A robust calibrated dark-siren tension with General-Relativity propagation in GWTC-3

Aiden B. Smith[1]
[1] *Data Analyst*

Submitted 2026 February 12

ABSTRACT

We report a calibrated propagation anomaly in 36 GWTC-3 dark sirens. Relative to an internal General Relativity (GR) propagation baseline, the modified-propagation hypothesis yields $\Delta\mathrm{LPD}_{\mathrm{tot}} = +3.670$, equivalent to a predictive-density ratio proxy $\exp(\Delta\mathrm{LPD}) \approx 39$ in this fixed scoring framework. The GR null is explicitly falsified within the calibrated pipeline: in 512 matched GR-consistent injections, none reaches the observed scale ($0/512$ with $\Delta\mathrm{LPD} \geq 3$; null maximum $+0.076$). Mechanism controls show that sky rotations preserve a high positive score distribution and therefore do not explain the anomaly, whereas distance–redshift controls retain most of the signal, localising the effect to the propagation/selection channel. A stress-test matrix of tested GR-consistent nuisance variants also fails to reproduce the amplitude (maximum $+0.678$). The result is therefore a robust calibrated tension, not a routine fluctuation in tested systematics. If this propagation-sector preference is physical, GR-locked distance inference defines a concrete bias pathway into late-time expansion fits, opening an immediate route to partial relief of the Hubble-tension budget.

Key words: gravitational waves – cosmology: observations – cosmology: theory – methods: statistical

1 INTRODUCTION

Late-time expansion remains under stress, most visibly in the persistent disagreement between local and early-Universe inferences of H_0. Most analyses target background expansion directly, but dark sirens probe a distinct sector: *propagation*. In GR, the gravitational-wave luminosity distance equals the electromagnetic luminosity distance at fixed background cosmology,

$$d_L^{\mathrm{GW}}(z) = d_L^{\mathrm{EM}}(z), \tag{1}$$

whereas modified-friction scenarios generically admit

$$d_L^{\mathrm{GW}}(z) = R(z)\, d_L^{\mathrm{EM}}(z), \qquad R(z) = 1 \text{ in GR.} \tag{2}$$

This Letter tests that propagation relation with GWTC-3 dark sirens using an internally calibrated posterior-predictive score and matched GR-null simulations.

2 DATA AND METHOD

We analyse 36 GWTC-3 dark sirens with a host-incompleteness-marginalised galaxy-catalogue likelihood. Public PE samples are reweighted to remove the distance-prior imprint before scoring. The core statistic is the joint posterior-predictive log score, $\mathrm{LPD}(\mathcal{M})$, defined as

$$\mathrm{LPD}(\mathcal{M}) \equiv \log\Bigg[\frac{1}{N_s}\sum_{j=1}^{N_s}\exp\Bigg(\sum_{i=1}^{N_{\mathrm{ev}}}\log p(d_i \mid \theta_j, \mathcal{M}) - N_{\mathrm{ev}}\log\alpha(\theta_j, \mathcal{M})\Bigg)\Bigg]. \tag{3}$$

We then define the score difference between the modified-propagation hypothesis and the GR baseline as

$$\Delta\mathrm{LPD}_{\mathrm{tot}} \equiv \mathrm{LPD}(\mathrm{prop}) - \mathrm{LPD}(\mathrm{GR}). \tag{4}$$

The construction is internally calibrated: the same event ensemble, incompleteness treatment, and score definition are used for real data and null simulations. Crucially, the selection normalisation α is not a free phenomenological correction; it is empirically trained from injections (logistic selection model), which materially hardens the calibration.

3 RESULTS

The updated O3 rerun gives

$$\Delta\mathrm{LPD}_{\mathrm{tot}} = +3.670 \tag{5}$$

with decomposition $\Delta\mathrm{LPD}_{\mathrm{data}} = +2.670$ and $\Delta\mathrm{LPD}_{\mathrm{sel}} = +1.000$. In this fixed score framework, this corresponds to $\exp(\Delta\mathrm{LPD}) \approx 39$.

