

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

Anthony Badea^{1,*}, Marcello Maggi², Christopher McGinn³, Austin Baty⁴, Yi Chen⁵, Hannah Bossi³, Gian Michele Innocenti³, Jingyu Zhang⁵, Yu-Chen Chen³, Tzu-An Sheng³, Yen-Jie Lee³ and The Electron-Positron Alliance[†]

¹ Enrico Fermi Institute, University of Chicago, Chicago, IL 60637, USA

² Istituto Nazionale di Fisica Nucleare, Bari Division, 70126 BA, Italy

³ Laboratory for Nuclear Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

⁴ Department of Physics, University of Illinois Chicago, Chicago, IL 60607, USA

⁵ Department of Physics, Vanderbilt University, Nashville, TN 37235, USA

* Correspondence: badea@uchicago.edu

[†] The authors dedicate this work to the memory of Prof. Mauro De Palma (University of Bari, Italy), who passed away recently. He coordinated the Bari group's activities in the ALEPH experiment and championed the use of ALEPH data to explore new research frontiers. This work continues the scientific vision he so enthusiastically supported.

How To Cite: Badea, A.; Maggi, M.; McGinn, C.; et al. Agentic AI—Physicist Collaboration in Experimental Particle Physics: A Proof-of-Concept Measurement with LEP Open Data. *Highlights in High-Energy Physics* **2026**, *2*(3), 10. <https://doi.org/10.53941/hihep.2026.100010>

Received: 12 June 2026

Revised: 17 September 2026

Accepted: 28 September 2026

Published: 30 September 2026

Abstract: We present an agentic measurement of the thrust distribution in e^+e^- collisions at $\sqrt{s} = 91.2$ GeV using ALEPH archived data. This work was carried out through an AI agent–physicist collaboration, using interactive natural-language prompting of Anthropic Claude and OpenAI Codex. Under physicist guidance, the agents wrote and executed all analysis code and drafted the majority of the text included in this paper. A fully-corrected thrust spectrum is obtained via Iterative Bayesian Unfolding and Monte Carlo based corrections. The principal contribution of this work is the demonstration of the agentic analysis, with thrust serving as a benchmark observable for the method. Dedicated precision measurements of thrust are reported in the community elsewhere. This work represents a step toward a theory–experiment loop in which AI agents assist with experimental measurements and theoretical calculations, and synthesize insights by comparing the results, thereby accelerating the cycle that drives discovery in fundamental physics. Our work suggests that precision physics, leveraging the open LEP data and advanced theoretical landscape, provides an ideal testing ground for developing AI agent systems for scientific applications.

Keywords: quantum chromodynamics; AI agents; electron-positron collisions; event shapes

1. Context of Electron-Positron Alliance

Modern collider datasets provide unprecedented statistical power, and experimental uncertainties on many Standard Model measurements are now at the percent level. In this regime, theoretical uncertainties increasingly limit the overall precision, and those associated with Quantum Chromodynamics (QCD) are among the dominant contributions. To achieve the physics goals of future programs, including the High-Luminosity LHC (HL-LHC) and Future Circular Collider (FCC-ee), theoretical predictions must therefore reach a precision comparable to the expected experimental uncertainties, in many cases at the sub-percent level, so that potential new physics anomalies can be interpreted reliably.

Precision QCD studies can be pursued at hadron colliders such as the LHC, but they are most cleanly and systematically tested in lepton collider environments. At present, however, there is no active high-energy lepton collider operating at Large Electron–Positron collider (LEP) energies or above. To address this, the Electron-Positron



Alliance (EPA) [1] works with LEP collaboration members to curate and reanalyse archived datasets, opening new opportunities enabled by three decades of theoretical advances and modern experimental techniques. The combination of ultra-precise e^+e^- data and state-of-the-art perturbative and non-perturbative theory tools provides a uniquely powerful environment for high-precision studies. The EPA has produced a series of precision measurements using these archived datasets, including two-particle correlations [2], jet energy spectra and substructure [3], long-range near-side correlations at WW threshold [4], energy-energy correlators [5], and an unbinned measurement of thrust [6].

More broadly, this program is a compelling opportunity for AI for Science in a domain where precision is paramount. While AI is increasingly used to assist theoretical calculations and generate predictions, in the physical sciences such predictions ultimately require experimental validation. In high-energy physics, this validation depends on collider datasets that are rarely publicly accessible. The e^+e^- environment, particularly at LEP energies, provides the cleanest setting for precision tests of the Standard Model (SM), with well-defined initial states and calculable dynamics. The successful curation and use of archived open data therefore establishes the conditions for a theory–experiment loop in which AI systems, guided by domain experts, can propose new calculations, observables, or improved models and directly confront them with high-precision data. This creates a concrete framework for developing AI systems that do not merely interpolate existing knowledge, but actively refine theoretical descriptions through direct comparison with experiment.

2. Introduction

Machine learning has been central to high-energy physics for decades, enabling reconstruction algorithms, event classification, and detector optimization across collider experiments. A new frontier has emerged with large language models (LLMs) developed by major AI laboratories, including Anthropic [7], OpenAI [8], Google DeepMind [9], and others, that can now assist directly with scientific reasoning, mathematical proof, and code generation. A striking asymmetry persists: AI is transforming the theoretical and computational side of science research, assisting derivations of new results in theoretical particle physics [10], designing novel protein structures for biology and health research [11], and simplifying complex calculations. Yet the experimental side of the AI-assisted research loop remains largely absent, especially in fundamental physics. To realize a future where AI rapidly accelerates the discovery of new physics, AI-assisted scientific workflows must be developed and validated across both theory and experiment.

A logical first step toward closing this gap is an agentic analysis of real experimental data with expert physicist oversight. This is precisely the approach pursued here. We communicated via natural language prompts with the Anthropic Claude and OpenAI Codex agents, brainstormed analysis goals, then iteratively prompted the AI agents to execute the full analysis workflow. No analysis code was written by hand. By analogy, the physicist’s role is similar to advising a graduate student, providing expert direction, reviewing results, and iterating. The agent was supplied with context comprising the published ALEPH thrust paper [12], prior ROOT-based ALEPH analysis macros covering event selection and thrust calculation, internal analysis notes documenting the miniDST data format and detector calibration procedures, and the archived ALEPH data [13]. This prompting strategy follows from our understanding in the AI literature that few-shot examples and supplied context improve AI-assisted outputs [14,15]. A detailed inventory of the inputs and prompting approach is given in Section 3.

This vision parallels the evolution of AI in neighboring scientific domains. In materials science, chemistry, and biology, AI work has progressed from computational prediction with post-hoc human validation toward AI-designed experiments executed by autonomous robotic systems, creating closed feedback loops where AI proposes and orchestrates experiments and learns from them. High-energy physics faces a parallel opportunity, with exciting early progress for theory calculations [16–23]. While we cannot deploy robots to construct particle experiments (yet), the emergence of open and curated data can serve as a virtual laboratory for developing and validating AI-assisted scientific workflows that operate across the precision theory–experiment interface [2,3]. This requires infrastructure on both sides: precise theoretical predictions and experimental measurements with rigorous uncertainty quantification, where AI-assisted reasoning is tested and held accountable to measurable constraints [20,24,25].

Precision QCD is a natural domain for this program, because e^+e^- annihilation offers both high-precision data and theory calculations. This matters because fine details are consequential: percent-level uncertainties, coupled to rigorous covariance estimates and end-to-end reproducibility, leave little room for an error to pass unnoticed. In hadronic or nuclear collision environments, by contrast, larger uncertainties can more easily conceal mistakes made by an AI system.

From both an AI and precision physics perspective, thrust is an ideal first observable for this program. Thrust probes both perturbative radiation and non-perturbative dynamics in a single distribution, is computable at high perturbative order, and has well-established connections to resummed calculations, making comparison to theory

controlled and interpretable [12, 26, 27]. Thrust, along with other event-shape observables measured at LEP, continues to anchor comparisons between perturbative calculations, generator modeling, and non-perturbative hadronization constraints. Recent thrust-based α_s extractions report $\alpha_s(m_Z) \simeq 0.1136 \pm 0.0012$, below the 2023 world-average value $\alpha_s(m_Z) = 0.1180 \pm 0.0009$ [28, 29]. This persistent tension, often referred to as the $e^+e^- \alpha_s$ puzzle, is of great interest because α_s is the least precisely measured coupling constant of the Standard Model. The associated uncertainties propagate to all measurements involving QCD, and the effective field theory tools used to extract α_s are widely employed throughout the field. Put differently, if something is missed in the simplest case of e^+e^- collisions, it is likely that similar effects will also be missed in precision studies of more complex systems. Therefore, testing the new AI technology on thrust in e^+e^- collisions opens the door to significant progress in precision physics.

The objective of the agent-physicist collaboration in this work was to measure the thrust spectrum by applying appropriate selections, unfolding, corrections, and uncertainty handling. The agent was asked to execute the full analysis loop: data and simulation processing, unfolding orchestration, systematic propagation, figure production, and note maintenance with traceable links to code and outputs. We utilize the thrust measurement as a testing ground to understand the capabilities of the agents, establish a software base to conduct agentic analysis of the LEP data, and identify areas for improvement. We note that areas for improvement include both the core capabilities of the AI and the interaction of the physicist with the AI. We envision using our work to pursue a next stage, where the same framework will be extended to additional event-shape observables (C-parameter and heavy-jet mass), and the approach of using AI-assisted scientific workflows for both theoretical and experimental analysis will be pursued in a loop between them.

The remainder of the document is structured as follows. Section 3 defines the AI-agent operating model used in this work. Section 4 summarizes the ALEPH detector. Section 5 describes data and simulation samples. Section 6 details reconstruction and selections, including particle-category and event-level validation. Section 7 presents thrust construction and the executable unfolding workflow. Section 8 discusses uncertainty measurements. Section 9 presents the results of the measurement. Section 10 reports the conclusions from the study.

3. AI Agent Usage in This Analysis

This work includes an explicit evaluation of AI-agent support for an experimental QCD analysis workflow. The central question is not only whether an agent can generate text or isolated scripts, but whether it can execute a full analysis loop under physicist supervision: data and Monte Carlo (MC) processing, unfolding, systematic propagation, figure production, and note maintenance with traceable links to code and outputs.

3.1. Scope and Operating Model

The agents used in this analysis are Anthropic Claude Code and OpenAI Codex, depending on the preference of the physicist. The operating model is human-in-the-loop: the physicist formulates goals and provides expert direction through iterative prompts; the agent executes the corresponding implementation steps; the physicist reviews the outputs and provides feedback; the process is then repeated. Physics decisions remain analyst-owned, while the agent provides implementation speed, consistency checks, and documentation rigor. In practical terms, the agent is used to:

- write, execute, and monitor analysis code,
- enforce consistent plotting and note formatting conventions for work communication,
- maintain an up-to-date understanding of the analysis code and configurations,
- iterate quickly on ideas and other time-consuming tasks.

Our operating model is intentionally conservative to see how the agent performs when used to accelerate the common physics engineering and documentation tasks, rather than the core scientific decision tasks.

3.2. Inputs and Context Supplied to Agents

The inputs supplied to the agents at the start of each session were the following.

- Context: the published ALEPH thrust measurement, prior ROOT-based analysis macros for event selection and thrust calculation, and internal analysis notes documenting the open data format, detector calibration, and object definitions [12, 13].
- Datasets: the ALEPH LEP1 data files for the 1994 running period at $\sqrt{s} = 91.2$ GeV, together with the corresponding ALEPH MC samples. Additional modern MC were given for theoretical uncertainty estimation.

- Physicist-written prompts: prompts specifying the physics objective (e.g., “propagate the hadronic-event correction as an independent systematic component with per-bin envelope uncertainties”), acceptance criteria (e.g., “refolded spectra must reproduce detector-level data within statistical precision”), and output-format expectations (figure style, table layout, note section to update).

3.3. Boundaries, Controls, and Reproducibility

To keep scientific accountability explicit, each high-impact step has a control point where physicist sign-off is required before results are treated as baseline. The analysis stages where control points were used are: event and object selection definition, unfolding regularization choice, systematic-combination policy, final figure and uncertainty-presentation choices, interpretation description of the work. Beyond these sign-off gates, the physicist performed concrete validation checks at each major analysis stage by inspecting the artifacts of the agents’ work. These checks are enumerated here.

- Event selection: Cut-flow yields were compared to the published ALEPH numbers [12] to verify selection efficiency. Distributions of $\cos \theta_{\text{thrust}}$ and φ were inspected for detector artifacts; this step uncovered a localized φ excess in the data that was traced to a detector region and checked before proceeding.
- Unfolding: Refolded spectra were required to reproduce the detector-level data within statistical precision. An iteration scan confirmed that the refolded-to-measured ratio is stable as the regularization parameter is scanned.
- Systematic uncertainties: Each variation’s impact on the unfolded distribution was inspected for large bin-to-bin fluctuations. The covariance and correlation matrices for individual components as well as the total were computed.
- Results: The unfolded thrust distribution was compared to the ALEPH 2004 published result [12] over the 42 reported bins. Per-bin uncertainty breakdowns and correlation matrices are reported.
- All figure captions, table entries, and interpretive statements in this work were reviewed by the physicist for physics accuracy. Agent-generated text was treated as a draft requiring human approval before inclusion in the baseline.

