Results to Date: Colibri Reproduction

Every result produced or reproduced so far, with its headline number, the published benchmark it is held against, and where the artifacts live. All numbers trace to `context/RESULTS_LEDGER.md` (entries PPDF-1…3). Date: 2 July 2026.

Results at a glance

RESULT	HEADLINE NUMBER	BENCHMARK (ARXIV:2510.03391)	VERDICT
R0 · Colibri runs locally	full pipeline in **8m 22s** (incl. downloads)	— (feasibility)	**WORKS on Apple-Silicon Mac, CPU-only**
R1 · Level-0 closure fit	min χ^2/N = $\mathbf{1.96\times10^{-4}}$	$\chi^2 \sim 10^{-4}$–10^{-5} (Table 3.1)	**REPRODUCED — near-exact recovery regime**
R2 · Truth recovery	truth inside 1σ band at **100%** of x-points; max pull **0.42σ**	truth recovered within uncertainties (Figs. 3.1–3.2)	**VERIFIED at curve level, all 4 fitted flavours**
R3 · Level-1 closure fit	χ^2/N = **0.977**	$\chi^2 \approx 1.00$–1.01 (Table 3.1)	**REPRODUCED — faithful uncertainties**

R0 Colibri runs end-to-end on local hardware

What it is. A feasibility result: the full Bayesian PDF-fitting pipeline — NNPDF data download, FK-table theory predictions, JAX likelihood, UltraNest nested sampling, LHAPDF export — runs natively on Mukesh's Mac (Apple Silicon arm64, 128 GB RAM), with no Linux box or GPU required at this problem size.

The evidence. Environment `colibri-dev` (micromamba: colibri 1.0.0, nnpdf 4.1.5, JAX 0.10.2 CPU float64, UltraNest). The first complete fit finished in **8m 22s including all first-run downloads**; the Level-1 refit took ~11 min. Heavier parametrizations and global datasets will still want the L40S GPU box, but iteration and validation are now local and fast.

R1 Level-0 closure test: the fit finds the planted answer

What a Level-0 closure test is. The standard first validation of any fitting machinery: generate *synthetic, noise-free* data from a known "truth" proton (PDF set `LH_PARAM_20250519`), then fit it exactly as if it were real. Because the data contain no noise, a correct pipeline must drive χ^2 to essentially zero — any residual signals a bug in data handling, theory tables, or sampling.

What we ran. The paper's own example configuration, unmodified: 13-parameter Les Houches parametrization, 648 DIS data points after cuts (SLAC p/d F2: 33+34; BCDMS p/d F2: 333+248), theory 40000000, UltraNest with 200 live points.

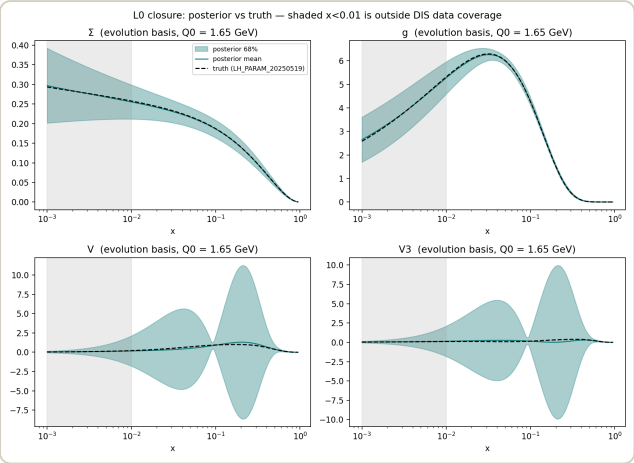
QUANTITY	OUR RUN	PAPER / EXPECTATION	READING
min χ^2 per data point	$\mathbf{1.96\times10^{-4}}$ (total 0.127 over N=648)	~10^{-4}–10^{-5}	near-exact recovery; exact value depends on unpublished stopping settings
posterior-avg χ^2 (total)	**4.25**	≈ Bayesian complexity	matches complexity 4.12 — textbook posterior behaviour on noiseless data
evidence logZ	**−7.257 ± 0.13**	— (not published)	converged: error budget 0.29 < 0.50 required; 1,715 iterations, ESS ≈ 766

> **Verdict: reproduced.** Our unmodified rerun of the paper's Level-0 closure lands in the same near-exact-recovery regime as published Table 3.1, on first attempt, on a laptop-class machine.