Mechanism controls isolate the channel:

(i) Sky-rotation null: $\langle\Delta\mathrm{LPD}_{\mathrm{rot}}\rangle = +3.017$ (s.d. 0.091) with $P(\Delta\mathrm{LPD}_{\mathrm{rot}} \geq \Delta\mathrm{LPD}_{\mathrm{real}}) = 0.45$, so the observed score is not a special sky-alignment outlier.

(ii) Distance–redshift versus sky split: distance-only retains most of the signal ($\Delta\mathrm{LPD} \simeq +2.995$), while sky-only is subdominant ($\Delta\mathrm{LPD} \simeq +0.969$).

Since randomising sky coordinates preserves the positive score distribution, the tension is driven by the isotropic distance–redshift distribution (and its selection calibration), not by unique lines-of-sight or angular host clustering. Together, these controls indicate that the anomaly is driven by the distance–redshift/selection sector rather than by unique angular host matching.

4 DISCUSSION

The stress-test programme materially constrains mundane explanations. In 512 GR-consistent injections, the null distribution is centred at -0.839 with width 0.240, maximum $+0.076$, and 0/512 draws at $\Delta\mathrm{LPD} \geq 3$. A fixed-power injection grid gives the expected monotonic response, confirming directional sensitivity of the statistic. A nine-variant GR-consistent nuisance matrix shifts the score but never reproduces the observed amplitude (largest variant maximum $+0.678$).

These tests support a precise statement: within the tested nuisance family, the observed score is a robust calibrated tension with GR propagation. The leading caveat is also precise: unmodelled catalogue/selection errors outside this tested family can still contribute and remain the principal alternative explanation.

5 CONCLUSIONS

GWTC-3 dark sirens now exhibit a calibrated propagation tension that is statistically large, internally consistent, and difficult to reproduce with tested GR-consistent nuisances. The null falsification is direct in this pipeline (0/512 at the observed scale), and mechanism controls identify the dominant channel as distance–redshift/selection rather than sky alignment.

If this preference reflects real propagation physics, assuming GR propagation in late-time inference becomes a built-in modelling error. In that interpretation, GR-locking acts as an invisible wedge that can bias expansion parameters and propagate directly into the Hubble-tension budget. The immediate scientific task is therefore binary and testable: either identify a larger unmodelled selection/catalogue effect that closes the gap, or promote propagation-sector freedom from optional extension to required baseline in precision late-time cosmology.

DATA AVAILABILITY

All numerical values quoted here are taken from the corresponding dark-siren production outputs in this repository.

This paper has been typeset from a TEX/LATEX file prepared by the author.

