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,♯] **Philip Harris**[1,2,♭]
[1]Department of Physics, Massachusetts Institute of Technology
[2] The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
[3] CERN

Abstract

Large language model-based AI agents can now autonomously execute substantial portions of a high energy physics (HEP) analysis pipeline. We present *Just Furnish Context* (JFC), a proof-of-concept framework that integrates coding agents, literature-based knowledge retrieval, broad-scope specification documents encoding HEP conventions, and multi-agent review. Given access to a HEP dataset, an execution framework, and a corpus of prior experimental literature, Claude Code succeeds in automating all stages of a typical analysis: event selection, background estimation, uncertainty quantification, statistical inference, and paper drafting. We demonstrate this by conducting analyses on open data from the ALEPH, DELPHI, and CMS experiments to perform electroweak, QCD, and Higgs boson measurements—all from short physics prompts with no human intervention beyond an unblinding approval gate. Rather than replacing physicists, these tools promise to offload the repetitive technical burden of analysis code development, freeing researchers to focus on physics insight, novel method development, and rigorous validation.

1 Introduction

A typical experimental high energy physics analysis is a years-long endeavor. For large collider experiments the process is algorithmic: a physicist identifies a measurement channel, studies existing literature, designs and validates an event selection procedure using Monte Carlo simulation, quantifies systematic uncertainties, and performs statistical hypothesis tests to extract measurements or limits. This process invariably involves writing thousands of lines of code, most of which is structurally similar across analyses within a collaboration. While useful for developing programming skills, it consumes a substantial fraction of a physicist's time and demands little physics reasoning or insight.

In this paper, we argue that the majority of this process can be delegated to AI agents. Modern coding assistants can autonomously write and execute code, access documentation, iteratively critique and debug their own output, and document their progress for human overseers [1, 2]. We show that, given a general framework distilling the processes used in practice at large HEP collaborations and an initial high-level physics prompt, Claude Code is able to produce a complete analysis: event selection, background estimation, systematic uncertainty evaluation, statistical inference, and a written report with publication-grade figures. The report quality is typically, at a cursory expert review, indistinguishable from one produced by a junior graduate student.

Recent work by Badea et al. [3] demonstrated that an AI agent, with iterative physicist supervision at each step, can contribute to a measurement using LEP open data. We build on this direction but take

[*]These authors contributed equally
[†]emoreno@mit.edu [‡]sambt@mit.edu [§]novaka@mit.edu [♯]dolores.garcia@cern.ch [♭]pcharris@mit.edu

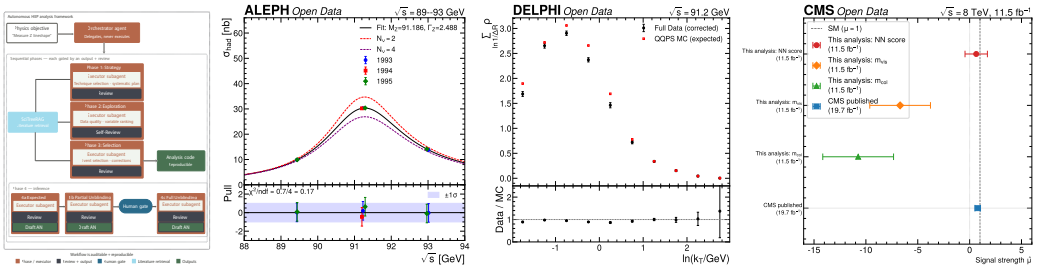

Figure 1: Left: A flow chart summarizing the JFC agent framework, including each analysis and self-review stage. Right (columns 2-4): Sample plots from analysis notes generated by Claude Code agents in the JFC framework. These include plots from the Z lineshape analysis with ALEPH data (left), the primary Lund jet plane analysis with DELPHI (middle), and the $H \to \tau\tau$ signal strength extraction with CMS Open Data (right).

a fundamentally different approach: rather than requiring continuous human feedback, JFC delegates review to specialized AI agents, concentrating human oversight at a single unblinding gate, yielding a nearly fully autonomous framework.

We do not present the analyses produced by JFC as legitimate scientific results. They are intended to showcase the level of HEP analysis that current agentic systems can produce entirely autonomously, based only on a short physics prompt and structured scaffolding for how a physics analysis is done. We believe this warrants rethinking how mainstream analysis work is done in HEP collaborations and how student labor is utilized.

