



中國科學院高能物理研究所
Institute of High Energy Physics, CAS



HEP-Skill
v0.10

HEP-Skill

Building a Skill-Oriented
AI Workflow for HEP Analysis



Shuo Han

Institute of High Energy Physics, CAS

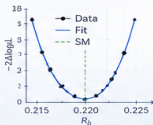
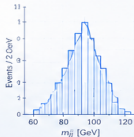
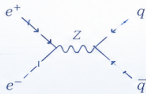
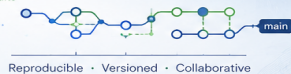


August 2026



Human Researchers

Scientific Authority · Judgment · Final Results



Domain Skills
HEP expertise built-in



Evidence-driven
Validated and traceable



Reproducible
From data to results



Collaborative
Human + AI, together

HUMAN DRIVER & REVIEWER
objective • intervene
direct • review

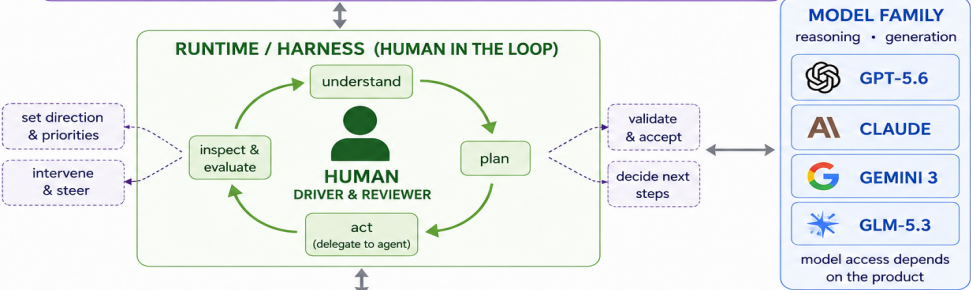
AGENT PRODUCT / SURFACE
interaction • context • permissions • tool routing

**CODEX**
multi-surface
project agent

**CLAUDE CODE**
terminal-first
coding agent

**CURSOR AGENT**
IDE-native
coding agent

**ANTIGRAVITY / GEMINI CLI**
agent platform +
terminal agent



MODEL FAMILY
reasoning • generation

**GPT-5.6**

**CLAUDE**

**GEMINI 3**

**GLM-5.3**

model access depends on the product

PROJECT ACTION + EVIDENCE

**tool calls**
        ...

repository • shell / compute • web / MCP • files / figures • version history

**code • results • reviewable evidence**

Why a Native Agent Alone Is Not Enough for Physics Analysis



Takeaway

Native agents are powerful executors, but physics analysis needs persistence, reproducibility, correctness, and shared memory. A skill-oriented workflow provides these foundations.

Representative Approaches to Automated HEP Analysis

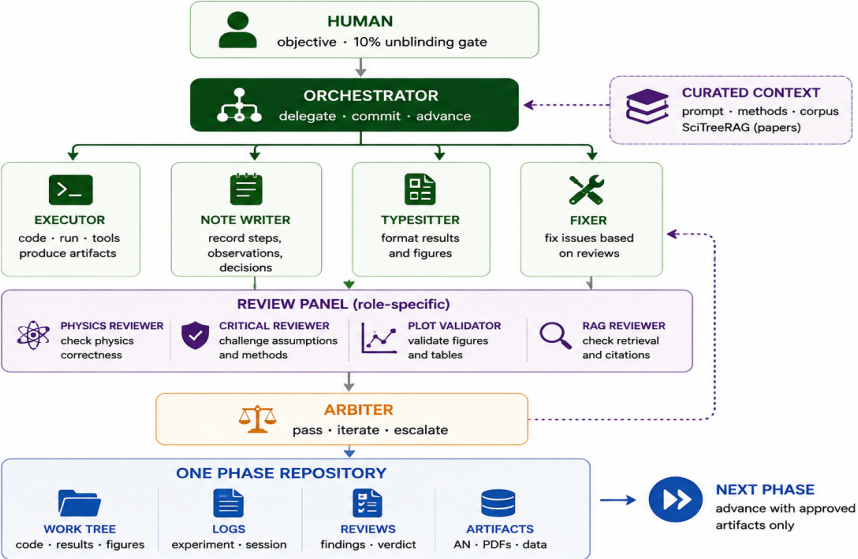
Work	Core mechanism	Demonstrated HEP task	Boundary and lesson
Dr.Sai arXiv:2604.22541	Dedicated runtime: Host delegates to Planner, Coder, Tester, and Reflector; built-in scheduling and memory	Ten J/ψ branching-fraction re-measurements in BESIII	Runtime owns orchestration, DSL/RAG, tools, data, and facility access
JFC Just Furnish Context arXiv:2603.20179	Orchestrator moves artifacts through executor, review, and arbiter roles; SciTreeRAG supplies papers	End-to-end studies with ALEPH, DELPHI, and CMS open data	Sequential phases; review at every handoff; human unblinding gate
Snakemake method Agent + Snakemake arXiv:2512.07785	Supervisor-coder loop inside a fixed five-stage Snakemake DAG	ATLAS Open Data $H \rightarrow \gamma\gamma$; success, errors, API calls, and cost	DAG fixes dependencies and order; the agent writes and repairs each stage

All three show that agents can execute long technical chains; they differ mainly in **where domain knowledge, execution structure, and collaboration live**.

Architectures, tasks, and validation boundaries summarized from the cited papers; arXiv links are in the first column.

JFC: One Phase Repository, Role-Specific Review

Orchestrator governs the workflow; human provides objective and final gate.



**End-to-End Automation**
Orchestrator drives the workflow across phases.

**Quality by Design**
Role-specific reviews ensure correctness and robustness.

**Traceable & Reproducible**
All artifacts, logs and reviews live in one phase repository.

**Safety with Human Gate**
Human sets objective and controls unblinding.

**Scalable**
Clear roles and context enable long, complex studies.

Which Path Should Govern the Analysis?

