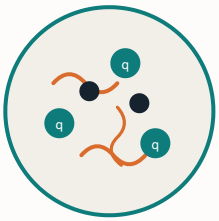
What's Inside the Proton?

Why the world's biggest physics experiments depend on a map nobody can calculate, how that map is drawn today — and how we're going to help draw it better, using Bayesian statistics.

The Large Hadron Collider smashes protons together hundreds of millions of times per second, hunting for new laws of nature. But here is the awkward secret: a proton is not a simple ball. It is a seething bag of quarks, antiquarks and gluons — "partons" — and when two protons collide, what actually hits is one constituent from each bag. To predict *anything* at the LHC, you first have to know how the proton's momentum is shared out among its contents.



the proton: a bag of partons

Quarks (teal), sea quarks (dark) and gluons (orange) each carry a slice of the proton's momentum; PDFs describe that sharing.

§1 The ingredient list nobody can compute

That momentum-sharing map is called a set of **Parton Distribution Functions (PDFs)**. For each type of particle inside the proton, the PDF tells you how likely you are to find it carrying a given fraction x of the proton's momentum. It is the LHC's ingredient list: every prediction — Higgs production rates, searches for new particles, precision tests of the Standard Model — is built on top of it.

The catch: the theory of the strong force (QCD) becomes so violently non-linear inside the proton that **nobody can calculate the PDFs from first principles**. The map cannot be computed. It has to be *measured* — or more precisely, reverse-engineered from collision data. Inferring what's inside the bag from the debris that comes out is a classic **inverse problem**, like reconstructing a piñata's contents from where the candy lands.

§2 A blurry X-ray, stitched from 4,600 snapshots

No single experiment can photograph the inside of a proton. Each measurement — an electron bouncing off a proton at HERA, a Z boson produced at LHCb, top-quark pairs at ATLAS — is one blurry, partial glimpse from one angle, constraining one corner of the map. A **global fit** is the algorithm that stitches thousands of these glimpses, collected over four decades by dozens of experiments, into a single consistent picture of the proton.

The state of the art is the **NNPDF collaboration**'s fit, "NNPDF4.0", which combines roughly 4,600 data points from 80+ datasets and reaches percent-level precision in the best-measured regions. Remarkably, the entire machinery — every dataset, every line of fitting code — is **public and open-source**. That open infrastructure is the foundation this project builds on.

§3 How the map is drawn today

Two ideas power the modern approach. First, since nobody knows what shape the PDFs should take, NNPDF lets a **neural network be the curve** — a flexible ruler with no preconceived shape, constrained by hard physics rules (momentum must add up; probabilities can't go negative). Second, to estimate *uncertainty*, they fit not once but **a thousand times**, each time to a copy of the data jittered within the experimental error bars. The spread of those 1,000 curves is quoted as the uncertainty on the proton's structure.

This "replica" method is clever, battle-tested — and, as page 2 explains, it rests on an assumption that deserves a hard look now that precision is the whole game.

THE GLOBAL FIT, IN NUMBERS

~4,600
experimental data points in NNPDF4.0, the state-of-the-art fit

80+
datasets, from 1980s fixed-target experiments to the LHC

~1%
precision reached on proton structure in the best-measured regions

1,000
separate neural-net fits run just to estimate the uncertainty

JARGON, DECODED

Parton
Anything inside the proton: quarks, antiquarks, gluons.

PDF f(x)
Probability that a parton carries fraction x of the proton's momentum.

Global fit
One simultaneous fit of all PDFs to all experiments.

Replica
A jittered fake copy of the data, used to estimate uncertainty.

THE OPPORTUNITY · AND OUR PLAN

Error Bars You Can Trust: the Bayesian Upgrade

§4 The catch in today's error bars

Both mainstream ways of quoting PDF uncertainty lean on approximations. The **Hessian** method (used by other groups) assumes the fit's quality landscape is a perfect bowl around the best answer — fine near the bottom, misleading if the valley is banana-shaped or has side valleys. The **replica** method is provably equivalent to the true statistical answer **only when the model is linear** — and a neural network is anything but. Recent work by Maria's group showed that for flexible models, jittered-photocopy fits can quietly produce biased error bars.