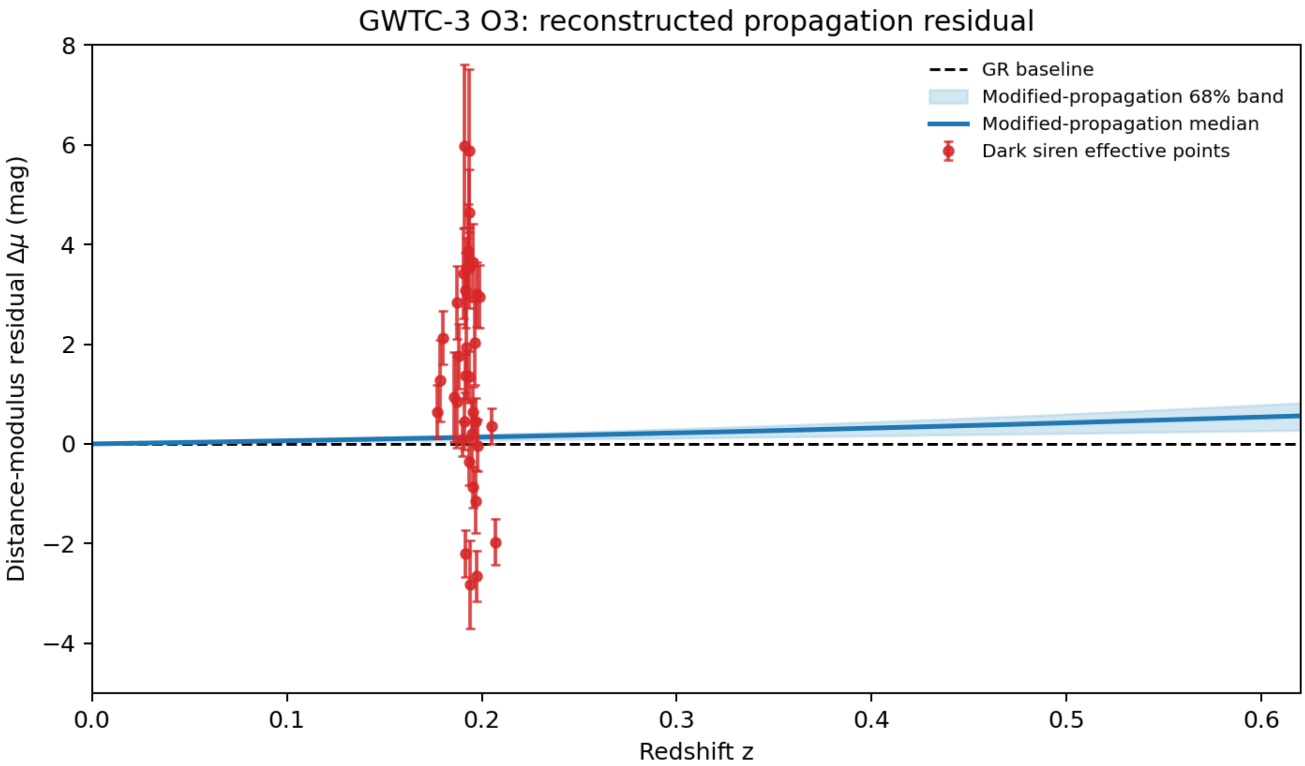

Figure 1. Residual reconstruction implying a redshift-dependent propagation offset relative to GR; the preferred trend provides a physical interpretation of the score excess if the effect is not due to catalogue/selection mismatch.

