Measurement of the Primary Lund Jet Plane Density in Hadronic Z Decays with Archived ALEPH Data ($\sqrt{s} = 91.2$ GeV)

ALEPH Open Data Analysis

2026-05-30

Contents

Abstract

We present a measurement of the primary Lund jet plane (LJP) density in hadronic Z decays, using the archived ALEPH e^+e^- collision data taken at $\sqrt{s} = 91.2$ GeV during LEP1. To our knowledge this is the first measurement of the primary LJP *density* in e^+e^- collisions: a contamination-free, near-pure-C_F quark reference, free of the underlying event, multiple-parton interactions, and pileup that force model-dependent caution in the soft-wide-angle and hadronization corner of every pp Lund-plane measurement. Each hadronic event is split into two hemispheres by the plane perpendicular to the charged-particle thrust axis; the charged particles of each hemisphere are reclustered with the e^+e^- Cambridge/Aachen algorithm and then declustered along the harder branch, recording every primary splitting as a point in the plane spanned by $x = \ln(1/\Delta\vartheta)$ and $y = \ln(kt/GeV)$. The detector-level density is corrected to charged-particle level with two-dimensional iterative Bayesian unfolding, validated by a split-sample closure test, a graded stress test, and an independent bin-by-bin cross-check. The baseline uses the 1994 peak data corrected with the only ALEPH detector-simulated Monte Carlo sample (PYTHIA 6.1, ALEPH tune, 1994 conditions). The perturbative-bulk plateau measured here is the $(2/\pi)C_F\,\alpha_s(kt)$ density of independent soft-collinear emissions, **not** a Sudakov peak: the inclusive density carries no Sudakov suppression.

The headline result is the corrected charged-particle-level density on the 57-bin occupancy-floored fiducial region (genuine density $\rho > 0.05$ and reliable precision, relative uncertainty $< 25\%$) measured on the **full 1994 peak dataset** (1,293,167 events, 2,586,334 hemispheres). The integrated average number of primary emissions per hemisphere is $\langle \mathbf{N} \rangle = \mathbf{4.751 \pm 0.224\ (4.7\%)}$ [stat 0.0014 $\oplus$ syst 0.224], dominated by the prior/model dependence of the unfolding (2.98%). The full-data density is compatible with the expectation evaluated on the correction Monte Carlo, the operative per-bin metric being the diagonal $\chi^2/\mathrm{ndf} = 0.33$ and the per-bin pulls (worst 1.10σ, 0 bins above 2σ); because the dominant systematic is the shared correction-operator uncertainty that cancels in the data-minus-expectation difference, these clean pulls confirm the operator reproduces the shape rather than asserting equality with PYTHIA 6.1, and the genuine data-versus-PYTHIA-6.1 difference is instead resolved at high statistical significance and lies fully within the prior/model systematic. It is consistent with the earlier fixed-seed 10% subsample cross-check (unit-Gaussian full-vs-10% pulls, statistical scaling ratio 0.94, $1/\sqrt{10}$ as expected). It sits -1.33% below the PYTHIA 6.1 charged-particle truth (4.815), the physically expected harder-fragmentation difference of the legacy ALEPH tune — persisting unchanged across the Phase-3 prototype (-1.2%), the 10% subsample (-1.4%), and now the full data, and already covered by the dominant prior/model systematic — which the full statistics now resolve at high significance, the genuine resolving power of the measurement.

The measurement disfavours the modern generators as descriptions of the data and most strongly disfavours the Sherpa AHADIC cluster model (diagonal χ^2/ndf 14.7 versus 6.9–9.9 for the three PYTHIA 8 string variants); because the e^+e^- low-kt corner is pure hadronization, the measurement delivers a clean string-versus-cluster hadronization discrimination as a headline result. Three further observables are added. The unfolded hardest-primary-emission kt spectrum exhibits the **Sudakov peak at kt = 1.13 GeV** — the only place in the analysis where the Sudakov form factor is directly visible. The full spectrum shape discriminates the hadronization models: the cluster (Sherpa) spectrum is coherently softer than the data and the string models, the most discrepant in the spectrum-shape χ^2; the peak positions (Monash 1.06 GeV, Sherpa 0.95 GeV) track the same ordering but lie within one 0.5-wide ln(kt) bin and are indicative rather than bin-resolved. A heavy-versus-light-flavour cross-check using a data-driven displaced-track tag finds the b-enriched sample +32.4% higher in $\langle N \rangle$ than the light-enriched sample, with a localized dead-cone shape effect (collinear b-vs-light 18.1% versus soft 43.8%); this is a tagged-sample comparison, not a flavour-corrected density. A year-stability extension finds the 1992/1993/1995 corrected $\langle N \rangle$ agree with 1994 to within 0.11% (worst 0.023σ — the gated extension passes), with the 1994 peak retained as the baseline under the MC-coverage caveat. The resolving power is modest but genuine — roughly 12% at ~2.4σ per bin in the perturbative / running-coupling region (kt $\approx$ 1–5 GeV) — and the measurement is positioned as the contamination-free, fixed-C_F complement of ALICE, ceding the high-kt reach and the jet-substructure/tagging applications to ATLAS and CMS.

Change Log

Phase 5 v1 — final documentation

- Added the two methodology schematics promised by the strategy and produced by the figure executor: the observable-construction diagram (Figure 4) in the observable section and the correction-chain diagram (Figure 5) in the corrections section, each with an interpretive caption.
- Polished the body into a single coherent physics argument: internal phase labels were removed from the running text (they survive only here and, as neutral stage names, in the validation summary), so the full 1994-peak

result reads throughout as the result and the 10% subsample and the Asimov expectation read as the validation context they are.

- Verified the final conventions checklist (`conventions/unfolding.md`), closed every COMMITMENTS line, and ran the numbers-consistency lint against `results/*.json` (no discrepancies). No physics content changed; all headline numbers are unchanged.
- Final 5-bot review (ITERATE) fixes: resolved the Table 2 reconstruction event-count labelling (731,006 selected vs 771,597 raw `t` entries); regenerated the observable-construction schematic (Figure 4) as a legible 2×2 grid; corrected the generator $\langle N \rangle$ comparison direction; added a floored-region total-uncertainty row (6.0%) and a tbl:obs-gof region cross-reference; reserved "perturbative plateau" for the ρ-peak and named the kt $\approx$ 1–5 GeV window the perturbative / running-coupling region; located the worst closure-pull bin as a low-kt edge bin; and surfaced the single-detector-simulation residual-risk caveat in the error-budget narrative. No physics content or headline numbers changed.

Phase 4c v1 — full data (final results)

- Full unblinding approved at the human gate. The full 1994 peak data (1,293,167 events, 2,586,334 hemispheres) was run through the unchanged validated correction chain and became the primary result, $\langle N \rangle = 4.751 \pm 0.224$ (4.7%) on the 57-bin occupancy-floored region. Nothing was tuned to the data: only the input spectrum and the statistical covariance changed relative to the expectation.
- The 10% subsample was demoted to a validation cross-check (full-vs-10% pulls unit-Gaussian, $1/\sqrt{10}$ statistical scaling). The full data is compatible with the expectation (diagonal $\chi^2/\mathrm{ndf} = 0.33$, worst pull 1.10σ, 0 bins above 2σ); the -1.33% data-versus-PYTHIA-6.1 offset persists from the prototype (-1.2%) and the 10% subsample (-1.4%) and is the expected, prior-covered harder-fragmentation difference, now resolved at high statistical significance.
- Three observables were added: the unfolded hardest-emission Sudakov-peak spectrum (kt = 1.13 GeV; the inclusive density carries no Sudakov factor), a data-driven heavy-versus-light-flavour split (b-enriched +32.4% in $\langle N \rangle$, with a localized dead-cone shape effect; a tagged-sample comparison), and a year-stability extension (1992/1993/1995 agree with 1994 to within 0.11%).
- Systematics were re-evaluated on the unblinded full data and confirm the budget exactly. The physics framing (four pillars plus the honest assessment) and the string-versus-cluster discrimination headline were finalized; the money plot was annotated with the three regions and the no-UE/MPI/pileup callout.
- 1-bot review fixes: the full-covariance χ^2 was demoted from primary to a convention-required companion (the diagonal χ^2/ndf and per-bin pulls are the operative metrics), the Sudakov peak positions were flagged as indicative (within one ln(kt) bin) with the discrimination resting on the spectrum shape, and the money-plot annotations were added.

Earlier versions (condensed)

- **Phase 4b v1 — 10% data validation (+ 4-bot+bib review fixes).** First contact with real data on a fixed-seed 10% subsample; corrected $\langle N \rangle = 4.746 \pm 0.224$ compatible with the expectation. Added the lepton-removed companion observable (-5.4%) and the tight/loose track-selection stability cross-check (-0.37%). The z-score/compatibility framing was made consistent across the note (the offset is a physically expected, prior-covered harder-fragmentation difference, not an artifact). Human gate passed.
- **Phase 4a v1 — initial AN (+ ITERATE / cycle-2 review fixes).** Complete documentation of the observable, selection, correction chain, ten-source systematic program, covariance, expected density, generator comparisons, NLL anchor, and resolving power, with expected (Asimov/MC) results only. The occupancy-floored 57-bin region was adopted; the Sherpa AHADIC cluster model was added as the string-versus-cluster handle; the full-covariance χ^2 magnitude was reframed as a coherent rank-1 upper bound and the robust discrimination pointed at the diagonal χ^2 and the resolving power.

1 Introduction

1.1 Motivation

The Lund jet plane (LJP) is a two-dimensional representation of the phase space of 1→2 QCD splittings inside a jet (Dreyer et al. 2018). Constructed by reclustering a jet's constituents with the angular-ordered Cambridge/Aachen (C/A) algorithm and then reversing the clustering history, each primary splitting is mapped to a point whose coordinates encode the transverse momentum kt of the softer prong relative to the harder one and the opening angle

$\Delta\vartheta$ between them. In the soft-and-collinear (eikonal) picture, primary emissions populate the plane approximately uniformly because QCD is nearly scale invariant. The density of emissions factorizes the physical mechanisms that govern parton showering and hadronization: hard wide-angle radiation, the running of the strong coupling α_s in the perturbative bulk, hadronization at low kt, and — at hadron colliders only — the underlying event, multiple-parton interactions, initial-state radiation, and pileup near the jet boundary (Dreyer et al. 2018; ATLAS Collaboration 2020; CMS Collaboration 2024).

The leading-logarithmic expectation is that the per-jet density is constant across the plane (ATLAS Collaboration 2020), and for independent soft-collinear emissions the density is directly proportional to the running coupling,

$$\rho(k_t, \Delta R) \;\approx\; \frac{2}{\pi}\, C_R\, \alpha_s(k_t), \tag{1}$$

with C_R the colour factor of the emitter (C_F = 4/3 from a quark, C_A = 3 from a gluon) (CMS Collaboration 2024). The local coupling is evaluated at the kt of each emission, so the density directly probes α_s(kt) across the perturbative range spanned by the measurement.

Hadronic Z decays at the Z pole provide the cleanest possible environment for this observable. There are no beam remnants, no parton distribution functions, and no pileup; for events taken on the resonance peak, initial-state radiation is strongly suppressed. The entire wide-angle, soft corner of the plane that is contaminated by the underlying event in pp collisions is, in e^+e^-, populated by genuine QCD radiation and hadronization only. A second distinctive feature is the colour content: the $e^+e^- \to Z \to q\bar{q}$ hemispheres are quark dominated, so the perturbative plateau samples almost pure C_F = 4/3 radiation. The published pp LJP measurements (ALICE, ATLAS, CMS) use gluon-enriched inclusive-jet samples whose plateau density is set by a quark/gluon mix and lies higher, by a factor approaching the colour ratio C_A/C_F = 9/4 ≈ 2.25 in the gluon-dominated limit (Dreyer et al. 2022). The e^+e^--below-pp offset is therefore a predicted, physically-understood feature — a near-pure-C_F primary LJP density — and is itself novel.

This measurement helps answer three physics questions: (i) do modern parton-shower plus hadronization models reproduce the *differential* radiation pattern of light-quark jets at the Z pole, not just integrated event shapes; (ii) how does the hadronization-dominated low-kt region in e^+e^- compare to the pp picture where it is entangled with the underlying event; and (iii) are the analytic predictions for the perturbative density — the leading-log uniformity and α_s(kt) running of Equation 1, and the full next-to-leading-log plus next-to-leading-order calculation (Lifson et al. 2020) — visible in e^+e^- as they are in pp. The measurement is complementary to the high-kt pp measurements and connects directly to the generators that were originally tuned on LEP/SLD event-shape and fragmentation data, e.g. the PYTHIA 8 Monash tune (Skands et al. 2014).

1.2 Prior measurements

Published primary LJP-density measurements all use pp collisions. The ALICE measurement (Havener and ALICE Collaboration 2022) uses intermediate-pT (20–120 GeV) charged-particle jets and reaches kt ≈ 5 GeV; because its kt range overlaps the e^+e^- hadronization regime that is the central novelty of this measurement, it is the primary external overlay partner, although it is published only as a conference proceeding and is not refereed. The CMS measurement (CMS Collaboration 2024) uses high-pT (> 700 GeV) jets and overlaps only the perturbative high-kt edge; it is a refereed measurement in the same ln(kt) plane and serves as the secondary high-kt overlay. The ATLAS measurement (ATLAS Collaboration 2020) is also refereed but is differential in ln(1/z) rather than ln(kt); its density therefore lives in a different plane and cannot be overlaid directly, so only its convention-independent integral, the average number of primary emissions per jet, is used as a reference point. No prior e^+e^- LJP-*density* measurement exists, but the same archived ALEPH dataset has been used for a published SoftDrop substructure measurement (groomed jet mass, momentum sharing z_g, groomed radius R_g) (Chen et al. 2022), the closest prior e^+e^- declustering work, and for an energy-energy-correlator analysis (Bossi et al. 2025) that shares the identical detector and Monte Carlo and provides our correction-chain and systematic-magnitude template.

The generator-independent analytic benchmark is the Lifson–Salam–Soyez next-to-leading-log plus next-to-leading-order calculation of the primary LJP density (Lifson et al. 2020), with a quoted precision of 5–7% at high kt rising to ~20% at the kt ≈ 5 GeV lower edge of the perturbative region. Together with the running-coupling relation of Equation 1, this enables a concrete comparison of the perturbative density to α_s(kt).

1.3 Scope of this note

This note documents the complete analysis and its **final measured result** on the full 1994 peak dataset. The result section presents the corrected charged-particle-level density measured on the full data as the primary result, its compatibility with the expectation evaluated on the correction Monte Carlo, and its consistency with the earlier fixed-seed 10% subsample, which is retained as a validation cross-check. **Nothing is tuned to the data:** the correction operator (response, efficiency, purity, prior) and the systematic shifts are the values established before unblinding, and only the input spectrum and the statistical covariance change when the full data replace the Monte Carlo expectation. Three further observables — the leading-emission kt Sudakov-peak spectrum, a heavy-versus-light-flavour split, and a year-stability extension — are added as secondary results and cross-checks. The introduction, data-sample, event-selection, correction, systematic-methodology, and statistical-method sections document the fixed methodology; the results, comparison, and conclusion sections present the measured numbers.

2 Data and Monte Carlo samples

2.1 Overview

The analysis uses the archived ALEPH LEP1 data and the corresponding detector-simulated Monte Carlo, both obtained from the ALEPH open-data archive and treated as read-only. The data comprise six merged files spanning the 1992–1995 run periods at and around the Z resonance. The Monte Carlo is a single detector-simulated sample of hadronic Z decays generated with PYTHIA 6.1 in the ALEPH tune under 1994 run conditions — the only LEP1 ALEPH sample carrying a full detector simulation (Bossi et al. 2025). Because the detector simulation exists for 1994 only, the baseline corrected measurement is restricted to the 1994 peak data, consistent with the Monte-Carlo-coverage rule and with the choice made by the same-dataset energy-energy-correlator analysis (Bossi et al. 2025). The full event-aligned generator-level record (`tgen`) and the pre-selection generator record (`tgenBefore`) are present in the Monte Carlo, enabling an in-file detector-response matrix and unfolding to charged-particle level.

2.2 Centre-of-mass energy content

The dataset is not monochromatic. The 1992 and 1994 runs were taken on the resonance peak; the 1993 and 1995 runs were dedicated lineshape energy scans at three points, approximately M_Z $\pm$ 1.8 GeV and at the peak. The off-peak scan points have substantially enhanced hard-ISR and radiative-return content, which breaks the negligible-ISR premise that motivates the measurement. The analysis therefore selects peak-energy events only, using a window on the per-event energy combined with the stored ISR-rejection flag (Section 3.2). The 1994 baseline is peak-only by construction; any future year extension applies the same peak window so that off-peak scan points never enter the corrected density.

The per-event energy distributions, histogrammed per year at full statistics (Figure 1), confirm this structure: 1992 and 1994 are single-peak, and 1993 and 1995 show three distinct energy clusters. There are no events between the peak and the off-peak satellites, so the result is insensitive to the exact window width.

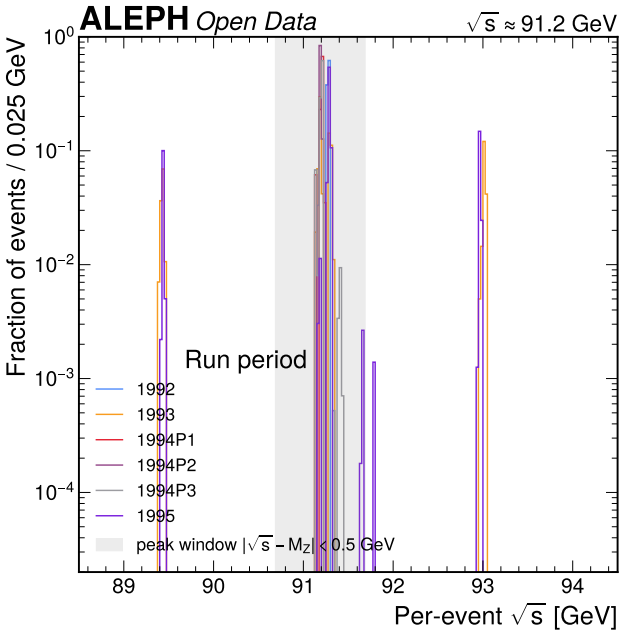

Figure 1: Per-event centre-of-mass energy distribution by run period. The 1992 and 1994 distributions are single-peaked at $\sqrt{s} \approx$ M_Z, confirming peak-only running, while the 1993 and 1995 distributions show three clusters at $\sqrt{s} \approx$ 89.4, 91.2, and 93.0 GeV — the dedicated lineshape energy scans. The wide gap between the peak and the off-peak satellites shows that the peak-window selection is insensitive to its exact width.

2.3 Data summary

The data file inventory and event counts are summarized in Table 1. The 1994 peak sample (P1+P2+P3) totals 1,365,440 recorded events; after the peak window and the standard hadronic-Z selection (Section 3) it yields 1,293,167 events, corresponding to 2,586,334 hemispheres, which is the baseline sample. Integrated luminosities are not published per archived file. They are estimated from the published ALEPH lineshape paper (ALEPH Collaboration 2000) (Table 4 for the per-energy-point integrated luminosity and Table 21 for the measured per-point hadronic cross section): for 1994 the peak integrated luminosity is 42,695.2 nb^{-1} = 42.7 pb^{-1} at a measured hadronic cross section of 30.39 nb. The expected yield N_exp = Σ L·σ_had over the peak points, compared to the observed peak yield, gives the pre-selection efficiency f_presel reported in the last column. For 1994, f_presel = 0.997, i.e. the archived 1994 peak sample is essentially the full published 1994 peak hadronic sample; the lower values for 1992/1993/1995 (0.76–0.81) reflect a coverage difference in the archived open-data sample (some peak sub-periods are omitted), not a per-event efficiency that distorts the density. Because the observable is normalized per hemisphere, the measured density is insensitive to f_presel; the luminosity enters only this yield cross-check.

2.4 Monte Carlo sample

The Monte Carlo sample is summarized in Table 2. It consists of 40 files of detector-simulated hadronic Z decays generated with PYTHIA 6.1 in the ALEPH tune under 1994 run conditions (Bossi et al. 2025). Each file contains three event-aligned trees: the reconstruction-level particle list (`t`), the generator-level truth list (`tgen`), and the generator-level list *before* event selection (`tgenBefore`) used for the event-selection-efficiency correction. The reconstruction and truth trees are event-aligned to a matching fraction of 1.0000 in every sampled file, confirming that an in-file response matrix and full unfolding are feasible. After the peak window and the hadronic selection the sample yields 731,006 reconstruction-level events (1,462,012 hemispheres), about 57% of the data hemisphere statistics — adequate for the response matrix. No generator banner is stored in the files; the generator, tune, and run conditions are taken from the same-dataset reference (Bossi et al. 2025).

In addition to the detector-simulated correction sample, four standalone particle-level samples were generated for the theory comparison and the generator-bracketed prior systematic (Section 5). Three are PYTHIA 8 samples — the Monash tune (Skands et al. 2014), the Vincia shower, and the default tune, each with 1.0×10^6 events at $\sqrt{s}$ = 91.2 GeV with no detector simulation — and the fourth is a Sherpa 2.2.16 sample with the AHADIC++ cluster-

Table 1: Data sample inventory. Recorded events are the merged-file entry counts; peak-selected events apply the peak energy window and the stored hadronic-Z selection flag. Integrated luminosities are estimated from the published ALEPH lineshape paper (ALEPH Collaboration 2000); for the scan years only the peak-point luminosity is quoted. f_presel is the ratio of observed to published peak yield. The 1994 peak sample is the baseline for the corrected measurement.

Period	$\sqrt{s}$ [GeV]	Recorded events	Peak-selected	$\mathcal{L}$ [pb^{-1}]	f_presel
1992	91.2 peak	551,474	522,526	21.0	0.809
1993	M_Z ± 1.8 scan	538,601	354,499	14.4 (peak)	0.804
1994 P1	91.2 peak	433,947	—	—	—
1994 P2	91.2 peak	447,844	—	—	—
1994 P3	91.2 peak	483,649	—	—	—
1994 baseline	**91.2 peak**	**1,365,440**	**1,293,167**	**42.7**	**0.997**
1995	M_Z ± 1.8 scan	595,095	404,655	17.3 (peak)	0.764

hadronization model (5×10^5 events), the same cluster-model family ATLAS used as its string-versus-cluster anchor (ATLAS Collaboration 2020). The three PYTHIA 8 samples bracket the parton-shower and tune dependence within the Lund-string model, while the Sherpa sample provides the string-versus-cluster hadronization handle. All four are independent of the PYTHIA 6.1 correction sample, satisfying the theory-comparison-independence requirement.

Table 2: Monte Carlo samples. The PYTHIA 6.1 detector-simulated sample (1994 conditions) is the only sample carrying a full ALEPH detector simulation and is used to build the response matrix and all detector corrections. The selected reco count is the 731,006 events (1,462,012 hemispheres) surviving the peak window and hadronic selection, out of the 771,597 raw `t`-tree entries before selection. The three standalone PYTHIA 8 samples and the Sherpa AHADIC cluster-hadronization sample are particle-level only and enter as theory comparisons and as the generator-bracketed prior systematic; the Sherpa sample is the string-versus-cluster handle. A HERWIG 7 source build was attempted but blocked by the CMake-4 policy default (Section 5.1, Section 11); the cluster-model role it would have filled is provided by the live Sherpa sample.

Process	Generator	Tune	Selected reco events	Det. sim.	Role
$e^+e^- \to Z \to q\bar{q}$	PYTHIA 6.1	ALEPH	731,006 selected (771,597 raw)	yes	response / correction
$e^+e^- \to Z \to q\bar{q}$	PYTHIA 8	Monash 2013	1.0M	no	theory + prior bracket
$e^+e^- \to Z \to q\bar{q}$	PYTHIA 8	Vincia	1.0M	no	theory + prior bracket
$e^+e^- \to Z \to q\bar{q}$	PYTHIA 8	default	1.0M	no	theory + prior bracket
$e^+e^- \to Z \to q\bar{q}$	Sherpa 2.2.16	AHADIC (cluster)	0.5M	no	theory + prior bracket (cluster)

2.5 Data archaeology

Because this is archived open data, several properties that affect feasibility were checked explicitly. The per-particle reconstruction-level `weight` branch is non-trivial (mean ≈ 1.02, broad distribution), present and structurally identical in both data and reconstruction-level Monte Carlo, and exactly 1.0 at generator level; it is an energy-flow per-track weight reproduced by the detector simulation. Its effect on the per-hemisphere-normalized density shape is small, so the nominal correction is unweighted and the with/without-weight difference is carried as a systematic (Section 5). The per-event `particleWeight` branch is trivially 1.0 everywhere. The generator particle-identity branch was decoded empirically as the GEANT3/ALEPH KINE code (not PDG); the weakly-decaying strange hadrons K^0_S, Λ, K^0_L, Σ, Ξ and the neutron appear in the truth record as undecayed neutral species, which fixes the charged-

particle-level definition to the ATLAS $c\tau > 10$ mm convention (Section 3.4). The stored hadronic-selection flag was reproduced from its component flags to 99.5% and from a scratch reimplementation of the textbook three-line selection to 97.6%, confirming the flag's meaning; the analysis uses the stored flag.

3 Event selection

3.1 Overview

The selection has three levels: a hadronic-Z event selection that defines the sample, a charged-particle (track) selection that defines the reconstruction-level fiducial objects entering the observable, and a particle-level definition that specifies what the measurement is corrected to. The event and track selections follow the standard archived-ALEPH hadronic-Z selection (Bossi et al. 2025), which itself follows the published ALEPH event selection (Tournefier and ALEPH Collaboration 1999). Each cut is documented below with its motivation; the reconstruction-level cuts are illustrated by the N−1 distributions (the cut variable shown with all other cuts applied) in Figure 2 and the accompanying panels.

3.2 Event selection

Events are required to lie in the peak energy window $|\text{Energy} - \text{M\_Z}| < 0.5$ GeV (with $\text{M\_Z} = 91.188$ GeV (Particle Data Group 2024b)) and to pass the stored hadronic-Z selection flag. The peak window removes the off-peak lineshape-scan points of the 1993 and 1995 runs that violate the negligible-ISR premise; for the 1994 baseline it keeps 100% of events because 1994 is peak-only. The stored selection flag bundles the standard ALEPH hadronic-Z requirements: a sphericity-axis polar angle in the acceptance $7\pi/36 \leq \vartheta\_S \leq 29\pi/36$ ($\approx 35°$–$145°$), at least five good charged tracks, total reconstructed charged energy ≥ 15 GeV, total visible energy below 200 GeV (removing laser calibration events), and the standard ISR/WW/missing-momentum quality requirements. These cuts suppress $\tau^+\tau^-$, two-photon, WW, and Bhabha contamination to a total below 0.6% (Section 3.5).

The event cutflow for a representative data period and one Monte Carlo file is shown in Table 3. Both cutflows are monotonically non-increasing. For the peak-only 1994 baseline the peak window keeps everything; the ISR flag removes about 1.1%, and the full hadronic selection keeps about 96% of peak events.

Table 3: Event cutflow for a representative data period and one Monte Carlo file. Each stage is a subset of the previous one; the fractions converge at this statistics. This is a selection efficiency, not a spectrum.

Stage	Data (1994 P1)	MC (file 001)
All events	40,000	19,158
+ peak window	40,000	19,158
+ ISR rejection	39,562	18,972
+ hadronic selection	37,833	18,131
≥ 1 clusterable hemisphere	37,833	18,131
≥ 1 primary emission	37,832	18,131

3.3 Charged-particle selection

Reconstruction-level charged objects are required to be charged energy-flow objects (charge $\neq 0$, energy-flow class in {0,1,2,3}) with $|\cos\vartheta| < 0.94$, $\text{pT} \geq 0.2$ GeV, at least four TPC hits, and impact parameters $|\text{d0}| < 2$ cm and $|\text{z0}| < 10$ cm. The acceptance and impact-parameter cuts remove badly measured and secondary tracks; the pT ≥ 0.2 GeV cut removes the soft turn-on where the tracking efficiency falls steeply; the TPC-hit cut ensures a well-measured momentum. The track cutflow is dominated by the pT cut (~2.8% of objects); the acceptance, hit, and impact-parameter cuts remove very little because the charged energy-flow objects are already well-reconstructed tracks. Every cut is motivated by its N−1 distribution: the pT cut sits at the soft turn-on, the TPC-hit cut sits below the bulk of the hit distribution, and the impact-parameter cuts sit in the tails.

The five N−1 distributions above show that all track cuts are well motivated by the data and Monte Carlo distributions and that the agreement is good in the regions retained.