6 / 24

Two routes, different trade-offs. Both deserve exploration.

Route A: Automated (e.g., JFC Style)

Agent-driven pipeline with structured roles



Consistent



Reproducible



Traceable



Scalable

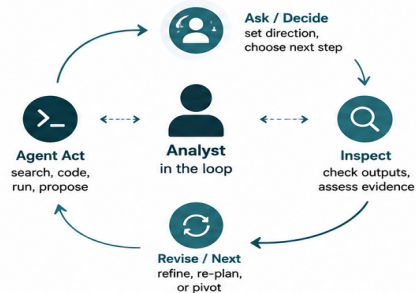


Choose the **right mix** for the task, the team, and the stakes.



Route B: Human-in-the-Loop (High-Touch)

Analyst drives the agent closely, step by step



Flexible



Adaptive



Creative



Judgment



Key Point

No one-size-fits-all. **Automate** routine execution; **escalate** scientific judgment.



+

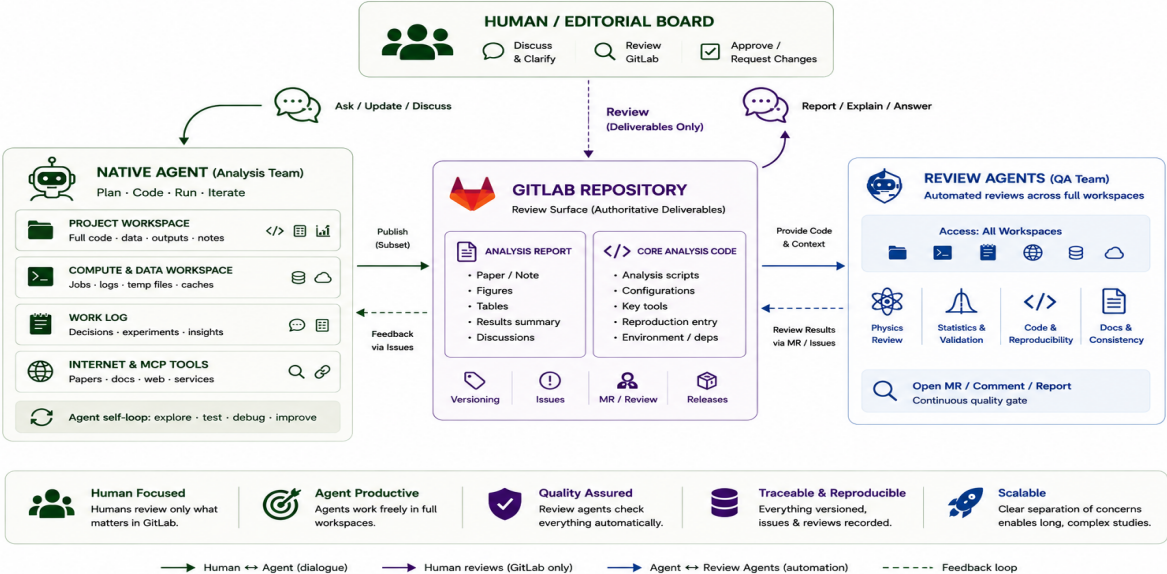


Best results come from humans and agents working **together**.

6 / 24

HEP-Skill Workflow: Humans Review GitLab, Agents Work Everything

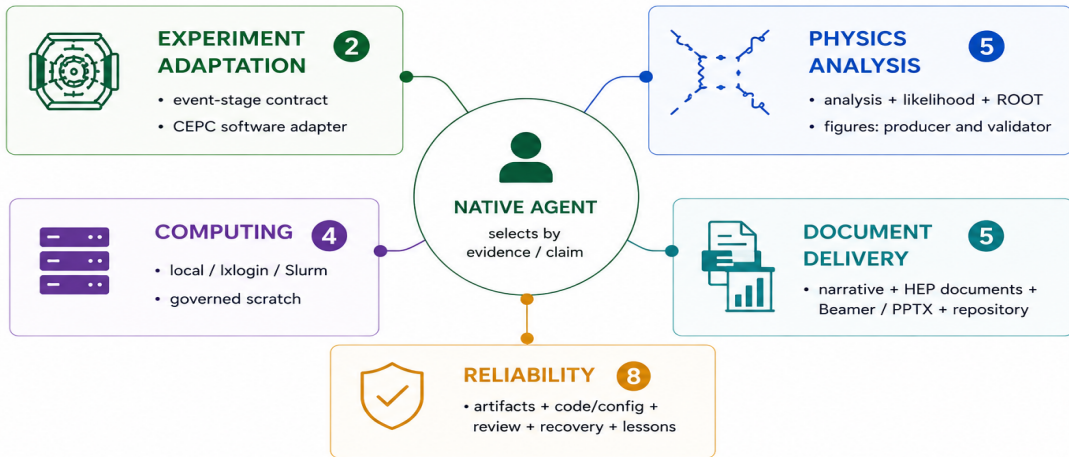
Human (Editorial Board) reviews the reviewable subset in GitLab via dialogue with the agent.



HEP-Skill: Five Composable Capability Areas

8 / 24

One bounded responsibility • checkable evidence • portable across projects • SKILL.md ≤ 300 lines



HEP-Skill consists of 24 responsibilities organized into the five families above.
Not a fixed DAG — skills are composed on demand by the native agent.

Experiment Adaptation and Compute Routing Stay Separate

9 / 24



EXPERIMENT ADAPTATION 2 Skills



HEP Event Processing Pipeline
(gen → sim → reco → ntuple)

hep-event-processing-pipeline



CEPC Software Workflow
(Key4hep · Gaudi · EDM4hep)

cepc-software-workflow



Project-specific Configuration
(paths · collections ·
detector conditions)



Stable Analysis Contract
(schema · provenance ·
event counts)

- Experiment layers are adapted to produce a stable analysis interface.