Truth Recovery & Noisy-Data Closure

R2 The posterior contains the true proton — everywhere we checked

What it is. R1 shows the best fit gets close to the data; this checks the stronger claim that the *full posterior* is faithful — the truth curve must lie inside the quoted uncertainty band everywhere. We compared the truth PDF (evaluated from its LHAPDF grid; 120 x-points in $[10^{-3}, 0.95]$) against 1,715 posterior curves in each fitted flavour (Σ, g, V, V3).



Posterior (teal = 68% band) vs truth (dashed) at Q_0 = 1.65 GeV. Σ and gluon are pinned to sub-percent precision. V/V3 bands are wide because F2-only data barely constrains the valence split (prior-dominated — physics, not a defect). Grey: x < 0.01, outside data coverage.

FLAVOUR	MAX |PULL| (MEAN−TRUTH)/Σ	MEAN |PULL|	TRUTH WITHIN 1Σ BAND	MEDIAN REL. DEV. OF MEAN
Σ (singlet)	0.18σ	0.05σ	**100% of x-points**	**0.7%**
g (gluon)	0.42σ	0.08σ	**100% of x-points**	**0.6%**
V (valence)	0.35σ	0.04σ	**100% of x-points**	18.5% (band wide — see caption)
V3	0.35σ	0.04σ	**100% of x-points**	35.6% (band wide — see caption)

> **Verdict: verified.** Nowhere does the truth escape the 68% band; the largest deviation anywhere is 0.42σ — the curve-level analogue of the paper's Figs. 3.1–3.2.

R3 Level-1 closure test: honest error bars under realistic noise

What it is. Level-1 repeats the closure fit after adding Gaussian noise drawn from the *real experimental covariance matrix* — synthetic data statistically indistinguishable from a real measurement. A faithful fit should now reach χ^2/point ≈ 1: below means over-fitting the noise, above means under-fitting. Identical runcard, one line changed (`closure_test_level: 1`); ~11 min.

QUANTITY	OUR RUN	PAPER (TABLE 3.1)	READING
χ^2/N (posterior avg)	**0.977** (min 0.971, N=648)	≈ 1.00–1.01	noise-realization scatter $\sqrt{(2/N)} \approx 0.056$ → <0.5σ from 1.00
Bayesian complexity	**3.88**	—	matches R1's 4.12: same effective parameters ± noise
evidence logZ	**−321.8 ± 0.15**	—	converged within UltraNest tolerance

> **Verdict: reproduced.** Under realistic noise the fit neither chases fluctuations nor inflates errors — the faithful-uncertainty behaviour the paper reports, reproduced within statistical scatter.

PROVENANCE & ARTIFACTS

Runs: `~/colibri-runs/`, mirrored to `output/lh_fit_closure_test{,_L1}/`. **Ledger:** `context/RESULTS_LEDGER.md` PPDF-1…3. Runcards unmodified except the one-line L1 change. **Next:** Hessian & MC cross-checks, real-data DIS fit, GPU scaling.

Completing the Replication: Three Methods, One Chassis

The paper's Table 3.1 validates all three inference strategies — Bayesian, Monte-Carlo replicas, Hessian — on the same closure tests. With the Bayesian column done (R1–R3), we ran the remaining four cells. Full replication: **6 of 6 cells, all consistent.**