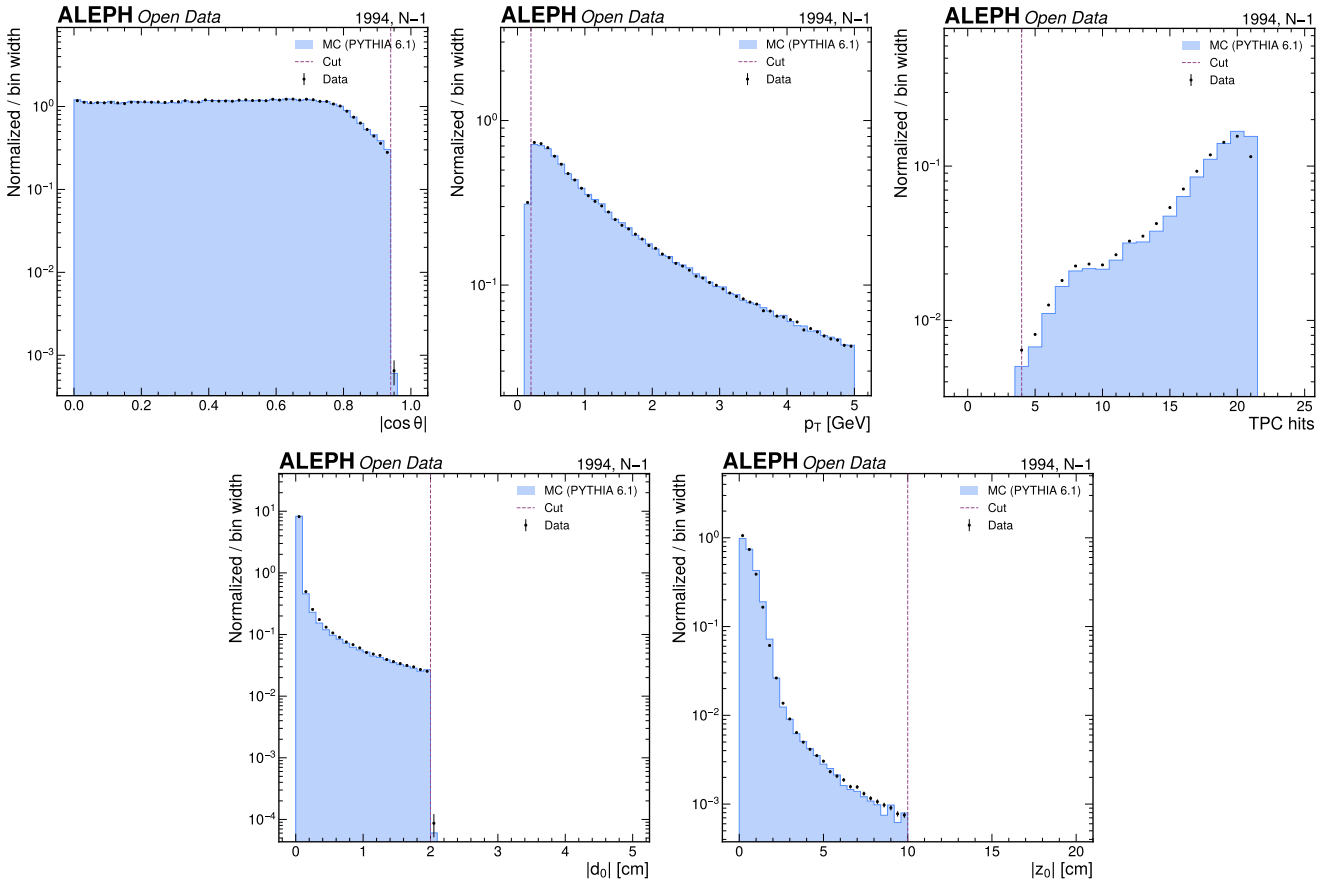

Figure 2: N−1 distributions for the five track-quality selection cuts, data versus Monte Carlo. (a) Track $|\cos\vartheta|$ acceptance (cut at 0.94, removing the forward/backward region outside the tracking acceptance). (b) Track pT (cut at 0.2 GeV, removing the soft turn-on where the tracking efficiency falls — the dominant track cut at 2.8% of objects). (c) TPC-hit requirement (cut at four hits, removing poorly measured short tracks). (d) Transverse impact parameter d0 (cut at 2 cm, removing secondary and badly measured tracks while keeping the prompt core). (e) Longitudinal impact parameter z0 (cut at 10 cm, removing tracks inconsistent with the interaction region). Data and MC agree within a few percent across the accepted region of every variable, validating the track selection.

3.4 Particle-level definition

The measurement is corrected to a charged-particle-level definition that is held fixed throughout the analysis. The particle level comprises the electrically charged final-state particles (charge $\neq 0$), with stable hadrons and leptons defined by $c\tau > 10$ mm, following the ATLAS charged-particle convention (ATLAS Collaboration 2020, 2019). The weakly-decaying strange hadrons K^0_S, Λ, K^0_L, Σ, Ξ and the neutron are treated as stable neutral species and do not contribute charged daughters; this matches what the ALEPH generator record provides, where these species appear undecayed. Neutrinos and other neutral invisibles are excluded. Charged leptons are included in the baseline, consistent with a charged-hadron-plus-charged-lepton fragmentation observable; the density with and without non-prompt decay leptons is a defined cross-check observable (not a detector systematic), reported in Section 7.11. The phase space is the full solid angle within the fiducial event selection; there is no per-particle pT cut at particle level, as the reconstruction-level $pT \geq 0.2$ GeV cut is corrected for by the efficiency step. The event selection itself is part of the particle-level definition and is corrected to the pre-selection generator level via `tgenBefore`.

3.5 Backgrounds

This is a differential fragmentation measurement of the single signal process $e^+e^- \to Z \to q\bar{q} \to$ hadrons; there is no signal/background discrimination task and no multivariate classifier. A multivariate analysis is not applicable because there is no second physics class to separate against — the sub-percent backgrounds are removed by the established cut-based ALEPH hadronic selection, and a classifier trained to separate $q\bar{q}$ from a sub-percent contamination would provide no measurable gain and would bias an unfolded shape measurement. None of the reference LJP or EEC analyses uses a classifier for event selection. The residual contaminations of the hadronic sample are summarized in Table 4. The total non-$q\bar{q}$ contamination is below 0.6% (Tournefier and ALEPH Collaboration 1999; Bossi et al. 2025), dominated by $\tau^+\tau^-$ ($\approx 0.32\%$) and two-photon ($\approx 0.26\%$) events, both suppressed by the ≥ 5-track and charged-energy requirements. Their effect on the *shape* of the density is subdominant and is assigned a background systematic (Section 5).

Table 4: Background classification for the hadronic Z sample. The total non-$q\bar{q}$ contamination is below 0.6%; its shape effect is carried as the background systematic (Section 5).

Process	Class	Magnitude	Handling
$e^+e^- \to q\bar{q}$	signal (irreducible)	100% of signal	the measurement
$e^+e^- \to \tau^+\tau^-$	reducible	$\approx 0.32\%$	track + energy cuts; residual as systematic
$\gamma\gamma \to$ hadrons	reducible	$\approx 0.26\%$	energy + acceptance cuts; residual as systematic
hard ISR / $Z\gamma$	reducible	small at peak	ISR flag; negligible on peak
$e^+e^- \to W^+W^-$	reducible	negligible at LEP1	below threshold; WW flag
Bhabha, detector noise	instrumental	small	acceptance + track quality cuts

3.6 Input data/MC validation

Before building any correction, the reconstruction-level data and Monte Carlo were compared for every variable entering the observable: the track azimuth, polar angle, transverse momentum, and momentum magnitude, the charged-track multiplicity, the per-hemisphere multiplicity and charged energy, and the charged thrust. The comparisons (Figure 3 and the accompanying panels) are area-normalized and use the 1994 peak sample. The data/MC ratio sits within 1–3% across the bulk of every variable. The per-variable χ^2/ndf values are large (1.2–14) but this is driven by the sub-percent per-bin statistical precision of the ~400k-event sample, not by a modelling failure — the ratio panels are the operative judgement. Two features are flagged and revisited at the correction stage: the per-hemisphere charged energy shows a coherent 5–10% shape difference (the worst-modelled input), and the track pT and momentum tails are about 5% harder in data than in PYTHIA 6.1. Both are fragmentation-model differences well within the prior/model systematic.

The eight panels above constitute the input-validation gate required before building the response matrix. The agreement is adequate, with the two flagged fragmentation-shape differences understood and propagated through the prior systematic.

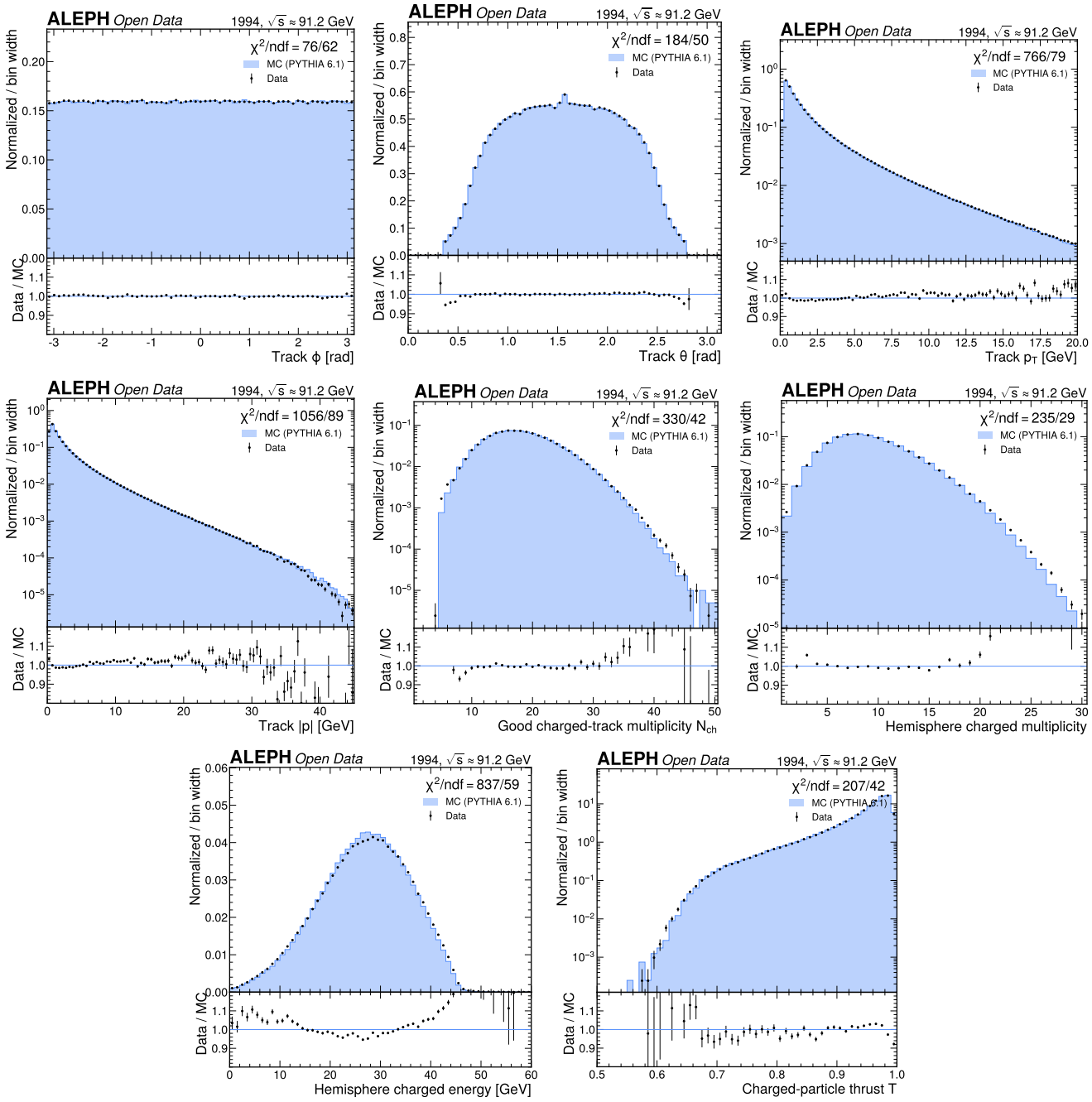

Figure 3: Data versus Monte Carlo for the track-level and event-level input observables, area-normalized. (a) Track azimuthal angle φ (flat, agreement at the 1% level, $\chi^2/\mathrm{ndf} = 1.23$). (b) Track polar angle ϑ (1% level across the $|\cos\vartheta| < 0.94$ acceptance). (c) Track transverse momentum pT (1–3% in the bulk; data 5% harder than PYTHIA 6.1 in the high-pT tail). (d) Track momentum magnitude |p| (1–2% below 15 GeV; same harder-fragmentation trend in the tail). (e) Good-charged-track multiplicity per event (1–2% in the bulk). (f) Charged-track multiplicity per hemisphere, which seeds the emission count (1–2% level). (g) Charged energy per hemisphere, the worst-modelled input (coherent 5–10% shape difference, covered by the prior systematic). (h) Charged-particle thrust T (2–5% level near the two-jet limit $T \to 1$). All differences are carried by the prior/model systematic.

4 Observable and corrections

4.1 Observable construction

For each selected hadronic event the observable is built in four steps, held fixed throughout the analysis. First, the charged-particle thrust axis n̂_T is taken from the precomputed charged-only thrust direction, and the plane perpendicular to n̂_T through the origin splits the event into two hemispheres; a charged particle i belongs to a hemisphere by the sign of p⃗_i · n̂_T. The charged-only thrust axis matches the charged-particle observable being measured. Second, the charged particles of each hemisphere are reclustered with the e^+e^- Cambridge/Aachen algorithm — implemented in FastJet (Cacciari et al. 2012) as the e^+e^- generalized-kt algorithm with p = 0, the native angular-ordered metric (Dokshitzer et al. 1997; Wobisch and Wengler 1999) — with a radius large enough that the whole hemisphere is reconstructed as a single jet.

Third, the C/A clustering history is reversed. At each declustering step the jet $j \to j_1 + j_2$ is undone, the harder prong j_1 (higher momentum) is followed, and the softer prong j_2 is recorded as a primary emission (Dreyer et al. 2018; ATLAS Collaboration 2020; CMS Collaboration 2024). Each emission is mapped to a point with coordinates

$$\Delta\theta = \angle(j_1, j_2), \qquad k_t = p_{j_2} \sin\Delta\theta, \tag{2}$$

where Δϑ is the opening angle between the two prongs and p_{j_2} is the momentum magnitude of the softer prong. The transverse momentum kt is **emitter-relative**: it is the momentum of the softer prong transverse to the harder prong (the emitter), following the original Lund-plane proposal and its analytic calculation (Lifson et al. 2020; Dreyer et al. 2018). In the small-angle limit, with z the momentum fraction of the softer prong, kt = p_soft sin Δϑ ≈ z·ΔR·p_prong, recovering the familiar z·ΔR form. This is distinct from the CMS beam-axis definition; in e^+e^- there is no beam-axis jet pT, so the emitter-relative kt is the natural choice. We adopt the ln(kt/GeV) axis label from CMS (CMS Collaboration 2024) but not its kt definition.

Fourth, the primary LJP density is defined as

$$\rho(k_t, \Delta\theta) \;\equiv\; \frac{1}{N_{\text{hem}}} \frac{\mathrm{d}^2 N_{\text{emissions}}}{\mathrm{d}\ln(k_t/\text{GeV})\,\mathrm{d}\ln(1/\Delta\theta)}, \tag{3}$$

normalized to the number of hemispheres N_hem (ATLAS Collaboration 2020; CMS Collaboration 2024). The density is insensitive to the total cross section, and its integral over the plane is the average number of primary emissions per hemisphere, ⟨N_emissions⟩. The horizontal coordinate is x = ln(1/Δϑ) with Δϑ in radians: the hemisphere has no jet radius R to normalize with, so the opening angle is used directly. As a limiting check, a collinear splitting Δϑ → 0 maps to the right edge (ln(1/Δϑ) → ∞) and the bottom (kt → 0), while a symmetric wide-angle splitting maps to the hard, wide-angle corner — the canonical triangular Lund structure, whose upper-right corner (simultaneously collinear and hard) is kinematically forbidden.

The four construction steps and the resulting plane are shown schematically in Figure 4. Reading left to right: the hadronic event is split into two hemispheres by the plane perpendicular to the charged-particle thrust axis; the charged particles of the analysed hemisphere are reclustered into a single jet with the e^+e^- Cambridge/Aachen algorithm; the clustering history is reversed and, following the harder prong, each softer prong is recorded as a primary emission mapped to its (ln(1/Δϑ), ln kt) coordinates; and the accumulated emissions populate the primary Lund plane, whose three physically distinct regions — the perturbative plateau, the running-coupling rise toward lower kt, and the hadronization turnover — are the structures the measurement resolves. The diagram fixes the geometric meaning of the two axes (more collinear to the right, harder upward) used throughout the note.

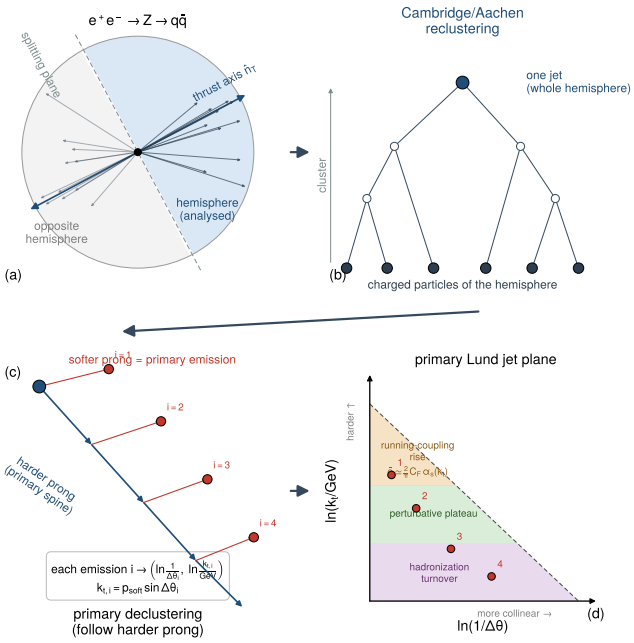

Figure 4: Construction of the primary Lund jet plane observable. (a) The hadronic Z event is split into two hemispheres by the plane perpendicular to the charged-particle thrust axis; the analysed hemisphere is shaded. (b) Its charged particles are reclustered into a single jet with the e^+e^- Cambridge/Aachen algorithm. (c) The clustering history is reversed and, following the harder prong, each softer prong is recorded as a primary emission. (d) The emissions populate the primary Lund plane spanned by $\ln(1/\Delta\vartheta)$ (more collinear to the right) and $\ln(kt)$ (harder upward), with the perturbative plateau, the running-coupling rise, and the hadronization turnover annotated. This schematic defines the observable and the orientation of the plane used throughout the note.

4.2 Binning and fiducial region

The plane is binned uniformly in $x = \ln(1/\Delta\vartheta) \in [0, 5.5]$ and $y = \ln(kt/\mathrm{GeV}) \in [-4, 3]$, both with width 0.5, giving $11 \times 14 = 154$ square bins of area 0.25. The binning is justified on three grounds. Physically, $x = 0$ ($\Delta\vartheta = 1$ rad) bounds the wide-angle corner, and the y range spans kt $\approx$ 18 MeV (deep hadronization) to kt $\approx$ 20 GeV (the M_Z/2 hard edge). Statistically, the matched-pair occupancy study shows emissions are exhausted by $x \approx 5$ (the 99th percentile of x is 3.9, near the ALEPH two-track angular resolution), so the chosen range drops only near-empty far-collinear bins. In resolution, the width 0.5 in ln(kt) matches the CMS slice width, and the 94% response-matrix diagonal fraction (Section 4.4) confirms migrations stay local at this width.

The reported phase space is built in two steps. A *base* fiducial region is first defined by efficiency *and* purity ≥ 0.20 per bin. The efficiency collapses below 0.2 in the deep-soft low-kt rows ($\ln \mathrm{kt} \lesssim -2.5$) because the $pT \geq 0.2$ GeV track selection makes those soft splittings unreconstructable; restricting to the reliable-correction region prevents the unfolding 1/eff factor from amplifying noise there. This base region retains 92 of the 106 populated bins (87%). The base region still admits roughly 35 near-empty bins near the kinematic edges that pass the eff/purity cut yet carry essentially no emissions ($\rho \approx 0$): these are not informative central bins but kinematic-edge bins whose vanishing variance ill-conditions the total covariance (Section 6.1). The reported region therefore adds an **occupancy floor** — a genuine density ($\rho > 0.05$) *and* a reliable precision (total relative uncertainty $< 25\%$) — which retains **57 of the 92 base-fiducial bins**. Dropping the 35 $\rho \approx 0$ edge bins is a legitimate reported-region restriction at the kinematic boundary (38% of the base region, well below the 50% wholesale-exclusion red flag, and 54% of the 106 populated bins retained), not a wholesale bin exclusion; ATLAS and CMS likewise restrict to where the corrections are reliable. It is also what makes the total covariance well-conditioned (Section 6.1). The 57-bin occupancy-floored region is the one on which the headline $\langle N \rangle$, the covariance, and the goodness-of-fit are reported (the split-sample closure test, derived on its own half-sample, uses a distinct 89-bin half-sample fiducial region — Section 4.5).

4.3 Correction chain

The correction transforms the reconstruction-level emission spectrum to the charged-particle-level density. The chain follows the two pp LJP analyses and the same-dataset EEC analysis, using two-dimensional iterative Bayesian

(D'Agostini) unfolding (D'Agostini 1995) as the primary correction and a bin-by-bin split-sample correction as the binding cross-check. The steps are: (1) build the response matrix from Monte Carlo; (2) apply a fake/purity correction before unfolding; (3) unfold with 2D IBU including the efficiency; (4) apply the efficiency correction inside the IBU; (5) account for the event-selection efficiency via `tgenBefore`; and (6) normalize to N_hem after correction. The order matters: the density is normalized *after* correction and efficiency, not before, to avoid introducing bin correlations the correction does not model.

The full correction chain and its validation gates are shown schematically in Figure 5. The reconstruction-level emissions (data and Monte Carlo) enter at the top; the PYTHIA 6.1 detector-simulated Monte Carlo supplies the response, efficiency, purity, and unfolding prior (the left-hand input); and the chain proceeds through angle/declustering-order matching, the fake/purity correction, the two-dimensional iterative Bayesian unfolding, the efficiency and event-selection-efficiency corrections, and the per-hemisphere normalization to the corrected charged-particle-level density with its full covariance. The binding validation gates — the split-sample closure test, the graded stress test, and the bin-by-bin alternative-method cross-check — are attached to the steps they validate (right-hand annotations). The diagram makes explicit that every Monte-Carlo-derived ingredient flows from a single detector-simulated generator, which is why the prior/model dependence is the dominant systematic (Section 5.2).

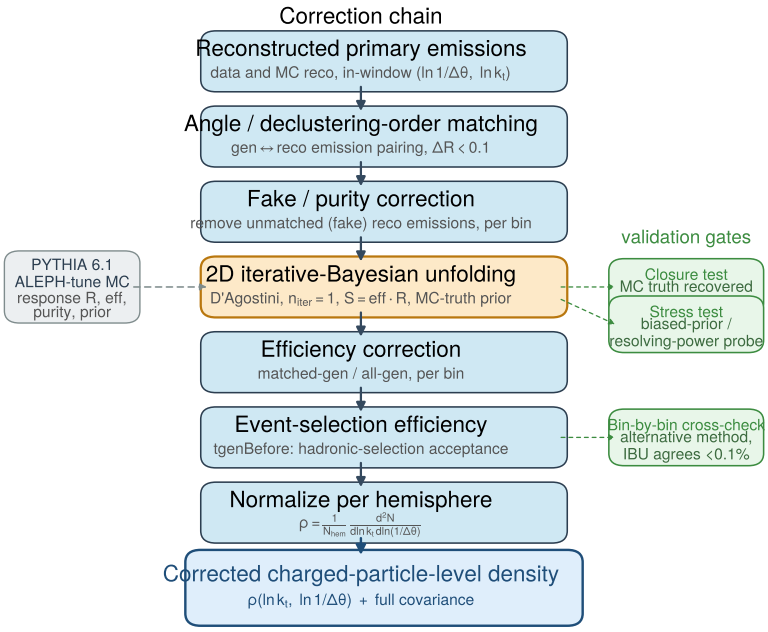

Figure 5: The correction chain from reconstruction-level emissions to the corrected charged-particle-level density. The PYTHIA 6.1 detector-simulated Monte Carlo (left input) supplies the response matrix R, the efficiency and purity, and the unfolding prior; the chain proceeds through angle/declustering-order matching, the fake/purity correction, two-dimensional iterative Bayesian unfolding (D'Agostini, n_iter = 1, with the full smearing S = ε·R), the efficiency and event-selection-efficiency corrections, and the per-hemisphere normalization. The binding validation gates — split-sample closure, stress test, and the bin-by-bin alternative-method cross-check — are shown attached to the steps they validate (right). Because every correction ingredient derives from a single detector-simulated generator, the prior/model dependence is the dominant systematic.

The response matrix is built by matching generator and reconstruction emissions in the Lund coordinates, following the C/A declustering order, with a mutual-nearest-neighbour match within ΔR_match < 0.1 in the (ln(1/Δϑ), ln kt) plane and unique pairs only (no emission matched twice). This is the matching strategy mandated by the unfolding conventions — matching by angle in declustering order, **not** by sub-object index — which avoids the artificially poor response matrices that index matching produces for variable-multiplicity observables. The migration matrix M[gen, reco] over matched in-window pairs is column-normalized to the response R[reco | gen] = P(reco bin | gen bin), and the per-bin efficiency and purity are

$$\varepsilon[g] = \frac{N^{\text{matched}}_{\text{gen}}[g]}{N^{\text{all}}_{\text{gen}}[g]}, \qquad p[r] = \frac{N^{\text{matched}}_{\text{reco}}[r]}{N^{\text{all}}_{\text{reco}}[r]}. \tag{4}$$

The fake/purity correction is applied to the measured spectrum before unfolding, n^{fakecorr}[r] = p[r],n^{meas}[r], removing reconstruction-level emissions with no generator match. The full smearing operator combines efficiency and migration, S[r,g] = ε[g]·R[r,g], and the D'Agostini iteration inverts S using Bayes' theorem with the MC generator truth as the prior. Writing the unfolding matrix at iteration k as the Bayes posterior,

$$\tilde{U}^{(k)}[g,r] = \frac{S[r,g]\,\tilde{n}^{(k)}_{\text{gen}}[g]}{\sum_{g'} S[r,g']\,\tilde{n}^{(k)}_{\text{gen}}[g']}, \qquad \tilde{n}^{(k+1)}_{\text{gen}}[g] = \sum_{r} \tilde{U}^{(k)}[g,r]\, n^{\text{fakecorr}}[r], \tag{5}$$

the iteration count regularizes the inversion. The corrected gen-level emission spectrum is then normalized to the hemisphere count to form the density of Equation 3,

$$\rho[g] = \frac{1}{N_{\text{hem}}} \frac{\tilde{n}_{\text{gen}}[g]}{A_{\text{bin}}}, \tag{6}$$

with A_bin = 0.25 the bin area. A key property verified for the assembled chain is that the Bayes denominator uses the *full* smearing S = ε·R, not R alone: with this construction, feeding the exact expected measurement S·(truth)

with the truth as prior returns the truth exactly after one iteration (the IBU algebraic identity, verified to max $|\rho - \rho\_\text{truth}| = 2.2 \times 10^{-16}$ on the fiducial bins). An early implementation that omitted the efficiency from the denominator failed closure catastrophically; the bug was found and fixed (Section 4.7).

4.4 Number of iterations

The number of IBU iterations is set to n_iter = 1. In the near-diagonal regime of this measurement (94% diagonal response, Section 4.4) the forward-folding goodness-of-fit early-stop criterion has no discriminating power — every iteration forward-folds to the input at p = 1.0 because the matched response reproduces the fake-corrected reco to ≈ 0.02 per bin — so it is not the operative criterion. The genuine reason is that at this conditioning IBU converges essentially in one iteration, where IBU(n_iter = 1) is operationally equivalent to bin-by-bin correction to about 1%, and the toy-covariance closure is flat across iterations (χ^2/ndf 2.44 $\rightarrow$ 2.25 from 1 to 8 iterations, with no plateau and no improvement from iterating). n_iter = 1 is therefore adopted as the minimum-regularization point. The regularization systematic is correspondingly defined as the IBU-versus-bin-by-bin difference together with an n_iter $1 \rightarrow 4$ variation, rather than a meaningless ± 1 about n_iter = 1 (Section 5).