COMPUTING ENVIRONMENT 4 Skills



Local Workstation

local-first-compute-routing
(default)



IHEP lxlogin

ihep-lxlogin-ssh



**Local-first
Compute Routing**

local-first-compute-routing



IHEP AI login / Slurm

ihep-ai-login-slurm



**Governed Scratch
Workspace**

agent-scratch-lifecycle

- Execution target is chosen on demand; analysis code stays the same.



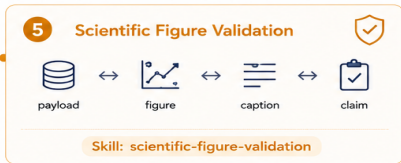
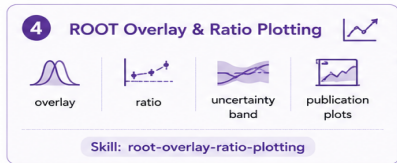
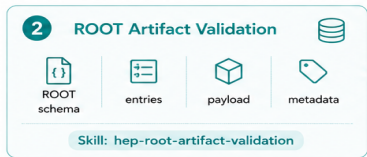
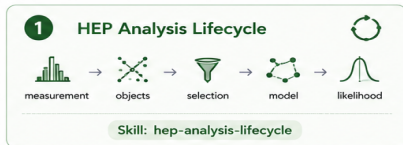
Physics code never depends on where it runs.

Adapt experiment layers via stable contracts; route compute via interchangeable targets.

Physics Analysis: Evidence-Oriented Analysis Skills

10 / 24

Five complementary capabilities ensure that analysis results are credible, reproducible, and publication-ready.



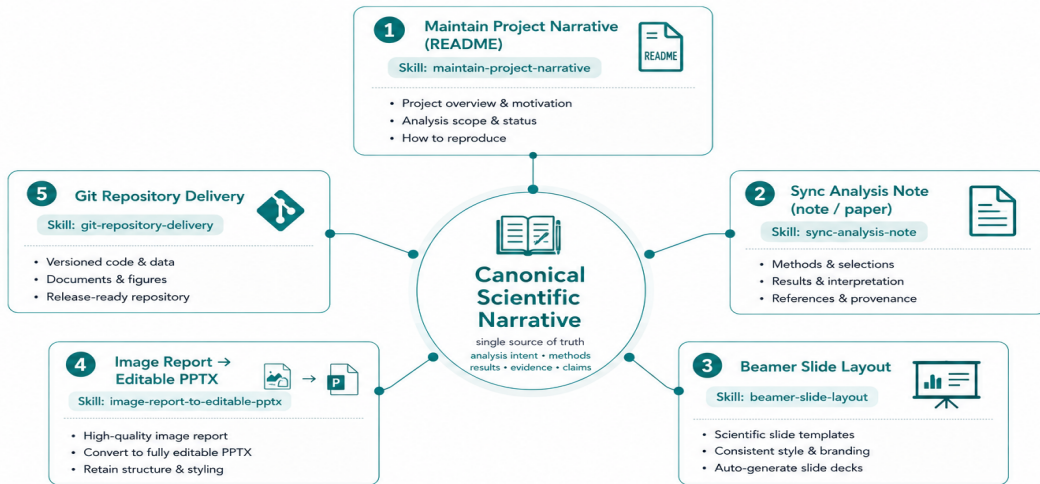
Every scientific claim is backed by machine-checkable evidence.

These five skills work together to transform analysis intent into validated, trustworthy results.

Documentation and Delivery: One Scientific Narrative

11 / 24

Five complementary skills turn the canonical narrative into all publishable deliverables.



One editable source, many publication formats.

A consistent narrative flows from README to paper, slides, reports, and repository — always in sync.

Process Reliability: Eight Skills Around Every Analysis

12 / 24

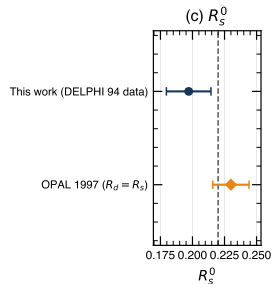
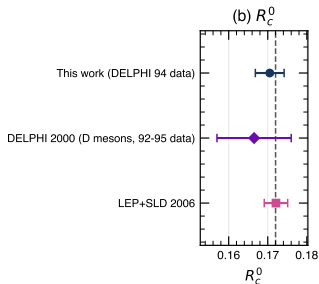
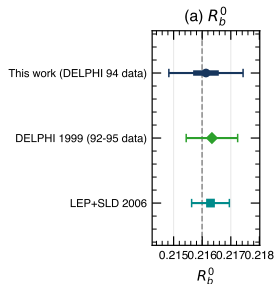
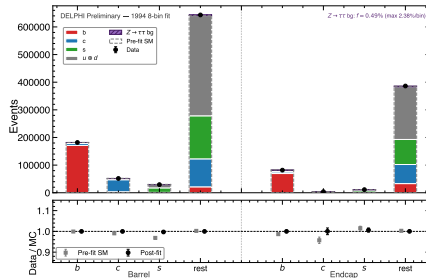
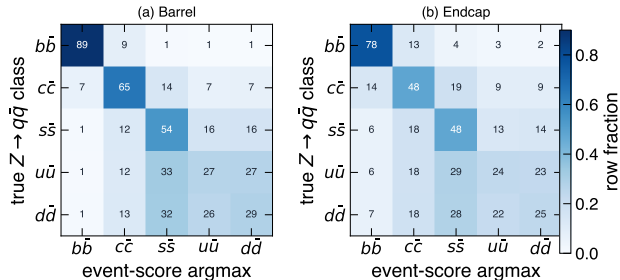
Eight complementary skills ensure that every analysis is reproducible, auditable, and continuously improved.



Reliability is accumulated through reusable governance skills.

These eight skills work together to make analysis processes trustworthy, maintainable, and continuously better.