> *When the whole field is chasing 1% precision, the error bar is the product. An error bar you can't fully trust is a problem worth solving.*

§5 Enter Colibri: the full landscape, not a shortcut

Colibri (Spanish for *hummingbird*) is a new open-source framework from the **HEP-PBSP group at Cambridge** — Maria's team. Instead of approximating, it computes the **exact Bayesian posterior**: the complete landscape of "candidate protons", each weighted by how well it explains the data and by what we assumed going in (the prior). A technique called **nested sampling** surveys this landscape systematically, contour by contour, with no bowl-shaped assumptions and no photocopies.

Two bonuses come free. First, nested sampling also measures the landscape's total volume — the **Bayesian evidence** — which lets you *fairly compare* different models of the proton, not just fit one. Second, Colibri is a **common chassis**: you can plug in any parametrization (classic polynomials, grids, neural nets) while the data, theory engine and statistics stay fixed — so differences you see are physics, not plumbing. It runs on JAX with GPU acceleration, and sits directly on NNPDF's public data and theory tables. In its validation tests, all methods recover a known "truth" proton with textbook-perfect statistics ($\chi^2 \approx 1.00$), confirming the machinery works.

§6 What we are trying to do

1. **Master the machinery.** Install Colibri, reproduce the paper's validation fits (closure tests on synthetic data), and verify we get the published statistics before touching anything real. No result counts until it's reproduced.
2. **Run a real Bayesian PDF fit.** Fit actual experimental data (starting with deep-inelastic scattering) through the full pipeline — NNPDF data in, posterior out — on our own GPU infrastructure.
3. **Exploit what Bayesian buys us.** Use the evidence to compare proton parametrizations head-to-head, study how prior assumptions shape the answer, and explore extensions worth discussing with Maria's group — the frontier where this project can contribute.

The near-term milestone is deliberately modest and verifiable: **a reproduced closure test, then a DIS-only Bayesian fit** whose posterior we can hold up against the published NNPDF results — the first rung on a ladder toward contributing to how the field draws its map of the proton.

THREE WAYS TO FIT A PROTON

	ASSUMES	YOU GET
Hessian	landscape is a perfect bowl	one best fit ± symmetric errors
Replicas	model behaves linearly	1,000 fits; spread = error
Bayesian	only your stated prior	full posterior + evidence for model comparison

THE TOOLBOX

Colibri
Bayesian PDF-fitting framework (HEP-PBSP, Cambridge).

UltraNest
Nested-sampling engine; maps the posterior, computes the evidence.

NNPDF data
Open archive of all usable measurements, full error correlations.

FK tables
Precomputed theory grids: candidate PDF → predictions in milliseconds.

JAX + GPU
Fast math engine; our L40S GPU box does the heavy sampling.

READ MORE

Colibri — arXiv:2510.03391, "*A new tool for fast-flying PDF fits*"
NNPDF4.0 — arXiv:2109.02653, "*The path to proton structure at 1% accuracy*"
Code — github.com/HEP-PBSP/colibri · github.com/NNPDF/nnpdf

PART 2 · 2 JULY 2026 · WHAT WE ACTUALLY DID

First Flight: Testing on a Planted Proton

§7 First, get it running

Everything on pages 1–2 was theory until today. We installed the full Colibri + NNPDF pipeline on the Mac and ran a complete fit end-to-end: it downloaded four decades of experimental measurements and the pre-computed theory tables, then sampled the full statistical answer — in **8 minutes 22 seconds, laptop-class hardware, no super-computer**. Practice laps are now free and fast.

§8 The fire drill: fitting a proton we invented

You don't point a brand-new telescope at an undiscovered galaxy first — you point it at something you already know. The equivalent here is a **closure test**: we planted a fake proton (a curve we know exactly), simulated the collision data it would produce at real experiments, and asked Colibri to find it back. Two drills, with opposite pass criteria:

Drill 1 — clean data. Synthetic measurements with zero noise. A correct machine must recover the planted proton essentially exactly. Ours did: mismatch of **0.0002 per data point** — the same near-zero regime the Colibri paper reports.

