

How well do we really know what is inside the proton?

A proton is not a single solid object — it is made of smaller particles, and the way they share the proton's momentum has to be **measured from experiments, not calculated from theory**. The result is a set of curves called **parton distribution functions**. Every prediction made at the Large Hadron Collider depends on them, and on the **error bars** attached to them.

WHAT THIS PROJECT IS ABOUT

Those error bars are not measured — they are **computed**, and there are **three standard ways** to compute them. The field assumes the three broadly agree. Maria asked me to check. **They do not.**

1 · THE CONTEXT

Who produces these proton maps, why Maria's group built a new tool, and what she asked me to do.

2 · THE PHYSICS

What is actually being measured, and where the error bar comes from.

3 · WHAT I RAN

Four weeks, 28 experiments — validating the tools and measuring the underlying problem.

4 · WHAT I FOUND

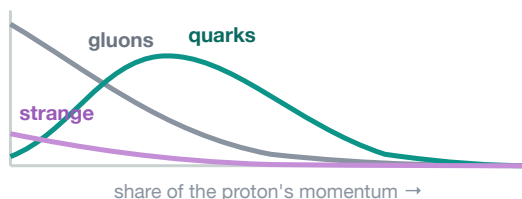
The three methods disagree sharply. One of them is right — and I can show which.

What I understand this programme to be

Proton structure, and how honestly we can state its uncertainty. Six points, in order:

1 What we are mapping

A proton is not one solid particle. It is a swarm of **quarks and gluons** sharing its momentum. A **parton distribution function** is simply the curve saying how likely each one is to carry a given share.



one curve per particle type · slide 7

2 Why anyone needs them

These curves **cannot be calculated** — the maths is intractable. They must be **fitted to thousands of experiments**. Every prediction at the Large Hadron Collider then uses them as an input.

PDF × collision = prediction · slide 8

3 Who actually makes them

Four collaborations worldwide. **NNPDF** (“Neural Network PDF”, Europe), **MSHT** (Martin–Stirling–Thorne–Watt, UK), **CT** (CTEQ–TEA, US), **JAM** (Jefferson Lab, US). Everyone else downloads their results.

I use the **MSHT** proton model · slide 4

4 The problem: the $\pm$ is a choice

Each fit must also state **how uncertain the curve is**. That number is not measured — it is computed, and there are **three accepted recipes**. Each group builds one recipe into its own software.

Hessian · Monte-Carlo · Bayesian · slide 10

5 What Maria's group built

Her Cambridge programme asks whether apparent “new physics” is really a **mis-modelled proton**. Her group released **Colibri** — the first program that can run **all three recipes on one identical fit**.

Colibri, 2025 · slide 5

6 What I am doing, and why it is hard

I put the real MSHT proton model into Colibri and ran all three recipes on the same fit. It is hard because the model has **more freedom than the data can pin down** — so the recipes are free to disagree.

we measure them differing by up to **144×** · slide 16

The curves are fitted to data, not calculated — by four groups worldwide

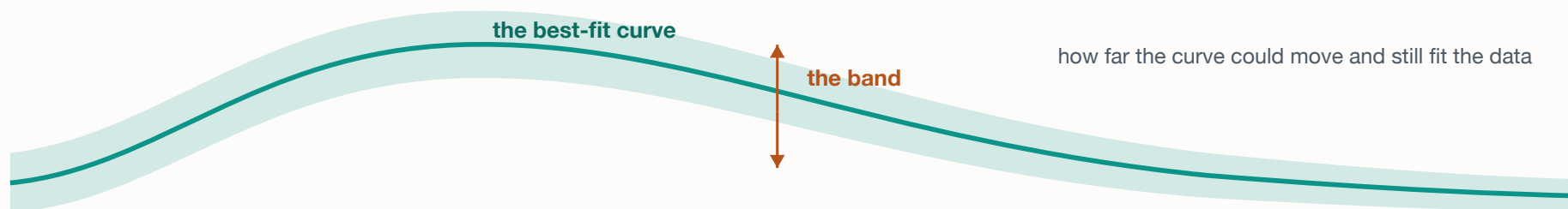
There is no equation that gives these curves from first principles: the maths of the strong force cannot be solved that way. So each curve is **fitted** — adjusted until it matches thousands of past collision measurements. Only a handful of groups do this. Everyone else uses their results.

GROUP	WHAT THE LETTERS STAND FOR	WHERE	HOW THEY DESCRIBE THE PROTON	WHAT THEY PUBLISH
NNPDF	Neural Network PDF	Europe	A neural network — a flexible computer model with no fixed shape	NNPDF4.0 — a central curve + ~100–1000 replicas
MSHT	Martin, Stirling, Thorne, Watt — four physicists' surnames	United Kingdom	A fixed algebraic formula the model I use	MSHT20 — a central curve + ~60 error members
CT	CTEQ–TEA — a US theory-and-experiment project	United States	A fixed algebraic formula	CT18 — a central curve + ~60 error members
JAM	Jefferson Lab Angular Momentum	United States	A fixed algebraic formula	JAM sets — a central curve + replicas

SO WHAT IS THE ACTUAL DELIVERABLE?

Not a program — a **data product**. Each group publishes a “**PDF set**”: numerical tables of the fitted curves, distributed through **LHAPDF**, the shared library every collider analysis reads. Notice the pattern in the last column — **the uncertainty recipe decides the form of what they ship**. Hessian groups ship a fixed set of error members; Monte-Carlo groups ship an ensemble of replicas.

What “the uncertainty band” is, and the three ways to compute it



SCHEMATIC

the band is the honest statement of doubt — and it is **computed**, not measured

RECIPE 1

one fit · cheapest

Hessian



Measure **how sharply the fit worsens** as you step away from the best point. Steep walls mean a small uncertainty; shallow walls mean a large one.

- **Assumes** one clean, well-defined best fit.
- **Breaks if** a direction is flat — the maths cannot be inverted.

RECIPE 2

~100 fits

Monte-Carlo replicas



Make about a hundred **noisy copies of the data**, re-fit every one, and take how much the answers scatter as the uncertainty.

- **Assumes** the scatter of re-fits equals the uncertainty.
- **Breaks if** re-fits wander to the edge of the allowed range.

