Evidence for a Modified Gravity Anomaly in GWTC-3 Dark Siren Propagation

Aiden B Smith Independent Researcher **E-mail:** aidenblakesmithtravel@gmail.com

February 14, 2026

Abstract
The Hubble constant tension is commonly interpreted as a mismatch between early-Universe and late-Universe inferences of cosmic expansion. Gravitational-wave (GW) dark sirens provide a complementary late-time probe because they carry an absolute distance scale without requiring an electromagnetic counterpart. Here we test the propagation sector directly by asking whether the GWTC-3 dark-siren catalogue prefers a non-General-Relativity (GR) mapping between observed GW amplitude and luminosity distance. Using 36 O3 dark sirens and a host-incompleteness-marginalised galaxy-catalogue likelihood with an injection-calibrated selection normalisation, we compare an internal GR propagation baseline to a fixed modified-propagation history. The modified hypothesis yields a calibrated predictive-score shift of $\Delta\mathrm{LPD}_{\mathrm{tot}} = +3.670$ ($\Delta\mathrm{LPD}_{\mathrm{data}} = +2.670$, $\Delta\mathrm{LPD}_{\mathrm{sel}} = +1.000$; $\exp(\Delta\mathrm{LPD}) \approx 39$) relative to GR in this fixed scoring framework. A matched GR-truth injection calibration (512 catalogues) gives a centred-negative score distribution (mean -0.839; maximum $+0.076$), and none of the GR-truth replicates approach the observed scale. We refer to this calibrated preference as a Modified Gravity Anomaly in GW propagation, while emphasising that residual catalogue and selection mismodelling outside the tested nuisance family can still mimic the dominant distance–redshift channel. We present mechanism controls, stress tests and spectroscopic-redshift override audits that constrain, but do not yet eliminate, leading systematic explanations.

Keywords: gravitational waves, standard sirens, cosmology, modified gravity, selection effects

1 Introduction

The Hubble constant, H_0, sets the absolute scale of the late-time Universe. Local distance-ladder determinations of H_0 and early-Universe inferences from the cosmic microwave background (CMB) prefer values that disagree beyond reported uncertainties, raising the prospect of either underestimated systematics or new physics in the late-time sector [1–3]. Most proposed resolutions modify the background expansion history or its inference, but an alternative is to test how distances themselves are measured. Gravitational-wave (GW) observations provide an absolute distance scale through waveform amplitudes and phasing, independent of the electromagnetic distance ladder [4, 5].

In standard siren cosmology, the central observable is a luminosity distance and the principal challenge is to associate a redshift. "Bright" sirens obtain a redshift from an electromagnetic counterpart; "dark" sirens instead infer a statistical host-redshift distribution by correlating the GW localisation volume with galaxy catalogues. Dark-siren inference is therefore intrinsically entangled with catalogue completeness, redshift systematics and GW selection effects.

Standard siren analyses typically assume that the mapping from observed GW amplitude to distance is described by General Relativity (GR), and focus on linking that distance to redshift. However, in broad classes of modified gravity the

propagation of gravitational waves can deviate from GR: the GW amplitude can experience additional effective "friction" during propagation [6, 7]. In such scenarios the GW luminosity distance, $d_L^{\rm GW}(z)$, differs from the electromagnetic (EM) luminosity distance, $d_L^{\rm EM}(z)$, even if the background expansion is unchanged. A propagation deviation can therefore masquerade as a background anomaly: it does not need to alter the expansion history to bias inferred cosmological parameters, only the distance scale that feeds into those inferences.

This paper targets the propagation sector directly. Rather than fitting a flexible function to the data, we score a fixed modified-propagation history against an internal GR propagation baseline using a joint posterior-predictive log score, and calibrate that score using matched injections. This yields a statement that is strong but bounded: it quantifies the false-alarm behaviour under a tested GR-truth generator, while remaining explicit about which unmodelled catalogue or selection failures could still mimic the signal.