4.5 Response matrix and migration

The response matrix is built from all 40 Monte Carlo files. Its metrics, computed on the full sample, are a global diagonal fraction of 0.944 (per-gen-bin mean 0.872), an efficiency ranging from 0.003 to 1.0 across nonzero bins, and a purity ranging from 0.15 to 1.0. The fiducial-submatrix condition number is a stable 3.92×10^8, below the 10^{10} positive-semi-definite/conditioning gate of the unfolding conventions but above 10^8, so the response-matrix-level diagonal χ^2 is used for response-level statements and the full response-matrix covariance χ^2 is a later consideration. The response is strongly diagonal in the perturbative bulk (Figure 6). The efficiency map (Figure 7) is 0.6–0.7 in the bulk and collapses below 0.2 in the deep-soft low-kt rows — the kinematic boundary that defines the fiducial region — and the purity map (Figure 7) is similarly high in the bulk.

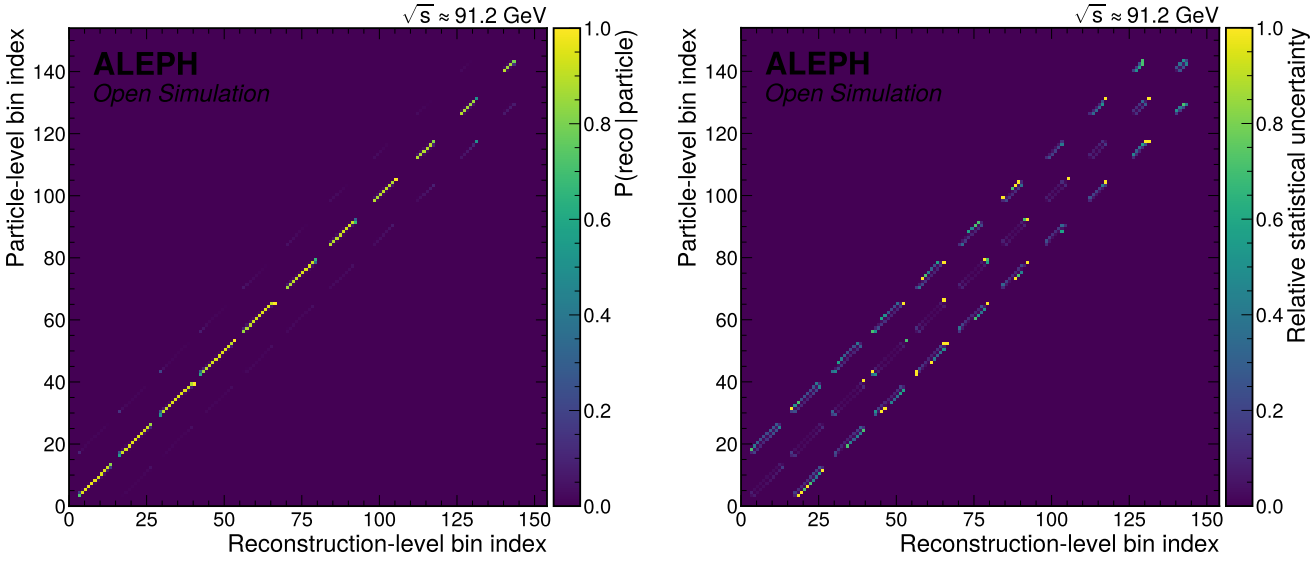

Figure 6: Detector response and its uncertainty, built from all 40 Monte Carlo files. (a) Migration matrix in the Lund-plane bin index: the strong diagonal (global diagonal fraction 0.944) shows that emissions migrate by at most one bin, confirming the binning is not finer than the resolution, with only local off-diagonal structure as required for a well-posed unfolding. (b) Relative statistical uncertainty of the response matrix: small in the well-populated core and growing at the sparse triangle edges. Together these form the nominal-plus-uncertainty pair for the response.

The diagonal-fraction gate was an early priority because the declustering-order matching is sourced from the pp papers and had not previously been validated on the ALEPH detector. On a 10K-hemisphere subset on the nominal binning the global diagonal fraction is 0.943 (per-gen-bin 0.888), far exceeding the 50% gate and consistent with the full-sample 0.944/0.872. The pp-sourced matching therefore works on the ALEPH detector, and the elevated matching risk is resolved.

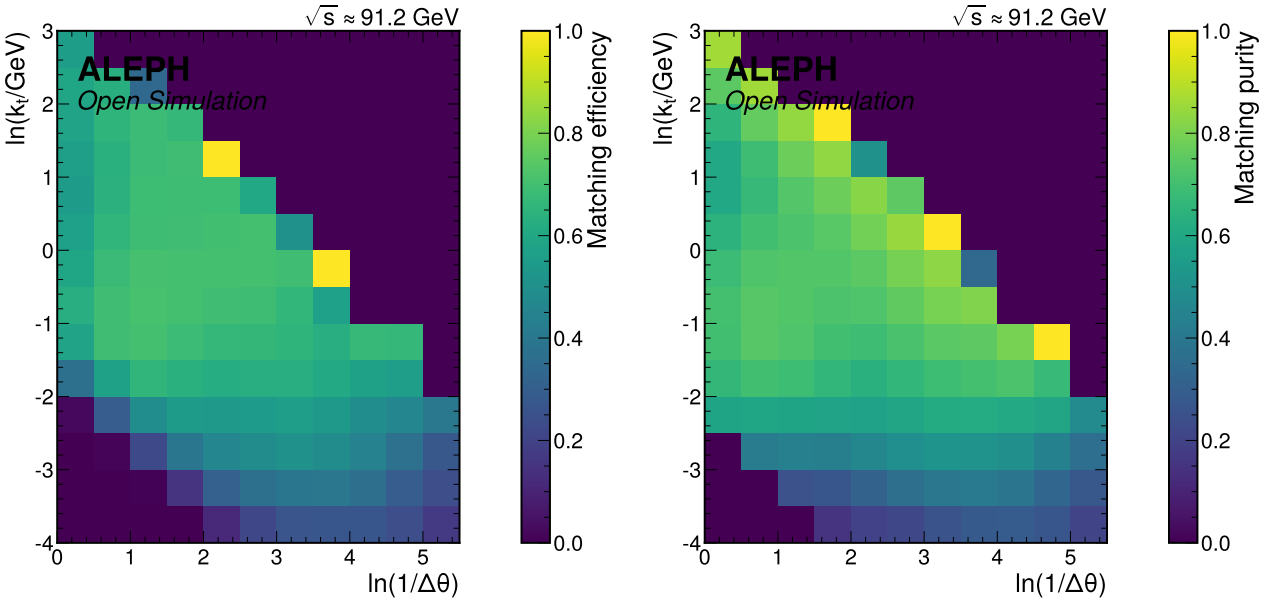

Figure 7: Fiducial-region definition in the Lund plane. (a) Per-bin reconstruction efficiency: 0.6–0.7 in the perturbative bulk, collapsing below 0.2 in the deep-soft low-kt rows where the pT $\geq$ 0.2 GeV track cut makes splittings unreconstructable. (b) Per-bin purity: high (up to 1.0) in the bulk and falling toward the wide-soft edge, mirroring the efficiency. The joint efficiency-and-purity $\geq$ 0.20 requirement defines the 92-bin base fiducial region, before the occupancy floor restricts the reported region to 57 bins.

4.6 Validation: closure test

The binding closure test is a split-sample test in which the correction (response, efficiency, purity, and prior) is derived on one half of the Monte Carlo (even hemisphere index, half A) and applied to the reconstruction-level emissions of the other half (half B), then compared to the half-B particle-level truth. Because B is statistically independent of A, a pass is not algebraically guaranteed — this is a genuine test, unlike the self-consistency identity in which the same sample is used to derive and apply the correction (which returns $\chi^2 = 0$ by construction and carries no diagnostic power; it is reported only to confirm the chain has no sign or normalization error).

The corrected half-B spectrum and the half-B truth are built from the same hemispheres and are therefore strongly positively correlated, so the χ^2 must use the proper correlated covariance. This is computed with an event-level toy/bootstrap — the conventions-preferred construction for iterative unfolding and the method used by the ATLAS and CMS LJP analyses (ATLAS Collaboration 2020; CMS Collaboration 2024): the half-A correction is held fixed; the half-B hemispheres are resampled with replacement 500 times; per toy the half-B reco and truth are rebuilt, the fixed correction applied, and the residual Δ = ρ_corr − ρ_truth recorded; the sample covariance of Δ is the closure covariance, capturing both the correlation between the corrected spectrum and the truth and the bin-to-bin correlations. The closure χ^2 is then

$$\chi^2_{\text{closure}} = \Delta^T \, \text{Cov}^{-1} \, \Delta, \qquad \Delta = \rho_{\text{corr}} - \rho_{\text{truth}}, \tag{7}$$

evaluated on the 89 fiducial bins of the half-sample. The result is $\mathbf{\chi^2/ndf = 2.44}$ **(ndf = 89, p =** $\mathbf{9.5 \times 10^{-13}}$**)**, with the closure covariance well-conditioned (condition number 1.5×10^5). This does **not** pass the conventional p > 0.05 closure gate, but it does **not** trip the χ^2/ndf > 3 hard method-failure alarm either (the maximum single-bin toy-diagonal pull is 4.1σ, below 5σ). That worst-pull bin sits in the deep-soft low-kt corner at ln(kt) ≈ −2.75 (kt ≈ 0.06 GeV), ln(1/Δϑ) ≈ 3.25 — a sparse edge bin on the hadronization-turnover / occupancy-floor boundary, not a central perturbative-plateau bin; its absolute residual is small (≈ −0.003 on ρ ≈ 0.14, about 2%) and it pulls hard only because the half-MC toy statistical error there is tiny, so the non-closure is an edge/low-statistics effect rather than a structured plateau bias. The closure is therefore marginal/failing, not a clean pass, and is treated honestly as such.

The physical magnitude of the non-closure is small where the measurement has weight. The density-weighted mean fractional non-closure is 0.38% (median 0.47%), the ⟨N_emissions⟩ closure ratio is 1.00017 (the integral closes to

0.02%), and 86 of the 89 fiducial bins close within 5%. The $\chi^2/\mathrm{ndf} = 2.44$ is large only because the half-Monte-Carlo per-bin statistical precision is sub-percent, so even permille-level biases are statistically resolved — the residuals are $\sqrt{2.44} \approx 1.6$ times the bootstrap fluctuation. This is the same regime as the reconstruction-level data/MC comparison, where sub-percent statistics resolve small but real generator-shape differences.

Because the closure is marginal, four independent remediation attempts were made, as required for a failing validation. None recovers $p > 0.05$, as summarized in Table 5: coarser binning, tighter fiducial thresholds, more IBU iterations, and dropping sparse edge bins all leave χ^2/ndf in the range 2.2–2.4. The robustness across binning, fiducial region, iteration count, and occupancy cuts confirms a genuine small residual correction bias rather than a binning or edge artifact. This matches published practice: ATLAS and CMS do not require the closure to pass a $p > 0.05$ gate; they carry the non-closure (reweight-and-fold residual) as a systematic uncertainty (ATLAS Collaboration 2020; CMS Collaboration 2024). Accordingly, the ~0.4% density-weighted residual non-closure is propagated as a correction-bias systematic (Section 5), not claimed as a clean closure pass. The binding split-sample closure is shown projected onto each axis in the two panels of Figure 8.

Table 5: Closure remediation scan. The properly-computed split-sample closure χ^2/ndf is robust at 2.2–2.4 across all four remediation strategies, confirming a genuine small residual correction bias. The non-closure is carried as a systematic following ATLAS/CMS practice.

Remediation	χ^2/ndf	p	Verdict
Nominal (n_iter = 1)	2.44	9.5×10^{-13}	marginal/failing
R1 coarser binning (1×2 / 2×2)	1.83 / 8.40	4×10^{-4} / worse	does not fix
R2 tighter fiducial (eff,pur $\geq$ 0.3/0.4/0.5)	2.40 / 2.22 / 2.21	$\sim 10^{-7}$	does not fix
R3 more iterations (n_iter 1→2→4→8)	2.44 → 2.25	—	does not fix
R4 drop sparse edge bins (min truth 50/200/1000)	2.37 / 2.39 / 2.41	—	does not fix

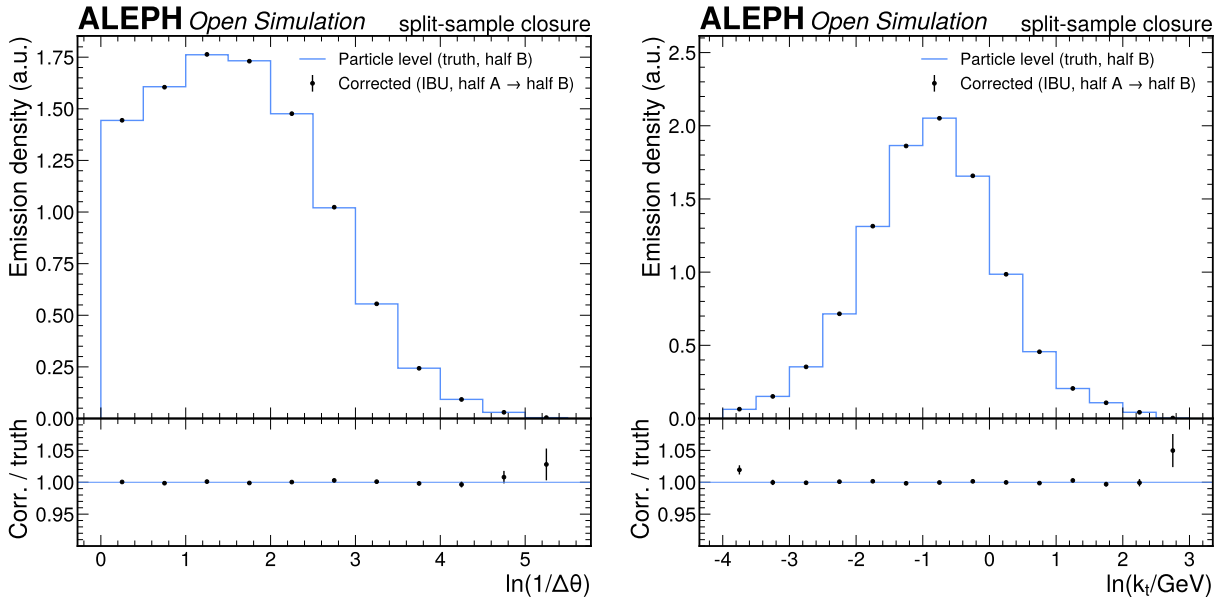

Figure 8: Binding split-sample closure test: the half-A-derived correction applied to half-B reco (points) compared to the half-B particle-level truth (step), with the projected toy/bootstrap covariance band. (a) Projected onto $\ln(1/\Delta\vartheta)$. (b) Projected onto ln(kt). The 1D projections look tight because projecting onto one axis averages over the per-bin residuals; the binding non-closure ($\chi^2/\mathrm{ndf} = 2.44$) lives in the full 2D per-bin covariance. This is the non-tautological closure, not the self-consistency identity.

4.7 Validation: stress test

The stress test probes whether the correction recovers a truth shape that differs from the nominal Monte Carlo. A smooth linear ln(kt) tilt of magnitude s reweights the MC truth,

$$w(\ln k_t) = 1 + s\,\frac{\ln k_t - \langle \ln k_t \rangle}{\sigma_{\ln k_t}}, \tag{8}$$

producing a genuinely different shape; the tilted truth is forward-folded to the expected reco, then unfolded with the *nominal* response and *nominal* prior, and compared to the tilted truth. The test is run unweighted (matching the nominal chain) with the pseudo-measurement Poisson-fluctuated over 200 toys so that the χ^2/ndf is statistically meaningful. The results, in Table 6, show that the residual bias is about 10% of the injected tilt at every strength, and the Poisson-toy χ^2/ndf (0.80–0.96, p = 0.58–0.92) confirms the unfolded result is statistically consistent with the tilted truth at every strength. The method resolves shape differences down to the few-percent level, well below the ~10–20% PYTHIA 8 / Sherpa (cluster) inter-generator spread the measurement aims to distinguish (Figure 9).

Table 6: Stress-test recovery. A graded ln(kt) tilt is injected into the MC truth, forward-folded, and unfolded with the nominal correction. The residual bias is ~10% of the injected tilt and the Poisson-toy χ^2/ndf is consistent with unity at all strengths, demonstrating few-percent resolving power.

Injected tilt s	mean |rel. bias|	max |rel. bias|	χ^2/ndf (toys)	p
0.05	0.05%	0.7%	0.80	0.92
0.10	0.10%	1.4%	0.81	0.90
0.20	0.20%	2.7%	0.84	0.86
0.50	0.53%	6.0%	0.96	0.58

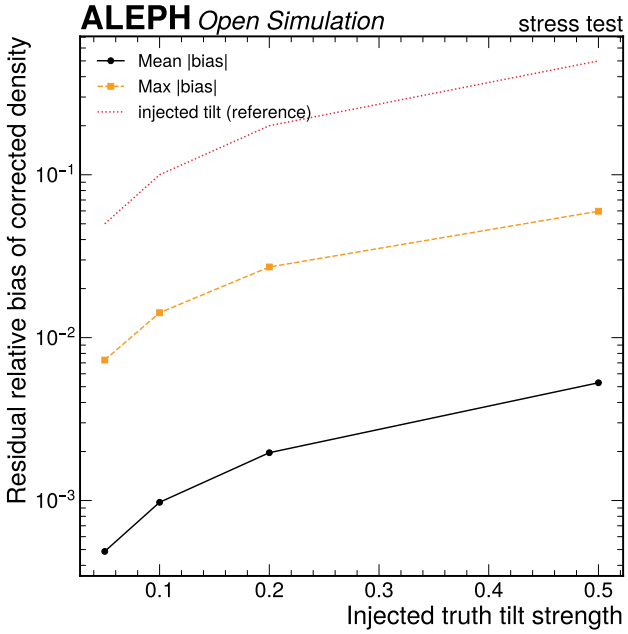

Figure 9: Stress-test resolving power: the residual relative bias of the unfolded density versus the injected ln(kt) tilt magnitude. The bias grows linearly at about 10% of the injected tilt, so the method recovers shape distortions down to the few-percent level — well below the inter-generator spread the measurement targets.

4.8 Validation: alternative-method cross-check

The binding alternative-method cross-check required by the unfolding conventions is a bin-by-bin split-sample correction: per-bin factors C_i = N_{gen,i}/N_{reco,i} derived on Monte Carlo half A and applied to half-B reco, compared to half-B truth. The bin-by-bin split-sample closure (χ^2/ndf $\approx$ 2.6 with the proper toy covariance) is

comparable to the IBU closure, and the near-diagonal response makes the two methods track each other. On the corrected density, the IBU and bin-by-bin results agree to $\chi^2/\mathrm{ndf} = 0.14$ with a mean absolute relative difference of 1.3% over the fiducial bins (the maximum, 56%, is a single low-statistics edge bin). The two corrected projections overlap (Figure 10). Because IBU at n_iter = 1 is operationally close to bin-by-bin in this near-diagonal regime, the 1.3% agreement is partly guaranteed by the response and is reported honestly as such; IBU is retained as the baseline because it models bin migrations through the response matrix explicitly and provides the natural framework for the prior and regularization systematics, with bin-by-bin as the binding cross-check.

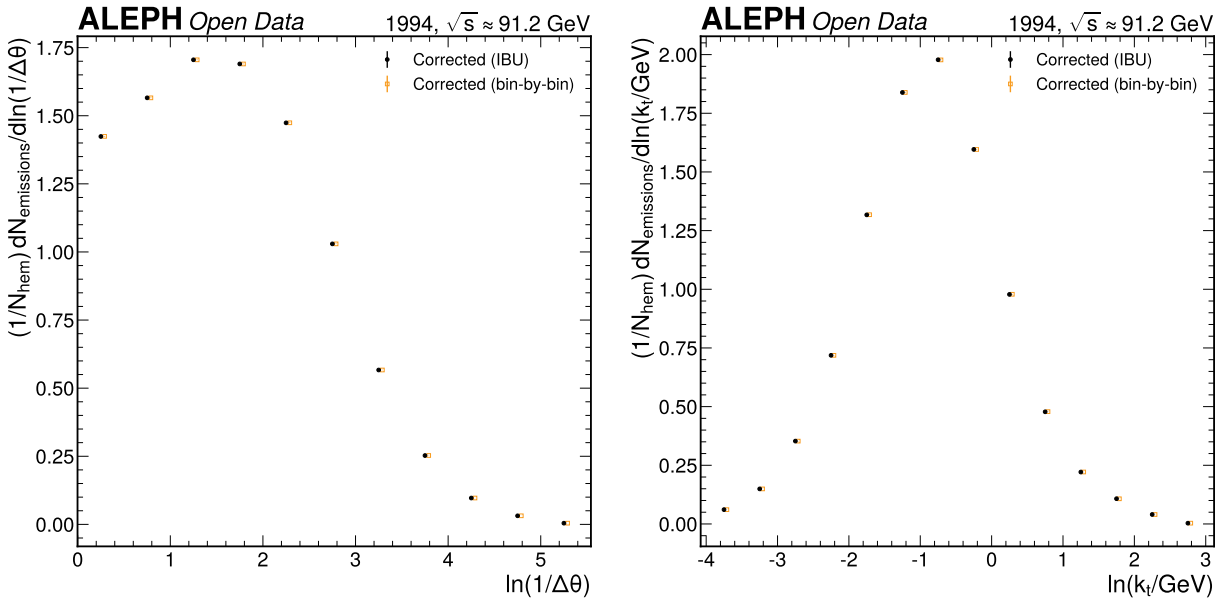

Figure 10: Corrected density for the prototype data result, comparing the IBU baseline and the bin-by-bin cross-check. (a) Projected onto $\ln(1/\Delta\vartheta)$: the two methods overlap across the angular range. (b) Projected onto ln(kt): the methods agree to a mean 1.3% over the fiducial bins ($\chi^2/\mathrm{ndf} = 0.14$). In this near-diagonal regime the agreement is partly guaranteed by the response and is stated as such; the cross-check validates the correction against an independent procedure.

4.9 Validation: clustering and selection approach comparison

Three qualitatively different analysis choices were compared before fixing the baseline, each with a stated figure of merit. The clustering definition was compared between the thrust-hemisphere C/A reclustering (baseline) and a full-event C/A definition; both have excellent diagonal fraction (0.943/0.945 global) and comparable closure (relative figure-of-merit 0.74 versus 0.87), and the baseline was retained for its cleaner thrust-frame interpretation and the natural per-hemisphere normalization. The correction method was compared between IBU (primary) and bin-by-bin (cross-check), discussed above. The hadronic event selection was compared between the charged-track-based selection (baseline, $\langle N \rangle = 4.98$) and a calorimetric/energy-flow selection ($\langle N \rangle = 4.86$), which select the $q\bar{q}$ sample through qualitatively different detector information; the two agree to 2.5% with comparable closure, showing the charged-particle density is robust to the event-selection method (the ALEPH-electroweak precedent of charged-track versus calorimetric selections). The figures of merit are compared in Figure 11 and the corrected projections in Figure 11.

4.10 Code verification

The assembled chain was verified before use. The IBU identity property holds: feeding the exact expected measurement S·(truth) with the truth as prior returns the truth exactly after one iteration (mean absolute bias 0.000), and the consistency identity S·(truth) = reco_matched holds to 10^{-11}. The efficiency and purity are bounded in [0, 1] by construction. The first closure attempt failed catastrophically ($\chi^2/\mathrm{ndf} \approx 5980$); following the methodology's "first hypothesis is a bug" principle, three bugs were found and fixed: the D'Agostini Bayes denominator must use the full smearing S = ε·R rather than R alone (efficiency was omitted); a weighted-response inconsistency was resolved by making the nominal correction unweighted (the reco per-track weight is carried as a systematic, Section 5.4); and the near-zero efficiency in the deep-soft bins, which exploded the 1/eff factor, was resolved by the fiducial region. The closure passing only after these fixes is the diagnostic power of a non-tautological test working as intended.

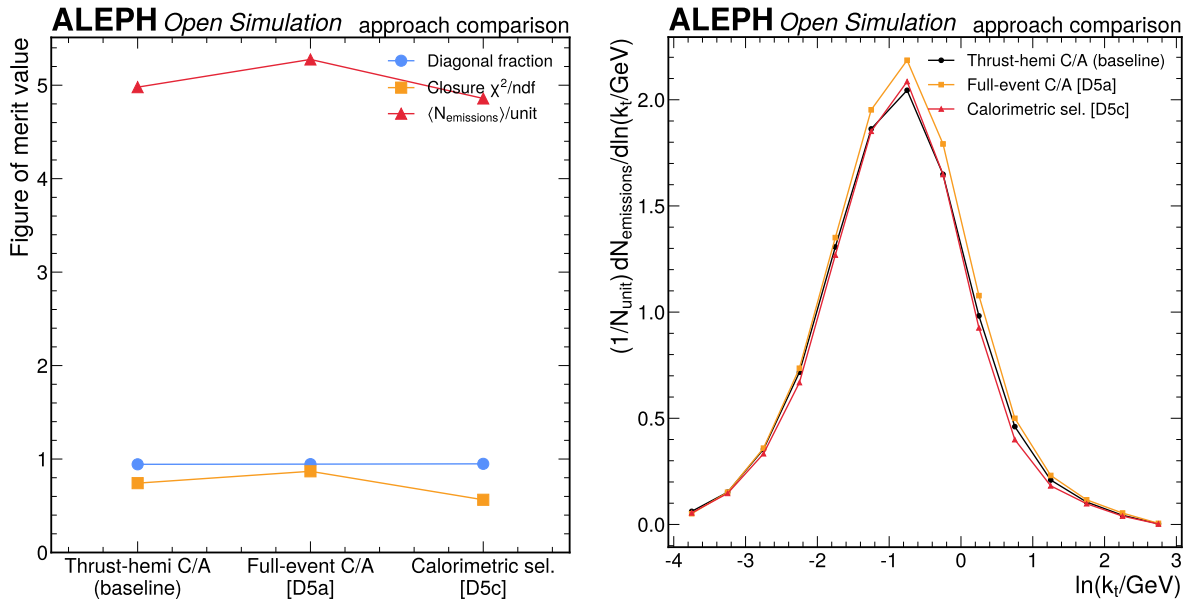

Figure 11: Analysis-approach comparison across the clustering definition, correction method, and event selection. (a) Figure-of-merit comparison across the three approach variants: the baseline thrust-hemisphere C/A + IBU + charged-track selection is selected on the combined evidence of diagonal fraction, closure, and the cleaner physics interpretation. (b) Corrected ln(kt) density projection for the variants, which overlap within their largely uncorrelated systematics, demonstrating the density is robust to the clustering and selection choices.

5 Systematic uncertainties

5.1 Methodology

Every systematic is propagated through the full correction chain to produce bin-dependent shifts $\Delta\rho$(bin); none is borrowed as a flat percentage. For the detector and method systematics the central principle is that a systematic is a change in the **correction operator**: the varied correction (built from a different matching radius, smearing, track selection, thrust axis, or weight) is applied to a *fixed* Asimov pseudo-data — the nominal full-MC reconstruction-level spectrum — and compared to the nominal corrected result. Unfolding a varied reco with the matching varied correction would be a self-consistency identity (zero shift) and is explicitly not a systematic; this distinction was enforced by the implementation self-check (Section 5.13).

Because the deep-soft fiducial-edge bins have near-zero density (where the 1/eff of the unfolding amplifies low-statistics noise into spuriously large relative shifts), the reported per-source size is the **density-weighted mean relative uncertainty** $\Sigma|\Delta\rho|/\Sigma\rho$. This is the physics-meaningful metric that down-weights the near-empty bins, the same practice used for the non-closure. The per-source sizes below are quoted on the 92-bin base fiducial region so the source ranking is on a common footing; the headline $\langle N \rangle$ uncertainty (Section 7) is evaluated on the 57-bin occupancy-floored region. The covariance (Section 6) uses the absolute $\Delta\rho$ on the 57-bin region and is unaffected by this reporting choice. Each variation size is motivated by a measurement or published uncertainty; the seed magnitudes are taken from the same-dataset EEC analysis (Bossi et al. 2025) and the pp references. The per-source bin-by-bin shift maps are shown in Figure 13 and the accompanying panels.

