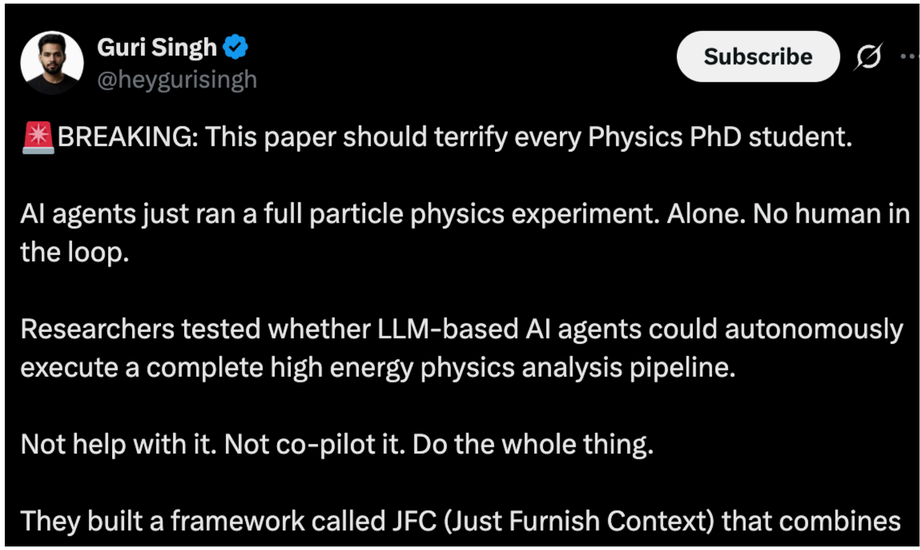




AI Agents Can Already Autonomously Perform Experimental High Energy Physics

NPTwins 2026

Based on paper: https://arxiv.org/abs/2603.20179

Eric A. Moreno, Andrzej Novak, Sam Bright-Thonney, Dolores Garcia, 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?

Everyday a new agent paper drops

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.

CPTNP-2026-012

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,

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,*

nthony.badea@cern.ch

Agentic Diagrammatica: Towards Autonomous Symbolic Computation in High Energy Physics

Tony Menzo[1,2]*; Alexander Roman[1], George T. Fleming[2], Sergei Gleyzer[1], Konstantin T. Matchev[1], Stephen Mrenna[2]

[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,

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

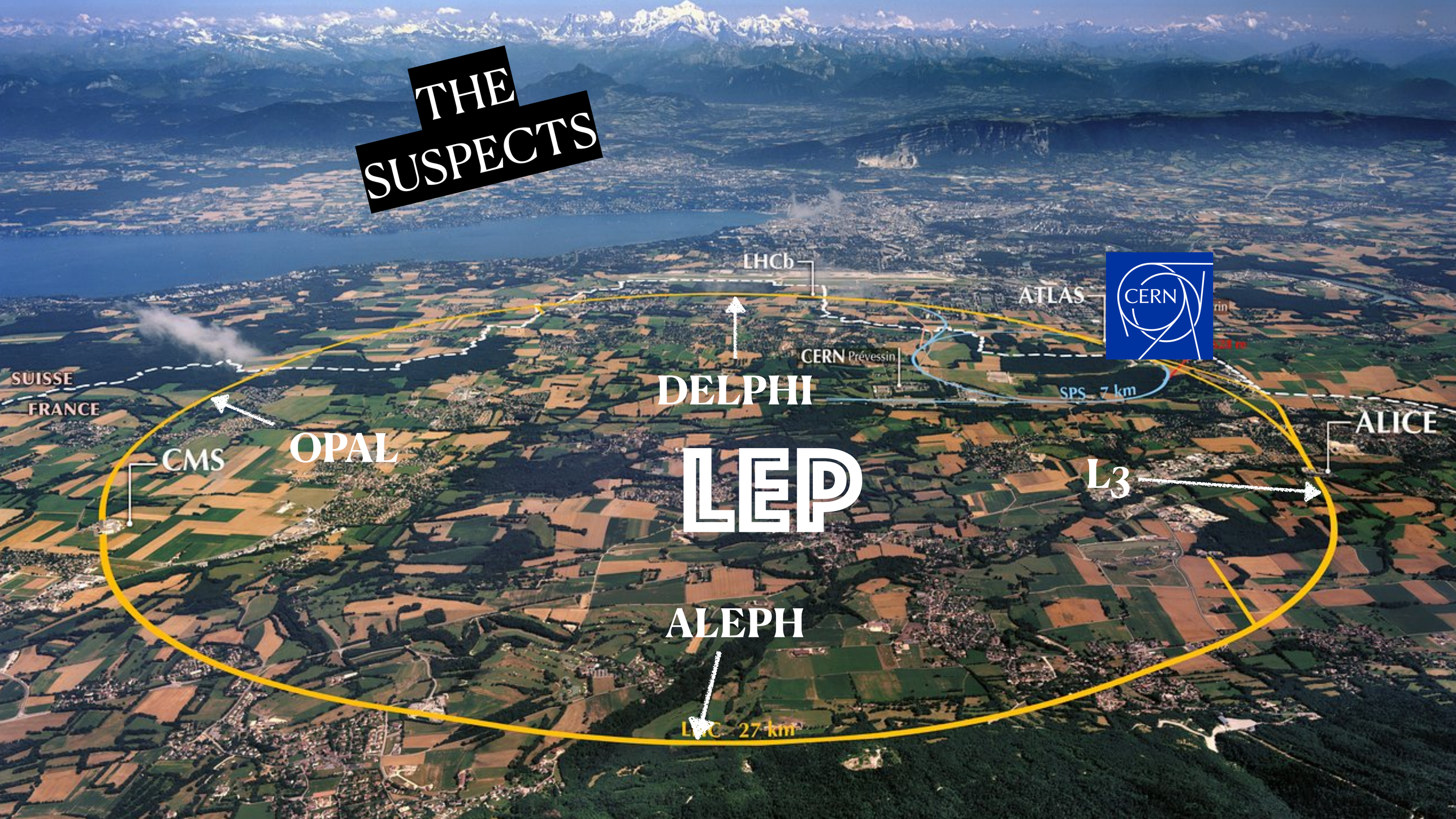
AI Agents Can Already Autonomously Perform Experimental High Energy Physics

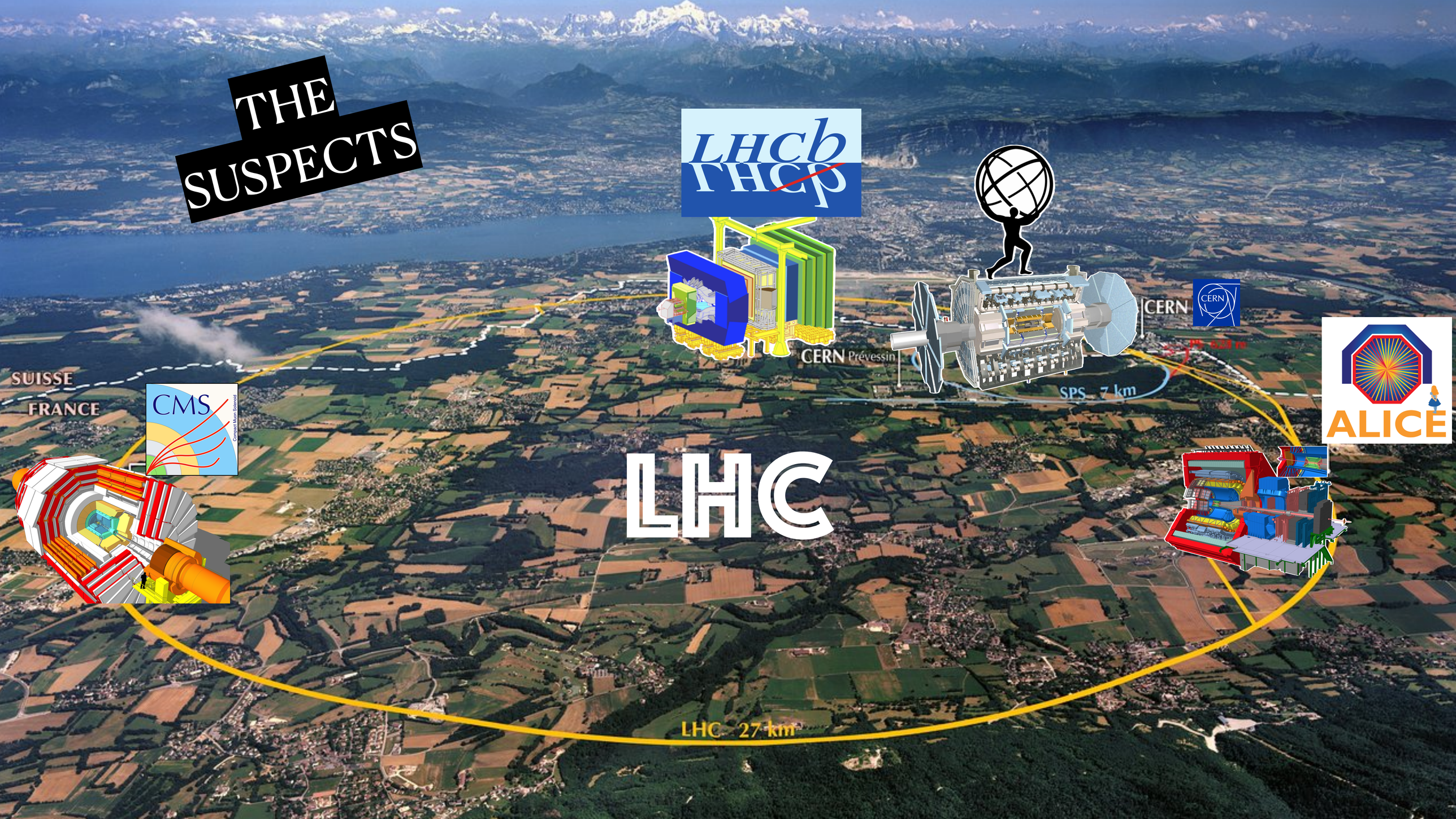
Eric A. Moreno*†[1,2], Samuel Bright-Thonney*‡[1,2], Andrzej Novak*§[1,2], Dolores Garcia♯[3], and Philip Harris♭[1,2]

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

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

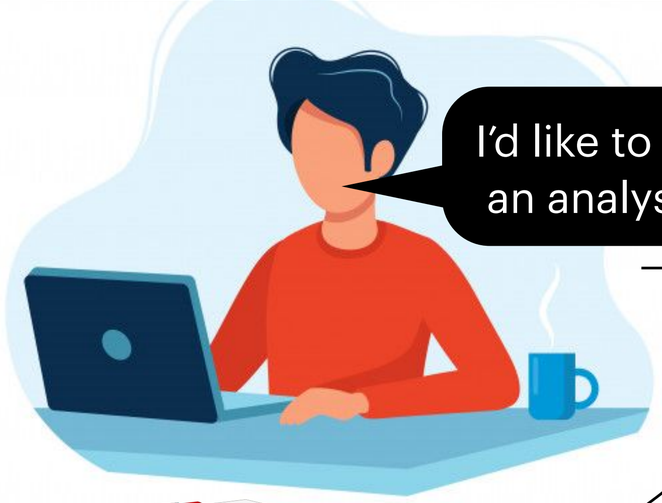
Lianhao Zhou[1] Hongyi Ling[1] Cong Fu[1] Yepeng Huang[2] Michael Sun[3] Wendi Yu[1] Xiaoxuan Wang[4] Xiner Li[1] Xingyu Su[1] Junkai Zhang[4] Xiusi Chen[5] Chenxing Liang[1] Xiaofeng Qian[6,7,8] Heng Ji[5] Wei Wang[4] Marinka Zitnik[2] Shuiwang Ji[1,6,9]*

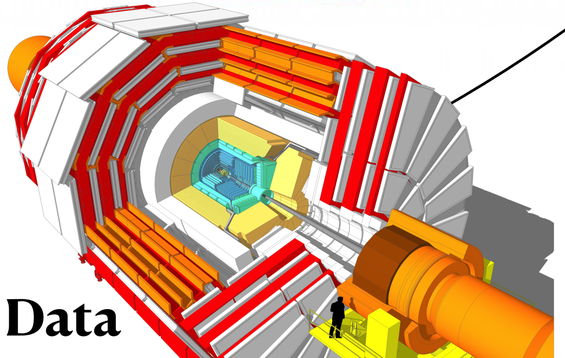

THE SUSPECTS
LHCb
ATLAS
CERN
CERN Prévessin
SUISSE
FRANCE
DELPHI
SPS 7 km
ALICE
CMS
OPAL
LEP
L3
ALEPH
LEP 27 km


THE SUSPECTS
LHCb
CERN Prévessin
CERN
SPS 7 km
SUISSE
FRANCE
CMS
Compact Muon Solenoid
LHC
ALICE
LHC 27 km

How a HEP Analysis Goes







Data

Analysis Framework

Office Feedback

Update Framework

Congrats!

1 year

Go back!

Collaboration Review

1 year

Convener Feedback

Update Framework

2-3 years

Subgroup Feedback

Update Analysis



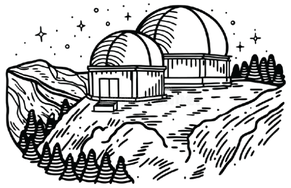
https://arxiv.org/abs/2510.21935

Samuel Bright-Thonney et al.

AutoSciDACT

How do we automate scientific discovery?

Observe
Take data & understand

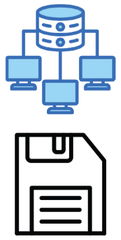


Inquire
Ask a question



Expert / LLM Input
What to look at?
What reference?

Research
Search for answers



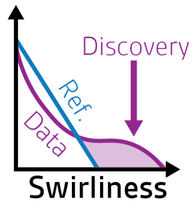
Simulation or reference

Data

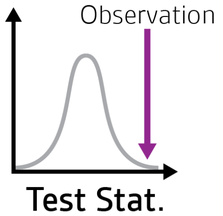
Hypothesize
Formulate a hypothesis



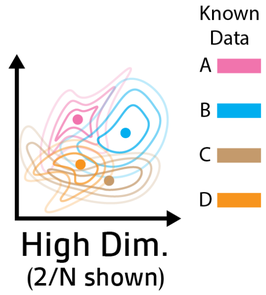
Test
Experiment



Conclude
Results

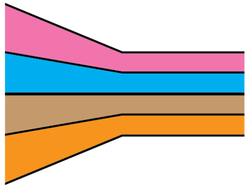


Reduce
Simplify with symmetry



Pre-trained model / foundation model

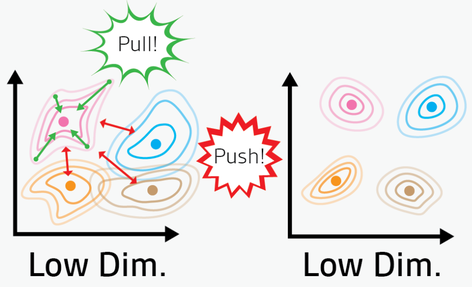
Isolate
Select a subset



Data selection

AutoSciDACT

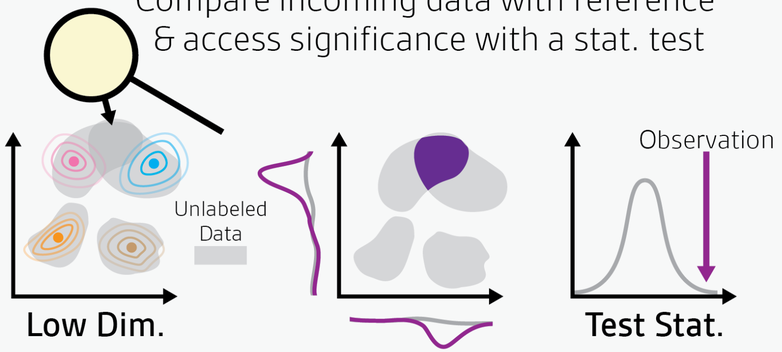
Learn a Space
Organize known signals to create a reference model



Supervised contrastive learning

Resulting ref. model

Anomaly Detection & Test
Compare incoming data with reference & access significance with a stat. test



Search for overdensity

Score anomalous regions

Quantify the significance

https://neurips.cc/virtual/2025/loc/san-diego/122964
https://ml4physicalsciences.github.io/2025/files/NeurIPS_ML4PS_2025_347.pdf

Andrzej Novak et al.