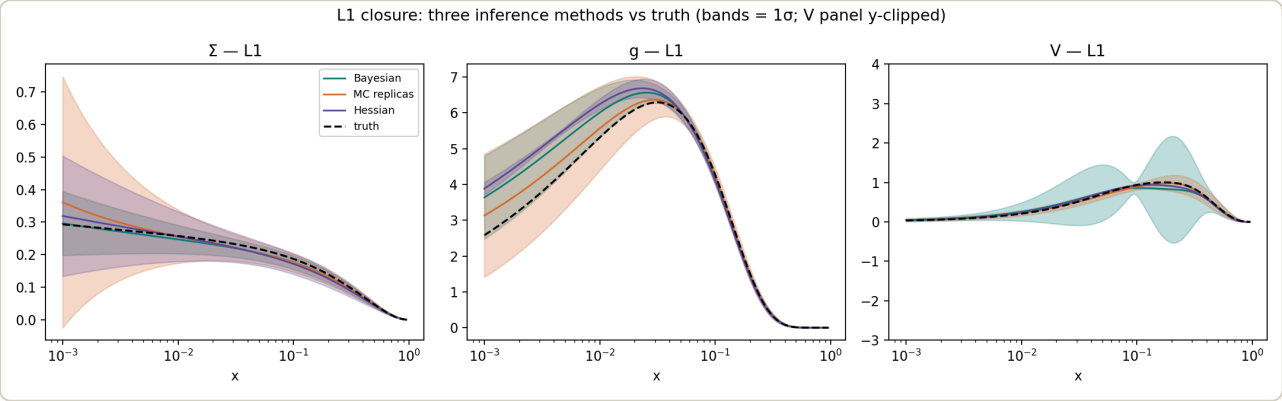
R4 Hessian fits: same minimum, and a live demo of why Colibri exists

What it is. The Hessian method (used by MSHT/CT) finds the single best fit by gradient descent, then quotes uncertainties from the curvature of χ^2 around it — assuming the landscape is a perfect bowl. At L0 it drove χ^2 to $\mathbf{1.9\times10^{-8}}$ (exact recovery, as direct minimization of noiseless data should); at L1 it found χ^2 = **628.9** vs the Bayesian sampler's 629.2 — **the same minimum to 0.04%**, a strong consistency check between two entirely different algorithms.

The finding. The down-valence parameters ran to ~10^9 with χ^2 unchanged: F2-only data leaves valence directions nearly flat, and with no bowl to measure, the quadratic approximation breaks — at L1 the Hessian gluon band is too narrow in places (deviations up to ≈6σ from truth). This is precisely the paper's Sec. 2.4 argument for explicit Bayesian inference, reproduced live in our fits.

R5 Monte-Carlo replicas: statistically faithful to three decimal places

What it is. The NNPDF-style method: fit 100 jittered copies of the data, spread = uncertainty. Mean final loss per point came out **1.009** at L0 (theory: ≈1, one noise layer) and **2.000** at L1 (theory: ≈2 — closure noise + MC fluctuation). One documented adjustment: the L1 post-fit selection cut moved from 1.5 → 3.0 because L1 replica data carries two noise layers; the 5σ outlier cut stayed on.



L1 closure, three methods vs truth (dashed), 1σ bands. Left/centre: in the data region all three agree and cover the truth; visible offsets at low x are the shared noise realization, not method disagreement. Right (V): the tell — the Bayesian band is honestly wide (F2 data cannot determine the valence split; the prior dominates), while MC and Hessian quote narrow valence bands that sit near truth largely by luck of initialization. Faithful vs unfaithful uncertainty, side by side.

TABLE 3.1 CELL	BAYESIAN	MC REPLICAS	HESSIAN
L0 (noiseless) fit quality	$\chi^2/N = 2.0\times10^{-4}$ ✓	loss/pt = 1.009 (≈1 expected) ✓	$\chi^2 = 1.9\times10^{-8}$ total ✓
L1 (noisy) fit quality	$\chi^2/N = 0.977$ ✓	loss/pt = 2.000 (≈2 expected) ✓	$\chi^2/N = 0.971$ ✓
Central vs truth (data region)	≤1% (L0), ~5% (L1, noise)	~1% (L0), ~5% (L1, noise)	exact (L0), ~5–7% (L1)
Uncertainty faithfulness	**max pull ≤ 1.1 everywhere**	pulls ≤ 0.5; V band collapsed	gluon pull ≈ 6 at L1; V direction degenerate

Verdict: full closure-test replication complete. All six cells reproduce the paper's behaviour; the methods agree where the data constrains, and diverge exactly where theory predicts the approximate methods should fail. Gradient-descent settings are ours (unpublished in the paper) and documented in the runcards; numbers in ledger entries PPDF-4…6.