RECIPE 3

slowest · most complete

Bayesian posterior



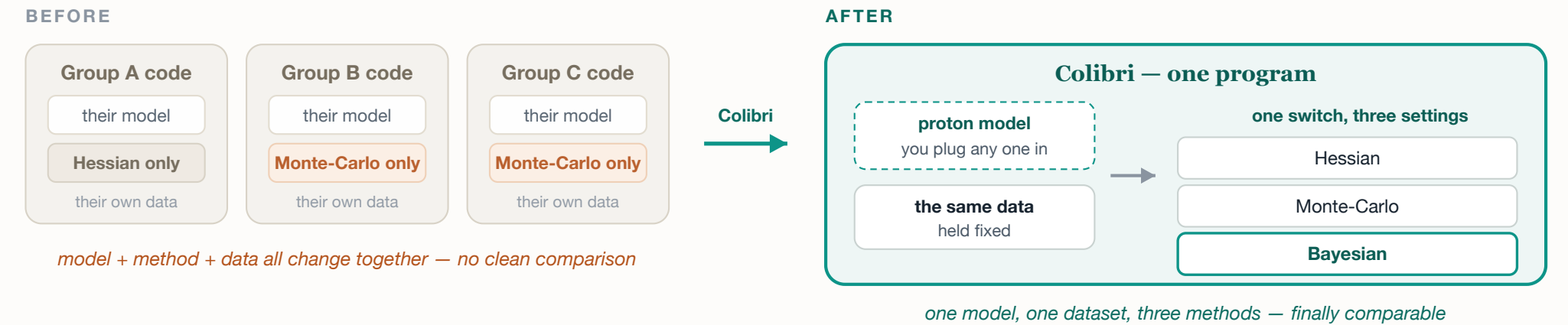
Map out **the whole range of settings** the data permits, and read the spread of the resulting curves straight off.

- **Assumes** a stated prior, and a sampler that finishes.
- **Breaks if** the sampler stops before exploring properly.

All three answer the same question by different logic — so when the data is decisive they agree, and when it is not, they need not.

One program that can run all three methods on the very same fit

Maria Ubiali leads a programme at Cambridge called **Physics Beyond the Standard Proton**, funded by the European Research Council. Its worry: a hint of new physics at the collider might really be the proton being mis-modelled. So the size of that band has to be trustworthy.



SCHEMATIC what changed · Colibri released 2025 by the Cambridge group (Ubiali et al.)

Now you can hold the model and the data completely fixed and change *only* the method — so any difference you see is caused by the method alone.

What Maria asked me to do

Her brief, from our meeting and follow-up email (7–8 July 2026). The last line is what this deck answers.

1

Move from a toy model to a real one

Colibri had only been demonstrated on the 13-parameter Les Houches toy. Implement **MSHT20** — “one of the most broadly used PDF sets in LHC phenomenology” — as a new Colibri model.

done

2

Start with data we already understand

Use the same DIS data as the earlier toy fits so the comparison is clean — then scale up.

done

3

Compare Hessian vs Monte-Carlo vs Bayesian uncertainties

Her stated **final goal** — the head-to-head that Colibri makes possible for the first time.

this deck

A real, widely-used parametrisation · real data · all three uncertainty methods under one roof.

What I did, and what came out

The published demonstration of Colibri used a simplified **13-setting toy proton**, and there all three methods agreed. I swapped in the **real 52-setting** proton form and compared the methods on real data. Everything after this slide is the detail behind these numbers.

THE MODEL

52 settings

the real MSHT proton form, re-implemented — not the 13-setting toy

THE DATA

3 092

real measurements of electrons and neutrinos scattering off protons

THE WORK

28 runs

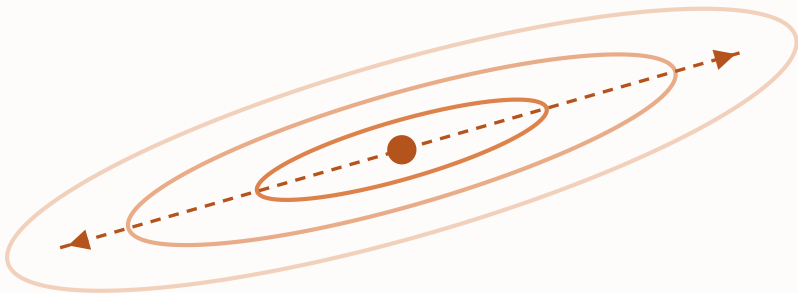
over four weeks — validating the tools, then the comparison itself

THE OBSTACLE

14 of 52

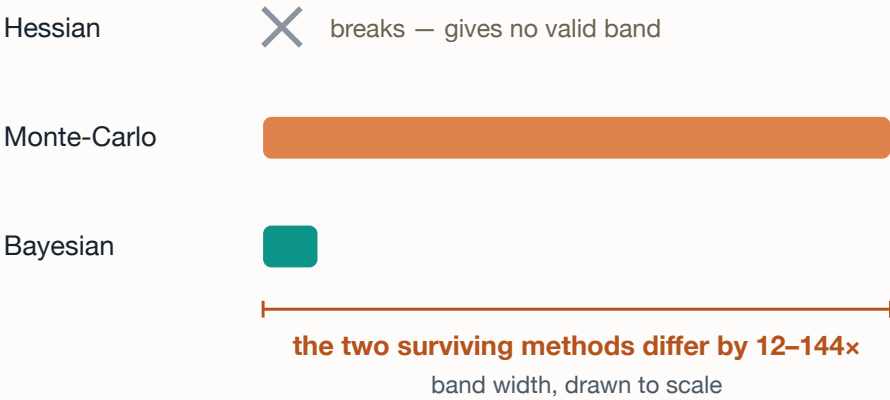
settings the data cannot pin down at all — mostly the strange quarks

WHY IT IS HARD



a “flat direction” — the fit slides freely, the data does not object
each ring = settings that fit the data equally well · worst for strange quarks

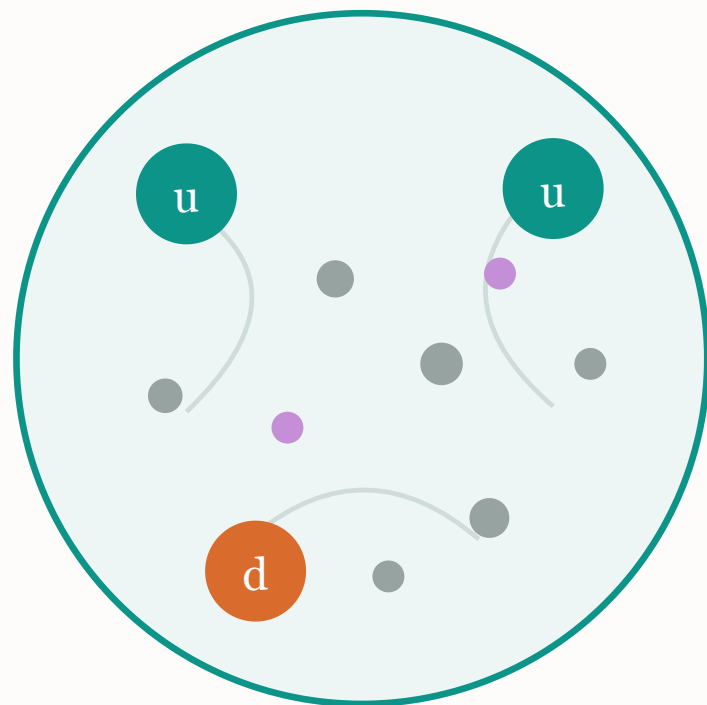
WHAT CAME OUT — SAME FIT, SAME DATA



SCHEMATIC + OUR RESULT left: why the question is ill-posed · right: the spread we measure on real data

Where the data is silent, the method — not the measurement — decides how big the uncertainty looks.