2 Related Work

Interest in applying LLM-based agents to HEP has grown significantly in recent years. A community vision paper [4] outlines grand challenges for an AI-native particle physics ecosystem. Gendreau-Distler et al. [5] automate a Higgs diphoton measurement with LLM agents, though multi-step task planning remains beyond their scope. Diefenbacher et al. [6] show agents can independently arrive at sensible anomaly detection strategies. Other efforts target simulation orchestration [7], collider phenomenology [8], and experiment design [9]. In the broader AI-for-science landscape, autonomous pipelines have been demonstrated in chemistry [10, 11] and general scientific discovery [12]. None of the existing HEP-focused systems combine autonomous multi-step planning, literature-based knowledge retrieval, and multi-agent review. These gaps motivate JFC.

3 The JFC Framework

JFC is an agentic framework for autonomous execution of HEP analysis pipelines[2], built on Claude Code as both the LLM backend and agent runtime, using Claude Opus (`claude-opus-4-6`, 1M-token context) for every executor, reviewer, and arbiter role. All analyses were run on a Claude Max subscription ($200/month at the time of writing). An orchestrator agent delegates execution and review work to subagents across sequential phases. Each phase must produce a written artifact and pass an independent review before the next phase can begin, providing structured task decomposition and fresh context management. The key architectural principle is that the agent operates with genuine autonomy: unlike prior systems that constrain the agent to code generation within pre-defined steps, the JFC analysis agent receives a high-level physics objective—such as "measure the hadronic cross-section of the Z boson using ALEPH data"—and must autonomously determine the full analysis strategy.

Task decomposition. The orchestrator decomposes the analysis into sequential phases, each gated by the production of a written artifact on disk and passage of an independent review: (1) **Strategy**: the executor agent queries the literature corpus, identifies signal and background processes, proposes event selection, defines the blinding variable, and outlines the systematic uncertainty program;

[2]Code and documentation: `https://jfc-mit.github.io/`

(2) **Exploration**: data reconnaissance, literature search, and reference analysis survey; (3) **Selection**: implementation of event selection, background estimation, closure and stress tests; (4a) **Expected results**: systematic uncertainties, statistical model construction, Asimov fits, and signal injection tests; (4b) **Partial unblinding**: the analysis is run on a 10% signal-region data subsample and validated against expected results, followed by a mandatory human approval gate; (4c) **Full unblinding**: executes only after human approval and runs the complete dataset. Documentation sub-phases after each inference phase produce a complete LaTeX analysis note. Context is passed between phases exclusively through written artifacts and an append-only experiment log, preventing context window exhaustion [13].

Scaffolding. The agent's autonomy is scaffolded by two complementary knowledge bases: (i) a *methodology specification*—a natural-language document defining analysis phases, required outputs, and review gates, expressed in prose so the agent exercises judgment within constraints; and (ii) a *conventions directory*—living documents encoding accumulated domain knowledge for specific techniques (standard systematic uncertainty sources, validation checks, known pitfalls). Where the methodology tells the agent *that* it must evaluate systematic uncertainties, the conventions tell it *which* sources are standard and why. Together, these constitute an operational context analogous to a graduate course curriculum.

Literature-based knowledge retrieval. A critical component is integration with SciTreeRAG [14], which indexes a corpus of published LEP papers using hierarchical document structure for contextually coherent retrieval. This is complemented by arXiv and INSPIRE-HEP queries for modern methodology and published inputs. The retrieval system is particularly important for event selection design, systematic uncertainty identification, and statistical method selection—the most expertise-intensive aspects of an analysis. Only the ALEPH analyses used SciTreeRAG; the DELPHI and CMS analyses relied on the agent's pre-trained knowledge plus arXiv/INSPIRE queries, and still succeeded.

Multi-agent review. Each phase undergoes automated review by specialized agents mirroring a real HEP collaboration's internal review. Seven agents participate: a *physics reviewer* (denied access to methodology documents, evaluating purely on physics); a *critical reviewer* (conventions compliance and regression detection); a *constructive reviewer* (improvements and alternatives); a *rendering reviewer* (PDF compilation checks); a *BibTeX validator*; a *plot validator* (programmatic physics sanity checks); and an *arbiter* synthesizing all findings into a PASS/ITERATE/ESCALATE verdict. The review protocol is binding: the analysis cannot proceed until all reviewers sign off.