Split-N-Fit

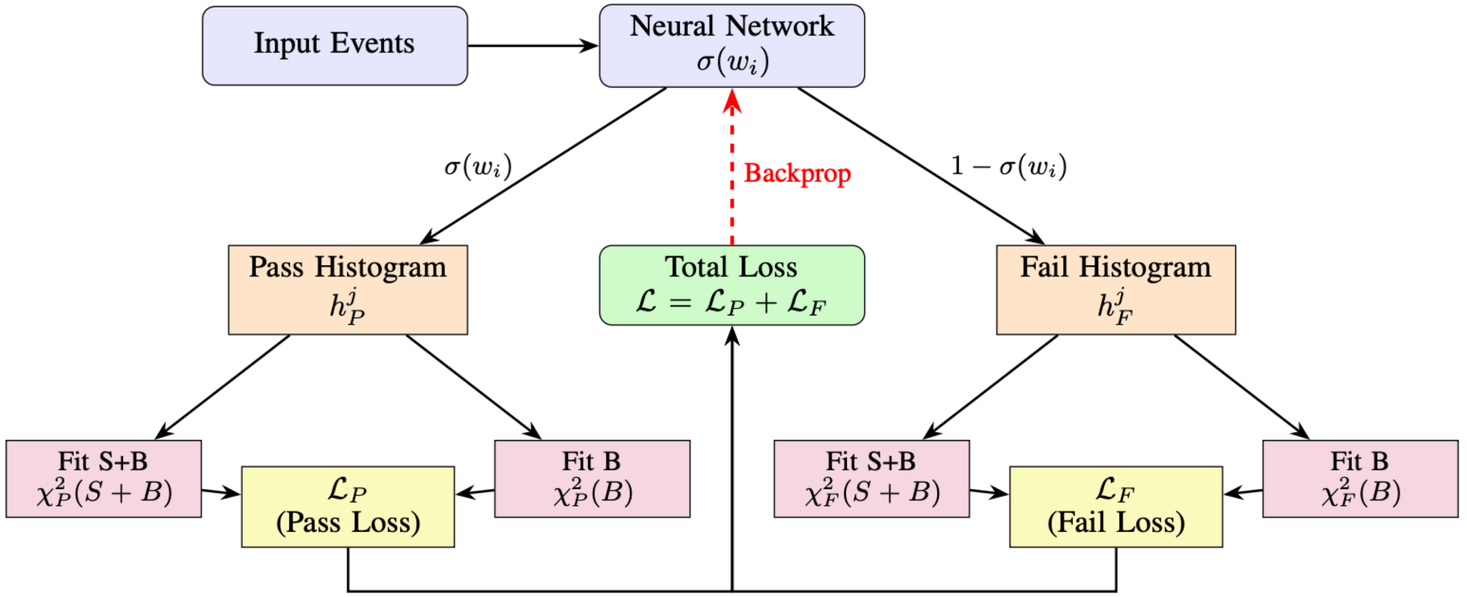
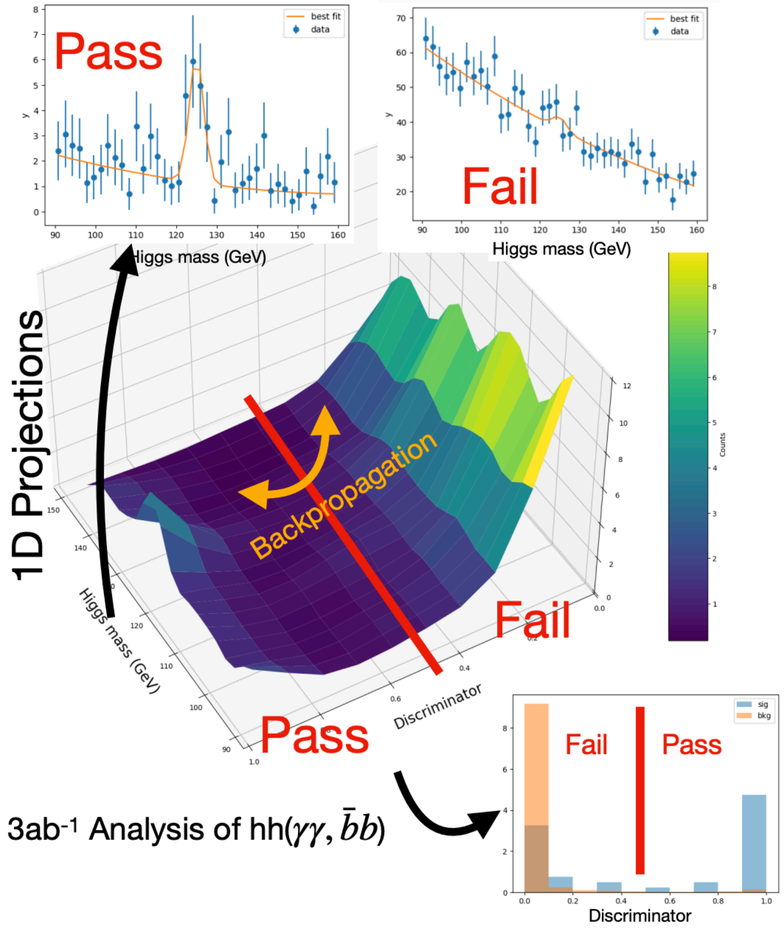


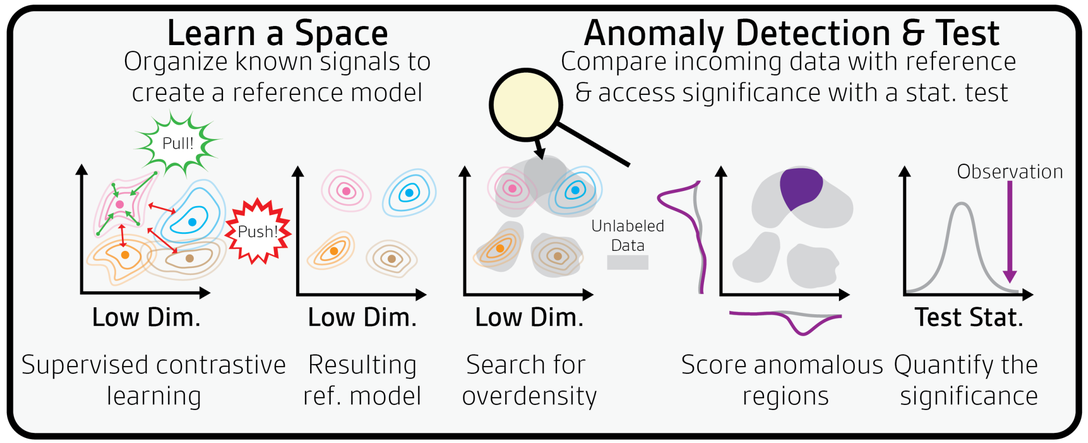
Figure 1: Schematic overview of the Split-N-Fit method. Events are split into pass and fail histograms. Both S+B and B functions are fit to each histogram, yielding χ^2 values that define individual loss terms $\mathcal{L}_P$ and $\mathcal{L}_F$, which combine to form the total differentiable loss.

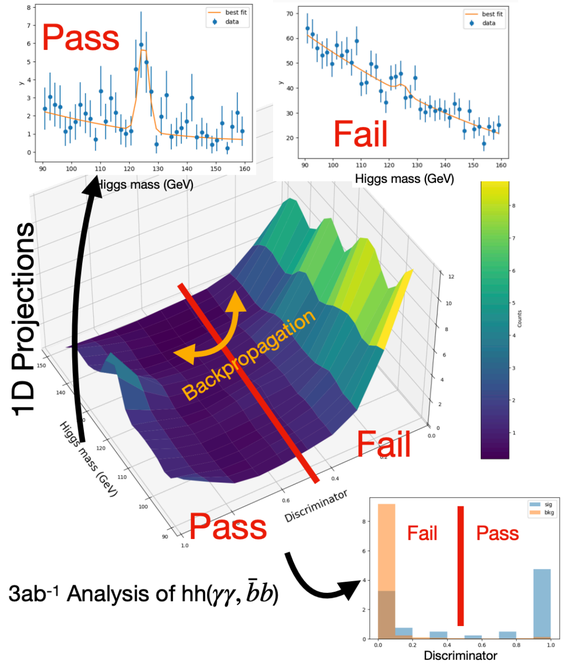


- Automate event selection using differentiable NN
- Instead of training a classifier and only later doing the physics fit, they merge the two into one end-to-end optimization

The Slop-ification Landscape

You can do something clever: **AutoSciDACT**





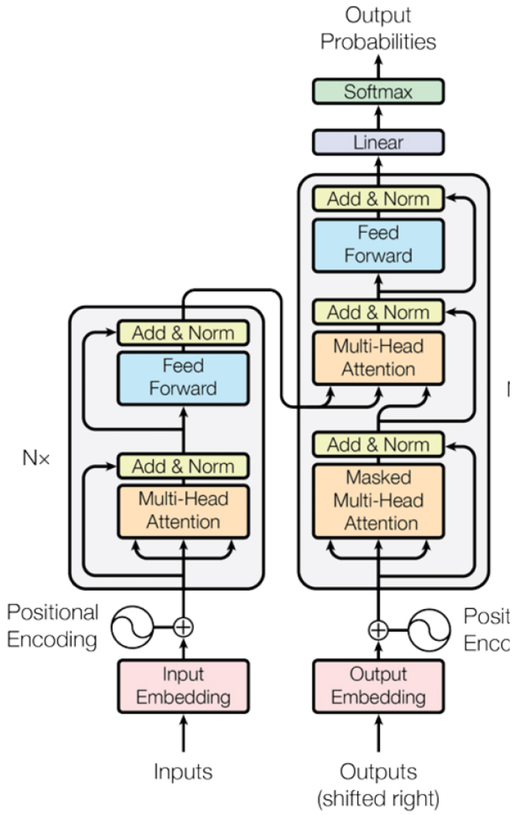
Or you can do something slop:

Claude ChatGPT Gemini

Is slop really the answer?

Is Slop Really the Answer?

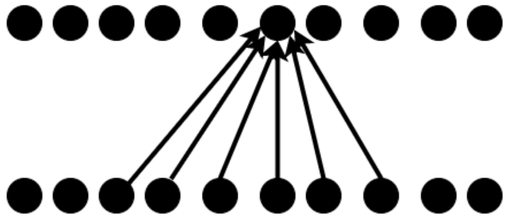
Claude ChatGPT Gemini



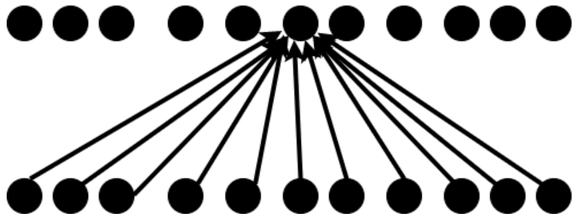
Paper: Attention is all you need
https://arxiv.org/abs/1706.03762

Each point only looks at a small nearby window

Convolution



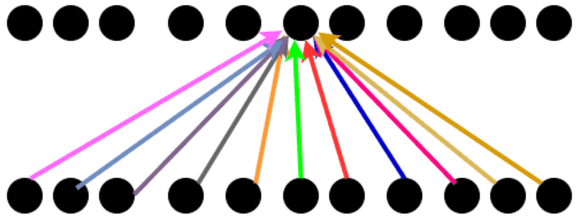
Fully Connected layer



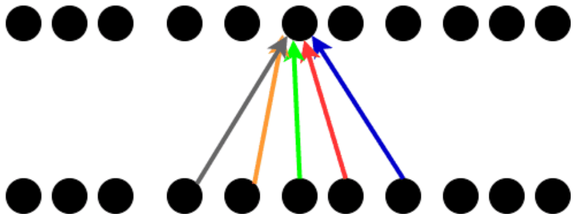
Each point connects to everything with fixed weights

Each point can look at **any position in the sequence** and give more weight to the most relevant ones.

Global attention



Local attention



Each point looks at a small neighborhood, but it can **selectively focus** within that neighborhood

Memory = Attention through time

Attention is a **dynamic weighted connection mechanism**. Instead of treating all inputs equally, the model learns where to “pay attention.”

↓

Attention means: for this token, **which other tokens matter most** right now?

↓

Memory means the model can **reach back to past information and selectively retrieve what matters for the present moment**. In that sense, **memory is just attention applied across previous states** or earlier tokens.

↓

Attention asks: what should I focus on right now?

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

*maybe I don't know what I'm talking about I'm not a neuroscientist

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

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

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

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

Humanity

$$a_t = f(x_t,\ \mathrm{Attn}(x_t, m_t),\ \mathbf{g_t})$$

- a_t = action at time t
- x_t = current input or situation
- m_t = memory/state
- $\mathrm{Attn}(x_t, m_t)$ = 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

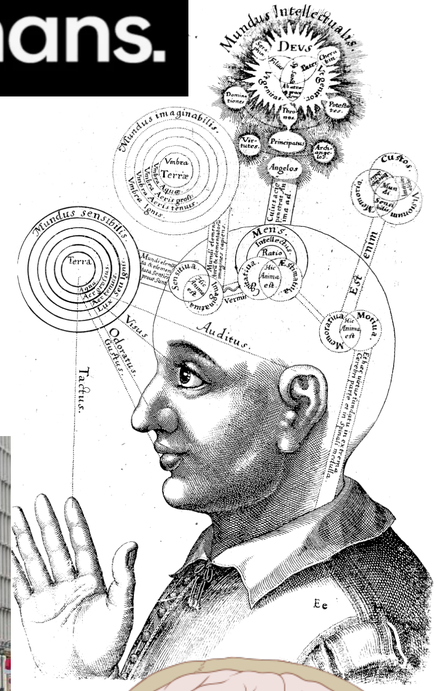
ChatGPT

Gemini

≈









So I'm Doing This Analysis...

A search for boosted light scalars decaying to two tau leptons



EXO-22-001



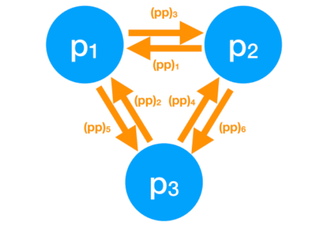
Eric A. Moreno, Andrzej Novak, Dylan Rankin, Jeffrey Krupa, Philip Harris, Cristina Mantilla Suarez
And the DAZSLE team

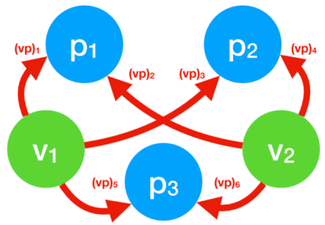
Neural Networks

- Three channels: $\tau_h\tau_h$, $e\tau_h$, $\mu\tau_h$
- Z-taggers:
  - Responsible for rejecting jets from Z → ee and Z → $\mu\mu$, similar to anti e/μ tau ID
  - One for ee vs. $e\tau_h$, one for $\mu\mu$ vs. $\mu\tau_h$
- Di-tau taggers:
  - Responsible for identifying jets coming from true ττ as opposed to fake backgrounds
  - One each for $\tau_h\tau_h$, $e\tau_h$, $\mu\tau_h$
  - Important backgrounds: $\tau_h\tau_h$ vs. QCD/W+Jets, $e\tau_h$ / $\mu\tau_h$ vs. $t\bar{t}$/W+Jets
- Mass regressions:
  - Responsible for correcting mass of jet for loss of neutrinos, similar to SVfit
  - One each for $\tau_h\tau_h$, $e\tau_h$, $\mu\tau_h$
- All based on interaction network architecture (fully connected graph neural network) [1909.12285]

Optimal NN working points for different channels and years

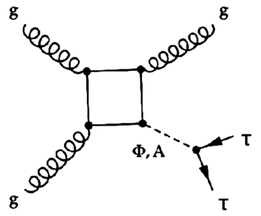
Year	$\tau_h\tau_h$ nn_disc	$\tau_h\mu$ nn_disc	$\tau_h e$ nn_disc
2016APV	0.99995	0.997	0.990
2016	0.99995	0.997	0.992

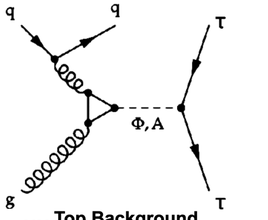


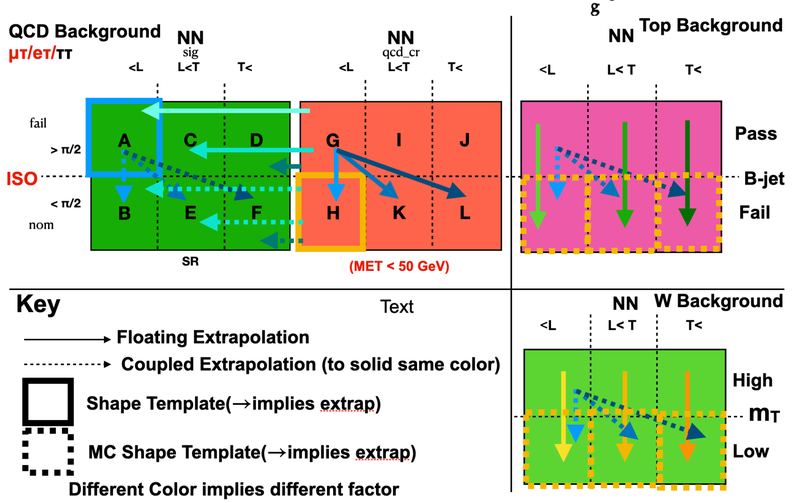


Introduction

- Light scalar searches at LHC are difficult
  - EXO-17-024 (φ→bb) covers some parts of phase space
  - Still much uncovered phase space (leptophilic, M< 50 GeV)
- EXO-22-001 is a novel analysis aimed at searching for boosted **low mass resonances** decaying to taus (φ→ττ)
  - Final state consists of boosted ττ system recoiling against ISR jet
  - Use of new ML and sample generation techniques
  - Targeting resonance mass between 30 GeV and 300 GeV

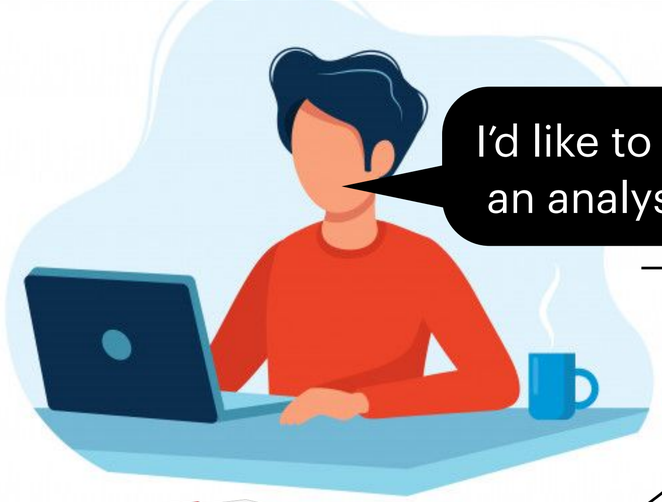


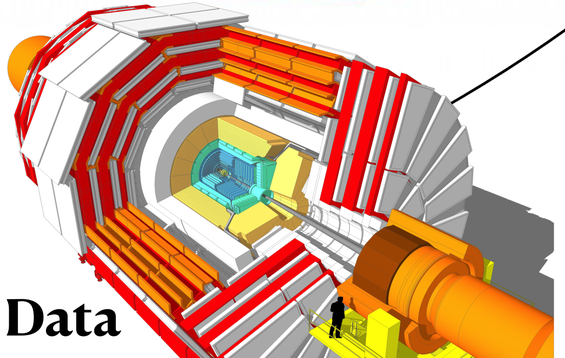




And I'm currently in this infinite loop:







Analysis Framework

Office Feedback

Update Framework

Collaboration Review

Congrats!