A proton is not one particle — it is a swarm of quarks and gluons



the proton

- 3 **valence quarks** (u u d) — these give the proton its identity
- **gluons** — they bind it together, and carry about half its momentum
- the **sea** — short-lived quark–antiquark pairs, including **strange**

collectively these are called *partons*

SCHEMATIC a picture of the constituents, not a measurement

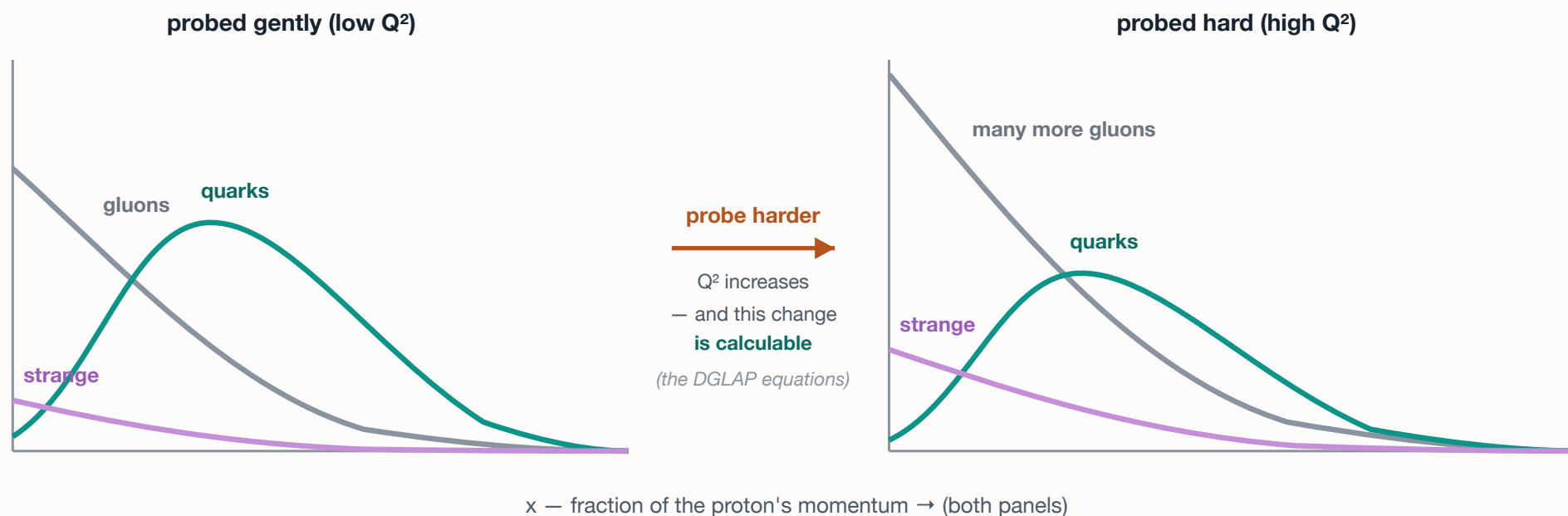
At LHC energies a proton–proton collision is really a collision between **one parton from each proton**. To predict any rate, you must know how likely each parton is to be there, carrying each share of the momentum.

So "what is inside the proton" is not background — it is an input to every prediction.

A parton distribution function is the curve that says how the momentum is shared

$f_i(x, Q^2)$ = how likely parton i carries fraction x of the momentum, when probed at energy Q

one curve per parton type — and the curve **changes** with the energy you probe at



SCHEMATIC

illustrative shapes · the real curves are fitted — that is what this project is about

A PDF depends on **two** things: the momentum share x , and the energy Q you probe at. Probe harder and you resolve more structure — many more low- x gluons. Crucially, **the change with Q is calculable** — that is what the **DGLAP equations** do. What must be **fitted** is only the shape at one starting energy; DGLAP then carries it to every other energy.

PDFs cannot be calculated from theory — they must be fitted to data

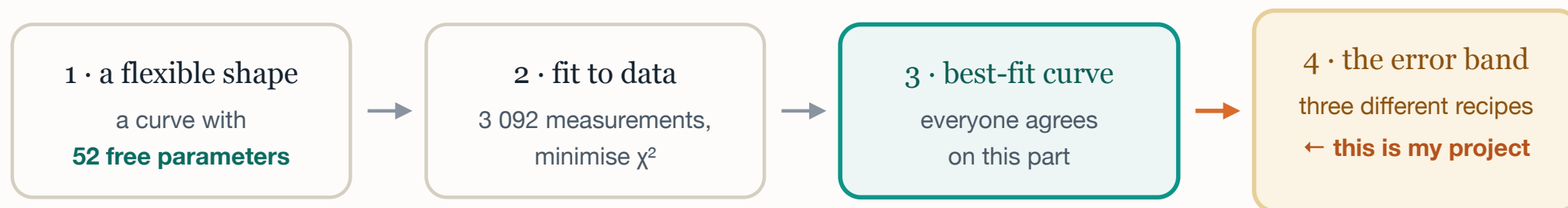


SCHEMATIC collinear factorisation — the structure of every LHC prediction

The PDF part is **non-perturbative**: there is no first-principles calculation. It is measured indirectly, by fitting flexible curves to thousands of scattering measurements. That fitted PDF — and its error bar — is then an input to **every** LHC prediction.

On many LHC measurements, the PDF uncertainty is the single largest theory uncertainty.

The fit gives a curve and a band. The band is where the disagreement lives.



SCHEMATIC the standard PDF-fitting chain, with my scope marked

THE SHAPE I USE — “MSHT20”

One of the standard published PDF forms (from the Martin–Thorne group). A flexible curve with **52 free parameters** — these are what the fit adjusts.