Reproducibility is ensured through multiple complementary mechanisms. All analysis steps are version-controlled, with code and configuration managed in a git repository so that any result can be traced to a specific commit and set of inputs. Workflow runs produce structured output artifacts, including summary files that capture the configuration used, the software versions, and caches of the exported spectra and covariance files. Physicist oversight at each review stage provides a layer of verification, with each step requiring explicit physicist approval before being promoted to the nominal result and the analysis to continue. These controls ensure that results are reproducible and that the physicist retains accountability for the scientific claims.

4. ALEPH Detector

ALEPH was a general-purpose LEP detector with near 4π coverage and high-precision charged-particle tracking in a 1.5 T solenoidal magnetic field [12,30,31]. The detector subsystems most relevant for this event-shape analysis are the central tracker, electromagnetic and hadronic calorimeters, and outer muon chambers.

The central tracking system combined a silicon vertex detector (VDET), an inner tracking chamber (ITC), and a large time projection chamber (TPC) with 21 pad rows covering $|\cos \theta| < 0.97$. Together these detectors provided precise track parameters and momentum measurements, and they define the object-quality observables used in this analysis (for example d_0 , z_0 , and N_{TPC} , defined in Section 6.2). The transverse-momentum resolution of the combined tracking system is $\delta p_T/p_T = 6 \times 10^{-4} p_T \oplus 0.005$ (GeV/c), the second term arising from multiple scattering and dominating at low p_T .

The calorimeter system comprised two main subsystems. The electromagnetic calorimeter (ECAL) reconstructed isolated photons with a relative energy resolution of $0.18/\sqrt{E [\text{GeV}]} + 0.009$, providing the primary response to photons and electromagnetic showers and anchoring the neutral-energy scale for event-shape reconstruction. The hadronic calorimeter (HCAL) with resolution $\sigma_E/E \simeq 0.85/\sqrt{E [\text{GeV}]}$ is essential for neutral-hadron energy flow and global visible-energy reconstruction. The muon system outside HCAL supports lepton categorization in the archived ALEPH object definitions.

A key feature of the ALEPH reconstruction is the energy-flow algorithm, which combines tracking and calorimetric information to form a set of energy-flow objects that represent both charged and neutral particle candidates in the event. Tracking information dominates for charged particles, where the momentum resolution is superior to the calorimeter energy resolution. Calorimetric information provides the primary input for neutral particles (photons and neutral hadrons), which leave no track in the TPC. The combined energy-flow reconstruction achieves an energy resolution of $\sigma(E) = (0.59 \pm 0.03)\sqrt{E/\text{GeV}} + (0.6 \pm 0.3) \text{ GeV}$ for hadronic events. This energy-flow approach, used in the ALEPH 2004 published event-shape analysis [12], is the particle-level input

adopted for the present analysis. The 1998 ALEPH event-shape analysis [32] unfolded to the charged particle only thrust spectrum to minimize model dependence on neutral-energy simulation; our nominal analysis follows the 2004 approach using both charged and neutral energy-flow objects, with a charged-only variation retained as a cross-check.

5. Data and Simulated Event Samples

The measurement uses archived ALEPH LEP1 data from 1994 at $\sqrt{s} \simeq 91.2$ GeV. This period is used because matched detector-level and generator-level archived MC inputs required for unfolding are available in this configuration. The 1994 sample corresponds to approximately 40 pb^{-1} after quality requirements and provides the largest fully synchronized archived data/MC configuration for this workflow. Table 1 summarizes the event counts in the nominal thrust workflow. The archived MC reco sample contains approximately 752k selected events compared to 1.33M in data. As a result of having fewer simulated events, MC statistical fluctuations are larger than those in data.

Table 1. Number of total and selected events in the archived data samples.

Sample	Total Events	Selected Events
1994 data reco	1,365,440	1,326,376
1994 MC reco	771,597	751,878
1994 MC gen (t_{gen})	771,597	771,597
1994 MC gen-before ($t_{\text{genBefore}}$)	973,769	973,769

Detector corrections are derived from archived ALEPH MC processed through the same analysis chain. The archived PYTHIA 6 [33] MC was produced with the standard ALEPH detector simulation software chain (JETSET 7.4 parton shower and hadronization with the ALEPH tune, GEANT 3 detector simulation), consistent with the simulation used in the published ALEPH event-shape analyses of the same dataset [12]. In the archived sample, the detector-level and generator-level event records are connected through a common event identifier, enabling event-matched response construction and post-unfolding hadronic-selection corrections.

The unfolding step consumes four synchronized inputs, each a ROOT TTree in the archived sample: measured data tree (t), detector-level MC tree (t), generator-level MC tree used for response matching (t_{gen}), generator-level MC tree before hadronic event selection ($t_{\text{genBefore}}$). The ratio $t_{\text{genBefore}}/t_{\text{gen}}$ is used to derive a hadronic event-selection correction applied after unfolding. Event matching between reco and gen levels is performed with the unique event identifier. To populate the response matrix for unfolding, generator entries are matched to the reconstructed selected-event identifier set, consistent with the matched-response construction described in Section 7.

In addition to the archived baseline MC, alternative modern generator predictions from PYTHIA 8.317, HERWIG 7.1.5, and SHERPA 2.2.6 are used in this work through event-level reweighting variations to estimate theory model dependence. These variations are propagated as dedicated systematic components in the final uncertainty model while preserving the same detector-response baseline.

6. Event Reconstruction and Selection

6.1. Object Definitions

The archived reconstruction builds energy-flow candidates from combined tracking and calorimeter information. Charged particle candidates rely on high-quality tracking measurements, while neutral particle candidates utilize the calorimeter. The objects are categorized with an integer label `pwflag` as summarized in Table 2.

Table 2. Object labels used in archived ALEPH events.

pwflag	Category
0	charged hadrons
1	electrons (tracker + ECAL category)
2	muons (tracker + muon category)
3	V^0 candidates
4	photons
5	neutral hadrons

V^0 candidates (`pwflag` = 3) include K_S and Λ from displaced-vertex reconstruction. These objects enter the thrust sum with their measured momentum vector rather than individual decay products, which is consistent with

treating them as single hadronic objects for event-shape purposes. Their contribution to the thrust numerator and denominator is captured by the neutral-response systematic variations.

6.2. Object and Event Selections

Table 3 lists the object-level selection. These criteria define which reconstructed objects are propagated into thrust and therefore which detector effects enter the response model. They are imposed on the reconstructed collection only; the generator-level particles of the same events are retained without them. Here d_0 and z_0 are the transverse and longitudinal impact parameters of the track with respect to the per-event reconstructed primary vertex. Both are taken from the archived reconstruction. N_{TPC} is the number of TPC coordinates associated with the track.

Table 3. Object selections applied to data and detector level MC.

Object Variable	Requirement
Charged category	$\text{pwflag} \in [0, 2]$
Charged acceptance	$ \cos \theta \leq 0.94$
Charged transverse momentum	$p_T \geq 0.2 \text{ GeV}$
Charged impact parameter	$ d_0 \leq 2 \text{ cm}$
Charged longitudinal impact	$ z_0 \leq 10 \text{ cm}$
Charged hit quality	$N_{\text{TPC}} \geq 4$
Neutral category	$\text{pwflag} \in [3, 5]$
Neutral energy	$E \geq 0.4 \text{ GeV}$
Neutral acceptance	$ \cos \theta \leq 0.98$
Use charged objects	$\text{keepChargedTracks} = \text{True}$
Use neutral objects	$\text{keepNeutralTracks} = \text{True}$

The hadronic event selection is summarized in Table 4, following the standard ALEPH selection [12, 32]. These selections enforce a phase space with high tracking quality, controlled forward acceptance, and stable visible-energy topology before entering event-shape reconstruction. The residual contamination from processes such as $e^+e^- \rightarrow \tau^+\tau^-$ is expected to be less than 0.26% for these event selections [32]. Approximately 1.36 (1.32) million e^+e^- collisions resulting in the decay of a Z boson to quarks are analyzed (selected). Note that the acceptance requirement based on the event shape variable sphericity introduces a slight sculpting of the thrust distribution. The cut removes approximately 2.3% of selected events, with the largest effect in the multijet region. In data, the ratio of the thrust distribution with the cut to that without rises monotonically from ≈ 0.85 at $T \approx 0.6\text{--}0.75$ to ≈ 0.95 at $T \approx 0.9$ and ≈ 0.99 as $T \rightarrow 1$.

Table 4. Nominal event selections applied to data and detector level MC to build the nominal thrust input particle set.

Event Variable	Requirement
Total charged energy	$E_{\text{trk}}^{\text{tot}} \geq 15 \text{ GeV}$
Sphericity-axis acceptance	$ \cos \theta_{\text{Sph}} \leq 0.82$
Charged multiplicity	$N_{\text{trk}} \geq 5$
Total multiplicity	$N_{\text{trk}} + N_{\text{neu}} \geq 13$

6.3. Data-MC Kinematic Checks

The agreement between the data and MC samples is assessed below. The particle and track level distributions are shown after all selections. The energy-flow candidates are categorized as charged tracks (Figure 1), charged lepton 1 (electron) candidates (Figure 2), charged lepton 2 (muon) candidates (Figure 3), neutral long-lived particles that decay to two charged tracks (V^0) (Figure 4), photons (Figure 5), and neutral hadrons (Figure 6) based on the reconstructed properties utilizing different detector subsystems. Event-level variables used for hadronic selection are shown in Figure 7. The selections on these observables define the phase space used to build detector response and unfolded event-shape spectra. The data-MC agreement is in general quite good with some mismodeling observed in particular for the neutral particles.

Several effects are visible in these distributions. An enhancement near $\cos \theta = 0$ appears in both data and simulation for charged tracks, as shown in Figure 1a. It originates from the fixed floating-point precision used when the archived sample was reprocessed and will be addressed in a subsequent reprocessing. Narrow peaks at $E \simeq |\vec{p}| \simeq 45 \text{ GeV}$ appear in the charged-track (Figure 1c,d,h) and V^0 (Figure 4c,d,h) spectra of the reconstructed

simulation but not at generator level. These are reconstruction-level effects, and their origins remain under study. The muon ϕ spectrum in Figure 3b shows an enhancement present only at generator level, which is believed to arise from the processing of the archived simulation. The N_{trk} oscillatory structure for the generator-level distributions in Figure 7a is an artifact of the following: charged particles are produced in pairs, so even bins are populated, while radiation effects populate odd bins. The impact of these effects on the thrust spectrum is negligible.