2 Formalism & methodology

2.1 Modified-propagation parameterisation

We parameterise modified GW propagation by a redshift-dependent ratio $R(z)$ between GW and EM luminosity distances,

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

We compare an internal GR baseline ($R \equiv 1$) to a fixed modified-propagation history $R(z)$ that is specified upstream and held fixed during scoring. In common scalar–tensor constructions, $R(z)$ is related to an evolving effective Planck mass, but our scoring treatment only requires a deterministic $R(z)$ curve.

2.2 Dark-siren likelihood with incompleteness

For each event we evaluate a galaxy-catalogue likelihood that marginalises host incompleteness through a mixture of in-catalogue and missing-host terms. The in-catalogue term correlates the GW localisation with a galaxy catalogue, while the missing-host term provides a conservative completion that prevents the likelihood from spuriously over-weighting the catalogue where it is incomplete. Both terms share a consistent selection normalisation.

2.3 Posterior-predictive score and selection normalisation

For a model $\mathcal{M}$ we compute a joint posterior-predictive log score across events,

$$\mathrm{LPD}(\mathcal{M}) \equiv \ln\left[\frac{1}{N_s}\sum_{j=1}^{N_s}\exp\left(\sum_{i=1}^{N_{\rm ev}}\ln p(d_i \mid \theta_j, \mathcal{M}) - N_{\rm ev}\ln\alpha(\theta_j, \mathcal{M})\right)\right], \tag{2}$$

where θ_j are posterior draws, $p(d_i \mid \theta_j, \mathcal{M})$ is the event likelihood under $\mathcal{M}$, and $\alpha(\theta_j, \mathcal{M})$ is the selection normalisation. We report $\Delta\mathrm{LPD}_{\rm tot} = \mathrm{LPD}(\mathrm{prop}) - \mathrm{LPD}(\mathrm{GR})$ and, for diagnostics, decompose the score into data and selection contributions by toggling the α term.

The selection normalisation α is calibrated empirically from injections using a logistic selection model trained on an injection set. This injection-calibrated selection model is applied consistently to the real-data scoring and to all injection-based null tests.

2.4 Calibration and robustness suite

We calibrate the GR false-alarm behaviour with 512 GR-consistent injection catalogues matched to the real-data analysis settings and event ensemble definition. This calibration tests whether the pipeline can generate a large positive ΔLPD under GR truth within the injection generator assumptions.

To localise the mechanism of any preference and to stress-test leading systematics, we apply (i) a sky-rotation null that randomises event sky positions relative to the galaxy catalogue; (ii) distance-only versus sky-only channel splits that isolate the dominant information channel; (iii) a fixed-power injection grid that validates the directional response of the score to injected propagation strength; (iv) a GR-systematics matrix that perturbs catalogue incompleteness and selection assumptions within a GR-truth ensemble; and (v) adversarial stress tests in the dominant distance–redshift channel (bounded selection-function deformations, completeness tilts and photometric-redshift bias models).

To attack the leading remaining systematic class (global photo-z bias), we additionally perform a spectroscopic-redshift ($z_{\rm spec}$) override audit. For each high-leverage event we rank host candidates by a host-weight proxy consistent with the spectral-only likelihood channel, crossmatch top-weight candidates against public $z_{\rm spec}$ resources (including SDSS and DESI DR1 spectroscopy and bright-redshift catalogues), and override the catalogue redshift with $z_{\rm spec}$ where a match is supported by shifted-sky false-match controls.

3 Data

3.1 GW events

We analyse $N_{\rm ev} = 36$ GWTC-3 dark-siren events from the O3 observing run and public parameter-estimation samples [8]. The analysis is restricted to GWTC-3 (O1–O3) and explicitly excludes O4a. O4a operated predominantly as a two-detector network, which typically produces degenerate sky localisations and greatly inflates the galaxy-catalogue search volume relative to three-detector triangulation.