THE DATA I FIT TO

Deep-inelastic scattering — electrons and neutrinos fired at protons. **19 published datasets, 3 092 numbers** (SLAC, BCDMS, NMC, HERA, CHORUS, NuTeV). Each number is one **binned cross-section** at a given (x, Q^2) — itself distilled from **millions of raw collisions** — and they come with a **3 092 × 3 092 covariance matrix** (~9.6 M entries) describing how their errors are correlated.

THE SOFTWARE — “COLIBRI”

A PDF-fitting framework written in JAX. I re-implemented MSHT20 inside it so the fit is **auto-differentiable** — which part 3 needs.

Steps 1–3 are settled — everyone fits the same way and gets essentially the same central curve. **Step 4 is not settled.**
Converting “how well did the data pin down those 52 parameters” into a band on the curve can be done three different ways.

They disagree when the model has more freedom than the data can pin down

Data pins it down



one answer → the three methods agree

A “flat direction”



many equally-good answers → the methods differ

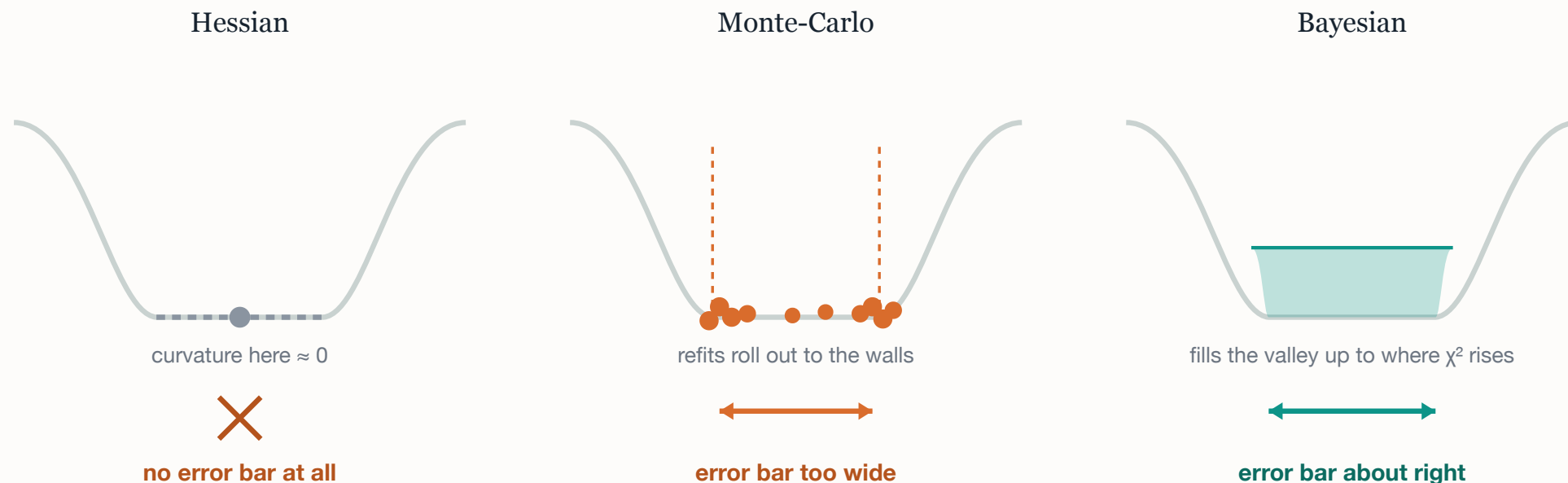
SCHEMATIC

each contour = parameter settings that fit the data equally well

DIS data cannot distinguish certain **combinations** of the 52 parameters — you can change them a lot and the fit quality barely moves. Along such a direction the error bar is no longer set by the data. **It is set by the method.**

DIS is nearly blind to the light sea and to strange quarks — that is exactly where the flat directions are.

The same flat valley, read three different ways



SCHEMATIC

the curve is the fit quality (χ^2) along a flat direction – low means “fits the data well”

All three methods are looking at the **same valley**. The Hessian only measures the **steepness at the bottom** — and a flat valley has none. Monte-Carlo lets noisy refits **roll along the floor** until the edge of the allowed range stops them. The Bayesian method **fills the valley** up to where the fit quality starts to degrade.

This one picture is the whole result: the disagreement is not a bug — it is three honest answers to an ill-posed question.

28 experiments in three stages, each with its success criteria fixed in advance

STAGE 1 · PPDF-1...11

validation

Check the tools

Reproduced a published closure-test paper, then ported the full MSHT20 parametrisation into Colibri.

- 96-check parity gate against the reference — passed to 10^{-9} .
- Caught a hidden factor in the reference that is invisible at published parameter values.

STAGE 2 · PPDF-12...16

real data

Fit real data — and find the problem

MSHT20 fits real DIS well ($\chi^2/N \approx 1.0$). But the fit turned out to be **degenerate**.

- An earlier rigid model floored at $\chi^2/N \approx 5.6$ — that floor was model rigidity, not the data.
- The flexibility that fixed it is what creates the degeneracy.

STAGE 3 · PPDF-18...28

the result

Test the three methods

Measured the degeneracy directly, compared all three uncertainty methods, and re-checked my own conclusions.

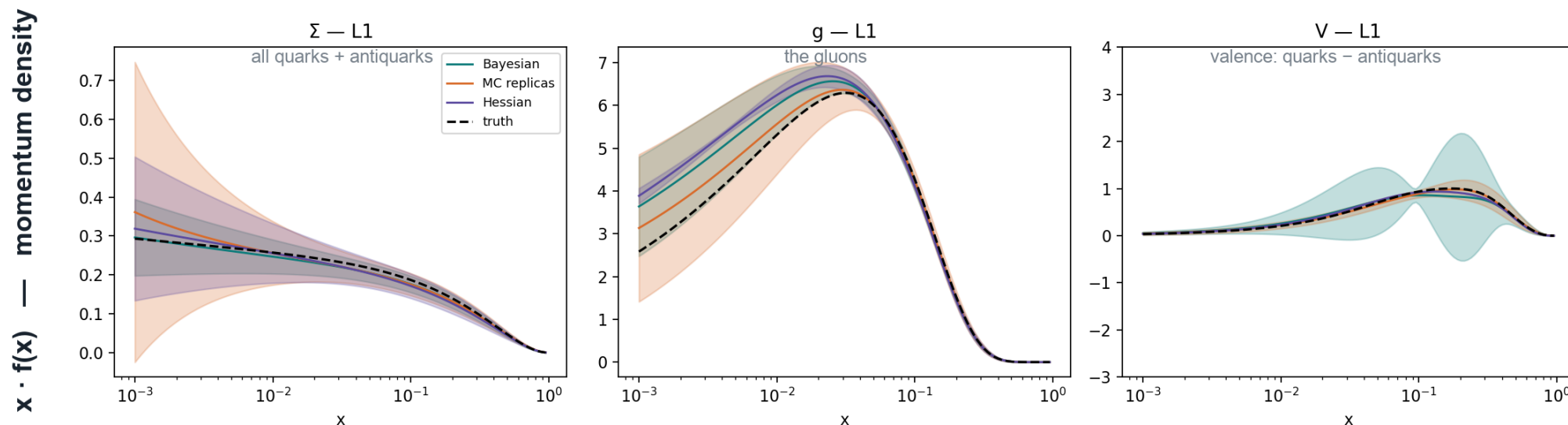
- Six pre-registered phases; 3 complete, 1 partial.
- One conclusion retracted after a convergence study.