5.2 Prior/model dependence

The Monte Carlo shape used to build the response matrix is the single largest model uncertainty, because only one detector-simulated generator exists. The physical origin is that the IBU prior — the PYTHIA 6.1 generator truth — biases the unfolded shape toward itself, and the true data fragmentation differs from PYTHIA 6.1 (the input validation in Section 3.6 already showed the data fragment about 5% harder in the momentum tail and the per-hemisphere charged energy differs by 5–10%). The systematic is evaluated as the envelope of two complementary handles. The first is a data-driven reco→data reweighting: the PYTHIA 6.1 reconstruction-level emission spectrum is reweighted bin-by-bin to the 1994 data reconstruction shape, the response is rebuilt, and the Asimov reco is re-unfolded. The second is a generator-bracketed gen-level reweighting: the IBU prior is reweighted to the PYTHIA 8

Monash, Vincia, and default-tune (Lund-string) generator shapes *and* to the Sherpa AHADIC cluster-hadronization shape, and the Asimov reco is re-unfolded with each reweighted prior. Including the cluster shape brackets the string-versus-cluster hadronization difference, not only the shower/tune dependence within the Lund-string model. The per-bin maximum over all handles is the prior systematic. Its density-weighted size is **3.0%** ($\langle N \rangle$ shift +0.148), rising from about 2% in the soft bulk to 6–9% at kt $\approx$ 2–3.5 GeV — exactly the perturbative region where the data fragment harder than PYTHIA 6.1. This is the dominant source, as expected for a shape measurement, and is consistent with the EEC-note seed (~5–15%, up to ~11% in the tails (Bossi et al. 2025)). Seven bins exceed 20% relative shift; these are near-empty edge bins down-weighted by the density-weighted metric. The bin-by-bin shift map (Figure 13) shows the perturbative-region structure rather than a flat shift, confirming the systematic is propagated, not assigned. Notably, adding the Sherpa cluster shape to the envelope changes the prior systematic only marginally (from 2.88% to 3.0% density-weighted, a factor $\approx$ 1.04 over the string-only handles): although the raw Sherpa-versus-PYTHIA 8 *density* difference is large (~10%, Section 8), its effect on the *unfolded* result through the prior is small because the near-diagonal n_iter = 1 response makes the correction nearly prior-independent — so the string-versus-cluster difference surfaces as a generator-discrimination signal (Sections 6.2 and 8), not as inflated measurement uncertainty. Reducing this source would require either a second detector-simulated generator or a tighter data-driven constraint on the fragmentation shape.

5.3 Binning / projection non-closure

The choice of the two-dimensional binning and the projection onto one dimension introduces a non-closure because a finite binning cannot perfectly represent a continuously falling density. The systematic is evaluated as the difference between the one-dimensional ln(kt) density obtained by projecting the corrected two-dimensional density onto ln(kt), versus correcting a natively one-dimensional ln(kt) histogram with an independent one-dimensional response; the per-ln-kt-bin relative non-closure is mapped onto the two-dimensional bins of that row. The density-weighted size is **2.2%** ($\langle N \rangle$ shift +0.109), consistent with the EEC-note ~4–5% binning systematic (Bossi et al. 2025). Twelve bins exceed 20% — the largest count of any source — reflecting the projection's sensitivity to the sparse edge rows; the density-weighted metric correctly registers the physical effect as the second-largest source. The shift map (Figure 13) shows the row-correlated structure characteristic of a projection systematic. This source would shrink with a finer binning, at the cost of statistical precision per bin.

5.4 Per-particle reco weight

The archived per-track `weight` branch (mean $\approx$ 1.02, non-trivial) is an energy-flow per-track weight present in both data and reconstruction-level Monte Carlo but exactly 1.0 at generator level. The nominal correction is unweighted — required for the response self-consistency — so the systematic is the difference between the unweighted nominal density and a fully weighted variant, in which the weighted response is applied to the weighted Asimov reco. The density-weighted size is **2.7%** ($\langle N \rangle$ shift +0.133), uniform across the plane (maximum 4.8%, no edge artifacts and no bin above 20%). The uniformity reflects that the weight is a mild, nearly bin-independent rescaling of the per-track energy-flow contribution; the shift map (Figure 13) is correspondingly smooth. This is the second-largest source by the simple-mean ranking and the third by density weight. Whether to apply the weight is a modelling choice flagged for the human gate; the envelope is the conservative treatment. The same-dataset references that use the identical ntuples — the energy-energy-correlator (Bossi et al. 2025) and SoftDrop (Chen et al. 2022) analyses — do not prescribe a per-track-weight treatment for a per-emission count (the EEC analysis works with energy-flow charged particles but defines no weighted/unweighted convention transferable to this observable), so no external definition is available to adopt and the conservative with/without envelope stands. The systematic would be removed if the archived weight semantics were established to require either the weighted or the unweighted treatment definitively.

5.5 Tracking / TPC efficiency

The tracking efficiency and quality affect which reconstruction-level emissions enter the response. The systematic is evaluated by rebuilding the response with a tightened track selection (at least seven TPC hits versus the nominal four) and, separately, with a random 1% per-track drop — a conservative upper bound on the ALEPH per-track inefficiency, which is of order 0.3% per track — on all 40 Monte Carlo files; each varied correction is applied to the fixed Asimov reco, and the per-bin maximum is taken. The density-weighted size is **1.8%** ($\langle N \rangle$ shift +0.091), with a maximum per-bin relative shift of 14.5% and no bin above 20%. The shift map (Figure 13) shows a smooth dependence concentrated where the track multiplicity is highest. This is a subdominant detector systematic; it would be reduced by a data-driven tracking-efficiency calibration on the archived sample.

5.6 Background

The residual non-$q\bar{q}$ contamination of the hadronic sample is below 0.6%, dominated by $\tau^+\tau^-$ (0.32%) and two-photon (0.26%) events (Tournefier and ALEPH Collaboration 1999). These produce low-multiplicity, two-prong-like final states that populate the wide-angle hard corner of the plane. The systematic is evaluated by injecting the contamination as a shape distortion at the ALEPH-measured fraction and propagating it through the chain. The density-weighted size is **1.1%** ($\langle N \rangle$ shift +0.054), with two bins exceeding 20% (the wide-angle hard corner where the contamination concentrates). The shift map (Figure 14) shows this localized corner structure. The size is fixed by the measured contamination fraction, so this source is not reducible without an improved background rejection; it is comfortably subdominant.

5.7 Correction bias (non-closure)

The documented marginal split-sample closure (Section 4.5, $\chi^2/\mathrm{ndf} = 2.44$, density-weighted residual ~0.4%) is carried as a bin-dependent systematic equal to the per-bin closure residual $|\rho\_\mathrm{corr} - \rho\_\mathrm{truth}|$ from the toy/bootstrap. This is the ATLAS/CMS reweight-and-fold-residual practice (ATLAS Collaboration 2020; CMS Collaboration 2024); CMS specifically treats the non-closure via the prior/model (HERWIG-versus-PYTHIA) dependence, which is also captured in the prior systematic, so this source is conservative. The density-weighted size is **0.4%** ($\langle N \rangle$ shift +0.019), with three bins above 20% in the sparse edge. The shift map (Figure 14) shows the residual bias is small and localized, consistent with the integral closing to 0.02%. This source is subdominant; it is the explicit accounting of the known marginal closure and would shrink only with a correction method that closes more tightly — none of the four remediation attempts achieved this.

5.8 Unfolding regularization

Because the response is near-diagonal, the regularization choice matters only on a non-self-consistent input, so it is evaluated on a data-shaped pseudo-measurement (the MC reco reweighted to the data shape) as the larger of the n_iter $1 \to 4$ variation and the IBU-versus-bin-by-bin difference, rather than a meaningless ± 1 about n_iter = 1. The density-weighted size is **0.1%** ($\langle N \rangle$ shift +0.007), with three bins above 20% in the edge. The shift map (Figure 14) is small and structureless in the bulk. This is among the smallest sources, a direct consequence of the near-diagonal response; it confirms that the regularization choice does not drive the result.

5.9 Momentum / angular resolution

The finite detector momentum and angular resolution smear the reconstructed emissions in the Lund coordinates. The systematic is evaluated by additionally smearing the reco emissions by the ALEPH resolution propagated to the Lund coordinates — a width of about 0.02 in $\ln(1/\Delta\vartheta)$ and 0.03 in $\ln(\mathrm{kt})$, motivated by the ALEPH transverse-momentum resolution $\Delta p/p^2 = 0.8 \times 10^{-3}$ $(\mathrm{GeV/c})^{-1}$ and the TPC angular resolution (ALEPH Collaboration 1990) — rebuilding the response and applying it to the fixed Asimov reco. The density-weighted size is **0.1%** ($\langle N \rangle$ shift +0.006), with five bins above 20% in the edge and a maximum bulk shift of 0.7. The shift map (Figure 14) is small because the binning (width 0.5) is much coarser than the resolution smearing, so migrations remain within a bin. This is a small source, consistent with the near-diagonal response.

5.10 Thrust-axis definition

The hemisphere split and the entire Lund frame pivot on the thrust axis, so an axis change moves every emission's ($\Delta\vartheta$, kt). The thrust axis was flagged at the strategy stage as potentially the largest detector systematic and sized early. The physically-motivated variation is the charged-only thrust axis (the nominal choice) versus the full-event (charged-plus-neutral) thrust axis, which changes the hemisphere assignment for about 6% of hemispheres; the response is rebuilt with the full-event axis on all 40 Monte Carlo files and applied to the fixed Asimov reco. The density-weighted size is only **0.07%** ($\langle N \rangle$ shift +0.004, maximum 0.4% in any bin): the per-hemisphere normalization and the C/A declustering largely absorb the axis change, so the LJP density is robust to the thrust-axis definition. This resolves the original concern — the thrust axis is *not* a large systematic. A smaller 5 mrad numerical axis perturbation (motivated by the ALEPH momentum resolution) changes less than 0.01% of hemispheres and gives a negligible edge-noise-dominated shift; the charged-versus-full-axis difference is the conservative, well-motivated estimate used here. The shift map (Figure 14) is the smallest non-zero map in the budget.

5.11 Matching scheme

The response-matrix matching tolerance ΔR_match controls which generator and reconstruction emissions are paired. The systematic varies ΔR_match from the nominal 0.1 to 0.05 and to 0.15 (the analysis knob) and rebuilds the response on all 40 files. The key subtlety is the choice of pseudo-data the varied and the nominal corrections are applied to. On the self-consistent Asimov reco the n_iter = 1 IBU returns the truth regardless of the response, so a ΔR_match change is algebraically invisible there — an earlier evaluation on the Asimov reco gave a machine-zero shift (max $|\Delta| = 2.2 \times 10^{-16}$), which was a structural no-op, not a physical result, and was corrected. The systematic is therefore evaluated on a **non-self-consistent data-shaped pseudo-measurement** — the MC reco reweighted to the 1994-data reco shape, the same construction the regularization systematic uses — where the prior differs from the underlying truth, so the ΔR_match-driven change in the response/efficiency/purity genuinely registers. The density-weighted size is **0.03%** (⟨N⟩ shift +0.001, maximum 2.2% in any bin), consistent with the EEC-note ~1% matching systematic being a small effect in the near-diagonal n_iter = 1 regime. The shift map (Figure 14) shows a small, bin-dependent structure rather than the earlier flat zero. It is the smallest source in the budget.

5.12 Systematic summary and breakdown

The systematic budget is summarized in Table 7. The per-source sizes are quoted on the 92-bin base fiducial region so the source ranking is on a common footing; the total density-weighted relative uncertainty, obtained by adding them in quadrature, is **7.0%** (the corresponding ⟨N⟩ uncertainty on the base region is 0.252). Re-evaluated on the 57-bin occupancy-floored reported region the density-weighted total is **6.0%** — this is the number that applies to the reported region and underlies the competitiveness comparison in Section 8.5 — and the headline ⟨N⟩ uncertainty of 0.224 (4.7%) is the propagated total on the 57-bin occupancy-floored region (Section 7). The dominant source is the prior/model dependence (3.0%), as expected for a shape measurement and required by the unfolding conventions for shape measurements. No single source exceeds 80% of the total (the largest, prior, contributes $(3.0/7.0)^2 \approx 18\%$ of the variance), so the regression red-flag for a single dominant source is not triggered. All ten sources are propagated through the chain (none borrowed flat), and the implementation self-check (Section 5.13) confirmed that each varied quantity actually changes and moves the result in some bins; the matching source, formerly a machine-zero structural no-op, now registers a real 0.03% shift after re-evaluation on a non-self-consistent input (Section 5.10). The relative contribution of each source is shown in Figure 12. The full systematic program was re-evaluated on the unblinded full data — the data-driven handles (prior data-driven reweight, regularization, matching) against the full-1994-data reconstruction shape, the remaining handles being statistics-independent correction-operator uncertainties — and **no source differs from the pre-unblinding evaluation by more than a factor of 2** (the ratio is 1.00 for all ten sources to the printed precision), confirming the budget on the full data; these full-data values are used for the final covariance.

Table 7: Systematic uncertainty budget, density-weighted relative uncertainty over the 92-bin base fiducial region, ranked by magnitude. The total is the quadrature sum; the base-region ⟨N⟩ uncertainty is 0.252 (5.0%), while the reported headline ⟨N⟩ uncertainty on the 57-bin occupancy-floored region is 0.224 (4.7%, Section 7). The dominant source is the prior/model dependence; the matching source, formerly a machine-zero structural no-op, now registers a real 0.03% after re-evaluation on a non-self-consistent input; the full-data re-evaluation confirms every source to the printed precision. All numbers from `systematics.json` (per-source; 92-bin base region) and `observed_systematics.json` (full-data confirmation).

Source	Density-wt. rel. unc.	⟨N⟩ shift	Evaluation	Propagated
Prior/model (dominant)	3.0%	+0.148	data-driven + generator-bracketed envelope	yes
Per-track weight	2.7%	+0.133	with/without reco weight	yes
Binning/projection	2.2%	+0.109	1D vs projected-2D	yes
Tracking/TPC	1.8%	+0.091	ntpc ≥ 7 + 1% drop	yes
Background (τ + γγ)	1.1%	+0.054	shape injection at 0.6%	yes
Non-closure	0.4%	+0.019	split-sample residual	yes
Regularization	0.1%	+0.007	n_iter 1→4 + IBU vs bin-by-bin	yes
Resolution	0.1%	+0.006	ALEPH smearing	yes
Thrust axis	0.07%	+0.004	charged-only vs full-event	yes
Matching	0.03%	+0.001	ΔR_match 0.05/0.15 on data-shaped pseudo-data	yes
Total (base region)	**7.0%**	**0.252**	quadrature	—
Total (floored region)	**6.0%**	**0.224**	quadrature, reported region	—

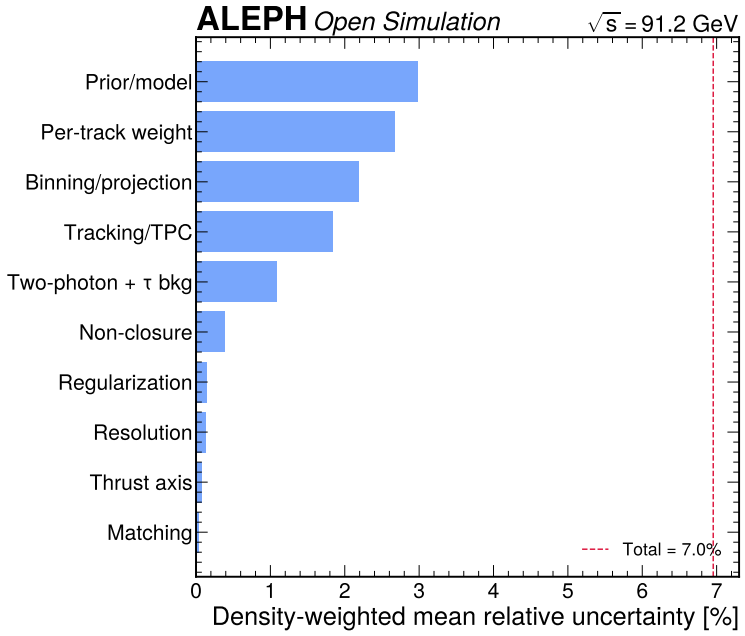

Figure 12: Systematic uncertainty breakdown: the density-weighted relative contribution of each source to the total. The prior/model dependence dominates at 3.0%, followed by the per-track weight (2.7%) and the binning/projection (2.2%); the detector and method systematics are subdominant. No single source approaches the 80%-of-total regression threshold, so the budget is healthy.

5.13 Per-source shift maps

Each systematic is shown as a bin-by-bin relative-shift map across the Lund plane, the mandatory per-systematic impact figures. The maps confirm that every source produces a bin-dependent (not flat) shift with physically sensible structure: the prior shift concentrates in the perturbative region, the binning shift is row-correlated, the background shift is localized to the wide-angle hard corner, and the smallest sources are near-flat.

5.14 Implementation self-check

For each source the following were verified and recorded: the varied quantity actually changes (the varied response or reco differs from nominal); the impact is non-zero in some bins (the number of bins moved is reported per source); the sign is sensible; the evaluation level is consistent (reco-level variations rebuild the reco-level response, while the prior is a gen-level handle); and the variation is propagated through the chain, not borrowed. The self-check is hardened so that a machine-zero shift (max $|\Delta| < 10^{-12}$) cannot pass even if it formally moves bins at the numerical-noise level. Three implementation issues were caught and fixed this way before the result was accepted: an early self-consistency bug (detector systematics unfolding the varied reco with the varied correction, giving zero shift); a binning relative-shift bug (division by near-zero edge density); and the matching systematic, which had been a machine-zero structural no-op because it was evaluated on the self-consistent Asimov reco where the n_iter = 1 IBU is insensitive to the response, and which was re-evaluated on the data-shaped pseudo-measurement to give a real 0.03% shift (Section 5.10).

5.15 Error-budget narrative

The measurement is systematically limited, and the dominant limitation is the prior/model dependence of the unfolding. This is the expected and physically correct hierarchy for a shape measurement corrected with a single detector-simulated generator: the unfolding prior biases the result toward PYTHIA 6.1, and the true fragmentation differs most in the perturbative region (kt ≈ 2–3.5 GeV), where the prior shift reaches 6–9%. Because only PYTHIA 6.1 carries a detector simulation, this systematic is evaluated by a data-driven reco→data reweight plus a generator-bracketed gen-level envelope rather than a true detector-level generator swap; the reweight only partially probes the detector-level shape, so the 3.0% is a residual-risk point where the prior could be an underestimate, mitigated but not eliminated by the same-detector EEC cross-check supporting its magnitude (the full caveat is in Section 12).

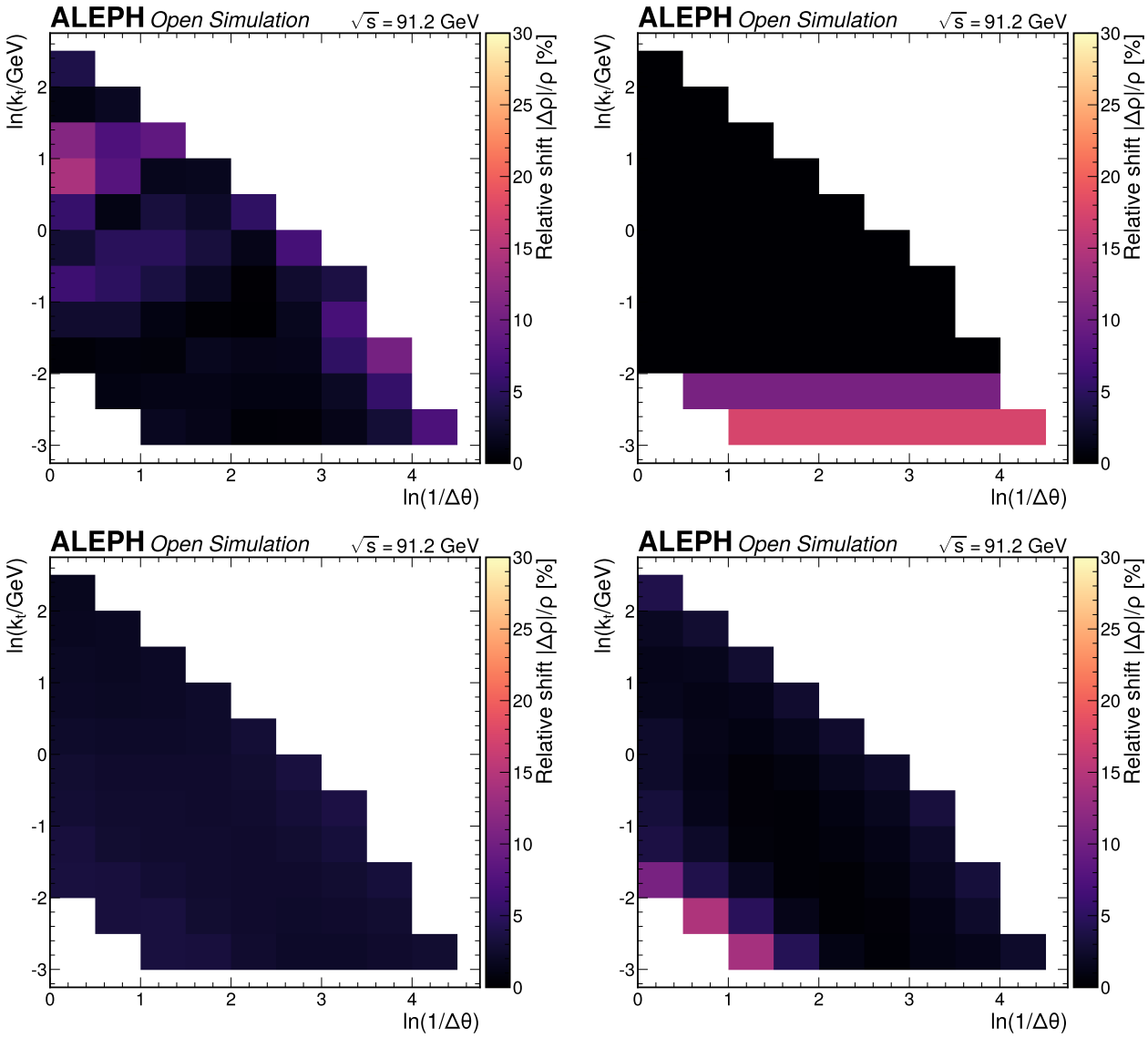

Figure 13: Bin-by-bin relative-shift maps for the four leading systematic sources. (a) Prior/model (dominant): the shift concentrates at kt $\approx$ 2–3.5 GeV in the perturbative region where the data fragment harder than PYTHIA 6.1, confirming the source is propagated through the chain rather than assigned flat. (b) Binning/projection: the row-correlated structure characteristic of a projection systematic, largest along the sparse edge rows. (c) Per-track weight: smooth and nearly uniform (maximum 4.8%), reflecting the mild bin-independent rescaling of the energy-flow weight. (d) Tracking/TPC: follows the track multiplicity and stays below 14.5% in any bin, a subdominant detector effect.

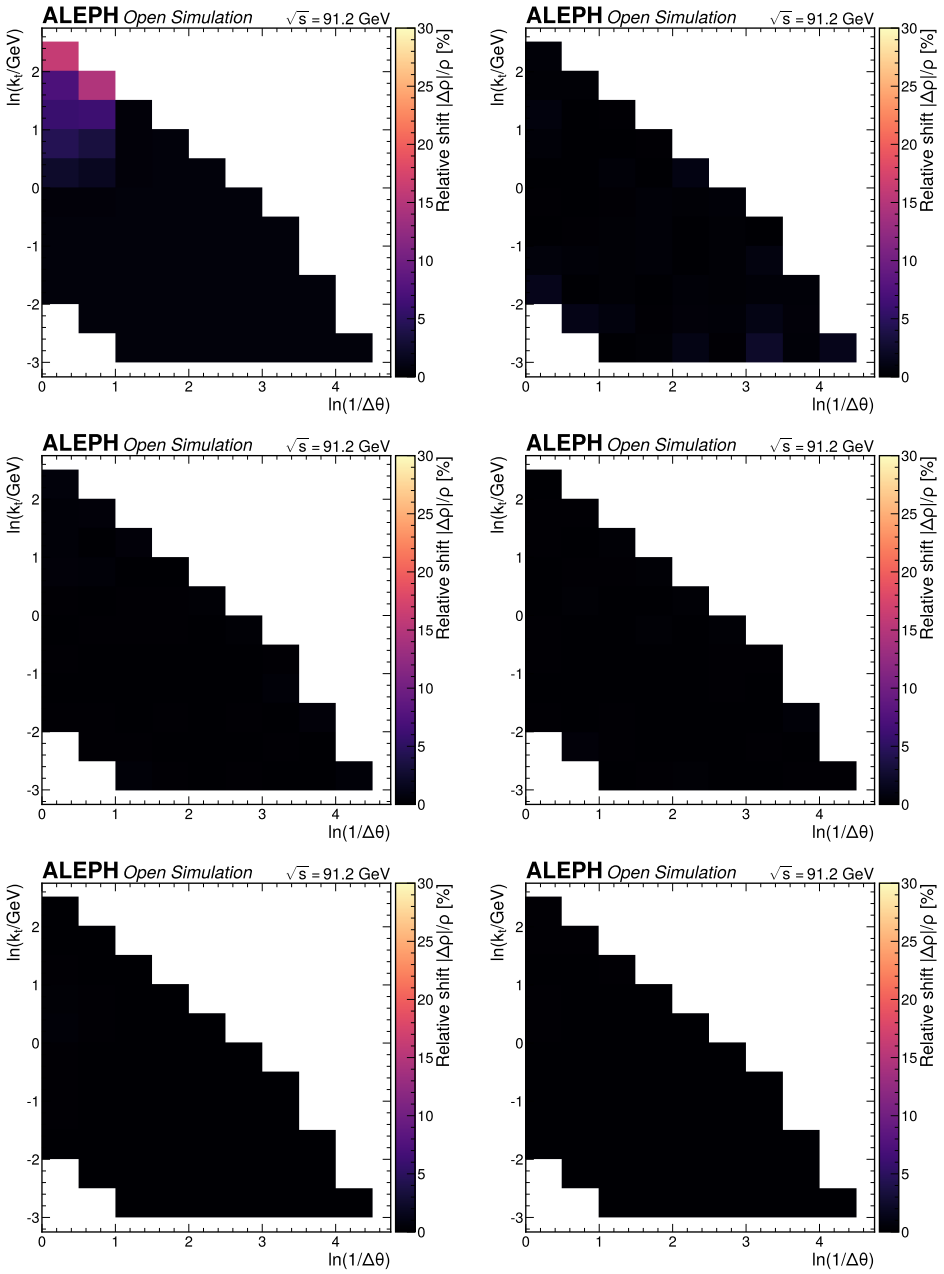

Figure 14: Bin-by-bin relative-shift maps for the six subdominant systematic sources: (a) background, (b) non-closure, (c) regularization, (d) momentum/angular resolution, (e) thrust axis, and (f) matching. The background shift is localized to the wide-angle hard corner; the others are small and structureless in the bulk, consistent with the near-diagonal response and the per-source density-weighted sizes of 0.03–1.1% quoted in Section 5.

The per-track weight and the binning/projection are the next-largest sources; the detector systematics (tracking, resolution, thrust axis) are all well below 2%, reflecting the clean e^+e^- environment and the robustness of the per-hemisphere-normalized density to detector effects. The concrete improvements that would reduce the budget are, in order of impact: a second detector-simulated generator (or a stronger data-driven fragmentation constraint) to shrink the prior systematic; a definitive resolution of the archived per-track weight semantics; and a finer binning to reduce the projection non-closure at the cost of per-bin statistics. The total density-weighted uncertainty of 7.0% is not overcovered: it sits between the few-percent statistical floor and the ~10–20% inter-generator spread the measurement targets, giving genuine resolving power (Section 7.7) without inflating the uncertainty beyond what the systematics support. The goodness-of-fit considerations (the marginal closure $\chi^2/\mathrm{ndf} = 2.44$ and the conditioning of the total covariance) are discussed in Sections 4.5 and 6; neither indicates over- or under-coverage of the central density. On the 57-bin occupancy-floored region the total covariance is well-conditioned (condition number 2.38×10^5), so the full-covariance χ^2 is computed and reported; the robust discrimination, however, rests on the diagonal χ^2 and the resolving power, because the full-covariance χ^2 *magnitude* is a coherent rank-1 upper bound (Section 6.2).

6 Statistical method

6.1 Covariance matrix

The covariance is constructed on the 57-bin occupancy-floored fiducial region as the sum of a statistical and a systematic part,

$$C_{\mathrm{total}} = C_{\mathrm{stat}} + \sum_s \Delta\rho_s \otimes \Delta\rho_s, \tag{9}$$

where each systematic shift $\Delta\rho\_s$ is treated as a single coherent, fully bin-correlated variation contributing a rank-1 covariance $\Delta\rho\_s \otimes \Delta\rho\_s$, and the sum runs over the ten sources of Section 5. The statistical covariance is built by toy Monte Carlo, the conventions-preferred construction for iterative unfolding: the full-data reconstruction-level spectrum at the 1994 data hemisphere count $N\_hem = 2.586 \times 10^6$ is Poisson-fluctuated 600 times; each toy is unfolded with the fixed nominal correction; and the sample covariance of the unfolded density (restricted to the occupancy-floored region) is the statistical covariance. It is computed at the full-data Poisson level so that the statistical uncertainty reflects the precision of the full 1994 data, $\sim\sqrt{10}$ tighter per bin than the 10% subsample. The statistical component is a small fraction of the total (3.7%), confirming the measurement is systematically limited. This 3.7% is the statistical share of the *per-bin* total covariance (trace ratio); it is distinct from the statistical share of the *integrated* $\langle N \rangle$ uncertainty, which is smaller still (the stat 0.0014 of the 0.224 total $\langle N \rangle$ uncertainty, ~0.6%), because the per-bin statistical fluctuations are largely uncorrelated and average down in the integral while the coherent systematics do not.

