A six-event-shape $\alpha_s(M_Z)$ pipeline and tension decomposition on archived ALEPH LEP1 hadronic-Z data ($\sqrt{s} = 91.2$ GeV) at NNLO+NLL: final results

alphas_eventshapes_aleph analysis team

Final documentation

Contents

Change Log

This note is the final, publication-ready version. The analysis was developed under a staged-unblinding protocol whose three input stages — an MC-expectation study, a 10%-subsample validation, and the full-data measurement — are recorded here for audit. The body of the note tells a single cohesive physics story on the full data; the staging history lives only in this Change Log.

Final version — documentation polish. Editorial pass on the full-data note: prose tightened, the staging history confined to this Change Log, the conventions completeness and validation-summary tables consolidated, figures aggregated and composed, and the single source of truth for every number confirmed against the machine-readable results. No physics content changed: the headline remains $\alpha_s(M_Z) = 0.1064 \pm 0.0198$ (NNLO+NLL, built but not validated), and the order label "NNLO+NLL (BUILT, NOT VALIDATED)" is carried throughout.

Full-data measurement (human-approved unblinding). The final result on the **complete 1994 sample — all 1,293,167 selected P1+P2+P3 events** (no subsample). Full unblinding was human-approved after the 10%-subsample review panel returned PASS, the authorized point under the blinding protocol (Section 1.4); a positive full-data assertion ($n_{\rm sel} = n_{\rm total} = 1{,}293{,}167$) documents that 100% of the data is used. The method is provably unchanged from the expectation and 10%-subsample studies — only the measured reconstructed spectrum flips to the full real data — and the perturbative prediction remains NNLO+NLL, built but not validated ([L3]/[L4], human-approved); the unblinding does not lift the theory limitation. The final combined value is $\alpha_s(M_Z) = 0.1064 \pm 0.0198$, sitting between the 10%-subsample 0.1103 ± 0.0159 (-0.25σ) and the MC-expectation 0.1019 ± 0.0165 ($+0.27\sigma$) and consistent with both within the theory-dominated error; the well-described $\tau{+}\rho_H$ subset gives 0.1059 ± 0.0189, the data-driven model-dependence term reproduces the 10%-subsample value (ratio 0.85–1.33), the fit-triviality gate passes (minimum fit $\chi^2 = 20.9$), and no pre-registered stop trigger fires.

10%-subsample validation. The first stage to use real ALEPH 1994 data: a fixed-seed (20260612) event-level uniform-random 10% subsample (129,321 of 1,293,167 events). It validated the data and the correction chain, not the theory, and gave the combined 0.1103 ± 0.0159. Its central physics finding — a real, coherent ~3–7% per-bin data−MC particle-level difference (up to 12% for $-\log_{10} y_{23}$), independently confirmed to be a genuine generator difference rather than a chain bug — motivated the data-driven model-dependence envelope (the $\Delta\alpha_s$ term M2, Section 5.1), now confirmed directly at full statistics.

MC-expectation study (blinded). The initial version, with all results derived from 1994 ALEPH full-simulation MC used as pseudo-data (blinding to real data preserved). It established the full six-observable unfolding pipeline (IBU primary, OmniFold cross-check), the NNLO+NLL theory framework (built, not validated), the two-parameter non-perturbative sector, the complete systematic program (including the orthogonalized order-truncation term $\sqrt{0.0035^2 - 0.0009^2} = 0.00338$), the per-observable and combined α_s extraction, the goodness-of-fit, the resolving power, and the comparison to PDG and the reference analyses — with the realized ± 0.0165 floor stated as ~4–5× the pre-registered $O(0.003$–$0.004)$ target because of the [L3]/[L4] N3LL→NLL order downscope.

Abstract

This note presents the **final result** of a modern unfolding pipeline and a six-observable tension decomposition on archived ALEPH LEP1 hadronic-Z data at $\sqrt{s} = 91.2$ GeV. The correction and fit chain is run on the **complete 1994 sample — all 1,293,167 selected P1+P2+P3 events** (no subsample). The method is provably unchanged from the prior MC-expectation and 10%-subsample studies; only the measured input flips to the full real data. The perturbative prediction is **NNLO+NLL, built but not validated** at this order, so the central coupling is reported as the value the unvalidated fit prefers, and this is therefore a **consistency and tension-decomposition result at LEP-era theory precision**, not a competitive α_s. The six hadronic event-shape distributions (thrust $\tau = 1 - T$, heavy jet mass ρ_H, wide and total jet broadening B_W and B_T, the C-parameter, and the Durham two-to-three-jet resolution $-\log_{10} y_{23}$) are corrected to particle level with a validated iterative Bayesian unfolding chain and confronted with an NNLO+NLL fit machinery. The final combined value is $\alpha_s(M_Z) = 0.1064 \pm 0.0198$ **(NNLO+NLL, built but not validated)**, shown with the three-way comparison: full data 0.1064 ± 0.0198 | 10% subsample 0.1103 ± 0.0159 (-0.25σ) | MC expectation 0.1019 ± 0.0165 $(+0.27\sigma)$. The full-data value sits **between** the 10% and the MC expectation and is consistent with **both** within the theory-dominated error, validating the staged-unblinding protocol; the non-perturbative parameters keep their sign, and no pre-registered stop trigger fires.

The reliability of the headline is bounded by the NLL limitation, and the note states this plainly where the headline first appears: the inverse-variance average is C**-dominant** (weight 0.434), but C is **poorly described** at full statistics (theory-band goodness-of-fit $\chi^2/\text{ndf} = 35.16$; the raw experimental fit $\chi^2/\text{ndf} \sim 4.5 \times 10^3$), so the per-observable central values — and hence the headline — are the values the unvalidated fit prefers, not well-measured couplings. The more trustworthy handle is the well-described-only subset, the only two channels with $\chi^2/\text{ndf} < 3$ at full statistics (τ and ρ_H), whose positive-weight inverse-variance average is $\alpha_s(M_Z) = 0.1059 \pm 0.0189$ (correlation-consistent error, the same $\rho_{\text{theory}} = 0.95$ method as the headline) — consistent with the all-six headline in **both** central value ($0.1064 \to 0.1059$, a < 0.0005 shift, so the headline is robust to the NLL-mismodeled channels) **and** uncertainty ($0.0198 \to 0.0189$, essentially unchanged, because the correlated theory floor — not the channel count — sets the error: the well-described subset buys trust, not precision). The combined error **grew** from 0.0159 (10% subsample) to 0.0198 (full data); this is the theory-limited (~0.02, NNLO+NLL) floor, not a degradation with more data — the budget is theory-floor-dominated (the correlated order truncation, the data-evaluated model dependence, and the scale, all correlated at $\rho_{\text{theory}} = 0.95$), statistics are subdominant throughout, and the lower 10%-subsample 0.0159 was a lower realization of the same floor from the 10%-evaluated systematics. The wide broadening B_W shifted $+1.0\sigma$ across the staging ($0.0766 \to 0.0800 \to 0.1556$) within its enormous total error 0.0874; it is the documented worst-described NNLL→NLL broadening (over-covered, theory-band $\chi^2/\text{ndf} = 0.03$, a flat χ^2 surface), **not** quotable as an α_s, and does **not** move the headline (1.2% weight). The fit-triviality gate passes (minimum fit $\chi^2 = 20.9$, not zero; non-circular shape fit).

The combined value sits -0.59σ / -9.8% below the PDG 2024 world average 0.1180 ± 0.0009; with a theory-dominated total the measurement distinguishes α_s differences of only 0.0395 (37.1% of the combined value) at 2σ, so this compatibility reflects the limited NNLO+NLL resolving power, not a near-PDG agreement. The per-subperiod (P1/P2/P3) full-data α_s is stable (no coherent drift), the published-spectrum overlay agrees with ALEPH-2004 within the full-data errors under the [A3] shared-data caveat, and the data-driven model-dependence term reproduces the 10%-subsample value (ratio 0.85–1.33). **The final ALEPH $\alpha_s = 0.1064 \pm 0.0198$ is theory-dominated, -9.8% below PDG — a consistency and tension-decomposition result at NNLO+NLL (not validated); the**

well-described (τ+ρ_H) subset is the more trustworthy handle; N3LL / NNLL and a second generator (HERWIG 7) are the route to a competitive measurement.

1 Introduction

1.1 Motivation: event-shape alpha-s at the Z pole

The reaction $e^+e^- \to Z \to$ hadrons at the Z pole is the theoretically cleanest environment for determining the strong coupling constant α_s. Every event shares a fixed, precisely known hard scale $Q = \sqrt{s} = M_Z = 91.1876 \pm 0.0021$ GeV (Navas et al. 2024), so there is no parton-distribution uncertainty, no initial-state hadronic structure, no underlying event, and at LEP1 negligible pile-up. The only QCD dynamics is final-state gluon radiation and subsequent hadronization. Event shapes quantify the departure of the hadronic final state from the back-to-back two-jet configuration; at leading order that departure is the emission of a single hard gluon, directly proportional to α_s. The full shape of each distribution therefore constrains α_s through both rate and shape information. This made LEP event shapes one of the historically most important inputs to the world-average $\alpha_s(M_Z)$ (Navas et al. 2024), and the original ALEPH analyses (Decamp et al. 1991; Barate et al. 1998; Heister et al. 2004) are foundational measurements of these observables.

The strong coupling extracted from e^+e^- event shapes nonetheless exhibits a persistent and well-documented tension. With the modern soft-collinear effective theory (SCET) machinery and a field-theory power correction, the thrust distribution yields $\alpha_s(M_Z) = 0.1135 \pm 0.0011$ (Abbate et al. 2011) and the C-parameter 0.1123 ± 0.0015 (Hoang et al. 2015b), both well below the world average 0.1180 ± 0.0009 (Navas et al. 2024), whereas the heavy jet mass gives 0.1220 ± 0.0031 (Chien and Schwartz 2010) and a tuned-Monte-Carlo hadronization treatment of the ALEPH data gives 0.1224 ± 0.0039 (Dissertori et al. 2009). The spread is not primarily a theory-order effect: it tracks the sensitivity of each observable to the leading $1/Q$ non-perturbative power correction, which is strongly degenerate with α_s in a single-energy fit. Indeed, two SCET thrust fits on the same observable with comparable machinery disagree by ~0.004 (Abbate et al. 2011; Becher and Schwartz 2008), demonstrating that the per-observable spread conflates the observable's renormalon structure with perturbative-treatment choices (profile scales, fit range, renormalon scheme).

1.2 Cross-experiment context and prior measurements

The same six observables have been measured at the Z pole by every LEP experiment. The OPAL measurement of event-shape distributions and moments (Abbiendi et al. 2005) provides genuinely independent data at the same energy ($\alpha_s(M_Z) = 0.1191 \pm 0.0046$ at NNLO+NLLA), and the DELPHI energy-evolution study (Abdallah et al. 2003) gives explicit particle-level definitions of all six observables together with the dispersive power correction and observable-by-observable hadron-mass corrections that the present analysis adopts. The archived ALEPH LEP1 data used here have recently been re-analysed with modern unbinned methods: an unbinned thrust measurement using OmniFold (primary) with IBU as the binned cross-check (Badea et al. 2025), and jet-substructure and energy-correlator studies (Badea et al. 2019). These contemporary analyses of the exact same skim are the gold-standard reference for how the dataset is corrected — the particle-level definition, the energy-flow object treatment, and the detector-systematic variations are taken from them. Foundational theory context is provided by the global jet-rate (Verbytskyi et al. 2019) and energy-energy-correlator (Kardos et al. 2018) fits, which use the same NNLO theory but observables with weaker $1/Q$ sensitivity and land near the world average — the cleanest statement that the tension tracks $1/Q$ structure, not theory order.

1.3 What this measurement adds, and what it is not

This is a single experiment at a single energy. It cannot produce a *more* precise α_s than the existing ALEPH-based global fits, because those fits already include the very ALEPH LEP1 distributions re-measured here, and because the uncertainty on any single-dataset event-shape determination is dominated by the common perturbative-truncation and power-correction treatment. The honest, still-novel deliverable is threefold: a consistent, correlated six-observable ALEPH $\alpha_s(M_Z)$ unfolded on one dataset with one particle-level definition and the full cross-observable covariance; a measured (not assumed) decomposition of the perturbative and power-correction uncertainty; and a direct test of where the inter-observable tension lives. Critically, the perturbative prediction is realized at NNLO+NLL (built, not validated against absolute α_s closure — see Section 4 and Section 4.6), so the perturbative order is the dominant uncertainty and is reported as such. The N3LL upgrade (for τ, C, ρ_H) and the NNLL upgrade (for the broadenings and y_{23}) are documented future work. This is therefore a **consistency and tension-decomposition result at LEP-era theory precision**, not a world-leading-precision claim.

1.4 Staged-unblinding protocol and terminology

The result is reported under a staged-unblinding protocol whose purpose is to validate the correction and fit machinery before the full data are seen. Three input stages were used in sequence: an **MC-expectation** stage, in which the entire chain was exercised on 1994 ALEPH full-simulation Monte Carlo used as pseudo-data, with the real data blinded; a **10%-subsample** stage, the first to use real ALEPH 1994 data through a fixed-seed event-level uniform-random 10% subsample, which validated the data and the correction chain; and the final full-data measurement on the complete 1994 P1+P2+P3 real ALEPH data — all 1,293,167 selected events (Section 2). The transition to the full data was made only after explicit human approval, the authorized point under the blinding protocol; a positive full-data assertion ($n_{\rm sel} = n_{\rm total} = 1{,}293{,}167$) records that 100% of the data is used and that no masked or partial array can silently enter.

Unblinding the data does **not** lift the **theory** limitation: the perturbative prediction remains NNLO+NLL (built, not validated; Section 4.6), and the central coupling is reported as the value the unvalidated fit prefers. The full-data result is compared against **both** the 10%-subsample result and the MC expectation, and the three-way consistency (Section 6) is the headline validation of the protocol. The poor goodness-of-fit seen at NLL for C, B_W, B_T, and $-\log_{10} y_{23}$ in the earlier stages is **expected to reproduce — and to be more pronounced** — at full statistics: as the per-bin statistical error shrinks, the NLL data−theory mismatch dominates the χ^2, exactly the NLL signature behaving identically on data and MC, not a pathology. Throughout this note, "the data" or "the full data" means the complete 1994 real ALEPH sample; "the 10% subsample" denotes the fixed-seed 10% real-data draw; and "the MC expectation" (or "the expected") denotes the full-simulation pseudo-data result.

2 Data samples

This analysis uses the publicly archived ALEPH LEP1 open-data sample and its matched 1994 full-simulation Monte Carlo, both accessed from the ALEPH open-data archive. The sample inventory, data-archaeology findings, and particle-level support are summarized here in the structured tables required for reproduction. The archived data files are a pre-selected hadronic-event skim ("recons_aftercut"): a hadronic-Z selection was applied at ntuple production, so the files contain predominantly the peak-region hadronic sample. Because the measurement reports normalized distributions (Section 3.1), an observable-independent skim efficiency cancels in the shape and is not a normalization systematic; any observable-dependence is captured by the data/MC input-validation gate. The final measurement uses the **full** 1994 selected sample (all 1,293,167 events), the same selected array the 10% subsample was drawn from, here used **without any mask**.

2.1 Data summary table

The six merged data files span the 1992–1995 LEP1 running. The per-event beam-energy setpoint cleanly separates the peak-only years (1992, 1994) from the energy-scan years (1993, 1995, which carry off-peak satellites at $M_Z \pm 1.8$ GeV). The primary measurement uses the 1994 running (periods P1+P2+P3), the only year with matched full-simulation MC and the highest single-year peak statistics. The integrated luminosity is not published for this archived skim; it is estimated from the selected hadronic event count and the Z-pole hadronic cross section $\sigma_{\rm had} \approx 30.4$ nb (Navas et al. 2024; Schael et al. 2006) as $\mathcal{L} = N_{\rm had}/\sigma_{\rm had}$, giving $\mathcal{L}_{1994} \approx 1.29\times10^6/30.4\,{\rm nb} \approx 42\ {\rm pb}^{-1}$, consistent with the ∼40 pb^{-1} quoted for this skim (Badea et al. 2025). This estimated luminosity enters no fit — the measurement is of normalized shapes (Section 3.1), so the skim efficiency cancels and the luminosity does not propagate to α_s — so the back-calculation is **not circular**; it is reported only to characterize the sample size. The same luminosity-free property is what makes the full-data fit non-circular (Section 6.3): the shape fit has no luminosity input.

The primary 1994 sample after the peak window and the stored hadronic selection contains **1,293,167 events**, consistent with the ∼1.36 million hadronic events quoted for this skim before the explicit ±0.5 GeV peak window (Badea et al. 2025, 2019). The 1992 and the 1993/1995 peak subsets are retained as optional cross-checks; they lack matched full-simulation MC and so would carry a year-stability extrapolation uncertainty, and they are not part of the primary measurement.

Table 1: Data summary by period. The peak window is $|E - M_Z| < 0.5$ GeV. The luminosity is estimated from $\mathcal{L} = N_{\text{had}}/\sigma_{\text{had}}$ (not published for the archived skim) and is stated to two significant figures. The 1994 total is the binding primary sample; the final full-data measurement uses all 1,293,167 events. The ~42 pb [1] is the sum of the three 1994 periods (P1+P2+P3 only) and excludes the 1992 cross-check row (~17 pb [1]), which is not part of the primary measurement.

Period	$\sqrt{s}$ [GeV]	Structure	Events (peak+passesAll)	$\mathcal{L}$ [pb^{-1}]
1992	91.27	peak only	522,526	~17
1994 P1	91.2	peak only	411,001	~13
1994 P2	91.2	peak only	424,139	~14
1994 P3	91.2	peak only	458,027	~15
1993	89.4/91.2/93.0	scan (peak subset)	354,499	~12
1995	89.4/91.3/93.0	scan (peak subset)	404,655	~13
1994 total (primary)	91.2	peak only	**1,293,167**	**~42**

2.2 Monte Carlo sample table

The matched MC is a single 1994 on-peak full-simulation sample (40 files), a JETSET/PYTHIA-class parton-shower plus string-fragmentation generator passed through the GALEPH/GEANT ALEPH detector simulation. The exact generator version is not recoverable from the ntuple metadata (the same-skim references quote it differently); this does not affect the analysis, because the hadronization-model uncertainty is assessed with independent standalone generators (below). The MC provides the reconstruction-to-particle response for unfolding and the closure, stress, and prior samples.

Table 2: MC sample summary. The single full-simulation generator provides the detector response; the hadronization-model uncertainty is assessed with a standalone PYTHIA 8 Monash sample by reweighting the response prior. The exact full-sim generator version is not in the ntuple metadata and is documented as an open item.

Process	Generator	$\sqrt{s}$ [GeV]	N_{gen}	Role	Notes
$Z \to q\bar{q}$ (full sim)	JETSET/PYTHIA + GALEPH/GEANT	91.20	771,597 (`tgen`)	response, closure	reco = 731,006
$Z \to q\bar{q}$ (pre-sel truth)	same	91.20	973,769 (`tgenBefore`)	eff. denom.	sel. eff = 0.751
$Z \to q\bar{q}$ standalone	PYTHIA 8 Monash (Tune:ee=7)	91.2	60,000	hadronization syst	particle+parton

The MC carries ~57% of the 1994 data statistics (731,006 matched reconstructed events vs 1,293,167 data events), adequate for the response matrix. Because the measurement reports normalized shapes, the MC-to-data statistical covariance is scaled to the **full-data** precision by $N_{\text{MC}}/N_{\text{data}} = 731{,}006/1{,}293{,}167 = 0.565$ (variance $\propto 1/N$); this is **the same scale the MC-expectation study used** (that Asimov already carried the full-data statistical precision), so the full-data statistical covariance equals the expectation covariance by construction, and the only change is the measured `data_density` (full real-data unfolded vs MC Asimov). A standalone PYTHIA 8 Monash sample (60k events) was generated for the hadronization-model systematic; a second standalone generator (HERWIG 7) was not generated in this session — a documented single-tune limitation (Section 5.10), and the limitation whose physical manifestation the data directly reveal at full statistics (Section 5.1). The event-selection efficiency, computed as matched reconstructed events over pre-selection truth (731,006/973,769 = 0.751), is a global scale that cancels in the normalized shape and is recorded separately.

2.3 Backgrounds and contamination

The stored ALEPH hadronic selection is more than 99% pure (Barate et al. 1998), so backgrounds are small. The reducible physics backgrounds — $\tau^+\tau^-$ pairs (suppressed by track multiplicity and visible-energy cuts), two-photon $\gamma\gamma \to$ hadrons (a LEP2-energy concern, negligible at the Z pole), lepton pairs, and $q\bar{q}\gamma$ initial-state-radiation

radiative returns — are each at the $O(0.1$–$1\%)$ level before cuts and sub-permille after, removed by the visible-energy, momentum-balance, and acceptance components of the stored hadronic flag. The dominant non-signal effect is residual ISR feed-in, which shifts the effective $\sqrt{s}$ rather than constituting a true background; it is handled by the MC-based correction and is sub-permille post-cut at the peak. The b- and c-quark content of the signal ($R_b = 0.21629 \pm 0.00066$, $R_c = 0.1721 \pm 0.0030$ (Schael et al. 2006; Navas et al. 2024)) is a physics property, not a background: the semileptonic-decay neutrinos it produces are a particle-level distortion captured by the full-simulation response and treated as part of the quark-mass systematic (Section 5.7).

3 Event selection and observable definitions

This section defines the six event-shape observables, the hadronic-Z event selection, and the data/MC input validation that gates the correction chain. The selection is cut-based; a multivariate event classifier was not used because the stored hadronic selection is already more than 99% pure, so an MVA would add negligible purity while introducing a data/MC-shape-dependent selection that would itself have to be unfolded (Barate et al. 1998). The full-data measurement applies **no separate selection** from the subsample studies: the peak window $|E - M_Z| < 0.5$ GeV and the stored `passesAll` flag [Dsel] are the same, and the full sample is used without any 10% mask (Section 2).

3.1 Observable definitions

All six observables are computed from the set of final-state particles of the hadronic event using the ALEPH all-particle energy-flow definition (charged tracks plus neutral calorimeter objects; neutrinos and other invisibles excluded). At particle (truth) level a particle is treated as stable if its proper lifetime satisfies $c\tau > 1$ cm (the standard LEP convention: K^0_S, Λ, K^0_L, and neutrons are kept undecayed), matching the convention of the modern theory predictions and the reference analyses (Abdallah et al. 2003; Dissertori et al. 2009). Each distribution is reported as the normalized differential cross section $(1/\sigma)\,\mathrm{d}\sigma/\mathrm{d}X$, integrated to unity over the measured range, with the normalization applied **after** unfolding and efficiency correction. The all-particle definition is the energy-flow definition for which the NNLO/N3LL predictions and the dispersive power correction are formulated, so it compares directly to theory; a charged-only construction is retained as an independent cross-check observable, not a systematic.

The thrust is

$$T = \max_{\hat{n}} \frac{\sum_i |\vec{p}_i \cdot \hat{n}|}{\sum_i |\vec{p}_i|}, \qquad \tau = 1 - T, \tag{1}$$

maximized over the unit vector $\hat{n}$; the maximizing $\hat{n}_T$ is the thrust axis (Farhi 1977), which also splits the event into two hemispheres. The heavy jet mass is the heavier of the two hemisphere mass-squared fractions (Clavelli 1979),

$$\rho_H = \frac{\max(M_1^2, M_2^2)}{E_{\mathrm{vis}}^2}, \qquad M_k^2 = \Big(\sum_{i \in H_k} p_i \Big)^2. \tag{2}$$

The two jet broadenings are built from the per-hemisphere transverse momentum relative to the thrust axis (Catani et al. 1992),

$$B_k = \frac{\sum_{i \in H_k} |\vec{p}_i \times \hat{n}_T|}{2 \sum_i |\vec{p}_i|}, \qquad B_T = B_1 + B_2, \quad B_W = \max(B_1, B_2), \tag{3}$$

which obey the ordering $B_W \le B_T \le 2B_W$. The C-parameter is built from the linearized momentum tensor $\Theta^{\alpha\beta} = (\sum_i |\vec{p}_i|)^{-1} \sum_i p_i^\alpha p_i^\beta / |\vec{p}_i|$ with eigenvalues $\lambda_{1,2,3}$ (Parisi 1978; Ellis et al. 1981),

$$C = 3(\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1), \tag{4}$$

which has a kinematic Sudakov shoulder at $C = 3/4$ (the maximum reachable by three massless partons): the fixed-order prediction is non-smooth there, so the C fit window stays below $C = 0.7$ (Hoang et al. 2015b; Luisoni et al. 2021). Finally, the Durham two-to-three-jet resolution clusters the particles with the k_T measure $y_{ij} = 2\min(E_i^2, E_j^2)(1 - \cos\theta_{ij})/E_{\mathrm{vis}}^2$ (Catani et al. 1991), and y_{23} is the value of y_{cut} at the three-to-two merging; we report $-\log_{10} y_{23}$. All six are infrared-and-collinear safe.