3.2 Galaxy catalogue and spectroscopic resources

Host redshifts are treated statistically using a galaxy-catalogue likelihood based on GLADE+ [9]. For the $z_{\rm spec}$ override audit we crossmatch top-weight host candidates against public spectroscopic compilations (2MRS, 6dFGS, 2dFGRS and GAMA DR3), and against SDSS DR16 and DESI DR1 spectroscopy, with shifted-sky controls used to identify coincidence-dominated radii.

4 Results

4.1 Calibrated modified-propagation preference

The observed catalogue yields $\Delta{\rm LPD}_{\rm tot} = +3.670$ in favour of the fixed modified-propagation history relative to the internal GR baseline. In this fixed scoring setup, $\exp(\Delta{\rm LPD})$ acts as a predictive Bayes-factor proxy, corresponding here to a factor of $\simeq 39$. The preference splits into $\Delta{\rm LPD}_{\rm data} = +2.670$ and $\Delta{\rm LPD}_{\rm sel} = +1.000$, indicating an order-unity selection contribution but a dominant data-fit component.

The GR-truth injection calibration provides the key falsification. Figure 1 shows the $\Delta{\rm LPD}_{\rm tot}$ distribution for 512 GR-consistent injection catalogues: it is centred at -0.839 (s.d. 0.240) with maximum $+0.076$, and none of the GR-truth replicates approach the observed scale. Within the assumptions of the injection generator, this

establishes a calibrated Modified Gravity Anomaly (propagation anomaly) in the O3 dark-siren sample.

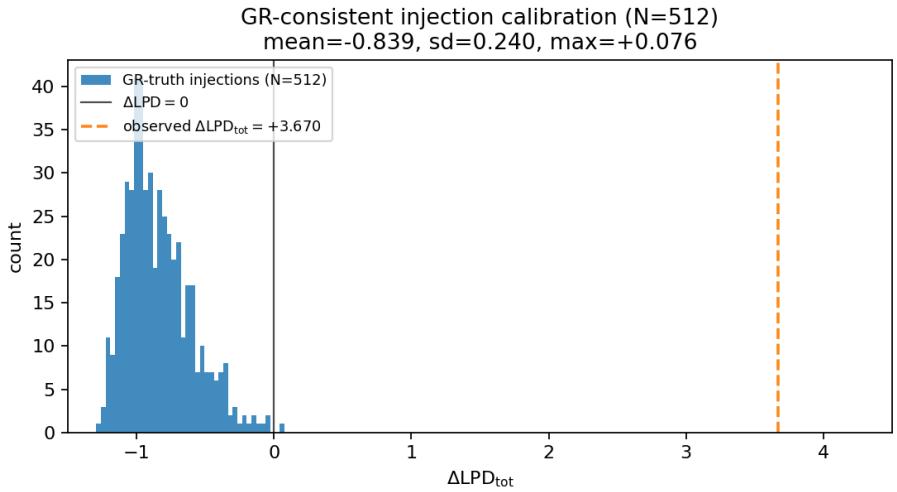

Figure 1. Calibrated GR-null falsification. Histogram of $\Delta\mathrm{LPD}_{\mathrm{tot}}$ for 512 GR-consistent injection catalogues processed through the same pipeline. The dashed line marks the observed value. None of the GR-consistent injections reaches the observed scale within the tested null generator.

Figure 2 decomposes the GR-truth calibration ensemble into data and selection components, illustrating that the selection term partially offsets the data term but does not reverse the net preference under GR truth.