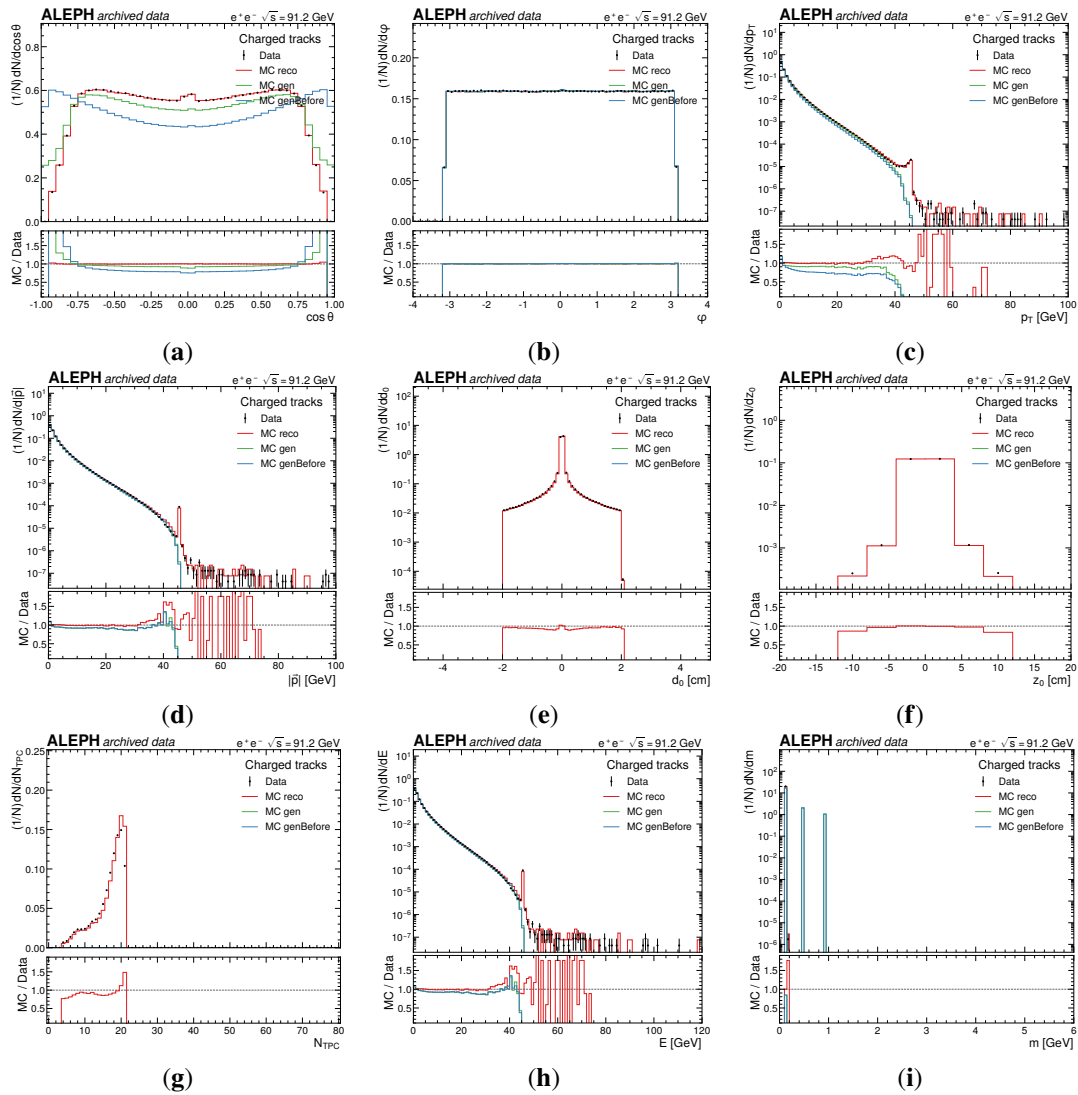


Figure 1. Selection variables for charged tracks: (a) $\cos \theta$, (b) ϕ , (c) transverse momentum p_T , (d) momentum magnitude $|\vec{p}|$, (e) transverse impact parameter d_0 , (f) longitudinal impact parameter z_0 , (g) TPC hit multiplicity N_{TPC} , (h) energy, and (i) mass.

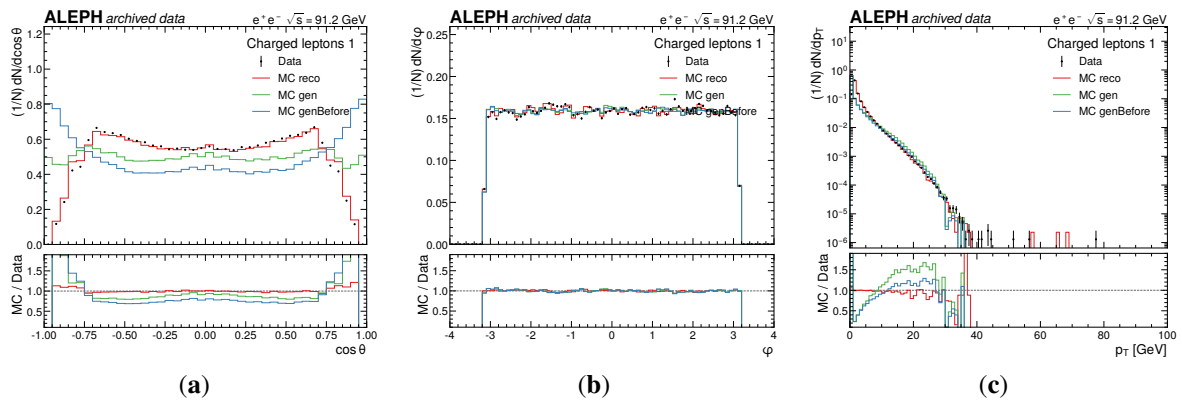


Figure 2. Cont.

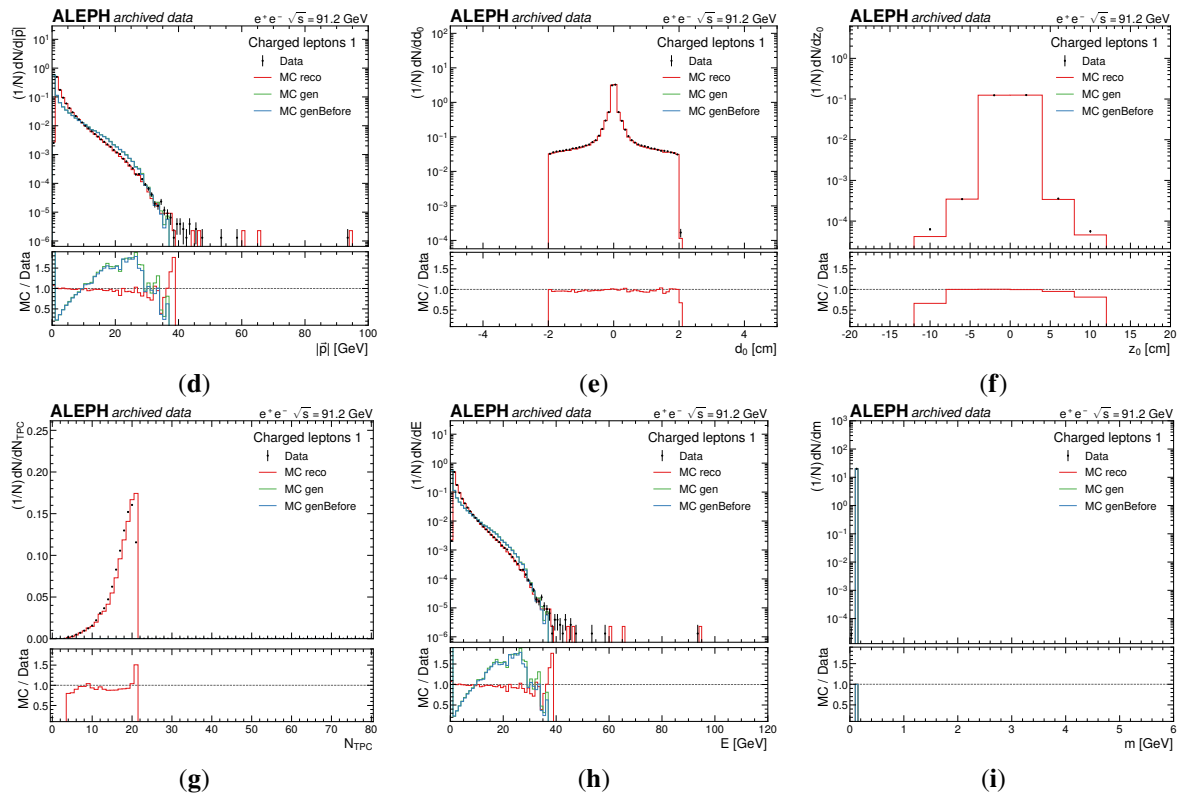


Figure 2. Selection variables for electrons: (a) $\cos \theta$, (b) ϕ , (c) transverse momentum p_T , (d) momentum magnitude $|\vec{p}|$, (e) transverse impact parameter d_0 , (f) longitudinal impact parameter z_0 , (g) TPC hit multiplicity N_{TPC} , (h) energy, and (i) mass.

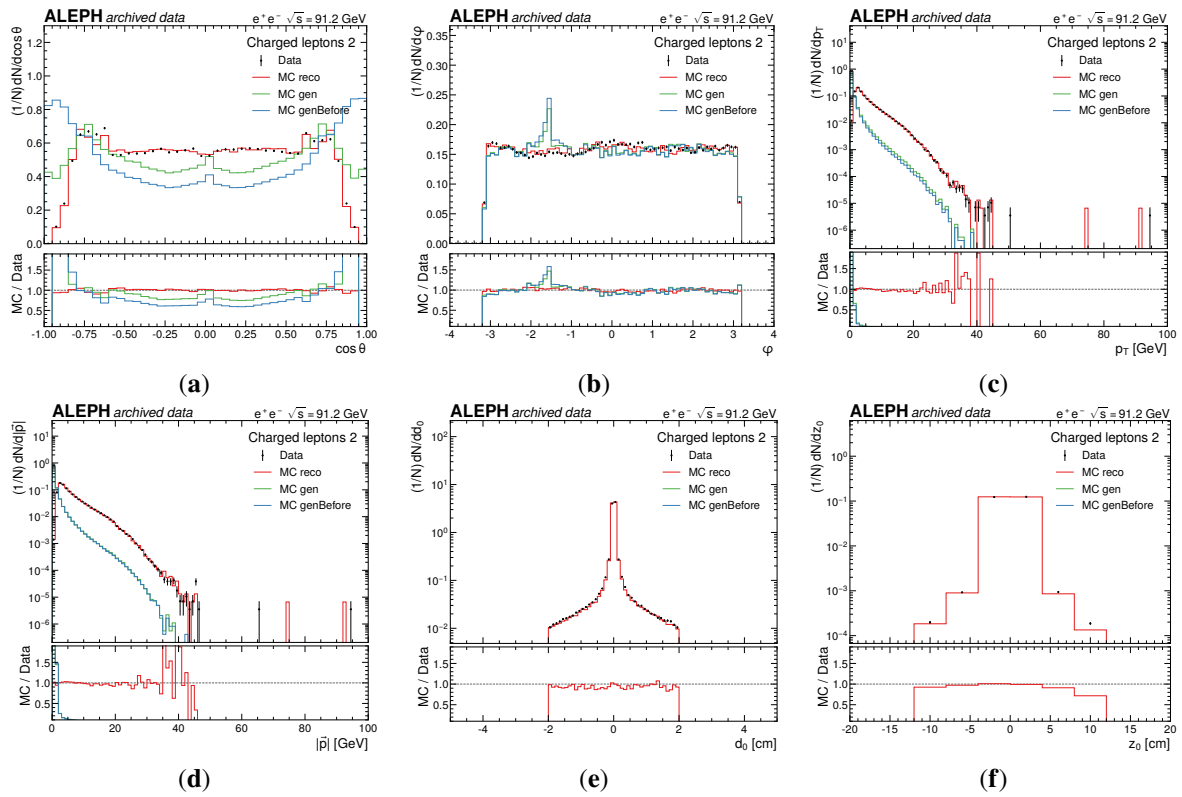


Figure 3. Cont.

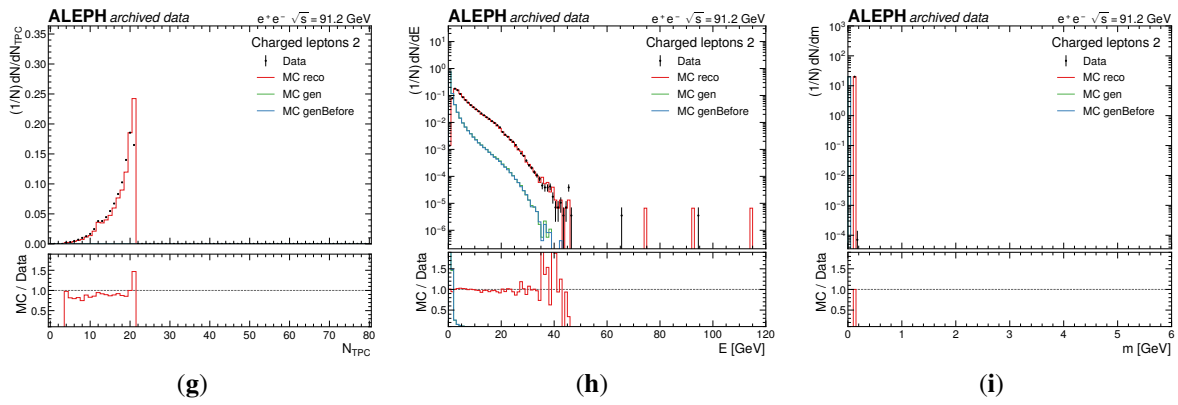


Figure 3. Kinematics for muons: (a) $\cos \theta$, (b) ϕ , (c) transverse momentum p_T , (d) momentum magnitude $|\vec{p}|$, (e) transverse impact parameter d_0 , (f) longitudinal impact parameter z_0 , (g) TPC hit multiplicity N_{TPC} , (h) energy, and (i) mass.