LEP Case Study: Joint $R_b/R_c/R_s$ Measurement



One Root README: The Human-Facing Narrative

Record independent layout review
RSS authored 8 minutes ago

o8Se8k6

Name	Last commit	Last update
analysis	Organize analysis code by responsibility	1 hour ago
config	Promote held-out 1994 response result	6 days ago
docs	Record independent layout review	1 hour ago
production	Prepare 1994 analysis for paper drafting	1 week ago
tests	Organize analysis code by responsibility	1 hour ago
gilegine	Add synchronized PRL-style paper draft	1 week ago
README.md	Organize analysis code by responsibility	1 hour ago
analysis-delivery.json	Record independent layout review	1 hour ago
analysis-result.json	Complete the source-constrained uncertal...	1 day ago
build-reports.txt	Organize analysis code by responsibility	1 hour ago
requirements.txt	Simplify repository around current 1994 an...	2 weeks ago

8 README.md

Direct R_0^b , R_0^c , and R_0^d from DELPHI open data

This repository contains the maintained analysis from DELPHI Nansao2 to electroweak flavor ratios, together with a recurring review task, an analysis note, and a PRL-format paper draft with Supplemental Material. A frozen five-flavor ParticleNet model supplies a whole-event representation; the physics result comes from an independently calibrated counting likelihood, not from classifier performance alone. Tagger training and generator-alternative data are external, identified inputs.

Scientific question and current status

The analysis asks whether the full particle content of a hadronic Z decay can be converted into a simultaneous, calibrated measurement of its bottom-, charm-, and strange-quark fractions. The broader methodological question is whether a modern whole-event representation can become an auditable electroweak observable, rather than remain only a flavor-tagging scores. This matters for future high-statistics electron-positron Z programs, where classifier separation is useful for precision physics only when its response can be constrained and propagated. The final LEP+SLD values, $R_0^b = 0.2183 \pm 0.0086$ and $R_0^c = 0.1721 \pm 0.0088$, provide the most precise heavy-flavor benchmarks; individual DELPHI measurements give closer historical DELPHI comparisons, although their tagging methods differ from the whole-event response fit used here. Direct information on R_0^d is much more limited. Particular compactness each event into five flavor probabilities, eight exclusive barrel/endcap counts then connect that representation to R_0^b , R_0^c , and R_0^d through an explicit response model. The light-flavor complement $R_0^u + R_0^s = 1 - R_0^b - R_0^c - R_0^d$ is derived without assuming $R_0^b = R_0^c$.

The current result is a reproducible 1994-only measurement candidate. It uses all selected 1994 data in the likelihood and an event-digest 1995 data sample to calibrate the detector and tag-response uncertainty. The result is systematically limited: the 1995 response calibration dominates R_0^b and R_0^c , while physics modeling, led by the independently constrained strange-quark response, dominates R_0^d . The four report artifacts rebuild from frozen inputs and share one machine-readable result.

The year roles follow the available detector and control information. Data from 1960–1990 precede DELPHI's double-sided silicon vertex detector installed in 1994; the local held-out simulation comparison shows its largest era-dependent alignment migration in charm, so those years need a separate response calibration. Since counting statistics do not dominate the 1994 result, 1995 is used as the independent response control. Reweigh it as measurement data would require disjoint control events or a joint likelihood that accounts for the shared-data covariance.

A separate ran-disjoint likelihood ratio test uses the predeclared 25% calibration and 75% measurement splits in both years. It gives $\mu = 1.447$ and $\sigma = 0.708$ $\pm$ total from 5000 null pseudoexperiments for common year-dependent (R_0^b, R_0^c, R_0^d) splits. This supports annual consistency but is not a 1994+1995 combination; the all-1994 result and covariance remain unchanged.

1. Produce ParticleNet inputs from NanoAOD

All global options precede the subcommand. The producer refuses to overwrite artifacts and can write a JSON receipt.

```
"$(LEP_PYTHON_BIN)" production/process_samples.py \
--output-root "$$(LEP_INPUT_ROOT)" \
--official-nansao "$$(LEP_OFFICIAL_NANSAO)" \
--tautau-nansao "$$(LEP_TAU_NANSAO)" \
--sample all \
--summary "$$(LEP_INPUT_ROOT)/manifests/features.json" \
features
```

A bounded smoke uses a new output root and an explicit sample:

```
"$(LEP_PYTHON_BIN)" production/process_samples.py \
--output-root /absolute/new-smoke-root \
--official-nansao "$$(LEP_OFFICIAL_NANSAO)" \
--sample data_1994 \
--summary /absolute/new-smoke-root/features-data94.json \
features --max-output 1000
```

production/make_particle_features.C applies the hadronic event and beam-error selections, maps particles to tracks, writes the model branches, and records dEdx/RICH validity for QA. The current result uses the original detector inputs; it does not apply a post-training validity mask.

2. Evaluate the frozen ParticleNet model

Run this step in a GPU allocation:

```
"$(LEP_WEAVER_BIN)/bin/python" production/process_samples.py \
--output-root "$$(LEP_INPUT_ROOT)" \
--sample all \
--summary "$$(LEP_INPUT_ROOT)/manifests/scores.json" \
scores \
--model-root "$$(LEP_MODEL_ROOT)" \
--weaver "$$(LEP_WEAVER_BIN)" \
--qpus 8 \
--batch-size 512
```

The checker requires the five `score_label_*` branches, finite probabilities in $[0,1]$, unit row sums, and exact `(run_num, event_num)` ordering relative to the feature tree. Each new score receipt also records the SHA-256 identity of the checkpoint, YAML data configuration and network definition. The maintained model bundle has one deterministic bundle digest in `analysis-delivery.json`; each scoring receipt keeps the three member hashes for diagnosis. The preserved training configuration records 100,000 events per flavor from 1994 official simulation plus 10,000 validation events per flavor. The exact event map is now part of the report input. All 550,000 development events are excluded from the response, leaving 941,000 held-out events, and the response artifact records zero duplicate keys, a disjoint train/validation split and no statistical scaling. The independent 1995 response control remains an out-of-year data/simulation check; it is not used to assess the simulation split.