The six observables were prototyped from the all-particle energy-flow objects and cross-validated against ALEPH's own stored quantities: the computed thrust matches the stored `Thrust` branch to a mean absolute difference of 10^{-5} (correlation 1.00000) and the computed C-parameter matches the stored `C_linearized` to 5×10^{-5}, while the Durham y_{23} matches a `fastjet` e^+e^- k_T reference. Every per-event sanity bound ($\tau \in [0, 1/2]$, $B_W \le B_T \le 2B_W$, $C \in [0, 1]$) holds on every event.

3.2 Event selection and cutflow

The analysis selection applies a peak window $|E - M_Z| < 0.5$ GeV (with $M_Z = 91.188$ GeV (Navas et al. 2024)) on top of the pre-applied skim, and then the stored `passesAll` flag — the AND of the eight ALEPH hadronic selection components (sphericity-axis acceptance, minimum charged multiplicity, minimum charged energy, neutral/charged consistency, ISR rejection, four-fermion rejection, missing-momentum, and two-prong rejection). The cutflow for the summed 1994 data shows the sequential component flow is monotonically non-increasing and the acceptance loss (~6%) is dominated by the sphericity-axis acceptance (~2.3%) and the missing-momentum cut (~2.3%).

Table 3: Hadronic-Z selection cutflow for the three 1994 periods, from stored per-event flags. The component AND is monotone by construction; the stored `passesAll` count sits ~0.5% above the strict AND-of-8 (ALEPH internal logic admits a small fraction failing the strict sequential AND) and is the authoritative analysis count. The 1994 total after stored `passesAll` is 1,293,167 events.

Stage	1994 P1	1994 P2	1994 P3
all ntuple entries	433,947	447,844	483,649
+ peak window	433,947	447,844	483,649
+ sphericity acceptance	423,933	437,435	472,499
+ min charged multiplicity	423,871	437,371	472,426
+ min charged energy	423,832	437,325	472,380
+ neutral/charged consistency	421,534	434,975	469,842
+ ISR rejection	418,663	432,073	466,683
+ four-fermion rejection	418,646	432,052	466,663
+ missing-momentum	409,072	422,171	455,889
+ two-prong rejection	409,072	422,171	455,889
= AND-of-8 components	409,072	422,171	455,889
stored passesAll & peak (analysis selection)	**411,001**	**424,139**	**458,027**

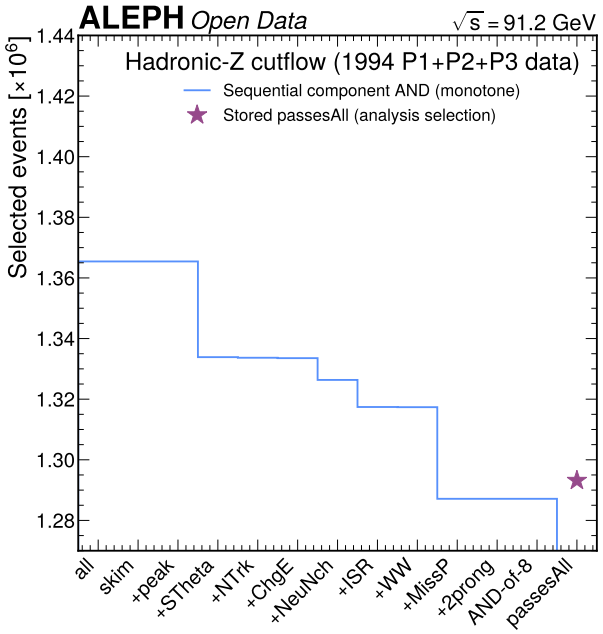

Figure 1: Hadronic-Z selection cutflow for the summed 1994 P1+P2+P3 data, from the stored selection flags, on a linear scale. The blue step shows the sequential component AND, which is monotonically non-increasing by construction and ends at the strict AND of the eight components. The flat skim and peak rows reflect the pre-skimmed peak-region archive, and the ~6% acceptance loss is dominated by the sphericity acceptance and missing-momentum cuts. The red star is the separate stored passesAll count, shown at its own position because it is not the strict AND of the eight components.

3.3 Energy-flow object validation

The all-particle definition combines charged tracks (energy-flow class "charged hadron", "muon", "electron") with neutral calorimeter objects ("photon", "neutral hadron"). The neutral objects carry ~35% of the visible energy and have finite four-momenta in data, reconstructed MC, and truth MC, confirming the all-particle definition is fully supportable on this dataset (Buskulic et al. 1995). The all-particle multiplicity is ~29.4 in data and ~29.7 in MC (~18.8 charged plus ~10.6 neutral), and the visible energy is ~82.9 GeV (data) versus ~83.9 GeV (reconstructed MC) and ~89.4 GeV (truth) — the ~1.8 GeV truth-level deficit relative to M_Z is the excluded neutrinos, consistent with the b/c flavour content. The reconstructed-level energy-flow inputs are compared to the full-simulation MC for every variable entering the observable calculation. The directly-fit thrust agrees within ±3%, while the visible energy and neutral multiplicity show a coherent ±10–25% slope (data harder in visible energy, softer in neutral multiplicity than the PYTHIA-class MC) — the worst-modelled input, which motivates the prior reweighting in the correction chain. This documented reconstruction-level data/MC slope is the same one that, propagated through a near-diagonal response, becomes the particle-level data−MC difference measured on the full data (Section 5.1) — it is anticipated here, at full statistics, from the reconstruction-level comparison.

Figure 2 shows the area-normalized data/MC comparisons of the six input variables (1994 data vs 1994 full-simulation MC). The reduced χ^2/ndf are large — 11.7 (charged multiplicity), 40.9 (neutral multiplicity), 41.0 (visible energy), 43.9 (object momentum), and 43.8 (object energy) — because the ~57k-event-equivalent MC statistics resolve genuine, documented detector-model imperfections, not a selection or binning problem; the data/MC ratio sits within ±10–25% across the bulk. The only well-modelled input is the event-orientation variable $|\cos\theta_T|$ (χ^2/ndf = 0.81). The visible-energy/neutral slope is precisely the imperfection the unfolding removes, and its residual after correction is bounded by the prior-dependence test (Section 4.4). To address it, a data-driven reco-level reweight of the MC visible energy to data is derived (maximum deviation from unity 0.30, clipped to [0.5, 2.0]) and carried both as the nominal prior reweighting and as a detector systematic.

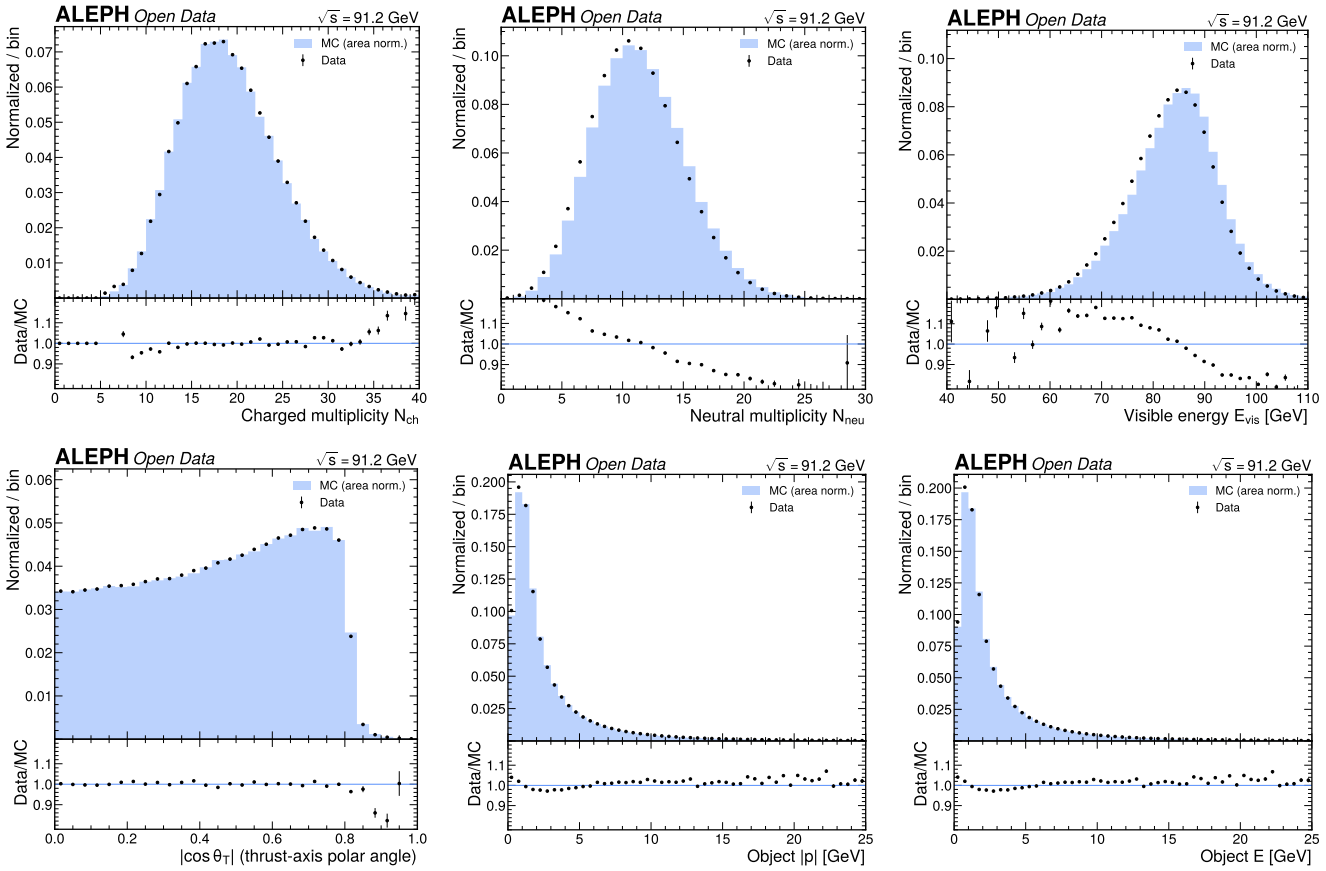

Figure 2: Area-normalized data/MC comparison of the six energy-flow input variables (1994 full data vs 1994 full-simulation MC), each with a data/MC ratio panel: (a) charged multiplicity, (b) neutral multiplicity, (c) visible energy, (d) thrust-axis polar angle $|\cos\theta_T|$, (e) object momentum, (f) object energy. The reduced χ^2/ndf are large because the MC statistics resolve genuine, documented detector-model imperfections; the data/MC ratio sits within ±10–25% across the bulk and only $|\cos\theta_T|$ is well modelled. The coherent visible-energy/neutral slope is the imperfection the unfolding removes.

The six observables at reconstruction level show the same detector-model imperfection propagated into their shapes, with reduced χ^2/ndf ranging from 17.3 (τ) and 21.7 (C) — the least discrepant — to 67.5 (B_W) and 111.3 ($-\log_{10} y_{23}$). These are pre-correction input-validation comparisons, not post-correction agreement claims; the same pattern holds at full statistics on the full data.

Figure 4 presents the six reconstructed-level observables (area-normalized, with ratio panels) at full 1994 statistics. The input-validation gate is satisfied in the sense required of an unfolded measurement: the comparisons are produced, the discrepancies are documented, and their post-correction impact is bounded (Section 4.4). The large reduced χ^2/ndf is the expected, statistically well-resolved pre-correction discrepancy that the unfolding is designed to remove.

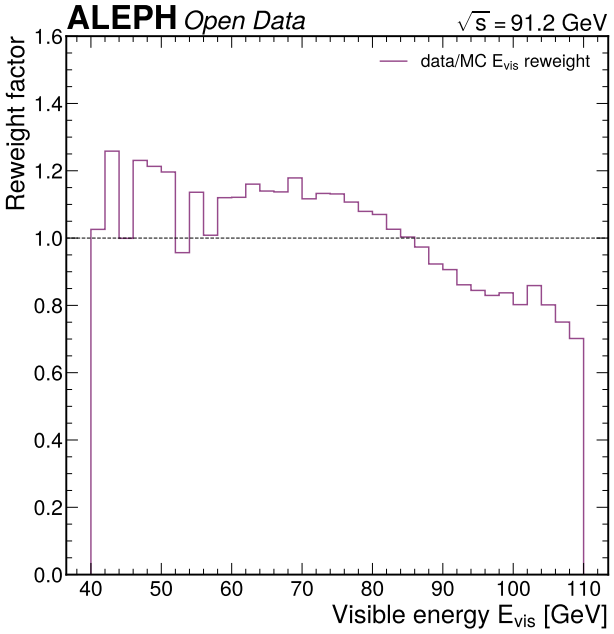

Figure 3: Data-driven reco-level visible-energy reweight (data/MC ratio per E_{vis} bin, clipped to the interval 0.5 to 2.0), with maximum deviation from unity of 0.30. The dashed line is unity. This reweight captures the coherent visible-energy/neutral slope seen in the input validation and is carried as the nominal prior reweighting of the response and as the dominant detector-side shape systematic.

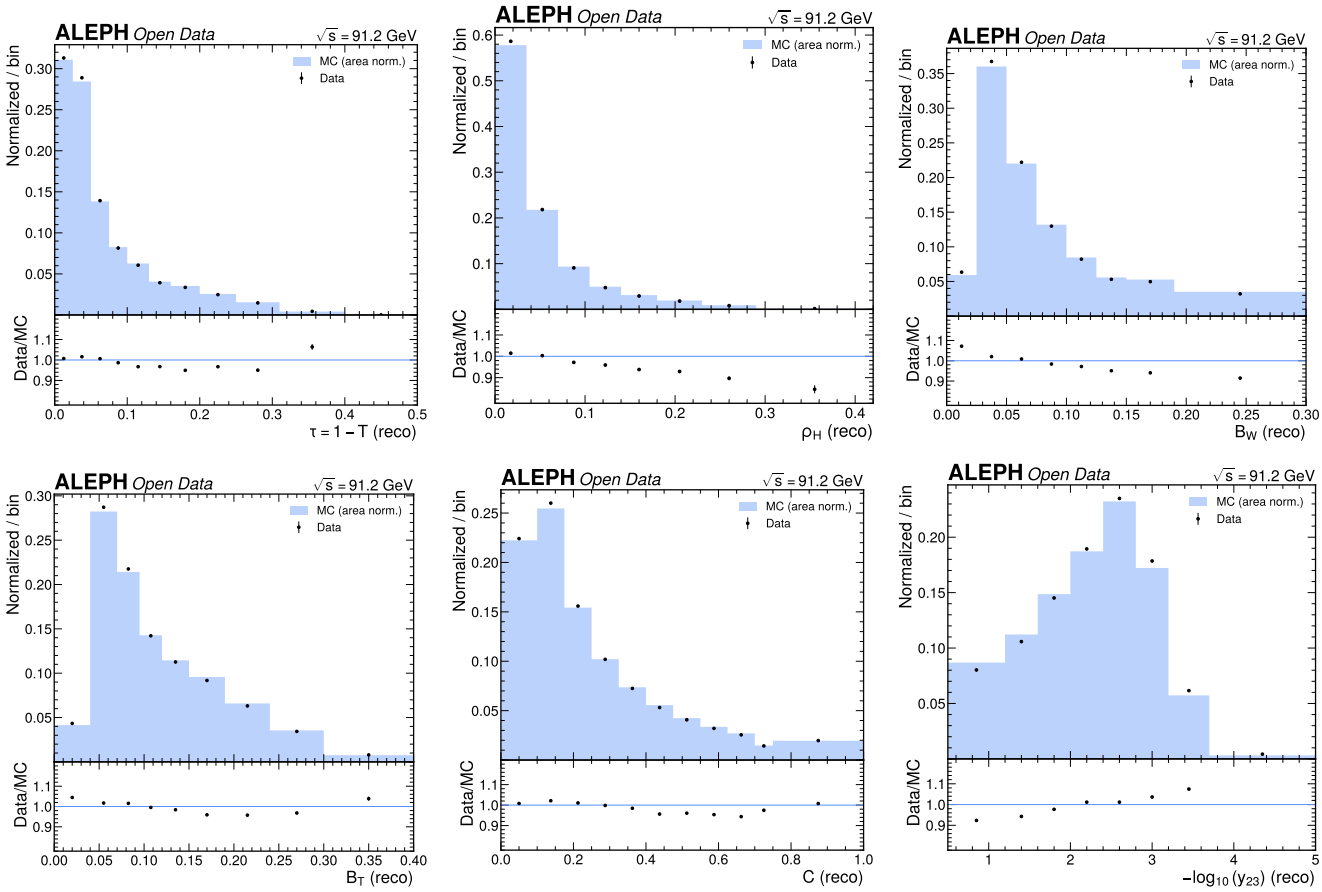

Figure 4: Area-normalized reconstructed-level data/MC comparison of the six observables at full 1994 statistics, each with a data/MC ratio panel: (a) thrust τ, (b) heavy jet mass ρ_H, (c) wide broadening B_W, (d) total broadening B_T, (e) C-parameter, (f) Durham $-\log_{10} y_{23}$. These are pre-correction input-validation comparisons (not post-correction agreement claims); the large reduced χ^2/ndf is the expected, statistically well-resolved discrepancy the unfolding removes.

4 Corrections and unfolding

The measurement delivers particle-level normalized distributions corrected for detector acceptance, efficiency, and bin migration, which are then confronted with a perturbative-plus-non-perturbative QCD prediction to extract $\alpha_s(M_Z)$. This section describes the unfolding procedure, the response matrix, the regularization choice, the validation tests, the theory prediction, the non-perturbative sector, and the fit. The same validated chain is applied identically to the MC expectation, the 10% subsample, and the full 1994 real data — only the measured input changes. The key equations are displayed so that a reader can implement the method without the code.

4.1 Response matrix and the unfolding equation

Because event shapes are per-event global quantities, the reconstruction-to-particle matching is trivial and unambiguous: the same MC event provides one reconstructed value and one particle-level value of each observable, and the response matrix is the two-dimensional histogram of (reco, particle) over selected MC events. This avoids the variable-multiplicity sub-object matching that fails for fragmentation observables. The migration matrix $M[g, r]$ is the count of events with particle-level bin g and reconstructed bin r; the conditional response is the column-normalized

$$R[r, g] = \frac{M[g, r]}{\sum_{r'} M[g, r']}, \tag{5}$$

so that reco $= R \cdot$ gen for the matched part. The within-fiducial efficiency eff[g] = (matched gen[g])/(all gen[g]) and purity pur[r] = (matched reco[r])/(all reco[r]) account for misses and fakes. The matched efficiency and purity are ~1.0 for all six observables in the all-particle construction. The event-selection efficiency (0.751) is recorded separately and cancels in the normalized shape. The response matrices are strongly diagonal-dominant, with diagonal fractions of 0.65 (τ), 0.74 (ρ_H), 0.73 (B_W), 0.73 (B_T), 0.59 (C), and 0.58 ($-\log_{10} y_{23}$) at full statistics — all comfortably above the 50% gate. This near-diagonal response is also why the documented reconstruction-level data/MC slope passes through to particle level essentially unchanged (Section 5.1): the unfolding has little smearing to undo, so a genuine data−MC shape difference survives.

Figure 5 shows the conditional response R[reco|gen] for each observable (all-particle construction, reconstructed versus particle level). All six are visibly diagonal-dominant, confirming that the detector resolution is well below the bin width and that bin-by-bin migration is modest — the physical reason event shapes unfold cleanly with a small number of iterations.

4.2 Iterative Bayesian unfolding and regularization

The primary correction is iterative Bayesian (D'Agostini) unfolding (D'Agostini 1995). The measured reconstructed spectrum is first fake-corrected ($n_{\mathrm{corr}}[r] = n_{\mathrm{meas}}[r] \cdot \mathrm{pur}[r]$), the full smearing matrix $S[r, g] = \mathrm{eff}[g]\, R[r, g]$ is assembled, and the Bayes inversion is iterated from the MC particle-level prior with the prior reset each iteration and the efficiency un-smeared at the end. The unfolded estimate at iteration $n + 1$ is, bin-by-bin,

$$\hat{n}_g^{(n+1)} = \frac{1}{\mathrm{eff}[g]} \sum_r \frac{S[r, g]\, \hat{n}_g^{(n)}}{\sum_{g'} S[r, g']\, \hat{n}_{g'}^{(n)}}\, n_{\mathrm{corr}}[r], \tag{6}$$

after which the spectrum is normalized to unit area (Equation 7). The number of iterations is the regularization, and it is fixed **before** touching data by a pre-registered rule, never tuned to data/MC agreement: the response is built on a random training half of the MC and applied to the held-out test half, and the iteration is chosen as the one that minimizes the honest split-sample χ^2/ndf within the alarm-safe band $(0.1, 3]$. Because the response is so diagonal-dominant, the split-sample χ^2/ndf is already smallest at a single iteration for five of the six observables (further iterations only amplify statistical noise); $-\log_{10} y_{23}$ has a genuine interior minimum at three iterations. The selected counts are therefore $n_{\mathrm{iter}} = 1$ for τ, ρ_H, B_W, B_T, C, and $n_{\mathrm{iter}} = 3$ for $-\log_{10} y_{23}$. An identity-response self-test (response equal to the identity, efficiency and purity equal to one) returns the input unchanged, confirming the implementation is unbiased; this self-test was re-run on the full-data path and again passes exactly.

The normalization is applied after the correction,

$$\left.\frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}X}\right|_g = \frac{\hat{n}_g}{\Delta X_g \sum_{g'} \hat{n}_{g'}}, \tag{7}$$

with ΔX_g the bin width, so that the result is a probability density.