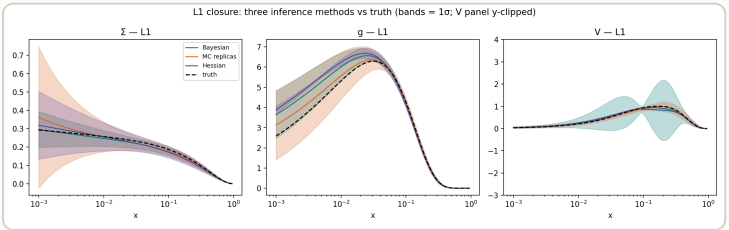
Drill 2 — realistic data. Same, but with noise injected exactly as real experiments fluctuate. Now perfection is *wrong*: the fit should match the data about as well as noise allows and no better — a score of ≈1 per point. Below 1 means it's fitting noise (hallucinating structure); above means it's missing signal. Ours: **0.977** — squarely in the honest zone, matching the paper.

§9 Did the band contain the truth? Everywhere.

A best-fit curve isn't enough — the whole point of Colibri is the *uncertainty band*. So we checked: at 120 momentum-fraction checkpoints in every fitted quark/gluon combination, does the planted truth stay inside the band? **Yes — 100% of checkpoints; worst excursion 0.42σ.** One honest wrinkle: for the valence combinations our four datasets genuinely carry little information, and the band there is huge. That's the machine saying "I don't know" — which, as the next section shows, is more than the older methods manage.

§10 The three-engine race

Colibri's party trick is running **three statistical engines on the identical problem**: the exact Bayesian sampler, the NNPDF-style 100-replica method, and the bowl-approximation (Hessian) method. Where the data constrains the proton, all three agree — the Bayesian sampler and the Hessian minimizer, two totally different algorithms, landed on the **same optimum to 0.04%**. Where the data is silent, they split: the Bayesian band is honestly wide; the other two quote confident narrow bands that are right largely by luck (the Hessian's internals even ran off to ~a billion along those directions). We didn't just read the paper's argument; **we watched the competitors fail in our own runs.**



Three engines vs the planted truth (dashed). Right panel: only the Bayesian band admits ignorance.

THE DAY IN NUMBERS

8m 22s
first complete fit, downloads included, on the Mac

648
real measurement points in the practice dataset (SLAC + BCDMS)

200+
individual fits run today (2 Bayesian, 200 replicas, 6 Hessian restarts)

6 / 6
cells of the paper's validation table reproduced

JARGON, DECODED

Closure test
Fit fake data made from a known "truth" — if you can't recover what you planted, fix the machine.

Level 0 / Level 1
Clean vs realistically noisy versions of that fake data.

Pull
A deviation measured in units of the quoted error bar; ≲1 is healthy.

Flat direction
A parameter combination the data doesn't constrain at all — where approximate methods quietly fail.

What We Have Now — and Why It Matters

§11 What we produced

A working proton-fitting laboratory on this Mac — the same open-source stack Maria's group and NNPDF use, installed, exercised, and fast enough to iterate on daily. **A validated machine**: the Colibri paper's entire closure-test table reproduced, six cells out of six, plus a truth-recovery study of our own design that goes a step beyond the paper's published checks. **A paper trail**: a four-page technical results report, a results ledger where every number links to the script and run that produced it, and the whole project now versioned on GitHub — nothing lives in anyone's head.

§12 Why it matters

Trust. The iron rule of measurement: never point an untested instrument at the real sky. Every claim we make from now on stands on today's planted-proton drills — we've earned the right to fit real data.

Understanding. The field is entering an era where the *error bar is the product*: LHC measurements are now so precise that a mis-stated proton uncertainty can fake or bury a discovery. Today we saw first-hand — not in a paper, in our own runs — where the standard error bars can quietly go wrong and how the Bayesian approach fixes it. That's the core intuition this whole project is built on.

Standing. We now operate Maria's group's stack end-to-end. That changes the conversations we can have: from "please explain" to "we ran it, here's what we found, and here's what we'd like to try" — with the evidence-based comparison of rival proton models as our most promising angle to contribute something new.

§13 What's next

1. **Fit real data — already running.** The same setup, pointed at the actual SLAC and BCDMS measurements instead of planted ones: our first genuine Bayesian proton fit, launched tonight on the Mac.
2. **Scale up.** Move to the GPU box for richer proton models (more parameters, more data) where a laptop stops being enough.
3. **Do new science.** Use the Bayesian evidence to compare competing proton parametrizations head-to-head — the capability Colibri uniquely offers, and our natural contribution lane with Maria's group.