Everything — including two bugs I caught and one claim I withdrew — is recorded in a single results ledger.

Control test: where the data is strong, all three methods agree with the truth

A **closure test**: I invent a proton with known curves, generate 3 092 fake measurements from it with realistic noise, then fit that fake data **three times** — once with each uncertainty method, using exactly the same code as the real fit.

L1 closure: three inference methods vs truth (bands = 1σ ; V panel y-clipped)

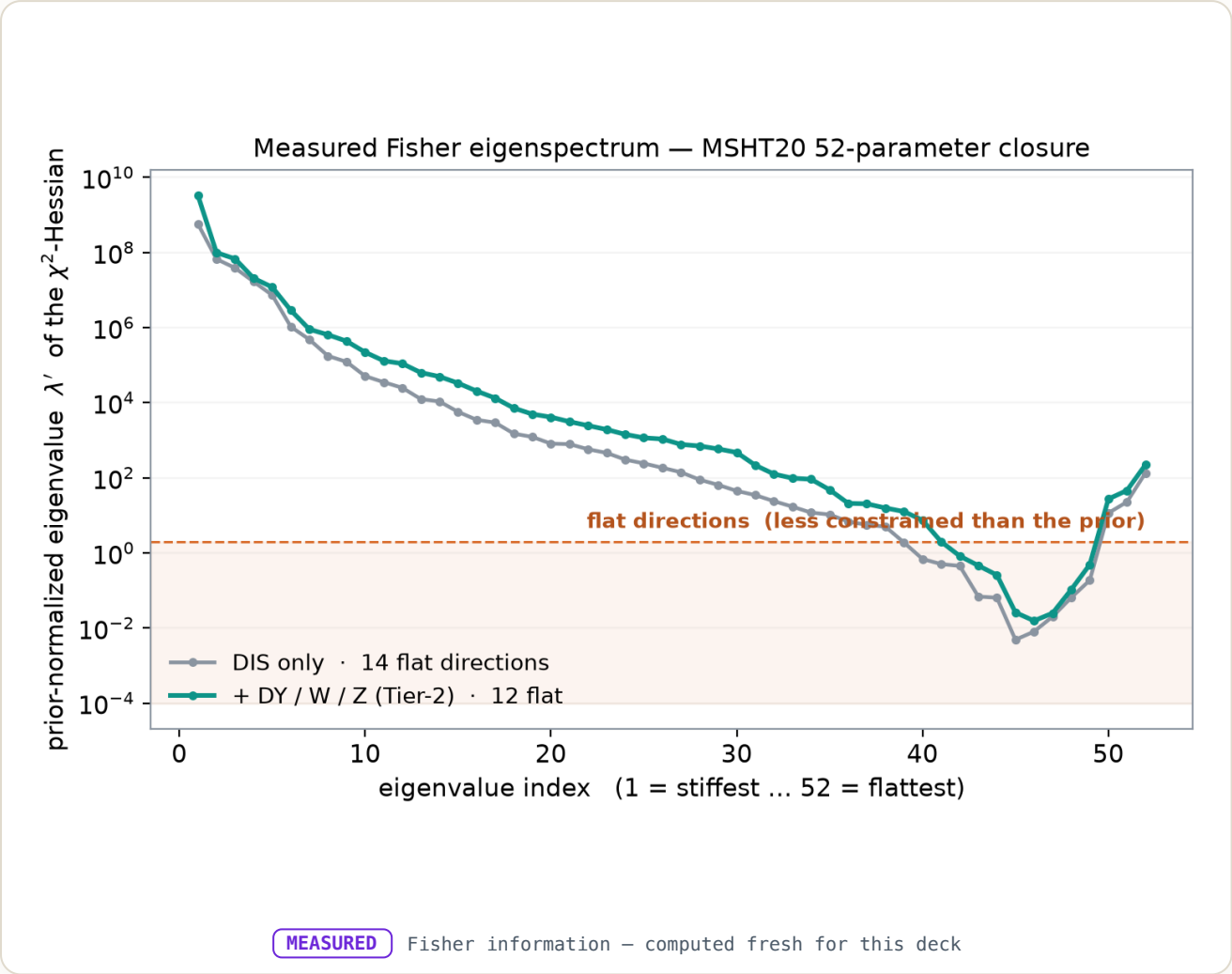


KNOWN-TRUTH TEST we invent a proton, generate fake data from it, and fit it — so the right answer is known. x = momentum share (log scale). **black dashed = the truth**. coloured bands = each method's 1σ error bar

Dashed line = the truth we planted. Bands = the three methods. For the quark singlet and the gluon, all three sit on the truth. **This proves the machinery is correct** — so any later disagreement is physics, not a bug. Notice the third panel: the bands already separate where the data runs out.

I measured how much of the model the data fails to constrain

WHAT I RAN A direct measurement of how much freedom the data actually removes , for each of the 52 parameters.	HOW Compute the curvature of χ^2 at the known truth, then find the combinations of parameters it constrains best and worst.	WHAT TO LOOK AT The tail on the right — combinations sitting in the shaded band are ones the data barely sees.
---	---	---



Every point is one **combination** of the 52 parameters, ranked from best- to worst-constrained. Points in the shaded band are ones the data barely sees at all.

14 of 52
directions unconstrained using DIS data alone

12 of 52
still unconstrained after adding Drell-Yan and W/Z data

More data helps — but only a little. And the worst-constrained directions are the **strange quarks**.

The standard sampler could not converge on this problem

WHAT I RAN

The standard Bayesian sampler (**nested sampling**) on the 52-parameter fit — the method the field normally reaches for.