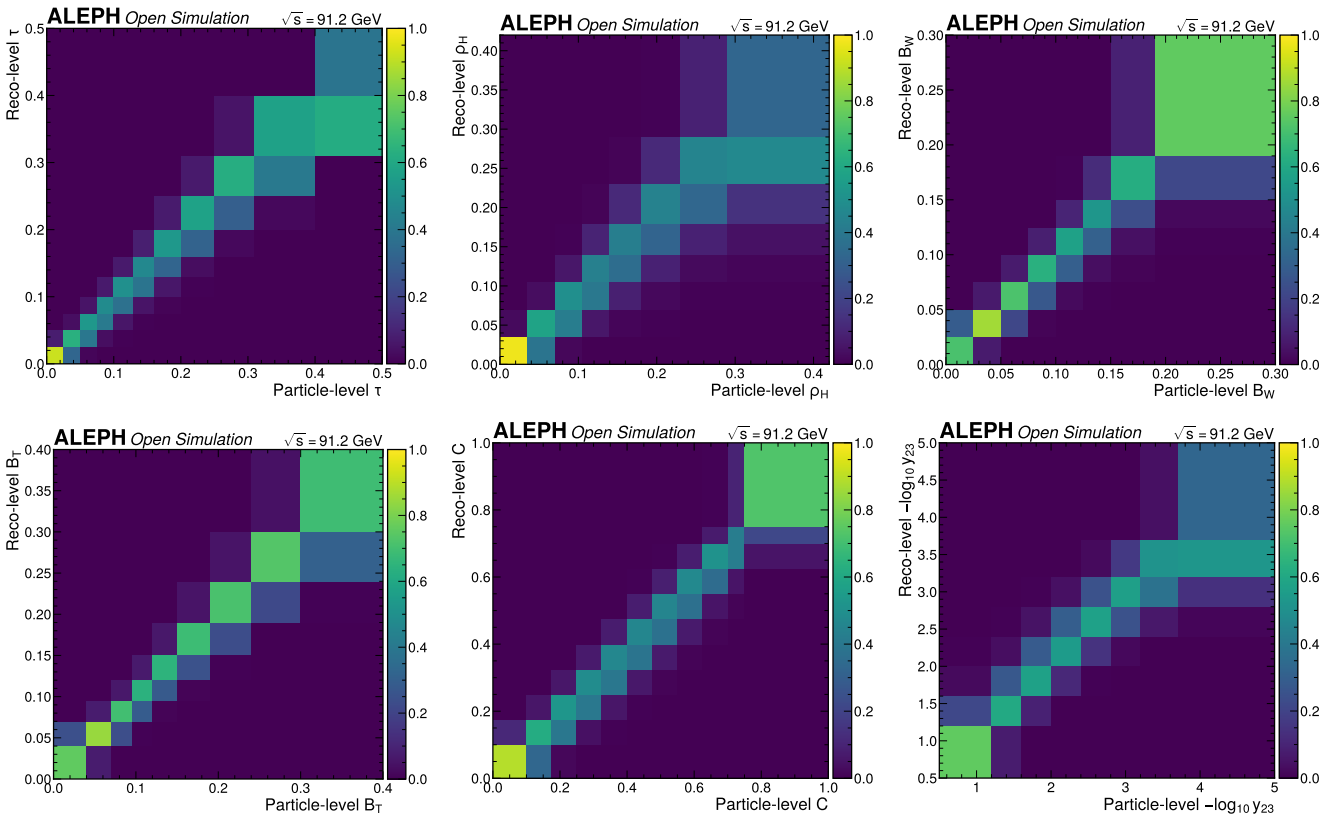

Figure 5: Conditional response $R[\text{reco}|\text{gen}]$ for each observable (all-particle construction, reconstructed versus particle level): (a) τ, (b) ρ_H, (c) B_W, (d) B_T, (e) C, (f) $-\log_{10} y_{23}$. All six are visibly diagonal-dominant, confirming the detector resolution is well below the bin width and bin-by-bin migration is modest — the physical reason event shapes unfold cleanly with few iterations.

4.3 Full-data construction

The MC-expectation (Asimov) target was the ALEPH-MC particle-level (all-particle) truth density on the fixed binning, recovered by running the identical IBU chain on the MC reconstructed-level pseudo-data (the unfolded-versus-truth residual was 10^{-16} to 10^{-4} per bin — the closure test of Section 4.4). For the final measurement the input is the **full real-data reco spectrum** (all 1,293,167 events): the same response, the same per-observable n_{iter}, the same normalization, and the same covariance structure are used — only the measured reconstructed input changes. The covariance is the authoritative toy-MC (per-observable) and event-level-bootstrap (joint 55×55) covariance, with the statistical variance scaled to the full-data precision by the ratio $N_{\text{MC}}/N_{\text{data}} = 731{,}006/1{,}293{,}167 = 0.565$ (variance scales as $1/N$). This is the **same scale the MC expectation used**, because that Asimov already carried the full-data statistical precision; therefore the full-data statistical covariance **equals** the expectation covariance by construction and the only change is the measured `data_density`. All covariances are positive-semi-definite; the per-observable and joint condition numbers reproduce the validated values exactly (τ 1.24×10^5, ρ_H 2.01×10^4, B_W 80.7, B_T 262.3, C 114.0, y_{23} 2.94×10^3; joint 7.46×10^5, well below the 10^8 cap).

4.4 Validation: closure, stress, prior dependence

Three validation tests anchor the correction chain, each on a statistically independent construction, and each with a figure showing the test result rather than a bare pass/fail. The closure test asks whether the method recovers a known truth from an independent sample; the response is built on a random training half of the MC and applied to the held-out test half (so the test is non-tautological), and Poisson toys make the χ^2 statistically meaningful. All six observables pass: the closure χ^2/ndf lies in $[0.38, 1.0]$ with $p > 0.05$ and the maximum pull is 1.81σ (Table 14). None is suspiciously good (the lowest, τ at 0.376, is well above the 0.1 alarm floor and reflects genuine sub-percent unfolded-versus-truth scatter). These tests bound the post-correction residual of the chain that is applied unchanged to the full data.

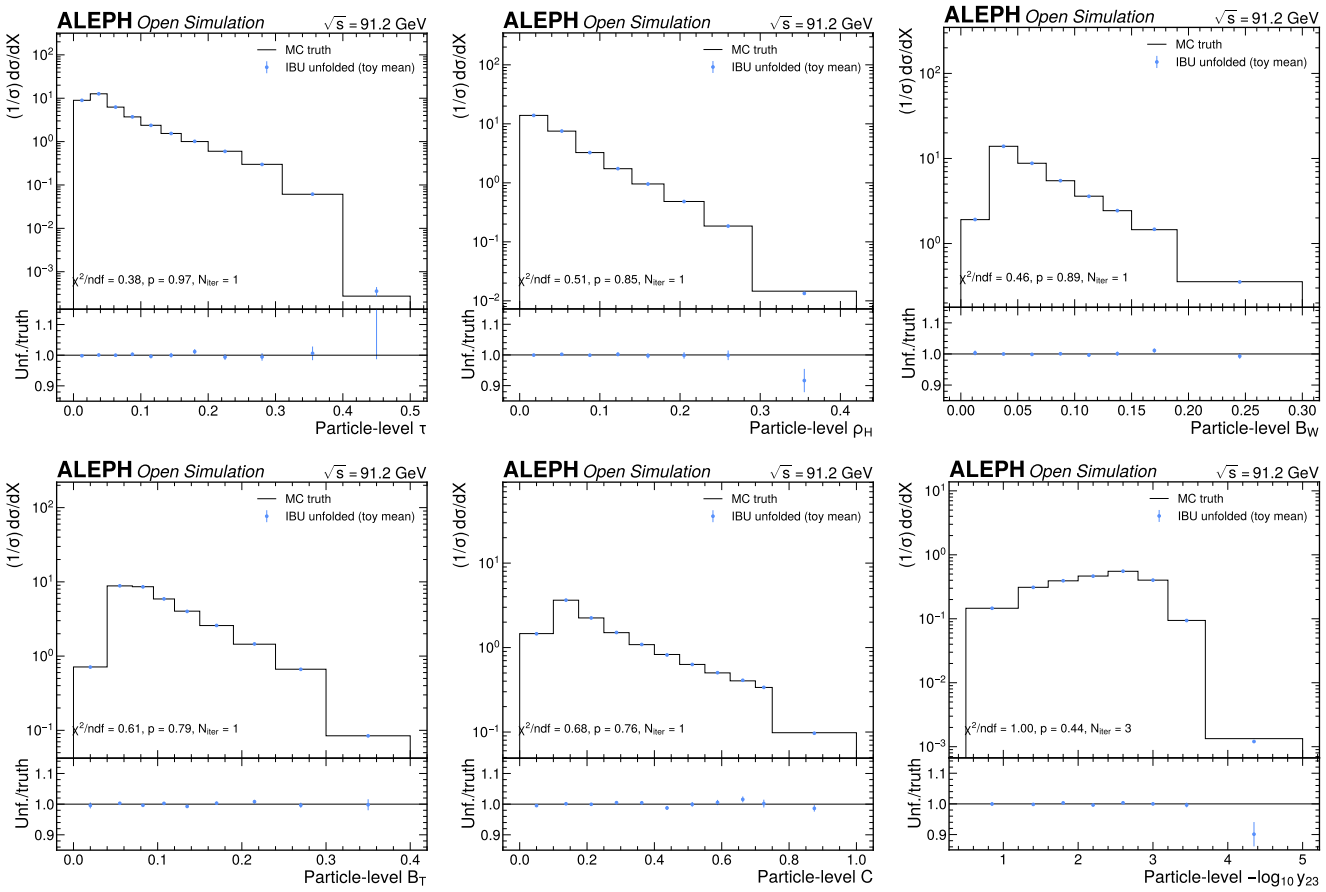

Figure 6: Split-sample IBU self-closure for each observable: the MC truth (black step) against the IBU unfolded toy mean (blue points), with the unfolded/truth ratio panel and the χ^2/ndf, p-value, and n_{iter} annotation: (a) τ, (b) ρ_H, (c) B_W, (d) B_T, (e) C, (f) $-\log_{10} y_{23}$. The response is built on one MC half and applied to the other; all six close with sub-percent residuals, confirming the unfolding is unbiased on an independent sample.

Figure 6 shows the split-sample IBU self-closure for each observable: the MC truth (black step) against the IBU unfolded toy mean (blue points), with the unfolded/truth ratio panel and the χ^2/ndf, p-value, and n_{iter} annotation. The response is built on one MC half and applied to the other; all six close with sub-percent residuals, confirming the unfolding is unbiased on an independent sample. Critically for the staged unblinding, this closure is an MC-recovers-MC test: it shows the chain does not *manufacture* a residual, so the coherent residual that appears on the **real** data (Section 6.2) is a genuine data−MC difference, not a chain artifact.

The stress test characterizes the method's resolving power. The MC truth is reweighted by a linear tilt of strength s in the observable,

$$w = 1 + s\,\frac{X - \langle X \rangle}{\sigma_X}, \qquad s \in \{5\%, 10\%, 20\%, 50\%\}, \tag{8}$$

forward-folded through the nominal response, Poisson-fluctuated, and unfolded with the **nominal** (untilted) prior; the recovered tilted truth is then compared to the input tilt. The mean relative bias of the recovered truth is far below the imposed tilt at every grade — at the largest 50% tilt the residual bias is at most 5.2% (ρ_H) and below 2.8% for the others (Table 4) — so the method resolves a 5% shape distortion with sub-percent bias, comfortably covering the few-percent data/MC differences this dataset carries (Section 6.2).

Table 4: Stress-test recovery: mean relative bias of the recovered truth versus the imposed linear-tilt strength (all-particle). The method resolves the smallest 5% tilt with sub-percent bias; the recovered bias is far below the imposed tilt at every grade.

Observable	$s = 5\%$	$s = 10\%$	$s = 20\%$	$s = 50\%$
τ	0.30%	0.58%	1.14%	2.77%
ρ_H	0.56%	1.10%	2.14%	5.15%
B_W	0.22%	0.43%	0.89%	2.56%
B_T	0.23%	0.46%	0.94%	2.62%
C	0.16%	0.32%	0.65%	1.84%
$-\log_{10} y_{23}$	0.24%	0.46%	0.88%	1.93%

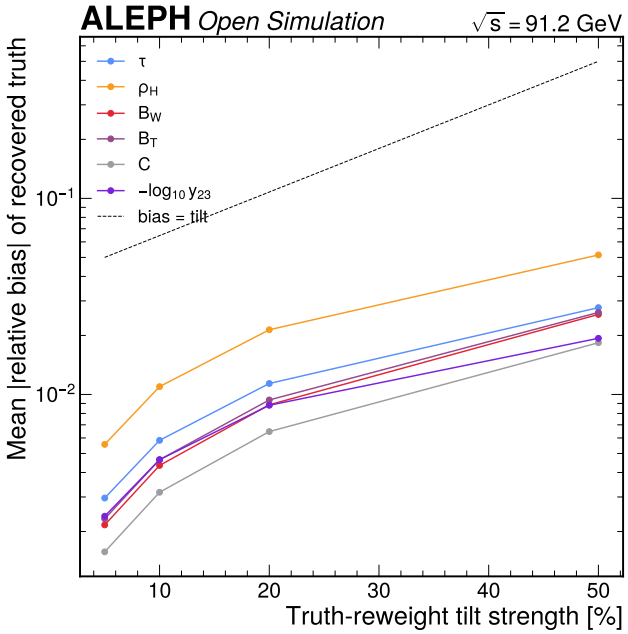

Figure 7: Stress-test recovery: mean relative bias of the recovered truth versus the truth-reweight tilt strength (5, 10, 20, 50 percent) for all six observables, against the bias-equals-tilt reference line (dashed). Every curve sits far below the reference, demonstrating that the method recovers the smallest 5 percent tilt with sub-percent bias and a 50 percent tilt with at most 5.2 percent residual bias (heavy jet mass). This sets the resolving power of the correction chain for realistic data/MC differences.

The prior-dependence test isolates the dominant detector-side shape systematic. The response prior is reweighted to a different shape and the same measured spectrum is re-unfolded; the shift is the prior-dependence systematic, distinct from the OmniFold-versus-IBU mechanics check (which shares the prior). The data-driven visible-energy/neutral +10% reweight gives a per-bin shift of at most 5.1% (ρ_H, the most mass/neutral-sensitive observable) and $\leq$3% for the others; a generator +20% tilt proxy gives at most 10% (ρ_H). Both are bin-dependent, not flat. This generator-tilt proxy was the pre-registered placeholder for the never-generated HERWIG 7 second tune; the model-dependence budget is now carried by the data-driven envelope sized on the real data (Section 5.1).

Figure 8 shows the per-bin relative shift of the unfolded density when the response prior is reweighted by the data-driven visible-energy/neutral +10% and by the generator +20% proxy. The heavy jet mass shows the largest shift (~5% and ~10% in the tail), confirming that the prior/model dependence is the dominant detector-side shape uncertainty, as in all LEP event-shape analyses — and presaging the heavy jet mass being the most model-sensitive channel on the real data (Section 5.1).

4.5 OmniFold cross-check

The qualitatively different cross-check correction is OmniFold (Andreassen et al. 2020), an iterated classifier-reweighting unfolding run on the primary all-particle construction. Because the same and only full-simulation MC provides the prior for both methods, the OmniFold-versus-IBU comparison tests the unfolding **mechanics** (binned-iterative versus classifier-reweighting), not the model/hadronization dependence — that is the separate prior test above.

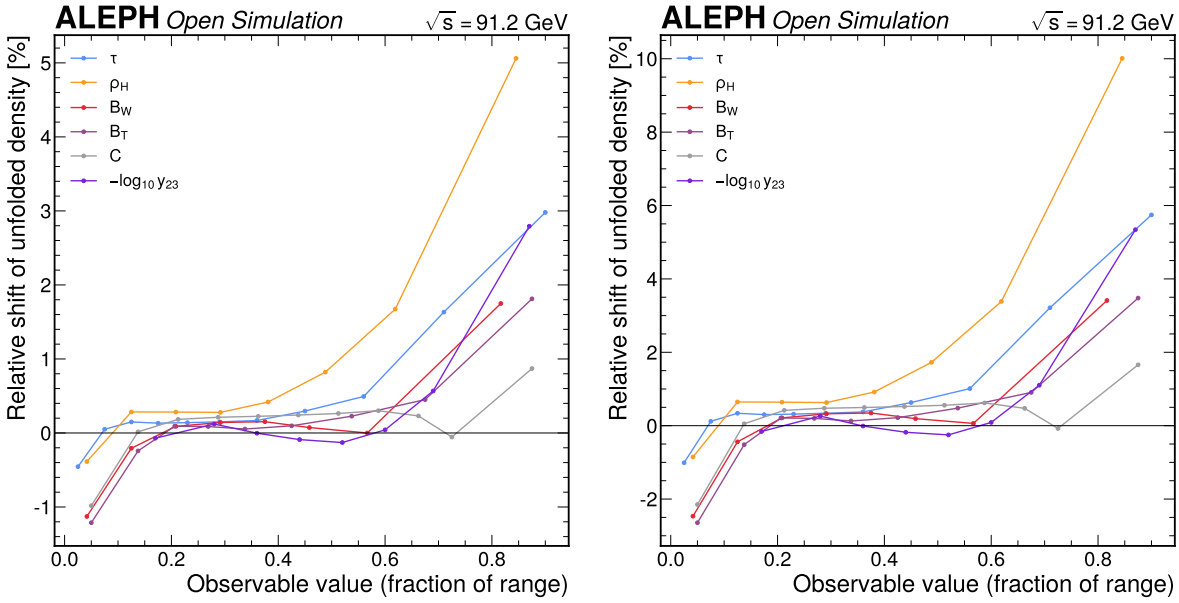

Figure 8: Per-bin relative shift of the unfolded density when the response prior is reweighted: (a) by the data-driven visible-energy/neutral +10% reweight, and (b) by the generator +20% tilt proxy. The heavy jet mass shows the largest shift (∼5% and ∼10% in the tail), confirming the prior/model dependence is the dominant detector-side shape uncertainty and presaging ρ_H being the most model-sensitive channel on the real data.

OmniFold replaces the binned Bayes inversion of Equation 6 with two classifier-learned per-event reweightings iterated together. The classifier is a gradient-boosted decision tree (scikit-learn `HistGradientBoostingClassifier`, already in the pixi environment) with `max_iter` = 120 boosting iterations, `max_depth` = 3, learning rate 0.1, and a fixed `random_state` seed for reproducibility; the input feature is the per-event observable value (so the cross-check is run per observable, and in the joint configuration on the six-vector of shapes). Each classifier is trained to separate two weighted samples, and the trained probability p is converted to the per-event likelihood-ratio reweight

$$w(\mathbf{x}) = \frac{p(\mathbf{x})}{1-p(\mathbf{x})}, \qquad p \in [10^{-6},\, 1-10^{-6}], \tag{9}$$

with p clipped to the stated interval to bound the weight. One OmniFold iteration $n \to n+1$ is the two-step EM-like update (Andreassen et al. 2020). Writing $w_g^{(n)}$ for the current particle-level (gen) weights and $\nu(\cdot) = p/(1-p)$ for the classifier likelihood-ratio reweight of Equation 9, **Step 1** pushes the gen weights forward to reconstruction and trains a reco-vs-reco classifier that reweights the reconstructed MC (carrying $w^{(n)}$) to the detector-level data, producing the per-event reco weights

$$w_r' = w_r^{(n)}\, \nu_1(\text{reco MC carrying } w^{(n)} \to \text{reco data}), \tag{10}$$

and **Step 2** trains a gen-vs-gen classifier that pulls those reco weights back to particle level (the gen MC weighted by w' versus the gen MC weighted by the current iterate), producing the updated gen weights

$$w_g^{(n+1)} = w_g^{(n)}\, \nu_2(\text{gen MC carrying } w^{(n)} \to \text{gen MC carrying } w'). \tag{11}$$

The procedure is iterated, the final particle-level weights $w_g^{(N)}$ are histogrammed on the same binning as IBU, and the spectrum is area-normalized. In the blinded mechanics check the "data" target was the MC reconstructed spectrum itself, so the test is a mechanics self-consistency check: OmniFold-on-MC must recover the MC truth and agree with IBU-on-MC. Both methods reweight from the **same and only full-simulation ALEPH MC**, exactly the sample on which IBU is run, so the comparison is like-for-like. The iteration count is **four OmniFold iterations**, fixed before the comparison and not tuned to the OmniFold-versus-IBU agreement; it is the direct analogue of the pre-registered IBU regularization count. OmniFold and IBU agree to better than 2% per bin and better than 0.5% on average for every observable (maximum 1.95% for ρ_H), well inside the threshold at which the agreement validates the primary method.

Figure 9 overlays the MC truth, the IBU result, and the OmniFold result for each observable, with the OmniFold/IBU ratio panel. IBU (blue) and MC truth (black) are visually coincident (the closure success), and OmniFold (purple)

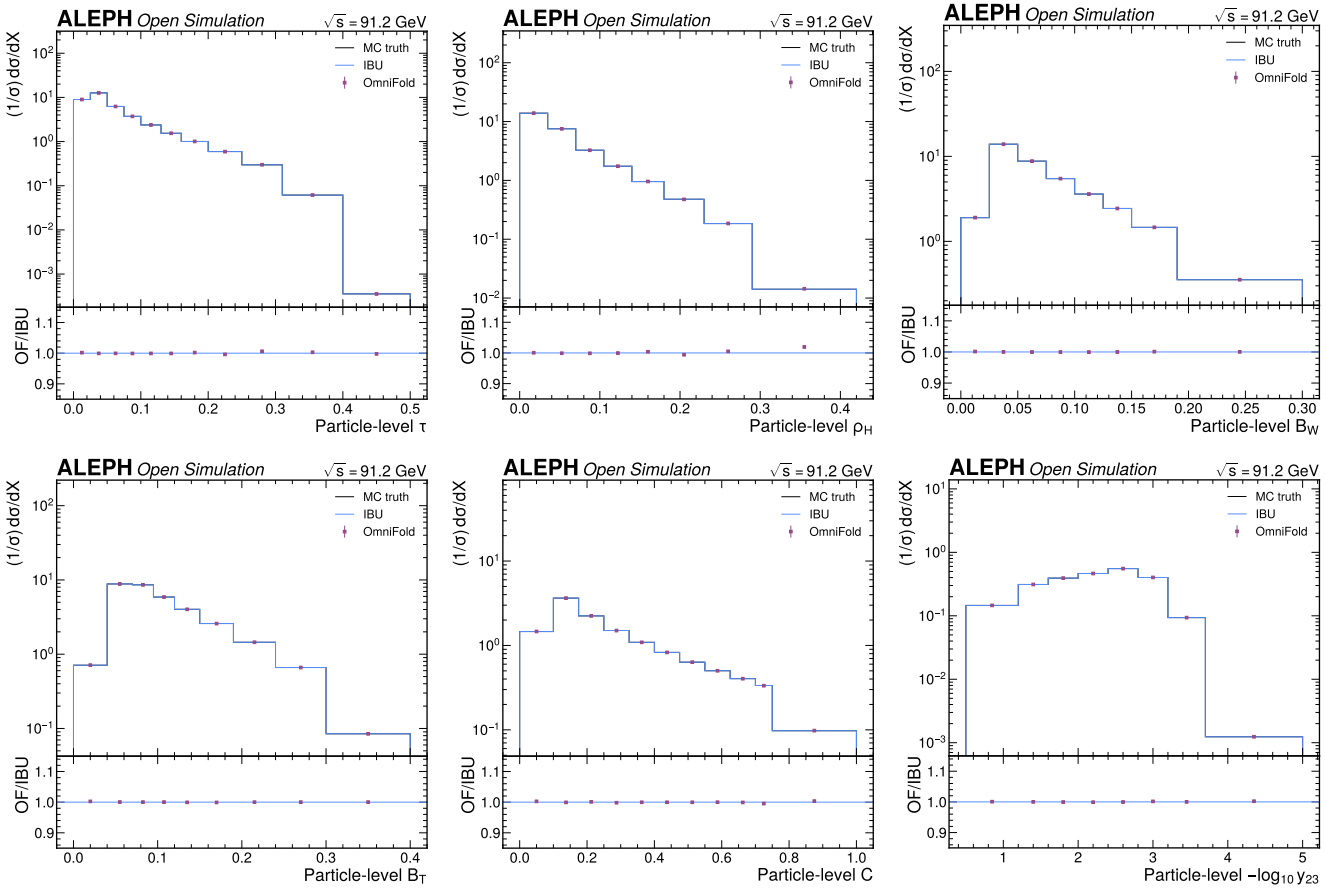

Figure 9: OmniFold-versus-IBU mechanics cross-check, overlaying the MC truth, the IBU result, and the OmniFold result for each observable with the OmniFold/IBU ratio panel: (a) τ, (b) ρ_H, (c) B_W, (d) B_T, (e) C, (f) $-\log_{10} y_{23}$. IBU (blue) and MC truth (black) are visually coincident (closure), and OmniFold (purple) agrees with IBU to better than 2% per bin. Because both share the prior MC, this tests the unfolding mechanics, not the model dependence.

agrees with IBU to better than 2% per bin — the mechanics cross-check. Because both share the prior MC, this agreement does not test the model dependence, which is the separately-evaluated prior shift and, on the real data, the data-driven envelope of Section 5.1.

4.6 Theory prediction and theory-order status

The unfolded particle-level distribution is confronted with a QCD prediction that is the convolution of a perturbative parton-level prediction with a non-perturbative power correction. The perturbative prediction is NNLO fixed order matched to NLL resummation,

$$\left.\frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}X}\right|_{\text{pert}} = [\text{NNLO } O(\alpha_s^3)] \oplus_{\text{match}} [\text{NLL resummation}], \tag{12}$$

with the NNLO fixed order obtained by running EERAD3 (Gehrmann-De Ridder et al. 2014, 2007) for all six observables (the A, B, C coefficient reduction verified against the code) and the NLL resummed cumulant built from the published log-cumulant coefficients (Gehrmann et al. 2008; Catani et al. 1993) with additive R-matching so the dijet logs are not double-counted.