3. Build fit caches and selection efficiencies

```
"$(LEP_PYTHON_BIN)" production/process_samples.py \
--output-root "$$(LEP_INPUT_ROOT)" \
--sample all \
--summary "$$(LEP_INPUT_ROOT)/manifests/scores.json" \
scores \
--model-root "$$(LEP_MODEL_ROOT)" \
--weaver "$$(LEP_WEAVER_BIN)" \
--qpus 8 \
--batch-size 512
```



NARRATIVE

question · result ·
scope



REPRODUCE

inputs · command
· checks



RESULT SOURCE

canonical result
artifacts

Core Code in Git: The Reproducible Measurement

Name	Last commit
..	
📁 crosschecks	Organize analysis code by responsibility
📁 reporting	Organize analysis code by responsibility
📁 validation	Organize analysis code by responsibility
🔗 configuration.py	Propagate beam energy and protect nominal reports
🔗 counting_model.py	Profile finite selection-efficiency statistics
🔗 energy_control.py	Write valid energy-control JSON
🔗 fit_engine.py	Use stable optimizer-coordinate toy intervals
🔗 fit_inputs.py	Profile finite selection-efficiency statistics
🔗 likelihood.py	Define source-explicit RooWorkspace likelihood
🔗 measurement_fit.py	Profile finite selection-efficiency statistics
🔗 response_calibration.py	Profile finite selection-efficiency statistics
🔗 response_priors.py	Profile finite selection-efficiency statistics
📄 run_fit.sh	Organize analysis code by responsibility
🔗 systematic_model.py	Propagate selection efficiency statistics in annual diagnostics
🔗 target_response_variations.py	Serialize target response run summary
🔗 tau_background.py	Complete and simplify the 1994 analysis chain
🔗 working_point_systematic.py	Profile finite selection-efficiency statistics



MODEL

likelihood · systematics · response



RUN

fit · validate · cross-check



BOUNDARY

code · config · tests in Git
data · caches external

Review Documents: One Result, Different Depths

Name	Last commit
..	
📁 reviews	Record independent layout review
📁 static	Simplify repository around current 1994 analysis
📄 analysis_note.pdf	Correct DELPHI comparison references
📄 analysis_note.tex	Correct DELPHI comparison references
📄 change-ledger.json	Record independent layout review
📄 paper_draft.pdf	Correct DELPHI comparison references
📄 paper_draft.tex	Correct DELPHI comparison references
📄 paper_supplement.pdf	Correct DELPHI comparison references
📄 paper_supplement.tex	Correct DELPHI comparison references
📄 review_talk.pdf	Correct DELPHI comparison references
📄 review_talk.tex	Correct DELPHI comparison references



Review talk

evidence · open decisions



Analysis note

full method · appendices



Paper + supplement

publication argument · detail



Review record

findings · decisions · re-checks

Paper Draft: Evidence, Review, Cross-check, Revise

(Dated: 11 August 2026)

We report a simultaneous determination of the bottom-, charm-, and strange-quark fractions in hadronic decays of the Z boson using 1994 DELPHI open data. A five-flavor ParticleNet representation is converted into eight mutually exclusive flavor-enriched and remainder counts. The score thresholds are fixed using simulated purity, while an independent, energy- and efficiency-matched 1995 data/simulation sample calibrates the full flavor-response uncertainty. A profiled Poisson likelihood gives $(R_b^0, R_c^0, R_s^0) = (0.21613, 0.17051, 0.19705)$ with total uncertainties $(0.00130, 0.00368, 0.01740)$, respectively, and $R_{\text{had}}^0 = 0.41631 \pm 0.01688$. The result determines R_b^0 without imposing $R_b = R_c$, and the R_c^0 precision is competitive with individual LEP-era determinations. For all three fitted ratios, systematic uncertainties exceed the statistical components.

The ratios

$$R_q^0 = \frac{\Gamma(Z \rightarrow q\bar{q})}{\Gamma(Z \rightarrow \text{hadrons})} \quad (1)$$

basic tests of the flavor structure of the neutral current. The final LEP+SLD combination gives the most precise heavy-flavor benchmarks, $R_b^0 = 0.21629 \pm 0.00066$ and $R_c^0 = 0.1721 \pm 0.0030$ [1]. The individual DELPHI measurements [2, 3] provide historical DELPHI comparisons, although their tagging methods differ from that used here. Direct information on light flavors is much scarcer: the OPAL strange-quark result used identified b-momentum hadrons and the theoretical condition $R_s = R_c$ [4].

ParticleNet introduced particle-cloud representations for tagging [5]; CEPC studies extended them to flavor tagging, jet-origin identification, and projected R_b^0 and R_c^0 sensitivity [6–8]. OzonFold uses classifier-based weighting for unbiased, high-dimensional unfolding [9]. ATLAS has applied it to a 24-observable differential measurement [10]. Finite-category applications at electron–positron colliders use event-level classifiers for hadronic-Higgs branching-fraction projections. A ParticleNet-based holistic analysis estimates the statistical precision for several decay modes [11], while the IFC Reference Detector TDR uses an event-level ParticleTransformer and a seven-class migration matrix [12].

Adapt ParticleNet from jets to whole events and use five flavor scores through a control-data-constrained energy response in an electroweak counting likelihood. It determines R_b^0 , R_c^0 , R_s^0 , and R_{had}^0 together, with $R_{\text{had}}^0 = R_b^0 + R_c^0 + R_s^0$; no light-flavor equality is imposed. We analyze DELPHI open data recorded near the Z pole in 1994 [13]. Data from 1992–1993 precede the double-sided silicon vertex detector installed in 1994 [14]. A simulation shows a corresponding change in the flavor response. We therefore do not combine the two detector eras without a separate response calibration. Because the measurement is systematics limited, 1995 data are reserved for the independent response control. The

disjoint control events or a joint treatment of the shared data covariance. After the maintained hadronic- and detector-quality selections, 1 389 506 events remain 906 345 with $|\cos\theta_{\text{had}}| < 0.70$ and 483 161 in the eras. Flavor-labeled $Zb\bar{b}$, $Zc\bar{c}$, $Zs\bar{s}$, $Zu\bar{u}$, and $Zd\bar{d}$ samples provide the nominal response; the hard process generated with KK2F [15]. Simulated $Z \rightarrow \tau^+\tau^-$ decays give a small fixed background. The event representation contains particle kinematics, impact-parameter information, and particle-identification observables. A frozen ParticleNet network returns five flavor probabilities. The simulated response sample is disjoint from network training and validation, and its finite-sample uncertainty propagated; measured data were not used in training.