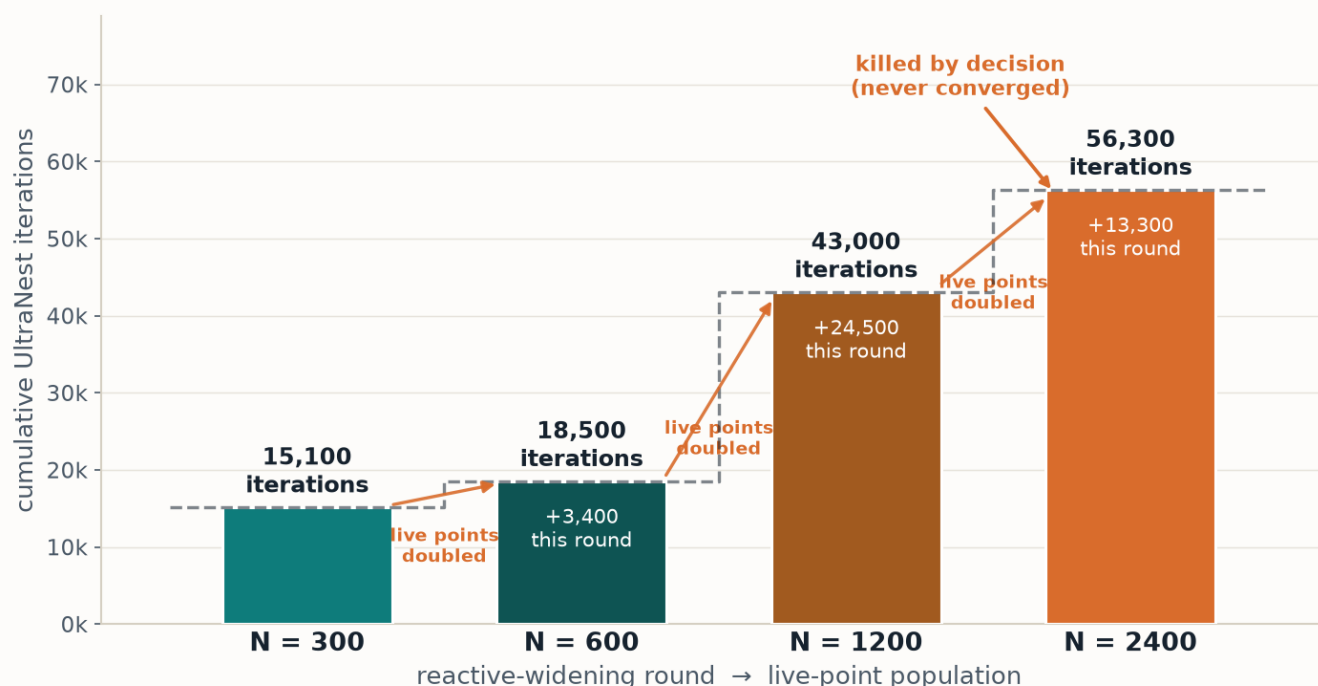
HOW

Each round **doubles the sampler's effort**. A healthy problem settles; the run is stopped when it clearly will not.

WHAT TO LOOK AT

The bars are cumulative work. They keep **climbing** instead of levelling off.

L0 discriminator: reactive widening never converged (zero-noise closure)



Full-DIS 52-parameter MSHT fit · escalating live points is the tell of a degenerate, flat-direction posterior — not a sampler setting to tune.

DIAGNOSTIC

our own run — nested sampling, effort doubled four times

Doubling the sampler's effort four times just made it run longer without settling. That is not a tuning problem — **an escalation that never converges is the signature of a flat posterior**. Switching to a gradient-based sampler, informed by the Fisher geometry, converged in about an hour.

On real data, the three methods give error bars 12–144× apart

WHAT I RAN

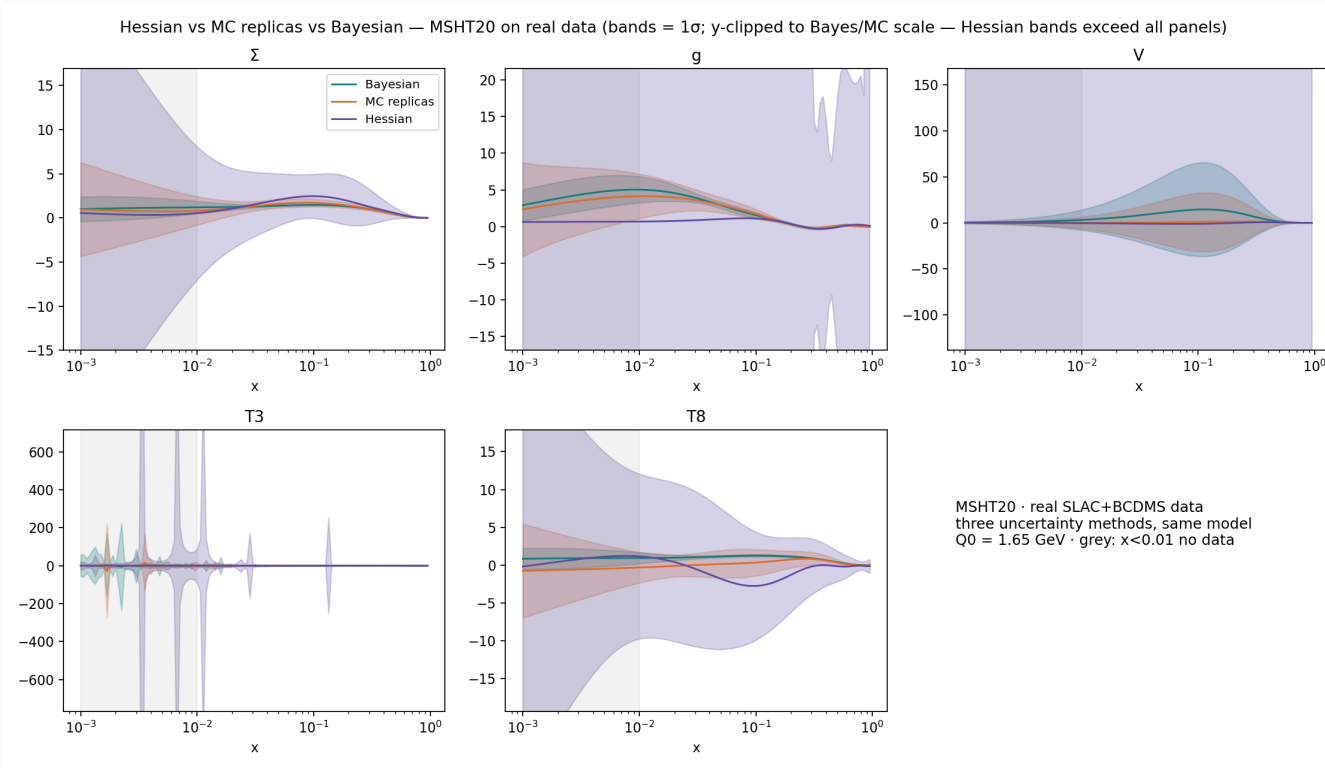
The **real** proton fit — actual DIS measurements, not simulated data, with all three uncertainty methods.