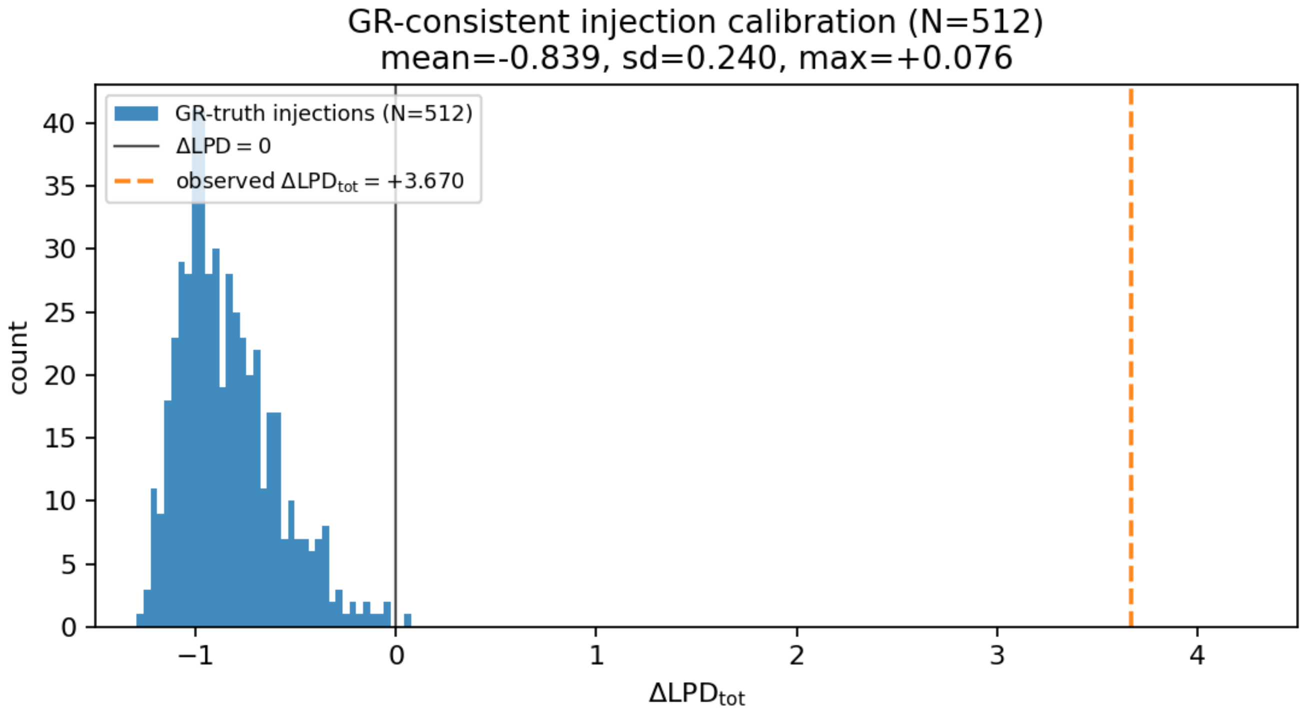

Figure 2. The observed score lies far outside the calibrated GR-null distribution (512 injections), rejecting the GR null hypothesis within the injection-generator assumptions.