Four exclusive categories are formed in each detector region. A b-tagged event has the largest b score above θ_b . Of the remaining events, the c-tag selects $s_c \geq \theta_c$ and $s_s \geq \theta_s$; the s-tag then requires $s_s \geq \max(s_{u1}, s_{d1})$ and $s_s \geq \theta_s$. All other events enter the remainder. The working points in both detector regions are fixed from the defined simulated-purity requirements before examining the 1994 measurement counts or data/simulation results.

Figure 1 shows the five-flavor event-score argmax for the exclusive thresholds are applied; the fit uses the thresholded exclusive-category response $p_{\text{b}k}$. The separate $u\bar{u}$ and $d\bar{d}$ rows retain light-flavor structure that combined only in the counting model.

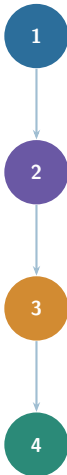
The thresholded flavor response enters a profiled counting likelihood. For detector region r , category k , and flavor q , the expected count is

$$\mu_{rk} = N_r \sum_{q=b,c,s,d} p_{\text{b}k}(\theta_q; 0, 1) f_q(R) + B_{rk}.$$

The likelihood is

$$\mathcal{L} = \prod_{r,k} \text{Pois}(n_{rk} | \mu_{rk}) \prod_q \mathcal{N}(\theta_q; 0, 1) \mathcal{N}(\kappa; \kappa_0, \sigma_\kappa).$$

Here $R = (R_b^0, R_c^0, R_s^0)$, with $R_{\text{had}}^0 = 1 - R_b^0 - R_c^0 - R_s^0$. The matrix N_r is the fixed background-subtracted



DRAFT

canonical result → journal story

REVIEW

physics · statistics · code · sources

CROSS-CHECK

test every claim-changing finding

REVISE

paper + supplement → re-review

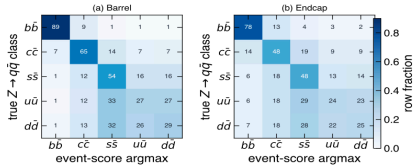


FIG. 1. Five-flavor ParticleNet event-score migration in the event-disjoint held-out 1994 simulation for (a) the barrel and (b) the endcap. Rows give the true $Z \rightarrow q\bar{q}$ event class, columns the event-score argmax, and entries are row-normalized percentages.

the nominal nuisance state, $p_{r,kq} = P(k | q, r)$; away from it, $p_{r,kq}$ is the corresponding response coefficient. The vectors θ_s are the response-source nuisance parameters, κ the charged-kaon response nuisance, and f_q the normalized selected flavor fraction, with $f_q \propto \epsilon_q R_q$. Here ϵ_q is the common-selection efficiency for flavor q , and its finite-simulation uncertainty is represented by three relative-flavor response modes among the θ_s . For each response source s , $C_s = L_s L_s^T$ and the eight response shifts are $\sum_s L_s \theta_s$. The likelihood contains the eight Poisson terms and all nuisance constraints used for the quoted covariance.

The response is constrained independently of the 1994 measurement. The 1995 low-, peak-, and high-energy data categories are interpolated to the nominal Z -pole energy. In each detector region, the simulated thresholds are adjusted to reproduce the 1994 exclusive b , c , and s efficiencies. The matched data/simulation difference defines a zero-centered covariance over all eight response columns. Control-sample statistics, finite 1994 simulation statistics in both the category response and common-selection efficiencies, hadronization, a separate 1994 detector/RICH variation, the LEP beam-energy calibration [16], and a charged-kaon response constraint are propagated in the likelihood. Purity-preserving threshold variations test the stability of the fixed category definition.

Figure 2 shows the eight observed counts and fitted expectations. The fit gives the pole ratios

$$\begin{aligned} R_b^0 &= 0.21613 \pm 0.00045_{\text{stat}} \pm 0.00122_{\text{sys}}, \\ R_c^0 &= 0.17051 \pm 0.00091_{\text{stat}} \pm 0.00357_{\text{sys}}, \\ R_s^0 &= 0.19705 \pm 0.00271_{\text{stat}} \pm 0.01718_{\text{sys}}. \end{aligned} \quad (4)$$

The production-to-pole conversion applies the standard

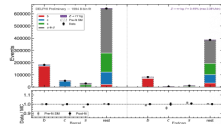


FIG. 2. Observed 1994 event counts and fitted expectations in the four exclusive categories in the barrel and endcap. The lower panel shows the data-to-fit ratios with statistical uncertainties from the observed counts.

shifts $\Delta R_b = +0.0002$ and $\Delta R_c = -0.0002$ and leaves the covariance and R_{ud}^0 unchanged [1]. The derived result is

$$R_{\text{ud}}^0 = 0.41631 \pm 0.00287_{\text{stat}} \pm 0.01664_{\text{sys}}. \quad (5)$$

The correlations of the quoted total covariance are $\rho_{bc} = -0.345$, $\rho_{bs} = 0.070$, and $\rho_{cs} = -0.255$. Figure 3 compares the three measurements with the corresponding single-experiment and combined LEP benchmarks.

A run-disjoint likelihood-ratio test of the 1994 and 1995 samples gives $q = 1.45$ and $p = 0.708$, consistent with a common (R_b^0, R_c^0, R_s^0) point. This control check does not alter the nominal all-1994 result or combine the years.