1 year

Go back!

1 year

Convener Feedback

Update Framework

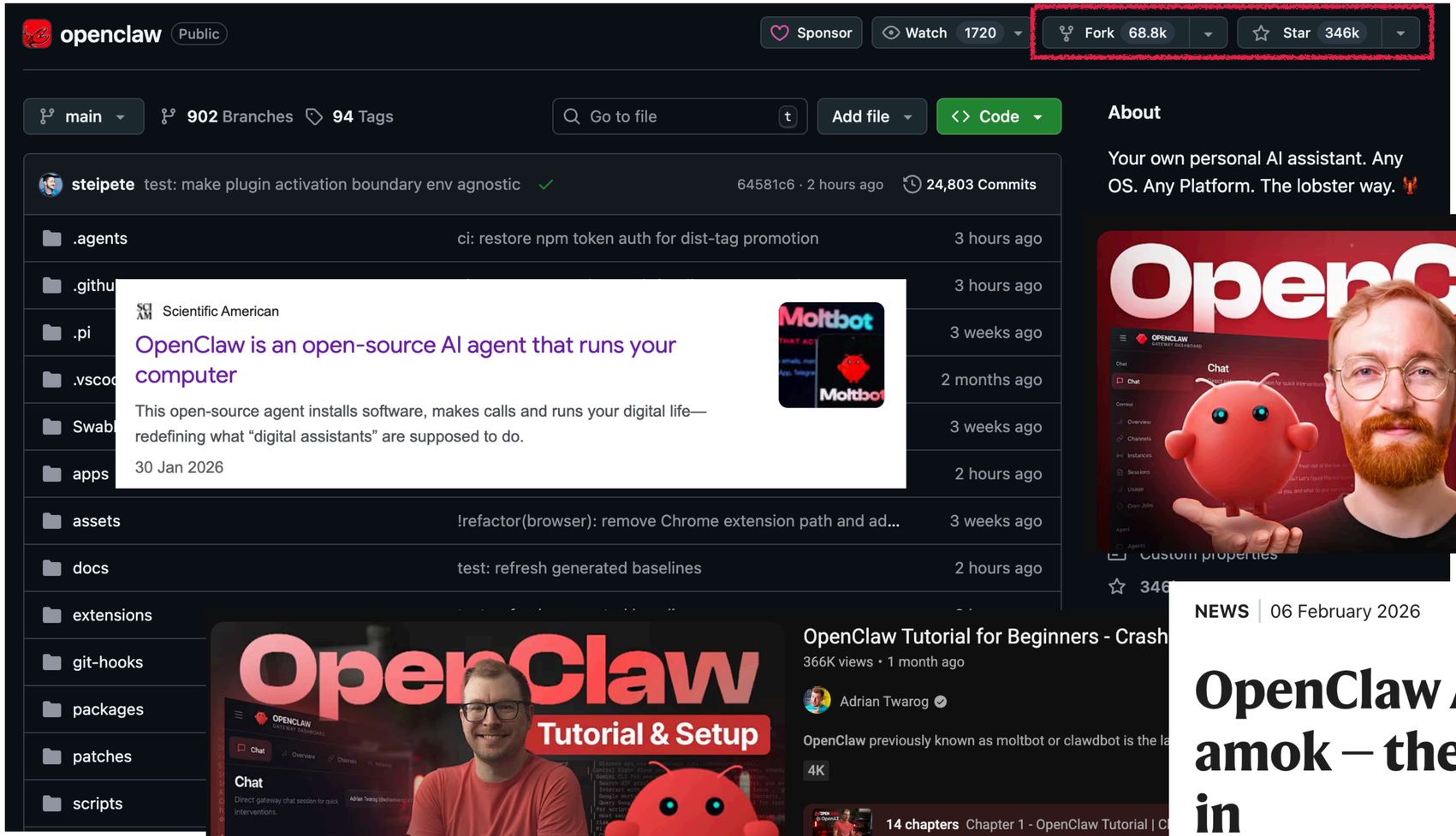
2-3 years

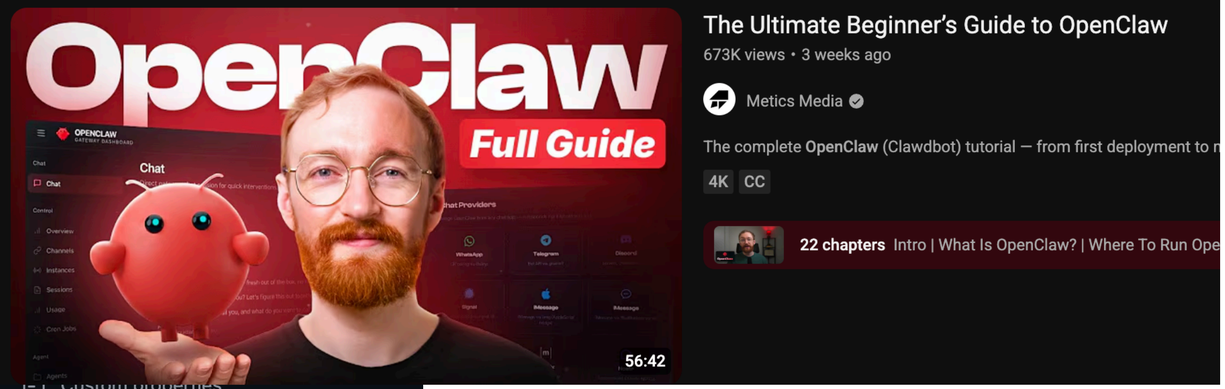
Subgroup Feedback

Update Analysis



And Meanwhile I'm Seeing This...



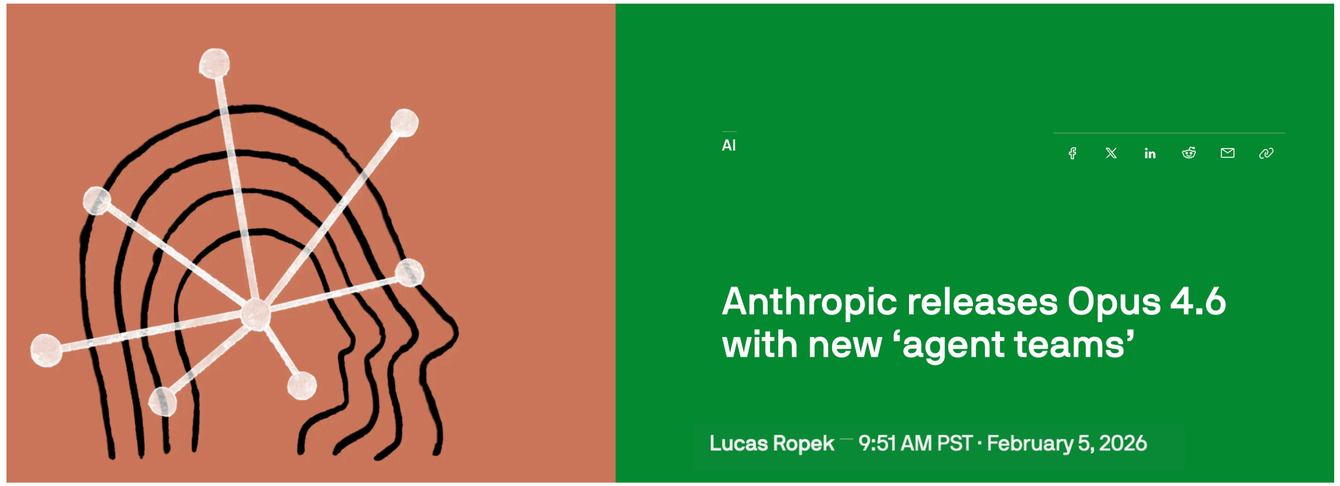


NEWS | 06 February 2026

OpenClaw AI chatbots are running amok — these scientists are listening in

Artificial-intelligence agents have their own social-media platform and are publishing AI-generated research papers on their own preprint server.

And This...



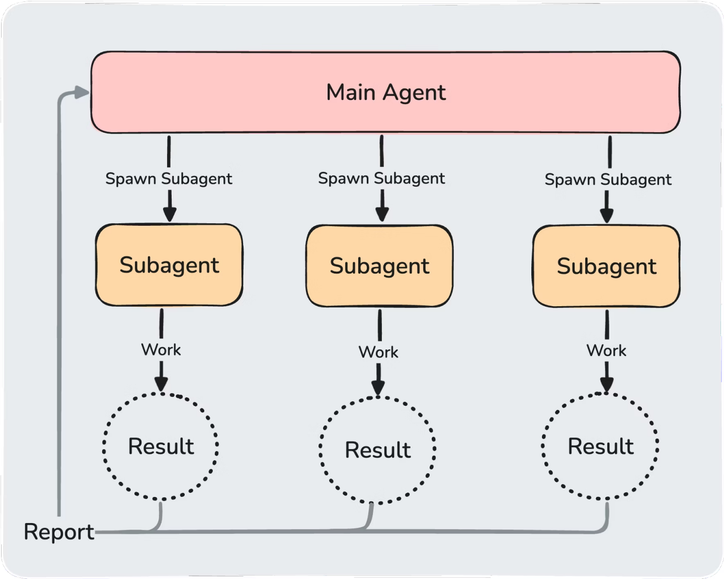
AI

Anthropic releases Opus 4.6 with new ‘agent teams’

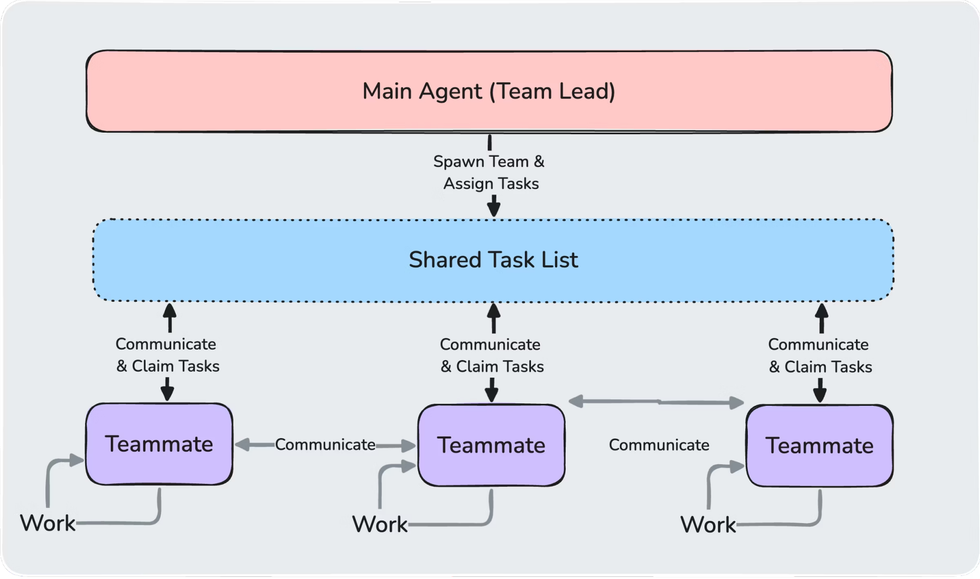
Lucas Ropek — 9:51 AM PST · February 5, 2026



Subagents

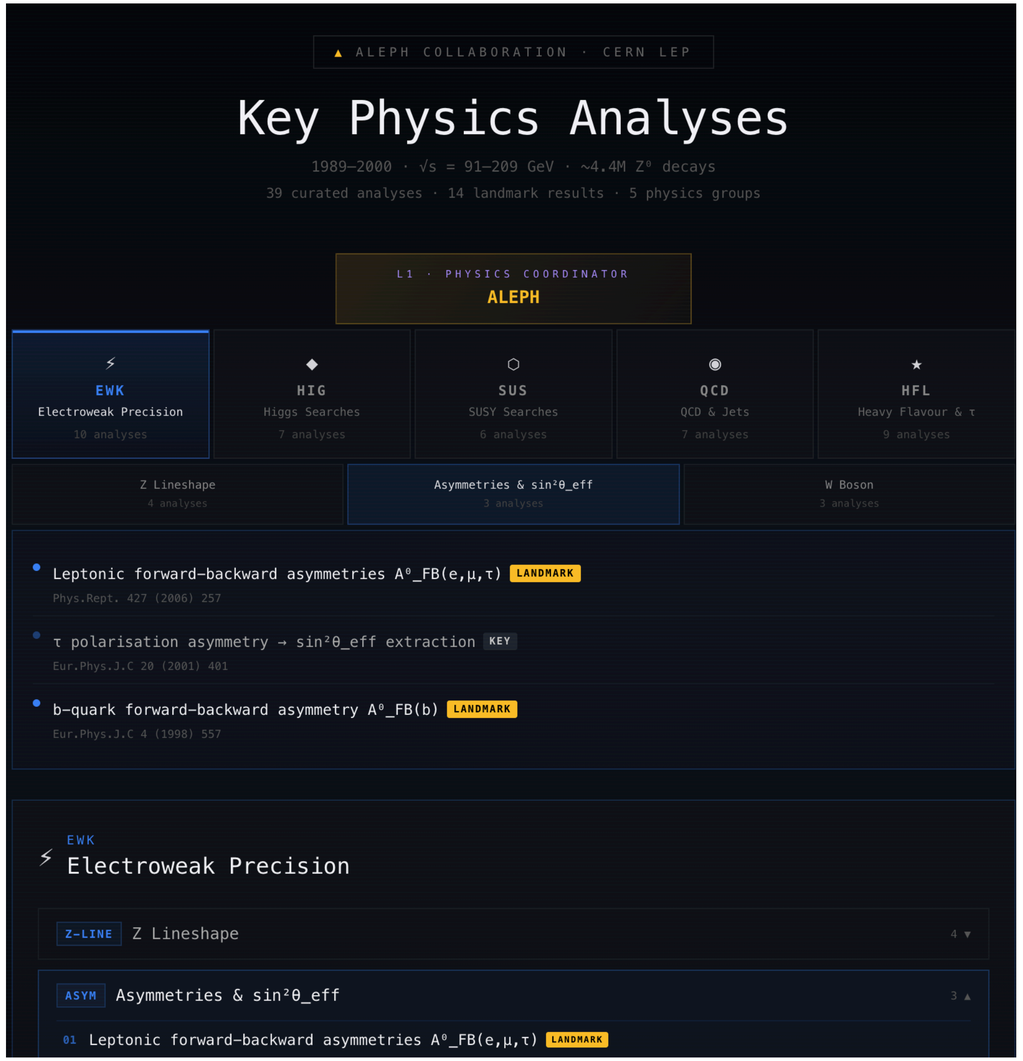


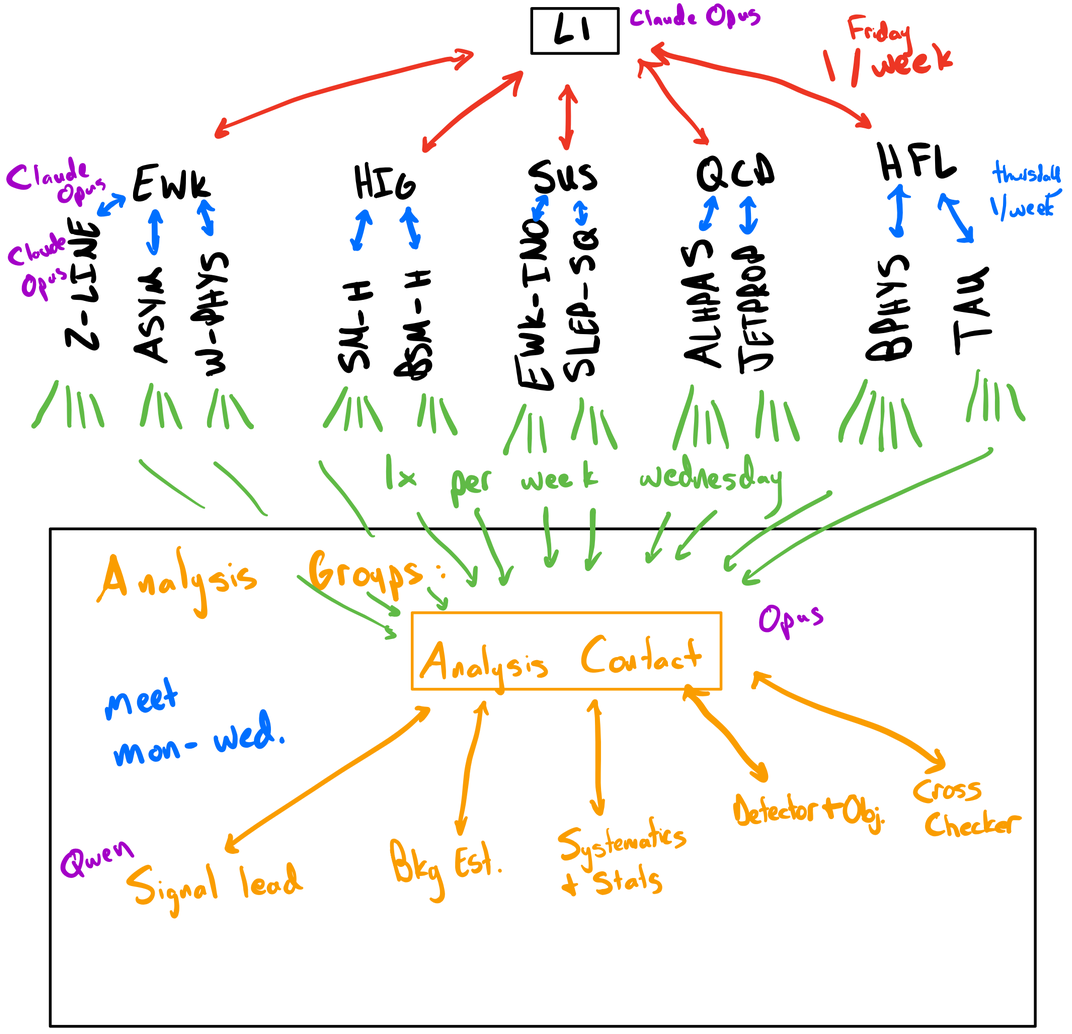
Agent Teams



The Grand Plan

Can I bot an entire collaboration with many subgroups/analyses?