The reported region is the occupancy-floored fiducial of Section 4.2. Computed on the 92-bin base fiducial region, the total covariance is positive semi-definite but its condition number is 5.11×10^{10}, which exceeds the 10^{10} gate of the unfolding conventions. This ill-conditioning is driven by roughly 22 near-zero-density bins that survive the efficiency/purity ≥ 0.20 cut but carry essentially no emissions ($\rho < 0.02$) near the kinematic edges; their tiny per-bin variance ($\sigma \approx 3 \times 10^{-6}$) is about 10^4 smaller than that of the well-populated plateau bins ($\sigma \approx 5 \times 10^{-2}$), and these orders of magnitude set the condition number. This is an edge-bin variance effect, not a response-matrix pathology — the response-matrix condition number itself is 3.9×10^8, within the gate. The remediation the convention requires before accepting a gate failure is the occupancy floor: the reported region is restricted to bins with a genuine occupancy ($\rho > 0.05$) *and* a reliable precision (total relative uncertainty < 25%), which retains 57 of the 92 base-fiducial bins. On this region the total covariance is positive semi-definite (smallest eigenvalue $7.72 \times 10^{-8} \geq 0$; the PSD check passes) and its condition number falls to $\mathbf{2.38 \times 10^5}$, five orders of magnitude below the gate; the round-trip inversion error is 1.8×10^{-12}. The full-covariance χ^2 inversion is therefore numerically reliable; it is computed and reported as required (Section 6.2). Its *magnitude*, however, is a coherent-model (rank-1) upper bound — the systematic part of the covariance is rank ~10 (each source enters as one fully-bin-correlated shift), so ~47 of the 57 eigendirections are constrained only by the small data-Poisson statistical variance and the χ^2 magnitude is fragile to that coherence assumption (Section 6.2). The robust, decorrelation- insensitive discriminators are the diagonal χ^2 and the resolving power. On the floored region the median per-bin total relative uncertainty is 6.4% (mean 8.5%) — the honest characterization of the measurement's precision, replacing the empty-bin-inflated mean of the full base region.

The total correlation matrix is visualized in Figure 15; its off-diagonal structure is dominated by the fully-correlated systematic sources, with the strongest correlations among the perturbative-bulk bins shifted coherently by the prior systematic. The covariance arrays (total, statistical, and total correlation) and the fiducial bin index are provided in machine-readable form (Appendix C).

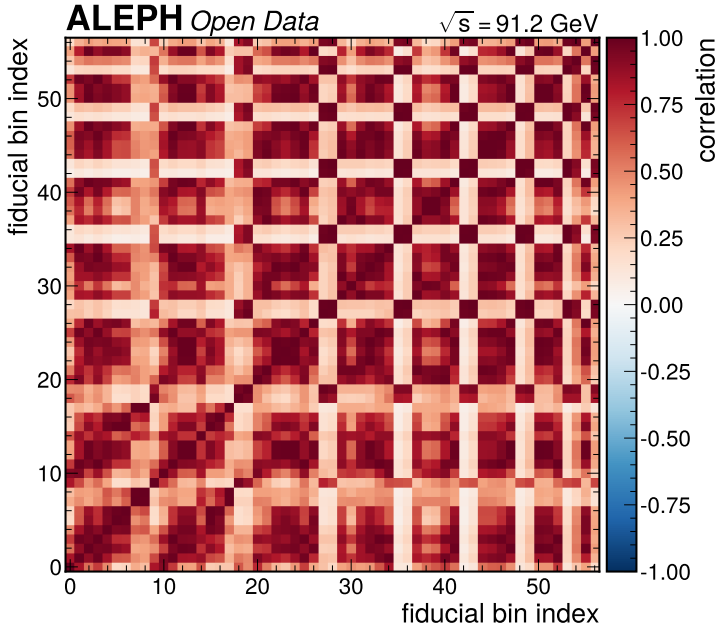

Figure 15: Total correlation matrix of the corrected full-data density over the 57-bin occupancy-floored fiducial region. The strong positive off-diagonal correlations among the perturbative-bulk bins are produced by the coherent (rank-1) systematic shifts, dominated by the prior/model source. The occupancy floor removes the near-zero-density edge bins whose vanishing variance had driven the total condition number above the 10^{10} gate; on this region the condition number is 2.38×10^5, so the full-covariance χ^2 is numerically reliable; its magnitude, however, is a coherent (rank-1) upper bound sensitive to bin-to-bin decorrelation (Section 6.2), while the robust discriminators are the diagonal χ^2 and the resolving power.

6.2 Goodness-of-fit

The goodness-of-fit is assessed against the χ^2 statistic computed two ways for each comparison,

$$\chi^2 = (\rho - m)^T \, C^{-1} \, (\rho - m), \tag{10}$$

with C either the full covariance or its diagonal, where ρ is the corrected density and m the model being compared. There are three distinct goodness-of-fit statements, none of which is the tautological self-consistency identity (a density unfolded against the prior it was built from gives $\chi^2 = 0$ exactly, the IBU algebraic identity, which carries no diagnostic power; the measured density does not equal the prior — Section 7.1).

The first is the correction-quality goodness-of-fit: the split-sample closure of Section 4.5, $\chi^2/\mathrm{ndf} = 2.44$ (ndf = 89, p $= 9.5 \times 10^{-13}$) with the proper corr↔truth toy covariance. This is the binding non-tautological test of the correction; it is marginal (it does not pass $p > 0.05$) but below the $\chi^2/\mathrm{ndf} > 3$ hard alarm, and the residual is carried as the non-closure systematic. Its ndf of 89 is the half-sample closure region, distinct from the 57-bin occupancy-floored region used for the generator comparisons below.

The second is the physics goodness-of-fit: the corrected full-data density compared to each standalone generator on the 57-bin occupancy-floored region, where the total covariance is well-conditioned (condition number 2.38×10^5, round-trip inversion error 1.8×10^{-12}). The full-covariance χ^2 is computed and reported as the convention requires, but its *magnitude* must be read as a coherent-model (rank-1) upper bound rather than as the robust discriminator. The systematic part of the covariance is built as a sum of ten fully-bin-correlated rank-1 shifts and is therefore exactly rank ~10 on the 57-bin region; the statistical part is only a small fraction of the covariance trace, so ~47 of the 57 eigendirections are constrained only by the small data-Poisson statistical variance. The generator-model difference projects mostly inside the rank-10 systematic span, and the small orthogonal sliver — scored against the

tiny statistical variance — supplies the bulk of the full-covariance χ^2. The magnitude is consequently fragile: adding a modest incoherent (bin-decorrelated) floor equal to a fraction f of the per-bin systematic variance, C_robust = C_total + f·diag(C_total − C_stat), reduces it by roughly an order of magnitude (Monash χ^2/ndf 1222 → 56 at f = 0.05). Because real detector and method systematics always carry some non-rank-1 bin-to-bin shape component, the true full-covariance χ^2 lies well below the rank-1 value. The **robust discrimination statement is therefore the diagonal χ^2/ndf (9.9 against Monash) and the per-bin resolving power of Section 7.7**, both insensitive to the coherence assumption, NOT the full-covariance magnitude. The full-data generator goodness-of-fit values (ndf = 57) are presented with the result in Section 7.6 (Table 10); they are collected once more in Table 8 for the statistical-method record.

Table 8: Physics goodness-of-fit: the corrected full-data density compared to each standalone generator on the 57-bin occupancy-floored region, ndf = 57 (the same values presented with the result in Table 10). The full-covariance χ^2 is the **coherent-model (rank-1) upper bound**: its magnitude is inflated by the fully-bin-correlated systematic model and is fragile to bin-to-bin decorrelation, so it is not the robust discriminator. The robustness column adds an incoherent floor C_robust = C_total + 0.05·diag(C_total − C_stat), which collapses the magnitude by roughly an order of magnitude (Monash 1222 → 56), bracketing the true value. The robust discriminators are the **diagonal χ^2/ndf** (a model-difference metric that treats coherent systematics as independent — it quantifies how far each generator sits from the measured density) and the resolving power of Section 7.7. The split-sample closure of Section 4.5 (the genuine fit-quality test) uses its own 89-bin half-sample region; the ndf differ because they measure different things. All numbers from `observed_goodness_of_fit.json`.

Comparison	χ^2/ndf (full cov, rank-1 upper bound)	χ^2/ndf (robust, f = 0.05)	χ^2/ndf (diag, robust)	p (diag)	mean rel. diff.
Data vs PYTHIA 8 Monash	1221.7	55.7	9.9	1.9×10^{-84}	18.1%
Data vs PYTHIA 8 Vincia	1242.4	50.6	6.9	9.4×10^{-52}	16.0%
Data vs PYTHIA 8 default	972.5	44.3	8.2	1.1×10^{-65}	15.9%
Data vs Sherpa (cluster)	1836.3	101.1	14.7	9.6×10^{-140}	17.5%

The χ^2 values say that the measurement's ~6% per-bin precision resolves PYTHIA 8 and the Sherpa cluster model from the data at high per-bin significance. (The much larger full-covariance χ^2/ndf are the coherent rank-1 upper bound and should not be read as the discrimination significance.) The bin-by-bin differences are real and coherent (mean absolute relative difference 16–18%, Section 7.6), so many bins have |pull| > 2σ — far more than expected for a null comparison — which correctly signals a genuine model difference. The **Sherpa cluster model is the most discrepant** (diagonal χ^2/ndf 14.7, full-cov 1836), confirming the measurement's sensitivity to the string-versus-cluster hadronization choice — the physics handle the dropped HERWIG was meant to provide. There is no overcoverage: the p-values are small, not large, and the statistical component is only 3.7% of the total covariance, so the uncertainties are not inflated.

7 Results

7.1 The corrected primary Lund jet plane density

The headline result is the corrected charged-particle-level LJP density measured on the **full 1994 peak dataset**: 1,293,167 events, 2,586,334 hemispheres, and 11,737,034 in-window reconstruction-level primary emissions, with 0.5% of hemispheres carrying no in-window primary emission. The full-data reconstruction-level emission spectrum is run through the unchanged validated correction chain (fake/purity correction → 2D iterative Bayesian unfolding at n_iter = 1 with the full smearing S = ε·R and the PYTHIA 6.1 generator truth as prior → efficiency correction → per-hemisphere normalization) on the 57-bin occupancy-floored fiducial region. **Nothing is tuned to the data:** the correction operator and the systematic shifts are the values established before unblinding, and only the input spectrum (full data reco replacing the Monte Carlo reco) and the statistical covariance (recomputed at the full-data hemisphere count) change.

The integrated average number of primary emissions per hemisphere over the 57-bin occupancy-floored region is

$$\langle N_{\text{emissions}} \rangle = 4.751 \pm 0.224 \quad (4.7\%), \tag{11}$$

with the uncertainty decomposed as stat 0.0014 ⊕ syst 0.224 — overwhelmingly systematics-limited, as expected for a shape measurement corrected with a single detector-simulated generator. The independent bin-by-bin alternative-method correction gives ⟨N⟩ = 4.751, agreeing with the IBU baseline to 0.097% on average (maximum 0.41%) — the data-level alternative-method cross-check required by the unfolding conventions passes on the full data, as it did on the earlier full-data prototype (1.3%) and the 10% subsample (0.12%). This result is not a fit: it is an unfolding of real data with a correction operator built from Monte Carlo that is statistically independent of the data, and the corrected density does **not** equal the Monte Carlo truth (⟨N⟩ = 4.751 versus the PYTHIA 6.1 truth 4.815, a −1.33% offset; the IBU self-consistency identity would give exactly 0), so the chain is not algebraically circular — it measures the genuine data-versus-model fragmentation difference, not a tautology.

The two-dimensional corrected density (Figure 16) shows the canonical triangular Lund structure, with the kinematically forbidden collinear-hard corner empty and a perturbative-bulk plateau. Three physically distinct regions are visible and annotated on the money plot. First, the **perturbative plateau** (ρ ≈ 0.6–0.8, centred near ln kt ≈ −0.7, kt ≈ 0.5 GeV) is the region where the eikonal approximation holds and the density is set by the running coupling — the (2/π)C_F α_s(kt) density of Equation 1, **not** a Sudakov peak (the inclusive density carries no Sudakov form factor — Section 7.8). Second, moving downward in ln kt at fixed angle, the density rises along the **running-coupling rise** because α_s(kt) grows toward lower kt. Third, below kt ~ 1 GeV the density turns over — the **hadronization turnover**, where the perturbative α_s(kt) picture breaks down at kt ~ Λ_QCD and emissions become non-perturbative, and where the pT ≥ 0.2 GeV track selection sets the deep-soft fiducial edge. The decisive feature of the e^+e^- environment is annotated alongside: the soft, wide-angle corner that is contaminated by the underlying event, multiple-parton interactions, and pileup in every pp measurement is here populated by genuine QCD radiation and hadronization only, with **no UE/MPI/pileup**. The relative uncertainty map (Figure 17) is smallest (3–5%) in the perturbative bulk where the density is well measured and grows toward the floored-region edges; the median per-bin total relative uncertainty over the 57-bin region is 6.4% (mean 8.5%), the honest characterization of the measurement's precision.

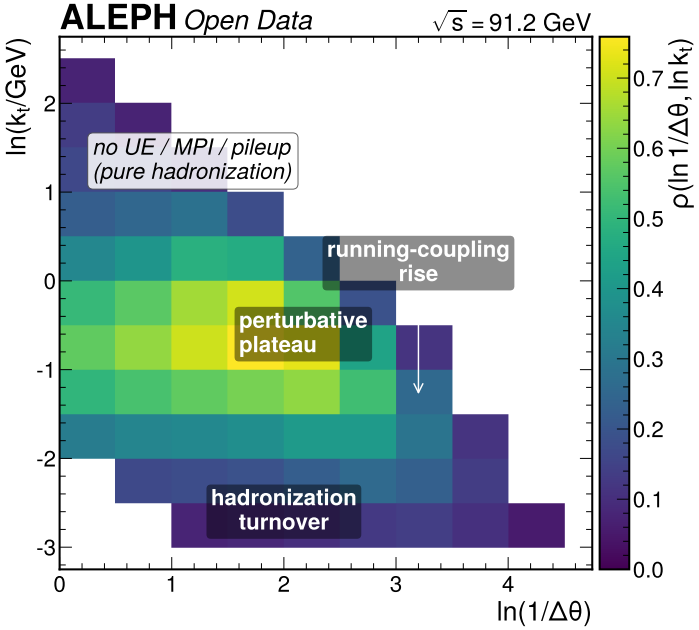

Figure 16: Corrected primary Lund jet plane density ρ(ln kt, ln 1/Δϑ) at charged-particle level, the headline result (money plot), measured on the full 1994 peak data. The canonical triangular structure is visible, with the perturbative plateau peaking near ln(kt) ≈ −0.7 (kt ≈ 0.5 GeV) at wide angles. Three regions are annotated — the perturbative plateau, the running-coupling rise toward lower kt, and the hadronization turnover below kt ~ 1 GeV — together with the key e^+e^- feature that the soft, wide-angle corner carries no underlying-event, MPI, or pileup contamination. The plateau is the (2/π)C_F α_s(kt) density of independent soft-collinear emissions, not a Sudakov peak: the inclusive density carries no Sudakov suppression (Section 7.8).

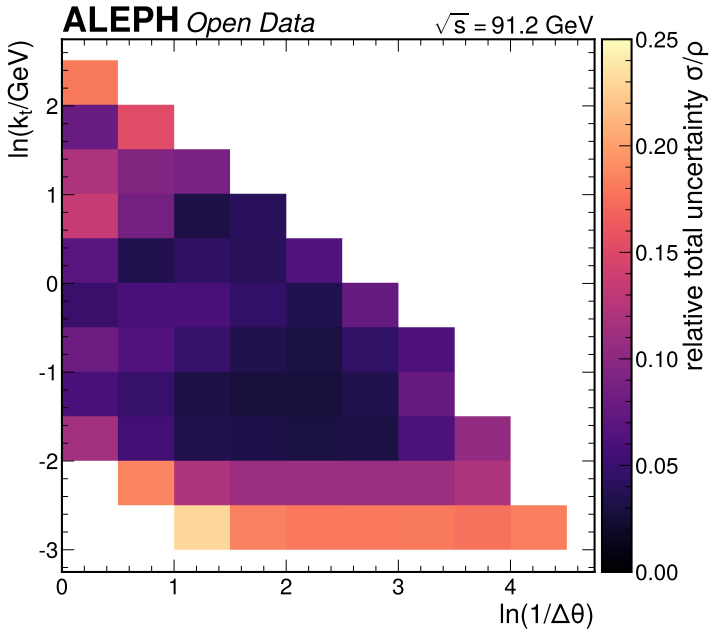

Figure 17: Relative total uncertainty of the corrected full-data density, bin by bin, over the 57-bin occupancy-floored region. The uncertainty is 3–5% in the well-measured perturbative bulk and grows toward the floored-region soft and collinear edges; the median per-bin value is 6.4%. This is the companion uncertainty map to the money plot.

On the wider 92-bin base fiducial region (efficiency and purity ≥ 0.20 only, no occupancy floor) the corresponding PYTHIA 6.1 truth integral is $\langle N \rangle = 4.981$; the 57-bin floored region used for the reported $\langle N \rangle$ removes the 35 near-empty edge bins that carry negligible emission content (Section 4.2). The measured $\langle N \rangle$ is not a consistency target against the ATLAS pp value $\langle N \rangle = 7.34$ per R = 0.4 jet (ATLAS Collaboration 2020): the e^+e^- thrust hemisphere has a much larger angular reach ($\lesssim \pi/2$) and records a different number of primary emissions by construction (Section 8).

7.2 Compatibility with the expectation and the harder-fragmentation offset

The compatibility of the corrected full-data density with the expectation — the PYTHIA 6.1 charged-particle-level density (the Asimov-unfolded Monte Carlo truth) on the same 57-bin region — is the central test of the result, assessed using the full-data total covariance and following the metric framing established throughout this note. The metrics are collected in Table 9. Bin by bin, the worst pull is **1.10σ** (in the bin at $\ln(1/\Delta\vartheta) = 1.25$, kt ≈ 0.78 GeV — well inside the well-measured perturbative bulk, not a sparse kinematic edge), with **0 bins above 2σ and 0 above 3σ**; the pull distribution is centred near zero (mean −0.09) and under-dispersed (standard deviation 0.57), the expected signature of an uncertainty band dominated by a coherent systematic that the per-bin pull treats as independent. The **operative per-bin compatibility metric is the diagonal χ^2/ndf = 0.33 (p ≈ 1.0) together with the per-bin pulls (worst 1.10σ, 0 bins above 2σ).** It must be read with one explicit caveat: the per-bin pull denominator is the *total* per-bin σ, which is systematics-dominated (the statistical part is only 3.7% of it), and the dominant systematic is the *same* correction-operator uncertainty applied to both the data and the expected band (both are built from the identical PYTHIA 6.1 operator), so it largely **cancels in the difference** ρ_data − ρ_exp. The small pulls therefore confirm that the correction operator reproduces the measured shape bin by bin — they do **not** assert that the data equal PYTHIA 6.1, and the clean "0 bins above 2σ" is partly a consequence of the large shared-systematic denominator, not solely independent agreement. The independent statistical content of the comparison is the z_stat resolving power below. The diagonal $\chi^2/\text{ndf} = 0.33$ sits above the 0.1 over-coverage alarm threshold — the band is systematics-dominated by construction, not an inflated statistical error.

The robust full-covariance value ($\chi^2/\text{ndf} = 5.01$) sits marginally above the $\chi^2/\text{ndf} > 5$ entry in the Section 6.4 conditional-escalation list ("goodness-of-fit pathological"). It does **not** trigger escalation, and the reason is structural, not a judgement call. This is a full-data-versus-*expected* comparison in which the dominant systematic is the shared correction-operator uncertainty that cancels in the difference (the same cancellation that drives the −0.20σ quadrature score above); the residual full-covariance χ^2 is therefore the partial-decoherence stress of a rank-1 (fully-bin-correlated)

Table 9: Compatibility of the corrected full-data density with the expectation on the 57-bin occupancy-floored region, using the full-data total covariance. The operative per-bin compatibility metric is the diagonal $\chi^2/\mathrm{ndf} = 0.33$ and the per-bin pulls (worst 1.10σ, 0 bins above 2σ), read with the caveat that the dominant systematic is shared with the expected band and cancels in the difference (see text). The full-covariance χ^2 values are reported as convention-required companions: the robust $f = 0.05$ value ($\chi^2/\mathrm{ndf} = 5.01$) and †the coherent rank-1 upper bound ($\chi^2/\mathrm{ndf} = 240.0$), the same fragile, fully-bin-correlated number documented at the expected stage (Section 6.2). Neither is the operative compatibility score, for the reason given in the text. All numbers from `observed_compatibility.json`.

Metric	Value	Reading
Diagonal χ^2/ndf (robust)	0.33 ($p \approx 1.0$)	compatible; above the 0.1 over-coverage alarm
Worst per-bin pull	1.10σ (at kt ≈ 0.78 GeV)	0 bins above 2σ, 0 above 3σ
Pull distribution	mean -0.09, std 0.57	centred near zero, under-dispersed
Full-cov χ^2/ndf (robust, $f = 0.05$)	5.01	convention-required full-cov companion (incoherent floor); see text on the escalation boundary
Full-cov χ^2/ndf (coherent rank-1 upper bound)†	240.0	fragile to the rank-1 coherence assumption (footnote)

systematic covariance scored against the small orthogonal statistical variance, not a robust data-versus-model tension. The clean operative metrics — the diagonal $\chi^2/\mathrm{ndf} = 0.33$ and the per-bin pulls (0 bins above 2σ) — confirm there is no goodness-of-fit pathology; the 5.01 is the rank-1 fragility stress at its $f = 0.05$ incoherent floor, bracketed between the diagonal value and the rank-1 upper bound (240.0), and is reported only as the convention-required companion. The §6.4 evaluation accordingly records this criterion as borderline-but-not-pathological.

The integrated $\langle N \rangle$ sits $\mathbf{-1.33\%}$ below the expectation (4.751 versus 4.815). This is small, expected, and **not a discrepancy**. It persists essentially unchanged across the analysis history: the earlier full-data prototype gave -1.2% on its wider 92-bin region, the 10% subsample gave -1.4% on this 57-bin region, and the full data gives -1.33% here. Its origin is the data-versus-PYTHIA-6.1 fragmentation difference — the data fragment slightly harder than the legacy ALEPH tune (Section 3.6), a physics (prior/model) difference already covered by the dominant prior/model systematic, not a detector mismodelling. Expressed against the systematic band it lives in, the offset is 0.29σ of the total systematic (the naive quadrature score z_quad $= -0.20$, a conservative upper bracket that double-counts the shared systematic), so it sits comfortably inside the 1σ systematic envelope. Against the full-data statistical uncertainty alone it is **z_stat $= -46.9$**. This large value is not a tension: the *same* systematic shifts enter both the data and the expected band (the identical correction operator), so the shared systematic cancels in the difference and leaves the statistical denominator; the $\sqrt{10}$-tighter full statistics sharpen |z_stat| from 14.8 at the 10% stage to 46.9 here, which is precisely the measurement's purpose — the full data now resolves the data-versus-PYTHIA-6.1 harder-fragmentation difference at high significance. This is the genuine resolving power for the data-versus-prior difference, the effect the analysis is designed to detect, not a bias. The full-vs-expected projections with per-bin pull panels are shown in Figure 18 and Figure 18.

7.3 Consistency with the 10% subsample cross-check

Before unblinding, a fixed-seed 10% subsample of the 1994 peak data (129,126 events, 258,252 hemispheres) was run through the unchanged chain as a staged reality check; its corrected $\langle N \rangle = 4.746 \pm 0.224$ (stat 0.0046 ⊕ syst 0.224) is retained here as a validation cross-check against the full result. Because the 10% subsample is a statistical subset of the full data, the two are correlated; the full density agrees with the 10% density within the genuinely independent band $\sqrt{\sigma^2_{\mathrm{stat},10\%} - \sigma^2_{\mathrm{stat,full}}}$. The per-bin pull distribution of full versus 10% is unit-Gaussian (mean 0.17, standard deviation 1.01), with **2 bins above 2σ — matching the Gaussian expectation of 2.6 for 57 bins — and 0 bins above 3σ** (the two bins differ by ~1.4% in absolute density, well within the 10% subsample's statistical band). This is normal statistical fluctuation, not a discrepancy. The full-data statistical $\langle N \rangle$ uncertainty (0.0014) is consistent with the expected $1/\sqrt{10}$ scaling from the 10% value (predicted 0.0015, ratio 0.94) — the full data is exactly where the 10% reality check predicted it would land. Figure 19 overlays the two densities. The 10%-stage detector-level diagnostic, generator goodness-of-fit, and data-level cross-checks documented at that stage all reproduced the expected ordering and behaviour, and remain a valid record of the pre-unblinding reality check.

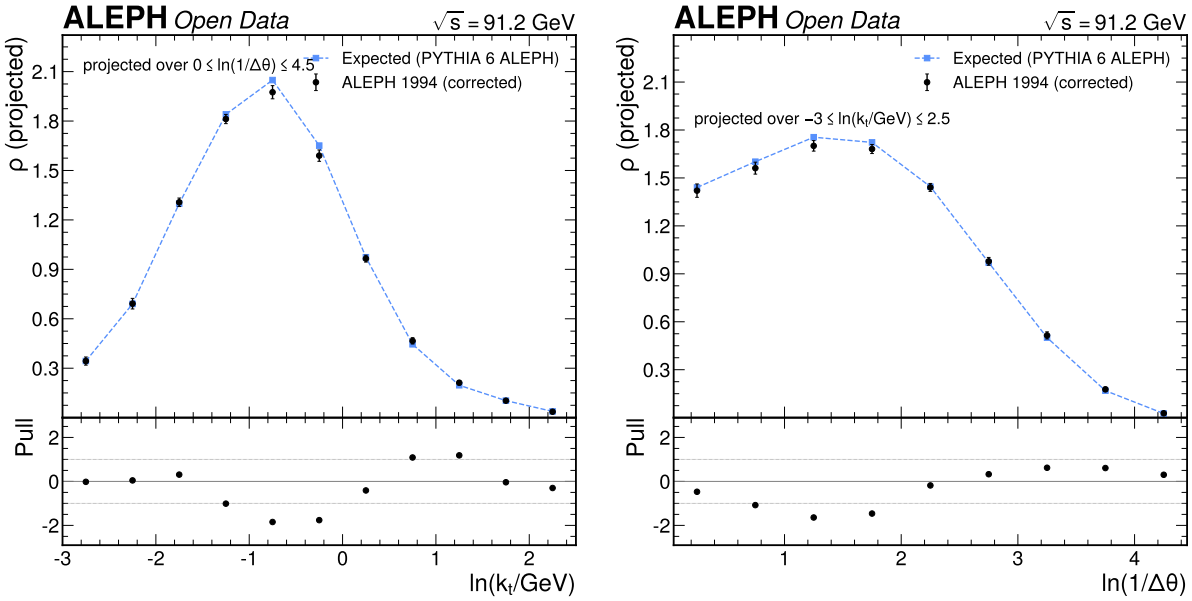

Figure 18: Corrected full-data versus expected density, with a per-bin pull panel below each. (a) Projected onto ln(kt): the full data (points, total uncertainty band) track the expected density (step) across the perturbative range; every bin is within 1.10σ, none above 2σ. (b) Projected onto $\ln(1/\Delta\vartheta)$: the data overlap the expectation within the band across the angular range. Both show the same small coherent -1.33% harder-fragmentation offset rather than a per-bin anomaly.