Tests with injected signals and independent pseudoex-

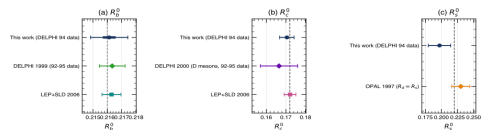


FIG. 3. One-dimensional comparisons for (a) R_b^0 , (b) R_c^0 , and (c) R_s^0 . Thick inner bars show the statistical uncertainty of the nominal 1994 measurement and thin outer bars show total one-standard-deviation uncertainties; dashed lines mark the Standard Model values. The DELPHI-only R_b^0 and R_c^0 results are historical R_b^0 comparisons, obtained with different tagging methods and partly overlapping the 1994 data used here [2, 3]. The LEP+SLD combinations are the most precise combined heavy-flavor benchmarks [1]; they include DELPHI and are not statistically independent of the DELPHI-only entries. The OPAL R_s^0 result assumes $R_d = R_s$, unlike the nominal measurement [4].

periments validate the fit response and reported intervals.

The total R_b^0 uncertainty, 0.00368, is smaller than that of the most precise standalone DELPHI determination, obtained from D -meson production and charm counting [3], although above the combined LEP+SLD precision [1]. For R_b^0 , R_c^0 , and R_s^0 , the systematic components exceed the counting uncertainties. Figure 4 shows the statistical-systematic variance split and the relative impacts of grouped uncertainty sources.

The result directly determines R_b^0 , R_c^0 , and R_s^0 without imposing down-strange flavor symmetry. At high-statistics Z programs planned for CEPC and FCC-ee [12, 17], larger flavor-control samples and detector-era calibration could reduce the response-calibration contribution.

This work is an external analysis of DELPHI open data and does not represent or imply endorsement by the DELPHI Collaboration. We thank the DELPHI data preservation effort and CERN Open Data Portal for preserving the data and software environment.

Data and code availability. The analysis code, machine-readable result, report sources, response and validation records, external-input bundle identities, and reproduction commands are provided in the accompanying GitLab repository. The DELPHI data are governed by the DELPHI open-data policy [13]; full event files and frozen model artifacts remain external and are identified by hashes and site locators.

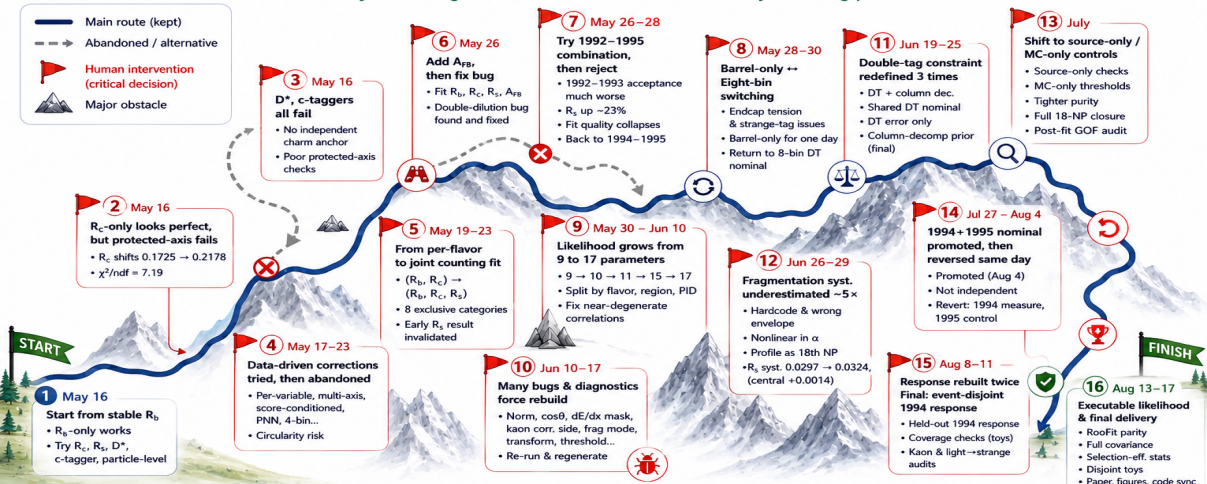
* These authors contributed equally to this work.

- [1] ALEPH, DELPHI, L3, OPAL, and SLD Collaborations, the LEP Electroweak Working Group, and the SLD Electroweak and Heavy Flavour Groups, “Precision electroweak measurements on the Z resonance,” *Phys. Rep.* **427**, 257 (2006).
- [2] DELPHI Collaboration, “A precise measurement of the partial decay width ratio $R_b^0 = \Gamma_{b\ell}/\Gamma_{had}$,” *Eur. Phys. J. C* **10**, 415 (1999).
- [3] DELPHI Collaboration, “Measurements of the Z partial decay width into $c\bar{c}$ and multiplicity of charm quarks per b decay,” *Eur. Phys. J. C* **12**, 225 (2000).

LEP ($R_b/R_c/R_s$) Analysis: A Long and Winding Road

19 / 24

From May 16 to Aug 18, 2026 • 386 commits • 16 major turning points



KEY TAKEAWAY

Not a straight path.
Many dead ends.
Many human decisions.
Built on hard lessons.



16

Major turning points



>10

Abandoned routes



>20

Critical human interventions



~ 10

Major bugs fixed



18

Nuisance parameters in final model

Current Final Uncertainties

$\sigma(R_b)$	$\sigma(R_c)$	$\sigma(R_s)$
0.00130	0.00368	0.01740

(Including selection-efficiency MC statistics)



Building and Improving the Skill System: Two Engines of Growth

Build with a Lighthouse Analysis and the Lessons from Pitfalls; Lift with Effective Skill Testing.
Together, they create a reusable, trustworthy, and continuously evolving Skill system.

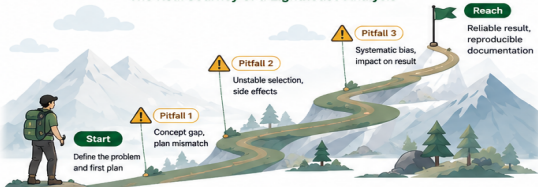


1 Build with a Lighthouse Analysis and Lessons from Pitfalls

Learn in real analysis, turn tacit experience into reusable Skills.