First Attempt: Hadronic Z-Line Shape

Claude

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) |
```

Does an analysis?

- 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

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 $\otimes\, \mathcal{O}(\alpha^2)$ 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 $\sigma^0_{\rm had} = 41.42 \pm 0.06$ 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

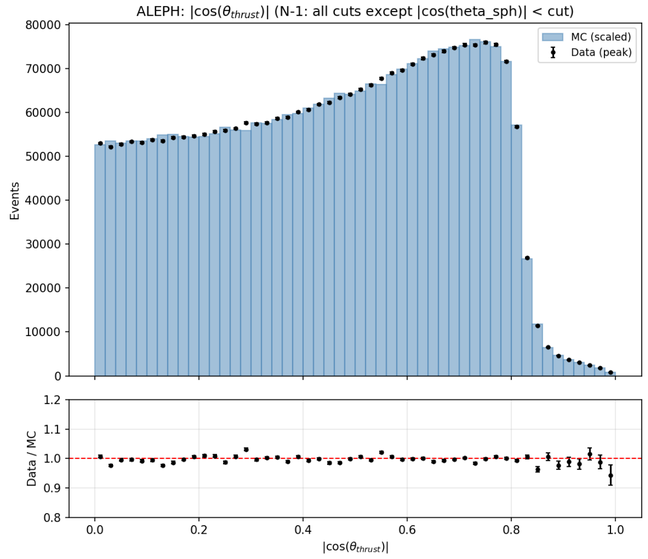
2 Data and Monte Carlo Samples 3
- 2.1 Collision Data . . . 3
- 2.2 Luminosity . . . 4
- 2.3 Monte Carlo Simulation . . . 5

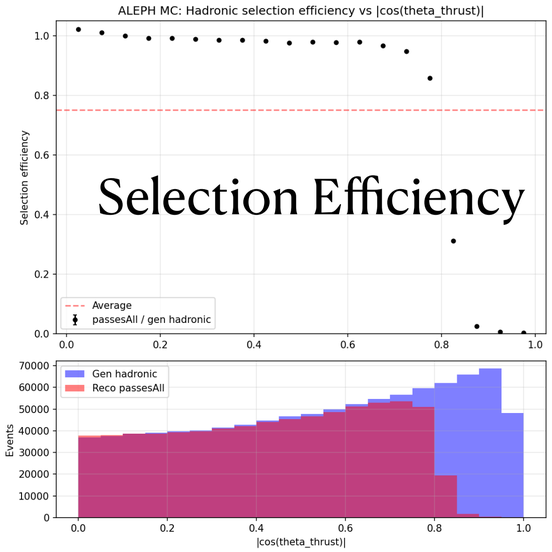
3 Event Selection 5
- 3.1 Good Track Definition . . . 5
- 3.2 Event-Level Selection . . . 6
- 3.3 Validation Against Pre-computed Flags . . . 6
- 3.4 Cut-Flow . . . 6
- 3.5 Selected Events per Energy Cluster . . . 7

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 $e^+e^- \to Z \to \tau^+\tau^-$ 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 $\pi^0 \to \gamma\gamma$ and initial-state radiation, can produce events that satisfy the hadronic selection criteria, particularly the charged multiplicity requirement $N_{\text{trk}} \geq 5$.

The tau-pair contamination is estimated as:

$$f_\tau = \frac{\sigma_\tau}{\sigma_{\text{had}}} \times \frac{\varepsilon_\tau}{\varepsilon_{\text{had}}} \tag{6}$$

where:

- $\sigma_\tau/\sigma_{\text{had}} = \Gamma_\tau/\Gamma_{\text{had}} = 84.08/1744.4 = 0.0482$ 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];
- $\varepsilon_{\text{had}} = 75.07\%$ is the hadronic selection efficiency (Section 5).

This gives:

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

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 $e^+e^- \to Z/\gamma^* \to$ hadrons follows the standard LEP parameterisation [LEP EWWG, Phys. Rept. 427 (2006) 257] with an s-dependent width:

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

where:

- The first term is the Z-exchange Breit-Wigner with running width $\Gamma(s) = \Gamma_Z \cdot s/M_Z^2$;
- $\sigma_{\text{int}}(s)$ 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 $\sigma_{\text{had}}^0 = (12\pi/M_Z^2) \cdot \Gamma_e\Gamma_{\text{had}}/\Gamma_Z^2$.

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 $\delta\sigma_{\text{had}}^0 = 0.010$ 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, $\delta\sigma_{\text{had}}^0 = 0.010$ nb.

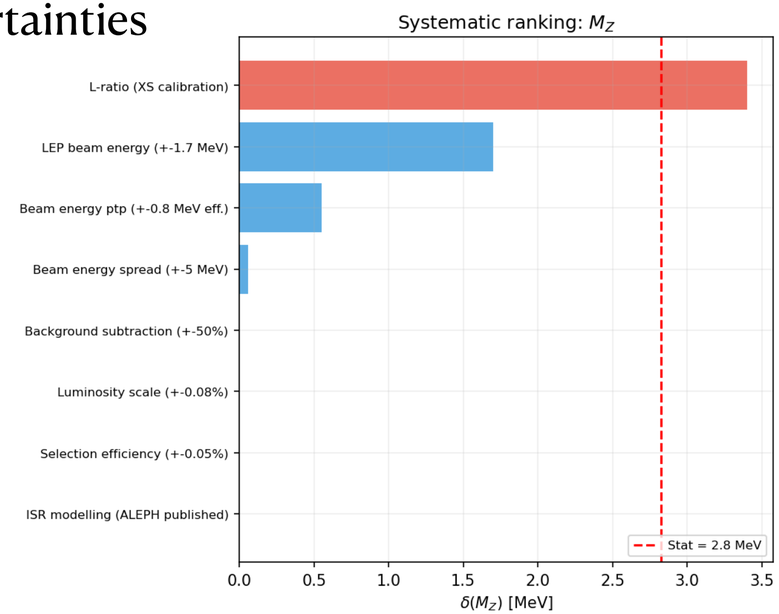
The ISR uncertainty has no effect on M_Z because the ISR convolution preserves the symmetry of the lineshape around the pole. It affects Γ_Z and σ_{had}^0 through the ISR correction to the peak cross-section.

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
$\mathcal{L}(\text{P}-2)/\mathcal{L}(\text{Peak})$	0.15511	± 0.00053 (0.34%)
$\mathcal{L}(\text{Intermediate})/\mathcal{L}(\text{Peak})$	0.00099	± 0.00001 (1.0%)
$\mathcal{L}(\text{P}+2)/\mathcal{L}(\text{Peak})$	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:

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

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}.

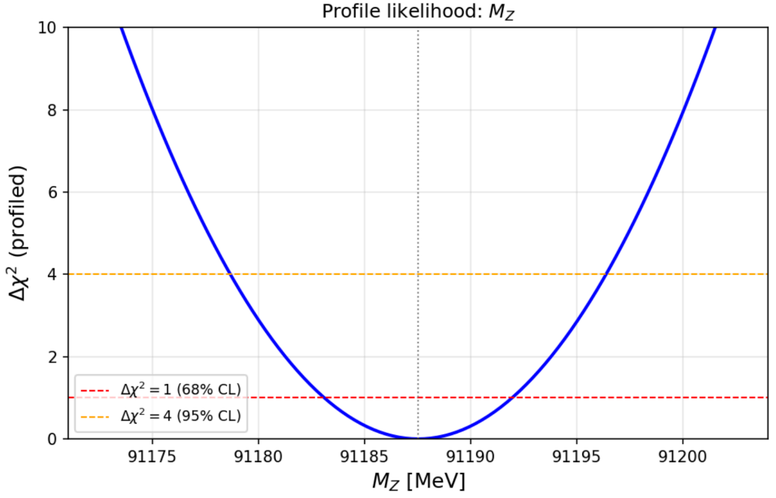
The minimisation is performed using the Nelder-Mead simplex algorithm (scipy `optimize.minimi` with convergence tolerances of 10^{-8} on the parameter values and 10^{-14} on the chi-square.

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

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

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 $\Delta\sigma^0_{\text{had}} = 5 \times 10^{-3}$ 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^0_{had}. The Asimov event counts are:

$$N_i^{\text{Asimov}} = \varepsilon_i \times \mathcal{L}_i \times \sigma^{\text{SM}}_{\text{had}}(\sqrt{s_i}) \tag{20}$$

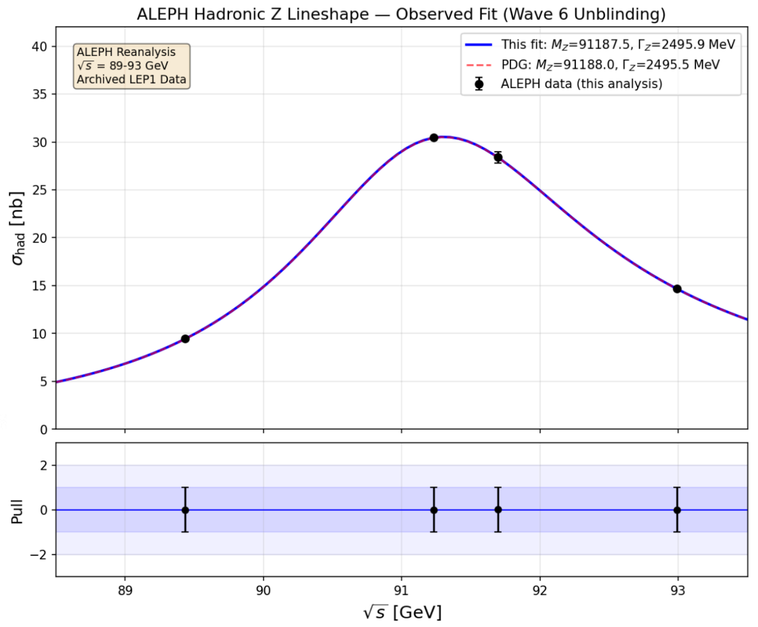
1. **Pre-unblinding review** (Wave 5): The analysis note is drafted using Asimov results. The full analysis chain is frozen: selection, backgrounds, efficiencies, systematics, statistical model.

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
$\delta\sigma^0_{\text{had}}$ (total)	0.055	0.057	0.037

Fit to Data

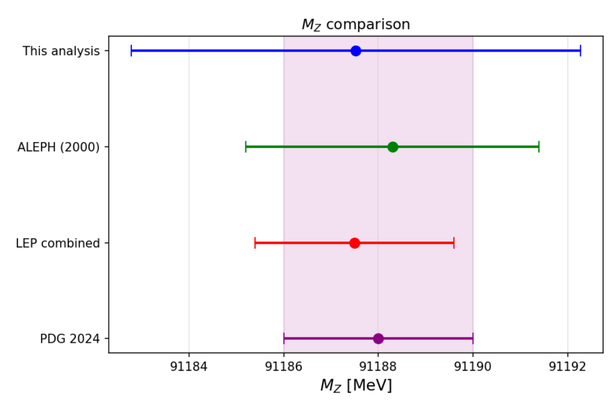


Fit Results & Comparison

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

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

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



Not Half Bad

Fitting Plan

8.4 Chi-Square Fit

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

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

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}.

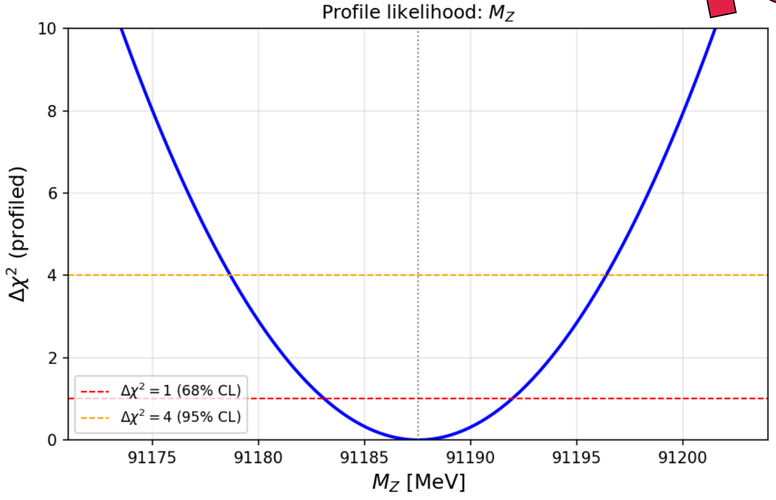
The minimisation is performed using the Nelder-Mead simplex algorithm (scipy `optimize.minimi` with convergence tolerances of 10^{-8} on the parameter values and 10^{-14} on the chi-square.

Statistical uncertainties on the fitted parameters are obtained from the inverse of t[illegible] matrix of the chi-square at the minimum:

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

where $H_{ij} = \partial^2\chi^2/\partial\theta_i\partial\theta_j$ is the Hessian, evaluated numer[illegible]lly usin[illegible]nite differences with step sizes $\Delta M_Z = \Delta\Gamma_Z = 10^{-5}$ GeV and $\Delta\sigma^0_{\text{had}} = 5 \times 10^{-3}$ 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 [illegible]rformed. Cut efficiencies are compared between data and MC, but no [illegible]neshape [illegible]ta is performed.

2. **Background estimation and valid[illegible]ion** (W[illegible]ve[illegible]: [illegible]g[illegible] [illegible]ractions are estimated from publish[illegible]d cross-sections [illegible]nd MC [illegible]ccept[illegible]nce. Data–MC comparisons are performed [illegible]kinemati[illegible]i[illegible]ib[illegible]t[illegible]ut no cr[illegible]s sections are computed from data.

3. **Syst[illegible]a[illegible]c [illegible]aluation** (Wave 4): All systematic uncertainties are e[illegible]ed[illegible]s a **Asimov dataset** — a pseudo-dataset construc[illegible]rom[illegible]S[illegible]ud[illegible]d[illegible]p[illegible]ions at each energy point with the PDG valu[illegible]of M_Z, [illegible]$_Z$ [illegible]d [illegible]. Asimov event counts are:

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

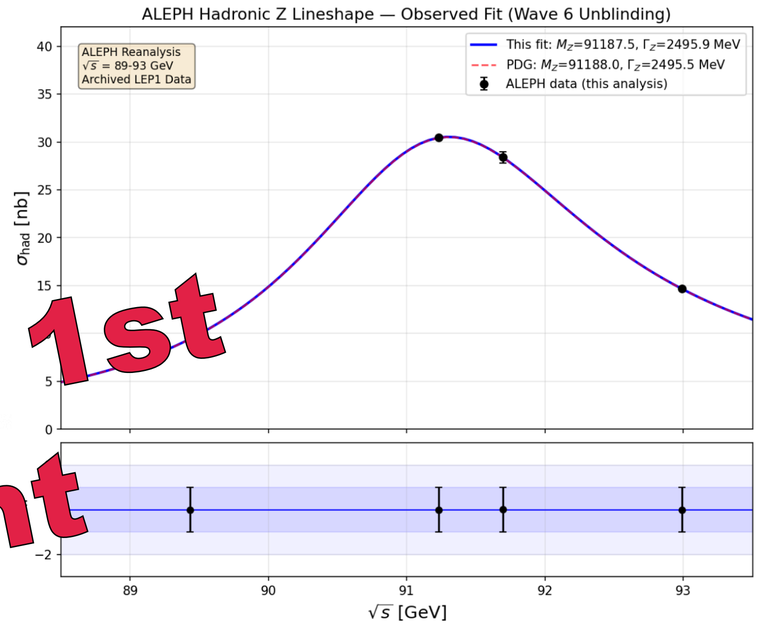
1. **Pre-unblinding review** (Wave 5): The analysis note is drafted using Asimov results. The full analysis chain is frozen: selection, backgrounds, efficiencies, systematics, statistical model.

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
$\delta\sigma^0_{\text{had}}$ (total)	0.055	0.057	0.037

Fit to Data

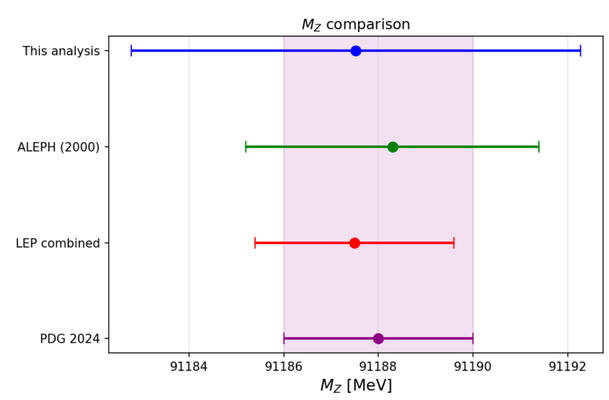


Fit Results & Comparison

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

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

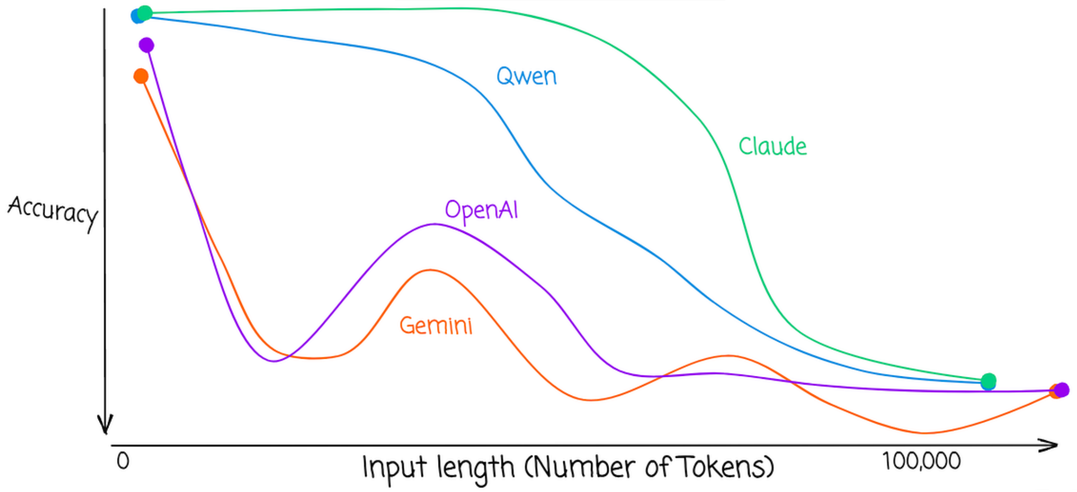
$\sigma^0_{\text{had}} = 41.507 \pm 0.029\ \text{nb}$



Phil's Take: Level of 1st Year Grad Student

Problems to Solve

Context Rot

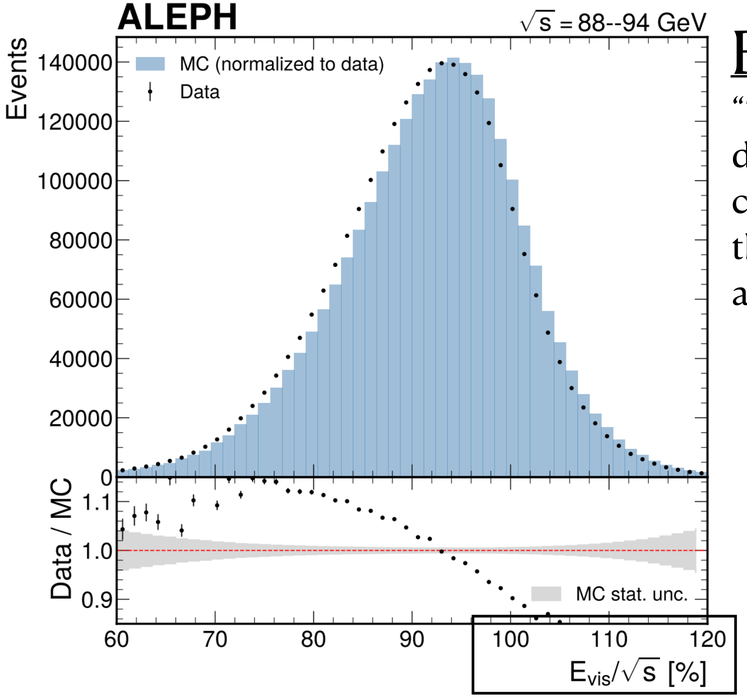


Data Leakage

Measurement of the Z Resonance Parameters at LEP

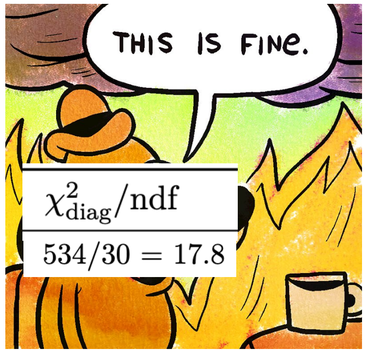
The ALEPH Collaboration

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 $M_Z = (91.1885 \pm 0.0031)$ GeV/c^2, $\Gamma_Z = (2.4951 \pm 0.0043)$ GeV, $\sigma^0_{\rm had} = (41.559 \pm 0.058)$ 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 $\alpha_s(M_Z) = 0.114 \pm 0.004 \pm 0.002_{\rm QCD} + 0.005 \log_{10}\left[\frac{M_{\rm H}}{150\,{\rm GeV}/c^2}\right]$. The lepton pair forward-backward asymmetry is measured to be $A^{0,\ell}_{\rm FB} = 0.0173 \pm 0.0016$ from which the effective weak mixing angle is derived: $\sin^2\theta^{\rm lept}_{\rm eff} = 0.23089 \pm 0.00089$. The measurement of the leptonic width $\Gamma_{\ell\ell} = 84.02 \pm 0.15$ MeV leads to a determination of the effective ρ parameter $\rho^{\rm lept}_{\rm eff} = 1.0064 \pm 0.0018$. The data support the Standard Model and favour a light Higgs.

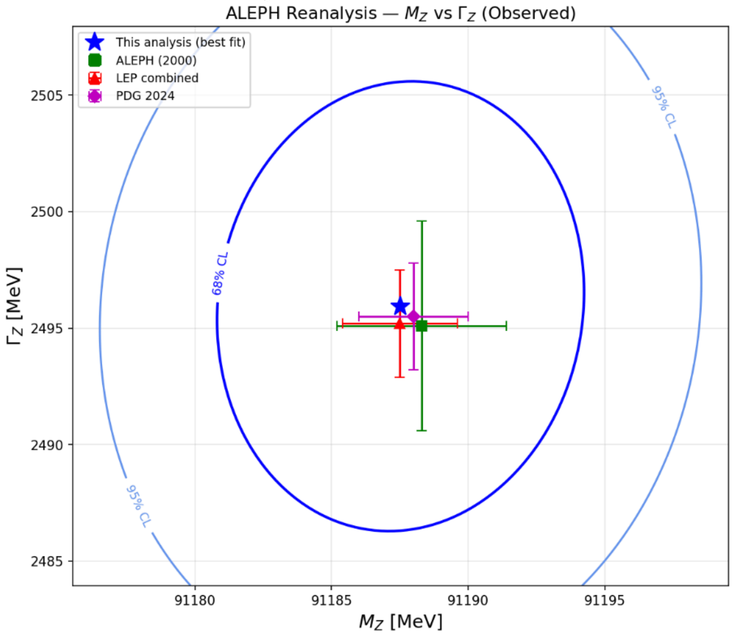


Bad Physics Intuition?

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

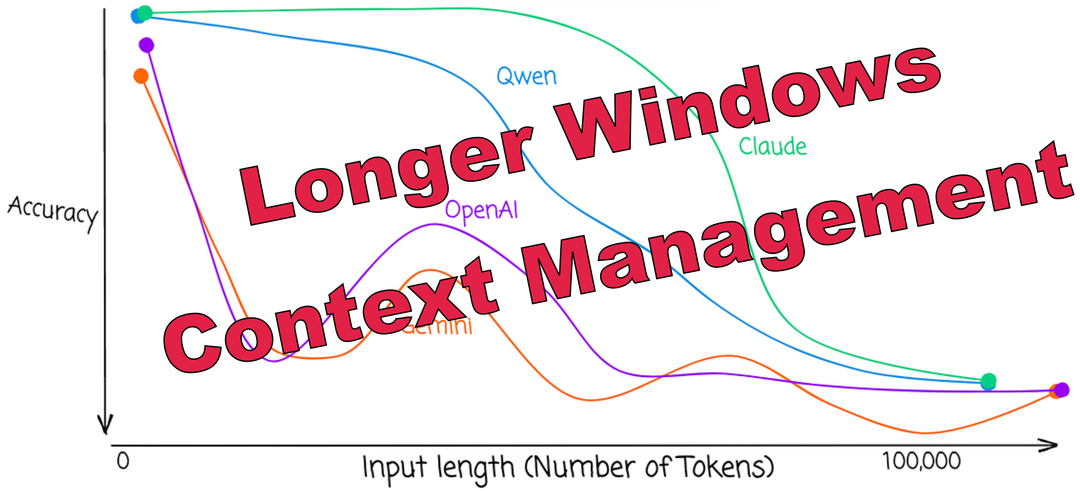


Worse Physics Performance



Problems to Solve

Context Rot



Longer Windows
Context Management

Data Leakage

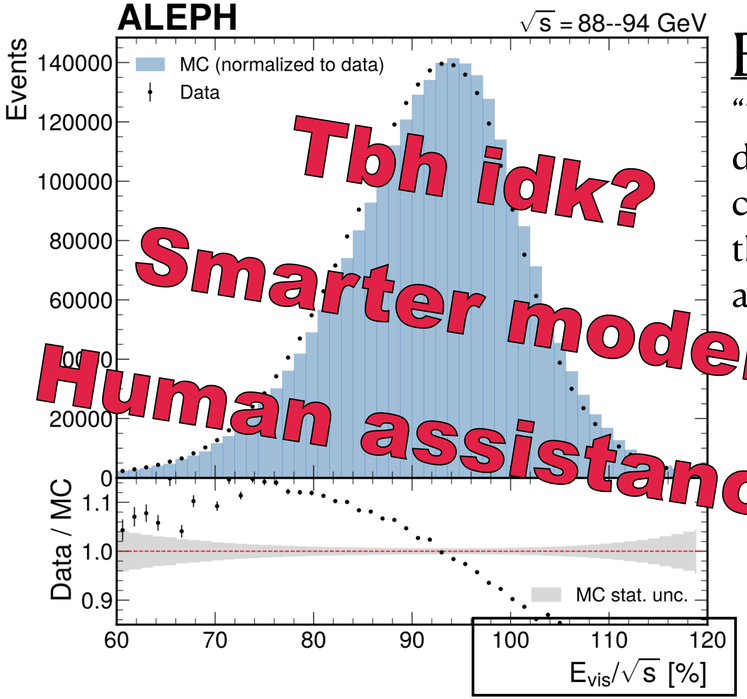
Measurement of the Z Resonance Parameters at LEP

The ALEPH Collaboration

The properties of the Z boson 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 $M_Z = (91.1885 \pm 0.0031)$ GeV/c^2, $\Gamma_Z = (2.4951 \pm 0.0043)$ GeV, $\sigma^0_{\rm had} = (41.559 \pm 0.028)$ 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 $\alpha_s(M_Z) = 0.1190 \pm 0.0027_{\rm exp} \pm 0.002_{\rm QCD} + 0.005 \log_{10}\left[\frac{M_{\rm H}}{150\,{\rm GeV}/c^2}\right]$. The lepton pair forward-backward asymmetry is measured to be $A^{0,\ell}_{\rm FB} = 0.0173 \pm 0.0016$ from which the effective weak mixing angle is derived: $\sin^2\theta^{\rm lept}_{\rm eff} = 0.23089 \pm 0.00089$. The measurement of the leptonic width $\Gamma_{\ell\ell} = 84.02 \pm 0.15$ MeV leads to a determination of the effective ρ parameter $\rho^{\rm lept}_{\rm eff} = 1.0064 \pm 0.0018$. The data support the Standard Model and favour a light Higgs.

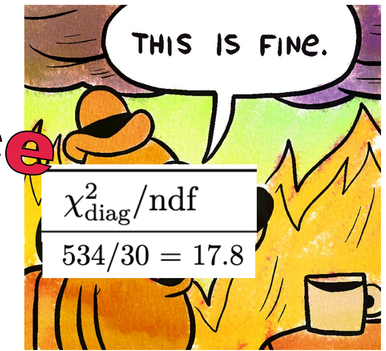
Novel Analyses

Bad Physics Intuition?

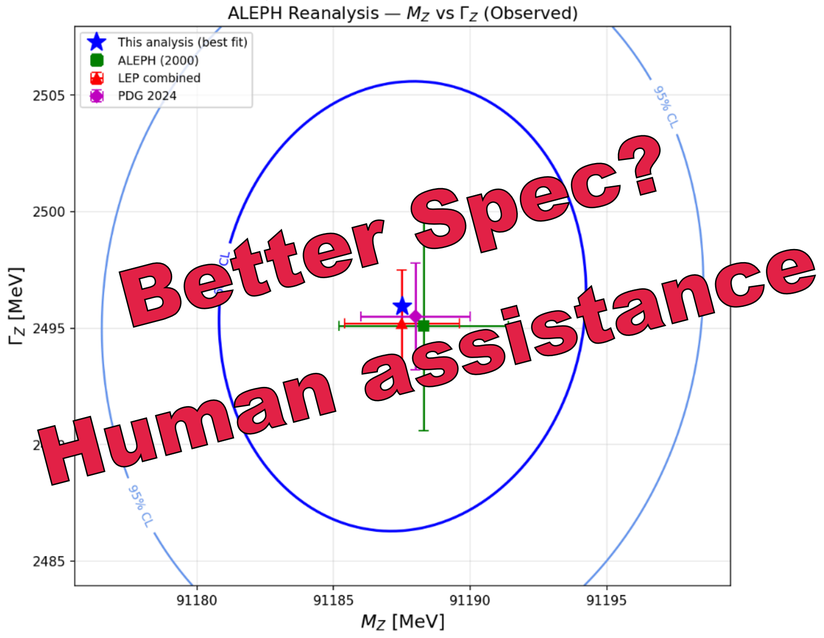


Tbh idk?
Smarter model
Human assistance

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



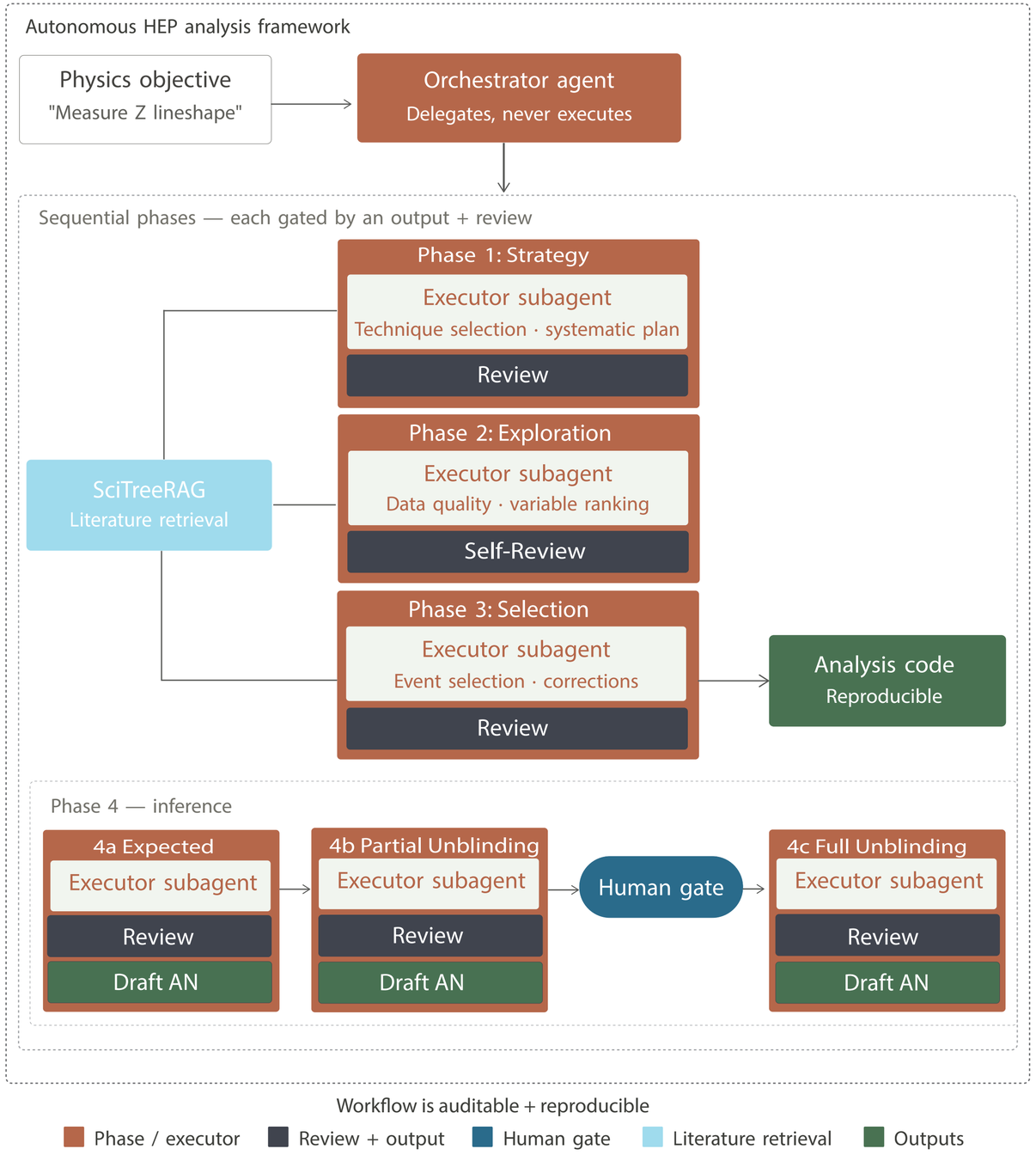
Worse Physics Performance



Better Spec?
Human assistance

https://arxiv.org/pdf/2603.20179

JFC: Just Furnish Context



Analysis Agents

lead-analyst, data-explorer, detector-specialist, theory-scout, signal-lead, background-estimator, systematics-fitter, systematic-source-evaluator, cross-checker, note-writer, ml-specialist

Phases (# Reviews)

Phase 1 — Strategy (4)
Phase 2 — Exploration (1)
Phase 3 — Selection and Background Modeling (1/channel)
Phase 4a — Expected Results (4)
Phase 4b — Partial Unblinding (4)
HUMAN GATE
Phase 4c — Full Unblinding (4)
Phase 5 - Documentation (5)

Literature Retrieval

SciTreeRAG[1]**:** 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 published numerical inputs

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

Reviewer agents

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

https://arxiv.org/pdf/2603.20179
https://github.com/jfc-mit/jfc

Spec & Prompting

- If you get a wrong answer at this point, your "spec" is wrong.
- 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]

https://arxiv.org/pdf/2603.20179
https://github.com/jfc-mit/jfc

Spec & Prompting

- Maybe we went too far?

Accumulated domain knowledge

conventions
- extraction.md
- int_plots.py
- postplotters_ed.py
- preamble.tex
- RESULTS.md
- search.md
- TEMPLATE.md
- unfolding.md

Probably need more supervision on this

Role definitions

agents
- arbiter.md
- bibtex_gal[illegible].md
- constructive_reviewer.md
- critical_reviewer.md