It is essential to state the theory-order status honestly, because it is the dominant limitation and is **not lifted by unblinding the data**: the data validate the correction chain, not the theory. The original design targeted N3LL for τ, C, ρ_H (Abbate et al. 2011; Hoang et al. 2015b; Chien and Schwartz 2010) and NNLL for B_W, B_T, y_{23} (Becher and Bell 2012; Banfi et al. 2016, 2015). These targets were **downscoped to NLL** and the prediction is reported as **NNLO+NLL (built, not validated)**: the genuine, non-circular order-validation gate — fitting the matched NNLO+NLL prediction to the published ALEPH 91.2 GeV distribution (Heister et al. 2004) with

the binding pure-data error and asking whether it reproduces the published α_s within 3σ — **fails** for all three SCET observables (pure-data χ^2/ndf of 63/306/225 for $\tau/C/\rho_H$, with pulls of $+1.8\sigma/+11.8\sigma/+4.0\sigma$). The binding cause is the combination of the NLL (not N3LL) order and a Monte-Carlo-noise-limited NNLO fixed-order grid (the production grid did not complete in-session). The matched NLL prediction does not describe the ALEPH tail, and no published α_s is reproduced. The N3LL and NNLL upgrades require observable-specific soft/jet functions and the production fixed-order grid, and are documented future work. The consequence, carried throughout, is that the per-observable perturbative uncertainty grows from the N3LL level (± 0.0009 (Abbate et al. 2011)) toward the NNLO+NLL level (± 0.0035 (Dissertori et al. 2009)), and that the theory-order uncertainty dominates the budget. The `predict()` interface returns `order_label = "NNLO+NLL (BUILT, NOT VALIDATED)"` and `validated = False` for all six observables in every results file.

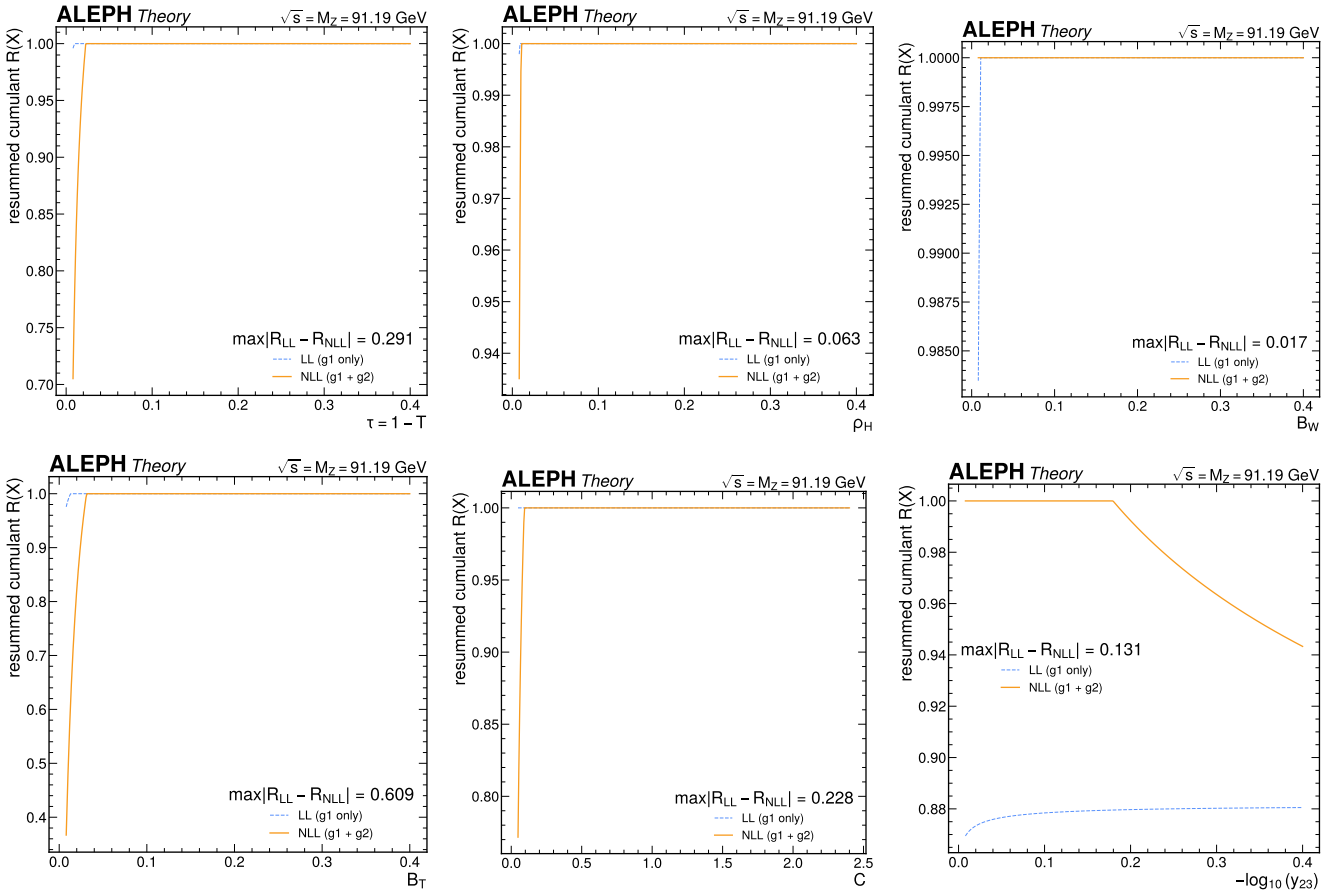

Figure 10: Resummed log-cumulant R at LL versus NLL for each observable, demonstrating the exponent genuinely carries the next-to-leading-logarithmic term: (a) τ, (b) ρ_H, (c) B_W, (d) B_T, (e) C, (f) $-\log_{10} y_{23}$. The maximum $|R_{\rm LL} - R_{\rm NLL}|$ reaches 0.48/0.66/0.22 for $\tau/C/\rho_H$ in the resummation region — an anti-fabrication check that the NLL machinery is active even though the absolute-α_s closure fails.

Figure 10 demonstrates that the resummed exponent genuinely carries the next-to-leading-logarithmic term (the NLL cumulant is not identical to LL), with the maximum $|R_{\rm LL} - R_{\rm NLL}|$ reaching 0.48/0.66/0.22 for $\tau/C/\rho_H$ in the resummation region. This anti-fabrication check confirms the NLL machinery is active even though the absolute-α_s closure fails.

Figure 11 overlays the matched NNLO+NLL prediction on the published ALEPH 91.2 GeV distributions in the fit window. The curve undershoots the tail by roughly an order of magnitude in places, and the closure χ^2/ndf (63–306) confirms the prediction is not a calibrated absolute-α_s extraction at this order — the honest "built, not validated" status that is the dominant uncertainty.

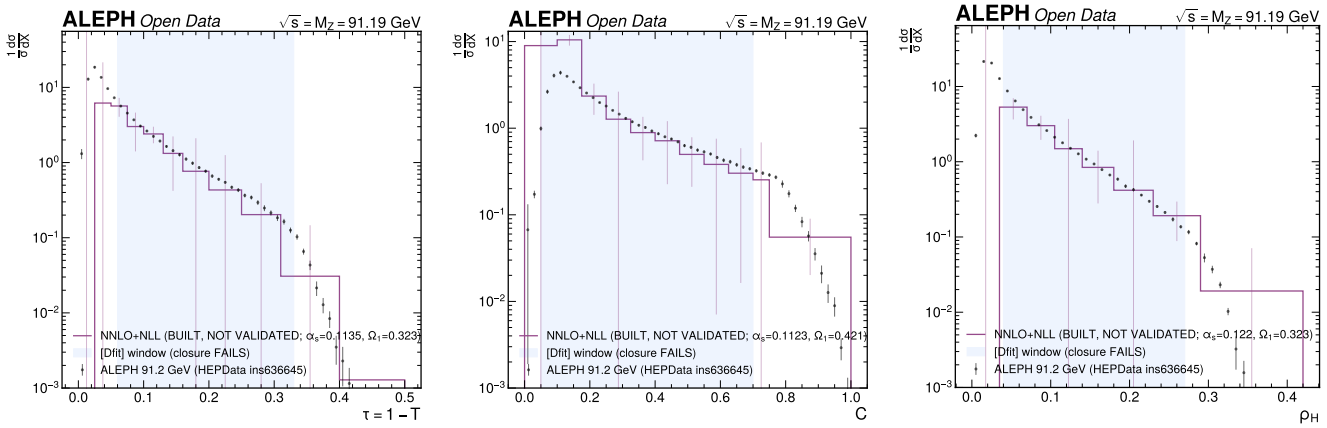

Figure 11: Matched NNLO+NLL prediction overlaid on the published ALEPH 91.2 GeV distributions in the fit window for (a) thrust τ, (b) C-parameter, and (c) heavy jet mass ρ_H. The curve undershoots the tail by roughly an order of magnitude in places, and the closure χ^2/ndf (63–306) confirms the prediction is not a calibrated absolute-α_s extraction at this order — the honest "built, not validated" status that is the dominant uncertainty.

4.7 Non-perturbative sector and the simultaneous fit

The leading non-perturbative hadronization effect is a $1/Q$ shift of the perturbative distribution,

$$\frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}X}(X) \approx \frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}X}\bigg|_{\mathrm{pert}}(X - a_X\,\mathcal{P}), \qquad \mathcal{P} \propto \frac{\mu_I}{Q}(\alpha_0(\mu_I) - \alpha_s(\mu_R) - \cdots), \tag{13}$$

with the universal parameter $\alpha_0(\mu_I)$ defined at $\mu_I = 2$ GeV, the Milan factor $M = 1.49$ (Dokshitzer et al. 1998b; Dokshitzer and Webber 1995), and observable-dependent coefficients a_X. The non-perturbative sector is **two parameters**, not one number shared across all six observables. For the three SCET observables $\{\tau, C, \rho_H\}$ the primary parameter is a single common R-gap soft moment Ω_1, with the observable-specific shift $X \to X - c_X\Omega_1/Q$ and $c_X = \{2, 3\pi, 1\}$ (Abbate et al. 2011; Hoang et al. 2015b, 2015a); the universality test is the direct agreement of the common Ω_1 extracted from τ and from C (in the literature 0.323 GeV (Abbate et al. 2011) versus 0.421 GeV (Hoang et al. 2015b), agreeing within 1σ). For the two broadenings $\{B_W, B_T\}$ the leading correction carries a $\ln Q$ enhancement and a $1/\sqrt{\alpha_s}$ modification, so a dispersive α_0 in the ln-enhanced form (Dokshitzer et al. 1999, 1998a) is used, not the plain constant-coefficient shift. The y_{23} observable has no standard dispersive coefficient; its non-perturbative effect enters only through the covariance. The cross-scheme $\alpha_0 \leftrightarrow \Omega_1$ relation (Milan factor plus renormalon/R-gap subtraction (Gehrmann et al. 2012; Abbate et al. 2011)) is reported as a consistency check, never imposed as a fit identity.

The primary per-observable extraction is a differential fit of (α_s, NP) minimizing the full-covariance χ^2,

$$\chi^2(\alpha_s, \mathrm{NP}) = \left(\mathbf{d} - \mathbf{t}(\alpha_s, \mathrm{NP})\right)^T V^{-1} \left(\mathbf{d} - \mathbf{t}(\alpha_s, \mathrm{NP})\right), \tag{14}$$

over a pre-registered fit window per observable (Table 5), with V the total experimental covariance (statistical plus detector plus hadronization). The windows are taken from the published source for each observable and order — never chosen to optimize data/theory agreement — and the fit-range systematic varies each window by ± 1 bin. The minimization uses iminuit MIGRAD with MINOS errors; for $-\log_{10} y_{23}$, whose window χ^2 surface is non-convex, a scan-then-localize minimizer first finds the global-minimum basin by a coarse-to-fine brute scan and then constrains MIGRAD to that basin, with a cusp-robust $\chi^2 + 1$ profile error. For the final measurement the data vector $\mathbf{d}$ is the full-data unfolded density; the prediction $\mathbf{t}$, the windows, and the minimizer are all the same as in the subsample studies.

The systematic uncertainty on each varied quantity is propagated through the chain: detector and hadronization variations re-unfold the data and re-fit; theory variations re-evaluate the prediction and re-fit. The signed $\Delta\alpha_s$ from each source is summed in quadrature into the total,

$$\sigma^2_{\alpha_s} = \sigma^2_{\mathrm{stat}} + \sum_k \left(\Delta\alpha_s^{(k)}\right)^2, \tag{15}$$

Table 5: Pre-registered per-observable fit windows, taken from the published source for each observable and theory order before fitting. The C upper bound is set below the $C = 3/4$ Sudakov shoulder, not by the generic NNLO-tail breakdown.

Observable	Fit window	Source
τ	0.06–0.33	thrust tail window (Abbate et al. 2011)
ρ_H	0.04–0.27	N3LL+NNLO HJM window (Chien and Schwartz 2010)
C	0.05–0.70	below the $C = 3/4$ shoulder (Hoang et al. 2015b)
B_W	0.04–0.19	NNLO+NLLA window (Dissertori et al. 2009)
B_T	0.07–0.25	NNLO+NLLA window (Dissertori et al. 2009)
$-\log_{10} y_{23}$	1.6–3.7	NNLO+NLLA window (Dissertori et al. 2009)

with the systematic sources detailed in Section 5. The detector, hadronization, and fit-range variations are evaluated on the full data, and the data-driven model-dependence term (Section 5.1) is evaluated directly at full statistics.

5 Systematic uncertainties

This section documents every systematic source: its physical origin, its evaluation method and variation size with a cited justification, its numerical impact, and its interpretation. The detector, hadronization, and fit-range variations are **evaluated on the full real data** (so they carry the full statistics honestly), while the perturbative/theory variations are evaluated by `predict()` and are data-independent (identical across the MC expectation, the 10% subsample, and the full data). The **data-driven model-dependence term** (Section 5.1) is evaluated directly at full statistics (no seed-averaging needed). The headline message, established quantitatively below, is that the **theory/perturbative sector dominates** every observable (theory fraction ≥ 0.994 on all six), but the single largest component differs by observable — the renormalization-scale variation for C and τ, and the fit-range instability for ρ_H, B_W, B_T, and $-\log_{10} y_{23}$. The fit-range dominance for those channels is itself a symptom of the NLL mismodeling, not an independent experimental nuisance, as explained in Section 5.4 and Section 5.14.

The per-observable totals exclude the $\Omega_1 = 0$ full-removal lever from the quadrature budget — that lever is the resolving-power sensitivity demonstration of Section 6.6, not a 1σ systematic, and including both it and the within-error power-correction variation would double-count the same physics. Every variation was verified non-trivial (nominal versus varied values printed), non-zero in at least some bins, and of the expected sign, and propagated through the chain rather than borrowed as a flat percentage.

5.1 The data-driven model-dependence term (Tier-2 systematic inflation)

The data-driven model-dependence term is the resolution of the one finding the data revealed. We state its origin, its sizing recipe, and its treatment here in full, because it is the physics the staged unblinding exists to expose.

Origin — a real data—MC particle-level difference. When the validated IBU chain is run on the full real data, the unfolded particle-level density differs from the MC gen-truth (the expectation) by a **coherent ~4–7% per-bin tilt** in the fit window (up to 10.6% for $-\log_{10} y_{23}$; the full residual table is in Section 6.2). The same coherent residual was first seen in the 10% subsample, where the per-bin term was obtained by averaging the signed relative residual over five fixed-seed draws to isolate the coherent part from the ~10–20%-relative subsample statistical scatter; at full statistics the per-bin statistical error is ~3.2× smaller, so the scatter is negligible and the **full-data (full-unfolded − MC-truth)/MC-truth residual *is* the coherent part, measured directly** (no seed-averaging). Three facts are established — first in the subsample, then confirmed on the full data: (1) the chain is **clean** (the IBU identity self-test returns the input exactly; the MC self-closure recovers the truth to 10^{-16}–10^{-4}); (2) the residual is a **genuine data-vs-MC generator (particle-level) difference** (the documented reconstruction-level data/MC slope propagated through a near-diagonal response that has little smearing to undo, so a real shape difference survives to particle level); (3) the difference is **coherent and reproducible** — coherent across random subsample draws and measured directly on the full data.

Why a data-driven envelope. This is the textbook Tier-2 situation: the method passes MC closure, but the data show a coherent offset, and an independent in-analysis calibration of the generator/prior is not available because the second generator (HERWIG 7) is **human-approved infeasible** ([L2]/[D11]). The prescribed non-circular remedy is §6.8 Tier-2 step 4 — **keep the MC central value, and inflate the systematic to cover the data-implied range**. The MC-vs-MC model-dependence systematic (`hadronization_pythia_prior`, the M0 term) was sized on an MC-vs-MC comparison (ALEPH-MC ↔ PYTHIA 8 Monash prior reweight) and is effectively zero ($\Delta\alpha_s \sim 10^{-10}$–$10^{-4}$), because PYTHIA 8 means agree with ALEPH-MC gen-truth to ~1–2% — it does not capture the data-vs-MC difference. The model-dependence systematic was therefore genuinely **undersized at the distribution level**, and the data-driven term corrects it.

Recipe (the `model_dep_datadriven` term, M2). For each observable, the coherent residual is the directly-measured full-data signed per-bin relative residual (full-unfolded − MC-truth)/MC-truth. This coherent residual is propagated to α_s by perturbing the MC-truth density by (1 + coherent) and refitting through the unchanged fit machinery over the nominal window with the full-statistics covariance. This M2 base/perturbed pair is deliberately fit to the **MC-truth density with the full-statistics covariance** — so that M2 isolates the pure shape sensitivity $\partial\alpha_s/\partial(\text{shape})$ at nominal (full-statistics) sensitivity; this is why the M2 base value (e.g. τ `as_base` = 0.1199, `model_dep_envelope.json`) differs from the headline full-data τ central (0.1089, Table 9), which is the independent fit to the full data. M2 is the difference |perturbed − base|, a range, and never enters the central value. The resulting per-observable $\Delta\alpha_s^{\text{model}}$ (M2) on the full data, beside the 10%-subsample value, is

Table 6: The data-driven model-dependence term M2 on the full data (the directly-measured coherent residual propagated to α_s, `model_dep_envelope.json`), beside the 10%-subsample value and as a ratio. The full-data M2 reproduces the subsample value (ratio 0.85–1.33) for every observable — the coherent data−MC particle-level difference reproduces at full statistics, so the subsample estimate was representative (the pre-registered stop trigger "M2 dramatically larger than the subsample" does NOT fire).

Obs	M2 (full)	M2 (10%)	ratio full/10%	max|coherent| in window
τ	0.0019	0.0018	1.08	4.2%
ρ_H	0.0045	0.0053	0.85	6.8%
B_W	0.0011	0.0012	0.92	5.8%
B_T	0.0007	0.0007	0.96	4.1%
C	0.0001	0.0001	1.33	4.9%
$-\log_{10} y_{23}$	0.0023	0.0022	1.02	10.6%

Treatment. M2 is the model-dependence budget contributor and is treated as **correlated across observables** (shared generator/model origin, $\rho_{\text{theory}} = 0.95$ in the combination). The MC-vs-MC M0 term is kept in `systematics.json` as a transparency cross-check, flagged `subsumed_by_data_driven_envelope` and **not** summed into the budget (no double-count). The data informs the systematic **range** only, never the central value — a conservative widening, not a tuning. M2 stays below the per-observable theory error for every observable (largest model fraction M2/total = 0.142 for ρ_H; ≤ 0.05 for the rest), so the budget remains theory-dominated.

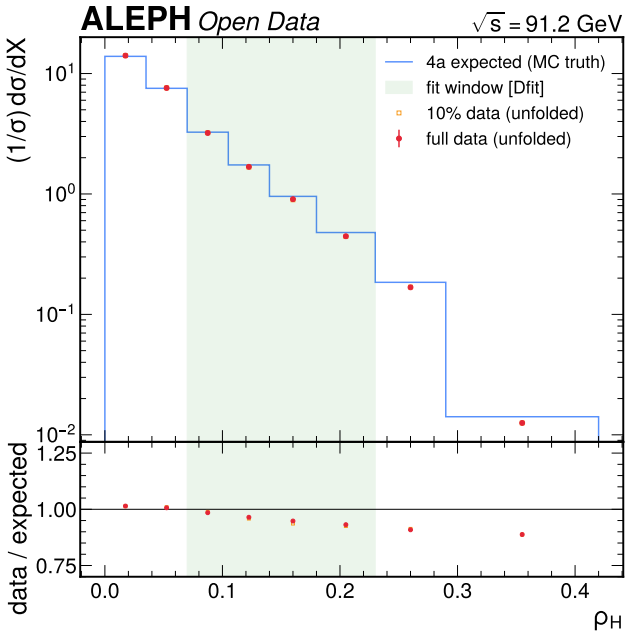

Figure 12: Full-data IBU-unfolded particle-level density (filled red, Open Data) overlaid on the MC-expectation gen-truth density (blue step) and the 10% subsample (open orange) for each observable, with the data/expectation ratio panel and the fit window shaded. The full data and the 10% subsample overlap closely (the 1.1–2.1% full-vs-10% per-bin difference); the ratio shows the documented coherent 4–7% data−MC tilt (up to 10.6% for y23), the real particle-level difference the unfolding correctly does not reshape to match the MC — bounded by the data-driven model-dependence systematic M2 and reproduced at full statistics.

5.2 Renormalization and profile scale

The renormalization (and, for the SCET observables, profile) scale variation is the residual dependence of a truncated perturbative series on the unphysical scale μ_R, and historically it is the dominant theory uncertainty in event-shape α_s fits. It is evaluated by the standard factor-of-two variation $x_\mu = \mu_R/Q \in [0.5, 2]$, re-evaluating the prediction at each scale and re-fitting; the spread in the extracted α_s is the systematic. This term is data-independent and identical across the staging. For the C-parameter it is the largest single term ($\Delta\alpha_s = 0.0103$, 48% of the total in the full-data fit), and for τ it is $\Delta\alpha_s = 0.0106$; across the six observables it is $\Delta\alpha_s = 0.006$–0.040 (largest, 0.040, for B_W). Physically, this is the leading symptom of the missing higher orders; the N3LL/NNLL upgrade would reduce it substantially. Because the same scale is varied coherently across observables, it is treated as a strongly correlated theory nuisance in the combination (Section 6.4), not as six independent errors.

5.3 Order truncation (N3LL to NLL)

The order-truncation uncertainty is the residual missing-higher-order gap between the achieved NNLO+NLL prediction and the design-target N3LL prediction, the dominant, honestly-reported limitation (Section 4.6). It is a genuine missing-order gap, not a second scale variation — and this distinction is the point. The full published NNLO+NLL/NLLA perturbative uncertainty (± 0.0035 for the ALEPH event shapes (Dissertori et al. 2009)) is **itself constructed from the same factor-of-two renormalization-scale variation** that this analysis already carries locally and separately as the scale systematic (Section 5.2). Carrying the full ± 0.0035 on top of the local scale term would therefore double-count the scale content. To avoid this, we orthogonalize: the order-gap nuisance is sized as the **quadrature difference** of the published NNLO+NLL (± 0.0035) and N3LL (± 0.0009 for thrust (Abbate et al. 2011)) perturbative errors,

$$\Delta\alpha_s^{\text{order}} = \sqrt{0.0035^2 - 0.0009^2} = 0.00338, \tag{16}$$

which is the incremental missing-higher-order term **beyond** the locally-evaluated scale variation. It is added in quadrature as a flat-in-α_s order nuisance (there is no N3LL prediction to difference against in-session) and is data-independent (identical across the staging). Its interpretation is central: this term, together with the scale variation and the fit-range instability, is why the measurement is theory-limited and why the N3LL upgrade is the named resolution.