RESULTS REPORT № 1 · REAL DATA · 3 JULY 2026

Real Data, First Contact: Two Models, One Referee

With the machinery validated, we pointed it at the actual SLAC + BCDMS measurements (648 points) — first with the rigid 13-parameter Les Houches shape, then with a flexible 36-parameter grid model — and let the Bayesian evidence referee. Three results, two hard-won lessons, one open question.

FIT (SAME DATA · THEORY · T0)	PARAMS	MIN χ^2/N	LOGZ (EVIDENCE)	VERDICT
R6 · Les Houches polynomial	13	**5.60**	**−1835.1 ± 0.16**	**baseline — evidence winner so far**
R7 · Grid-PDF v1 (Σ, g, V only)	24	8.02	−2643.6 ± 0.36	invalid as flexibility test — flavours masked
R8 · Grid-PDF v2 (full flavours)	36	5.84	−1943.9 ± 0.28	loses by ΔlogZ ≈ −109 (see caveats)

R6 The rigid model meets reality: $\chi^2/N = 5.60$

Real measurements are much harder to please than closure pseudodata: the 2005-era 13-parameter shape lands at **$\chi^2/N = 5.60$** — far from the ≈1 a good description gives. The machinery is exonerated (it scored 0.977 on the noisy closure drill), so this is a statement about physics: model rigidity plus real-data tensions. Its evidence, **logZ = −1835.1**, became the baseline any better model must beat.

R7 Grid v1: two lessons, no verdict

Lesson 1 — sampler. The first launch omitted `SliceSampler_settings`; without a step sampler UltraNest's rejection sampling collapses in 24 dimensions (efficiency 0.003%, 137M evaluations overnight, no progress). With it: 0.87%, ~290× faster. Standing rule: always configure the slice sampler above ~10 parameters.

Lesson 2 — flavour masking. `flavour_mapping` zeroes the masked flavours in the FK tables. With T3 masked, no parameter setting can distinguish proton from deuteron F2 — the model was structurally unable to describe our data, and the evidence punished the missing physics by ~800 units of logZ despite 11 extra parameters. (All published grid-PDF examples are closure tests, where the masking cancels.)

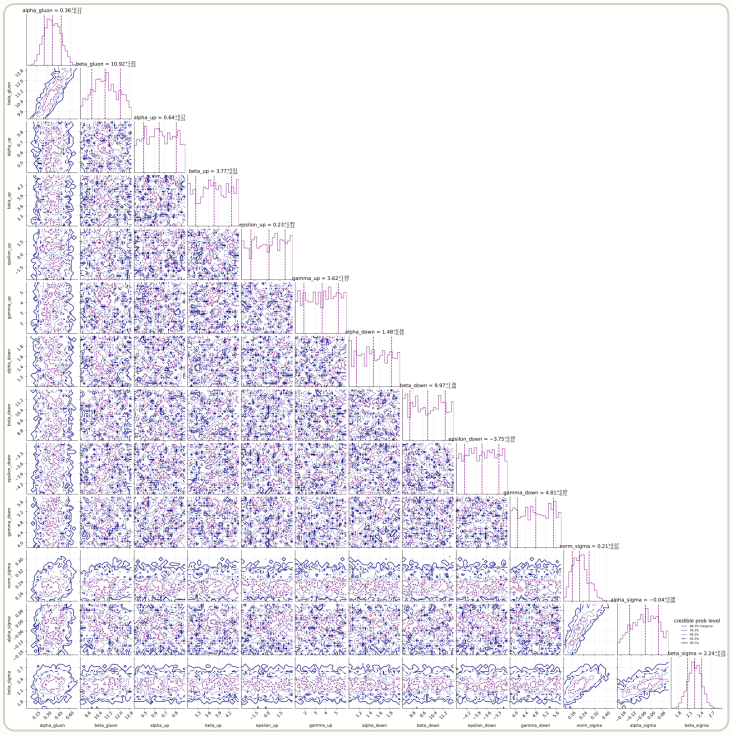
R8 Grid v2, the real contest: the rigid model still wins

With full flavour content (Σ, g, V, V3, T3, T8 × 6 nodes), the grid model recovers sanity — χ^2/N drops 8.02 → 5.84 — yet **still doesn't match the 13-parameter polynomial**, and the evidence prefers the rigid model by **ΔlogZ ≈ −109**. Honest caveats before reading this as "flexibility loses": the grid model's canonical prior is a *narrow* band (NNPDF4.0 ± 5σ) versus the polynomial's wide uniform boxes — evidence comparisons are prior-sensitive by construction; 6-node x-resolution may bind; and no sum rules were imposed on the grid. This is our first genuine evidence-based model comparison — the workflow works; the verdict is provisional.