- executor.md
- fixer.md
- investigator.md
- note_writer.md
- physics_reviewer.md
- plot_validator.md

Unclear if we need this

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
- [illegible].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

Too verbose

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

Maybe the phases are too rigid?

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]

Too limited?

https://arxiv.org/pdf/2603.20179
https://github.com/jfc-mit/jfc

Spec & Prompting

- Maybe we went too far?

6.3.1 Evidence-Based Review (mandatory)

Every verification claim in a review must cite specific evidence. The following patterns are NOT accepta

- "Verified" / "Confirmed" / "Checks out" — without a number or reference
- "Looks reasonable" / "Appears correct" — without stating what was checked
- "Consistent with expectations" — without stating the expectation and the comparison metric

Acceptable patterns:

- "Closure chi2/ndf = 1.3/36 (p = 0.24) from results/closure.json — PASS"
- "Tracking systematic: 0.95% (artifact Table 7) matches 0.95% (results/systematics.json) — consist
- "Figure 12 ratio panel: all bins within [0.85, 1.15] — acceptable data/MC agreement"

This requirement applies to ALL reviewer roles (physics, critical, constructive, rendering, plot validator). fine" without evidence is a failed review — it provides no information to the arbiter and no audit trail for

Reviewers who find themselves unable to cite evidence for a "verified" claim should recognize this as a verified the claim, and should go back and check.

6.3.2 Reviewer Investigation Subagents

Reviewers may spawn focused investigation subagents when a concern requires deeper analysis than r cases:

- Tracing a computed value through multiple scripts to verify correctness
- Checking whether a systematic variation is correctly propagated through the analysis chain (readin
- Verifying that intermediate outputs are consistent (e.g., same efficiency used across scripts)

Phase 4a (Expected Results) review focus

Systematic completeness: Every source from conventions + references implemented or formally downscoped? Signal injection/closure passes? Operating point stable (Category A if not)? If MC covers only a subset of data periods: is the scaling documented, are uncertainties inflated, are MC stat correlations between shared periods handled correctly (§12)?

External input audit: Is the external input table present (§2.5)? For each borrowed published value, is data-driven measurement genuinely infeasible? Could the analysis reformulate to avoid the external input (e.g., ratio extraction instead of absolute if total cross-section is missing)? If >50% of total uncertainty is from external inputs, flag.

Phase 2 traceability: Every Phase 2 commitment implemented or formally downscoped. Cross-check COMMITMENTS.md (populated at Phase 2, updated at each phase boundary).

Data quality presumption: If the executor claims data is unusable, the reviewer must verify: (1) what specific technical failure was demonstrated, (2) what calibration/correction was attempted, (3) whether published analyses used the same data. Concluding "data is poor" when published papers extracted results from it is Category A — extraordinary claims require extraordinary evidence (see §1, §12).

Closure test alarm bands: chi2/ndf < 0.1 is Category A (suspicious); chi2/ndf > 3 or any pull > 5-sigma is Category A (failure); `passes: false` in JSON while text claims acceptable is Category A (misrepresentation).

Systematic implementation audit: For each systematic, verify the variation input actually changes, the impact is non-zero and has the expected sign, and the evaluation level is consistent (gen vs reco).

Zero-impact diagnostic requirement. When any systematic has zero (or numerically negligible) impact across ALL bins, the executor must produce two concrete artifacts before the result can be accepted:

1. **Overlay figure:** Nominal and varied distributions on the same axes with a ratio panel (varied/nominal). This figure must be present in the phase artifacts — a prose claim of "verified" without the figure is Category A.
2. **Maximum absolute difference:** Report max|varied_i – nominal_i| across all bins. This number determines the verdict:
   - Exactly 0.0 (machine precision) → the variation was NOT applied. This is a code bug: missing branch, identical files, no-op weight. **Category A — investigate and fix.**
   -) or smaller → genuine numerical negligibility. Document the max|diff| value in the artifact. Acceptable.

6.4.2 Figure Review

Figure review has two components that catch different failure modes:

Code linter (batch, automated). A script or pixi task that greps all plotting scripts for mechanical violations of `appendix-plotting.md`. Catches: absolute `fontsize=N` values, wrong colorbar patterns, missing `hspace=0`, `figsize` violations, `ax.set_title()`, `data=Fals` `label`/`text` stacking. This runs before the review agent and produces a machine-readable report. Any violation is auto-Category A. linter does NOT read rendered figures — it reads code.

Visual validator (agent, reads rendered PNGs). The plot validator agent must **read every rendered figure image** (PNG) and assess vis quality as a human referee would. This is NOT a code review — the agent looks at the picture. The visual checklist:

- [ ] All text (labels, legends, tick marks, annotations) legible at the rendered size in the AN? Text that is technically present but unreadal at 0.45\linewidth is Category A.
- [ ] Legend overlaps with data points, fit curves, or other content? Category A.
- [ ] Any text elements collide with each other (e.g., experiment label stacking, e.g., "ExperimentSimulation Expected")? Category A.
- [ ] Axis labels and legend entries use publication-quality names, not code variable names ("efficiency_energy_dep" → Category A)?
- [ ] Subplot layout suits the content? (Horizontal bar charts need width; narrow panels with long labels are unreadable.)
- [ ] Axis ranges appropriate? (No excessive whitespace, no clipped content, data fills the plot area.)
- [ ] Ratio panel gap? (Any visible gap between main and ratio panels is Category A.)
- [ ] **Physics content check (Category A if failed).** The visual validator and ALL prose reviewers must look at every figure and check whether the physics makes sense — not just whether the formatting is correct. Concrete red flags:
  - Data/MC normalization mismatch: MC visibly above or below data by more than ~20% across the bulk of the distribution. This I indicates a normalization bug, wrong cross-section, or missing background — investigate before attributing to systematics.
  - Postfit data/MC worse than prefit: if the fit is supposed to improve agreement but the postfit plot shows larger residuals, the fit broken.
  - Empty bins where events are expected, or vice versa.
  - Ratio/pull panel systematically above or below 1.0/0.0 across all bins (global offset, not a shape effect). Diagnose the pattern:
    - **Uniform offset in y (MC consistently above/below data):** likely a cross-section prediction issue, or a missing scale factor

6.7 Phase Regression

Regression must be triggered, not avoided. Do not paper over upstream problems. Concrete triggers (must not be rationalized away):

- Data/MC disagreement on observable or MVA inputs
- Closure test failure (p < 0.05)
- Stress test failure without successful remediation (see §3 validation failure remediation)
- Operating point instability
- Unexplained dominant systematic (a single source exceeding 80% of total uncertainty warrants investigation, not acceptance)
- MC used for periods without corresponding simulation
- Result > 3-sigma from a well-measured reference value (see §6.8)
- Result with relative deviation > 30% from a well-measured reference value, regardless of pull (see §6.8 — triggers calibration-first investigation)
- GoF toy distribution inconsistent with observed chi2 (the observed chi2 falls outside the 95% interval of the toy distribution)
- Wholesale bin exclusion: flat-prior gate or similar criterion excluding > 50% of measurement bins (indicates binning or method problem, not a systematic effect — see §3 Phase 3)
- Two distributions that should be independent appear visually identical (indicates a bug or tautological comparison)
- Systematic double-counting: two or more systematic sources produce numerically identical per-bin shifts (indicates they are the same underlying effect with different names — merge and document)
- Correlated combination error: when combining results from multiple observables or methods, shared systematic sources (e.g., scale variation, generator choice) must be treated as correlated. Treating correlated uncertainties as uncorrelated artificially reduces the combined uncertainty.

Procedure:

1. Document issue, identify origin phase
2. Classify as regression trigger in review artifact
3. Orchestrator spawns **Investigator**: reads trigger → reads origin artifact → traces forward to identify affected phases → produces `REGRESSION_TICKET.md`

Physics Referee Review (Sonnet)

You are a journal referee. Assess whether the physics makes sense.

Look at the data shape FIRST, then read axis labels and reason about expectations.

- Does the observable behave as expected?
- Are values in a physically reasonable range?
- If there is a functional fit, does it describe the data? Is the functional form appropriate? Can you read the fitted parameters from the plot either from the legend or extra text box (at least the key ones)?
- If there is an expectation displayed (MC simulation, theory values), does the data match the expectation? In general? Is the shape right, but something is subtly off like constant yield offset or both look gaussian, but not aligned?
- If there's a pull/residual/ratio panel
  - Are pulls randomly distributed around 0?
  - Are ratios randomly distributed around 1?
  - If not do you see any obvious patterns? Constant offset? Particular structure, say a wave or a slope?
- Any red flags? Does it look like the authors missed a calibration or made a mistake?

Output format

Separate your feedback into two sections:

ANALYSIS ISSUES

Problems with the physics/data itself that cannot be fixed by editing the plot script. These indicate bugs in the upstream analysis — wrong functional form, suspicious values, failed fits, miscalibrations, missing corrections, etc. If none, say "None."

PLOT ISSUES

Problems with how the physics is presented that CAN be fixed in the plot script. Missing error bars, wrong axis labels, missing pull panel, fit results not shown, etc. If none, say "None."

https://arxiv.org/pdf/2603.20179
https://github.com/jfc-mit/jfc

Spec & Prompting

- If you get a wrong answer at this point, your “spec” is wrong.
- 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.
```