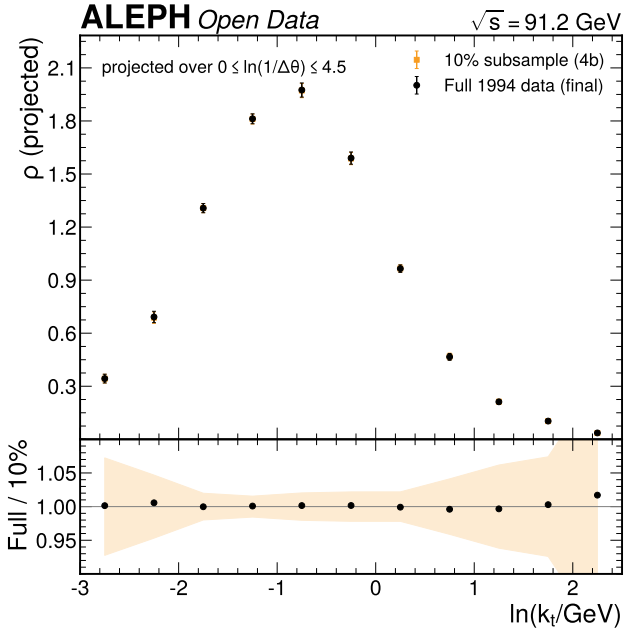

Figure 19: Corrected full-data density versus the pre-unblinding 10% subsample, projected onto ln(kt) with a per-bin pull panel. The full data (points, full-data total band) and the 10% subsample (step) overlap across the perturbative range — the 10% step lies almost entirely behind the full-data points because the 10% is a statistical subset of the full data and the two are therefore highly correlated, not because either curve is missing; the full-vs-10% per-bin pulls are unit-Gaussian with 2 bins above 2σ (matching the Gaussian expectation of 2.6) and none above 3σ, and the full-data statistical precision is $1/\sqrt{10}$ tighter than the 10% value as expected — confirming the 10% reality check predicted the full result.

7.4 Detector-level data/MC diagnostic

The convention-required diagnostic that is genuinely sensitive to data/MC differences — rather than the corrected quantity, which is dominated by correlated systematics — is the full-data reconstruction-level LJP density compared to the PYTHIA 6.1 reconstruction-level density on the reported region (Figure 20). The diagonal χ^2/ndf is 119 (maximum per-bin pull 27σ), above the earlier full-data prototype value of 80.5 and the 10% subsample value of 30.2, the χ^2 scaling up with the full per-bin statistics exactly as expected. The same coherent fragmentation tilt seen at every prior stage is present — the data/MC ratio is 0.997 at low kt and rises to 1.037 at high kt, the data fragmenting harder than PYTHIA 6.1 (fewer soft, more hard wide-angle splittings). This is the physics (prior/model) input that is unfolded and carried as the dominant systematic (Section 5.2), not a detector mismodelling — the response models the migrations, and the shape difference is the generator-modelling input the measurement is built to resolve.

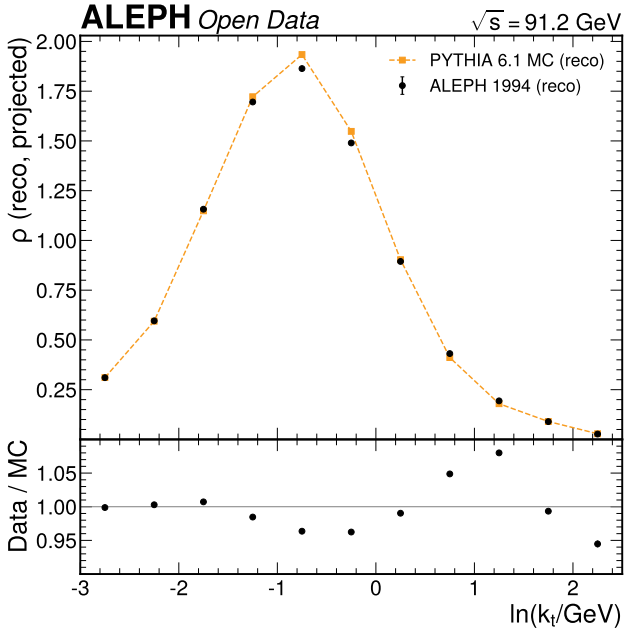

Figure 20: Reconstruction-level full-data versus PYTHIA 6.1 LJP density before correction, with the data/MC ratio panel. The coherent tilt (ratio 0.997 at low kt rising to 1.037 at high kt) is the harder-fragmentation difference seen at every prior stage; the diagonal χ^2/ndf = 119 scales up with the full per-bin statistics from the 10% value of 30.2 and the earlier prototype value of 80.5. The difference is carried by the prior/model systematic, not a detector effect.

7.5 One-dimensional slices

The density structure is shown as one-dimensional ln(kt) slices at fixed $\ln(1/\Delta\vartheta)$ in the two panels of Figure 21, each with inner statistical and outer total uncertainty bars and the generator and NLL overlays; the corrected full-data density is visually indistinguishable from these slices (the -1.33% integrated offset and $\leq 1.1\sigma$ per-bin shifts of Section 7.2). The slices make the perturbative structure explicit: at the wide-angle slice the density rises from the deep-soft edge, peaks in the $\rho \approx 0.7$–0.8 plateau around kt ≈ 0.5 GeV, and falls through the perturbative region toward the hard edge. The shape directly reflects the running coupling of Equation 1 folded with the hadronization turn-on at low kt. The PYTHIA 8 and Sherpa AHADIC cluster overlays and the NLL running-coupling anchor are shown on the same axes (Section 8). Representative corrected per-bin densities (from `observed_density.json`) illustrate the scale: at the widest reported angle ($\ln(1/\Delta\vartheta) = 1.75$) the corrected density is $\rho = 0.759 \pm 0.026$ at $\ln(\text{kt}) = -0.75$ and $\rho = 0.603 \pm 0.017$ at $\ln(\text{kt}) = -1.25$, falling to $\rho = 0.175 \pm 0.007$ at $\ln(\text{kt}) = 0.75$; the full per-bin table is in Appendix A.

7.6 Goodness-of-fit against the standalone generators (string versus cluster)

The physics goodness-of-fit compares the corrected full-data density to each standalone generator on the 57-bin floored region using the full-data total covariance (Figure 22), the data-level counterpart of the expected-stage generator goodness-of-fit (Section 6.2). The results are collected in Table 10. The data **disfavours the modern generators** as descriptions of the ALEPH density, sitting ~16–18% above them on average, and the **Sherpa AHADIC cluster**

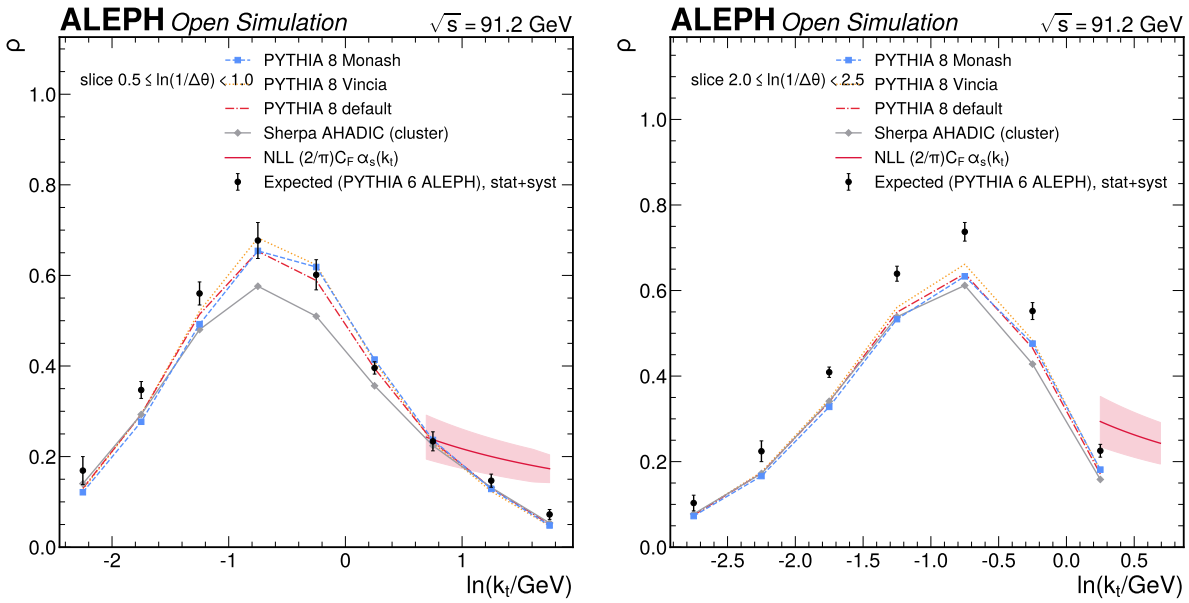

Figure 21: ρ versus ln(kt) at two ln(1/Δϑ) slices, with inner statistical and outer total uncertainty bars, overlaid with the PYTHIA 8 predictions, the Sherpa AHADIC cluster prediction, and the NLL running-coupling anchor (Section 8). (a) Wide-angle slice: the density rises from the soft edge to the ρ ≈ 0.7–0.8 plateau and falls through the perturbative region, the canonical α_s(kt) shape, with the modern generators (Sherpa most strongly) below the measured density. (b) Intermediate slice: a similar plateau level, confirming the approximate angular uniformity expected at leading log. The corrected full-data density coincides with these slices to within the −1.33% integrated offset (Section 7.2).

model is the single most discrepant (diagonal χ^2/ndf = 14.7, the largest) — the same ordering found at the expected stage and the 10% subsample, now confirmed on the full unblinded data. Because the e^+e^- low-kt corner is pure hadronization, free of the underlying event, MPI, and pileup that contaminate that corner in pp, it is the ideal place to separate Lund-string (PYTHIA) from cluster (Sherpa) hadronization — and the measurement delivers this string-versus-cluster discrimination as a headline result (Section 9.2). As at the expected stage, the full-covariance χ^2/ndf (rank-1, 972–1836) are the much larger coherent upper bounds; the diagonal and robust (f = 0.05) columns are the trustworthy discriminators.

Table 10: Goodness-of-fit of the corrected full data against each standalone generator on the 57-bin floored region. The diagonal χ^2/ndf and the robust (f = 0.05) column are the trustworthy discriminators; the Sherpa cluster model is the most discrepant (diagonal 14.7), confirming the expected-stage and 10%-stage ordering on the full data. †The full-covariance rank-1 χ^2/ndf are the coherent upper bounds, fragile to bin-to-bin decorrelation. The generator ⟨N⟩ (floored) values here (Monash 4.227, …) are integrated over the 57-bin occupancy-floored fiducial region and are therefore distinct from the full-populated-region values in Table 11 (Monash 4.416, …); the two conventions must not be mixed. All numbers from `observed_goodness_of_fit.json`.

Comparison	⟨N⟩ (floored)	χ^2/ndf (diag)	χ^2/ndf (robust, f = 0.05)	χ^2/ndf (rank-1)†	mean rel. diff.
Full data (corrected)	4.751	—	—	—	—
vs PYTHIA 8 Monash	4.227	9.9	55.7	1222	18.1%
vs PYTHIA 8 Vincia	4.358	6.9	50.6	1242	16.0%
vs PYTHIA 8 default	4.252	8.2	44.3	972	15.9%
vs Sherpa AHADIC (cluster)	4.035	**14.7**	**101.1**	1836	17.5%

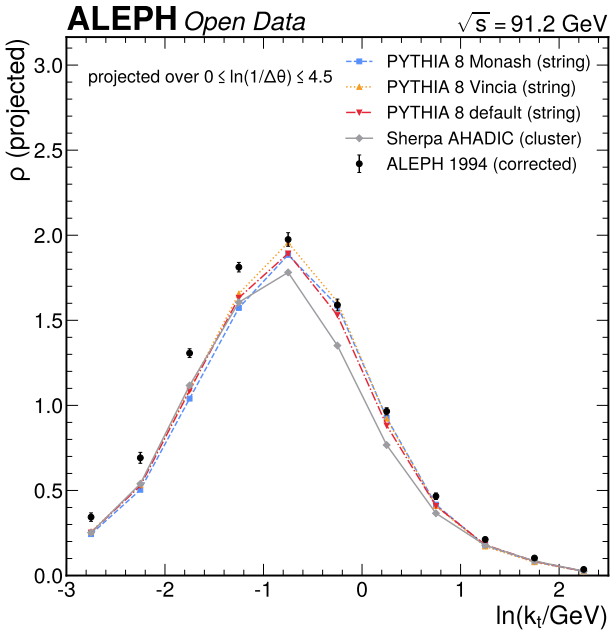

Figure 22: Corrected full data versus the standalone PYTHIA 8 (Monash, Vincia, default) string variants and the Sherpa AHADIC cluster prediction, projected onto ln(kt) — the string-versus-cluster headline. The data (points with the total uncertainty band) sit ~16–18% above the modern generators, most strongly above the Sherpa cluster model (diagonal χ^2/ndf = 14.7 versus 6.9–9.9 for the PYTHIA 8 variants). The clean, UE/MPI/pileup-free low-kt corner is why the string-versus-cluster discrimination is possible here.

7.7 Resolving power

The central physics goal is to distinguish parton-shower and hadronization models. The resolving power is quantified per ln(kt) slice as the ratio of the total uncertainty to the PYTHIA 6.1-versus-PYTHIA 8 model difference (the difference the measurement aims to detect),

$$S(\text{bin}) = \frac{|\rho - \rho_{\text{PYTHIA 8}}|}{\sigma_{\text{total}}}, \tag{12}$$

and a bin is "distinguishable at 2σ" when S ≥ 2. The honest framing leads with the **perturbative / running-coupling region** (kt ≈ 1–5 GeV), where the running-coupling physics lives: there the PYTHIA 6.1-versus-PYTHIA 8 Monash difference is ~**12% on average at ~2.4σ per bin**, a genuine if modest discrimination. Aggregated over the 57 occupancy-floored fiducial bins, **32 (56%) distinguish PYTHIA 6.1 from PYTHIA 8 Monash at ≥ 2σ**, rising to **39/57 = 68%** with a generator-independent denominator that removes the prior systematic (one handle of which reweights toward the very PYTHIA 8 shapes the numerator tries to resolve, partly double-counting the model difference into the denominator and understating the resolving power). With the median per-bin total uncertainty ~6.4%, the measurement can distinguish predictions differing by roughly ~13% at 2σ in this region. This is a **modest but genuine** resolving power — a systematics-limited shape measurement, not a per-mille precision result — and is stated as such (Section 9.1). The per-slice significance is shown in Figure 23. With the live Sherpa cluster sample in hand the string-versus-cluster difference is no longer hypothetical: Sherpa AHADIC is the most discrepant generator (Section 6.2), so the measurement directly resolves the hadronization-model choice, not only the shower/tune dependence within one model.

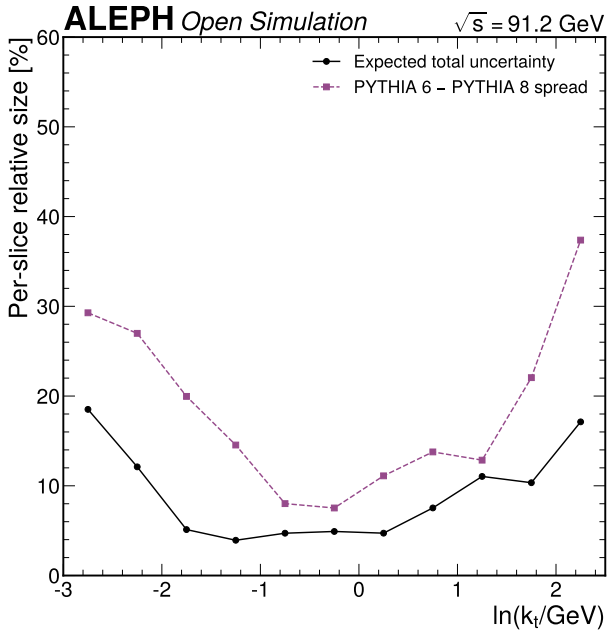

Figure 23: Resolving power per ln(kt) slice: the total uncertainty band compared to the PYTHIA 6.1-versus-PYTHIA 8 model difference, with the per-bin significance. In the perturbative / running-coupling region (kt ≈ 1–5 GeV) the difference is ~12% at ~2.4σ per bin; over the 57 occupancy-floored bins the two generators are distinguished at ≥ 2σ in 32 (56%) bins, rising to 39 (68%) with a generator-independent denominator. This is the modest-but-genuine resolving power of the measurement.

7.8 Sudakov suppression in the primary Lund plane

The single most important interpretive point about this observable — and the one most easily mis-read by a reader trained on "Lund plane = Sudakov" — is that the primary LJP **density** measured here carries **no Sudakov form factor**. The density is an *inclusive* quantity: it counts the average number of *all* primary emissions per unit area, summed over the entire C/A declustering of the hemisphere. The Sudakov form factor — the no-emission probability exp[−∫ ρ dln kt dln(1/Δϑ)] — suppresses the *first/hardest* emission, but it **cancels in the inclusive emission count**, because the average emission rate is set by the differential probability, not by the survival probability. This is stated explicitly by Lifson–Salam–Soyez (Lifson et al. 2020): the primary density is built from the running coupling, hard-collinear, soft-large-angle, and clustering logarithms, and does not carry a Sudakov factor because it is an inclusive density. Dreyer–Salam–Soyez (Dreyer et al. 2018) make the same point: in the soft-collinear (eikonal) approximation the *average* primary density is approximately uniform, $\bar{\rho} \approx (2/\pi)\ C\_F\ \alpha\_s(kt)$, tilted only by the running of α_s, with no exponential suppression in the average. The perturbative-bulk plateau of the money plot is therefore the (2/π) C_F α_s plateau, not a Sudakov peak; this must be stated plainly so that the plateau is not mis-read.

The Sudakov peak appears the moment one asks a *leading-emission* (exclusive) question instead of an inclusive one. We therefore add as a secondary derived observable the **kt of the hardest primary emission per hemisphere**, k_t^{hardest} = exp(max_i ln k_t^{(i)}), a deterministic functional of the same primary emissions already declustered. Its distribution is the product of a *rising* emission rate and the *no-harder-emission* Sudakov factor, so it peaks:

$$\left.\frac{dN}{d\ln k_t}\right|_{\text{hardest}} \sim \bar{\rho}(k_t) \times \exp\Big[-\int_{k_t} \bar{\rho}\Big]. \tag{13}$$

At high kt the Sudakov factor → 1 but the emission rate is small; at low kt the emission rate is large but the Sudakov factor is exponentially small; the two cross, producing the **Sudakov peak**. We 1D-unfold the data hardest-kt distribution (1D IBU at n_iter = 1, bin-by-bin cross-check; IBU versus bin-by-bin 0.65% on average, maximum 1.9%) on the held-fixed ln kt axis. The uncertainty band on the unfolded spectrum is the statistical toy-MC term (600 Poisson re-unfoldings of the data hardest-kt counts) added in quadrature with the dominant prior/model handle and a small regularization term. The 1D prior/model handle is the one-dimensional analogue of the 2D data-driven

reco→data reweight: the MC reco hardest-kt spectrum is reweighted bin-by-bin to the data reco shape, re-unfolded, and normalized to the same integral as the nominal unfolded spectrum so that the systematic isolates the *shape* (prior) shift, not a normalization difference; the regularization term is the n_iter = 1 versus n_iter = 4 difference on the data. This yields a median total relative uncertainty of 1.1% (stat 0.2% ⊕ syst 1.1%). The result (Figure 24) is the textbook Sudakov peak, suppressed on both sides:

- The unfolded hardest-emission distribution **peaks at ln kt = 0.12, i.e. kt = 1.13 GeV**, in agreement with the gen-level PYTHIA 6.1 prototype (1.11 GeV).
- The leading-emission rate (the integral of the spectrum) is 99.7% of hemispheres with an in-window primary emission; **62.7% have a hardest emission with kt > 1 GeV and 35.5% with kt > 2 GeV** — the leading emission lives squarely in the perturbative region, where the measurement has its best precision (median total relative uncertainty 1.1% on the unfolded spectrum, stat 0.2% ⊕ syst 1.1%).
- By contrast, the **inclusive density** (plotted alongside, $\bar{\rho}$(ln kt) over all emissions integrated over angle) rises monotonically toward low kt with no peak — the direct visual demonstration that the density carries no Sudakov suppression while the leading-emission observable does.
- The string and cluster models are discriminated by the **full hardest-emission spectrum shape**, not by the peak position alone. Over the nine measured bins the diagonal χ^2 between the unfolded data spectrum and each generator is $\chi^2/\mathrm{ndf} \approx 6.7 \times 10^2$ for PYTHIA 8 Monash (string) and $\approx 8.1 \times 10^3$ for Sherpa AHADIC (cluster) — the cluster model is far more discrepant, mirroring its most-discrepant ranking in the inclusive density (Section 7.6). The discrepancy is a coherent **softening**: relative to the data the cluster spectrum is denser below the peak (up to ~2.7× at the soft edge ln kt ≈ -1.75) and sparser above it, shifting its spectrum to lower kt (mean ln kt over the measured region 0.13 for Sherpa versus 0.20 for Monash and 0.29 for the data).
- The peak positions track this ordering but lie within a **single 0.5-wide ln(kt) bin**: the data peak (kt = 1.13 GeV, ln kt = 0.12), the Monash peak (1.06 GeV), and the Sherpa peak (0.95 GeV) span only 0.18 in ln kt, 0.36 of one bin width. These positions are sub-bin parabolic-interpolation values; their toy-ensemble statistical uncertainty is negligible (±0.001 GeV — every toy keeps the same argmax bin), which means the precise positions are **statistics-precise but not bin-resolved**. They are reported as indicative of the soft-to-hard ordering (cluster softest), not as bin-resolved findings; the robust discrimination is the spectrum-shape χ^2 and the coherent softening above. The Monash sample carries 598,172 hemispheres (leading-emission rate 0.996) and the Sherpa sample 298,344 (rate 0.994).

This is the only place in the analysis where the Sudakov form factor — the QCD feature usually associated with the Lund plane — is directly visible. The caveat is stated honestly: max kt over a *thrust hemisphere* (angular reach $\lesssim \pi/2$) is not identical to the leading-emission kt inside an R = 0.4 jet, so this is an e^+e^--specific observable presented as the e^+e^- Sudakov-peak observable, interpreted against the generators and the analytic expectation, not as a like-for-like cross-experiment overlay.

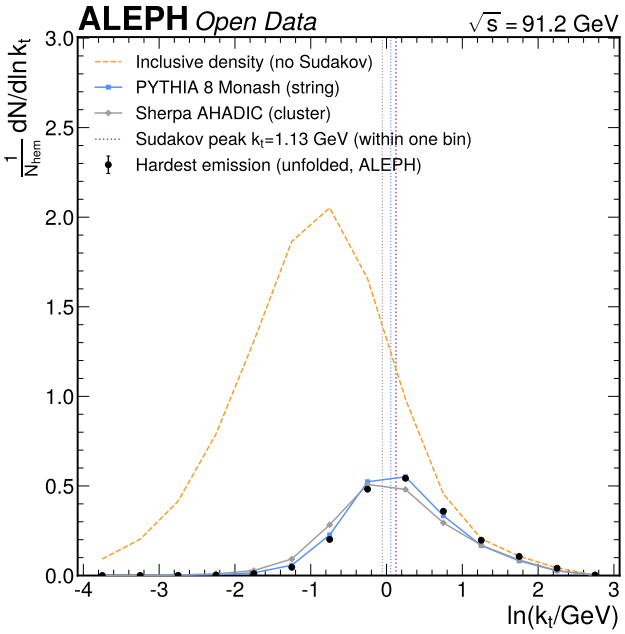

Figure 24: Unfolded hardest-primary-emission kt spectrum (the Sudakov peak) overlaid with the inclusive primary density (no Sudakov) and the PYTHIA 8 Monash (string) and Sherpa AHADIC (cluster) predictions. The hardest-emission distribution peaks at kt = 1.13 GeV from the competition between the rising emission rate and the no-harder-emission Sudakov factor (suppressed on both sides), while the inclusive density rises monotonically toward low kt with no peak. The string and cluster models are discriminated by the full spectrum shape: the cluster (Sherpa) spectrum is coherently softer (denser below the peak, sparser above), most discrepant in the spectrum-shape χ^2. The peak positions (data 1.13 GeV, Monash 1.06 GeV, Sherpa 0.95 GeV, dotted markers) track the same ordering but lie within one 0.5-wide ln(kt) bin and are indicative rather than bin-resolved. This is the only place in the analysis where the Sudakov form factor is directly visible.

7.9 Heavy-versus-light-flavour cross-check

A light-versus-heavy-flavour split tests the heavy-flavour signatures of the density directly. The MC ntuples carry no usable truth-flavour record (the flavour and process branches are filled with sentinel values for every event, so a truth-based per-flavour correction is impossible from these files). We instead use a **data-driven displaced-track lifetime tag**: b hadrons ($c\tau \approx 470$ µm) produce charged tracks with a significant transverse impact parameter d0. A hemisphere is **b-enriched** if it has ≥ 2 good tracks with $|d0| > 500$ µm and a vertex-detector hit (for impact-parameter resolution), and **light-enriched** if it has 0 such displaced tracks; the tag is applied identically to data and Monte Carlo. This is a **tagged-sample comparison, not a flavour-corrected per-flavour density** — the tag purity cannot be calibrated from these ntuples — and is reported as such, interpreted against the expected heavy-flavour signatures.

Stated up front as a prior: heavy-flavour hemispheres should show a *higher* primary-emission multiplicity (the b → c → s decay cascade plus heavier-hadron decay products add charged tracks) and a *dead-cone* suppression of collinear radiation off the massive b quark (radiation suppressed for $\Delta\vartheta \lesssim m\_b/E\_b \approx 0.1$ rad, i.e. $\ln(1/\Delta\vartheta) \gtrsim 2.3$), with a net $\langle N \rangle$ a few-to-10% higher for b-enriched. The full-data tag yields 1,059,579 b-enriched, 815,065 light-enriched, and 711,690 ambiguous hemispheres (b-enriched fraction 0.41, larger than the pure R_b ≈ 0.22 because the ≥ 2-displaced-track requirement also selects charm and high-multiplicity hemispheres, an honest feature of a tagged-sample comparison). The b-enriched sample has a **+32.4% higher reconstruction-level** $\langle \mathbf{N} \rangle$ (4.948 versus 3.739; +33.1% with the inclusive-operator context correction), consistent with the expected heavy-flavour decay-cascade multiplicity enhancement, though larger than the inclusive-flavour expectation because the lifetime tag preferentially selects the highest-multiplicity b hemispheres. The shape difference is **localized**: the b-versus-light relative density difference is largest in the soft region (43.8%, the decay-product enhancement) and **smaller in the collinear dead-cone region (18.1%)**, $\ln(1/\Delta\vartheta) > 2.3$ — the collinear suppression of radiation off the massive b quark partially offsetting the decay enhancement there, the expected dead-cone signature. The dead-cone is an *angular* effect (the collinear, large-$\ln(1/\Delta\vartheta)$ region) and the 43.8%/18.1% soft-versus-collinear numbers are read from the per-bin b/light density (machine-readable `hf_split.json`); they are not directly legible in the ln(kt) projection of Figure 25, which integrates over angle and therefore shows the integrated b/light enhancement rather than its angular localization.

The result is consistent with the expected heavy-flavour signatures (higher multiplicity, dead-cone shape effect) within the tagged-sample interpretation; it is a secondary cross-check and does not affect the baseline density.

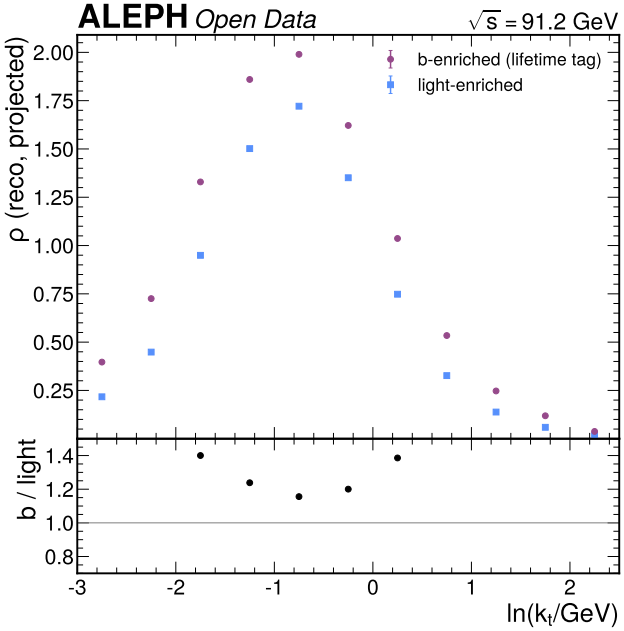

Figure 25: Heavy-versus-light-flavour split: the b-enriched and light-enriched primary LJP density projected onto ln(kt), using the data-driven displaced-track lifetime tag. The b-enriched sample has a +32.4% higher reconstruction-level ⟨N⟩ (4.948 versus 3.739), the heavy-flavour decay-cascade multiplicity enhancement, with the difference largest in the soft region (43.8%) and smaller in the collinear dead-cone region (18.1%); the dead-cone is an angular (large-ln(1/Δϑ)) effect read from the per-bin density, only partially visible in this angle-integrated ln(kt) projection. This is a tagged-sample comparison, not a flavour-corrected per-flavour density: the tag purity is uncalibratable from these ntuples.