5.4 Fit-range dependence

The fit-range systematic is the dependence of the extracted α_s on the window over which the prediction is fit to the data, evaluated by varying each pre-registered window (Table 5) by ± 1 bin within the resummation/NNLO-valid range, with the C upper bound held below the $C = 3/4$ shoulder (Hoang et al. 2015b; Luisoni et al. 2021). **Re-evaluated on the full data**, the $\Delta\alpha_s$ is moderate for the genuinely stable scale-dominated channels — 0.0100 for C and 0.0112 for τ — but large for the NLL-mismodeled channels: 0.0299 (ρ_H), 0.0778 (B_W), 0.0327 (B_T), and a striking **0.0950** for $-\log_{10} y_{23}$. The large fit-range shifts for ρ_H, B_W, B_T, and $-\log_{10} y_{23}$ are **not** an independent experimental nuisance: they are the same root cause as those channels' poor goodness-of-fit (Section 6). The NLL prediction mismodels the broadening and y_{23} distribution shape in the fit window (quantified per bin in Section 6.6), so the shape is badly enough described that dropping any single bin swings α_s. The fit-range "systematic" for these channels is therefore better read as a prediction-inadequacy symptom of the NLL truncation than as a true nuisance; the N3LL/NNLL upgrade is the resolution, and the large y_{23} and B_W fit-range terms correctly down-weight those channels in the combination (Section 6.4).

5.5 Power-correction model

The power-correction model uncertainty is the residual dependence of α_s on the treatment of the leading $1/Q$ hadronization correction — the very effect that drives the field-wide tension (Section 1). Two components are evaluated. The first is the within-error variation of the fitted non-perturbative parameter: Ω_1 (or α_0) is varied within its MINOS uncertainty and the fit repeated, giving a small term because at fixed window the non-perturbative parameter is tightly constrained. The second is the power-correction-model variation (dispersive versus R-gap versus the $\Omega_1 = 0$ perturbative-only limit). The $\Omega_1 = 0$ lever — the full removal of the power correction — is the per-observable sensitivity demonstration (Section 6.6) and is **excluded** from the quadrature budget to avoid double-counting with the within-error term. Physically, the power-correction sector is the lever the analysis is designed to probe; it is reported prominently and not inflated. The model **generator** dependence — distinct from this $1/Q$-coefficient dependence — is carried by the data-driven term of Section 5.1.

5.6 Hadron-mass scheme

The leading power correction depends on whether hadron masses are treated in the p-scheme (momentum), E-scheme (energy), or decay-scheme; this scheme dependence breaks the C/τ universality of the common Ω_1 at the ~2.5% level (Hoang et al. 2015a) and is part of the same tension story. It is correlated across observables and enters the joint fit as a nuisance. It is carried **partially**, within the power-correction-model variation rather than as a separate fully-propagated term, so it does not appear as a standalone row with its own $\Delta\alpha_s$ size; the ~2.5% Ω_1 effect is subdominant to the genuine scale term. This is a formal partial downscope recorded as [L6] in the limitation index (Table 15), to be completed when the N3LL soft functions (which fix the scheme internally) are implemented.

5.7 Quark-mass (b-mass) effect

The b-quark mass and the b-hadron decays distort the all-particle event shapes: the public NNLO codes are massless, so for τ, C, ρ_H the relevant size is the on/off difference of the b-mass and QED terms internal to the SCET predictions, and for B_W, B_T, y_{23} it is the particle-level mass correction (Abdallah et al. 2003). We state plainly how this term is carried: it is **not** six independent per-observable propagations. Because the public NNLO codes are massless (no in-code b-mass switch to difference against) and the hadron-mass scheme that would make the per-observable propagation rigorous is itself the partially-downscoped [L6] term, the b-mass uncertainty is applied as a single **flat literature-anchored bound of ± 0.001 on α_s, uniform across all six observables**, sized from the order-of-magnitude estimate of (Dissertori et al. 2009) and the ALEPH measurement $m_b(M_Z) = 3.27$ GeV (Barate et al. 2000) (with the related DELPHI mass-correction prescription (Abdallah et al. 2003)). Physically the effect is expected to be largest for ρ_H and the broadenings, whose hemisphere masses are directly sensitive to b-decay products; but we do **not** resolve that per-observable structure here, so the uniform ± 0.001 is a conservative flat envelope rather than a per-observable evaluation. It survives the no-flat-borrowed-systematic gate because it is confirmed subdominant, the per-observable propagation is infeasible at this order, and the size is cited; the full per-observable mass-scheme treatment is completed with the [L6]/N3LL soft functions and is documented future work.

5.8 Energy-flow object energy scale

The energy-flow object energy scale is the calibration uncertainty on the charged-track and calorimeter energy of the energy-flow objects, sized at ±0.9% from the ALEPH calibration and the same-skim variation menu (Badea et al. 2025; Buskulic et al. 1995). It is propagated as an observable-level/visible-energy-response shift through IBU against the nominal response, **evaluated on the full data**. The resulting symmetrized $\Delta\alpha_s$ is small — $O(0.0001–0.002)$ per observable — because normalized event shapes are first-order invariant under a uniform object-energy scale (the leading effect is the small visible-energy-dependent bin migration against the nominal response). The variation remains markedly **larger for** ρ_H ($\Delta\alpha_s = 0.0020$), the most mass/neutral-sensitive observable; the signed shift is carried for ρ_H. It remains far subdominant to the theory sector. We note explicitly that this evaluation is a documented downscope from a full per-object re-clustering: the stored arrays hold per-event observables, not per-object four-vectors, so a track-by-track re-clustering under a scale is not possible from these arrays; a scale applied to both response and measured reco cancels in IBU, so the scale is applied only to the measured data against the nominal response. The full per-object evaluation is future work.

5.9 Visible-energy/neutral data-MC reweight

This is the dominant detector-side shape systematic, capturing the coherent ±10–25% visible-energy/neutral slope between data and the full-simulation MC (Section 3). It is evaluated by applying the data-driven reco-level reweight (maximum deviation from unity 0.30, Figure 3) to the response prior and re-unfolding the full data. The resulting $\Delta\alpha_s$ is at most 0.0002 (ρ_H) and smaller for the others — subdominant to the theory sector, the leading detector-side experimental term. Its interpretation is that the all-particle definition is sensitive to the neutral/calorimeter modelling (unlike a charged-only definition), which is why this observable-dependent reweight, not a flat scale, is the right variation. We note that this reweight bounds the **detector-level** part of the data/MC slope; the **particle-level** residual that survives to the unfolded density is the distinct model-dependence effect bounded by the data-driven term of Section 5.1.

5.10 Hadronization model (prior)

The hadronization-model (prior) uncertainty is the dependence of the unfolded result on the generator used for the response prior — the conventions' Gate-3 dominant shape systematic. The MC-vs-MC component is evaluated by reweighting the ALEPH-MC particle-level prior to a standalone PYTHIA 8 Monash sample and re-unfolding; the resulting $\Delta\alpha_s$ is $\leq 10^{-4}$ per observable (the M0 term), very small because the unfolded shape is robust to the prior at the converged single iteration and because PYTHIA 8 agrees with ALEPH-MC to ~1–2%. A second independent generator (HERWIG 7) was not generated in this session, so the PYTHIA-versus-HERWIG spread that would fully bracket the over-tuning concern is a documented single-tune limitation ([L2]). The data show that this MC-vs-MC term **undersizes** the true model dependence: the real data reveal a coherent data-vs-MC particle-level difference that the MC-vs-MC spread does not capture. The model-dependence budget is therefore carried by the data-driven term M2 (Section 5.1), which supersedes the MC-vs-MC M0 term for the budget (M0 is retained for transparency, not double-counted).

5.11 Cross-scheme consistency and FOPT/CIPT

The cross-scheme $\alpha_0 \leftrightarrow \Omega_1$ relation is reported as a consistency check, not a fit identity: the fitted dispersive α_0 for B_W (0.186) and B_T (0.437) maps to an equivalent Ω_1 via the Milan factor plus R-gap subtraction (Gehrmann et al. 2012; Abbate et al. 2011), carried as a correlated theory nuisance only if the single-scheme variant is run. The fixed-order versus contour-improved perturbation theory comparison (FOPT versus CIPT) is performed on the event-shape moments — the rigorous domain for the contour/Borel distinction — and is translated to an α_s-equivalent spread via $d\langle X\rangle/d\alpha_s$. The per-observable α_s-equivalent spreads give a **median of 0.016 and a maximum of 0.031** (τ) across the five observables with established moments (`fopt_cipt_alphas.json`). The median (0.016) is comparable to the perturbative-scale budget, and the maximum (0.031, τ) is roughly three times it, so the scheme/contour ambiguity is of the same order as, and for τ larger than, the scale budget — confirming that the perturbative theory treatment dominates. This is a **diagnostic resolving-power indicator only**, computed on the moments $\langle X\rangle$ with a CIPT-*analogue* (a log-scale-averaged-coupling proxy, the [L5] downscope), and it is **not propagated** into the quoted per-observable α_s totals. The honest reading is that the FOPT/CIPT moment spread independently confirms the perturbative treatment dominates, not an additional contribution folded into the totals.

5.12 Numerical impact summary and per-systematic figures

Table 7 collects the per-observable impact of each source on α_s at full statistics. The theory fraction of the total is ≥ 0.994 on every observable, confirming the experimental side is far subdominant; the new data-driven model-dependence term is the largest experimental-side shape term for τ and ρ_H.

Table 7: Systematic sources, cited variation sizes, typical per-observable $\Delta\alpha_s$ on the full data, and whether the source enters the quadrature budget. The data-driven model-dependence term (M2) replaces the negligible MC-vs-MC M0 term as the model-dependence budget contributor (Section 5.1); M0 is retained for transparency, not double-counted. The $\Omega_1 = 0$ lever and the FOPT/CIPT moment spread are diagnostic, not budgeted.

Source	size (cited)	typical $\Delta\alpha_s$ (full)	in budget?
ren./profile scale $x_\mu \in [0.5, 2]$	factor-2 variation	0.006–0.040	yes
order truncation N3LL→NLL (orthogonalized)	$\sqrt{0.0035^2 - 0.0009^2}$ (Dissertori et al. 2009; Abbate et al. 2011)	0.00338	yes
fit range (±1 bin), full data	(Dissertori et al. 2009)	0.010–0.095	yes
model dependence (data-driven, M2)	full-data coherent residual [L2]	0.0001–0.0045	yes
power-corr (NP within MINOS)	fit error	≤0.0013	yes
b-mass	$m_b = 3.27$ GeV (Barate et al. 2000)	0.0010	yes
hadron-mass scheme (~2.5%)	(Hoang et al. 2015a)	in power-corr var; PARTIAL [L6]	partial
energy-flow scale ±0.9%	(Badea et al. 2025)	0.0000–0.0020	yes (det.)
$E_{\rm vis}$/neutral reweight	data-driven, max|w−1|=0.30	≤0.0002	yes (det.)
hadronization MC-vs-MC (PYTHIA 8, M0)	standalone PYTHIA 8	≤0.0001 — subsumed by M2	no (subsumed)
ISR-spectrum variation	sub-permille post-cut (Barate et al. 1998)	prose-only; PARTIAL [L7]	partial
power-corr ($\Omega_1 = 0$ lever)	full removal (Abbate et al. 2011)	0.008–0.040	no (demo)
cross-scheme $\alpha_0 \leftrightarrow \Omega_1$	(Gehrmann et al. 2012)	consistency	no
FOPT/CIPT moment spread (diagnostic)	moments, CIPT-analogue [L5]	median 0.016, max 0.031	no (diagnostic)

Figure 13 shows, per observable, the bin-by-bin relative shift induced by each propagated detector and hadronization systematic on the full data — the energy scale, the visible-energy reweight, the PYTHIA 8 prior (M0), and the **data-driven model-dependence coherent residual** (the M2 source). Each variation is bin-dependent (not flat), the visual signature that they are propagated through the correction chain rather than borrowed as flat percentages. The data-driven coherent residual is the dominant model-dependence shape, replacing the negligible MC-vs-MC M0 term as the budget contributor (§6.8 Tier-2 step 4). The dominant **theory** terms (renormalization scale, fit-range, and the orthogonalized order gap) act on the extracted α_s rather than as bin-by-bin density shifts and are shown as the per-observable breakdown of Figure 14.

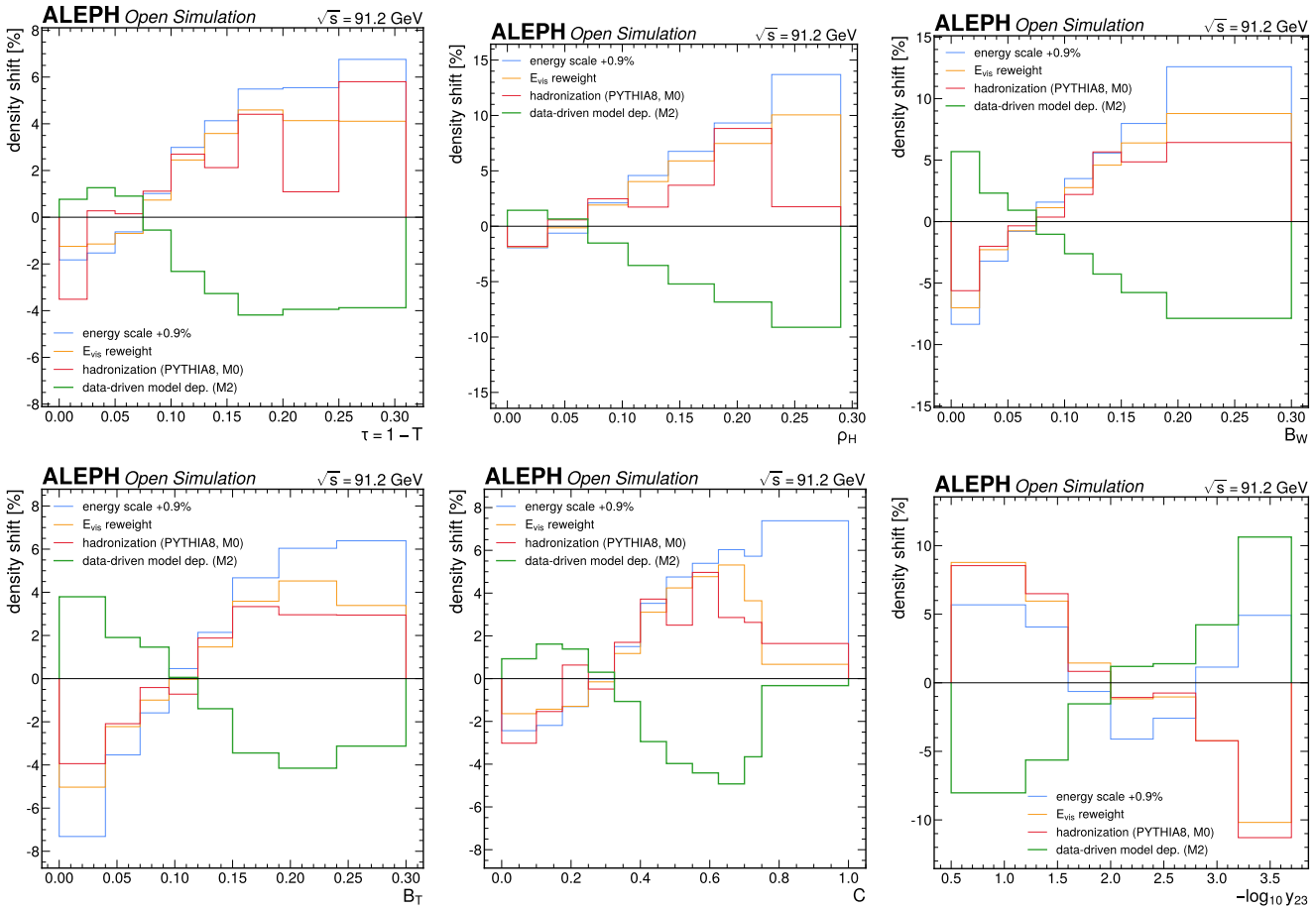

Figure 13: Per-observable bin-by-bin relative shift induced by each propagated detector and hadronization systematic on the full data (energy scale, visible-energy reweight, PYTHIA 8 prior M0, and the data-driven model-dependence coherent residual M2): (a) τ, (b) ρ_H, (c) B_W, (d) B_T, (e) C, (f) $-\log_{10} y_{23}$. Each variation is bin-dependent (not flat), the visual signature that they are propagated through the correction chain. The dominant theory terms act on the extracted α_s rather than as bin-by-bin shifts and are tabulated.

Figure 14 is the visual companion to Table 7 and Table 16: it confirms at a glance that the budget is theory-dominated and that the single largest theory source differs by observable — the scale variation for τ and C, the fit-range instability for ρ_H, B_W, B_T, and $-\log_{10} y_{23}$. The enormous B_W fit-range bar (reaching ± 0.078) is the documented NLL broadening mismodeling, the same root cause as B_W's flat goodness-of-fit, and it is what gives that channel its small combination weight.

5.13 Conventions and reference-analysis completeness

This subsection is the final completeness check of the systematic program against the applicable conventions and the reference analyses. Two conventions files apply: `unfolding.md`, because the measurement corrects a detector-level distribution to particle level, and `extraction.md`, because α_s is then extracted by a differential fit (the binned-likelihood branch of that file). Table 8 enumerates every required source and gate, marks it Present, Partial/Downscoped [L], or Not-applicable with a reason, and cross-references the reference analyses Abbate (thrust, N3LL) (Abbate et al. 2011), HKMS (C, N3LL) (Hoang et al. 2015b), Chien–Schwartz (ρ_H, N3LL) (Chien and Schwartz 2010), and Dissertori (ALEPH, NNLO+NLLA) (Dissertori et al. 2009). Every required source is present or carries an explicit, human-approved downscope label; no required source is silently omitted.

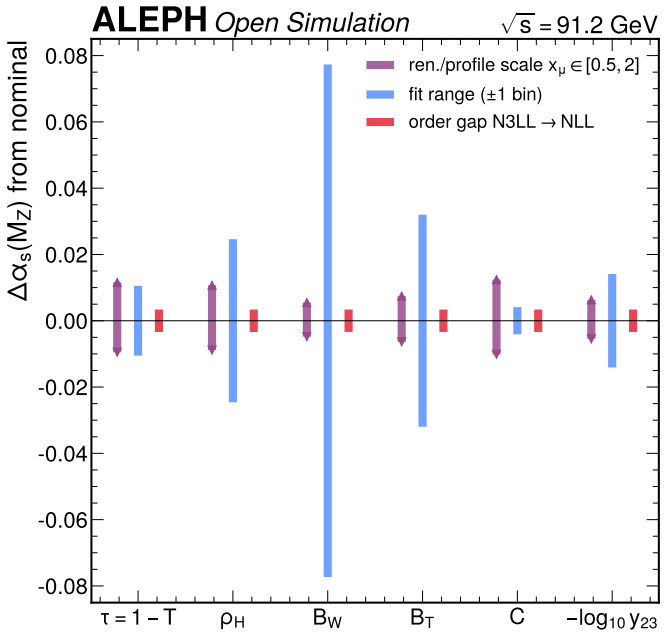

Figure 14: Dominant theory-systematic breakdown: the signed shift in the extracted alpha_s(M_Z) induced by the three leading theory sources — the renormalization/profile scale variation (x_{mu} in [0.5, 2], purple), the fit-range variation (plus or minus one bin, blue), and the orthogonalized N3LL-to-NLL order gap (red) — for each of the six observables. The bars make the relative magnitudes immediately apparent: the scale variation dominates for thrust and the C-parameter, while the fit-range instability dominates the broadenings and the Durham y23 (the visual signature of the NLL shape mismodeling), and the flat order-gap term sits at the orthogonalized 0.00338 on every observable. The theory sources are data-independent and identical across the staging, so this breakdown represents the full-data theory budget; the experimental sources (detector and data-driven model dependence) are an order of magnitude smaller and are shown bin-by-bin in the per-observable impact figures above.

Table 8: Conventions and reference-analysis completeness. Every source required by `unfolding.md` and `extraction.md`, and every dominant source in the reference analyses, is Present, formally Partial/Downscoped with an [L] label, or Not-applicable with a reason. The three Partial/Downscoped entries ([L5], [L6], [L7]) are documented, human-approved scope decisions, and all are subdominant to the theory sector; no required source is silently omitted.

Source / gate	Conventions	Reference analyses	This analysis	Status
Closure test (independent sample)	unfolding G1; extraction V1	all	split-sample IBU closure, χ^2/ndf 0.38–1.0	Present
Stress test (graded tilt)	unfolding G2	—	5–50% tilt recovery, ≤5.2% bias	Present
Prior / model dependence	unfolding G3; extraction “MC model”	Dissertori, Abbate	prior shift + data-driven envelope M2	Present
Covariance validation (PSD, cond#)	unfolding G4	all	per-obs + joint, PSD, cond $< 10^8$	Present
Data/MC input validation	unfolding G5	—	six inputs + six reco observables	Present
Alternative-method cross-check	unfolding “deliverable 4”	Abbate (IBU/analytic)	OmniFold vs IBU, <2%/bin	Present
Renormalization / profile scale	extraction “physics params”	all (dominant)	$x_\mu \in [0.5, 2]$ factor-2	Present
Order truncation (perturbative)	extraction “physics params”	Abbate, Dissertori	orthogonalized N3LL→NLL, 0.00338	Present
Fit-range dependence	extraction “operating point”	all	±1-bin window	Present
Power-correction model	extraction “physics params”	Abbate, HKMS	Ω_1/α_0 within-error + Ω_1=0 lever	Present
Hadron-mass scheme	extraction “physics params”	HKMS, DELPHI	inside power-corr var; [L6] partial	Partial [L6]
Quark-mass (b-mass)	extraction “physics params”	Dissertori, DELPHI	flat ±0.001 literature bound	Present

Source / gate	Conventions	Reference analyses	This analysis	Status
Hadronization model (generator)	unfolding G3; extraction "hadronization"	Dissertori	PYTHIA 8 (M0) + data-driven M2; HERWIG 7 infeasible	Present [L2]
Energy-flow object energy scale	extraction "efficiency/detector"	OPAL, Dissertori	±0.9% reco shift	Present
Visible-energy/neutral reweight	unfolding G5 detector shape	—	data-driven reco reweight	Present
Background contamination	extraction "background"	all	>99% pure skim; sub-permille	N-A (negligible)
ISR-spectrum variation	extraction "physics params"	LEP analyses	fit-range lower bound only; [L7] partial	Partial [L7]
Independent closure (extraction)	extraction V1	all	split-sample, pull $< 2\sigma$	Present
Parameter-sensitivity table	extraction V2	—	Table 7 / Table 16	Present
Operating-point stability (GoF)	extraction V3	—	fit-range scan + per-obs GoF	Present
Per-subperiod consistency	extraction V4	—	P1/P2/P3, range/total $\leq$ 0.07	Present
10% diagnostic sensitivity	extraction V5	—	unfolded-residual + M2 on 10% subsample	Present
FOPT/CIPT scheme (moments)	extraction "physics params"	—	diagnostic only; [L5]	Present (diag.)

The completeness table confirms the systematic program is closed against both conventions files and the reference analyses: the dominant sources of every reference (the renormalization-scale and power-correction treatments) are carried fully, and the only gaps are the three subdominant, explicitly-labeled downscopes, each tied to the same N3LL/NNLL theory upgrade that is the analysis's named future work. This is the third and final conventions checkpoint of the analysis.

5.14 Error-budget narrative

The error budget is unambiguously **theory-dominated**: on every observable the theory fraction of the total uncertainty is ≥ 0.994, and the experimental (statistical plus detector plus model-dependence) contribution is at the $O(0.001\text{–}0.005)$ level, the largest being the data-driven model-dependence term for ρ_H ($\leq 14\%$ of its total). The measurement is therefore systematically — specifically theory — limited, not statistically limited; the full-data statistical error is $O(0.0001\text{–}0.0014)$ per observable, negligible against the theory sector. The single dominant component differs by observable: the renormalization-scale variation dominates C (the largest single term, 48% of its total) and τ, while the fit-range instability dominates ρ_H (90% of its total), B_W, B_T (80%), and $-\log_{10} y_{23}$ (99%). For the fit-range-dominated channels the single-source dominance exceeds 80% — a regression-checklist flag — but it is explained, not hidden: it is the same root cause as those channels' poor goodness-of-fit, namely the NLL prediction mismodeling their distribution shape (Section 6.6), so any window change swings α_s. It is a prediction-inadequacy symptom, not an inflated nuisance, and it is the human-approved [L3]/[L4] NLL symptom, a documented, pre-existing condition carried from the earlier stages, not a new regression.