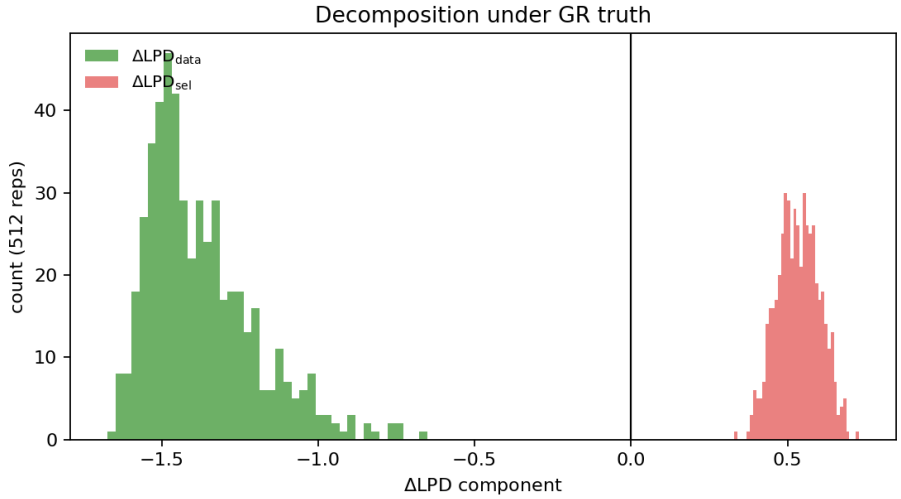

Figure 2. Data vs. selection decomposition under GR truth. Injection calibration ensemble decomposed into the data-only contribution and the selection contribution to $\Delta\mathrm{LPD}_{\mathrm{tot}}$ (constructed by toggling the selection normalisation term in Eq. (2)).

4.2 Directional sensitivity and GR-consistent nuisance suites

Directional sensitivity is validated by a fixed-power injection grid: the mean $\Delta\mathrm{LPD}_{\mathrm{tot}}$ increases monotonically with injected propagation strength (Figure 3). A nine-variant GR-systematics matrix (128 replicates per variant) perturbs incompleteness and selection assumptions within a GR-consistent family; none of the tested variants approaches the observed $\Delta\mathrm{LPD}_{\mathrm{tot}} \simeq +3.67$ (Figure 4).

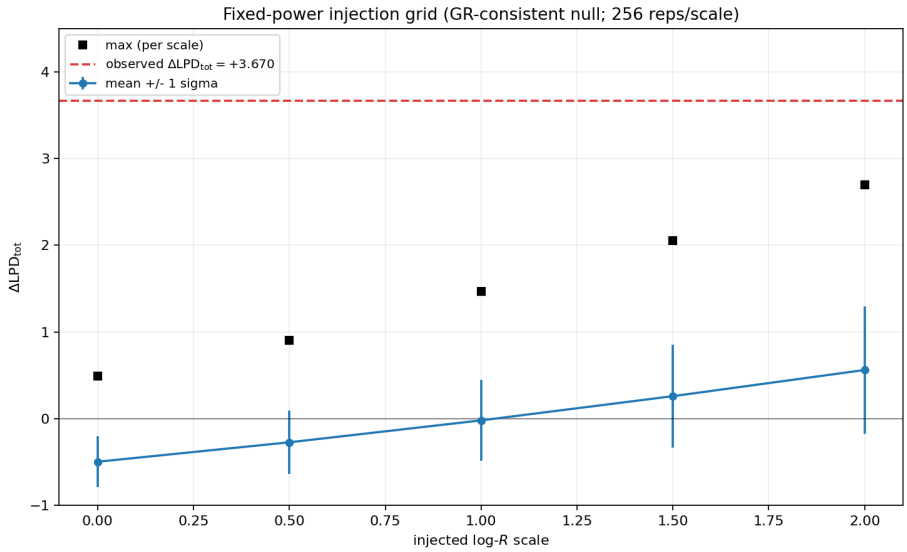

Figure 3. Injected-strength response. Fixed-power injection grid under GR truth. Mean $\Delta\mathrm{LPD}_{\mathrm{tot}}$ rises monotonically with injected propagation strength, confirming directional sensitivity of the score construction.