Unblinding protocol. The framework implements a three-stage blinding protocol (Asimov $\rightarrow$ 10% data $\rightarrow$ full data) with a mandatory human gate before full unblinding, mirroring the blinding protocols used by major HEP collaborations. This is the single point where autonomous execution is deliberately suspended.

4 Results and Discussion

We demonstrate JFC by reproducing several published measurements using archived LEP data and CMS Open Data [15]. Each analysis was initiated with a short natural-language prompt specifying only the measurement goal and data location; the agent autonomously determined the full analysis strategy. Table 1 summarizes the analyses produced with JFC, including links to anonymized repositories storing all analysis artifacts (code, plots, and final analysis notes).

Analysis quality. The agent-produced analyses are structurally sound, methodologically standard, and honestly documented. In all cases, the agent's analysis strategy is recognizably similar to the corresponding published analysis, consistently making the same high-level choices (e.g., iterative Bayesian unfolding for QCD measurements, HistFactory likelihood for the Higgs search, Breit–Wigner lineshape fitting for N_ν). Every note includes a systematic completeness table comparing considered sources against published analyses. The multi-agent review caught concrete, nontrivial errors—including $Z \rightarrow \mu\mu$ template contamination in $H \rightarrow \tau\tau$, a rank-deficient covariance matrix in the EEC unfolding, and an inflated α_s from NLO-only theory inputs—precisely the class of issues internal collaboration review is designed to surface.

Analysis	Data	Time	Repo
$H \to \tau\tau$ signal strength (μ_{τ_h}, $\sqrt{s} = 8$ TeV)	CMS	6h 30m	link
Z lineshape (M_Z, Γ_Z, σ^0_{had}) + α_s from R_ℓ	ALEPH	4h 58m	Link
Primary Lund jet plane density	ALEPH	7h 14m	link
R_b, R_c, and A^b_{FB}	ALEPH	4h 06m	link
Two-point energy-energy correlator + AEEC	ALEPH	3h 19m	link
N_ν from Γ_{inv}	DELPHI	5h 15m	link
Primary Lund jet plane density	DELPHI	6h 59m	link
Two-point energy-energy correlator	DELPHI	5h 23m	link

Table 1: Summary of analyses produced with JFC. Each was run autonomously from a short initial prompt, with no human intervention beyond the unblinding approval gate. Wall-clock times include idle periods waiting for usage-limit resets.

The main weaknesses are inherent to the prototype scope rather than errors in the agent's physics reasoning. The agent tends to be conservative with selection cuts, preferring high purity over efficiency, and exhibits a pattern of correct diagnosis but deferred treatment—identifying its own limitations but not always closing the loop on fixing them. This suggests the review-and-iterate cycle is where tighter human-agent collaboration would add the most value. A further limitation is inherent to open data: calibration inputs, luminosity, and reconstruction corrections are frequently incomplete or released only for an event subset, inflating statistical and systematic uncertainties independently of agent performance.

Limitations. Agents can produce superficially correct analyses with subtle physics errors. They excel at reproducing established strategies but are less reliable for genuinely novel approaches. On the execution side, agents are prone to ignoring instructions buried in large context windows, outputs are inherently stochastic, and they tend to rely on APIs from their training data rather than consulting up-to-date documentation. Visual inspection is a particular weakness: agents reliably generate publication-style figures mechanically but occasionally produce plots that are stylistically malformed or physically nonsensical without noticing, making automated plot validation essential.

Implications. These results suggest that a significant fraction of the technical work in a standard HEP analysis can be automated. The role of the physicist shifts from implementer to architect and critic: defining what should be measured, assessing whether the agent's approach is sensible, and catching subtle errors. This has the potential to dramatically increase the throughput of experimental programs and allow graduate training to emphasize physics insight over implementation. A particularly compelling application is systematic reanalysis of legacy datasets (LEP, Tevatron, B-factories): agents could produce first-pass results across archived data at a scale currently precluded by human-resource constraints, and could also be used to automatically re-execute the computational pipelines of existing publications as a reproducibility audit. However, technical proficiency in programming must remain a critical skill—physicists must be able to read and understand AI-produced code to validate it. The community should develop benchmarks for physics analysis tasks, standardized review protocols for AI-produced analyses, and training for AI-assisted workflows.