HOW

One fit, one best-fit curve. Then each method is applied to that same fit to produce its own error band.

WHAT TO LOOK AT

The **width** of each coloured band, panel by panel — same quantity, three very different answers.



REAL DATA

real proton DIS measurements · same model, same fit, three uncertainty methods

Teal = Bayesian (narrow). Orange = Monte-Carlo (wide). Purple = Hessian, which runs off the edge of every panel. Quote the Bayesian and you say $\pm 4\%$ on the gluon; quote Monte-Carlo and you say $\pm 300\%$ for the same quantity.

The Hessian gives no valid answer at all — its curvature matrix has 10 negative eigenvalues.

The bands are converged — more computing does not change them

THE WORRY

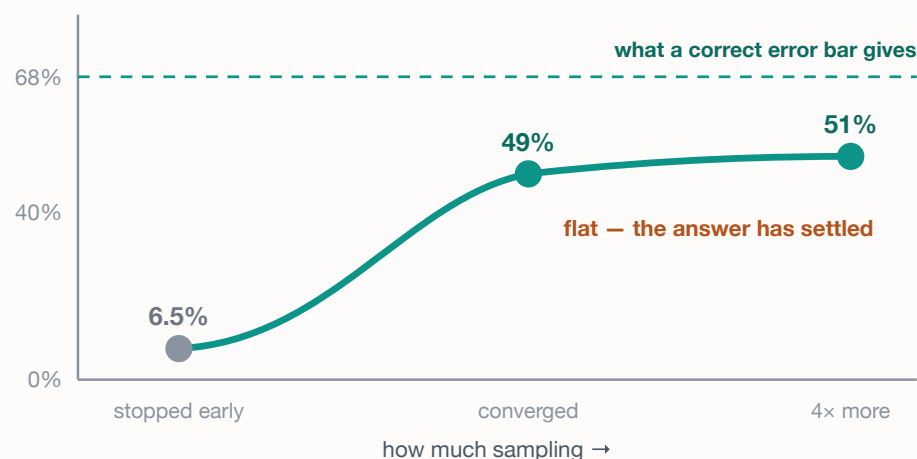
A Bayesian band is only trustworthy if the sampler explored the whole space. Mine may have stopped early.

THE TEST

Re-run the same fit with **4× more sampling** and watch whether the answer keeps moving or settles.

WHAT TO LOOK AT

If the curve **flattens**, more computing cannot widen the bands — what is left is real.



WHAT THE CHECK SHOWS

verified

Coverage climbs to ~58% and then **stops moving**: quadrupling the sampling shifts it by less than two points. The bands have **settled**.

WHY THIS MATTERS

An unfinished sampler reports bands that are **too narrow** — the dangerous direction to fail in. Having ruled that out, the disagreement between the methods is a property of **the fit itself**, not of how long we ran the computer.

The measured disagreement is real — not an artefact of insufficient computing.

Tested against a known truth, the Bayesian method is the reliable one

In a closure test the true curve is known, so we can simply count how often each method's error bar contains it. A correct 1σ band should contain the truth about **68%** of the time.

Bayesian

Contains the truth **58%** of the time — close to what it should be

about right

Monte-Carlo

Error bars **2–5× too wide** — it overstates the uncertainty

too cautious

Hessian

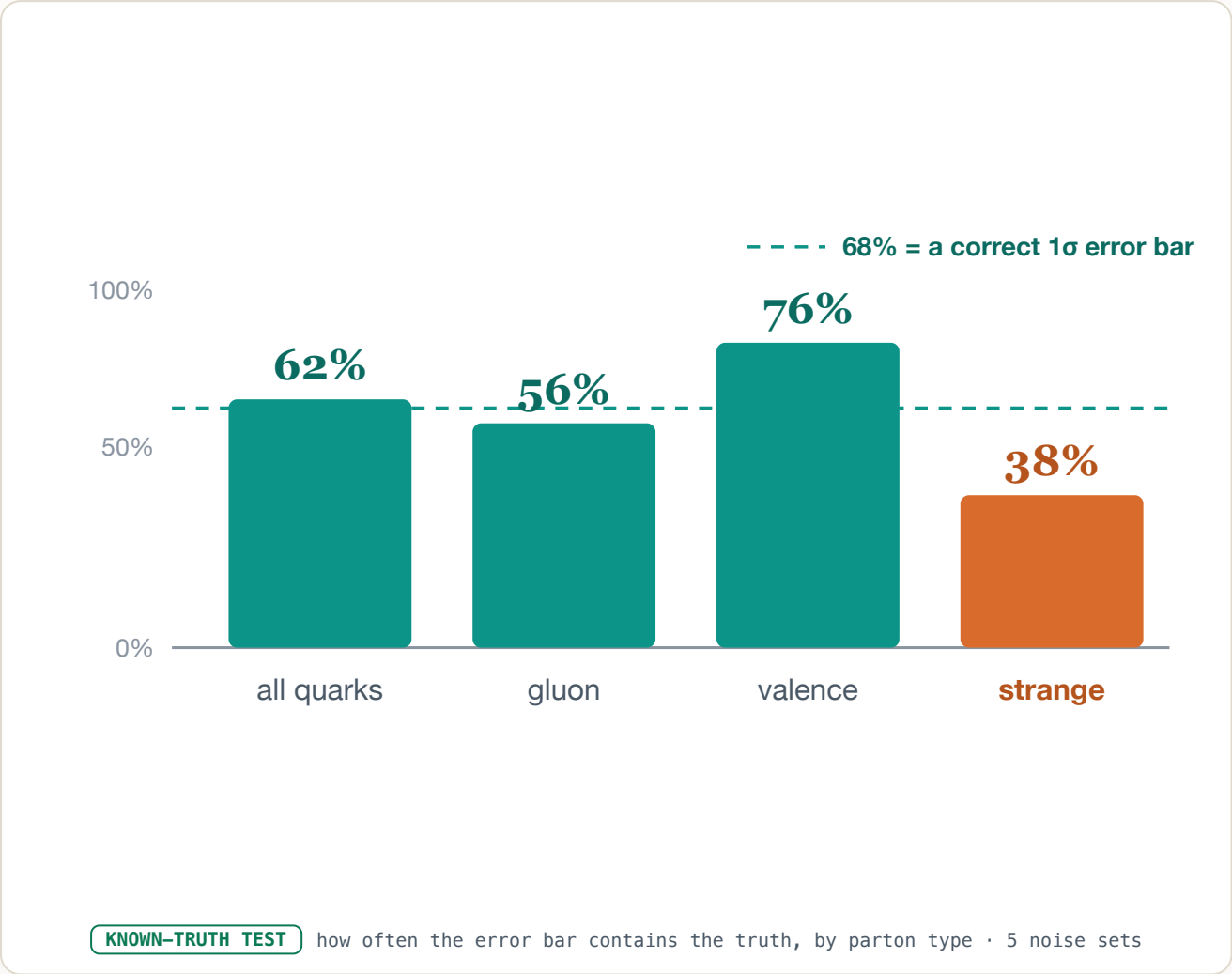
Gives **no valid error bar** in this regime — the mathematics breaks down

unusable

So the spread is not “nobody knows”. We can say which answer to use.