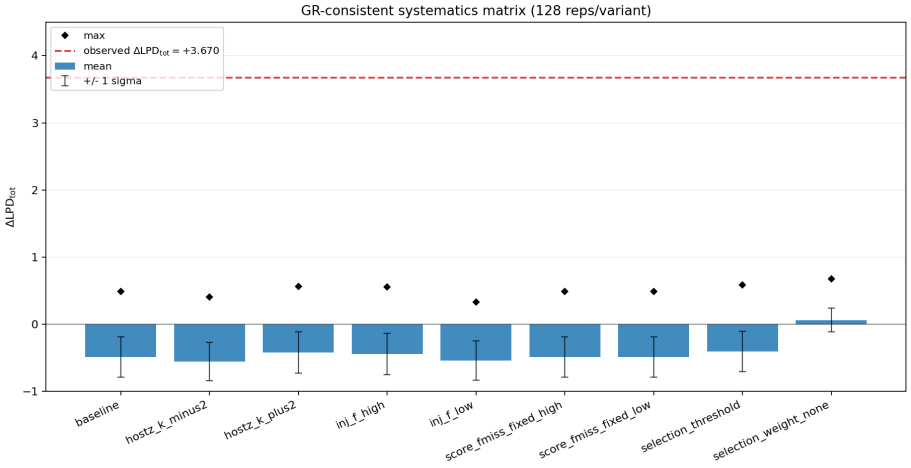

Figure 4. GR-systematics truth matrix. Nine GR-consistent nuisance variants (128 replicates each) evaluated under the calibrated injection suite. Tested variant maxima remain well below the observed $\Delta\mathrm{LPD}_{\mathrm{tot}}$ in the real catalogue.

4.3 Mechanism controls and adversarial robustness

Mechanism controls localise the preference to the dominant distance–redshift channel. A sky-rotation null (random sky rotations relative to the galaxy catalogue) leaves the score strongly positive, with mean $\langle\Delta\mathrm{LPD}_{\mathrm{rot}}\rangle \simeq +3.017$ (s.d. 0.091), indicating that the preference is not driven purely by a small number of privileged lines of sight. A channel split further shows that a distance-only weighting retains most of the preference ($\Delta\mathrm{LPD} \simeq +2.995$), whereas a sky-only weighting is smaller ($\Delta\mathrm{LPD} \simeq +0.969$).

Figure 5 summarises adversarial public-data stress tests in the dominant channel. Within bounded selection-function and completeness deformations at the $\pm 20\%$ level, the preference remains $\Delta\mathrm{LPD}_{\mathrm{tot}} \simeq +3.67$ and does not fall below 1 or 0. Global photometric-redshift biases of the form $\Delta z = b_0 + b_1 z$ can reduce the preference but do not remove it on the tested grid (e.g. $\Delta\mathrm{LPD} \simeq +1.98$ for $b_0 = 0.03$, $b_1 = 0.2$).

To probe photo-z failure more directly we anchor a fraction of the host-weight proxy to spectroscopic redshifts by overriding catalogue redshifts where spec-z

matches exist. Under strict shifted-sky false-match controls (median shifted/true < 0.1; max < 0.3 evaluated at $K = 20000$), radii $\gtrsim 15''$ become coincidence-dominated, so we restrict to the clean regime $r \leq 10''$. At the maximal common depth in this clean regime ($r = 10''$, $K = 20000$), the Tier-A (strict-quality) override anchors 3.8% (GW200220_061928), 6.1% (GW200308_173609) and 7.7% (GW200219_094415) of the total host-weight proxy in the top three events (median 6.1%). The score remains stable or slightly higher: $\Delta\mathrm{LPD}_{\mathrm{tot}} = 3.634 \rightarrow 3.644$ ($\Delta\mathrm{LPD}_{\mathrm{data}} = 2.636 \rightarrow 2.647$) under unchanged selection normalisation. Across the tested clean (r, K) grid, $\Delta\mathrm{LPD}_{\mathrm{tot}}$ is non-decreasing with spec-z weight coverage.