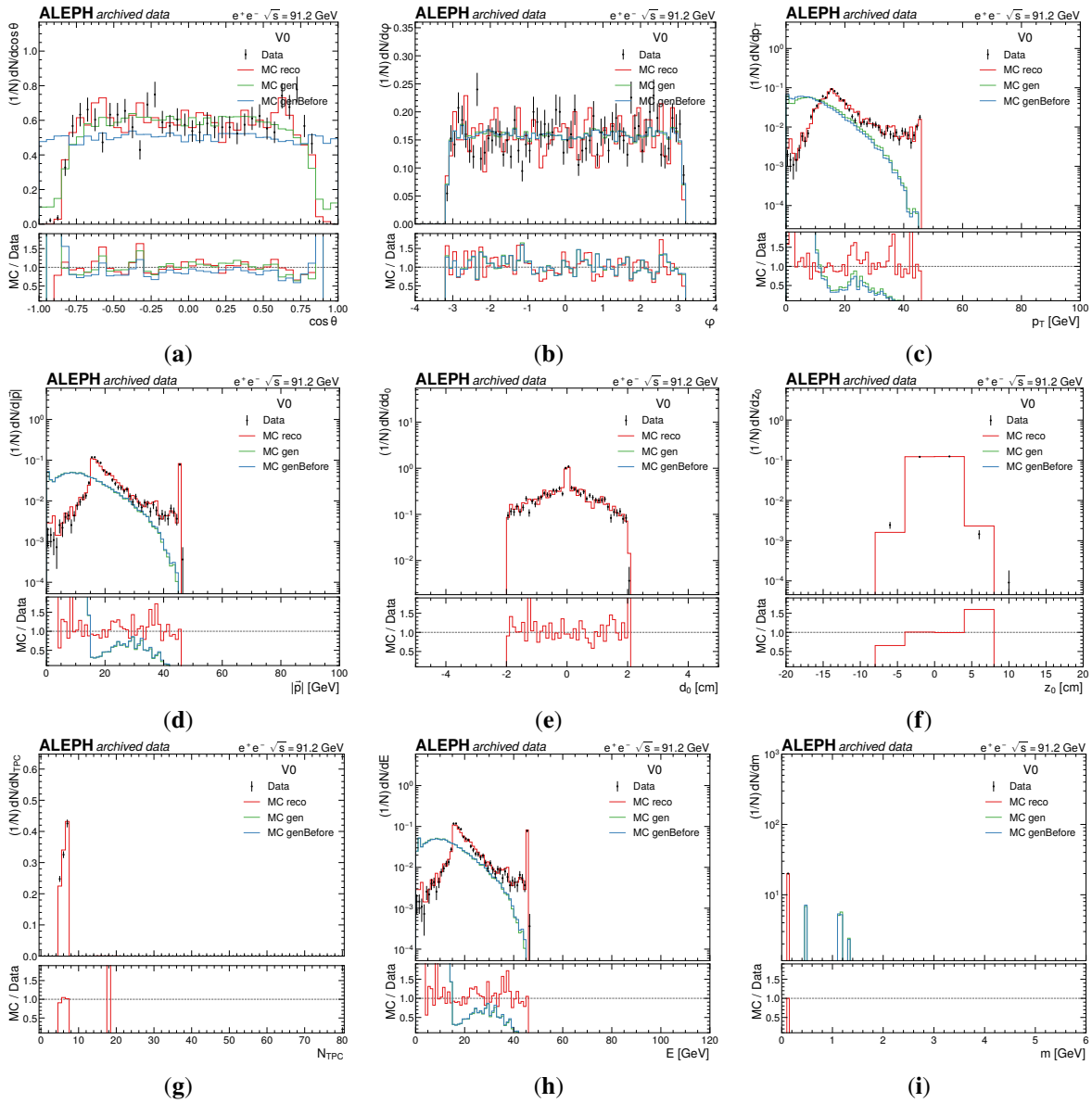


Figure 4. Kinematics for V^0 candidates: (a) $\cos \theta$, (b) ϕ , (c) p_T , (d) $|\vec{p}|$, (e) d_0 , (f) z_0 , (g) N_{TPC} , (h) energy, and (i) mass.

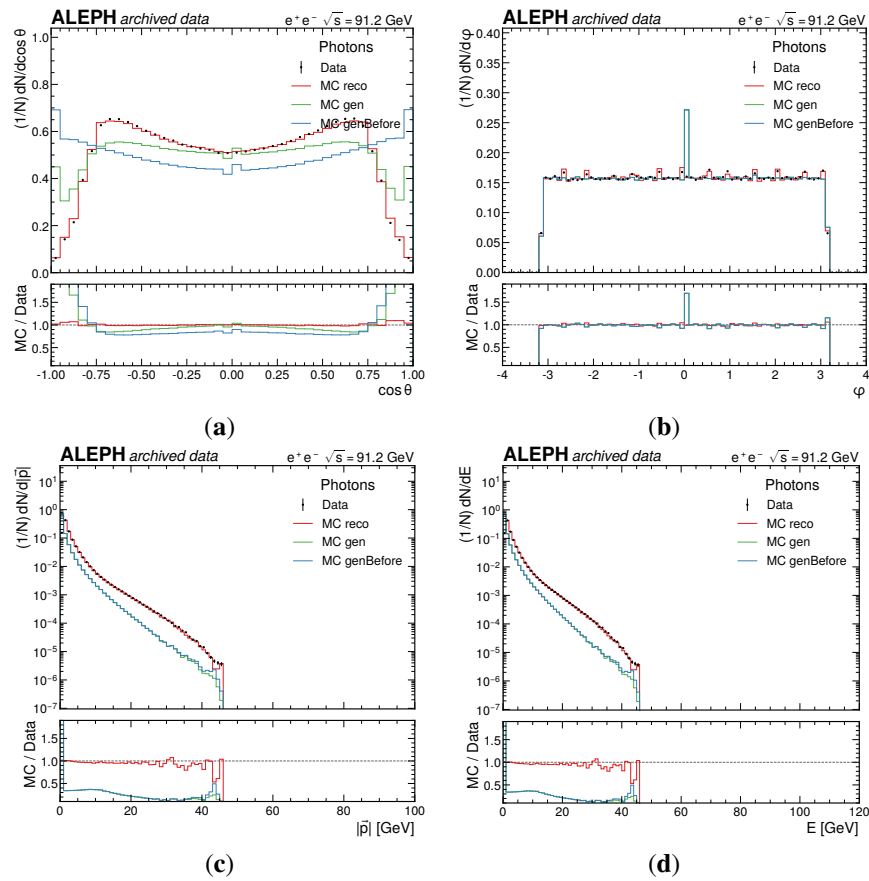


Figure 5. Kinematics for photons: (a) $\cos \theta$, (b) ϕ , (c) $|\vec{p}|$, and (d) energy.

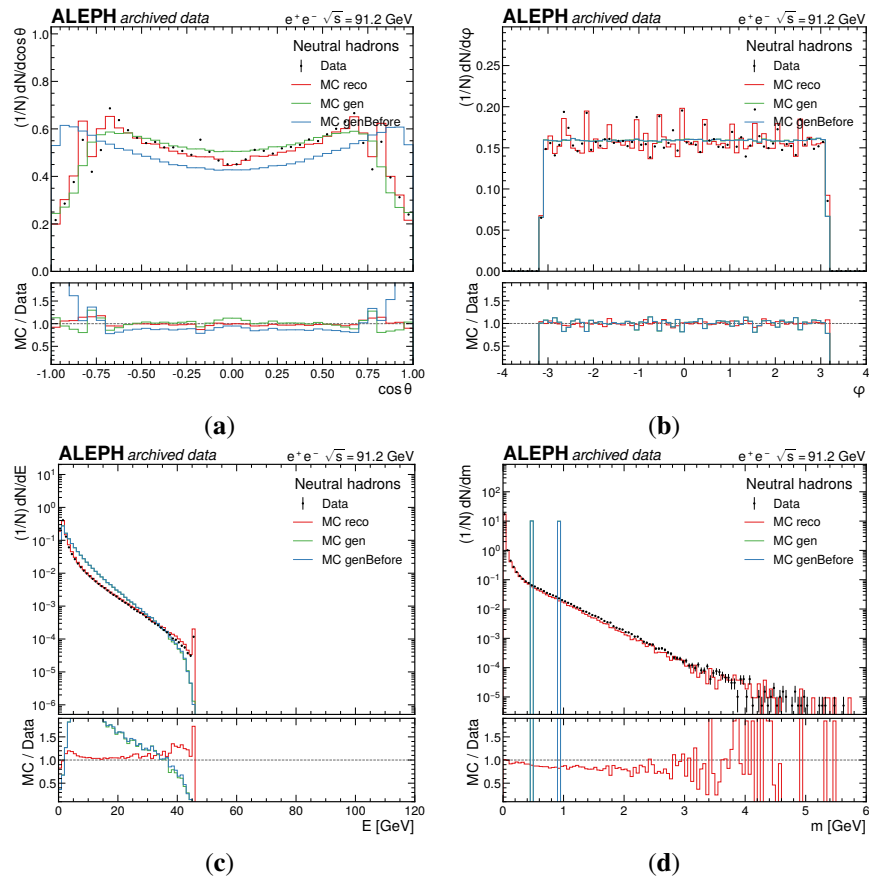


Figure 6. Kinematics of neutral hadrons: (a) $\cos \theta$, (b) ϕ , (c) energy, and (d) mass.

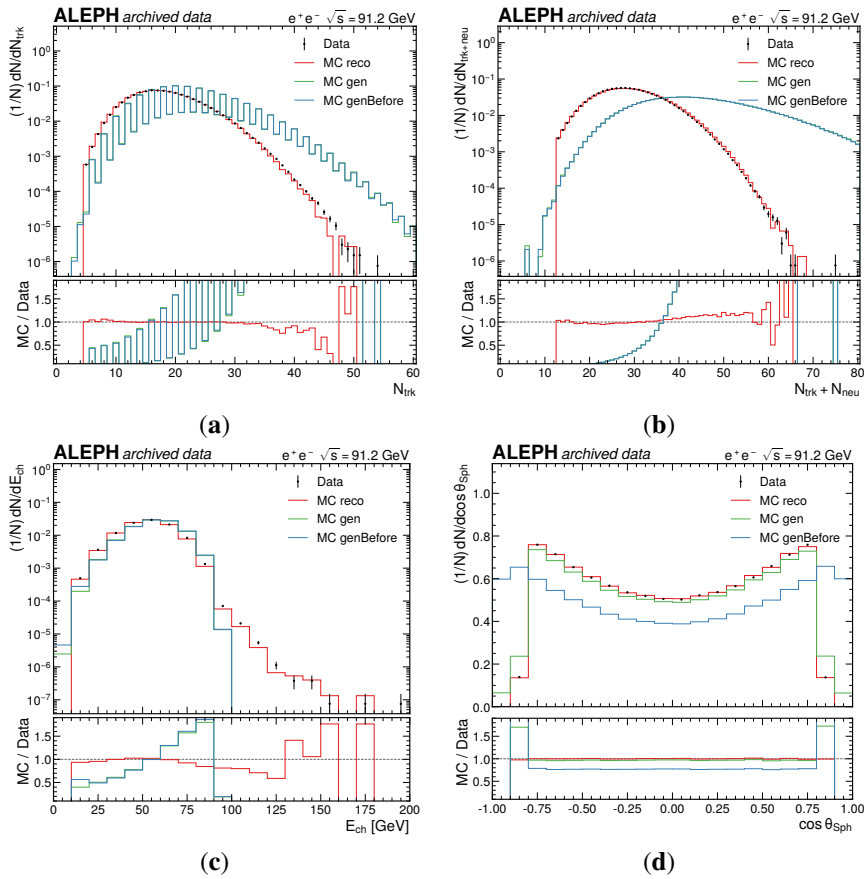


Figure 7. Event-level observables used in the nominal hadronic selection; legends include detector-level data/MC and generator-level references. The (a) number of charged tracks, (b) number of charged tracks plus neutral particles, (c) total charged energy, and (d) cosine of the sphericity axis polar angle are shown. The MC Gen and MC GenBefore ratios display an even-odd staggering: generator-level charged multiplicity has an approximately 6:1 even-to-odd ratio (arising from the Z-decay net-charge constraint and the specific ALEPH Pythia 6 particle record), while reconstruction efficiency losses smear individual tracks one at a time and equalise the even/odd balance in data and MC Reco. MC Reco is the relevant comparison for the analysis.

7. Analysis Method

7.1. Observable Definition and Construction

The thrust observable is defined as [26]

$$T = \max_{\hat{n}} \frac{\sum_i |\vec{p}_i \cdot \hat{n}|}{\sum_i |\vec{p}_i|}, \quad (1)$$

with all selected charged and neutral reconstructed objects from Section 6 entering the sum. The thrust is built using selected charged and neutral particles. Those objects are input to an optimization algorithm that iteratively determines the thrust axis. The algorithm we use is a clone of the HERWIG++ implementation [34–37], which is an exact calculation rather than one based on heuristics as is done by others [38]. The resolution is characterized using matched reco–gen event pairs. Figure 8 shows $(T_{\text{reco}} - T_{\text{gen}})$ as a function of T_{gen} . The resolution narrows toward $T \rightarrow 1$. To enable direct comparison, our work adopts the same binning as the previous ALEPH analyses [12,32], with 42 uniform bins over $0.58 \leq T < 1.0$ (bin width $\Delta T = 0.01$).