The Real Journey of a Lighthouse Analysis



From Pitfalls to Reusable Skills

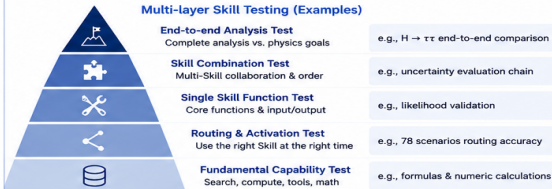


2 Lift with Effective Skill Testing

Use scientific testing to find issues, measure capability, and drive improvement.



Multi-layer Skill Testing (Examples)



Key Metrics We Care About



A Virtuous Cycle That Accelerates Both Truth and Reuse

Lighthouse analysis reveals real problems and insights.
Testing ensures quality and boundaries.
Together, they drive the Skill system to evolve and earn trust.



Skill Activation Across Three Runtime/Model Setups

ROUTING CORPUS

78 CASES

78 reviewed scenarios cover 24 Skills and all 24 routes.
Expected sets contain **107 required Skill instances**;
multi-Skill cases add more than one.

ACTIVATION SUBSET

6 TASKS / 9 READS

- 1 Generic likelihood validation
- 2 Mixed AI/HEP evidence + artifact chain
- 1 Local/remote compute choice
- 2 CEPC workflow + event-processing chain
- 2 Uncertainty closure + likelihood validation
- 1 CEPC-software vocabulary only

SAME SKILLS, THREE SETUPS

	Codex GPT-5.6 Sol medium	Codex GLM-5.3 high	Claude Code GLM-5.3 high
Required-Skill recall	107/107	107/107	104/107
Fully correct cases	67/78	71/78	65/78
Forbidden selections	0/83	0/83	2/83
Required Skills opened	9/9	9/9	9/9
Tasks with all reads	6/6	6/6	6/6
Median to all reads	24.0 s	17.3 s	17.2 s
Range	15.9–33.4 s	14.7–25.9 s	9.1–105.4 s
Required order	1/2	2/2	1/2
Full-task repeat	9/9	9/9	8/9

Routing recall and full-case correctness use different denominators. A Skill read is not rule compliance or task quality.

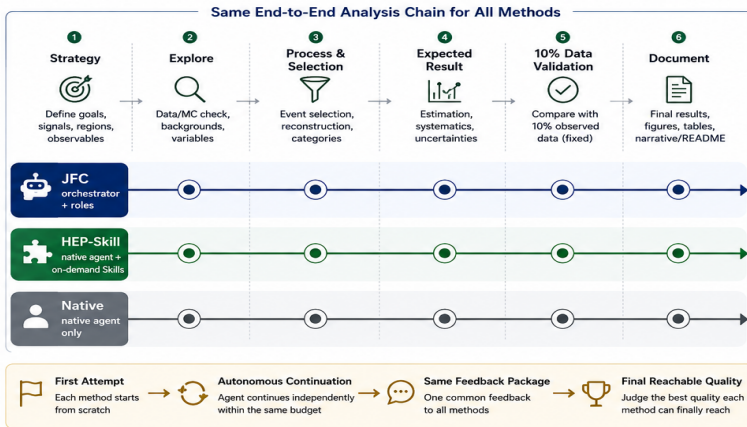
Full-Analysis Test Design

CMS Open Data $H \rightarrow \tau\tau$ ($\mu\tau_h$), 8 TeV

One physics task. Same resources. Only the analysis method changes.

 **Frozen Conditions**

-  CMS 2012 Open Data
-  $H \rightarrow \tau\tau$ ($\mu\tau_h$) 8 TeV
-  Same inputs, tools & versions
-  Same compute & time budget
-  Same data release rule



 **What Do We Compare?**

-  **Completion & Correctness**
How far it gets and how right it is
-  **Validation & Safety**
Soundness of checks, risk control
-  **Recovery & Supervision**
Can it recover from failures? How much human help is needed?
-  **Adaptability**
Can it adapt to new findings or changing conditions?
-  **Efficiency**
Time/compute used vs. quality achieved



Critical failures never average away.
Any critical failure leads to a low score regardless of other scores.



No ideal predecessor is inserted.
No group is given a pre-made "perfect result" to continue from.



Only 10% observed data is opened in the primary test.
The remaining 90% requires separate authorization.

Each method is run under identical conditions. We compare what they can achieve, how reliably they get there, and how safely and efficiently they operate.

Summary

1

A Skill-based HEP workflow

Agents do the work; Humans retain bounded scientific responsibility.

2

A lighthouse analysis

Many real pitfalls became reusable lessons.

3

A layered evaluation system

We are testing Skill activation and full-analysis performance.

4

Still under development

We expect Skill-based analysis to become increasingly common.

Release, Repositories, and Installation



HEP-Skill

PRODUCT

<https://code.ihep.ac.cn/shuohan/hep-skill>



HEP-Skill Evaluation

EVIDENCE

<https://code.ihep.ac.cn/shuohan/hep-skill-evaluation>



LEP Open Data

CASE STUDY

https://code.ihep.ac.cn/shuohan/LEP_legacy_Zqq



Install: check → dry run → backed-up apply

```
git clone https://code.ihep.ac.cn/shuohan/hep-skill.git
cd hep-skill
python3 scripts/manage_skill_system.py check
python3 scripts/manage_skill_system.py install (dry run)
python3 scripts/manage_skill_system.py install
--backup-dir ../hep-skill-install-backup-0.10.0-rc.16 --apply
```



Current candidate

0.10.0-rc.16 · source 6d7115b
24 Skills · Python 3
101/101 instruction contracts · 24/24 routes
10/10 script smokes · no source/runtime drift
Codex + Claude Code; **Codex recommended.**
✉ shuohan@ihep.ac.cn