Even the Bayesian method understates the uncertainty on strange quarks

WHAT I RAN The closure test again , but scored separately for each type of parton, and repeated on 5 independent noisy datasets.	HOW For each one, count how often the quoted error bar actually contains the known truth, then average over the 5.	WHAT TO LOOK AT Which bars reach the dashed line . Three do. Strange does not .
--	--	---



For most of the proton the Bayesian error bar is about the right size. For **strange quarks it is too small** — the method is overconfident exactly where the data is weakest.

- Fails in **4 of 5** noise sets.
- Still 32% even in the best-converged run.
- It is the **same strange sector** the Fisher analysis flagged.

What this establishes

ESTABLISHED

evidence in this deck

- The three uncertainty methods **genuinely disagree** — by 12–144×, depending on the flavour — on real proton data.
- The cause is a **measured degeneracy**: 14 of 52 parameter directions are unconstrained by DIS.
- Tested against a known truth, the **converged Bayesian is the reliable choice**; Monte-Carlo overstates; the Hessian is invalid.
- The tools are validated, and one over-claim of mine has been found and withdrawn.

OPEN

next work

- Whether the **strange under-coverage** is physical or residual under-sampling.
- Whether it **generalises** beyond the MSHT20 functional form.
- How a **regularised** Hessian — the version production fits use — would score.
- How much of the flat-direction spread is **prior-driven** rather than data-driven.

A methodology result that is solid, self-corrected, and points at a specific piece of physics: the strange quark.

The run ledger

Every result in this deck with its modality stated. Nothing computed is presented as measured, and nothing from a closure test is presented as a real-data result.

CLOSURE TESTS — TRUTH IS KNOWN		our sim
PPDF-1...6	Reproduced a published closure paper — level-0 recovery to $\chi^2/N \approx 10^{-4}$, level-1 to 0.98; three-method comparison against truth	
PPDF-10/11	MSHT20 ported to JAX; 96-check parity gate to 10^{-9} ; closure passes with coverage 65–100%	
PPDF-19/20	Fisher degeneracy measured: 14 flat directions (DIS) $\rightarrow$ 12 (+DY/W/Z); flattest are the strange sector	
PPDF-24/25	Toy validation + the decisive convergence study over 5 noise sets — coverage 58% mean, strange 38%	
PPDF-26	Degeneracy sweep — under-converged, reported inconclusive	

REAL PROTON DATA		real data
PPDF-7...9	First real-data fits with simpler models — $\chi^2/N \approx 5.6$ floor; two bugs found and logged	
PPDF-12	MSHT20 on real DIS: $\chi^2/N \approx 1.0$, $\Delta\log Z \approx +1450$ — the floor was model rigidity	
PPDF-13/14	Monte-Carlo ensemble (82 replicas) and Hessian — Hessian indefinite, 11/52 negative	
PPDF-15/23	Three-method comparison on real data — the 12–144× spread ; the figure shown in part 3I	
PPDF-28	Converged real-data Bayesian, blind metric pre-declared — reproduces the earlier band independently	

What this does not establish

Scope

- DIS-only data — **not a global fit**. Statements are about methods at this data-to-parameter ratio.
- One parametrisation (MSHT20). A neural-network form is untested.
- Real-data verdict is **transferred** from closure; no external anchor yet.

Method fairness

- The Hessian was tested in its **raw** form, not the regularised version production fits use.
- The degeneracy sweep under-converged, so the cure conclusion rests on the Fisher measurement.
- Prior-width sensitivity has not been scanned.

The strange result

- Strange parameters have the **lowest sampling quality**, so physical vs numerical is not yet fully separated.
- The Fisher spectrum used finite differences and a prior normalisation — the flat count is threshold-dependent.

None of these overturns the main finding — but each bounds what it means.

Where this goes next — and what I need from you

STEP 1

~3 h compute

Finish the sweep

Repeat the coverage test at several degeneracy levels, properly converged, to get the trend rather than two points.

I can do this myself.

STEP 2

quick

A fair Hessian rematch

Re-test the Hessian in the regularised, dynamic-tolerance form that production fits actually use.

I can do this myself.

STEP 3

needs your steer

Add strange-sensitive data

Neutrino-DIS dimuon and W+charm directly constrain strange. This would test whether pinning strange shrinks both the degeneracy *and* the method disagreement.

This is the ask — it needs FK-table integration.

Steps 1–2 tighten the methodology. Step 3 turns it into a physics result about the strange quark.

Setup & references

THE SETUP, IN FULL

model	MSHT20 functional form, 52 free parameters, 4 analytic sum rules
framework	Colibri — JAX, auto-differentiable likelihood
data	full-DIS: 19 datasets, ≈3 092 points (SLAC, BCDMS, NMC, HERA NC/CC, CHORUS, NuTeV)
theory	FK tables, theoryid 40000000, t ₀ covariance
closure	level-1 (realistic noise), planted truth = MSHT20 central values
metric	pointwise 1σ coverage for Σ, g, V, T8 at x > 5×10 ⁻⁴ ; pass ≥55%
sampler	numpyro NUTS, dense Fisher-preconditioned mass matrix, warmup 2500 / 3000 samples

REFERENCES & PROVENANCE

MSHT20	Bailey, Cridge, Harland-Lang, Martin, Thorne — Eur. Phys. J. C 81 (2021) 341
Colibri	PBSP JAX PDF-fitting framework
NNPDF4.0	comparison reference for global-fit quality
ledger	RESULTS_LEDGER.md — every number in this deck
lab log	HESSIAN_EIGENSPECTRUM_LOG.md — Fisher method and iterations
blind	real-data metric committed to version control before unsealing
full report	21-page written companion to this deck