7.2. Unfolding Setup and Regularization Choice

Detector correction is performed with the IBU method of D’Agostini [39,40], as implemented in RooUnfold [41],

$$U_j^{(n+1)} = \frac{1}{\epsilon_j} \sum_i M_i \frac{R_{ij} \pi_j^{(n)}}{\sum_k R_{ik} \pi_k^{(n)}}, \quad (2)$$

where M_i is the measured detector-level spectrum, R_{ij} is the response, and $\pi_j^{(n)}$ is the iteration prior. The prior for the first iteration is taken as the generator-level MC truth distribution $\pi_j^{(0)} = N_{\text{gen}}(j)$, the standard RooUnfold default. This choice initializes the unfolding close to the expected truth shape and avoids any flat-prior amplification of high-frequency fluctuations in the first iteration. The response matrix is filled with MC events that pass the detector-level selection and the corresponding generator-level event as shown in Figure 9. A selected event whose reconstructed thrust falls outside the unfolding range enters instead through RooUnfold's miss term, which supplies the efficiency factor $1/\epsilon_j$ appearing below.

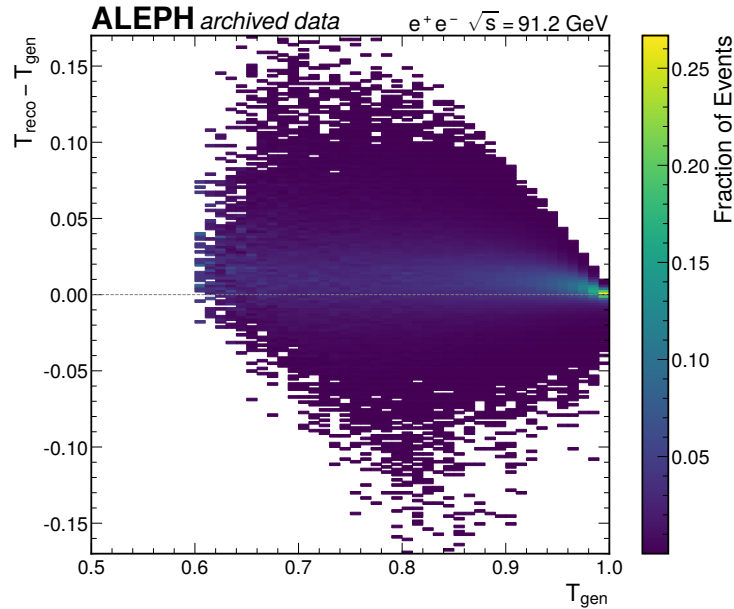


Figure 8. Thrust resolution ($T_{\text{reco}} - T_{\text{gen}}$) versus T_{gen} for matched MC event pairs.

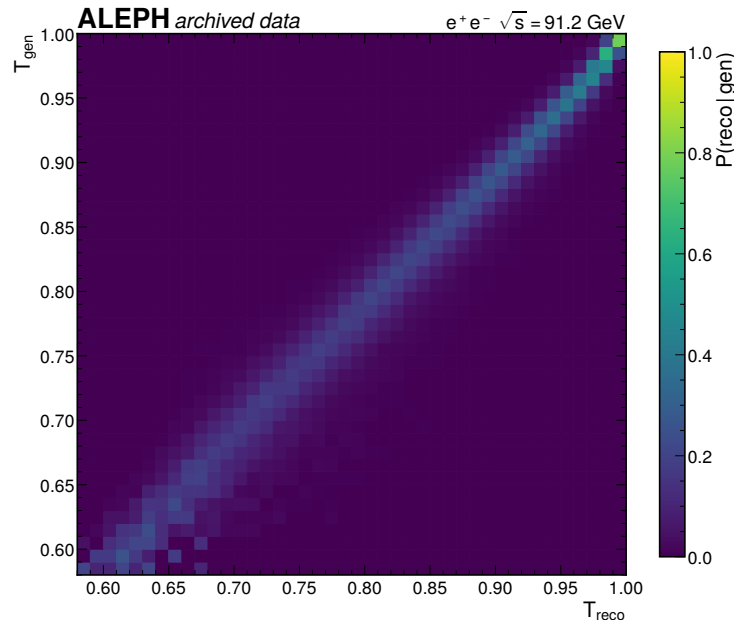


Figure 9. Nominal response probability matrix for thrust.

The IBU regularization term for the number of iterations is kept at a nominal value of $N_{\text{iter}} = 5$ as found by previous studies [2], with $N_{\text{iter}} = 4$ and 6 propagated as unfolding-regularization systematic branches. The iteration-choice diagnostics, including refolded spectra at $N = 4, 5, 6, 10, 100$ compared with detector-level data, confirm that the closure-shape χ^2/ndf is stable for $N_{\text{iter}} \in \{4, 5, 6\}$. The spread in the unfolded thrust spectrum across this range is below 1% in 37 of the 42 reported bins (median 0.3%), rising to at most 6.6% near the two-jet edge ($T \lesssim 0.62$) where the steeply falling spectrum amplifies regularization sensitivity.

The response matrix has a diagonal fraction of $f_{\text{diag}} = 0.33$, meaning approximately two-thirds of events migrate out of their generated thrust bin at reconstruction level. This level of bin-to-bin migration is expected for

thrust at ALEPH given the finite charged-particle momentum resolution and the sensitivity of thrust to soft particles near the event boundary. It implies that the unfolding correction is non-trivial and that the choice of regularization strength meaningfully affects the result. This also motivates the need to measure the final bin-by-bin correlation to assess the overall statistical strength.

The miss term, the generator-level events with no matched reconstructed counterpart, is shown in Figure 10. These are events that fail the reco-level hadronic event selection despite being generated within the fiducial volume. The miss rate is concentrated at low thrust ($T \lesssim 0.7$) and in events with low charged multiplicity, consistent with the hadronic event selection removing poorly reconstructed events. RooUnfold accounts for miss events via $\text{response.Miss}(x_{\text{gen}})$, which corrects the unfolded spectrum for this acceptance effect.

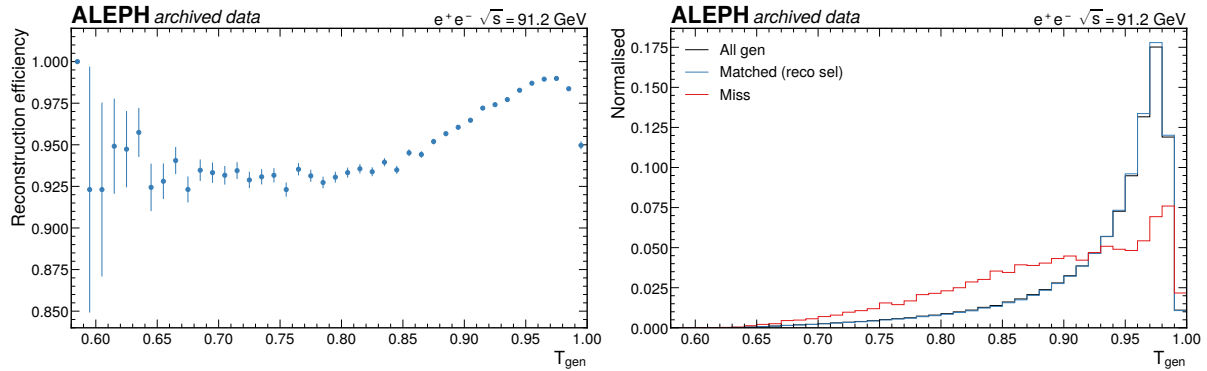


Figure 10. Left: reco efficiency (matched/all gen) as a function of T_{gen} , with binomial uncertainties. Right: thrust distributions for all gen, matched, and miss events (normalized to unit area). Miss events are concentrated at low thrust and low multiplicity, consistent with events failing the hadronic event selection.

7.3. Hadronic and QED ISR Correction

After unfolding, two multiplicative corrections are applied. A correction for the hadronic event-selection,

$$C_{\text{had}}(j) = \frac{N_{\text{genBefore}}(j)}{N_{\text{gen}}(j)}, \quad (3)$$

accounts for the efficiency of the selection at the generator level. The correction factor is derived from the archived PYTHIA 6 sample [33] and is applied so that the final spectrum restores the yield lost by the hadronic selection.

Then a correction for electromagnetic (QED) initial-state radiation (ISR) is applied,

$$C_{\text{ISR}}(j) = \frac{N_{\text{no ISR}}^{\text{gen}}(j)}{N_{\text{ISR}}^{\text{gen}}(j)}. \quad (4)$$

This factor is applied to enable comparisons with state-of-the-art theory calculations which do not include QED ISR. The correction is derived from standalone PYTHIA 8.317 samples generated with ISR switched off and on [42]. PYTHIA 8.317 is chosen here for simplicity; a higher-order generator such as KKMCEE 5.00.02 would be preferred for a full thrust measurement [43]. The ratio histograms are pre-computed once into a cache file and applied identically to the nominal and all systematic variation branches, so that the ISR correction cancels in the systematic shifts and its residual uncertainty (from finite MC statistics on the ratio) is propagated as a separate component Δ_{ISR} (Section 8.6).

For the nominal result, the data statistical covariance is computed by RooUnfold using analytical error propagation (`RooUnfold::kCovariance`), which propagates Poisson uncertainties in the measured data through the IBU iteration chain without toy MC sampling. The response matrix's own finite statistics are not included by that call and are propagated separately by a Poisson bootstrap over the simulated events, as described in Section 8.1. The resulting covariance matrices are read from the unfolding output and transformed to the normalized differential-thrust basis used in this work. Non-statistical component covariances are built with a configurable model, and all covariance products are exported over the reported bins.

8. Systematic Uncertainties

Systematic uncertainties arise from several components: data statistical, MC statistical, unfolding-regularization, experimental variations, theory unfolding prior, hadronic event-selection correction, and ISR correction. The sections below describe each component.

8.1. Statistical Uncertainty

The data statistical uncertainty is propagated analytically by RooUnfold through the IBU iteration chain, without toy-MC sampling, yielding a full statistical covariance matrix $\text{Cov}_{ij}^{(\text{stat})}$. The per-bin fractional uncertainty $\Delta_{\text{stat}}(j) = \sqrt{\text{Cov}_{jj}^{(\text{stat})}}/U_j^{\text{corr}}$ is typically below 2% in the central range and nearly uniform across the spectrum. The uncertainty is shown in Figure 11.

MC statistical uncertainties enter the budget through the two corrections derived from finite simulated samples: the MC statistical uncertainty of the ISR-off and ISR-on samples is propagated onto $C_{\text{ISR}}(j)$, and the archived-MC statistical uncertainty of $N_{\text{genBefore}}$ and N_{gen} is propagated onto the hadronic event-selection correction $C_{\text{had}}(j)$. The finite statistics of the response matrix itself are propagated separately, by giving every matched simulated event a Poisson weight and re-unfolding the data over 200 such replicas; the sample covariance of the resulting spectra is taken as $\text{Cov}_{ij}^{(\text{MCstat})}$. This is required because RooUnfold::kCovariance transports the measured data covariance only, and because the simulated sample is smaller than the data sample, so the response statistical term exceeds the data statistical term in every bin.

8.2. Unfolding Regularization Uncertainty

The unfolding regularization uncertainty is closely related: two additional unfolding passes at $N_{\text{iter}} = 4$ and $N_{\text{iter}} = 6$ quantify the sensitivity of the result to the regularization choice. The iteration envelope $\Delta_{\text{iter}}(j) = \max(|\delta_{\text{iter4}}(j)|, |\delta_{\text{iter6}}(j)|)$ is sub-percent in the bulk of the spectrum, confirming that the nominal $N_{\text{iter}} = 5$ sits on a stable plateau and that regularization is not a limiting source of uncertainty. Figure 11 shows the signed per-bin shifts for both components across the measurement range. The iteration shifts change sign across the range because $N_{\text{iter}} = 4$ and 6 bracket the nominal, and the bin-to-bin alternation at low thrust reflects the anti-correlation that regularization induces between neighboring bins.