Good Prompt (Inject Physics Knowledge)

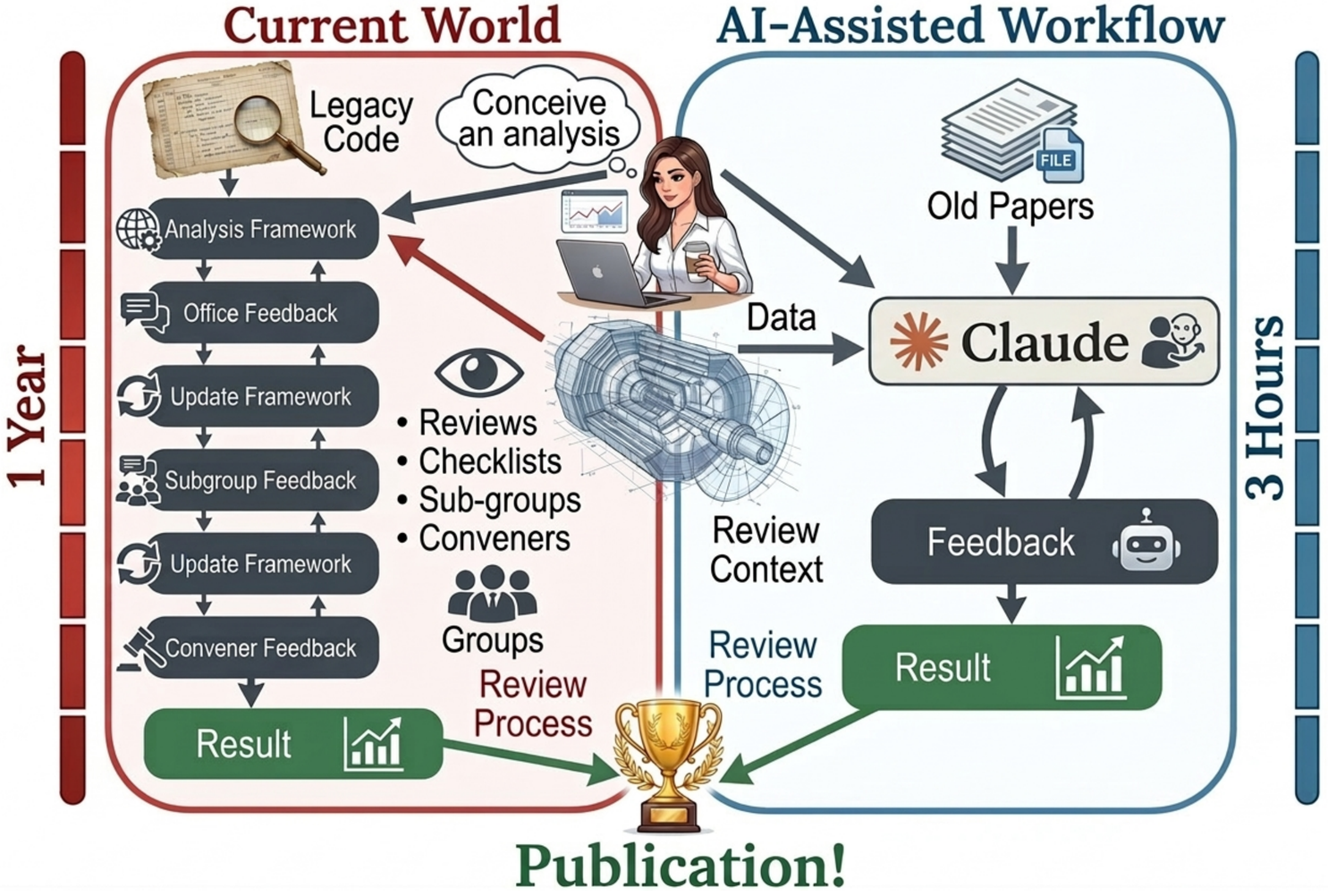
```
# Analysis Title: H→ττ Analysis with CMS Open Data (μτ_h final state)

# 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 basline 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
```

- 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 money!

https://arxiv.org/pdf/2603.20179

JFC: Just Furnish Context



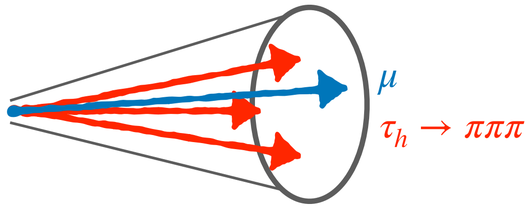
Analysis	Initial prompt (abridged)	Data	Time	Ref.
$H \to \tau\tau$	Measure the Higgs boson signal strength in the $\mu\tau_h$ final state at $\sqrt{s} = 8$ TeV	CMS	6h 30m	A/Github
Z lineshape + α_s	Measure Z lineshape parameters (M_Z, Γ_Z, σ^0_{had}), N_ν from invisible width, and extract $\alpha_s(M_Z)$ 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^b_{FB}	Measure R_b, R_c, and A^b_{FB} 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

E.g. H → ττ Open Data Analysis

e.g. boosted $\mu\tau_h$ https://arxiv.org/pdf/2603.20179

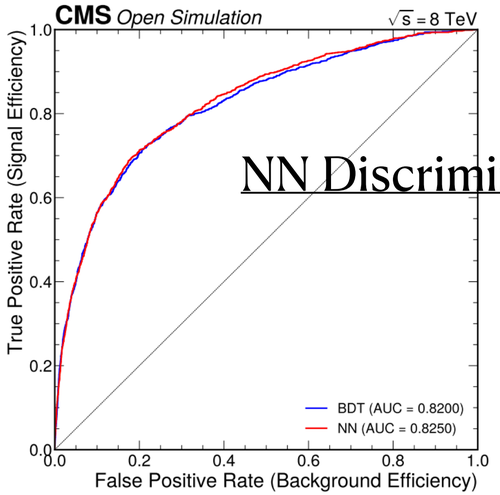


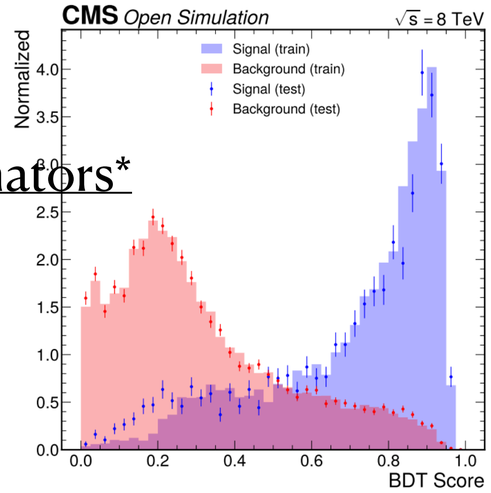
Topology-Driven Categorization: 1 jet, 3 jet, VBF event topologies*

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

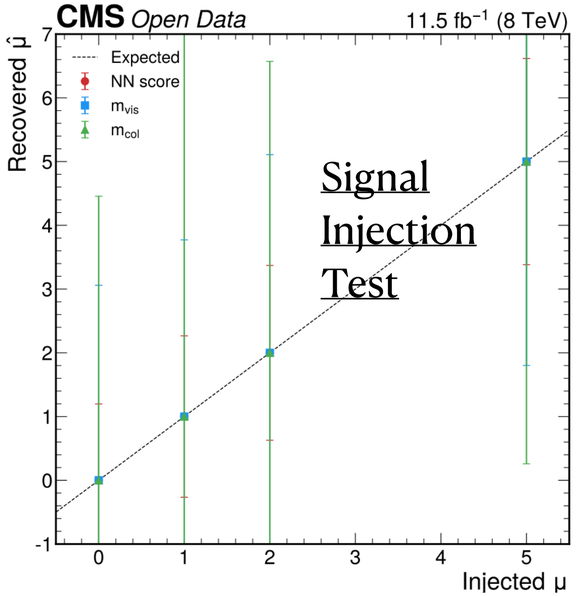
*Similar strategy to ATLAS 7/8TeV!