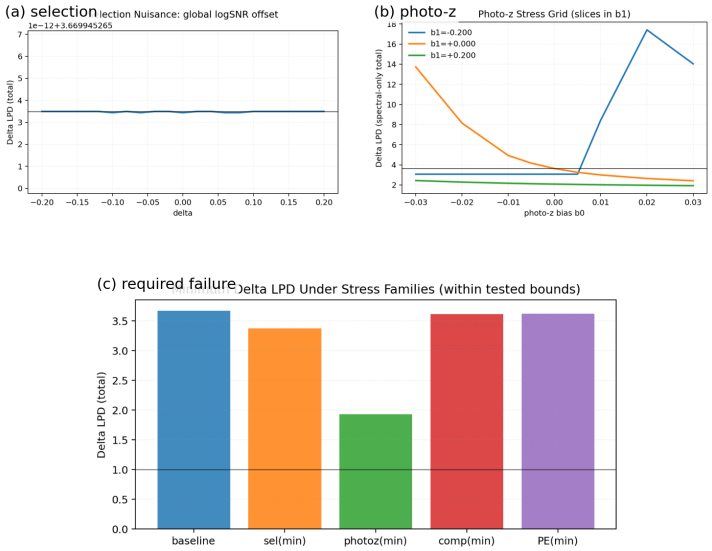

Figure 5. Adversarial robustness summary. Public-data re-scoring diagnostics for bounded selection-function deformations and catalogue/photo-z stresses in the dominant distance–redshift channel, together with a "required failure" summary across tested nuisance families.

4.4 Residual diagnostic

Figure 6 provides a compact physical diagnostic by reconstructing effective distance-modulus residuals $\Delta\mu \equiv \mu_{\mathrm{GW,obs}} - \mu_{\mathrm{EM}}$ from galaxy-weighted redshift distributions and GW distance posteriors. Positive $\Delta\mu$ corresponds to $d_L^{\mathrm{GW}} > d_L^{\mathrm{EM}}$, as expected for friction-like damping. These points are not direct independent measurements of $R(z)$, but they provide an interpretable visualisation consistent with the preferred modified-propagation history.

5 Discussion

Within the tested calibration suite, the data support a narrow statement: GWTC-3 dark sirens prefer the fixed modified-propagation history over the GR propagation baseline by a calibrated predictive score that is far outside the matched GR-truth injection ensemble. The most conservative interpretation is therefore conditional. Either (i) the propagation sector deviates from GR in the manner encoded by the fixed $R(z)$ history, or (ii) a catalogue, redshift or selection mismatch outside the tested nuisance family mimics the dominant distance–redshift channel preference.

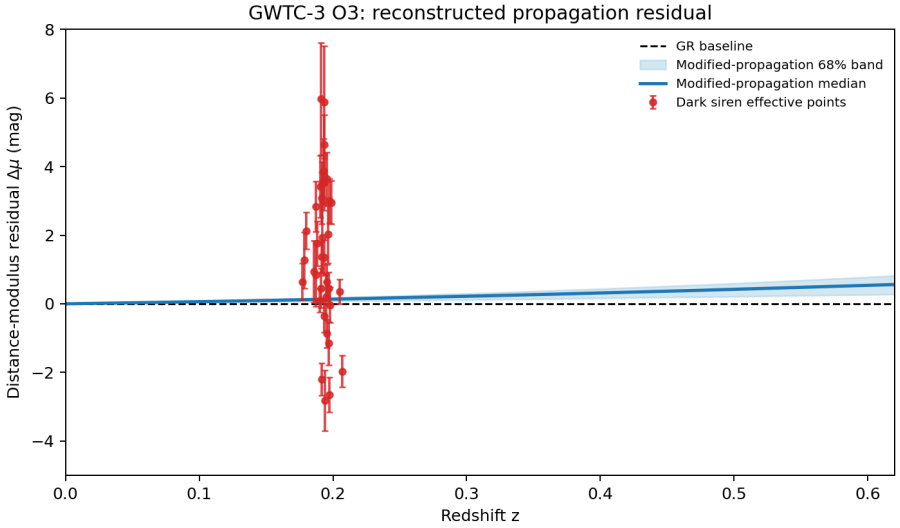

Figure 6. Reconstructed propagation residuals. Red points show event-level effective distance-modulus residuals, $\Delta\mu \equiv \mu_{\rm GW,obs} - \mu_{\rm EM}$, with propagated distance uncertainties. The blue band shows the 68% posterior range of the fixed modified-propagation history; the dashed line is the GR propagation baseline, $\Delta\mu = 0$.