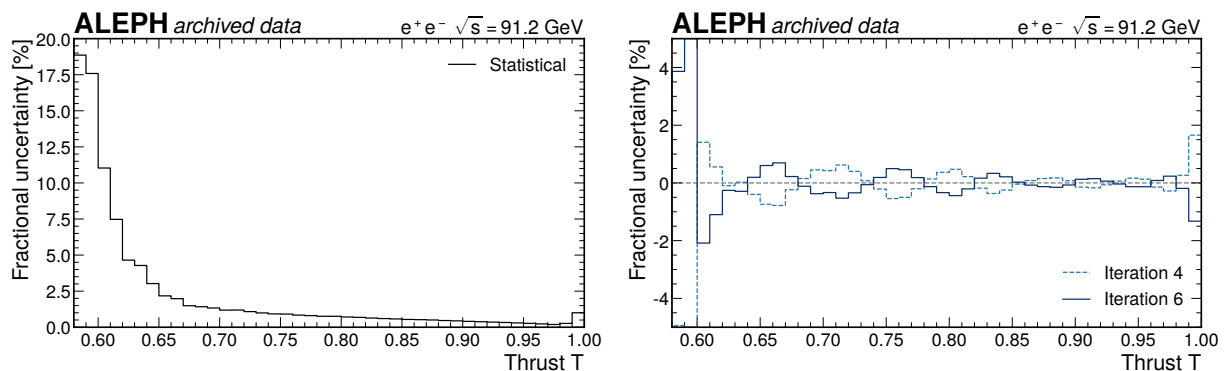


Figure 11. Signed per-bin shifts for the statistical (**left**) and unfolding-iteration (**right**) uncertainty components. The statistical component is nearly flat and small across the spectrum. The iteration panel shows $N_{\text{iter}} = 4$ and 6 relative to the nominal $N_{\text{iter}} = 5$.

8.3. Experimental Systematic Uncertainties

Experimental systematic uncertainties are handled following the standard ALEPH procedure [12] to vary individual selections, rerunning the full analysis chain, and measuring the difference with respect to the nominal as the uncertainty. The explored systematics are defined below.

- iter4/iter6: unfolding passes at $N_{\text{iter}} = 4$ and 6.
- ntpc7: tightened TPC hit requirement ($N_{\text{TPC}} > 7$, nominal > 4).
- pt04: raised charged- p_T threshold (> 0.4 GeV, nominal > 0.2 GeV).
- ech10: loosened charged-energy cut (> 10 GeV, nominal > 15 GeV).
- with_met: missing transverse energy added as a pseudo-particle.
- nes_up/nes_down: neutral energy scale varied up and down.
- ner: neutral objects randomly removed with 2.5%, probing acceptance efficiency.

Figure 12 shows the signed per-bin shifts for all variations. The largest contributions arise at low thrust ($T \lesssim 0.62$) where the steeply rising spectrum amplifies small absolute shifts; across the central range the experimental envelope is at the few-percent level.

The neutral-object response variations are chosen in preference to the standard ALEPH approach of removing all neutral particles within 1–2 GeV from the event [12]. The reason is that removing neutrals entirely redefines the

observable: the resulting distribution is no longer the energy-flow thrust T_{EF} but a different quantity, charged-only thrust T_{ch} , computed from a different particle-level definition. Treating the difference $T_{\text{EF}} - T_{\text{ch}}$ as a systematic uncertainty on the energy-flow measurement would conflate two distinct observables and produce an uncertainty that is unphysically large. Instead, `nes_up`, `nes_down`, and `ner` vary the detector response to neutral objects while holding the particle-level definition fixed, providing a physically meaningful probe of calorimeter modeling uncertainties without changing what is being measured.

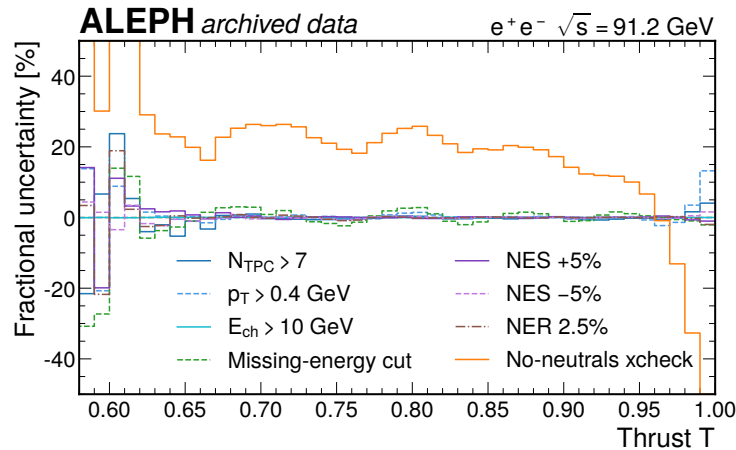


Figure 12. Signed per-bin shifts for the seven experimental variation branches: the four track/event branches (`ntpc7`, `pt04`, `ech10`, `with_met`) and the three neutral-response branches (`nes_up`, `nes_down`, `ner`). The `no_neutrals` curve is a cross-check and does not enter the uncertainty.

To quantify this distinction, the full unfolding chain was repeated with neutral particles (`pwflag` $\in \{4, 5\}$) removed, yielding a charged-only thrust distribution as shown in Figure 13. Both the energy-flow and charged-only results are processed through the complete analysis chain. The reconstructed particle collection differs between the two, but both are unfolded to the same energy-flow particle-level definition. Both distributions are shown with statistical uncertainties only; the full systematic suite is not propagated for this diagnostic comparison, and the charged-only branch is excluded from the nominal total uncertainty combination. The charged-only spectrum is systematically broader than the energy-flow, as expected from the missing neutral-energy deposits.

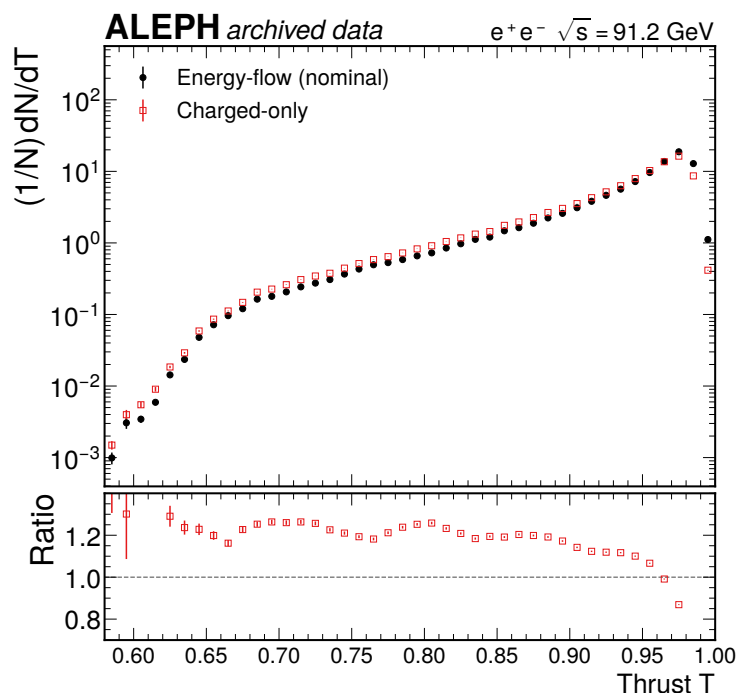


Figure 13. Comparison of the charged-only (red open circles) and energy-flow (black filled circles) unfolded thrust distributions. Error bars represent statistical uncertainties only.

The track and neutral object variations are combined as described below. The full experimental envelope is then combined in quadrature.

$$\Delta_{\text{track/event}}(j) = \sqrt{\sum_{v \in \{\text{ntpc7, pt04, ech10, with_met}\}} |\delta_v(j)|^2}. \quad (5)$$

$$\Delta_{\text{neutral}}(j) = \sqrt{\max(|\delta_{\text{nes_up}}(j)|, |\delta_{\text{nes_down}}(j)|)^2 + |\delta_{\text{ner}}(j)|^2}. \quad (6)$$

$$\Delta_{\text{exp}}(j) = \sqrt{\Delta_{\text{track/event}}^2(j) + \Delta_{\text{neutral}}^2(j)}. \quad (7)$$

8.4. Theory Systematic Uncertainties

Theory systematic variations probe the residual model dependence introduced by the Monte Carlo generator used in the unfolding. The nominal response matrix is constructed from PYTHIA 6 [33], the ALEPH simulation. Three additional generators are used to reweight the generator-level thrust distribution event-by-event, producing alternative response matrices that span a wide range of parton-shower algorithms and hadronization models: PYTHIA 8.317 with Monash tune, dipole-ordered shower [42], HERWIG 7.1.5, angular-ordered shower with cluster hadronization [35,44–46], and SHERPA 2.2.6 with AHADIC++ hadronization [47]. The theory envelope is taken as the maximum absolute shift across the three generators:

$$\Delta_{\text{theory}}(j) = \max(|\delta_{\text{pythia8}}(j)|, |\delta_{\text{herwig}}(j)|, |\delta_{\text{sherpa}}(j)|). \quad (8)$$

The three generators diverge progressively at lower thrust, where multi-jet configurations and hard, wide-angle gluon emission dominate and the parton-shower and matrix-element matching differ between generators; there the reweighted generator-level spectra move coherently and by similar amounts. The envelope also rises in the last bin, $0.99 \leq T < 1.00$, where the reweighted spectra are least well constrained: the bin sits at the edge of the two-jet limit, and the shift there is not smoothly connected to its neighbours. The theory envelope is the dominant uncertainty contribution across the central measurement range, as shown in Figure 14.

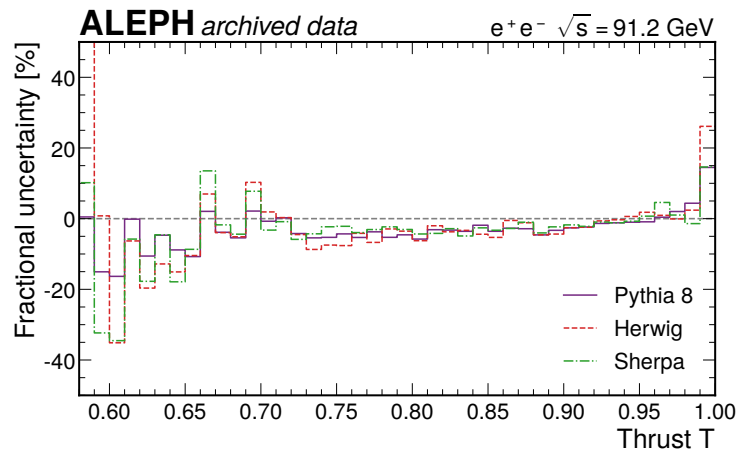


Figure 14. Signed per-bin shifts for the theory prior variations from PYTHIA 8.317, HERWIG 7.1.5, and SHERPA 2.2.6.

8.5. Hadronic Event-Selection Correction

After unfolding, a bin-wise hadronic event-selection correction is applied to account for the fraction of hadronic events lost to the generator-level selection:

$$C_{\text{had}}(j) = \frac{N_{\text{genBefore}}(j)}{N_{\text{gen}}(j)}, \quad U_j^{\text{corr}} = U_j C_{\text{had}}(j). \quad (9)$$

This correction maps the unfolded spectrum from the generator-level hadronic-event-selected phase space to the full inclusive hadronic phase space. The uncertainty on this correction is propagated as an independent component Δ_{had} . The correction is computed from two generator-level histograms: $t_{\text{genBefore}}$ (events before hadronic selection) and t_{gen} (events passing hadronic selection). Statistical uncertainties on both histograms are propagated through the ratio using standard error propagation, yielding symmetric bin-wise up and down variations

$C_{\text{had}}^{\pm}(j)$. These variations are applied to the unfolded spectrum and normalized to the differential cross-section basis, producing up and down shifted spectra.

8.6. QED ISR Correction

The QED ISR correction removes the contribution of events where the e^+e^- center-of-mass energy is reduced by photon radiation before the hard scatter, mapping the measured distribution from the ISR-inclusive sample to the pure $\sqrt{s} = M_Z$ distribution:

$$C_{\text{ISR}}(j) = \frac{N_{\text{no ISR}}^{\text{gen}}(j)}{N_{\text{ISR}}^{\text{gen}}(j)}. \quad (10)$$

Both published ALEPH event-shape analyses apply this correction [12,32]. The 1998 ALEPH analysis [32] defines it as the ratio of generator-level thrust distributions produced with and without ISR, using JETSET 7.4. The 2004 analysis [12] uses KORALZ, which provides a more accurate treatment of ISR, to generate the corrected distribution.

The ISR correction is applied bin-by-bin after the hadronic correction using standalone PYTHIA 8.317 samples with ISR disabled (`isr0`) and enabled (`isr1`), as described in Section 7.3. The uncertainty on this correction is derived from the finite Monte Carlo statistics. Up and down variations $C_{\text{ISR}}^{\pm}(j) = C_{\text{ISR}}(j) \cdot (1 \pm \sigma_{\text{rel}}(j))$ are applied to the unfolded spectrum and the envelope is taken as $\Delta_{\text{ISR}}(j)$. The uncertainty is purely MC-statistical in nature for simplicity, but should be assessed in closer detail for future work.

8.7. Covariance Evaluation

For the nominal run, the statistical covariance is computed by RooUnfold using analytical error propagation (`RooUnfold::kCovariance`), which propagates Poisson uncertainties in the data and response histograms through the IBU iteration chain without toy MC sampling. The resulting covariance matrix is read from the unfolding output and transformed to the normalized differential-thrust basis used in this work.