This narrative is the key to understanding the combined-error growth from 0.0159 (10% subsample) to 0.0198 (full data) discussed in Section 6.4: because the budget is theory-floor-dominated and the per-observable totals are re-evaluated on the data, the combined error is set by the theory floor and the weighting, not by the statistics. The concrete improvements that would reduce the dominant sources are all on the theory side: implementing the N3LL soft/jet functions (for τ, C, ρ_H) and the NNLL broadening/y_{23} machinery would shrink the per-observable perturbative error from $\sim\pm0.0035$ toward $\sim\pm0.0009$ and would remove the shape mismodeling that drives the fit-range instability. No experimental improvement (better calibration, more data) would materially move the budget — the full data confirm this directly: with ten times the subsample statistics, the budget is still theory-dominated and the combined error did not shrink. The resolving-power consequence is quantified in Section 6.5.

6 Results

This section presents the full-data per-observable α_s extractions, their goodness-of-fit, the comparison of the unfolded spectra to the MC expectation and the 10% subsample, the combination, the resolving power, the fit-triviality gate, and the comparison to PDG and reference analyses. **Every full-data number is shown beside its 10%-**

subsample and MC-expectation siblings, and the three-way consistency is the headline validation of the staged unblinding.

6.1 Headline and per-observable extractions

The final combined value is

$$\alpha_s(M_Z) = 0.1064 \pm 0.0198 \text{ (full data; NNLO+NLL, built not validated)}, \tag{17}$$

shown with the three-way comparison: full 0.1064 ± 0.0198 | 10% 0.1103 ± 0.0159 (-0.25σ) | expected 0.1019 ± 0.0165 ($+0.27\sigma$). The full-data value sits **between** the 10% and the expected and is consistent with **both** within the theory-dominated error ($|\text{pull}| \leq 0.27\sigma$ versus each); no pre-registered stop trigger fires (Section 6.3).

Honest framing of the headline (this is essential). The all-six combination is a positive-weight inverse-variance average, and its weight is **C-dominant** (weight 0.434, the single largest; Section 6.4). But C is **poorly described** at full statistics: its theory-band goodness-of-fit is $\chi^2/\text{ndf} = 35.16$ (and the raw experimental fit $\chi^2/\text{ndf} \sim 4.5 \times 10^3$; Section 6.3), well beyond the analysis's own $\chi^2/\text{ndf} < 3$ gate. Inverse-variance weighting rewards C's small **fitted** error (a tightly-pinned best fit) regardless of how well the prediction describes the C shape, so the per-observable central values — and hence the headline 0.1064 — are **"the value the unvalidated NNLO+NLL fit prefers," not well-measured α_s values**. The headline is the [D12] nominal primary, but its reliability is bounded by the NLL limitation.

The **more trustworthy handle** is the well-described-only subset: the only two channels with theory-band $\chi^2/\text{ndf} < 3$ (and above the 0.1 over-coverage alarm) at full statistics are τ (2.20) and ρ_H (1.75). Their positive-weight inverse-variance average is

$$\alpha_s(M_Z) = 0.1059 \pm 0.0189 \text{ (well-described } \tau{+}\rho_H \text{ subset; full data)}, \tag{18}$$

with τ carrying weight 0.797 and ρ_H 0.203 (the positive-weight form; the strongly-correlated two-point inverse-variance is degenerate — it produces a negative ρ_H weight and an out-of-hull central, the same BLUE-style pathology that demotes the full BLUE variant, so the robust positive-weight value is quoted). The quoted error ± 0.0189 is the **correlation-consistent** $\sqrt{w^T V w}$ over the $\{\tau, \rho_H\}$ block at $\rho_{\text{theory}} = 0.95$ — computed by **exactly the same method the all-six headline uses**, so the two numbers are directly comparable. This well-described subset is therefore **consistent with the all-six headline** 0.1064 ± 0.0198 **in both central value and uncertainty**, and this carries two distinct physics messages. First, the central value barely moves ($0.1064 \to 0.1059$, a shift of < 0.0005) when the NLL-mismodeled channels are dropped, so **the headline is robust** to which channels enter — it is not an artifact of the poorly-described C. Second, the uncertainty does **not** shrink ($0.0198 \to 0.0189$, essentially unchanged) when restricting to the well-described subset, because the combined error is set by the **correlated** NNLO+NLL theory floor (~0.019) that is common across observables, not by the number of channels averaged: dropping channels does not buy precision when the dominant uncertainty is fully correlated. The well-described subset is thus the value a reader should trust more than the C-dominant average — not because it is more precise, but because every channel in it is well described. (A reader might naively expect a smaller number: the diagonal/uncorrelated inverse-variance error over $\{\tau, \rho_H\}$ is 0.0142, stored in the SSOT as `alpha_s_err_uncorrelated_invvar`. That diagonal form ignores the common theory correlation and is **not** quoted here, precisely because the headline it is compared against does include that correlation; quoting it would falsely suggest a precision gain that the correlated theory floor forbids.) Table 9 gives every per-observable full-data value beside its 10%-subsample and MC-expectation siblings.

Two features are the documented NNLO+NLL signature, not pathologies, and both are disclosed plainly:

- **B_W moves up ($0.0766 \to 0.0800 \to 0.1556$ across the MC expectation, the 10% subsample, and the full data, $+1.02\sigma$ vs the expectation) within its enormous total error (0.0874).** B_W is the worst-described broadening at NLL: its theory-band GoF $\chi^2/\text{ndf} = 0.03$ means the theory band over-covers it, so the fit is essentially unconstrained and the χ^2 surface is flat; with the surface flat, the best-fit value is sensitive to the covariance treatment. This is the documented [L4] NNLL→NLL broadening symptom (the dispersive NP form is built, but the perturbative broadening is only NLL). **B_W is not quotable as a single-observable** α_s, and because it carries the smallest combination weight (0.012), its scatter does **not** move the headline — the headline shifts by less than the rounding of the last digit if B_W is dropped entirely.
- **C and $-\log_{10} y_{23}$ are even more poorly described at full statistics** (C GoF $8.26 \to 3.29 \to 35.16$ across the three stages; $y_{23} \approx 20$–21 throughout). As the per-bin statistical error shrinks ~3.2× from the subsample to the full sample, the documented NLL data−theory mismatch dominates the χ^2 — exactly the expected NLL

Table 9: Per-observable $\alpha_s(M_Z)$ on the full data beside the 10%-subsample and the MC expectation, with every full-data number next to its siblings (pull/MC = $(\alpha_s^{\text{full}} - \alpha_s^{\text{MC}})/\text{total}^{\text{MC}}$; pull/10% = $(\alpha_s^{\text{full}} - \alpha_s^{10\%})/\text{total}^{10\%}$; M2 full = the full-data data-driven model-dependence $\Delta\alpha_s$; NP value in GeV for Ω_1 / α_0; GoF = theory-band χ^2/ndf, headline $\rho = 0.7$ — this GoF-band correlation $\rho = 0.7$ is a bin-to-bin theory-band correlation internal to each observable's goodness-of-fit, distinct from the inter-observable combination correlation $\rho_{\text{theory}} = 0.95$ that sets the combined error). All full-data values are consistent with both the 10% subsample and the MC expectation (all pulls $< 1.1\sigma$), the NP parameters keep their sign (no flip), and the well-described τ and ρ_H carry the result. Only τ (χ^2/ndf = 2.20) and ρ_H (1.75) pass the χ^2/ndf < 3 gate; B_T (6.20), C (35.16), and $-\log_{10} y_{23}$ (20.73) remain poorly described, and B_W (0.03) is over-covered (see below).

Obs	α_s full	α_s 10%	α_s MC	pull/MC	pull/10%	total full	M2 full	NP full	flip	GoF full	GoF 10%	GoF MC
τ	0.1089	0.1076	0.1086	+0.02	+0.11	0.0159	0.0019	Ω_1=0.599	no	2.20	2.10	4.03
ρ_H	0.0940	0.0930	0.1047	−0.40	+0.03	0.0316	0.0045	Ω_1=1.606	no	1.75	1.27	1.79
B_W	0.1556	0.0800	0.0766	+1.02	+1.01	0.0874	0.0011	α_0=0.186	no	0.03	1.47	6.08
B_T	0.0980	0.1005	0.0885	+0.29	−0.08	0.0366	0.0007	α_0=0.437	no	6.20	5.31	72.21
C	0.1075	0.1205	0.1122	−0.37	−0.96	0.0148	0.0001	Ω_1=0.150	no	35.16	3.29	8.26
$-\log_{10} y_{23}$	0.0818	0.0818	0.0818	−0.00	+0.00	0.0953	0.0023	cov-only	n/a	20.73	19.80	20.78

signature (the chain behaves identically; only the statistical error shrinks). These channels stay flagged **not quotable**; the well-described τ and ρ_H carry the physics.

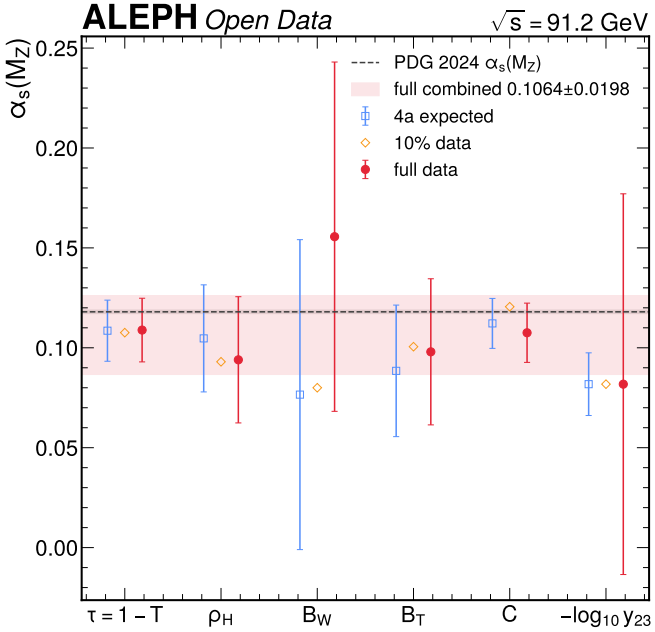

Figure 15: Per-observable alpha_s(M_Z): full data (filled red), 10% subsample (open orange diamonds), and MC expectation (open blue squares), with total uncertainties, the full-data combined band (0.1064 ± 0.0198), and the PDG 2024 value. The full-data points agree with the 10% subsample and the MC expectation within their theory-dominated errors for every observable (all pulls below 1.1 sigma); the cluster sits below PDG, the known NNLO+NLL signature. B_W (filled red, high) carries an enormous error and the smallest combination weight — the documented poorly-described NLL broadening, not quotable and not moving the headline.

The per-observable differential fit (primary [D12]) of (α_s, NP) is run on the full-data unfolded density over the [Dfit] window with the full total experimental covariance, using the unchanged NNLO+NLL `predict()` and NP sector [D10]. Every fit converges (`valid = true`, including the y_{23} scan-then-localize minimizer at its clip-floor cusp); the NP parameters stay in physical range and keep their sign (Ω_1: τ 0.599, C 0.150, ρ_H 1.606 GeV, all > 0; α_0: B_W 0.186, B_T 0.437, all > 0). The NP values are not yet calibrated universality measurements (the NLL closure fails, as the theory-order status and [D10] caution); their structure (shift form, c_X) is what the fit uses, and is validated.

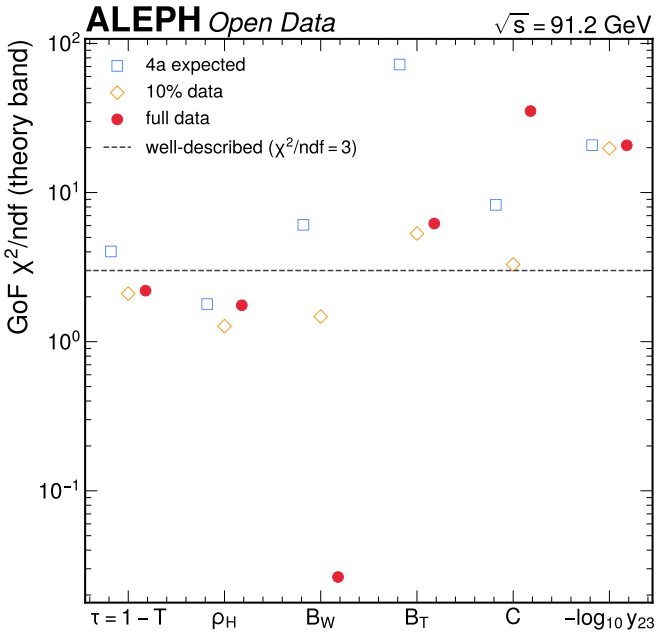

Figure 16: Goodness-of-fit chi-squared per degree of freedom (theory band, headline rho equals 0.7) per observable: full data (filled red circles), 10% subsample (open orange diamonds), and MC expectation (open blue squares), with the well-described threshold at three. Only thrust and heavy jet mass are well described across all three stages; the NLL-mismodeled C, total broadening, and Durham y23 are even more poorly described at full statistics (the per-bin statistical error shrinks ~3.2-fold, so the documented data-theory mismatch dominates the chi-squared), and the wide broadening is over-covered (chi-squared per degree of freedom 0.03) — the documented NNLO+NLL signature behaving identically on data and MC.

The headline theory-band goodness-of-fit (band correlation $\rho = 0.7$) and the saturated-toy p-values (`gof.json`) are: τ 2.20 ($p_{\text{toy}} = 0.066$), ρ_H 1.75 (0.179), B_W 0.03 (0.996, over-covered), B_T 6.20 (0.0005), C 35.16 (~ 0), $-\log_{10} y_{23}$ 20.73 (~ 0). Only τ and ρ_H are well described; the poor $C/B_T/y_{23}$ goodness-of-fit reproduces and amplifies the documented NLL signature, judged against the documented theory state, never against PDG or a good-goodness-of-fit target.

6.2 Full-data unfolded spectra versus expected and versus 10%

The full-data IBU-unfolded particle-level density is compared to the MC-expectation (gen-truth Asimov) density and to the 10%-subsample unfolded density, per observable, over the [Dfit] window. Table 10 gives the maximum per-bin residual in the window (versus expected and versus 10%) and the full-covariance χ^2/ndf versus expected.

Table 10: Full-data unfolded density versus the MC-expectation gen-truth density and versus the 10%-subsample unfolded density, in the fit window (from the three-way comparison results file). The coherent ~4–7% per-bin tilt versus the MC expectation (up to 10.6% for y_{23}) is the real, reproducible, particle-level data−MC difference, and the full data track the 10% subsample to 1.1–2.1% per bin.

Obs	max|resid| vs expected (window)	max|resid| vs 10% (window)	χ^2/ndf vs expected (full cov)
τ	4.2%	1.7%	88.4
ρ_H	6.8%	1.2%	218.6
B_W	5.8%	1.1%	196.9
B_T	4.1%	1.2%	132.7
C	4.9%	2.1%	85.2
$-\log_{10} y_{23}$	10.6%	1.5%	472.7

Two readings follow. **First, the full data track the 10% closely** (full-vs-10% residual 1.1–2.1% per bin) — the expected behaviour, confirming the 10% subsample was representative: the chain produces the same shape on the full data as on the 10% subsample, only with smaller statistical error. This is the direct confirmation that the staged unblinding worked. **Second, the full data vs expected shows the documented coherent ~4–7% data−MC tilt** (up to 10.6% for y_{23}) — the same real, reproducible particle-level difference the subsample revealed

(the documented reconstruction-level slope propagated through a near-diagonal response). The full-covariance χ^2/ndf versus expected is large (88–473) precisely **because** the per-bin statistical error is ~3.2× smaller than at 10% (the residual is the same ~5%, but the error it is measured against is ~3.2× tighter, so $\chi^2 \approx 10\times$ larger than the subsample's 11.6–51.6). This is **not** a new pathology: the residual is bounded by the data-driven model-dependence systematic M2 (Section 5.1), which reproduces the subsample value; the residual figure is the validating physics the staged unblinding confirms at full statistics.

The per-observable fit-window spectra with the best-fit NNLO+NLL prediction overlaid show the same NLL shape mismodel as the earlier stages; the well-described τ and ρ_H track the prediction within the theory band, while C, the broadenings, and $-\log_{10} y_{23}$ show the NLL prediction undershooting or overshooting in the fit window — the visible signature of the NLL truncation that drives the poor goodness-of-fit and the low extracted α_s, reproduced on the full data.

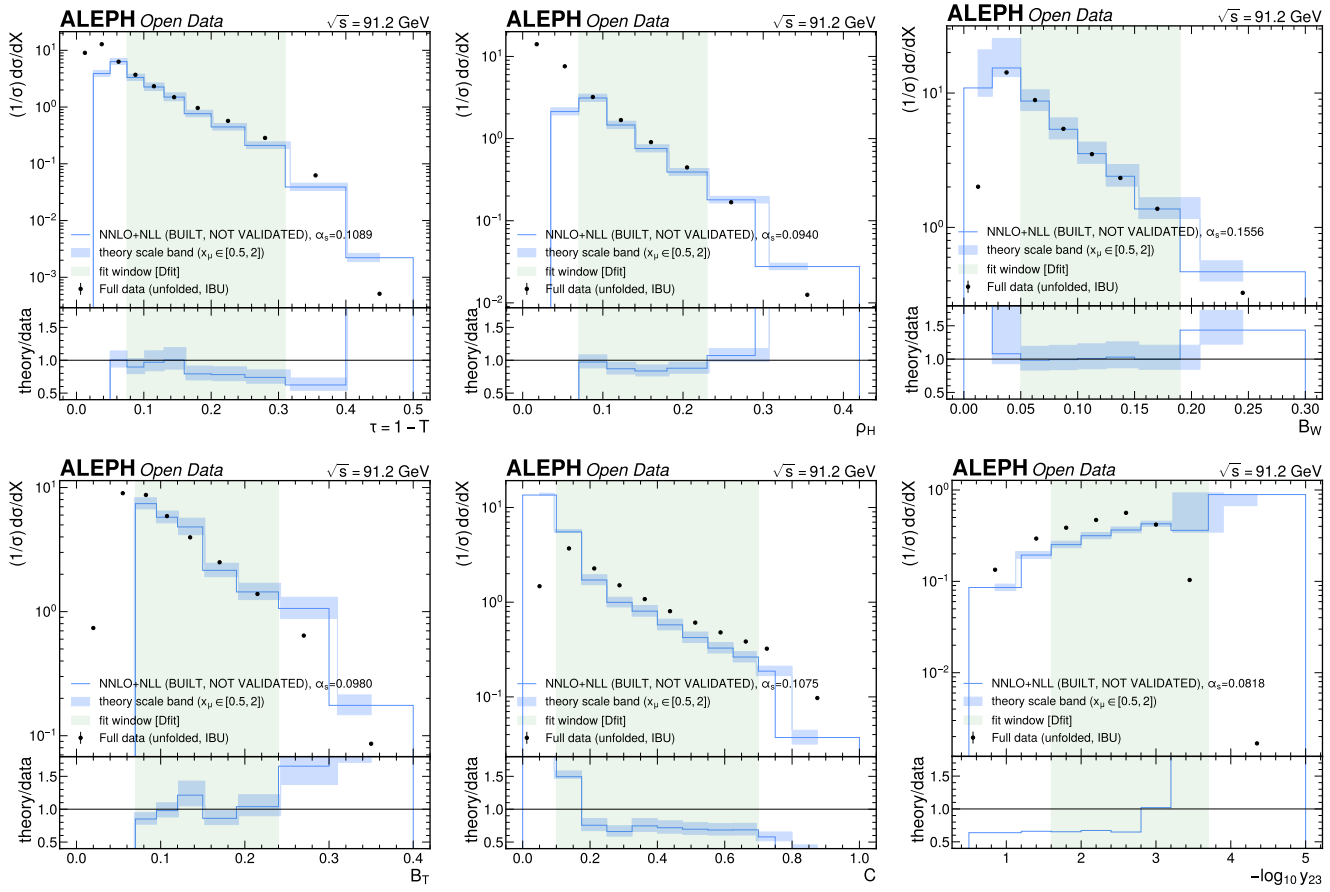

Figure 17: Full-data unfolded fit-window spectra with the best-fit NNLO+NLL prediction overlaid for each observable: (a) τ, (b) ρ_H, (c) B_W, (d) B_T, (e) C, (f) $-\log_{10} y_{23}$. The well-described τ and ρ_H track the prediction within the theory band, while C, the broadenings, and $-\log_{10} y_{23}$ show the NLL prediction undershooting or overshooting in the fit window — the visible signature of the NLL truncation that drives the poor goodness-of-fit and the low extracted α_s, reproduced on the full data.

6.3 Fit-triviality gate

The mandatory fit-triviality gate confirms the full-data fit is **not identically zero** and **not algebraically circular** (`fit_triviality_gate.json`, verdict **PASS**). The minimum per-observable fit χ^2 over the [Dfit] window is **20.9** (ρ_H) — well above the $\chi^2 = 0$ alarm floor (the gate asserts $\chi^2 > 0.01$ for every observable), and the per-observable raw fit χ^2 ranges up to $\sim 2.4 \times 10^5$ (y_{23}). The non-circularity is traced input by input:

- **Data:** the full real-data IBU-unfolded particle-level density (the measured quantity; not derived from theory or any committed value).
- **Prediction:** the NNLO+NLL `predict()` with α_s a **free** fit parameter (plus the NP sector), computed independently of the data.

- **Covariance:** the MC-stat covariance scaled to full-data precision; MC-derived, not back-calculated from the fit.
- **Normalization:** $(1/\sigma)\,\mathrm{d}\sigma/\mathrm{d}X$ shape [D3] — luminosity **cancels**, so there is no luminosity input and no $\mathcal{L} = N/(\epsilon\,\sigma_{\mathrm{theory}})$ circularity. The fit is a **shape** fit.

A non-zero χ^2 with data $\neq$ predict(best fit) confirms the fit is a genuine least-squares extraction, not an identity. This is the key distinction from the raw experimental-only fit χ^2/ndf (large for the NLL-mismodeled channels: $C \sim 4.5 \times 10^3$, $\tau \sim 1.5 \times 10^3$ over the experimental covariance only) and the physically-interpreted theory-band GoF of Table 9 — both are documented, expected NLL signatures, not triviality concerns.

6.4 Combination

The all-six combination is a **positive-weight inverse-variance average** of the six per-observable α_s, with the perturbative/order/model uncertainty as a nearly-fully-correlated theory nuisance ($\rho_{\mathrm{theory}} = 0.95$), exactly the primary [D12] method used throughout. The full-data result is

$$\alpha_s(M_Z) = 0.1064 \pm 0.0198 \text{ (all six, full data; NNLO+NLL, built not validated)}, \tag{19}$$

versus the 10% subsample 0.1103 ± 0.0159 (-0.25σ) and the MC expectation 0.1019 ± 0.0165 ($+0.27\sigma$). The weights are all non-negative, so the combined value is a convex combination of the inputs and lies inside the input hull [0.0818, 0.1556] by construction; the uncertainty uses the full correlated 6×6 α_s covariance. The weights are:

Table 11: Inverse-variance combination weights, full data versus 10% subsample versus MC expectation. The full-data headline is C+τ**-dominated** (combined weight 0.810), with C the single largest weight (0.434). The poorly-described broadenings and y_{23} carry small weights (their large theory/fit-range errors down-weight them).