NN Discriminators*

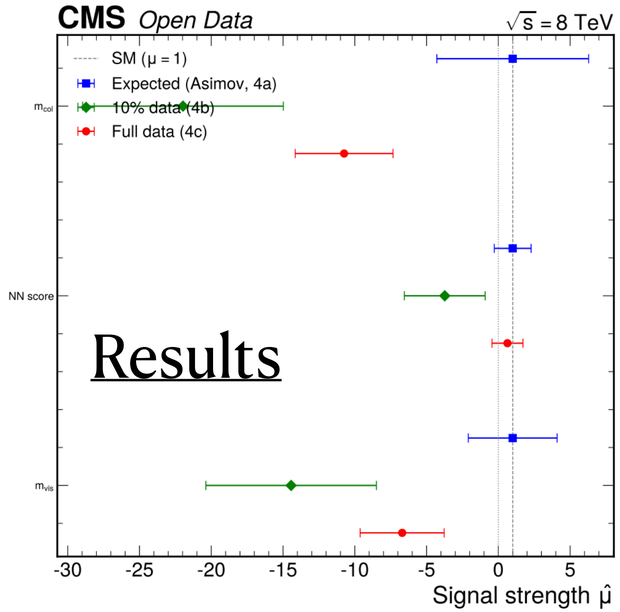


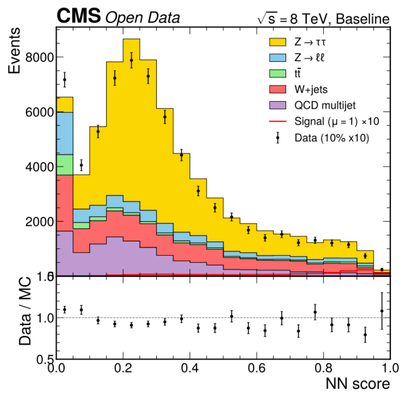


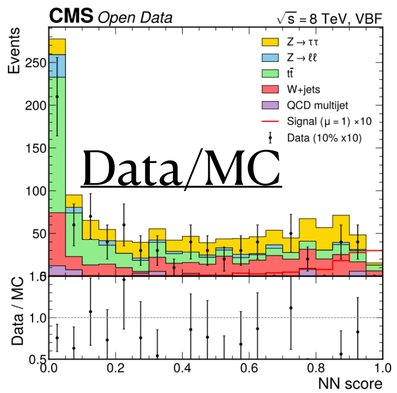
Signal Injection Test



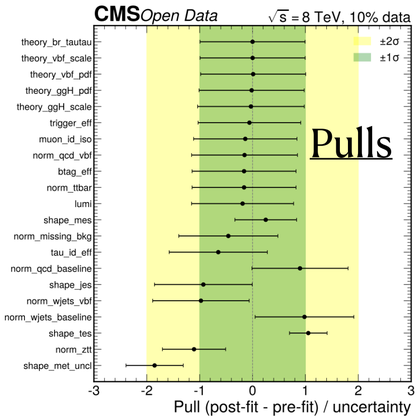
Results



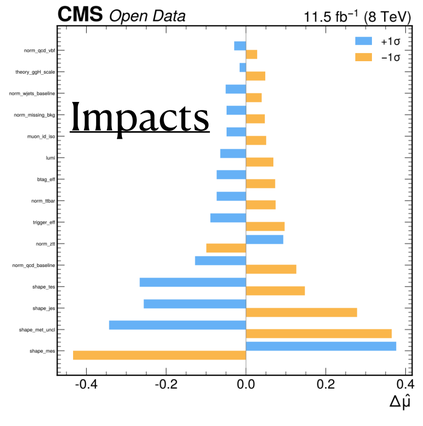




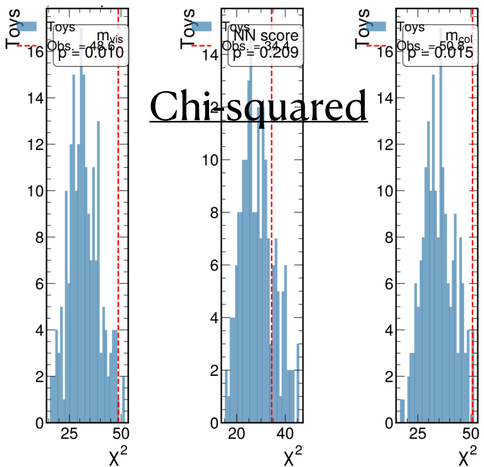
Data/MC



Pulls



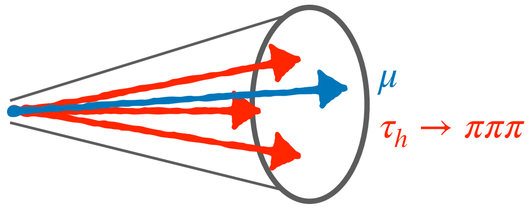
Impacts



Chi-squared

E.g. H → ττ Open Data Analysis

e.g. boosted $\mu\tau_h$ https://arxiv.org/pdf/2603.20179

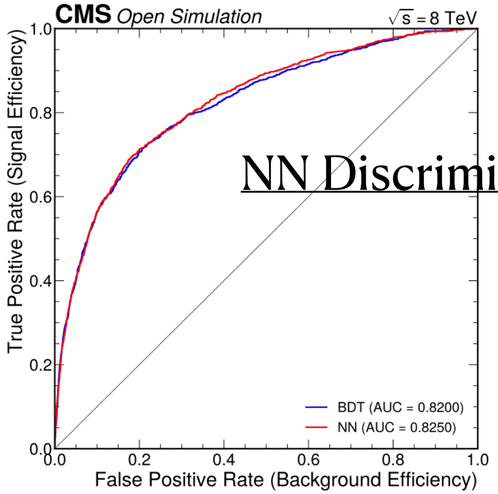


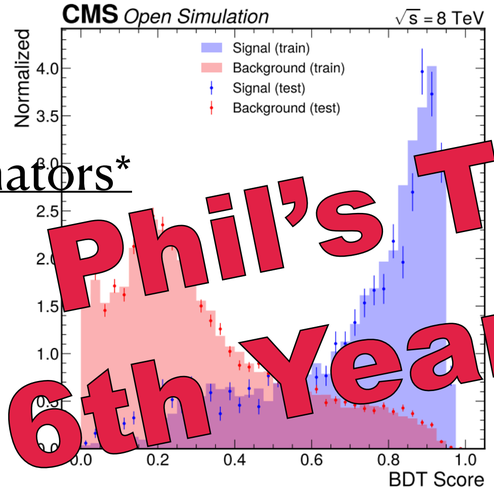
Topology-Driven Categorization: 1 jet, 3 jet, VBF event topologies*

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

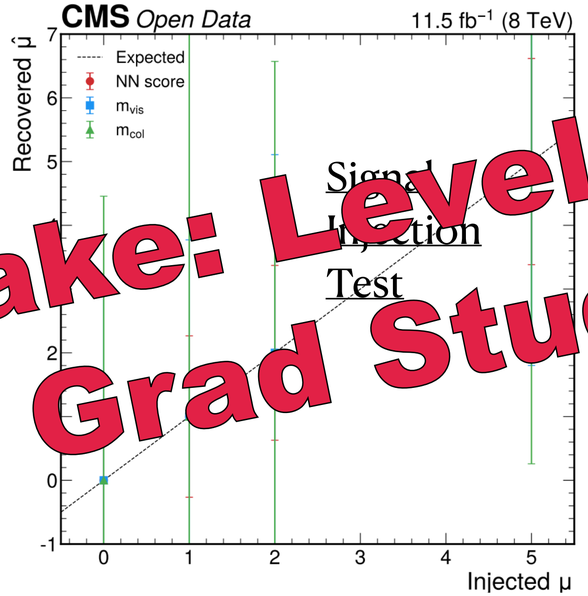
*Similar strategy to ATLAS 7/8TeV!

NN Discriminators*

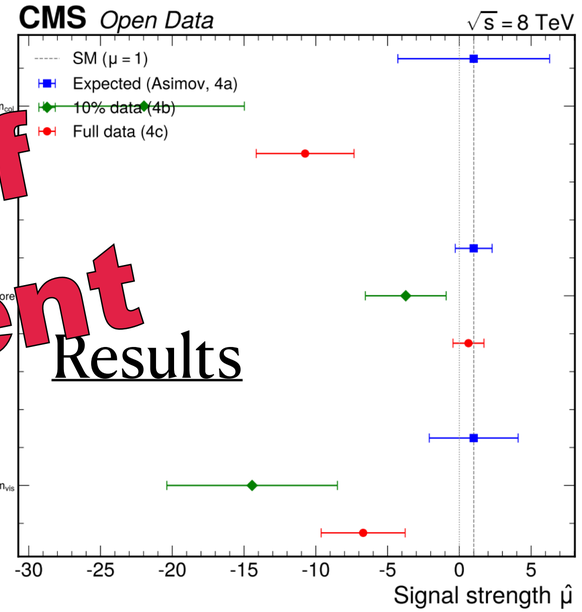




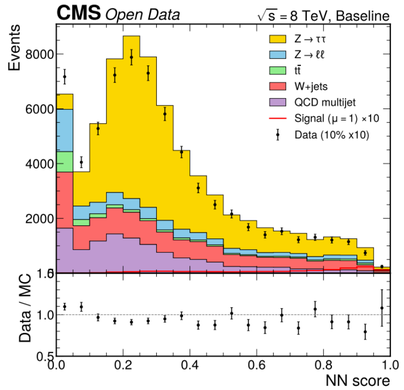
Signal Injection Test

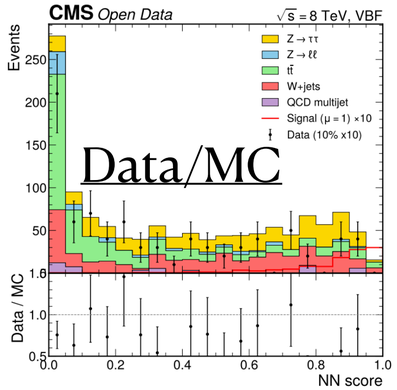


Results

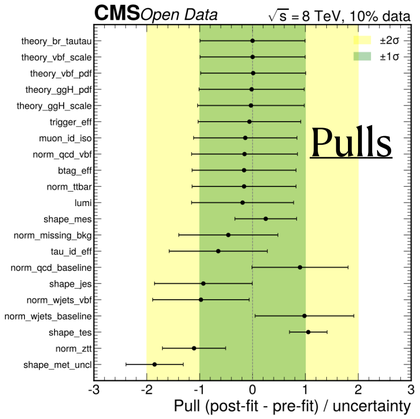


Phil's Take: Level of 6th Year Grad Student

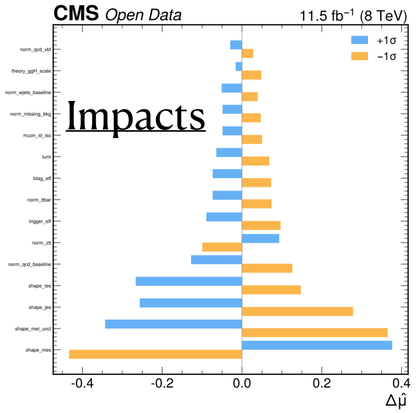




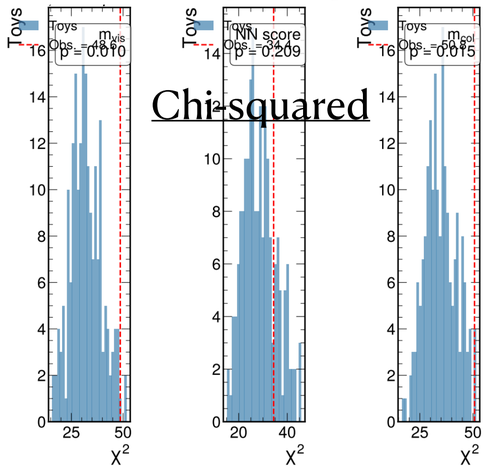
Data/MC



Pulls



Impacts



Chi-squared

Haters arrive

[redacted] 300 pages dumped by an llm

#← Also sent to the channel

[redacted]

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)

[redacted] the CMS example seems to be ~3k lies of physics code and `phase4/scripts` in https://github.com/jfc-mit/higg

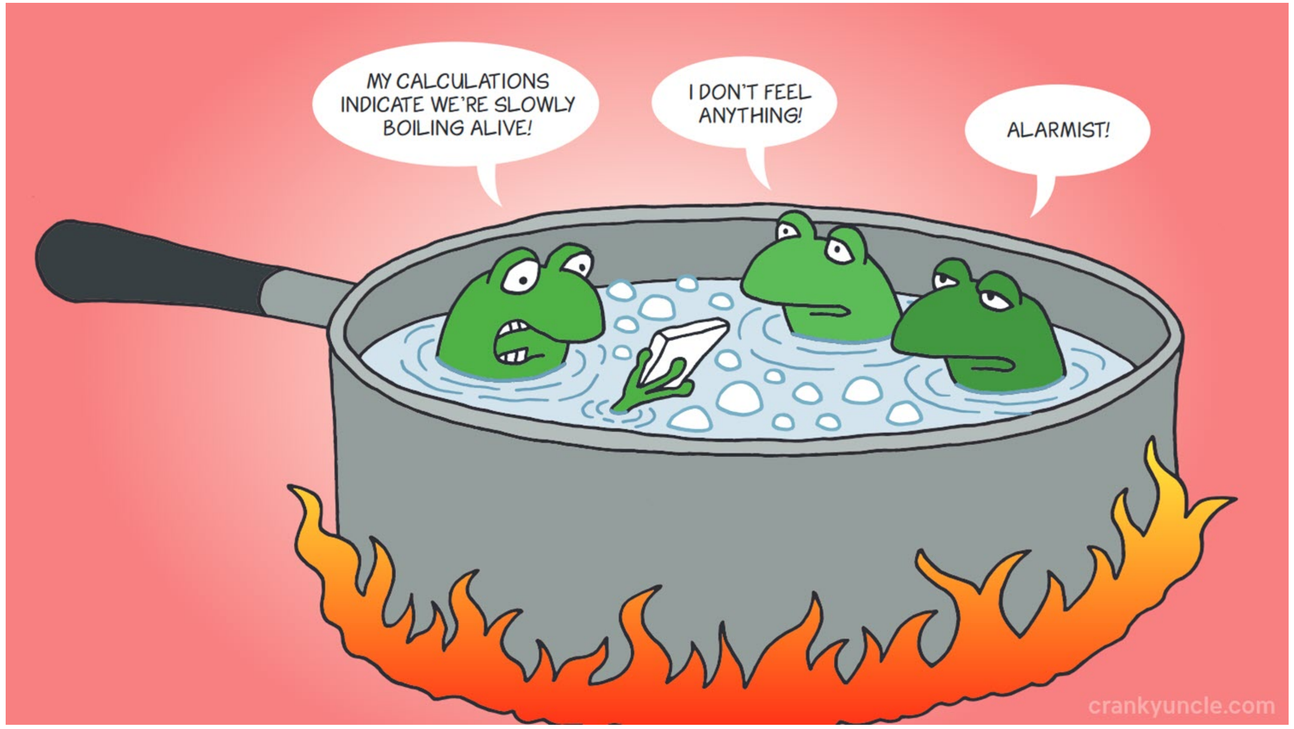
[redacted]

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

[redacted]

This paper is un reviewable

I don't want to see it



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

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.

Where you fit in!

- This does not mean that human physicists are no longer needed - quite the contrary. The role of the **physicist shifts from implementer to architect and critic.**
- Time not spent programming can be **dedicated to broadening theoretical and statistical foundations, building instrumentation expertise, planning projects, keeping up with the literature, and writing up results.**
- **Technical proficiency in programming must remain a critical** and carefully developed skill for all research practitioners.
- Take your research from 1 project -> 5 projects instantaneously through automation & agentic AI (or not)
- Alternatively: Finish 1 project in 1/5 the time! (or not)

Problems →Improvements

Model Laziness

11 Future directions

The following concrete steps are required to bring this analysis to publication quality:

11.1 Production-scale data processing

- Process all 427 LEP1 run directories ($ $3.5 million hadronic Z decays) using SLURM array jobs (estimated $ $4 hours).
- Process all 275+ QQPS MC files ($ $275,000 events; estimated$ $1 hour).
- Evaluate compatibility of additional QQPS versions (v99e, v98e) to increase MC statistics to $ $965,000 events.

11.2 NNLO+NLLA theory implementation

- Implement NNLO fixed-order coefficients from Gehrmann, Glover, and Heinrich (JHEP 0812, 2008).
- Implement NNLL resummation matched to NNLO for ρ_H, B_W, B_T, C, y_{23} from Dissertori, Gehrmann, Glover, and Luisoni (JHEP 0908, 2009).
- Implement N3LL thrust predictions from Abbate, Fickinger, Hoang, Mateu, and Stewart (Phys. Rev. D83, 2011).
- Compare FOPT and CIPT prescriptions.

Model Diversity

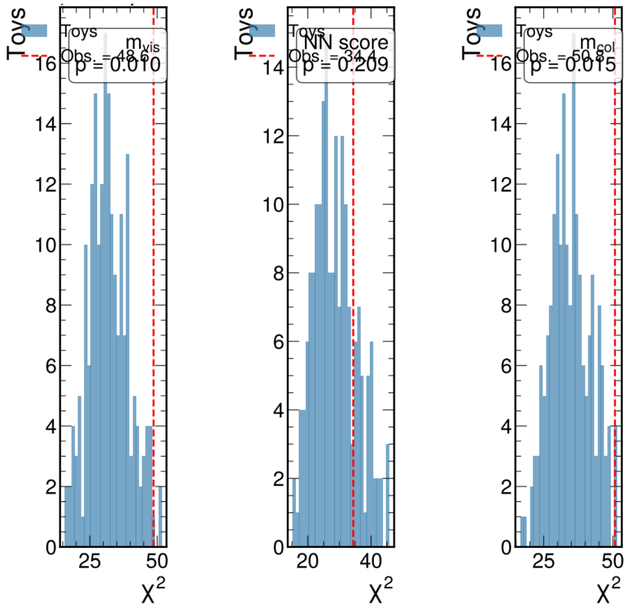
Claude

ChatGPT

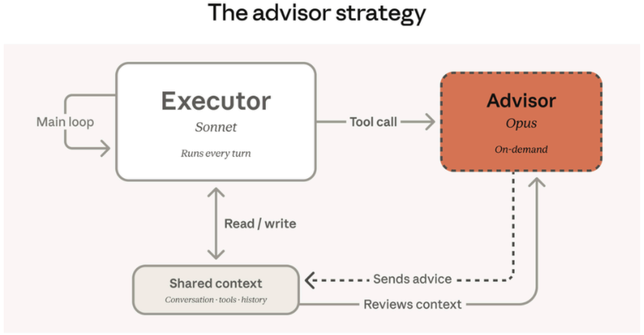
Gemini

Gemma 4

Stupid Plots



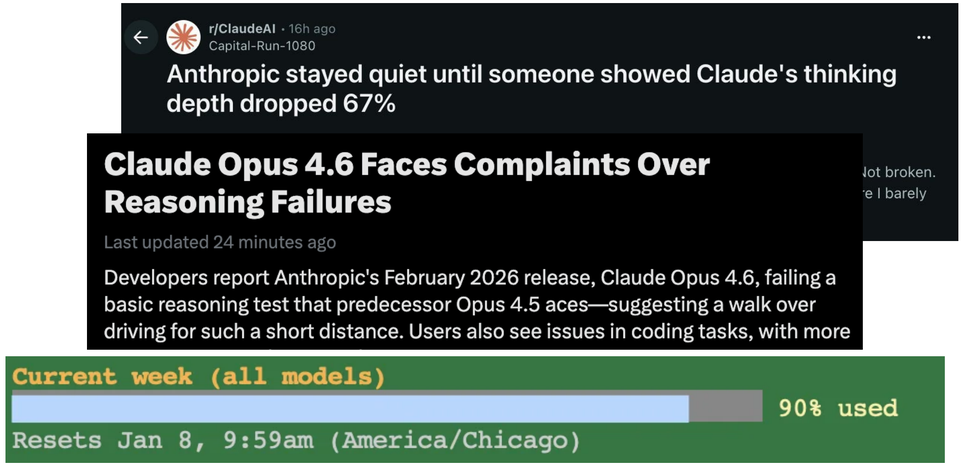
Emerging Technologies/Specs





https://github.com/obra/superpowers
https://github.com/getcompanion-ai/feynman

Claude NERFed due to demand



https://red.anthropic.com/2026/mythos-preview/

Claude Capybara in the room: Mythos

Project Glasswing

Securing critical software for the AI era

aws · ANTHROPIC · cisco · CROWDSTRIKE · THE LINUX FOUNDATION · Microsoft · BROADCOM · Google · JPMorganChase · NVIDIA · paloalto NETWORKS

