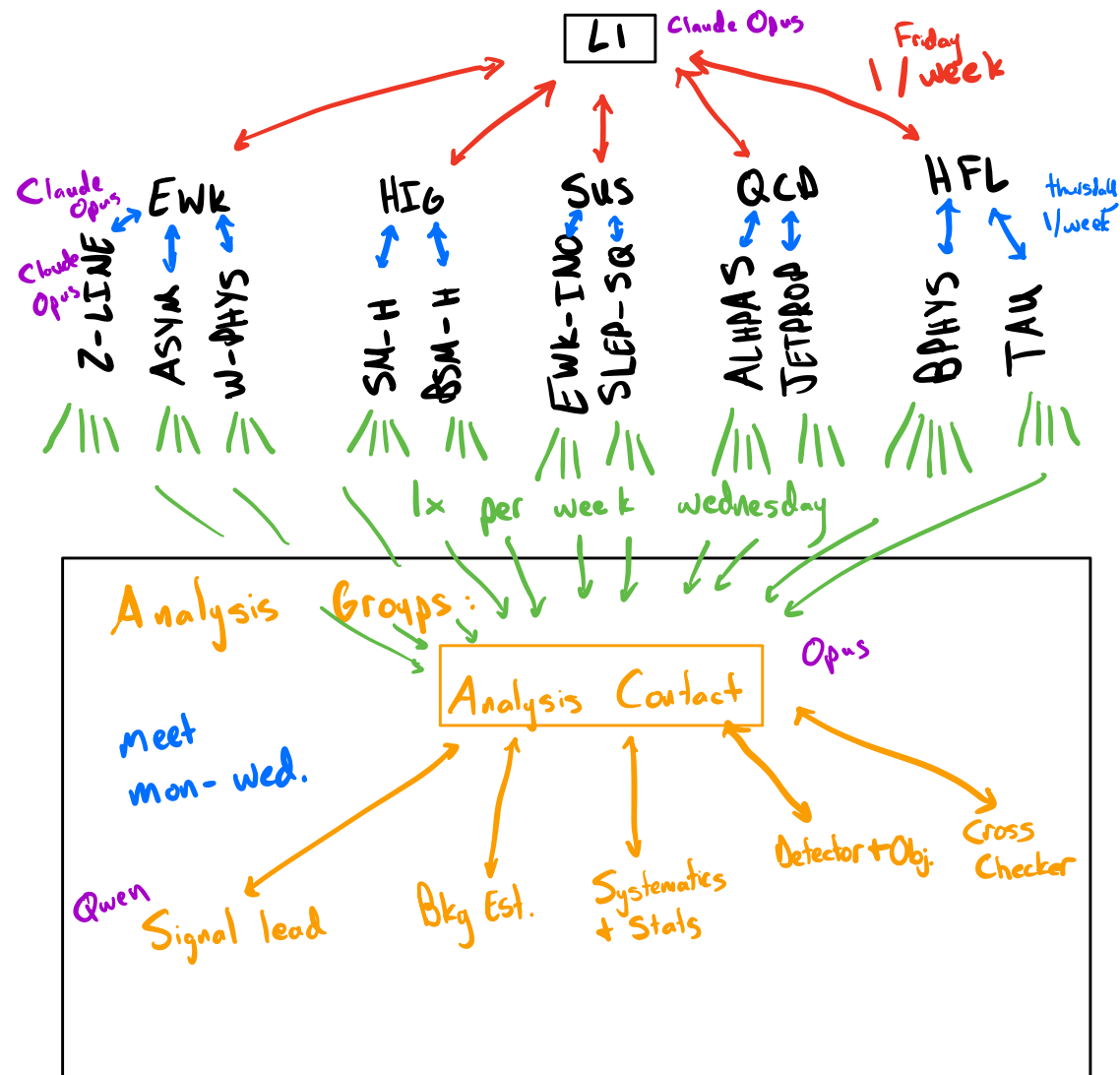
- Agentic AI harness frameworks are able to analyze HEP data in an experimental/phenocontext
- Full tests from us up to Opus 4.8 show successful analyses completing with **less precision** vs. collaboration publications, but **novel analyses are possible**, showing that models don't just copy results.
- **Successively newer models trend upwards in analysis quality** (qualitative assessment)
- Benchmarks/RL Environments are essential for trust within the community and for providing to frontier labs for scientific fine-tuning campaigns.

* Claude Prospects with Mythos / Fable

- Pros:
 - This could be the senior grad student -> post doc intelligence bump
 - Anthropic has dedicated fine-tuning teams maximizing scientific performance / improving reasoning
- Cons:
 - Might be prohibitively expensive to run (10 trillion parameter model)
 - Entirely reliant upon closed models
 - Brain drain in the field?



Future Plans!



- Preserve the LEP results with a **global fit of the Standard Model**
- Bring this to the LHC Collaborations as a proof of concept of fully-automated analysis workflow
- Scrutinize the current CMS analysis-paradigm