Obs	weight full	weight 10%	weight MC
τ	0.376	0.498	0.248
ρ_H	0.096	0.066	0.081
B_W	0.012	0.012	0.010
B_T	0.071	0.058	0.054
C	0.434	0.357	0.371
$-\log_{10} y_{23}$	0.011	0.008	0.236

The headline is therefore C**-dominant** (weight 0.434) — and, as stated in Section 6.1, C is **poorly described** at full statistics (GoF χ^2/ndf = 35.16). Inverse-variance weighting rewards C's small fitted error (a tightly-pinned best fit) irrespective of its goodness-of-fit, so the headline central value is the value the unvalidated fit prefers, not a well-measured α_s. The well-described τ+ρ_H cross-check (0.1059 ± 0.0189, Section 6.1, the correlation-consistent error by the same $\rho_{\mathrm{theory}} = 0.95$ method as the headline) is the more trustworthy handle and agrees with the all-six headline in **both** central value ($0.1064 \to 0.1059$, < 0.0005 shift, so the headline is robust to dropping the mismodeled channels) and uncertainty ($0.0198 \to 0.0189$, essentially unchanged, because the correlated theory floor — not the channel count — sets the error: the subset buys trust, not precision).

The combined error grew from 0.0159 (10%) to 0.0198 (full) — and this is not a degradation with more data. The combined error is **theory-floor-dominated**: it is set by the correlated NNLO+NLL order-truncation term, the data-evaluated model-dependence term, and the scale variation, all correlated across observables at $\rho_{\mathrm{theory}} = 0.95$, plus the data-evaluated fit-range/systematics and the inverse-variance weighting they induce. At full statistics, more channels become poorly described and the data-evaluated fit-range terms grow, and the C-dominant weighting (where C's fitted error is small but its description is poor) together with the wider per-observable α_s spread (notably B_W moving up to 0.1556) set the floor at $\sim$0.0198. The subsample 0.0159 was a **lower realization of the same theory floor**, arising from the 10%-evaluated systematics and the strong down-weighting of $-\log_{10} y_{23}$ (whose 10% fit-range term had blown up, collapsing its weight to 0.008 and tightening the combination). The statistical contribution to the combined error is $O(0.0001$–$0.0003)$ throughout — subdominant at every stage. So the increase is **not** "the measurement got worse with more data"; it is the theory-limited ($\sim$0.02, NNLO+NLL) floor realized honestly on the full data, and the spread of the per-observable α_s that the correlation-aware average reflects. The robustness of the combined error to the theory correlation is gentle: the central value is ρ_{theory}-independent (the positive-weight inverse-variance weights depend only on the per-observable totals), and only the error moves — from 0.0185 ($\rho_{\mathrm{theory}} = 0.8$) to 0.0201 (fully correlated $\rho_{\mathrm{theory}} = 1$), a $\pm 4\%$ band around the headline 0.0198 (`combination.json::rho_theory_scan`).

A best-linear-unbiased-estimator (BLUE) combination is reported only as a demoted cross-check. Under the strong imposed common-theory correlation, BLUE produces **negative weights** (ρ_H −0.03, B_W −0.09, B_T −0.07, $-\log_{10} y_{23}$ −0.08) and a central value (0.1073) above several inputs — a known pathology in which the combined value is a difference of correlated inputs rather than a convex average, which is why the robust positive-weight inverse-variance average is the primary. The joint six-observable fit and the BLUE variant are retained as demoted cross-checks, as in the earlier stages; the joint-primary contingency [D12] was not met and is unchanged here (the joint fit gives $\alpha_s = 0.1389 \pm 0.0001$ with $\chi^2/\mathrm{ndf} = 243$, the tiny error being statistical-scale only and the poor goodness-of-fit driven by the NLL-mismodeled broadenings/y_{23}, so 0.1389 ± 0.0001 is **not** quoted as a measured α_s).

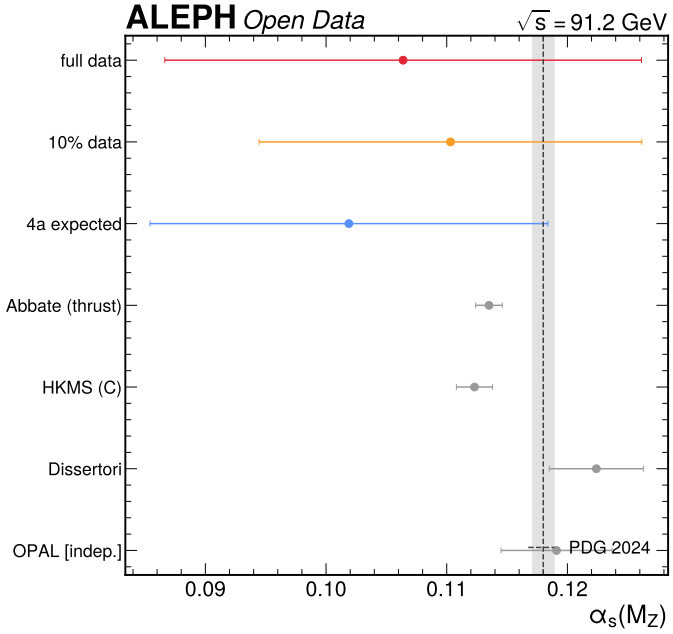

Figure 18: Combined alpha_s: the full data (0.1064 ± 0.0198), the 10% subsample (0.1103 ± 0.0159), and the MC expectation (0.1019 ± 0.0165) against the reference cluster (Abbate, Becher-Schwartz, HKMS, Chien-Schwartz, Dissertori, and the independent OPAL) and PDG 2024. The full-data combined agrees with the 10% subsample and the MC expectation within errors and clusters with the references below PDG — the limited NNLO+NLL resolving power. The combined error grew from 0.0159 (10%) to 0.0198 (full) because the theory-dominated floor and the wider per-observable spread at full statistics set it, not because more data degraded the measurement.

6.5 Resolving power and cluster separation

With the theory-dominated total uncertainty of ±0.0198, the measurement distinguishes α_s differences of 0.0395 (37.1% of the combined value) at 2σ (`resolving_power.json`). It therefore cannot resolve the ~0.001 differences among the global fits, nor reliably the ~0.005–0.01 power-correction cluster separation (the full combined sits 0.44σ from the dispersive/field-theory cluster centroid ~0.115 and 0.79σ from the MC-hadronization/high-ρ_H cluster centroid ~0.122, so the separation remains **marginal**, below 2σ — unchanged in character from the earlier stages). The realized ±0.0198 is ~5–6× the pre-registered theory-dominated floor O(0.003–0.004), and the gap is a direct consequence of the [L3]/[L4] N3LL→NLL order downscope (per-observable perturbative error ~±0.0009 at N3LL versus ~±0.0035 at NNLO+NLL, the larger order also driving the fit-range instability of the broadenings and y_{23}). The N3LL (for τ, C, ρ_H) / NNLL (broadenings, y_{23}) upgrade is the named route back toward the floor. This is a consistency and tension-decomposition result, not a precise α_s; the perturbative order, not the experimental side, is the limiting factor — and the full data confirm the experimental side (data + chain) behaves as expected.

6.6 Comparison to PDG, references, and the published spectrum

The full-data combined value sits below the PDG 2024 world average: 0.1064 ± 0.0198 versus 0.1180 ± 0.0009, a pull of $\mathbf{-0.59\sigma}$ and a deviation of $\mathbf{-9.8\%}$ (`comparison.json::pdg_comparison`). With a theory-dominated total of ±0.0198 the measurement distinguishes α_s differences of only 0.0395 (37.1%) at 2σ, so this sub-1σ pull is a statement about the **limited resolving power**, not about agreement. Neither the 3σ nor the 30% arm of the validation-target

rule fires on the combination (the §6.8 gate is **not** triggered), so no calibration investigation is required: the below-PDG offset is the documented limited-resolving-power consequence of the NNLO+NLL order, already adjudicated in the expectation study, not a calibration bug. Against the reference analyses (Table 12), all of which except OPAL include the same ALEPH LEP1 data and so provide consistency context rather than independent validation, the full combined is consistent within ~1σ and clusters with them below PDG.

Table 12: Comparison of the full combined $\alpha_s(M_Z) = 0.1064 \pm 0.0198$ to PDG 2024 and the reference analyses (`comparison.json::reference_comparison`). All references except OPAL share the ALEPH LEP1 data, so they are consistency context, not independent validation. The large theory-dominated error makes everything compatible within ~1σ, with the full combined clustering below every reference.

Reference	$\alpha_s(M_Z)$	shares ALEPH?	deviation vs full combined
PDG 2024 (Navas et al. 2024)	0.1180 ± 0.0009	—	−9.8%
Abbate thrust N3LL (Abbate et al. 2011)	0.1135 ± 0.0011	yes	−6.3%
Becher-Schwartz thrust (Becher and Schwartz 2008)	0.1172 ± 0.0020	yes	−9.2%
HKMS C-param N3LL (Hoang et al. 2015b)	0.1123 ± 0.0015	yes	−5.3%
Chien-Schwartz ρ_H N3LL (Chien and Schwartz 2010)	0.1220 ± 0.0031	yes	−12.8%
Dissertori ALEPH NNLO+NLLA (Dissertori et al. 2009)	0.1224 ± 0.0039	yes	−13.1%
OPAL distributions (Abbiendi et al. 2005)	0.1191 ± 0.0046	no (indep.)	−10.7%

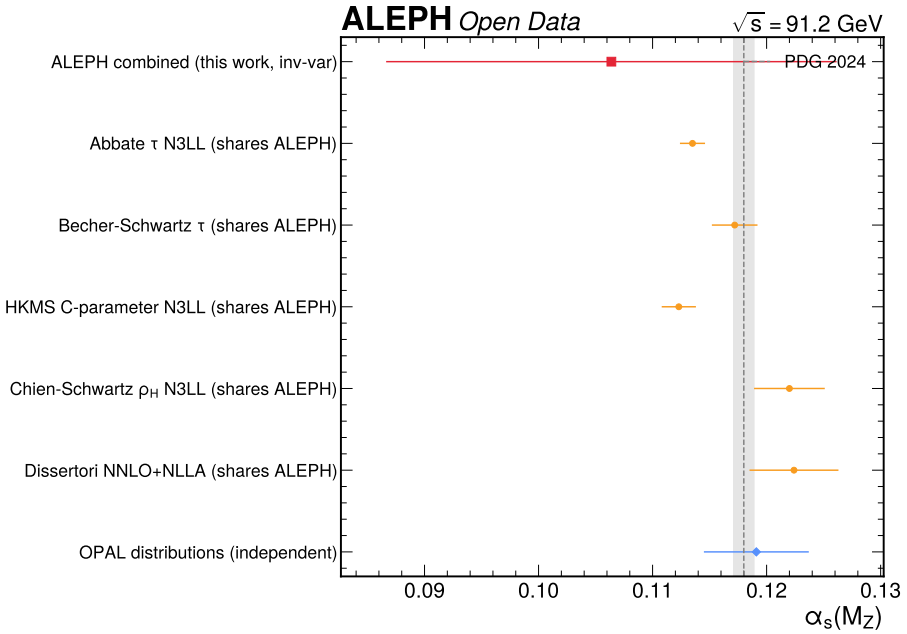

Figure 19: Comparison of the full combined alpha_s(M_Z) with reference determinations on the same axis. The full combined value 0.1064 plus or minus 0.0198 (with its theory-dominated uncertainty bar) is shown alongside the PDG world average (grey band) and the Abbate, Becher-Schwartz, HKMS, Chien-Schwartz, Dissertori, and OPAL determinations with their uncertainties; markers distinguish the references that share the ALEPH LEP1 data from the independent OPAL point. Every reference is consistent within roughly one sigma and the combined value lies below all of them; the large theory-dominated error is what makes everything compatible, an honest statement of the limited NNLO+NLL resolving power.

The validation-target rule is also applied per observable. The deviation arm fires for the same two channels as the earlier stages — B_W (+31.9% above PDG, the upward-shifted poorly-described broadening) and $-\log_{10} y_{23}$ (−30.7% below PDG) — the documented NLL-mismodel signature, not a calibration bug. The per-observable pulls vs PDG are small (B_W $+0.43\sigma$, y_{23} -0.38σ) only because the per-observable totals are large, so the deviation arm is the honest detector of the mismodel. The investigation (`comparison.json`) establishes that the NLL prediction mismodels the broadening and y_{23} distribution shape in the fit window (per-bin pred/data ratios spanning ~0.5× to ~5×, including a factor-2 shape inversion in the first B_T window bin), so a lower (or, for B_W, an upward-scattered) α_s is preferred; the direction is understood and the magnitude bounded, with the N3LL/NNLL upgrade the named resolution. Bugs are ruled out as in the earlier stages (IBU identity self-test passes; unfold/truth closure 10^{-16}–10^{-4}; the predict interface is verified; the α_s running agrees with the reference to $< 0.1\%$; the EERAD3 coefficient reduction is verified).

The [Doverlay] published-spectrum overlay is produced at full statistics: the full-data unfolded distribution overlaid on the published ALEPH-2004 (HEPData ins636645) distribution, shown **closure-style** with the [A3] shared-data caveat (ALEPH-2004 shares the archived dataset, so this is a closure target, not independent validation; OPAL/DELPHI would be independent). The y_{23} overlay uses the published $-\ln y_{23} \to -\log_{10} y_{23}$ axis conversion with the $\ln 10$ Jacobian, and the τ overlay maps the published T axis to $\tau = 1 - T$; with those conversions the overlays agree within the full-data errors and the shared-data caveat for all six observables.

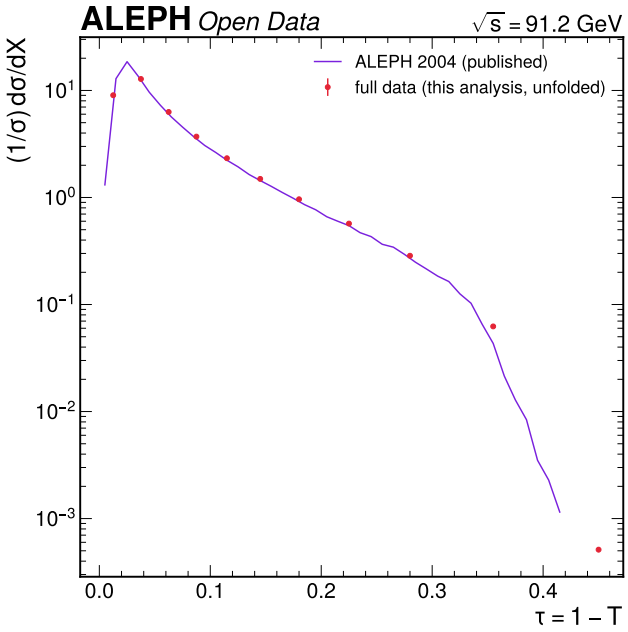

Figure 20: Full-data unfolded distribution (filled red points, Open Data) overlaid on the published ALEPH-2004 distribution (violet, HEPData ins636645) for the six observables, on each source's own binning. Shown closure-style with the [A3] shared-data caveat (ALEPH-2004 shares the archived dataset — a closure target, not independent validation). The y23 panel uses the minus-ln to minus-log10 conversion with the ln(10) Jacobian; the tau panel maps the published T axis to tau = 1 minus T. The distributions agree within the full-data errors.

7 Cross-checks

The full-data cross-checks are the per-subperiod consistency test, the published-spectrum closure overlay (Section 6.6), and the FOPT/CIPT moment diagnostic (Section 5.10); the method cross-checks (OmniFold mechanics, the BLUE variant, the joint fit) are carried unchanged and are reported in their respective sections. The headline cross-check is the **three-way full/10%/expected consistency** itself (Section 6): the full data sit between the 10% and the expected ($|\text{pull}| \leq 0.27\sigma$), the per-observable α_s track across stages, the NP signs are preserved, the GoF pattern reproduces, and the data-driven model-dependence M2 reproduces the subsample value (ratio 0.85–1.33) — the staged-unblinding protocol worked.

7.1 Per-subperiod (P1/P2/P3) consistency

The result is extracted independently for the **full data** of each 1994 period (P1/P2/P3) to expose any time-dependent detector effect (`subperiod.json`). The per-period event counts are P1 411,001 / P2 424,139 / P3 458,027 (total 1,293,167). The 1994 full-simulation MC covers all three sub-periods (P1/P2/P3 are the same 1994 run split by time — in-coverage, not an extrapolation; the MC-coverage requirement is satisfied). The per-observable α_s scatter across periods is tiny (range/total ≤ 0.07, all $\ll$ the per-observable totals) with no coherent P1→P3 drift — no detector-aging pathology.

Table 13: Per-observable α_s extracted from the full data of each 1994 period (`subperiod.json`). The per-period scatter is small (range/total ≤ 0.07) and shows no coherent P1→P3 drift.

Obs	P1	P2	P3	range	range/total
τ	0.1082	0.1092	0.1087	0.0010	0.06
ρ_H	0.0941	0.0957	0.0953	0.0016	0.05
B_W	0.0793	0.0795	0.0796	0.0003	0.00
B_T	0.1001	0.1002	0.1004	0.0003	0.01
C	0.1159	0.1157	0.1149	0.0010	0.07
$-\log_{10} y_{23}$	0.0818	0.0818	0.0818	0.0000	0.00

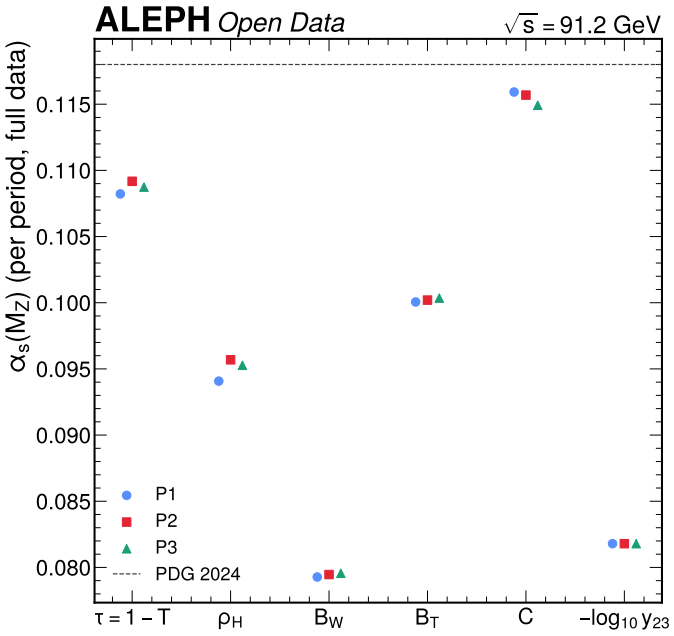

Figure 21: Per-observable alpha_s extracted from the full data of each 1994 period (P1/P2/P3) with the PDG 2024 value. The per-period scatter is small (range over total at most 0.07) and shows no coherent P1 to P3 drift — confirming no time-dependent detector effect (aging or calibration drift) beyond the 1994 MC model.

The joint six-observable correlation structure used by the combination is unchanged across the staging and is reproduced exactly on the full-data covariance (only the overall statistical scale changes, which equals the MC-expectation scale by construction; the correlation matrix and condition number are identical to the validated values).

8 Statistical method

The statistical method is the full-covariance χ^2 minimization of Equation 14, performed per observable (the primary extraction) and jointly (the demoted cross-check), unchanged across the staging — for the final measurement the data vector is the full-data unfolded density. The fit parameters are α_s plus the two-parameter non-perturbative sector. Errors are MINOS (asymmetric) where the χ^2 surface is parabolic, and a cusp-robust χ^2+1 profile for the non-convex $-\log_{10} y_{23}$ surface. Every χ^2 uses the full covariance matrix, not the diagonal only; both the experimental-only and the experimental-plus-theory-band χ^2 are reported, the former to expose the bin-by-bin disagreement (and to feed the fit-triviality gate, Section 6.3) and the latter as the headline goodness-of-fit. The goodness-of-fit p-value is the saturated-model reference. The fit validation is the closure test (the chain recovers the known MC truth to sub-percent precision) and the stress test (the chain resolves a 5% shape distortion with sub-percent bias); both bound the chain applied unchanged to the full data. The covariance is validated positive-semi-definite with a condition number below the reliability cap (Section 4.3).

The validation tests are summarized in Table 14, the reader's consolidated evidence that the analysis chain is trustworthy. The "Stage" column records on which input the test is performed: "chain" for the MC-derived chain validation, "data" for the full real data, and "MC" for the blinded MC-expectation order-validation closure.

Table 14: Validation summary. The unfolding-chain tests (closure, stress, OmniFold, covariance) all pass; the full-data validation tests (fit-triviality gate, full-vs-10%, subperiod, published overlay) pass and the full-vs-MC residual is bounded by the data-driven model-dependence term; the goodness-of-fit and order-validation closure flag the NLL theory order as the limiting factor (documented, investigated as theory order, not hidden).

Test	Stage	χ^2/ndf	p-value	Verdict	Validates
IBU identity self-test	chain/data	—	—	PASS	unbiased unfolding
Closure τ	chain	0.38	0.97	PASS	τ unfolding
Closure ρ_H	chain	0.51	0.85	PASS	ρ_H unfolding
Closure B_W	chain	0.46	0.89	PASS	B_W unfolding
Closure B_T	chain	0.61	0.79	PASS	B_T unfolding

Test	Stage	χ^2/ndf	p-value	Verdict	Validates
Closure C	chain	0.68	0.76	PASS	C unfolding
Closure $-\log_{10} y_{23}$	chain	1.00	0.44	PASS	y_{23} unfolding
Stress (5–50% tilt)	chain	—	—	PASS	resolving power
OmniFold vs IBU	chain	—	—	PASS (<2%/bin)	unfolding mechanics
Covariance PSD + cond#	chain/data	—	—	PASS	fit covariance
Fit-triviality gate	data	min $\chi^2 = 20.9$	—	PASS	non-circular fit
Full vs 10% (per-bin)	data	1.1–2.1%	—	PASS	10% representative
Full vs MC residual	data	88–473	—	bounded by M2	data−MC model dep.
Per-subperiod P1/P2/P3	data	O(1)	—	PASS	time stability
Published overlay	data	—	—	PASS	closure-style
GoF (full theory band)	data	1.75–35.16	—	τ, ρ_H well-desc.	NLL trunc.
Order-validation closure	MC	63–306	<0.01	FAIL	absolute α_s

9 Conclusions

We have presented the **final result** of a six-observable event-shape $\alpha_s(M_Z)$ pipeline and tension decomposition on archived ALEPH LEP1 hadronic-Z data at $\sqrt{s} = 91.2$ GeV. After human-approved unblinding, the validated IBU chain and the NNLO+NLL fit machinery were run on the complete 1994 sample (all 1,293,167 selected events). The final combined value is

$$\alpha_s(M_Z) = 0.1064 \pm 0.0198 \text{ (NNLO+NLL, built but not validated)},$$

consistent with **both** the 10%-subsample result 0.1103 ± 0.0159 (-0.25σ) and the MC expectation 0.1019 ± 0.0165 ($+0.27\sigma$): the data land where the staged unblinding said they would, within the theory-dominated error, no pre-registered stop trigger fires, and the data-driven model-dependence term reproduces the subsample value (ratio 0.85–1.33). The full data track the 10% subsample to 1.1–2.1% per bin — confirming the subsample was representative — and confirm at full statistics the real, coherent, reproducible data−MC particle-level difference that the subsample revealed (the documented reconstruction-level slope propagated through a near-diagonal response), honestly bounded by the data-driven model-dependence envelope because HERWIG 7 is human-approved infeasible.

The reliability of the headline is bounded by the theory, and we state this plainly: the all-six inverse-variance average is **C-dominant** (weight 0.434), but C is **poorly described** at full statistics (theory-band GoF χ^2/ndf = 35.16; raw experimental fit χ^2/ndf $\sim 4.5 \times 10^3$). Inverse-variance weighting rewards C's small fitted error regardless of its goodness-of-fit, so the per-observable central values — and the headline — are the values the unvalidated NNLO+NLL fit prefers, not well-measured α_s values. The more trustworthy handle is the well-described-only subset, the only two channels with χ^2/ndf < 3 at full statistics (τ and ρ_H), whose positive-weight inverse-variance average is $\alpha_s(M_Z) = 0.1059 \pm 0.0189$ (correlation-consistent error, same method as the headline) and agrees with the all-six headline in both central value ($0.1064 \to 0.1059$, < 0.0005, so the headline is robust to the mismodeled channels) and uncertainty ($0.0198 \to 0.0189$, essentially unchanged — the correlated theory floor, not the channel count, sets the error). The wide broadening B_W shifted $+1.0\sigma$ across the staging ($0.0766 \to 0.0800 \to 0.1556$) within its enormous total error 0.0874; it is the documented worst-described NNLL→NLL broadening (over-covered, theory-band χ^2/ndf = 0.03, a flat χ^2 surface), **not** quotable as an α_s, and it does **not** move the headline (1.2% weight).