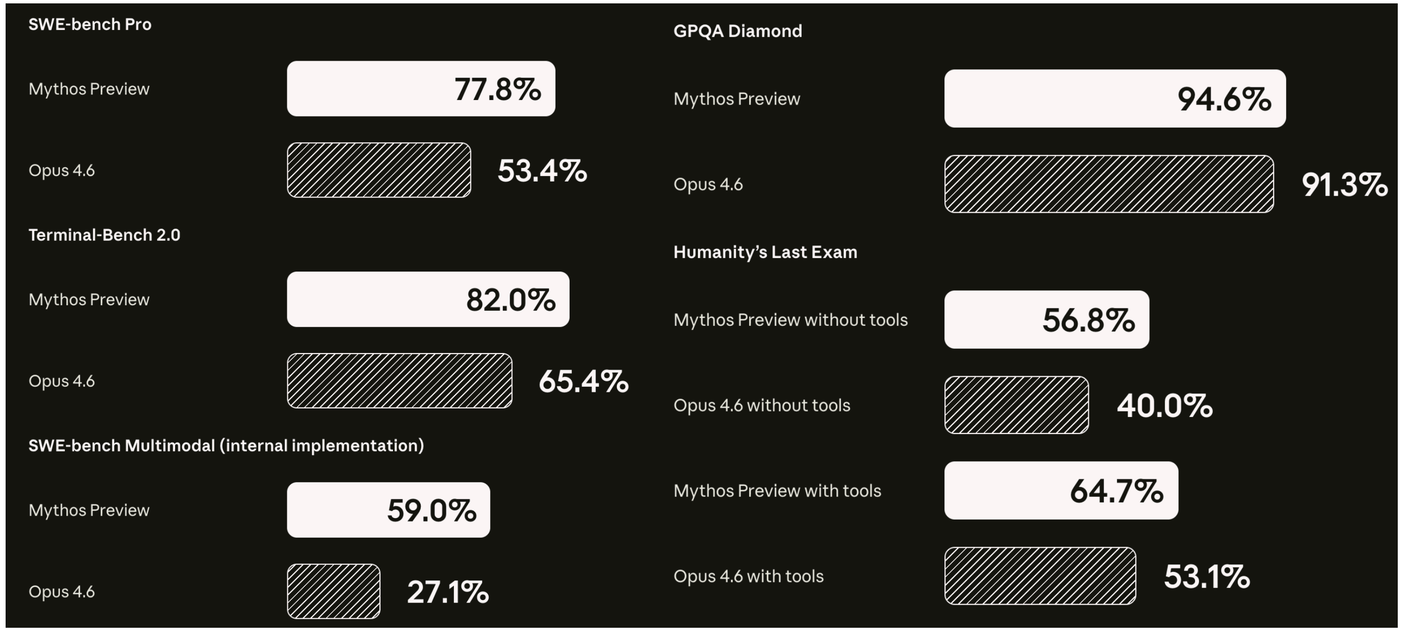
Anthropic's 'Claude Mythos' model sparks fear of AI doomsday

Claude Mythos Is Everyone's Problem

What happens when AI can hack everything?

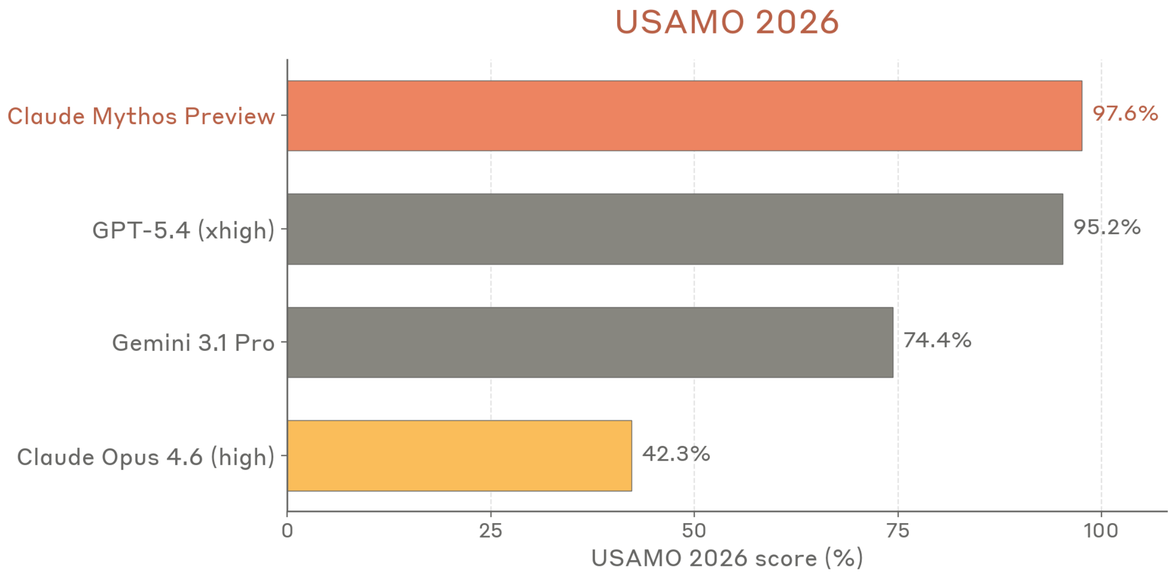
Behind the Curtain: AI's looming cyber nightmare

What Is Claude Mythos—And Why Anthropic Won't Let Anyone Use It

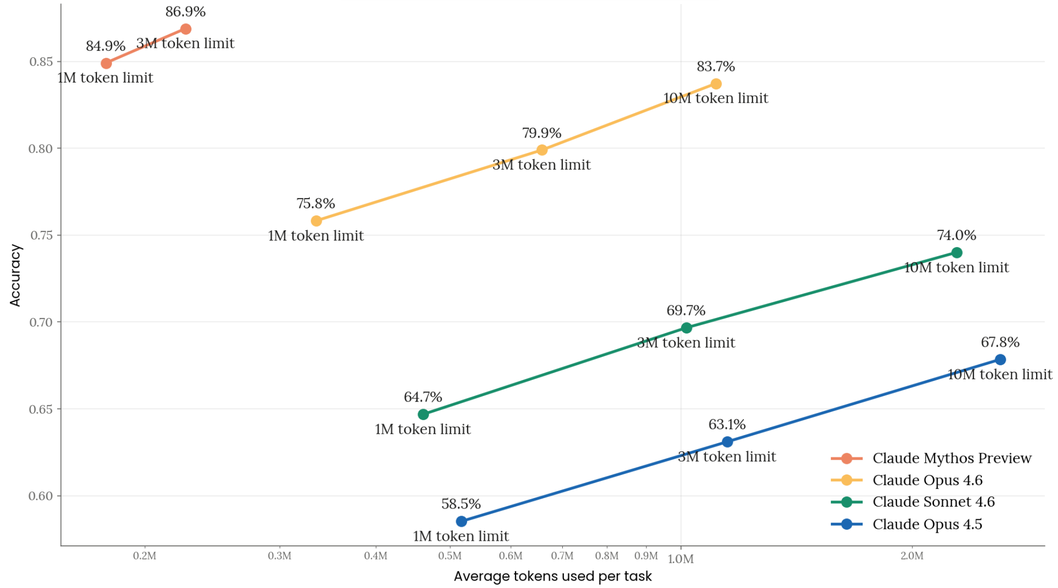


Claude Capybara in the room: Mythos

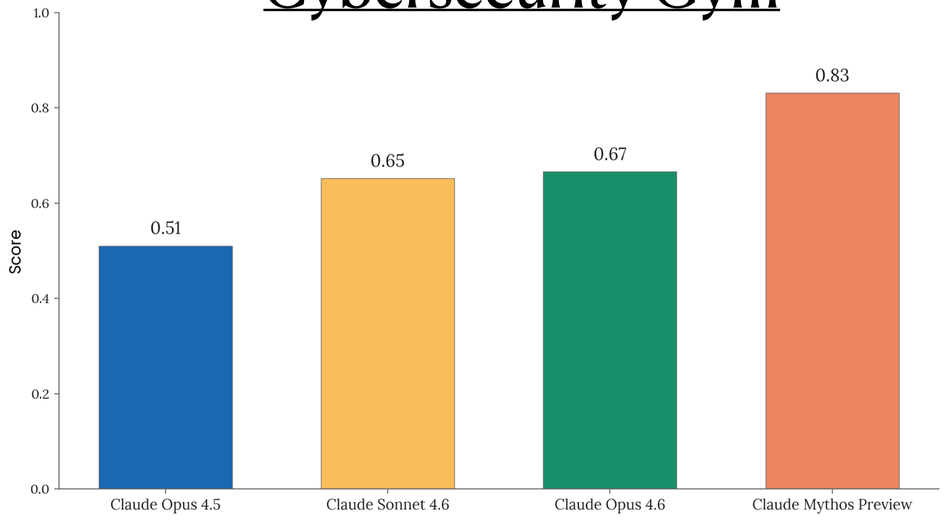
Math Olympiad



Web Research



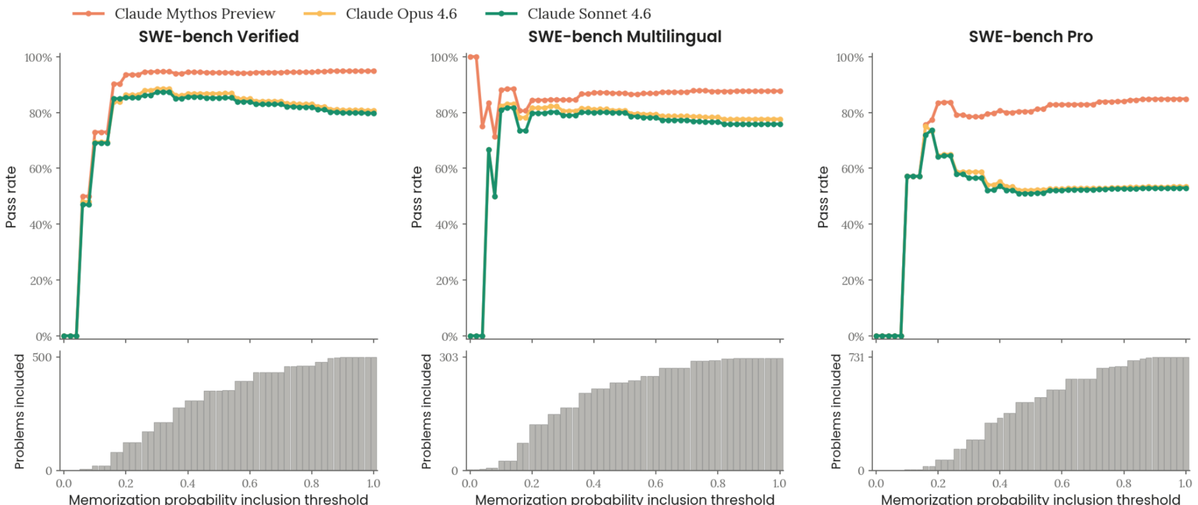
Cybersecurity Gym



Data Leakage

Pass rate vs memorization-filtering threshold

Curves show pass rate on the subset of problems whose max memorization probability is ≤ the x-axis value. Right edge of each curve = full-set score; bars below show how many problems remain at each threshold.



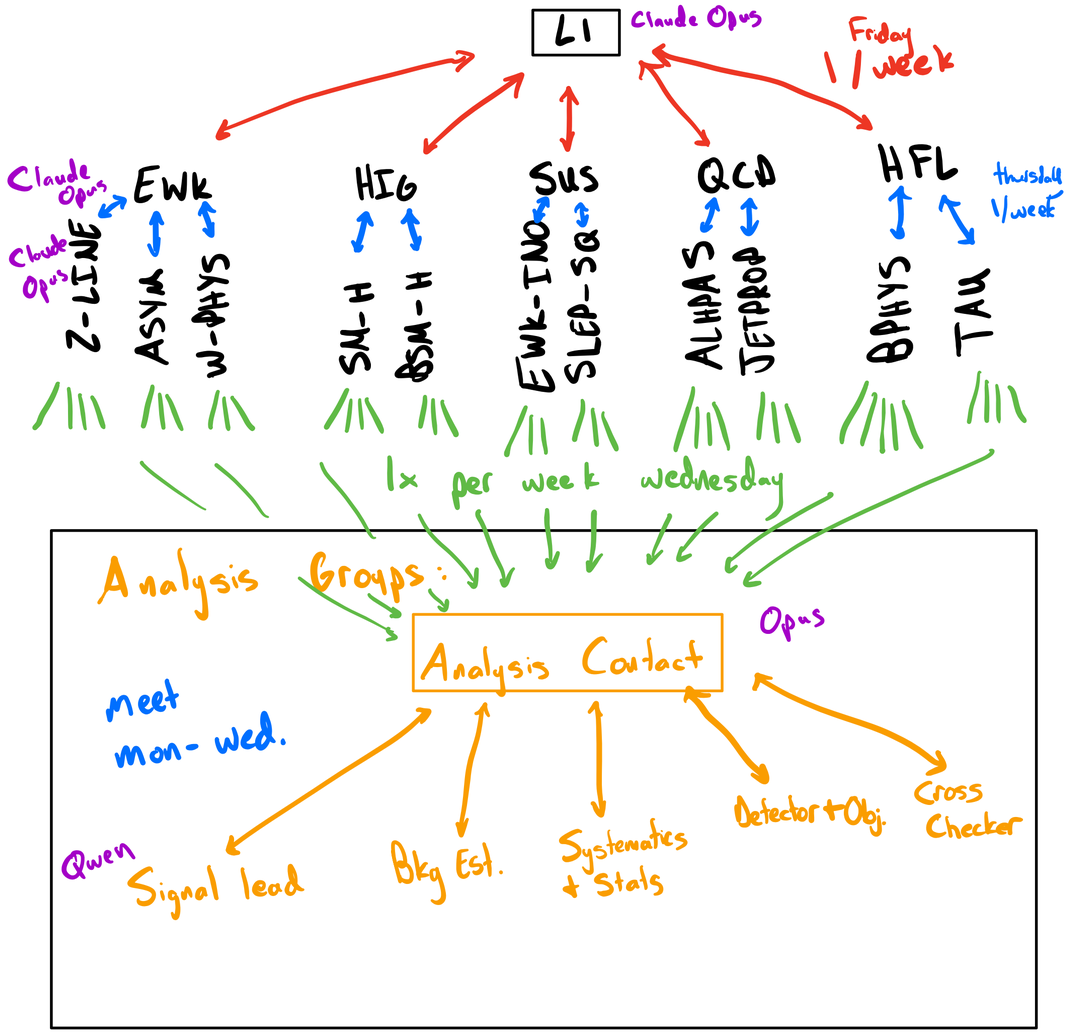
Claude

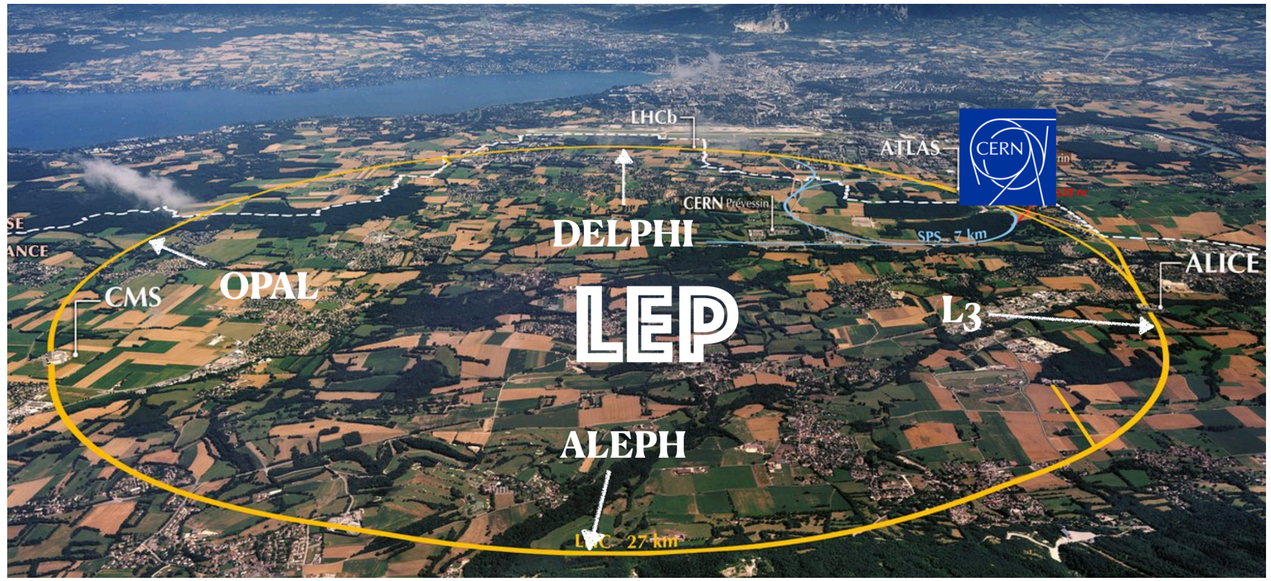
Prospects with Mythos

- 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)
  - Unclear when it will become available (if ever)
  - Untested by the public (maybe its not the big leap forward we think it is)



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