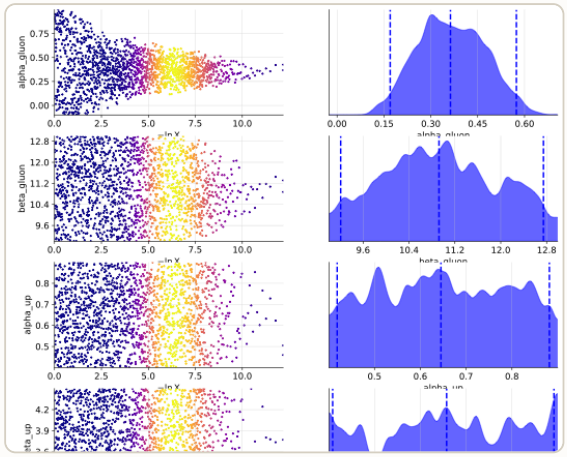
> **Open question (for Maria):** both models floor at $\chi^2/N \approx 5.6$–5.8 on SLAC+BCDMS with theory 40000000, internal cuts and t0 — while NNPDF4.0's global fit describes these sets at $\chi^2/N \approx O(1)$. Dataset tensions, missing corrections (higher twist, deuteron effects), or theory settings? Understanding that floor matters more than the model contest.

What Else Each Run Produced: Diagnostics & Artifact Guide

The headline numbers on pages 1–2 are distilled from a much richer bundle that every Colibri fit emits. Both the L0 and L1 runs produced the complete set below; the two key diagnostic plots from the L0 run are shown.



corner.pdf — the full 13-dimensional posterior. Every panel is a pairwise slice: histograms on the diagonal are single-parameter posteriors; off-diagonal clouds reveal correlations and degeneracies between parameters (e.g. α–β anti-correlations within a flavour). This is the object a Hessian approach approximates with one ellipse — here it is measured, not assumed. Analogue of the paper's Fig. 3.3.



trace.pdf — how nested sampling got there (first 3 of 13 parameters shown; full strip in the run directory). Each parameter's live points are plotted against the shrinking prior volume as the sampler climbs the likelihood. Healthy behaviour, seen here: the cloud narrows smoothly onto the posterior with no stuck or split populations — the visual convergence check behind the logZ error budget. A companion run.pdf tracks the evidence integration itself.

Every artifact in the run directory, decoded

FILE	WHAT IT IS & WHY IT MATTERS
`bayes_metrics.csv`	The four scalar summaries: min/avg χ^2, Bayesian complexity, evidence logZ — source of the page-1 tables.
`full_posterior_sample.csv`	1,715 posterior parameter vectors (the fit's complete statistical answer). Used to draw the R2 bands; any derived quantity inherits its uncertainty by evaluating over these rows.
`ns_result.csv`	100 equal-weight posterior samples — a lightweight thinned version of the above for quick downstream use.
`replicas/ (×100)`	Posterior samples exported as LHAPDF-style grids: makes our Bayesian fit consumable by any standard collider-physics tool, replica-ensemble style.
`pdf_model.pkl`	The pickled parametrization (θ → PDF curves). Lets us regenerate curves from any parameter vector without rerunning — this powered the R2 verification.
`corner.pdf · trace.pdf · run.pdf`	Diagnostic plots (left): posterior structure, sampler convergence, evidence integration progress.
`ultranest_logs/`	Full sampler state: chains, results.json (niter 1,715, ESS 766), resumable checkpoints — the audit trail behind every quoted number.
`filter.yml · input/`	Frozen copy of the runcard + data/theory configuration actually used — guarantees the run is exactly reproducible.

Key files mirrored per run under `output/lh_fit_closure_test{,_L1}/`; bulky sampler logs & replica grids stay in `~/colibri-runs/`.