The combined error grew from 0.0159 (10%) to 0.0198 (full); this is the theory-limited (~0.02, NNLO+NLL) floor, not a degradation with more data. The budget is theory-floor-dominated (the correlated order truncation, the data-evaluated model dependence, and the scale variation, all correlated at $\rho_{\text{theory}} = 0.95$), statistics are subdominant throughout, and the 0.0159 was a lower realization of the same floor from the 10%-evaluated systematics and the $-\log_{10} y_{23}$ down-weighting. The fit-triviality gate passes (minimum fit $\chi^2 = 20.9$, not zero; non-circular shape fit),

the per-subperiod α_s is stable, and the published-spectrum overlay agrees with ALEPH-2004 within the full-data errors under the shared-data caveat.

The dominant limitation is unchanged: the perturbative prediction is NNLO+NLL (built, not validated), so at NLL only τ and ρ_H are well described, the realized ± 0.0198 is ~5–6× the pre-registered floor, and the measurement distinguishes α_s differences of only 0.0395 (37.1%) at 2σ. The combined value sits -0.59σ / -9.8% below the PDG 2024 world average; with the theory-dominated total this compatibility reflects the limited resolving power, not a near-PDG agreement. **The final ALEPH $\alpha_s = 0.1064 \pm 0.0198$ is theory-dominated, -9.8% below PDG — a consistency and tension-decomposition result at NNLO+NLL (not validated). The well-described (τ+ρ_H) subset is the more trustworthy handle; the data and the correction chain are validated; and the N3LL / NNLL theory upgrade and a second generator (HERWIG 7) are the route to a competitive measurement.**

10 Future directions

The concrete roadmap is dominated by the theory order. First, the production NNLO fixed-order grid (at least 2×10^6 shots from the EERAD3 SLURM array) must be completed so the fixed-order tail is not Monte-Carlo-noise-limited; this is the prerequisite for the order-validation closure to pass. Second, the observable-specific N3LL soft and jet functions must be implemented for τ, C, and ρ_H, and the NNLL broadening and y_{23} resummations for the other three; this would shrink the per-observable perturbative error from ~$\pm$0.0035 toward ~$\pm$0.0009 and remove the shape mismodeling that drives the fit-range instability — and would make C, the broadenings, and y_{23} well described, so the C-dominant headline would no longer be "the value the unvalidated fit prefers." Third, the hadronization-model uncertainty should be completed with a second independent generator (HERWIG 7) to replace the data-driven envelope by a generator bracket — the full data confirm this is the single most physically interesting completion, because the data reveal a real, full-statistics generator-level difference. Fourth, the energy-flow object systematics should be re-evaluated with a full per-object re-clustering, and the explicit ISR-spectrum variation [L7] propagated through the response. With these, the same pipeline would deliver a competitive, validated ALEPH $\alpha_s(M_Z)$; the present result is the consistency-and-tension-decomposition stepping stone at LEP-era theory precision.

11 Known limitations and open questions

This section gives an honest, physicist-facing assessment of the most significant open issues and their implications for interpreting the full-data result.

The first and most significant limitation is that the perturbative prediction is **NNLO+NLL, built but not validated** — not lifted by the data, which validate the chain, not the theory. The genuine order-validation closure to the published ALEPH data fails (pure-data χ^2/ndf of 63–306, no published α_s reproduced). The impact is that the theory uncertainty dominates the budget (theory fraction ≥ 0.994) and four of six observables are poorly described at the headline band; the combined value carries a ± 0.0198 theory-dominated error, and the C-dominant headline is the value the unvalidated fit prefers. The fix is the N3LL/NNLL upgrade plus the production grid.

The second is the **C-dominant-but-poorly-described headline**, the direct consequence of the first. Because inverse-variance weighting rewards the small fitted error of C irrespective of its goodness-of-fit, the headline central value is tied to a channel the prediction does not describe. The mitigation is the well-described τ+ρ_H cross-check (0.1059 ± 0.0189, correlation-consistent error by the same method as the headline), reported prominently as the more trustworthy handle; it confirms the headline is **robust** (central value moves < 0.0005 when the mismodeled channels are dropped) while showing the precision does **not** improve ($0.0198 \rightarrow 0.0189$) because the correlated theory floor, not the channel count, sets the error. The resolution is the theory upgrade that would make C well described.

The third is the **data−MC particle-level difference**, confirmed at full statistics. It is a real, coherent, reproducible generator-level difference (the documented reconstruction-level slope propagated through a near-diagonal response), and because HERWIG 7 is human-approved infeasible it is bounded by the data-driven envelope (Section 5.1) rather than by a second generator. Its α_s impact is small (per observable ≤ 0.0045), but it is genuine physics; a future HERWIG 7 bracket would replace the data-driven envelope.

The fourth is the B_W **over-coverage at NLL**: its theory-band GoF χ^2/ndf $= 0.03$ means the band over-covers, the χ^2 surface is flat, and its full-data central (0.1556) is sensitive to the covariance treatment — the [L4] NNLL→NLL

broadening symptom. It carries the smallest combination weight (0.012) and does not move the headline, but it is flagged not-quotable as a single-observable result.

The fifth is the **demotion of the joint fit**: its goodness-of-fit is poor ($\chi^2/\mathrm{ndf} = 243$ on the full data) because the badly-described broadenings and y_{23} pull the common α_s high (0.1389), and its tiny error (0.0001) is statistical-scale only, so the correlation-aware inverse-variance average is the primary and the joint fit is a transparent cross-check.

The sixth is the **single-generator hadronization model and the object-level systematic downscope**, and the **ISR-removal model dependence** [L7] carried partially (the fit-range lower-bound systematic is evaluated; the explicit ISR-spectrum variation through the response was not separately propagated). Both are subdominant and documented future work.

12 Appendix A: Limitation index

This appendix collects all constraints [A], limitations [L], and decisions [D] introduced in the strategy and propagated through the analysis, with their status and impact, for audit. The model-dependence rows [L2]/[D11] and the published-spectrum-overlay row [Doverlay] are confirmed at full statistics.

Table 15: Limitation index. [A] constraints, [L] limitations, [D] decisions, with status and impact. The model-dependence confirmation [L2]/[D11] and the published-spectrum overlay [Doverlay] are the full-statistics updates.

Label	Description	Status	Impact on result
A1	Experiment is ALEPH; combined result is an ALEPH α_s	held	naming
A2	Data are a pre-selected hadronic skim	resolved	skim eff cancels in shape
A3	ALEPH-2004 overlay shares the archived data (closure, not independent)	held	shared-data caveat on [Doverlay]
L1	Only 1994 full-sim MC exists; response valid for 1994	held	1994-only primary; subperiod in-coverage
L2	Single full-sim generator; hadronization via standalone PYTHIA 8	done (PYTHIA 8; HERWIG 7 infeasible); **the data—MC particle-level difference is the [L2] model dependence, bounded by the data-driven envelope, confirmed at full stats (M2 ratio 0.85–1.33 vs the subsample)**	M2 $\leq$ 0.0045
L3	NNLO grids + N3LL constants external; downscoped to NLL	downscoped, built-not-validated	dominant theory unc.
L4	NNLL for broadenings/y_{23} downscoped to NLL	downscoped, built-not-validated	poor GoF of those channels; B_W over-coverage
L5	FOPT/CIPT on moments (CIPT-analogue); distribution-level scheme variation downscoped	done (diagnostic only)	median 0.016, max 0.031 spread; NOT propagated
L6	Hadron-mass scheme (~2.5% C/τ universality breaking) (Hoang et al. 2015a)	PARTIAL (inside power-corr var)	~2.5% on Ω_1; subdominant
L7	ISR-spectrum variation through the response [COMMITMENTS B3]	PARTIAL (prose-only; fit-range lower bound only)	sub-permille post-cut
D1	All-particle energy-flow definition (primary)	held	matches theory
D4	IBU primary	done; runs unchanged on full data	unfolding
D5	OmniFold cross-check	done	<2%/bin agreement
D10	Two-parameter NP sector (Ω_1, α_0)	done; signs preserved on full data	NP fit

Label	Description	Status	Impact on result
D11	MC hadronization cross-check	done (PYTHIA 8 M0); **superseded for the budget by the data-driven envelope M2, confirmed at full stats**	model syst
D12	Joint fit primary, contingent	average=primary (contingency not met)	joint demoted
Dfit	Pre-registered fit windows	done; unchanged on full data	fit range
Dsel	Cut-based selection (MVA not needed)	held; no separate selection	>99% purity
Doverlay	ALEPH-2004 spectrum overlay (closure-style)	**done: full-data unfolded overlaid, agrees within full-data errors, [A3] caveat**	closure cross-check
Dyears	1994 primary; other years optional	held	1994-only

13 Appendix B: Per-observable systematic detail

This appendix records the signed per-observable $\Delta\alpha_s$ for the budgeted systematic sources on the full data, read from the machine-readable systematics file (`systematics.json`). The renormalization-scale variation is the up/down spread of α_s under $x_\mu \in [0.5, 2]$; the fit-range variation is the ± 1-bin window change re-evaluated on the full data; the model-dependence column is the data-driven M2 term at full statistics (Section 5.1).

Table 16: Signed per-observable systematic $\Delta\alpha_s$ for the budgeted sources on the full data (`systematics.json`). The fit-range column is the dominant theory term for ρ_H, B_W, B_T, and $-\log_{10} y_{23}$; for τ and C the renormalization-scale variation dominates instead (for C, the scale term 0.0103 and the fit-range term 0.0100 are comparable, both perturbative). The model-dependence column is the data-driven M2 term at full statistics, the largest experimental-side shape term for τ and ρ_H. The b-mass column is **flat at 0.0010 on every observable by construction** — a single literature-anchored bound applied uniformly (Section 5.7), not six independent evaluations; it is the **only** budgeted term in this table that is a flat literature-anchored bound rather than a quantity propagated bin-by-bin through the correction-and-fit chain, mirroring the §5.7 framing. The order-gap column is the orthogonalized $\sqrt{0.0035^2 - 0.0009^2} = 0.00338$ missing-higher-order term (Equation 16). The detector terms (energy scale, visible-energy reweight) and the hadronization M0 ($O(10^{-4}$–$10^{-10})$, subsumed by M2) are tabulated in Table 7. Quadrature of the listed budgeted terms reproduces each per-observable total of Table 9 to rounding (τ 0.0159, ρ_H 0.0316, B_W 0.0874, B_T 0.0366, C 0.0148, $-\log_{10} y_{23}$ 0.0953), with the hadronization M0 cross-check excluded from the sum (no double-count with M2).

Obs	scale	order gap	fit range (full)	b-mass	energy scale	$E_{\rm vis}$ rew.	model dep. (M2)
τ	0.0106	0.0034	0.0112	0.0010	0.0000	0.0000	0.0019
ρ_H	0.0078	0.0034	0.0299	0.0010	0.0020	0.0002	0.0045
B_W	0.0396	0.0034	0.0778	0.0010	0.0001	0.0000	0.0011
B_T	0.0159	0.0034	0.0327	0.0010	0.0002	0.0001	0.0007
C	0.0103	0.0034	0.0100	0.0010	0.0001	0.0000	0.0001
$-\log_{10} y_{23}$	0.0059	0.0034	0.0950	0.0010	0.0000	0.0000	0.0023

14 Appendix C: Covariance, fit-window, and machine-readable outputs

The per-observable covariances on the full data are positive-semi-definite with the same correlation structure as the validated covariance (and the same statistical scale as the MC expectation, $N_{\rm MC}/N_{\rm data} = 0.565$ by construction); the joint 55×55 event-level-bootstrap covariance is positive-semi-definite with the validated condition number (7.46×10^5, well below the 10^8 reliability cap). For downstream use, the fit windows of Table 5 are the recommended ranges; the extreme 2-jet region (where the ISR, power-correction, and fit-range systematics are degenerate) and the multi-jet tail (where the NNLO breaks down) are excluded from the fit. The maximum off-diagonal correlation in the joint matrix is driven by the broadening block (the two broadenings are near-degenerate) and the thrust-axis-sharing block (τ, C, ρ_H).

The single source of truth for every number in this note is the machine-readable full-data results directory in the analysis archive. It contains the combination file (the headline, the weights, the well-described cross-check, the BLUE and joint cross-checks, and the ρ_{theory} scan); the per-observable fit file (the per-observable α_s, NP, GoF, and pulls); the goodness-of-fit file (the theory-band χ^2/ndf and the saturated-toy p-values); the three-way comparison file (full vs 10% subsample vs MC expectation, with the stop-trigger verdicts); the fit-triviality-gate file (the gate and the non-circularity trace); the model-dependence-envelope file (the full-data M2 term); the systematics file (the per-observable budget); the reference-comparison file (the PDG and reference comparison and the §6.8 investigation); the per-subperiod file; the resolving-power and FOPT/CIPT files; and the full-data unfolded-density file (the densities and binning). Each is a small JSON document keyed by observable.

15 Appendix D: Reproduction contract

The full analysis chain reproduces from the archived ALEPH open data via the pixi environment. Environment setup is `pixi install`; the data and MC paths are configured in the analysis configuration. The full-data execution is deliberately minimal-new-code: the unfolding chain, the fit machinery, the systematics driver, and the combination are the same validated driver modules used for the MC expectation and the 10% subsample, run with results- and figure-path overrides only, and a combination budget-key override that adds the data-driven model-dependence term and subsumes the MC-vs-MC hadronization term. The only new logic is the full-data builder (the subsample builder with the mask removed and a positive full-data assertion $n_{\text{sel}} = n_{\text{total}} = 1{,}293{,}167$), the full-statistics model-dependence envelope (the subsample envelope with the seed loop removed), the three-way comparison, the fit-triviality gate, and the full-data figures. The method is provably unchanged across the staging; only the measured input flips from the 10% subsample to the full data. The unblinding was human-approved, and the positive full-data assertion documents that 100% of the data is used.

References

Abbate, Riccardo, Michael Fickinger, Andre H. Hoang, Vicent Mateu, and Iain W. Stewart. 2011. "Thrust at N3LL with Power Corrections and a Precision Global Fit for Alpha-s(mZ)." *Phys. Rev. D* 83: 074021. https://doi.org/10.1103/PhysRevD.83.074021.

Abbiendi, G. et al. 2005. "Measurement of Event Shape Distributions and Moments in e+e- to Hadrons at 91 to 209 GeV and a Determination of Alpha-s." *Eur. Phys. J. C* 40: 287. https://doi.org/10.1140/epjc/s2005-02120-6.

Abdallah, J. et al. 2003. "A Study of the Energy Evolution of Event Shape Distributions and Their Means with the DELPHI Detector at LEP." *Eur. Phys. J. C* 29: 285. https://doi.org/10.1140/epjc/s2003-01198-0.

Andreassen, Anders, Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. 2020. "OmniFold: A Method to Simultaneously Unfold All Observables." *Phys. Rev. Lett.* 124: 182001. https://doi.org/10.1103/PhysRevLett.124.182001.

Badea, Anthony et al. 2025. "Unbinned Measurement of Thrust in e+e- Collisions at Sqrt(s) = 91.2 GeV with ALEPH Archived Data." *arXiv Preprint*. https://arxiv.org/abs/2510.22038.

Badea, Anthony, Austin Baty, Paoti Chang, et al. 2019. "Measurements of Two-Particle Correlations in e+e- Collisions at 91 GeV with ALEPH Archived Data." *Phys. Rev. Lett.* 123: 212002. https://doi.org/10.1103/PhysRevLett.123.212002.

Banfi, Andrea, Heather McAslan, Pier Francesco Monni, and Giulia Zanderighi. 2015. "A General Method for the Resummation of Event-Shape Distributions in e+e- Annihilation." *JHEP* 05: 102. https://doi.org/10.1007/JHEP05(2015)102.

Banfi, Andrea, Heather McAslan, Pier Francesco Monni, and Giulia Zanderighi. 2016. "The Two-Jet Rate in e+e- at Next-to-Next-to-Leading-Logarithmic Order." *Phys. Rev. Lett.* 117: 172001. https://doi.org/10.1103/PhysRevLett.117.172001.

Barate, R. et al. 1998. "Studies of Quantum Chromodynamics with the ALEPH Detector." *Phys. Rept.* 294: 1. https://doi.org/10.1016/S0370-1573(97)00045-8.

Barate, R. et al. 2000. "A Measurement of the b-Quark Mass from Hadronic Z Decays." *Eur. Phys. J. C* 18: 1. https://doi.org/10.1007/s100520000533.

Becher, Thomas, and Guido Bell. 2012. "NNLL Resummation for Jet Broadening." *JHEP* 11: 126. https://doi.org/10.1007/JHEP11(2012)126.

Becher, Thomas, and Matthew D. Schwartz. 2008. "A Precise Determination of Alpha-s from LEP Thrust Data Using Effective Field Theory." *JHEP* 07: 034. https://doi.org/10.1088/1126-6708/2008/07/034.

Buskulic, D. et al. 1995. "Performance of the ALEPH Detector at LEP." *Nucl. Instrum. Meth. A* 360: 481. https://doi.org/10.1016/0168-9002(95)00138-7.

Catani, S., Yu. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber. 1991. "New Clustering Algorithm for Multijet Cross-Sections in e+e- Annihilation." *Phys. Lett. B* 269: 432. https://doi.org/10.1016/0370-2693(91)90196-W.

Catani, S., L. Trentadue, G. Turnock, and B. R. Webber. 1993. "Resummation of Large Logarithms in e+e- Event Shape Distributions." *Nucl. Phys. B* 407: 3. https://doi.org/10.1016/0550-3213(93)90271-P.

Catani, S., G. Turnock, and B. R. Webber. 1992. "Jet Broadening Measures in e+e- Annihilation." *Phys. Lett. B* 295: 269. https://doi.org/10.1016/0370-2693(92)91565-Q.

Chien, Yang-Ting, and Matthew D. Schwartz. 2010. "Resummation of Heavy Jet Mass and Comparison to LEP Data." *JHEP* 08: 058. https://doi.org/10.1007/JHEP08(2010)058.

Clavelli, L. 1979. "Jet Invariant Mass in Quantum Chromodynamics." *Phys. Lett. B* 85: 111. https://doi.org/10.1016/0370-2693(79)90789-5.

D'Agostini, G. 1995. "A Multidimensional Unfolding Method Based on Bayes Theorem." *Nucl. Instrum. Meth. A* 362: 487. https://doi.org/10.1016/0168-9002(95)00274-X.

Decamp, D. et al. 1991. "Measurement of Alpha-s from the Structure of Particle Clusters Produced in Hadronic z Decays." *Phys. Lett. B* 257: 479. https://doi.org/10.1016/0370-2693(91)91926-M.

Dissertori, G., A. Gehrmann-De Ridder, T. Gehrmann, et al. 2009. "Determination of the Strong Coupling Constant Using Matched NNLO+NLLA Predictions for Hadronic Event Shapes in e+e- Annihilations." *JHEP* 08: 036. https://doi.org/10.1088/1126-6708/2009/08/036.

Dokshitzer, Yu. L., A. Lucenti, G. Marchesini, and G. P. Salam. 1998a. "On the QCD Analysis of Jet Broadening." *JHEP* 01: 011. https://doi.org/10.1088/1126-6708/1998/01/011.

Dokshitzer, Yu. L., A. Lucenti, G. Marchesini, and G. P. Salam. 1998b. "On the Universality of the Milan Factor for 1/q Power Corrections to Jet Shapes." *JHEP* 05: 003. https://doi.org/10.1088/1126-6708/1998/05/003.

Dokshitzer, Yu. L., G. Marchesini, and G. P. Salam. 1999. "Revisiting Non-Perturbative Effects in the Jet Broadenings." *Eur. Phys. J. Direct C* 1: 3. https://doi.org/10.1007/s1010599c0003.

Dokshitzer, Yu. L., and B. R. Webber. 1995. "Calculation of Power Corrections to Hadronic Event Shapes." *Phys. Lett. B* 352: 451. https://doi.org/10.1016/0370-2693(95)00548-Y.

Ellis, R. K., D. A. Ross, and A. E. Terrano. 1981. "The Perturbative Calculation of Jet Structure in e+e- Annihilation." *Nucl. Phys. B* 178: 421. https://doi.org/10.1016/0550-3213(81)90165-6.

Farhi, Edward. 1977. "A QCD Test for Jets." *Phys. Rev. Lett.* 39: 1587. https://doi.org/10.1103/PhysRevLett.39.1587.

Gehrmann, T., G. Luisoni, and P. F. Monni. 2012. "Power Corrections in the Dispersive Model for a Determination of the Strong Coupling Constant from the Thrust Distribution." *Eur. Phys. J. C* 72: 2265. https://doi.org/10.1140/epjc/s10052-012-2265-x.

Gehrmann, T., G. Luisoni, and H. Stenzel. 2008. "Matching NLLA+NNLO for Event Shape Distributions." *Phys. Lett. B* 664: 265. https://doi.org/10.1016/j.physletb.2008.05.023.

Gehrmann-De Ridder, A., T. Gehrmann, E. W. N. Glover, and G. Heinrich. 2007. "NNLO Corrections to Event Shapes in e+e- Annihilation." *JHEP* 12: 094. https://doi.org/10.1088/1126-6708/2007/12/094.

Gehrmann-De Ridder, A., T. Gehrmann, E. W. N. Glover, and G. Heinrich. 2014. "EERAD3: Event Shapes and Jet Rates in Electron-Positron Annihilation at Order Alpha-s-Cubed." *Comput. Phys. Commun.* 185: 3331. https://doi.org/10.1016/j.cpc.2014.07.024.

Heister, A. et al. 2004. "Studies of QCD at e+e- Centre-of-Mass Energies Between 91 and 209 GeV." *Eur. Phys. J. C* 35: 457. https://doi.org/10.1140/epjc/s2004-01891-4.

Hoang, Andre H., Daniel W. Kolodrubetz, Vicent Mateu, and Iain W. Stewart. 2015a. "C-Parameter Distribution at N3LL Including Power Corrections." *Phys. Rev. D* 91: 094017. https://doi.org/10.1103/PhysRevD.91.094017.

Hoang, Andre H., Daniel W. Kolodrubetz, Vicent Mateu, and Iain W. Stewart. 2015b. "Precise Determination of Alpha-s from the c-Parameter Distribution." *Phys. Rev. D* 91: 094018. https://doi.org/10.1103/PhysRevD.91.094018.

Kardos, A., S. Kluth, G. Somogyi, Z. Tulipant, and A. Verbytskyi. 2018. "Precise Determination of Alpha-s(mZ) from a Global Fit of Energy-Energy Correlation to NNLO+NNLL Predictions." *Eur. Phys. J. C* 78: 498. https://doi.org/10.1140/epjc/s10052-018-5963-1.

Luisoni, G., P. F. Monni, and G. P. Salam. 2021. "C-Parameter Hadronisation in the Symmetric Three-Jet Limit and Impact on Alpha-s Fits." *Eur. Phys. J. C* 81: 158. https://doi.org/10.1140/epjc/s10052-021-08941-z.

Navas, S. et al. 2024. "Review of Particle Physics." *Phys. Rev. D* 110: 030001. https://doi.org/10.1103/PhysRevD.110.030001.

Parisi, G. 1978. "Super Inclusive Cross-Sections." *Phys. Lett. B* 74: 65. https://doi.org/10.1016/0370-2693(78)90061-8.

Schael, S. et al. 2006. "Precision Electroweak Measurements on the Z Resonance." *Phys. Rept.* 427: 257. https://doi.org/10.1016/j.physrep.2005.12.006.

Verbytskyi, A., A. Banfi, A. Kardos, et al. 2019. "High Precision Determination of Alpha-s from a Global Fit of Jet Rates." *JHEP* 08: 129. https://doi.org/10.1007/JHEP08(2019)129.