Non-statistical component covariances are built with a configurable model. The current nominal release uses the `template_correlated` option, where each non-statistical component covariance is built as

$$\text{Cov}_{ij}^{(c)} = \Delta_c(i) \Delta_c(j) \rho_{ij}^{\text{stat}}, \quad (11)$$

with ρ_{ij}^{stat} taken from the statistical unfolding covariance correlation template. This keeps component magnitudes tied to the measured per-bin shifts while avoiding a purely diagonal systematic-covariance approximation. For cross-check studies, `diagonal` and `block` component models are also available [41].

8.8. Combination Model

The total bin uncertainty combines all seven components in quadrature:

$$\Delta_{\text{tot}}(j) = \sqrt{\Delta_{\text{stat}}^2(j) + \Delta_{\text{iter}}^2(j) + \Delta_{\text{exp}}^2(j) + \Delta_{\text{theory}}^2(j) + \Delta_{\text{had}}^2(j) + \Delta_{\text{ISR}}^2(j)}. \quad (12)$$

The quadrature sum is used since the components originate from distinct sources and the covariance cross terms are expected to be small.

8.9. Uncertainty Breakdown

Figure 15 shows the absolute fractional uncertainties. Theory is the dominant source across the central range, where the experimental envelope is smaller by roughly a factor of two. The statistical and iteration components remain sub-percent in the bulk.

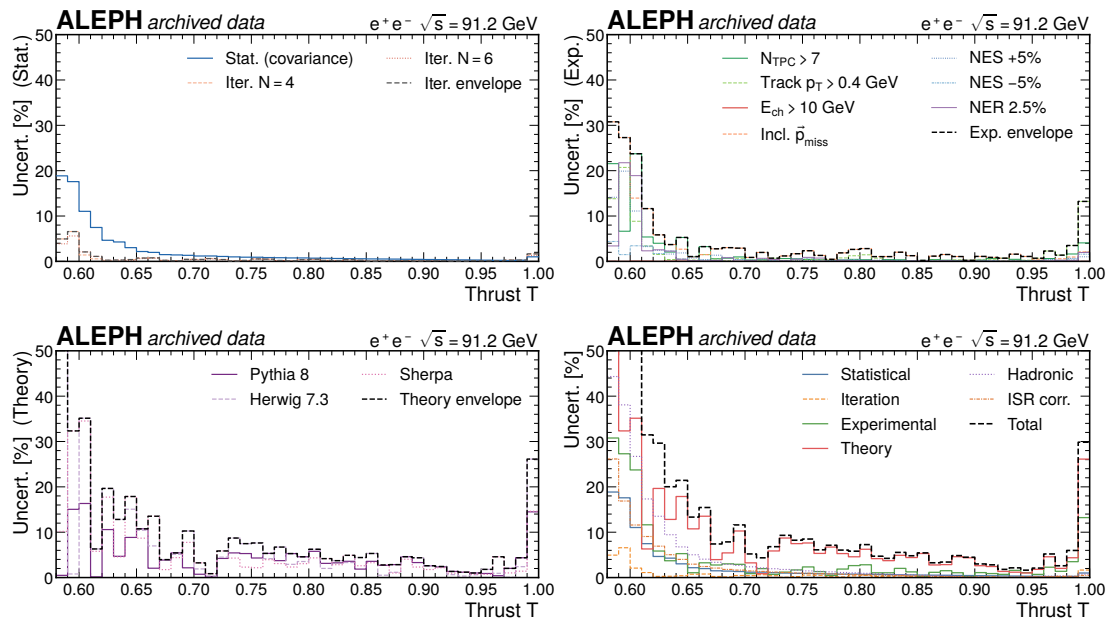


Figure 15. Fractional uncertainties: (**top left**) statistical and unfolding-iteration components, (**top right**) experimental variation branches with group envelope, (**bottom left**) theory reweighting branches with theory envelope, (**bottom right**) all groups overlaid.

9. Results

Figure 16 shows the unfolded spectrum with the total uncertainty band and comparison to the published ALEPH 2004 result [12]. Agreement is good across the bulk of the spectrum ($0.62 \leq T < 0.99$), with the largest discrepancy in the highest thrust bin. This is physically expected: the dijet limit is sensitive to MC modeling of soft hadrons, and the hadronic correction $C_{had}(j)$ and ISR correction $C_{ISR}(j)$ carry growing uncertainties there. Potential dataset differences between the current analysis and the ALEPH 2004 result can also contribute at the endpoint.

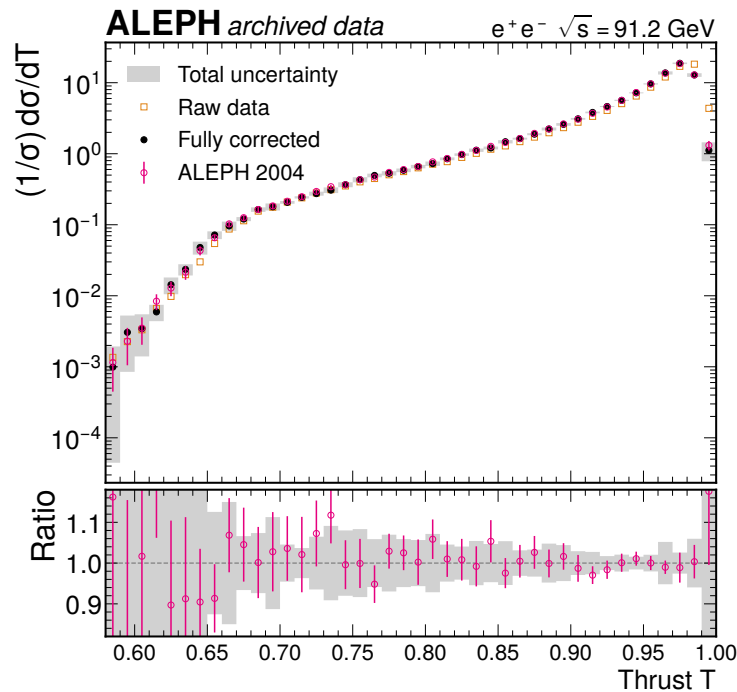


Figure 16. Fully corrected thrust distribution with total uncertainty band compared to the published ALEPH 2004 result.

The per-component correlation matrices are derived from the individual component covariance matrices $\text{Cov}_{ij}^{(c)}$ described in Section 8.7. For each component c , the correlation is

$$\rho_{ij}^{(c)} = \frac{\text{Cov}_{ij}^{(c)}}{\sqrt{\text{Cov}_{ii}^{(c)} \text{Cov}_{jj}^{(c)}}}. \quad (13)$$

Figure 17 shows all seven components simultaneously, enabling direct comparison of their correlation structures. The components differ mainly in correlation length, and the ordering follows directly from whether the underlying variation is bin-local or coherent across the spectrum. The data statistical component exhibits the narrow correlation band characteristic of IBU: neighboring bins are positively correlated through the unfolding propagation, but the correlation falls off quickly with bin separation. The response matrix MC statistical component behaves the same way, since both originate in bin-local Poisson fluctuations spread over neighboring bins by the unfolding. The ISR correction is nearly diagonal, consistent with its sub-percent size and with the bin-local character of the MC statistical uncertainty on the $C_{\text{ISR}}(j)$ ratio, and the hadronic correction is modestly correlated, commensurate with its $\sim 2\%$ magnitude in the central range. The iteration component is somewhat broader, as the regularization responds coherently to low-frequency features of the spectrum rather than to individual bins. Broader still is the experimental component, driven by the neutral-energy response shifts (`nes_up`, `nes_down`), which displace the spectrum in the same direction over a wide range of bins. The theory component has the strongest and most extended off-diagonal correlations: generator reweighting alters the parton-shower and hadronization description, and therefore the overall spectral shape, rather than any localized region. It is the dominant source of off-diagonal structure.

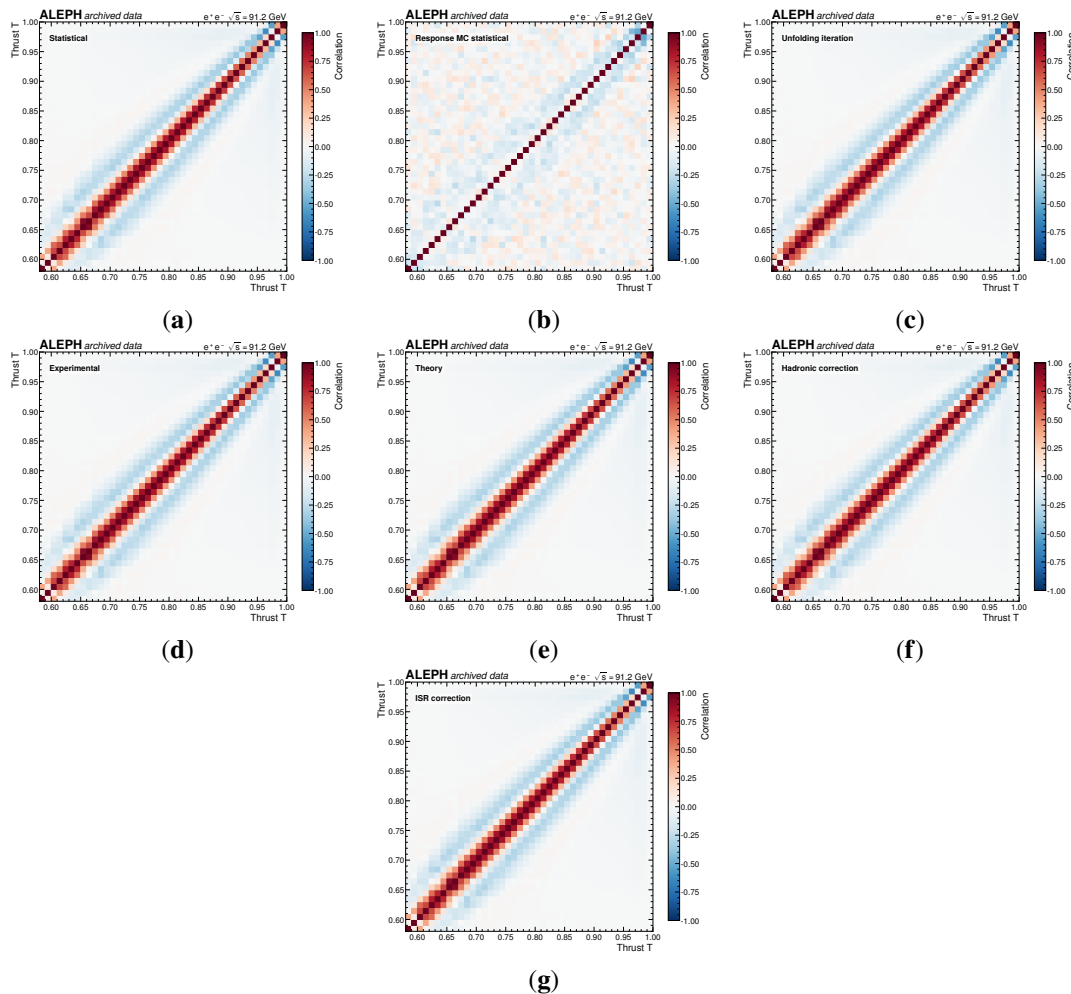


Figure 17. Correlation matrices for the (a) data statistical, (b) response matrix MC statistical, (c) unfolding iteration, (d) experimental variation, (e) theory prior, (f) hadronic event selection correction, and (g) QED ISR correction components as indicated.

10. Conclusions

AI agents have successfully carried out a complete thrust measurement using archived ALEPH 1994 data at $\sqrt{s} = 91.2$ GeV under physicist guidance and collaboration. The analysis propagates uncertainty components arising from statistical, experimental, and theoretical sources and combines them in quadrature. The corresponding

correlation matrices are derived from the covariance structure. The implementation of the ISR correction, following the prescription of both published ALEPH event-shape analyses [12,32], brings the measured thrust distribution into good agreement with the published ALEPH 2004 result, with the only notable discrepancy at the $T \rightarrow 1$ dijet endpoint. The AI-agent operating model demonstrated in this cycle, executable-workflow orchestration under physicist supervision with full provenance tracking, provides a scalable framework for precision QCD as a testing ground for closed theory–experiment loops in fundamental physics.

Lessons from AI–Physicist Collaboration

Carrying out this analysis has also produced practical insights into how AI agents and physicists can work together effectively. Our workflow followed a constrained human-in-the-loop cycle in which the physicist defined physics objectives, acceptance criteria, and interpretation authority, while the AI agent owned implementation and execution. Crucially, experienced physicists contribute domain context that cannot yet be reduced to formal specifications alone, including judgments about physically meaningful binning, acceptable regularization behavior, uncertainty communication, and when apparent anomalies warranted deeper investigation. A major future goal identified in this work is to translate this expert knowledge into explicit, code-based verification layers.