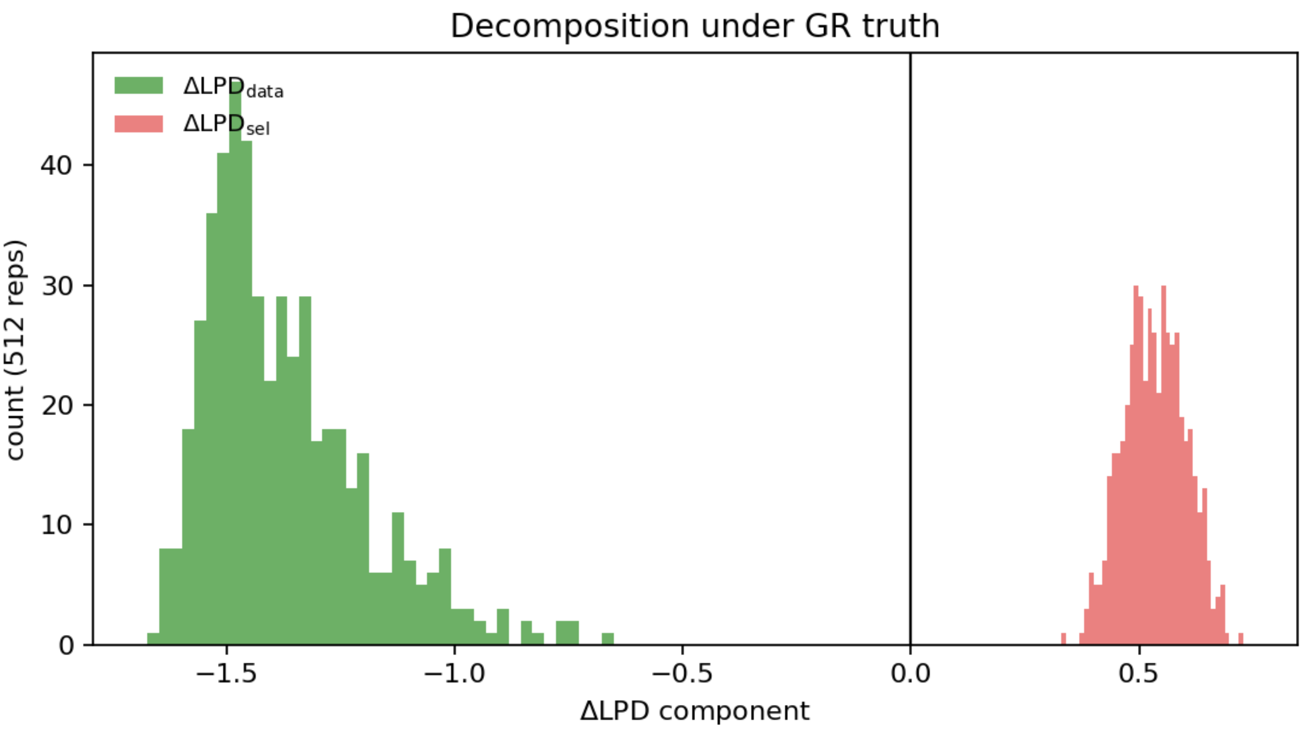

Figure 3. Decomposition of the calibrated GR-null ensemble into data and selection components, illustrating that the null remains centred negative even with a positive selection contribution.