> *The machine works, the error bars are honest, and the next fit is real data.*

THE SHELF — WHERE EVERYTHING LIVES

This guide
output/tldr_proton_pdf.pdf

Technical report
output/results_report.pdf — every result with numbers and provenance

Fit outputs
output/lh_fit_closure_test/ (+ _L1/) — posteriors, plots, exported PDF sets

Method shoot-out
output/crosscheck_methods/ — three-engine comparison

Ledger & scripts
context/RESULTS_LEDGER.md · scripts/

Repository
github.com/mukesh-bansal/53-proton-pdf (private)

SCORECARD

Clean drill
0.0002 per point — near-exact recovery ✓

Noisy drill
0.977 ≈ 1 — honest, no over-fitting ✓

Truth in band
100% of checkpoints, worst 0.42σ ✓

Engine agreement
same optimum to 0.04% ✓

Bonus finding
flat directions expose the bowl approximation ✓

PART 3 · 3 JULY 2026 · FIRST CONTACT WITH REAL DATA

The Data Meets the Models — and Referees Them

§14 Real data bites

We pointed the validated machine at the **actual SLAC and BCDMS measurements** — no more planted protons. The classic 13-parameter shape scored **5.6 per data point**, nowhere near the ≈1 of a good description. That's a finding, not a failure: yesterday's drills prove the machinery is sound, so the misfit is physics — a 2005-era formula is too stiff for 648 real measurements, and the data itself may carry tensions no smooth proton can resolve. Its evidence score, **logZ = –1835**, became the number to beat.

§15 Two stumbles, two lessons

The stalled climber. Our first flexible-model run sat overnight making almost no progress. Cause: we'd left out the sampler's "climbing gear" (its slice-stepping mode), so in 24 dimensions it was randomly throwing darts — 137 million throws, 0.003% kept. One config block later it ran **290× faster**. Lesson filed: always pack the climbing gear above ~10 parameters.

The blind taster. That first flexible model was also missing an ingredient: the quark combination that makes a proton differ from a deuteron (a proton–neutron pair). Since our data is precisely proton *and* deuteron measurements, the model was structurally blind to the thing being measured — and the referee noticed instantly, docking it **~800 evidence points** despite its extra parameters. **The evidence punishes missing physics, not just complexity.** Existing examples never hit this because they only ever fit simulated data generated from the same blinkered model.

§16 The referee's first real verdict

Fixed and refitted: a **36-parameter flexible grid** (no assumed shape at all) versus the **13-parameter classic**, same data, same theory. Result: the flexible model improved hugely (8.0 → 5.8 per point) but **still lost — the evidence prefers the simple model by ~109**. One big caveat before crowning a winner: the flexible model competed in a straitjacket — its standard prior confines it to a narrow band around NNPDF's published proton, while the classic model roamed freely. Evidence verdicts always depend on the prior; loosening it fairly is a design question worth taking to Maria. What's not provisional: **the model-comparison workflow works end-to-end** — exactly the capability this project exists to exercise.

THE REAL-DATA SCOREBOARD

5.60
misfit per point — classic 13-param shape (logZ –1835, leader)

5.84
flexible 36-param grid, full ingredients (logZ –1944)

8.02
flexible grid, missing ingredient (logZ –2644, disqualified)

290×
speed-up from one sampler config line

JARGON, DECODED

Evidence (logZ)
One number scoring a whole model: fit quality, honestly penalized for wasted parameter space. Higher wins.

Prior sensitivity
Evidence verdicts depend on what each model was allowed to try. Fair contests need fair priors.

Flavour content
Which quark/gluon combinations a model can vary. Missing one = structural blindness.

QUESTIONS FOR MARIA

The 5.6 floor
Why do both models floor at ~5.6/point on SLAC+BCDMS alone, when the global fit gets ≈1? Cuts, deuteron corrections, theory settings?

Fair priors
What's the recommended prior for grid models on real data?

Sum rules
Should the grid model carry momentum/valence sum rules in this setup?