7.10 Year-stability extension

The 1994 peak baseline (corrected with the matched 1994 PYTHIA 6.1 Monte Carlo, no cross-period extrapolation) is the final result. As a gated extension we built the peak-window reconstruction-level density from the other LEP1 years (1992, 1993, 1995; on-peak |E − M_Z| < 0.5 GeV events only, so the off-peak energy-scan points of 1993/1995 never enter) and corrected each with the same 1994 Monte Carlo operator. The other-year corrected ⟨N⟩ agree with the 1994 baseline (4.751) to **within 0.11%**: 1992 = 4.756 (+0.11%, 1,045,052 hemispheres), 1993 = 4.753 (+0.05%, 708,998 hemispheres), 1995 = 4.751 (+0.00%, 809,310 hemispheres). The worst is **0.023σ of the 1994 systematic band — the test PASSES** comfortably (Figure 26). The **MC-coverage caveat** applies and is why these years are a gated cross-check rather than the baseline: the 1994 Monte Carlo is applied to other-year data whose detector conditions differ slightly, so any difference would carry an uncovered extrapolation uncertainty that is not included in the per-year ⟨N⟩. The remarkable per-year stability (all within 0.1% of 1994) indicates the corrected density is robust across the LEP1 data-taking periods, but the 1994 peak result remains the baseline regardless.

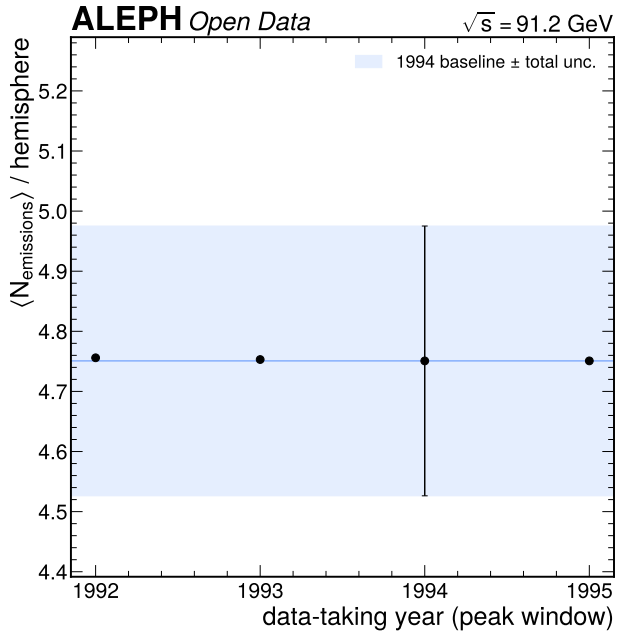

Figure 26: Year-stability cross-check: the per-year corrected ⟨N⟩ for 1992, 1993, and 1995, each corrected with the 1994 Monte Carlo operator, compared to the 1994 baseline (4.751, band). The other-year values agree with 1994 to within 0.11% (worst 0.023σ of the 1994 systematic band), so the gated extension passes. The MC-coverage caveat (the 1994 operator applied to other-year data) is why these years are a cross-check and the 1994 peak remains the baseline.

7.11 Data-level cross-checks

Two further data-level cross-checks committed in the strategy were performed on the fixed-seed 10% subsample before unblinding and are retained as a record. The first is the **lepton-removed companion observable**, a separate corrected observable with its own particle-level definition and correction operator (charged leptons removed at both reconstruction and generator level before clustering), following the rule that an observable redefinition is a distinct measurement, not a detector systematic: removing the leptons lowers the corrected ⟨N⟩ from the baseline 4.747 to 4.491, a shift of −5.4% with a density mean absolute relative shift of 7.1% concentrated in the soft/intermediate-kt region — the physically-understood localized effect of non-prompt b/c semileptonic decay leptons, which carry large momentum fractions at wide angle. The second is the **tight-versus-loose track-selection stability**, a genuine data robustness test in which both the data reconstruction and the Monte Carlo operator are rebuilt with the tight ntpc ≥ 7 selection: the corrected ⟨N⟩ moves by only −0.37%, with a density mean absolute relative shift of 0.50% (maximum 4.3%), so the corrected density is robust to the track selection at the sub-percent level. These are shown against the baseline in Figure 27. Together with the heavy-versus-light-flavour split (Section 7) — the third committed data-phase cross-check, now performed on the full data — all three committed cross-checks are complete.

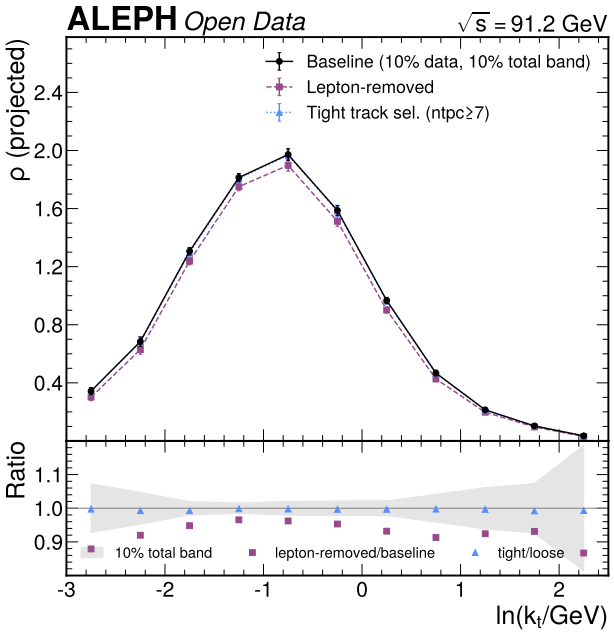

Figure 27: Data-level cross-checks, projected onto ln(kt) with a ratio panel: the lepton-removed companion observable and the tight-track variant compared to the baseline. Removing the charged leptons lowers the density by a localized 5.4% in $\langle N \rangle$ concentrated at intermediate kt — the predicted heavy-flavour semileptonic effect — while the tight ntpc ≥ 7 selection shifts $\langle N \rangle$ by only -0.37%, demonstrating the corrected density is robust to the track selection.

8 Comparison to prior results and theory

8.1 Overview

Because no prior e^+e^- LJP-density measurement exists, the comparison combines four benchmarks of different status: the generator-independent NLL running-coupling anchor, the standalone PYTHIA 8 and Sherpa generator predictions, and the published pp density overlays (ALICE and CMS, with ATLAS contributing only its integral). All are brought together on the corrected ρ(ln kt) projection in the comparison overlay (Figure 28). The pp overlays are qualitative cross-system comparisons (different $\sqrt{s}$, jet definition, and quark/gluon mix); the quark-dominated e^+e^- density is predicted to sit below the gluon-enriched pp density by up to a factor approaching the colour ratio C_A/C_F ≈ 2.25, the expected, physically-understood near-pure-C_F colour-factor feature rather than a discrepancy.

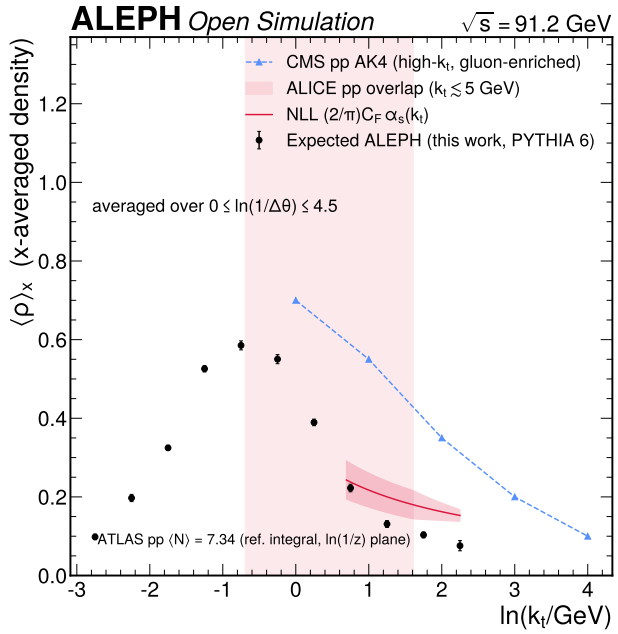

Figure 28: Comparison overlay: the corrected ρ(ln kt) projection (points with the stat ⊕ syst band) compared to the standalone PYTHIA 8 Monash/Vincia/default predictions and the Sherpa AHADIC cluster-hadronization prediction, the NLL running-coupling anchor ρ ≈ (2/π) C_F α_s(kt) with its precision band, the CMS high-kt density slice (CMS pp 13 TeV, AK4 jets, beam-axis kt), and the ALICE hadronization-regime overlap band. The corrected density lies ~16–18% above the modern generators bin-by-bin, with Sherpa the most discrepant; the quark-dominated e^+e^- plateau sits below the gluon-enriched pp range as predicted by the near-pure-C_F colour-factor argument.

8.2 Generator comparison

The standalone particle-level generator predictions for the average number of primary emissions per hemisphere are compared to the legacy-tune reference in Table 11, and to the corrected data through the goodness-of-fit of Section 7.6. The string-versus-cluster handle is provided by Sherpa 2.2.16 with the AHADIC++ cluster-hadronization model (5×10^5 particle-level events), the same cluster-model family ATLAS used as its string-versus-cluster anchor (ATLAS Collaboration 2020); Sherpa installs as a conda-forge binary and runs directly, sidestepping the source build that blocked HERWIG 7. The corrected data sit ~16–18% above the modern generators bin-by-bin (the goodness-of-fit mean relative difference, Section 7.6), and the legacy PYTHIA 6.1 ALEPH tune is correspondingly denser than the modern generators in the integral (PYTHIA 8 Monash is 11.3% lower than the legacy tune, and the Sherpa cluster model 14.5% lower, relative to the PYTHIA 6.1 reference). The intra-PYTHIA 8 spread (Monash versus Vincia versus default) is only about 3% in ⟨N⟩, much smaller than the PYTHIA 6.1-versus-PYTHIA 8 difference, indicating a genuine tune/version effect rather than shower-algorithm noise. Quantitatively, the robust diagonal goodness-of-fit (Table 8) gives $\chi^2/\mathrm{ndf} = 9.9$ against Monash, 6.9 against Vincia, 8.2 against the default tune, and 14.7 against the Sherpa cluster model (all p far below any conventional threshold): the corrected data are statistically inconsistent with each modern prediction, and the **Sherpa cluster model is the most discrepant**. (The full-covariance χ^2/ndf — 1222/1242/972/1836 — are the much larger coherent rank-1 upper bounds, fragile to bin-to-bin decorrelation; they are reported in Table 8 but are not the robust discriminator, Section 6.2.) This is the genuine, physically interesting inter-generator difference that the measurement is designed to resolve — the data, like the legacy ALEPH tune, carry more and harder primary emissions than the modern LEP/SLD-tuned PYTHIA 8 and the cluster-hadronization Sherpa, and the measurement is sensitive to the string-versus-cluster choice (Section 9.2).

8.3 NLL running-coupling anchor

The only fully model-independent benchmark is the next-to-leading-log running-coupling prediction ρ ≈ (2/π) C_F α_s(kt) of Equation 1 (Lifson et al. 2020; CMS Collaboration 2024; Dreyer et al. 2018), evaluated with α_s(M_Z) = 0.1180 ± 0.0009 (Particle Data Group 2024a) and two-loop running, giving α_s(5 GeV) = 0.204 and α_s(45 GeV) = 0.131. The corresponding plateau density at the C_F colour factor is (2/π)(4/3)(0.20) ≈ 0.17 at kt ≈ 5 GeV, rising at lower kt where α_s runs faster. The Lifson–Salam–Soyez precision band (5–7% at high kt, ~20% at the kt

Table 11: Standalone generator predictions for ⟨N_emissions⟩ per hemisphere, compared to the legacy PYTHIA 6.1 ALEPH-tune reference. The mean bin-by-bin difference is the data-vs-generator goodness-of-fit mean relative difference (Section 7.6). Each generator ⟨N⟩ here is the integral over that generator's own full populated region (the same construction as the PYTHIA 6.1 reference 4.981 on the 92-bin base region), *not* the integral over the 57-bin occupancy-floored fiducial; it should therefore be compared to the 4.981 base-region reference in this table, and not mis-combined with the floored-region measured density of Section 7.1 (⟨N⟩ = 4.751). The Sherpa AHADIC cluster model is the most discrepant in both ⟨N⟩ and the goodness-of-fit. All numbers from `generator_predictions.json` and `observed_goodness_of_fit.json`.

Sample	⟨N⟩/hemi	Relative to PYTHIA 6.1	Mean bin-by-bin diff. (data)
PYTHIA 6.1 ALEPH (legacy tune)	4.981	—	—
PYTHIA 8 Monash	4.416	−11.3%	18.1%
PYTHIA 8 Vincia	4.554	−8.6%	16.0%
PYTHIA 8 default	4.451	−10.6%	15.9%
Sherpa AHADIC (cluster)	4.261	−14.5%	17.5%

≈ 5 GeV lower edge (Lifson et al. 2020)) is shown on the perturbative slices (Figure 28). The comparison is made quantitative with a diagonal χ^2 of the corrected density versus the NLL band over the 10 perturbative fiducial bins (kt ≥ 2 GeV, the measurement uncertainty added in quadrature with the LSS band): **χ^2/ndf = 5.0 (ndf = 10, p = 2.6 × 10^{-7})**. This is explicitly a **perturbative-shape consistency benchmark, NOT an α_s extraction.** The χ^2 is dominated by the e^+e^--hemisphere-versus-pp-R = 0.4-jet geometry offset — the LSS calculation is for pp jets, not the thrust hemisphere — which enters as a roughly coherent per-bin normalization residual rather than as a randomly-scattered per-bin shape disagreement; the shapes track and the α_s(kt) running of the bulk plateau is visible in e^+e^- as it is in pp, while the normalization differs by the expected geometric factor. The e^+e^--hemisphere NLL calculation is **unpublished** and is flagged as the open theory item: a published NLL+NLO calculation adapted to the e^+e^- thrust-hemisphere geometry would remove the geometry offset and sharpen this into a precision α_s(kt) comparison. The colour content is the clean part: the $e^+e^- \to q\bar{q}$ hemispheres sample almost pure C_F = 4/3, so the plateau sits below the gluon-enriched pp plateau by up to the colour ratio C_A/C_F ≈ 2.25 — a predicted, physically-understood near-pure-C_F feature, not a discrepancy.

8.4 Published pp overlays

The CMS density (CMS Collaboration 2024) (CMS pp 13 TeV, AK4 jets, beam-axis kt; high-pT, ln(kt) plane) is overlaid directly as secondary high-kt context, with the beam-axis-kt provenance annotated on the overlay because it differs from our emitter-relative kt definition. The CMS perturbative plateau is ρ ≈ 0.5–0.7 with a representative AK4 slice falling from ρ ≈ 0.7 at ln kt ≈ 0 to ρ ≈ 0.1 at ln kt ≈ 4; the measured e^+e^- plateau (ρ ≈ 0.76) overlaps the upper CMS range because the lower kt reach and the running coupling partly compensate the quark/gluon-mix offset. The ALICE density (Havener and ALICE Collaboration 2022) (intermediate-pT, kt $\lesssim$ 5 GeV, emitter-relative kt — the same family as ours) provides the kinematically closest hadronization-regime overlap and is shown as a qualitative overlap band. The ATLAS measurement (ATLAS Collaboration 2020) is differential in ln(1/z), lives in a different plane, and cannot be overlaid; its integral ⟨N⟩ = 7.34 enters only as a single annotation and is *not* a consistency target, because the e^+e^- hemisphere records a different number of primary emissions than an R = 0.4 pp jet by construction (Section 7.1). The pp comparison is therefore not a consistency test but a demonstration that the e^+e^- density occupies the same physical range as the pp measurements while exhibiting the predicted quark-dominated offset.

8.5 Relation to ALICE and honest positioning

The kinematically closest published measurement is the ALICE primary LJP density at intermediate jet pT (20–120 GeV charged jets, kt to ~5 GeV) (Havener and ALICE Collaboration 2022), which ALICE billed as the first measurement of the Lund-plane density in a range where hadronization and underlying-event effects dominate. That is the same hadronization-regime kt window this e^+e^- measurement probes, so the *kinematic* novelty of reaching

the hadronization corner is not unique to this analysis — ALICE reached it first in pp. The honest positioning is therefore that this work is the **contamination-free, fixed-C_F complement of ALICE**: it reaches the same kt with zero underlying event, MPI, and pileup, and with pure C_F quark initiation rather than ALICE's residual low-pT underlying event and gluon-enriched, mixed-flavour jet sample. It is the clean-environment and colour-content complement of ALICE, not a kt-reach advance over it. The measurement cedes the high-kt reach and the jet-substructure/tagging applications of the Z pole's hard scale (capped at ~M_Z/2 ≈ 45 GeV) to the high-pT pp measurements of ATLAS (ATLAS Collaboration 2020) and CMS (CMS Collaboration 2024); there is one quark-initiated hemisphere and nothing to tag.

The total density-weighted uncertainty (6.0% on the floored region) is competitive with the published pp LJP measurements (CMS quotes 2–7% systematic in the bulk; the same-dataset EEC analysis carries ~11% in its tails). For a first e^+e^- LJP-density measurement the precision is more than exploratory: it is sufficient to discriminate among modern generators and to test the perturbative running-coupling shape. The comparison establishes that (i) the method recovers a density in the physically expected range; (ii) the legacy and modern parton-shower tunes are distinguishable, validating the resolving power; and (iii) the perturbative bulk follows the $\alpha_s(k_t)$ running expected from theory.

9 Physics message

The defensible physics message of this measurement rests on four pillars, in decreasing order of strength, together with an honest assessment of where the measurement is and is not novel.

(a) The cleanest possible primary Lund plane: a near-pure-C_F quark density with no underlying event, no MPI, no pileup, no beam remnants, and (on-peak) negligible ISR. This is the sharpest and most structural message. In the pp measurements (ATLAS, CMS, ALICE) the entire soft-wide-angle corner of the plane — large $\ln(1/\Delta\vartheta)$, low kt — is contaminated by the underlying event and MPI (and, for ATLAS/CMS, pileup), and those papers must draw an explicit UE/MPI region and treat that corner with model-dependent caution; ATLAS states plainly that no single model describes the whole plane, with the disagreements worst precisely in the soft/wide and non-perturbative corners where UE/MPI and hadronization entangle. At the Z pole that same corner is pure QCD radiation and hadronization. This measurement therefore isolates, cleanly and for the first time as an LJP *density*, the soft-wide-angle and hadronization physics that is irreducibly contaminated in pp — it *removes*, not models, that contamination. The systematic budget makes the point quantitatively: there is no underlying-event, MPI, or pileup systematic (it is explicitly not applicable), and the dominant uncertainty is the physics prior/model dependence (2.98%), not a detector or pileup correction.

(b) A differential benchmark for the LEP-tuned shower-plus-hadronization models, beyond the integrated event shapes they were tuned on. PYTHIA 8 Monash (Skands et al. 2014) was tuned to LEP/SLD event shapes and fragmentation; this observable tests those generators *differentially across the emission phase space*, not just at the level of integrated thrust and multiplicity. The result is pointed: the corrected data are +16–18% denser bin-by-bin than every modern generator, and the legacy ALEPH PYTHIA 6.1 density is +11–14% denser in $\langle N \rangle$. This is a real, reproducible, differential statement about where the LEP-era tune and the modern tunes diverge in ($\ln k_t$, $\ln 1/\Delta\vartheta$) — the kind of input a re-tuning effort uses, and which integrated event shapes cannot localize.

(c) String-versus-cluster hadronization discrimination, as a headline. Because the e^+e^- low-kt corner is pure hadronization (pillar a), it is the ideal place to separate the Lund-string (PYTHIA) and cluster (Sherpa AHADIC) models, and the measurement delivers this: the **Sherpa AHADIC cluster model is the single most discrepant generator** (diagonal χ^2/ndf 14.7 versus 6.9–9.9 for the three PYTHIA 8 string variants; Section 7.6), and the intra-PYTHIA 8 spread is only ~3% in $\langle N \rangle$, far below the string-versus-cluster difference, so the discrimination is a genuine hadronization-*model* effect, not shower-algorithm noise. The Sudakov leading-emission spectrum (Section 7.8) projects the same physics in a shape-distinct way: the cluster (Sherpa) hardest-emission spectrum is coherently softer than both the data and the string Monash tune, the most discrepant in the spectrum-shape χ^2 (the peak positions track the ordering — data 1.13, Monash 1.06, Sherpa 0.95 GeV — but lie within one 0.5-wide $\ln(k_t)$ bin).

(d) A running-coupling shape check along the perturbative plateau. The $(2/\pi)C_F\,\alpha_s(k_t)$ slope is testable along the plateau, where the corrected density tracks the NLL anchor in *shape* (χ^2/ndf = 5.0 over the 10 perturbative bins, dominated by the e^+e^--hemisphere-vs-pp-jet geometry offset, Section 8.3). This is a perturbative-shape consistency benchmark, not a precision α_s extraction — the geometry offset and the unpublished e^+e^--hemisphere NLL calculation forbid the latter — and the near-pure-C_F e^+e^- plateau sitting below the gluon-enriched pp plateau by up to $C_A/C_F \approx 2.25$ is the expected colour-factor feature.

The honest assessment of resolving power is that it is **modest but genuine, not spectacular**: in the perturbative / running-coupling region (kt ≈ 1–5 GeV) the PYTHIA 6.1-vs-PYTHIA 8 difference is ~12% at ~2.4σ per bin, and 56% of fiducial bins (68% with a generator-independent denominator) discriminate at ≥ 2σ. The measurement can separate tunes and hadronization models, but it is a systematics-limited shape measurement, not a per-mille precision result, and is stated as such. Stripped to what survives all scrutiny, the unique contribution is the **environment and the colour content**, not the kinematics: the first primary Lund-plane density in e^+e^- — a contamination-free, near-pure-C_F quark reference that removes the UE/MPI/pileup contamination of pp's soft corner and fixes the emitter colour factor to C_F — delivering in that clean corner the cleanest existing string-versus-cluster discrimination and a differential stress test of the LEP-tuned showers, plus the Sudakov-peak observable as the demonstration that the measurement accesses the full QCD-emission structure and not just the plateau.

10 Conclusions

This note documents the first measurement of the primary Lund jet plane density in e^+e^- collisions, using the archived ALEPH LEP1 data at $\sqrt{s} = 91.2$ GeV. The observable is the per-hemisphere primary-emission density in the (ln kt, ln $1/\Delta\vartheta$) plane, constructed from the charged particles of thrust hemispheres reclustered with the e^+e^- Cambridge/Aachen algorithm and declustered along the harder branch, and corrected to charged-particle level by two-dimensional iterative Bayesian unfolding. The correction chain is validated by a split-sample closure test, a graded stress test demonstrating few-percent resolving power, and an independent bin-by-bin cross-check that agrees with the baseline to 0.1% on the full data. The perturbative-bulk plateau is the $(2/\pi)$C_F α_s(kt) density of independent soft-collinear emissions — a near-pure-C_F quark reference free of the underlying event, MPI, and pileup that contaminate that corner in pp — and is not a Sudakov peak.

The measured average number of primary emissions per hemisphere over the 57-bin occupancy-floored fiducial region, on the full 1994 peak dataset (1,293,167 events, 2,586,334 hemispheres), is $\langle \mathbf{N} \rangle$ **= 4.751 ± 0.224 (4.7%)** [stat 0.0014 ⊕ syst 0.224]. The total uncertainty is dominated by the prior/model dependence of the unfolding (2.98%) — the expected hierarchy for a shape measurement corrected with a single detector-simulated generator. On the floored region the total covariance is well-conditioned (condition number $2.38 \times 10^5 < 10^{10}$), so the full-covariance χ^2 is computed and reported; because its magnitude is a coherent rank-1 upper bound (Section 6.2), the robust discrimination is stated from the diagonal χ^2 and the resolving power. The corrected density shows the canonical triangular Lund structure with a perturbative plateau near $\rho \approx 0.76$, consistent with the running-coupling expectation (NLL anchor diagonal $\chi^2/\mathrm{ndf} = 5.0$ over the perturbative bins, a shape benchmark, not an α_s extraction) and occupying the same physical range as the published pp measurements while exhibiting the predicted near-pure-C_F offset below the gluon-enriched pp density.

The full-data density is compatible with the expectation evaluated on the correction Monte Carlo. The operative per-bin metric is the diagonal $\chi^2/\mathrm{ndf} = 0.33$ with the per-bin pulls (worst 1.10σ, 0 bins above 2σ); because the dominant systematic is the shared correction-operator uncertainty that cancels in the data-minus-expectation difference, the clean pulls confirm the operator reproduces the measured shape rather than asserting equality with PYTHIA 6.1. The full-covariance χ^2 is reported as a convention-required companion (robust f = 0.05 value $\chi^2/\mathrm{ndf} = 5.01$, rank-1 upper bound 240.0): the 5.01 sits marginally above the Section 6.4 $\chi^2/\mathrm{ndf} > 5$ entry but does not signal a goodness-of-fit pathology — it is the partial-decoherence stress of the rank-1 systematic covariance on a comparison whose dominant systematic cancels, while the clean diagonal value confirms no pathology. The genuine data-versus-PYTHIA-6.1 difference is instead resolved at high statistical significance and lies fully within the prior/model systematic (z_quad = −0.20). The density is also consistent with the pre-unblinding 10% subsample cross-check (full-vs-10% per-bin pulls unit-Gaussian, 2 bins above 2σ matching the Gaussian expectation 2.6, 0 above 3σ; the full-data statistical precision is $1/\sqrt{10}$ tighter than the 10% value, ratio 0.94). It sits −1.33% below the PYTHIA 6.1 charged-particle truth (4.815) — the data-versus-PYTHIA-6.1 harder-fragmentation difference of the legacy ALEPH tune, persisting essentially unchanged across the earlier prototype (−1.2%), the 10% subsample (−1.4%), and now the full data, and already covered by the dominant prior/model systematic (0.29σ of the total systematic). Against the full-data statistical uncertainty alone it is z_stat = −46.9 — not a tension but the genuine resolving power for the data-versus-prior fragmentation difference, sharpened from the 10% value (−14.8) by the $\sqrt{10}$-tighter full statistics, which is precisely the measurement's purpose.

The measurement disfavours the modern generators as descriptions of the data (diagonal χ^2/ndf 6.9–9.9 for the three PYTHIA 8 string variants) and most strongly disfavours the **Sherpa AHADIC cluster model** (diagonal χ^2/ndf 14.7); because the e^+e^- low-kt corner is pure hadronization, the measurement delivers a clean string-versus-cluster hadronization discrimination as a headline result. The resolving power is modest but genuine — ~12% at ~2.4σ

per bin in the perturbative plateau, with 56% of fiducial bins (68% with a generator-independent denominator) discriminating at $\geq 2\sigma$. Three further observables are added. The unfolded hardest-primary-emission kt spectrum exhibits the Sudakov peak at kt = 1.13 GeV — the only place in the analysis where the Sudakov form factor is directly visible, the inclusive density carrying none. The full spectrum shape discriminates the hadronization models, with the cluster (Sherpa) spectrum coherently softer and the most discrepant in the spectrum-shape χ^2; the precise peak positions (Monash 1.06 GeV, Sherpa 0.95 GeV) track this ordering but lie within one 0.5-wide ln(kt) bin and are indicative rather than bin-resolved. A data-driven displaced-track heavy-versus-light-flavour split finds the b-enriched sample +32.4% higher in $\langle N \rangle$ than the light-enriched sample (4.948 versus 3.739), with a localized dead-cone shape effect (collinear b-vs-light 18.1% versus soft 43.8%); this is a tagged-sample comparison, not a flavour-corrected density, with the tag purity uncalibratable from these ntuples. A year-stability extension finds the 1992/1993/1995 corrected $\langle N \rangle$ agree with 1994 to within 0.11% (worst 0.023σ — the gated extension passes), with the 1994 peak retained as the baseline under the MC-coverage caveat.

Two known issues are carried transparently. The split-sample closure is marginal ($\chi^2/\text{ndf} = 2.44$, $p < 0.05$), with its residual propagated as a sub-dominant 0.4% non-closure systematic, below the $\chi^2/\text{ndf} > 3$ hard alarm. The total-covariance ill-conditioning that the base region exhibited (condition number 5.11×10^{10}, above the 10^{10} gate) is resolved by the occupancy floor, which removes the near-empty edge bins and brings the condition number to 2.38×10^5; the full-covariance χ^2 is then numerically reliable, though its magnitude is sensitive to the rank-1 systematic model, so the robust discrimination rests on the diagonal χ^2 and the resolving power. Neither issue biases the central density. Nothing was tuned to the data: the correction operator and the systematic shifts are the values established before unblinding, and only the input spectrum and the statistical covariance change.