5 Conclusion

AI agents based on large language models, combined with a relatively small set of structured prompts, can already autonomously execute substantial portions of a standard HEP analysis pipeline. We have demonstrated this by reproducing published ALEPH, DELPHI, and CMS measurements with agents that plan their own analysis strategy, retrieve domain knowledge from the literature, execute the full analysis chain, undergo automated multi-agent review, and produce complete written reports. The technology to perform analysis at the graduate student level is here and warrants serious attention from the experimental HEP community. We are not suggesting that AI agents should replace physicists, but that they can take over much of the technically demanding implementation work, freeing physicists to focus on developing physical insight, asking creative questions, and exercising the expert judgment that no AI system can yet replicate.

References

[1] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? 2023.

[2] Josh Achiam et al. GPT-4 Technical Report. 2023.

[3] Anthony Badea, Yi Chen, and Yen-Jie Lee. Agentic AI–Physicist Collaboration in Experimental Particle Physics: A Proof-of-Concept Measurement with LEP Open Data. 2026.

[4] Thea Klaeboe Aarrestad et al. Building an AI-native Research Ecosystem for Experimental Particle Physics: A Community Vision. 2 2026.

[5] Eli Gendreau-Distler, Joshua Ho, Dongwon Kim, Luc Tomas Le Pottier, Haichen Wang, and Chengxi Yang. Automating High Energy Physics Data Analysis with LLM-Powered Agents. In *39th Annual Conference on Neural Information Processing Systems: Includes Machine Learning and the Physical Sciences (ML4PS)*, 12 2025.

[6] Sascha Diefenbacher, Anna Hallin, Gregor Kasieczka, Michael Krämer, Anne Lauscher, and Tim Lukas. Agents of Discovery. 9 2025.

[7] Tony Menzo, Alexander Roman, Sergei Gleyzer, Konstantin Matchev, George T. Fleming, Stefan Höche, Stephen Mrenna, and Prasanth Shyamsundar. HEPTAPOD: Orchestrating High Energy Physics Workflows Towards Autonomous Agency. 12 2025.

[8] Shi Qiu, Zeyu Cai, Jiashen Wei, Zeyu Li, Yixuan Yin, Qing-Hong Cao, Chang Liu, Ming-xing Luo, Xing-Bo Yuan, and Hua Xing Zhu. An End-to-end Architecture for Collider Physics and Beyond. 3 2026.

[9] Justin Hill and Hong Joo Ryoo. GRACE: an Agentic AI for Particle Physics Experiment Design and Simulation. 1 2026.

[10] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. ChemCrow: Augmenting Large-Language Models with Chemistry Tools. 2023.

[11] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi:10.1038/s41586-023-06792-0.

[12] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. 2024.

[13] Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts llm performance. Technical report, Chroma, July 2025. URL `https://research.trychroma.com/context-rot`.

[14] James McGreivy, Blaise Delaney, Anja Beck, and Mike Williams. Seeing the Forest Through the Trees: Knowledge Retrieval for Streamlining Particle Physics Analysis. 9 2025.

[15] CERN open data, 2026. URL `https://opendata.cern.ch/`.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions: a proof-of-concept framework for autonomous HEP analysis, with explicit caveats that results are not presented as legitimate scientific measurements.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 discusses limitations including subtle physics errors, complexity ceiling, prompt nondeterminism, and instruction compliance issues.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include theoretical results or proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The framework specification, agent definitions, and all analysis code are publicly available, with links to anonymized GitHub repositories provided in the text. The analyses use publicly available open data from CERN.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The framework and all agent-produced analyses are publicly available on GitHub. The data used is from the CERN Open Data portal.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3 describes the full framework architecture, agent configuration, and execution details. Table 1 reports wall-clock times.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper presents a framework demonstration, not a statistical comparison of methods. The physics analyses produced are explicitly not presented as scientific results.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All analyses were run using a Claude Max subscription. Wall-clock times are reported in Table 1.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. The paper explicitly discusses responsible use and the need for human oversight.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 4 discusses implications for analysis workflows, graduate training, and the risks of AI-produced analyses requiring human validation.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The framework produces physics analyses, not models or datasets with misuse risk. The paper explicitly warns against treating AI-produced results as scientific measurements without human validation.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All data used is from the CERN Open Data portal, which provides open access.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The framework code is released on GitHub with documentation.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing or human subjects research.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This work does not involve human subjects research.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM usage is central to this work and is described throughout. The framework is built on Claude Code with the Claude Opus model.