From this experience, several recurring challenge classes emerged. Implicit physics priors, expert expectations about acceptable binning, regularization behavior, or uncertainty presentation, were often assumed rather than explicitly stated, requiring iterative clarification. Figure formatting and style choices (label offsets, legend placement, pad geometry) required many short feedback loops, with cosmetic shortcomings sometimes affecting interpretability and therefore physics conclusions. Synchronization across rapidly evolving scripts, outputs, and note text demanded continuous provenance refreshing. These lessons motivate creation of reusable tools, stronger configuration contracts, clearer channels for encoding expert heuristics into machine readable constraints in future iterations of this program, and new verification layers throughout the autonomous analysis flow.

Author Contributions

A.B. (Anthony Badea): Conceptualization, primary analyzer, data curation, methodology, implementation, draft preparation. M.M., Y.-J.L.: Data curation, review, feedback. C.M., A.B. (Austin Baty), Y.C., H.B., G.M.I., J.Z., Y.-C.C., T.-A.S.: Data curation. E.-P.A.: Experimental collaboration under which all authors contribute significantly to data curation, physics analysis, draft preparation, and review. All authors have read and agreed to the published version of the manuscript.

Funding

Y.C. and J.Z. are supported in part by the U.S. National Science Foundation under ID 2514008. A. Baty is supported by the U.S. Department of Energy, Grant No. DE-FG02-94ER40865. H.B., Y.-C.C., G.M.I., C.M., Y.-J.L., and T.-A.S. are supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics under grant Contract Number DE-SC0011088. A. Badea is supported by the Schmidt Sciences Foundation and Enrico Fermi Institute.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. The analysis code is made available through the Electron-Positron Alliance git organization: <https://github.com/orgs/Electron-Positron-Alliance>.

Acknowledgments

The authors would like to thank Roberto Tenchini and Guenther Dissertori from the ALEPH Collaboration for their useful comments and suggestions on the use of ALEPH data. We thank all the members of the Electron-Positron Alliance for their tremendous efforts to resurrect the LEP dataset, ultimately enabling the studies presented in this paper.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used Anthropic Claude and OpenAI GPT (ChatGPT and Codex) to carry out the experimental analysis and draft the paper. After using this tool/service, the authors reviewed

and edited the content as needed and take full responsibility for the content of the published article."

References

1. The Electron-Positron Alliance. Available online: <https://ee-alliance.org/home/> (accessed on 18 February 2026).
2. Badea, A.; Baty, A.; Chang, P.; et al. Measurements of Two-Particle Correlations in e^+e^- Collisions at 91 GeV with ALEPH Archived Data. *Phys. Rev. Lett.* **2019**, *123*, 212002. <https://doi.org/10.1103/physrevlett.123.212002>.
3. Chen, Y.; Badea, A.; Baty, A.; et al. Jet energy spectrum and substructure in e^+e^- collisions at 91.2 GeV with ALEPH Archived Data. *J. High Energy Phys.* **2022**, 2022, 8. [https://doi.org/10.1007/jhep06\(2022\)008](https://doi.org/10.1007/jhep06(2022)008).
4. Chen, Y.C.; Chen, Y.; Badea, A.; et al. Long-range near-side correlation in e^+e^- collisions at 183–209 GeV with ALEPH archived data. *Phys. Lett. B* **2024**, *856*, 138957. <https://doi.org/10.1016/j.physletb.2024.138957>.
5. Bossi, H.; Chen, Y.; Chen, Y.C.; et al. Energy Correlators from Partons to Hadrons: Unveiling the Dynamics of the Strong Interactions with Archival ALEPH Data. *arXiv* **2025**, arXiv:2511.00149.
6. Badea, A.; Baty, A.; Bossi, H.; et al. Unbinned measurement of thrust in e^+e^- collisions at $\sqrt{s} = 91.2$ GeV with ALEPH archived data. *arXiv* **2025**, arXiv:2510.22038.
7. Anthropic. Accelerating Scientific Research. Available online: <https://www.anthropic.com/news/accelerating-scientific-research> (accessed on 15 January 2026).
8. OpenAI. Advancing Science and Math with GPT-5.2. Available online: <https://openai.com/index/gpt-5-2-for-science-and-math/> (accessed on 11 December 2025).
9. Google DeepMind. Gemini Deep Think: Accelerating Mathematical and Scientific Discovery. Available online: <https://deepmind.google/blog/accelerating-mathematical-and-scientific-discovery-with-gemini-deep-think/> (accessed on 11 February 2026).
10. OpenAI. GPT-5.2 Derives a New Result in Theoretical Physics. Available online: <https://openai.com/index/new-result-theoretical-physics/> (accessed on 13 February 2026).
11. Google DeepMind. AlphaProteo: Generating Novel Proteins for Biology and Health Research. Available online: <https://deepmind.google/blog/alphaproteo-generates-novel-proteins-for-biology-and-health-research/> (accessed on 13 February 2026).
12. ALEPH Collaboration. Studies of QCD at e^+e^- centre-of-mass energies between 91 and 209 GeV. *Eur. Phys. J. C* **2004**, *35*, 457–486. <https://doi.org/10.1140/epjc/s2004-01891-4>.
13. ALEPH Documentation. Available online: <https://aleph-new.docs.cern.ch> (accessed on 16 September 2026).
14. Brown, T.; Mann, B.; Ryder, N.; et al. Language Models are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Virtual Event, 6–12 December 2020; vol. 33, pp. 1877–1901.
15. Wei, J.; Wang, X.; Schuurmans, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; vol. 35, pp. 24824–24837.
16. Abolhasani, M.; Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth.* **2023**, *2*, 483–492. <https://doi.org/10.1038/s44160-022-00231-0>.
17. MacLeod, B.P.; Parlane, F.G.L.; Morrissey, T.D.; et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **2020**, *6*, eaaz8867. <https://doi.org/10.1126/sciadv.aaz8867>.
18. Chen, Z.; Chen, S.; Ning, Y.; et al. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. *arXiv* **2024**, arXiv:2410.05080.
19. Lu, C.J.J.; Lu, C.; Lange, R.T.; et al. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv* **2024**, arXiv:2408.06292.
20. Woodruff, D.P.; Cohen-Addad, V.; Jain, L.; et al. Accelerating Scientific Research with Gemini: Case Studies and Common Techniques. *arXiv* **2026**, arXiv:2602.03837.
21. Bubeck, S.; Coester, C.; Eldan, R.; et al. Early science acceleration experiments with GPT-5. *arXiv* **2025**, arXiv:2511.16072.
22. Guevara, A.; Lupsasca, A.; Skinner, D.B.; et al. Single-minus gluon tree amplitudes are nonzero. *arXiv* **2026**, arXiv:2602.12176.
23. Schwartz, M.D. Resummation of the C-Parameter Sudakov Shoulder Using Effective Field Theory. *arXiv* **2026**, arXiv:2601.02484.
24. Wang, H.; Fu, T.; Du, Y.; et al. Scientific discovery in the age of artificial intelligence. *Nature* **2023**, *620*, 47–60. <https://doi.org/10.1038/s41586-023-06221-2>. Correction in: *Nature* **2023**, *621*, E33. <https://doi.org/10.1038/s41586-023-06559-7>.
25. Messeri, L.; Crockett, M.J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **2024**, *627*, 49–58. <https://doi.org/10.1038/s41586-024-07146-0>.
26. Farhi, E. Quantum Chromodynamics Test for Jets. *Phys. Rev. Lett.* **1977**, *39*, 1587–1588. <https://doi.org/10.1103/physrevlett.39.1587>.
27. Abbate, R.; Fickinger, M.; Hoang, A.H.; et al. Thrust at N^3LL with Power Corrections and a Precision Global Fit for $\alpha_s(m_Z)$. *Phys. Rev. D* **2011**, *83*, 074021. <https://doi.org/10.1103/physrevd.83.074021>.

28. Benitez, M.A.; Hoang, A.H.; Mateu, V.; et al. On Determining $\alpha_s(m_Z)$ from Dijets in e^+e^- Thrust. *arXiv* **2024**, arXiv:2412.15164.
29. Particle Data Group; Workman, R.L.; Burkert, V.D.; et al. Review of Particle Physics. *Prog. Theor. Exp. Phys.* **2022**, 2022, 083C01. <https://doi.org/10.1093/ptep/ptac097>.
30. Decamp, D.; Deschizeaux, B.; Lees, J.P.; et al. ALEPH: A detector for electron-positron annihilations at LEP. *Nucl. Instrum. Methods Phys. Res. A* **1990**, 294, 121–178. [https://doi.org/10.1016/0168-9002\(90\)91831-u](https://doi.org/10.1016/0168-9002(90)91831-u).
31. ALEPH Collaboration; Buskulic, D.; Casper, D.; et al. Performance of the ALEPH detector at LEP. *Nucl. Instrum. Methods Phys. Res. A* **1995**, 360, 481–506. [https://doi.org/10.1016/0168-9002\(95\)00138-7](https://doi.org/10.1016/0168-9002(95)00138-7).
32. ALEPH Collaboration; Barate, R.; Buskulic, D.; et al. Studies of quantum chromodynamics with the ALEPH detector. *Phys. Rep.* **1998**, 294, 1–165. [https://doi.org/10.1016/s0370-1573\(97\)00045-8](https://doi.org/10.1016/s0370-1573(97)00045-8).
33. Sjöstrand, T.; Edén, P.; Friberg, C.; et al. High-energy-physics event generation with Pythia 6.1. *Comput. Phys. Commun.* **2001**, 135, 238–259. [https://doi.org/10.1016/s0010-4655\(00\)00236-8](https://doi.org/10.1016/s0010-4655(00)00236-8).
34. Brandt, S.; Dahmen, H.D. Axes and Scalar Measures of Two-Jet and Three-Jet Events. *Z. Phys. C* **1979**, 1, 61–70. <https://doi.org/10.1007/bf01450381>.
35. Bähr, M.; Gieseke, S.; Gigg, M.A.; et al. Herwig++ physics and manual. *Eur. Phys. J. C* **2008**, 58, 639–707. <https://doi.org/10.1140/epjc/s10052-008-0798-9>.
36. Buckley, A.; Butterworth, J.; Grellscheid, D.; et al. Rivet user manual. *Comput. Phys. Commun.* **2013**, 184, 2803–2819. <https://doi.org/10.1016/j.cpc.2013.05.021>.
37. Buckley, A. Thrust.hh Source File. RIVET Hep Forge 2011. Available online: https://rivet.hepforge.org/code/1.7.0/a00541_source.html (accessed on 25 February 2023).
38. Belle II Collaboration. Belle II Software: Thrust Class. Available online: https://software.belle2.org/prerelease-10-00-00a/doxygen/classBelle2_1_1Thrust.html#a82233fb167d2bdefb9b3b9e2450c30a9 (accessed on 14 July 2025).
39. D’Agostini, G. A Multidimensional unfolding method based on Bayes’ theorem. *Nucl. Instrum. Methods Phys. Res. A* **1995**, 362, 487–498. [https://doi.org/10.1016/0168-9002\(95\)00274-x](https://doi.org/10.1016/0168-9002(95)00274-x).
40. D’Agostini, G. Improved iterative Bayesian unfolding. *arXiv* **2010**, arXiv:1010.0632.
41. Abye, T. Unfolding algorithms and tests using RooUnfold. In Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, Geneva, Switzerland, 17–20 January 2011; pp. 313–318.
42. Bierlich, C.; Chakraborty, S.; Desai, N.; et al. A comprehensive guide to the physics and usage of PYTHIA 8.3. *SciPost Phys. Codebases* **2022**, 8. <https://doi.org/10.21468/scipostphyscodeb.8>.
43. Jadach, S.; Ward, B.F.L.; Wąs, Z.; et al. Multi-photon Monte Carlo event generator KKMCEE for lepton and quark pair production in lepton colliders. *Comput. Phys. Commun.* **2023**, 283, 108556. <https://doi.org/10.1016/j.cpc.2022.108556>.
44. Bellm, J.; Gieseke, S.; Grellscheid, D.; et al. Herwig 7.0/Herwig++ 3.0 release note. *Eur. Phys. J. C* **2016**, 76, 196. <https://doi.org/10.1140/epjc/s10052-016-4018-8>.
45. Bellm, J.; Gieseke, S.; Grellscheid, D.; et al. Herwig 7.1 Release Note. *arXiv* **2017**, arXiv:1705.06919.
46. Bewick, G.; Ferrario Ravasio, S.; Gieseke, S.; et al. Herwig 7.3 Release Note. *Eur. Phys. J. C* **2024**, 84, 1053. <https://doi.org/10.1140/epjc/s10052-024-13211-9>.
47. Bothmann, E.; Singh Chahal, G.; Höche, S.; et al. Event generation with Sherpa 2.2. *SciPost Phys.* **2019**, 7, 034. <https://doi.org/10.21468/scipostphys.7.3.034>.