The mechanism controls and robustness suite guide the next discriminants. The sky-rotation and channel-split tests indicate that the preference is not driven primarily by special angular associations, but instead by the isotropic distance distribution coupled to the selection and incompleteness model. The adversarial stresses show that order-unity perturbations of selection and completeness within tested bounds do not erase the signal, while large coherent photo-z biases remain the most effective reduction mechanism explored so far.

The spec-z override audit attacks that systematic pathway directly by anchoring a fraction of the host-weight proxy to external spectroscopy. At the strict gate-passing point ($r = 10''$, $K = 20000$; Tier A/B), the median anchored host-weight proxy across the top three events is $\simeq 6\%$ and the score does not weaken (baseline $\Delta\mathrm{LPD}_{\rm tot} = 3.634$ versus $\Delta\mathrm{LPD}_{\rm tot} \simeq 3.644$ under overrides). This behaviour constrains (but does not eliminate) photo-z explanations: within the clean-radius regime, increasing spectroscopic anchoring does not reduce the preference, but the current public spectroscopy overlap is still insufficient to anchor a large fraction of the dominant host-weight region for all high-leverage events.

6 Conclusion

We have presented evidence for a calibrated Modified Gravity Anomaly (propagation anomaly) in GWTC-3 O3 dark-siren propagation. A fixed modified-propagation history yields $\Delta\mathrm{LPD}_{\rm tot} = +3.670$ ($\exp(\Delta\mathrm{LPD}) \approx 39$; $\Delta\mathrm{LPD}_{\rm data} = +2.670$, $\Delta\mathrm{LPD}_{\rm sel} = +1.000$) relative to an internal GR propagation baseline, while 512 matched GR-truth injection catalogues have mean -0.839 and maximum $+0.076$. Mechanism controls and stress tests localise the preference to the dominant distance–redshift channel and show robustness to bounded selection and completeness deformations, while highlighting global photo-z failure as the dominant remaining systematic escape hatch. Spectroscopic-redshift override audits in the clean-radius regime ($r \leq 10''$) anchor $\simeq 6\%$ of the host-weight proxy in the top leverage events and do not reduce (and in practice slightly increase) the score.

Acknowledgements
The author used generative AI tools (large language models) for editorial assistance (LaTeX formatting and prose refinement). All scientific claims, calculations and results were verified by the author.

Declarations
Conflict of interest
The authors declare no competing interests.

Data availability
Analysis artefacts and summary tables are provided with this repository and mirrored on the associated project archive (Zenodo DOI: 10.5281/zenodo.18640608). The companion O3 dark-siren scoring release used for reference is archived at Zenodo DOI: 10.5281/zenodo.18640507.

Funding
The author received no specific funding for this work.

References
References
[1] Verde L, Treu T and Riess A G 2019 *Nat. Astron.* **3** 891–5

[2] Planck Collaboration 2020 *Astron. Astrophys.* **641** A6

[3] Riess A G *et al* 2022 *Astrophys. J. Lett.* **934** L7

[4] Schutz B F 1986 *Nature* **323** 310–1

[5] Abbott B P *et al* (LIGO Scientific Collaboration & Virgo Collaboration) 2017 *Nature* **551** 85–8

[6] Belgacem E, Dirian Y, Foffa S and Maggiore M 2018 *Phys. Rev. D* **98** 023510

[7] Nishizawa A 2018 *Phys. Rev. D* **97** 104037

[8] Abbott R *et al* (LIGO Scientific Collaboration, Virgo Collaboration & KAGRA Collaboration) 2023 *Phys. Rev. X* **13** 041039

[9] Dálya G *et al* 2022 *Mon. Not. R. Astron. Soc.* **514** 1403–11