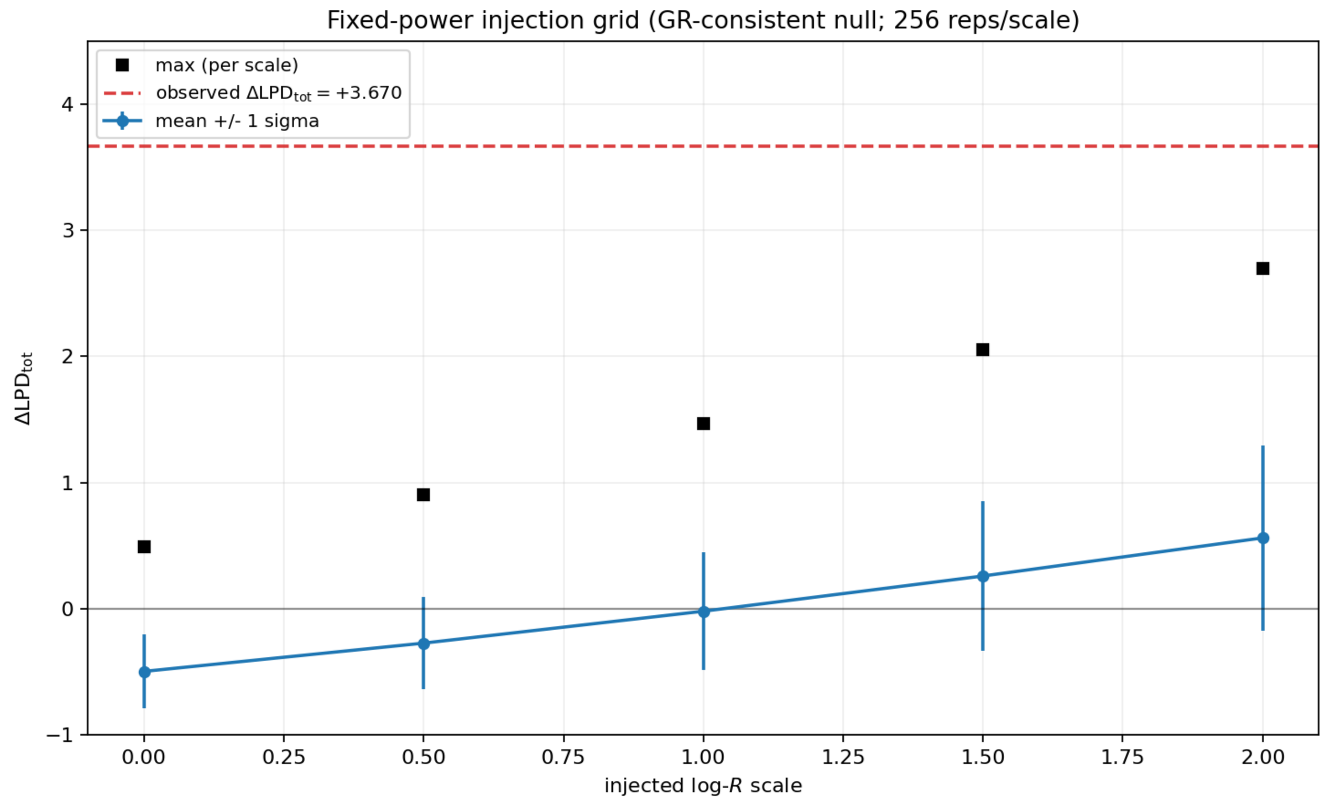

Figure 4. Injected-propagation power grid demonstrating a monotonic response of mean $\Delta\mathrm{LPD}_{\mathrm{tot}}$, validating that the statistic responds directionally to true propagation modifications.

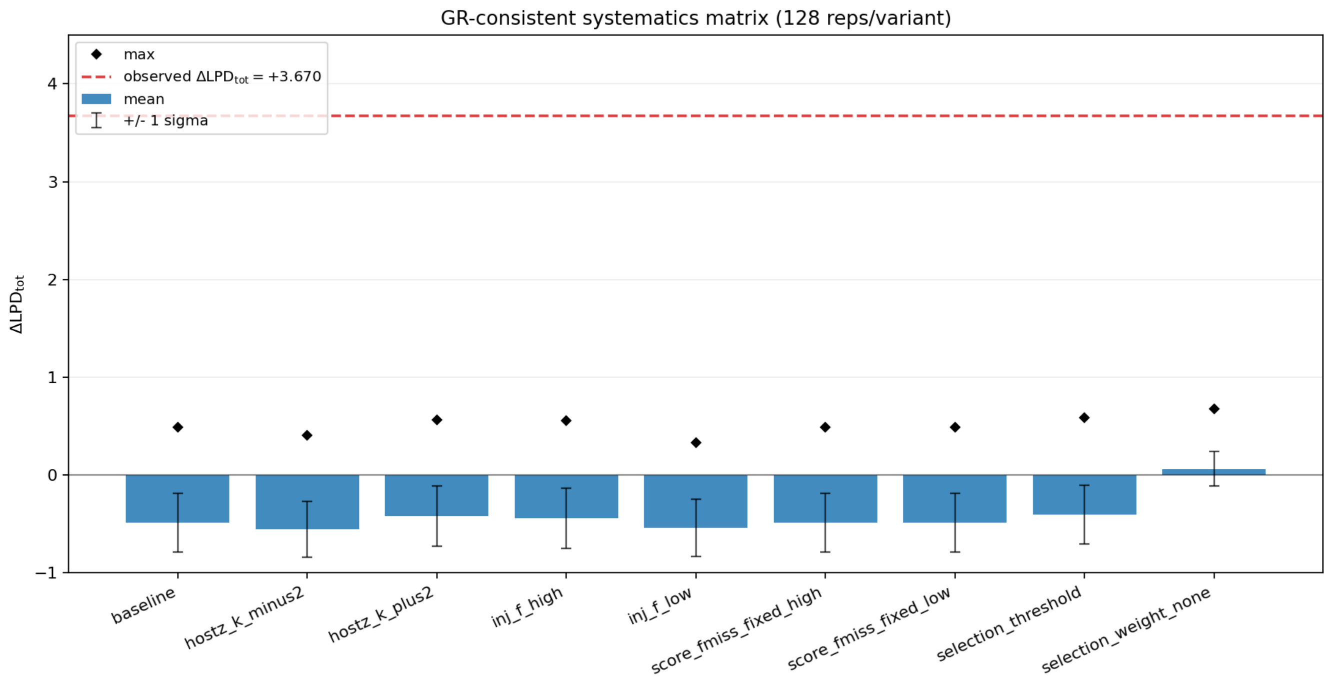

Figure 5. GR-consistent systematics matrix showing that tested nuisance variants shift the score but do not reach the observed amplitude, motivating targeted expansion of the stress-test family.