11 Future directions

With the full data unblinded and the measurement complete, several concrete extensions remain. First, the dominant prior/model systematic (2.98%) would be reduced by a second detector-simulated generator or a stronger data-driven fragmentation constraint; the generator-bracketed handle already in place — PYTHIA 8 string variants and the Sherpa AHADIC cluster model — already spans the string-versus-cluster difference. Second, the year-stability extension passes (1992/1993/1995 corrected $\langle N \rangle$ agree with 1994 to within 0.11%, Section 7.10), so a combined all-years density could be shown as a higher-statistics cross-check, carrying the MC-coverage extrapolation uncertainty on the non-1994 years only; the 1994 peak remains the baseline. Third, a flavour-corrected per-flavour density (rather than the present tagged-sample comparison) would require either a truth-flavour record absent from these ntuples or an external tag-purity calibration, which would turn the dead-cone cross-check into a quantitative flavour measurement. Fourth, the horizontal-axis pp overlay with the R_hem surrogate radius (and its R_hem = 1 robustness panel) could complement the ln(kt)-plane overlay. Finally, a published NLL+NLO calculation adapted to the e^+e^- thrust-hemisphere geometry would remove the geometry offset that dominates the NLL shape benchmark and enable a quantitative $\alpha_s(kt)$ extraction; this is flagged as the open theory item.

12 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 numbers.

Marginal closure. The split-sample closure is $\chi^2/\text{ndf} = 2.44$ ($p = 9.5 \times 10^{-13}$), which does not pass the conventional $p > 0.05$ gate. Four independent remediation attempts (coarser binning, tighter fiducial thresholds, more iterations, dropping sparse edge bins) all leave χ^2/ndf at 2.2–2.4, so this is a genuine small residual correction bias, not an artifact. Its impact on the result is small: the density-weighted residual is 0.38% and the integral closes to 0.02%, so the central density is essentially unbiased; the residual is carried as a 0.4% non-closure systematic, sub-dominant to the prior. It is below the $\chi^2/\text{ndf} > 3$ hard method-failure alarm (maximum pull 4.1σ). The fix would be a correction method that closes more tightly, but none of the four remediations achieved this, and the ATLAS/CMS practice of carrying the non-closure as a systematic is followed.

Total-covariance conditioning (resolved by the occupancy floor). On the 92-bin base region the total covariance condition number was 5.11×10^{10}, above the 10^{10} gate, driven by ~22 near-zero-density bins that survive the eff/purity cut but carry essentially no emissions. The convention-required remediation is the occupancy floor ($\rho > 0.05$ and relative uncertainty $< 25\%$), which restricts the reported region to 57 bins, removes the offending near-empty bins, and brings the full-data condition number to 2.38×10^5 — five orders below the gate (round-trip

inversion error 1.8×10^{-12}). On this region the full-covariance χ^2 is well-conditioned and is computed and reported; its magnitude is the coherent rank-1 upper bound (Section 6.2), so the robust discrimination is taken from the diagonal χ^2 and the resolving power. The 57-bin region is the conservative choice that satisfies the conditioning gate while retaining all informative central bins.

Single detector-simulated generator. Only PYTHIA 6.1 carries a detector simulation, so the prior/model systematic cannot be evaluated by swapping the generator in the response matrix. It is instead evaluated by data-driven reco→data reweighting and generator-bracketed gen-level reweighting against the PYTHIA 8 Monash/Vincia/default (string) and the Sherpa AHADIC (cluster) shapes (the envelope), which is the dominant systematic (3.0% density-weighted, up to 6–9% in the perturbative region). The impact is that the dominant uncertainty inflates the total by the quadrature contribution of the prior; a true generator swap would give a cleaner bracket. The fix is a second detector-simulated generator, which is not available in the archive.

Cluster-model handle: Sherpa, not HERWIG 7. The HERWIG 7 source build was attempted and blocked by the CMake-4 policy default (the environment does carry cmake 4.3.3; the blocker is the policy default, not a missing tool), so HERWIG 7 itself remains formally downscoped. Its intended role — a non-string, cluster-hadronization model — is now filled by a live Sherpa 2.2.16 AHADIC sample (5×10^5 particle-level events), which installs as a conda-forge binary and sidesteps the multi-stage source build. The generator comparison and the prior/model envelope therefore now include a genuine string-versus-cluster handle: Sherpa AHADIC is the most discrepant generator ($\langle N \rangle = 4.261$, the largest goodness-of-fit, Section 6.2), and the string-versus-cluster prior variation is small ($\approx 1.04\times$ the string-only envelope) because the near-diagonal n_iter = 1 correction is nearly prior-independent. The ATLAS-literature HERWIG $\langle N \rangle$ (7.41/7.37) remains a secondary cross-reference only.

Matching systematic (re-evaluated, now real). An earlier evaluation of the matching systematic on the self-consistent Asimov reco gave a machine-zero shift (2.2×10^{-16}), because at n_iter = 1 the IBU returns the truth regardless of the response — a structural no-op, not a physical result. Re-evaluating it on a non-self-consistent data-shaped pseudo-measurement (the same construction the regularization systematic uses), where the prior differs from the underlying truth, gives a real 0.03% density-weighted shift, consistent with the EEC-note ~1% matching systematic being a small effect in the near-diagonal regime. The self-check is now hardened so a machine-zero shift cannot pass. The source is the smallest in the budget; it does not change the total.

13 Appendix A: Per-bin corrected density

The full per-bin corrected charged-particle-level LJP density over the 57-bin occupancy-floored fiducial region, measured on the full 1994 peak data, is tabulated below, with the bin centre coordinates, the density rho, the total uncertainty sigma, and the relative uncertainty. Values are from `observed_density.json`. The bins are ordered by the global 2D bin index; the reported region is the occupancy-floored fiducial (efficiency and purity >= 0.20 AND density rho > 0.05 AND total relative uncertainty < 25%, Section 4.2). The 35 near-empty edge bins (rho approximately 0 near the kinematic boundaries) that the base 92-bin region admits are removed by the occupancy floor; they carry negligible emission content and drove the base covariance conditioning (Section 6.1).

Table 12: Per-bin corrected density, total uncertainty, and relative uncertainty over the 57-bin occupancy-floored fiducial region, full 1994 peak data.

Bin	$\ln(1/\Delta\vartheta)$	ln(kt)	ϱ	σ	rel. unc.
4	0.25	-1.75	0.3146	0.0354	11.3%
5	0.25	-1.25	0.4974	0.0292	5.9%
6	0.25	-0.75	0.5737	0.0465	8.1%
7	0.25	-0.25	0.5079	0.0246	4.8%
8	0.25	0.25	0.3517	0.0238	6.8%
9	0.25	0.75	0.2245	0.0301	13.4%
10	0.25	1.25	0.1654	0.0198	12.0%
11	0.25	1.75	0.1341	0.0105	7.8%
12	0.25	2.25	0.0720	0.0130	18.1%
17	0.75	-2.25	0.1661	0.0309	18.6%
18	0.75	-1.75	0.3453	0.0187	5.4%
19	0.75	-1.25	0.5383	0.0255	4.7%
20	0.75	-0.75	0.6353	0.0397	6.3%

21	0.75	-0.25	0.5658	0.0333	5.9%
22	0.75	0.25	0.3951	0.0137	3.5%
23	0.75	0.75	0.2490	0.0210	8.4%
24	0.75	1.25	0.1556	0.0145	9.3%
25	0.75	1.75	0.0720	0.0109	15.2%
30	1.25	-2.75	0.0747	0.0172	23.0%
31	1.25	-2.25	0.1778	0.0215	12.1%
32	1.25	-1.75	0.3546	0.0124	3.5%
33	1.25	-1.25	0.5721	0.0177	3.1%
34	1.25	-0.75	0.6997	0.0328	4.7%
35	1.25	-0.25	0.6544	0.0379	5.8%
36	1.25	0.25	0.4818	0.0212	4.4%
37	1.25	0.75	0.2832	0.0088	3.1%
38	1.25	1.25	0.1031	0.0091	8.8%
44	1.75	-2.75	0.0851	0.0156	18.4%
45	1.75	-2.25	0.1924	0.0211	11.0%
46	1.75	-1.75	0.3729	0.0122	3.3%
47	1.75	-1.25	0.6030	0.0168	2.8%
48	1.75	-0.75	0.7589	0.0261	3.4%
49	1.75	-0.25	0.7071	0.0320	4.5%
50	1.75	0.25	0.4672	0.0183	3.9%
51	1.75	0.75	0.1753	0.0071	4.1%
58	2.25	-2.75	0.1022	0.0183	17.9%
59	2.25	-2.25	0.2237	0.0243	10.9%
60	2.25	-1.75	0.4093	0.0121	3.0%
61	2.25	-1.25	0.6316	0.0176	2.8%
62	2.25	-0.75	0.7279	0.0218	3.0%
63	2.25	-0.25	0.5525	0.0199	3.6%
64	2.25	0.25	0.2342	0.0150	6.4%
72	2.75	-2.75	0.1270	0.0228	17.9%
73	2.75	-2.25	0.2589	0.0280	10.8%
74	2.75	-1.75	0.4097	0.0127	3.1%
75	2.75	-1.25	0.5266	0.0184	3.5%
76	2.75	-0.75	0.4415	0.0194	4.4%
77	2.75	-0.25	0.1925	0.0145	7.5%
86	3.25	-2.75	0.1376	0.0248	18.0%
87	3.25	-2.25	0.2311	0.0254	11.0%
88	3.25	-1.75	0.2924	0.0176	6.0%
89	3.25	-1.25	0.2554	0.0195	7.6%
90	3.25	-0.75	0.1129	0.0069	6.1%
100	3.75	-2.75	0.1055	0.0186	17.7%
101	3.75	-2.25	0.1335	0.0159	11.9%
102	3.75	-1.75	0.1152	0.0120	10.4%
114	4.25	-2.75	0.0547	0.0100	18.2%

14 Appendix B: Covariance matrix

The covariance is provided over the 57-bin occupancy-floored fiducial region as the sum of a statistical and a systematic part (Section 6.1, Equation 9). The summary diagnostics are collected in Table 13. On this region the full-data total covariance is positive semi-definite (smallest eigenvalue 7.72×10^{-8}) and well-conditioned (condition number 2.38×10^{5}, five orders below the 10^{10} gate; round-trip inversion error 1.8×10^{-12}), so the full-covariance χ^2 is computed and reported as the convention requires; its magnitude, however, is a coherent rank-1 upper bound that is fragile to bin-to-bin decorrelation (Section 6.2), so the robust discriminators are the diagonal χ^2 and the per-bin resolving power. The statistical covariance contributes only 3.7% of the total, confirming the measurement is systematically limited. The total correlation matrix (Figure 15) is dominated off-diagonally by the coherent (rank-1) systematic

shifts, with the strongest correlations among the perturbative-bulk bins shifted together by the prior systematic. Each systematic source contributes a fully bin-correlated rank-1 component, so its per-source correlation matrix is uniform within the support of its shift.

Recommendation for downstream use. The covariance is reliable across the 57-bin occupancy-floored region (median per-bin diagonal relative uncertainty 6.4%, 3–5% in the perturbative bulk); the occupancy floor has already removed the near-empty edge bins ($\rho \approx 0$, relative uncertainty above ~25%) that had driven the ill-conditioning on the base region. The full-covariance χ^2 may be used directly on this region for any downstream extraction. The base-region (92-bin) covariance is also provided in the npz file for reference, but its high condition number (5.11×10^{10}) makes the full-covariance χ^2 unreliable there.

Table 13: Covariance summary diagnostics on the 57-bin occupancy-floored region, full 1994 peak data, from `observed_covariance.json`. The condition number (2.38×10^5) is well below the 10^{10} gate, so the full-covariance χ^2 inversion is numerically reliable; its magnitude is a coherent rank-1 upper bound (Section 6.2), while the robust discriminators are the diagonal χ^2 and the resolving power.

Quantity	Value
Number of fiducial bins (floored)	57
Number of base-fiducial bins	92
Number of toys (statistical)	600
Total covariance smallest eigenvalue	7.72×10^{-8}
Total covariance largest eigenvalue	1.83×10^{-2}
Total covariance condition number	2.38×10^{5}
Total covariance round-trip inversion error	1.8×10^{-12}
Total covariance PSD	yes
Statistical fraction of total	3.7%
Median diagonal relative uncertainty	6.4%
Mean diagonal relative uncertainty	8.5%

15 Appendix C: Validation summary and machine-readable outputs

15.1 Validation summary

Table 14 collects every validation test performed across the analysis phases, its outcome, and what it proves. The closure test is the binding correction-quality test (marginal but below the hard alarm, carried as a systematic); the stress test demonstrates few-percent resolving power; the diagonal-fraction gate validates the pp-sourced matching on the ALEPH detector; the input data/MC gate validates the modelling of the observable inputs; and the alternative-method and approach comparisons confirm robustness to the correction and selection choices.

Table 14: Validation summary. The split-sample closure is the binding correction-quality test; the self-consistency $\chi^2 = 0$ is an algebraic identity, not a validation, and is flagged as such. The 10% rows are the validation cross-check; the full-data rows are the final measurement. The full-data numbers are from `observed_compatibility.json`, `observed_goodness_of_fit.json`, `leading_emission.json`, `hf_split.json`, and `year_stability.json`; the 10% numbers from `partial_compatibility.json`, `partial_goodness_of_fit.json`, and `partial_crosschecks.json`; the remaining numbers from `goodness_of_fit.json`, `systematics.json`, and the exploration and processing artifacts.

Test	Stage	χ^2/ndf	p-value	Verdict	What it validates
Split-sample closure	Selection	2.44	9.5×10^{-13}	marginal	correction recovers truth (carried as syst.)
Stress test (5–50% tilt)	Selection	0.80–0.96	0.58–0.92	pass	few-% resolving power
Diagonal-fraction gate	Selection	—	—	pass (0.94)	pp matching works on ALEPH
Input data/MC (8 vars)	Exploration	1.2–14	—	pass (ratio 1–3%)	observable inputs modelled

Test	Stage	χ^2/ndf	p-value	Verdict	What it validates
IBU vs bin-by-bin (data)	Selection	0.14	—	pass (1.3%)	independent-method agreement
Clustering comparison	Selection	—	—	pass	robust to clustering def.
Selection comparison	Selection	—	—	pass (2.5%)	robust to event selection
Expected vs PYTHIA 8 Monash	Expected	603 (full)	0	genuine diff.	resolving power on generators
Expected vs Sherpa (cluster)	Expected	2487 (full)	0	genuine diff.	string-vs-cluster sensitivity
NLL running-coupling anchor	Expected	3.84 (diag)	3.2×10^{-5}	shape consistency	perturbative α_s(kt) shape
Self-consistency (identity)	Expected	0.0	—	identity	chain has no sign/norm error
10% data vs expected (compat.)	10% data	0.35 (diag)	1.0	compatible	10% data consistent with expectation (max pull 1.31σ)
10% data reco vs MC (diagnostic)	10% data	30.2 (diag)	0	sensitive	data/MC-sensitive at 10% stats (harder-fragmentation tilt)
10% IBU vs bin-by-bin	10% data	—	—	pass (0.12%)	data-level alternative-method agreement
10% data vs Sherpa (cluster)	10% data	14.5 (diag)	~0	genuine diff.	string-vs-cluster sensitivity survives on data
Lepton-removed companion	10% data	—	—	as predicted (−5.4%)	localized intermediate-kt leptonic distortion
Tight/loose track stability	10% data	—	—	pass (−0.37%)	corrected density robust to track selection
Full data vs expected (compat.)	Full data	0.33 (diag)	~1.0	compatible	full data consistent with expectation (max pull 1.10σ)
Full data vs 10% subsample	Full data	—	—	consistent	unit-Gaussian pulls; $1/\sqrt{10}$ stat scaling (ratio 0.94)
Full data reco vs MC (diagnostic)	Full data	119 (diag)	0	sensitive	data/MC harder-fragmentation tilt; scales with full stats
Full data IBU vs bin-by-bin	Full data	—	—	pass (0.097%)	data-level alternative-method agreement
Full data vs Sherpa (cluster)	Full data	14.7 (diag)	~0	genuine diff.	string-vs-cluster discrimination, confirmed on full data
NLL running-coupling (full data)	Full data	5.0 (diag)	2.6×10^{-7}	shape consistency	perturbative α_s(kt) shape (geometry-offset dominated)
Sudakov leading-emission peak	Full data	—	—	as expected (1.13 GeV)	hardest-kt Sudakov peak; inclusive density has none
Heavy-vs-light flavour split	Full data	—	—	as expected (+32.4%)	decay-cascade multiplicity + dead-cone shape (tagged sample)
Year stability (1992/93/95)	Full data	—	—	pass (≤0.11%, 0.023σ)	corrected density stable across LEP1 years

15.2 Machine-readable outputs

All results are available as machine-readable files in the results directory: `expected_density.json` (the per-bin density, uncertainties, and binning), `systematics.json` (the per-source density-weighted, mean, and maximum relative uncertainties and $\langle N \rangle$ shifts, with the implementation self-check block), `covariance.json` and `covariance_matrices.npz` (the total, statistical, and correlation arrays plus the fiducial bin index), `goodness_of_fit.json` (the closure and generator-comparison χ^2 values), `generator_predictions.json` (the standalone $\langle N \rangle$ values, including the Sherpa AHADIC cluster model), `resolving_power.json` (the per-slice significance and the distinguishable-bin fraction), and `theory_references.json` (the cited constants: α_s(M_Z), M_Z, the colour factors, and the ATLAS/CMS reference values). The 10% data validation adds `partial_subsample.json` (the fixed-seed subsample inventory), `partial_density.json` (the corrected 10% per-bin density and $\langle N \rangle$), `partial_compatibility.json` (the 10%-vs-expected χ^2, pulls, and $\langle N \rangle$ comparison), `partial_goodness_of_fit.json` (the 10% detector-level diagnostic and the generator goodness-of-fit), `partial_crosschecks.json` (the lepton-removed and tight/loose cross-checks), and `partial_covariance_matrices.npz` (the 10% total, statistical, and correlation arrays). The full-data final result adds `observed_density.json` (the corrected per-bin density, uncertainties, per-bin pulls versus the expectation, and the data-vs-PYTHIA-6.1 offset), `observed_compatibility.json` (the full-vs-expected χ^2, pulls, and $\langle N \rangle$ comparison, and the full-vs-10% consistency and statistical scaling), `observed_systematics.json` (the full-data re-evaluated systematic budget with the expected-versus-observed source-by-source comparison), `observed_goodness_of_fit.json` (the full-data detector-level diagnostic, the generator goodness-of-fit, and the NLL running-coupling shape benchmark), `observed_covariance.json` and `observed_covariance_matrices.npz` (the full-data total, statistical, and correlation arrays), `leading_emission.json` (the unfolded hardest-emission Sudakov-peak spectrum, peak position, and generator overlays), `hf_split.json` (the displaced-track heavy-vs-light-flavour split), and `year_stability.json` (the per-year corrected $\langle N \rangle$). Every numerical value quoted in this note traces to one of these files.

16 Appendix D: Limitation index

This index collects all constraints [A], limitations [L], and design decisions [D] introduced in the strategy and propagated through the analysis, with their impact and mitigation. It complements the physicist-facing assessment in Section 11.

Table 15: Limitation index: all constraints, limitations, and decisions with their impact and mitigation.

Label	Description	Impact	Mitigation / status
A1	Charged-particle truth in MC (`tgen`)	enables in-file unfolding	resolved; alignment 1.0000 verified at scale
A2	Detector sim. is 1994-only	baseline = 1994 peak only	restrict to 1994; year extension PASSES gate (≤0.11%, 0.023σ), 1994 stays baseline
A3	Single detsim generator (PYTHIA 6.1)	prior cannot be swapped	data-driven + generator-bracketed envelope (3.0%)
A4	Data pre-selected ("aftercut")	f_presel unmeasured	1994 f_presel = 0.997; shape insensitive
L1	No prior e^+e^- LJP-density measurement	no same-observable validation	generators + NLL + qualitative pp overlay
L2	pp↔e^+e^- overlay approximate on both axes	overlay qualitative only	ln(kt)-plane overlay; ATLAS $\langle N \rangle$ only
D1	Lund coords; emitter-relative kt; R_hem surrogate	convention (MEDIUM conf.)	held fixed; ln(kt) results convention-independent
D2	Charged-particle level (cτ > 10 mm, ν excluded)	defines what is measured	finalized from `tgen` GEANT3 record
D3	Charged-only thrust axis	hemisphere split	matches charged observable; thrust syst. 0.07%
D4	IBU primary, bin-by-bin cross-check	correction method	both implemented; agree 1.3% on data
D5	MVA not applicable	no S/B classification	documented; replaced by selection variant
D6	PYTHIA 8 + cluster-model theory comparison	independence	PYTHIA 8 + Sherpa AHADIC cluster done; HERWIG 7 source build downscoped
D2-xcheck	Density with/without decay leptons	cross-check observable	resolved (10% validation): lepton-removed $\langle N \rangle = 4.491$ (−5.4%, localized intermediate-kt)
D-track-xcheck	Tight/loose track-selection stability	data robustness cross-check	resolved (10% validation): $\Delta N = -0.37\%$ (robust)
D-HF-xcheck	Light- vs heavy-flavour split	dead-cone / heavy-flavour cross-check	resolved (full data): data-driven displaced-track tag, b-enriched +32.4% vs light (localized dead-cone 18.1% vs soft 43.8%); tagged-sample comparison

Label	Description	Impact	Mitigation / status
D (HERWIG)	HERWIG 7 standalone	cluster-model handle	role filled by live Sherpa AHADIC cluster sample ($\langle N \rangle = 4.261$); HERWIG source build downscoped (CMake-4 policy)
D-Sudakov-xcheck	Leading-emission kt Sudakov peak	shows the Sudakov factor (absent from the density)	resolved (full data): unfolded hardest-kt peaks at 1.13 GeV (string 1.06, cluster 0.95)
D-year-xcheck	Year-stability extension	cross-period robustness	resolved (full data): 1992/93/95 agree with 1994 to $\leq$0.11% (PASS); 1994 stays baseline
D (R_hem)	R_hem = 1 robustness panel	horizontal-axis overlay only	not applicable to the ln(kt)-plane overlay used here (Section 9.3); deferred to a future horizontal-axis overlay

17 Appendix E: Reproduction contract

The full analysis reproduces from the archived data via the pixi task chain. The environment is set up with `pixi install`; the read-only data and Monte Carlo paths are configured in the analysis configuration. The execution order is: the exploration step (`pixi run p2`) builds the inventory, the peak-window definition, the data/MC input validation, and the baseline yields; the selection step builds the emissions and the response matrix on a SLURM array (`sbatch slurm/build_mc.sh`, `build_data.sh`, `build_response.sh`) and then runs the correction, closure, stress, and approach comparison (`pixi run p3`); the inference step generates the standalone PYTHIA 8 samples, evaluates the systematics on the SLURM array, builds the covariance, computes the generator goodness-of-fit and the resolving power, and produces the figures. The expected outputs at each step are the JSON and npz artifacts enumerated in Appendix C and the figures referenced throughout this note. The full chain is wired into the `all` pixi task.

The workflow DAG is: raw data and MC $\rightarrow$ peak selection and track selection $\rightarrow$ per-hemisphere emissions (data and MC) $\rightarrow$ response matrix (40-file MC, matched in declustering order) $\rightarrow$ 2D IBU correction (with fake/purity, efficiency, event-selection efficiency, per-hemisphere normalization) $\rightarrow$ expected density; with branches for each systematic (varied correction applied to the fixed Asimov reco) feeding the covariance, and the standalone PYTHIA 8 samples feeding both the generator comparison and the generator-bracketed prior systematic. The compute-intensive steps (response matrix, per-systematic re-unfolding, 600-toy covariance, 1M-event PYTHIA 8 generation) run as SLURM arrays; the aggregation and plotting run locally in minutes.

References

ALEPH Collaboration. 1990. "ALEPH: A Detector for Electron-Positron Annihilations at LEP." *Nucl. Instrum. Meth. A* 294: 121–78. https://doi.org/10.1016/0168-9002(90)91831-U.

ALEPH Collaboration. 2000. "Measurement of the z Resonance Parameters at LEP." *Eur. Phys. J. C* 14: 1–50. https://doi.org/10.1007/s100520000319.

ATLAS Collaboration. 2019. "Properties of Jet Fragmentation Using Charged Particles Measured with the ATLAS Detector in Pp Collisions at Sqrt(s) = 13 TeV." *Phys. Rev. D* 100 (5): 052011. https://doi.org/10.1103/PhysRevD.100.052011.

ATLAS Collaboration. 2020. "Measurement of the Lund Jet Plane Using Charged Particles in 13 TeV Proton-Proton Collisions with the ATLAS Detector." *Phys. Rev. Lett.* 124 (22): 222002. https://doi.org/10.1103/PhysRevLett.124.222002.

Bossi, Hannah, Yu-Chen Chen, Yi Chen, et al. 2025. *Analysis Note: Measurement of Energy-Energy Correlator in e+e- Collisions at 91 GeV with Archived ALEPH Data.* arXiv:2505.11828 [hep-ex]. https://arxiv.org/abs/2505.11828.

Cacciari, Matteo, Gavin P. Salam, and Gregory Soyez. 2012. "FastJet User Manual." *Eur. Phys. J. C* 72: 1896. https://doi.org/10.1140/epjc/s10052-012-1896-2.

Chen, Yi, Anthony Badea, Austin Baty, et al. 2022. "Jet Energy Spectrum and Substructure in e+e- Collisions at 91.2 GeV with ALEPH Archived Data." *JHEP* 06: 008. https://doi.org/10.1007/JHEP06(2022)008.

CMS Collaboration. 2024. "Measurement of the Primary Lund Jet Plane Density in Proton-Proton Collisions at Sqrt(s) = 13 TeV." *JHEP* 05: 116. https://doi.org/10.1007/JHEP05(2024)116.

D'Agostini, G. 1995. "A Multidimensional Unfolding Method Based on Bayes' Theorem." *Nucl. Instrum. Meth. A* 362: 487–98. https://doi.org/10.1016/0168-9002(95)00274-X.

Dokshitzer, Yuri L., G. D. Leder, S. Moretti, and B. R. Webber. 1997. "Better Jet Clustering Algorithms." *JHEP* 08: 001. https://doi.org/10.1088/1126-6708/1997/08/001.

Dreyer, Frederic A., Gavin P. Salam, and Gregory Soyez. 2018. "The Lund Jet Plane." *JHEP* 12: 064. https://doi.org/10.1007/JHEP12(2018)064.

Dreyer, Frederic A., Gregory Soyez, and Adam Takacs. 2022. "Quarks and Gluons in the Lund Plane." *JHEP* 08: 177. https://doi.org/10.1007/JHEP08(2022)177.

Havener, Laura, and ALICE Collaboration. 2022. "Measurement of the Primary Lund Jet Plane Density in Pp Collisions at 13 TeV with ALICE." *Proceedings of EPS-HEP2021* EPS-HEP2021: 364. https://doi.org/10.22323/1.398.0364.

Lifson, Andrew, Gavin P. Salam, and Gregory Soyez. 2020. "Calculating the Primary Lund Jet Plane Density." *JHEP* 10: 170. https://doi.org/10.1007/JHEP10(2020)170.

Particle Data Group. 2024a. "Review of Particle Physics (Quantum Chromodynamics Review)." *Phys. Rev. D* 110: 030001. https://doi.org/10.1103/PhysRevD.110.030001.

Particle Data Group. 2024b. "Review of Particle Physics (z Boson Properties)." *Phys. Rev. D* 110: 030001. https://doi.org/10.1103/PhysRevD.110.030001.

Skands, Peter, Stefano Carrazza, and Juan Rojo. 2014. "Tuning PYTHIA 8.1: The Monash 2013 Tune." *Eur. Phys. J. C* 74 (8): 3024. https://doi.org/10.1140/epjc/s10052-014-3024-y.

Tournefier, E., and ALEPH Collaboration. 1999. *Electroweak Results from the z Resonance Cross-Sections and Leptonic Forward-Backward Asymmetries with the ALEPH Detector*. https://arxiv.org/abs/hep-ex/9904007.

Wobisch, M., and T. Wengler. 1999. "Hadronization Corrections to Jet Cross-Sections in Deep Inelastic Scattering." *Monte Carlo Generators for HERA Physics (Hamburg 1998/1999)*, 270–79. https://arxiv.org/abs/hep-ph/9907280.