A calibrated dark-siren tension with General-Relativity propagation

Aiden B. Smith
Data Analyst

The Hubble constant tension reflects a persistent mismatch between early-Universe and late-Universe inferences of cosmic expansion. Gravitational-wave "dark sirens" offer a complementary route to late-time inference because they carry an absolute distance scale without requiring an electromagnetic counterpart. Here we ask a sharper question than standard distance–redshift fitting: does the observed GWTC-3 catalogue prefer a non-General-Relativity propagation law in which gravitational-wave amplitudes decay differently with redshift? We report a calibrated propagation anomaly in 36 GWTC-3 dark sirens. Relative to an internal GR baseline, a fixed modified-propagation history yields a joint predictive-score shift of $\Delta\mathrm{LPD}_{\mathrm{tot}} = +3.670$ (a Bayes-factor proxy of $\exp(\Delta\mathrm{LPD}) \approx 39$ in this fixed scoring framework), while 512 matched GR-consistent injection catalogues never approach the observed scale. If physical, this preference implies that GR-locked distance inference can act as a hidden systematic in late-time cosmology and must be tested as seriously as conventional selection and calibration effects.

The Hubble constant, H_0, anchors the absolute scale of the late-time Universe. Multiple analyses now report that local determinations of H_0 and early-Universe inferences from the cosmic microwave background (CMB) prefer values that disagree well beyond quoted 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 important alternative is to test how distances themselves are measured. Gravitational-wave (GW) observations are useful here because compact-binary coalescences provide an absolute distance scale through their amplitude and waveform evolution, independent of the cosmic distance ladder.[4,5]

In standard siren cosmology, the central observable is a distance, and the main challenge is to assign a redshift. "Bright" sirens achieve this with an electromagnetic counterpart; "dark" sirens instead infer a statistical redshift distribution by correlating a GW sky localisation volume with galaxy catalogues. Dark sirens are therefore intrinsically entangled with catalogue completeness, photometric and spectroscopic redshift quality, and selection effects in the GW detector network. These complications are not a weakness if they can be calibrated: they create an opportunity to build an internally audited pipeline in which the false-alarm rate is measured, not assumed.

Standard siren cosmology typically treats gravitational waves as a clean distance indicator and focuses on linking distance to redshift. However, in many broad classes of modified gravity, the propagation of gravitational waves can deviate from General Relativity (GR): the GW amplitude can experience additional effective "friction" as it travels cosmological distances.[6,7] In such scenarios, the GW luminosity distance, $d_L^{\mathrm{GW}}(z)$, differs from the electromagnetic (EM) luminosity distance, $d_L^{\mathrm{EM}}(z)$, even if the background expansion is unchanged. This transforms dark sirens into a direct propagation test: if $d_L^{\mathrm{GW}}(z) \neq d_L^{\mathrm{EM}}(z)$, then any inference that assumes GR propagation is biased by construction.

This distinction matters because propagation and background tests fail differently. Background tests ask which expansion history best explains distances and redshifts under the assumption that distance inference itself is correct. Propagation tests ask whether the mapping from observed GW amplitude to distance is consistent with GR once the same background is held fixed. A propagation anomaly 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.

Dark sirens are GW events without a confirmed EM counterpart. They provide a distance measurement, but their redshift is uncertain and must be inferred statistically, for example by cross-correlating the GW sky localisation volume with galaxy catalogues. This makes dark-siren analyses especially sensitive to selection and catalogue systematics. The key question is therefore not whether a modified-propagation model can fit a particular catalogue better, but whether an observed preference survives the best available internal calibration and stress testing.

Here we adopt a deliberately conservative strategy: we evaluate a fixed modified-propagation history against a fixed GR baseline using a predictive score, and we calibrate that score using matched injections. This produces a statement that is both strong and bounded. It is strong because the probability of obtaining the observed score under the tested GR-truth generator is directly measured to be extremely small. It is bounded because the calibration can only falsify the particular GR-null generator that was tested; it cannot exclude all possible unmodelled catalogue and selection failures. The scientific value is therefore diagnostic: it quantifies how far current dark-siren analyses can be pushed before propagation freedom becomes a necessary ingredient of the model family.

A calibrated propagation anomaly

We analyse 36 GWTC-3 dark sirens from the O3 observing run[8] with a host-incompleteness-marginalised galaxy-catalogue likelihood using GLADE+.[9] Rather than treating the propagation sector as a free phenomenological function, we score a fixed modified-propagation history against an internal GR baseline. The comparison is performed using a joint posterior-predictive log score over the full event ensemble, and we report the score difference $\Delta\mathrm{LPD}_{\mathrm{tot}}$ between the modified-propagation hypothesis and GR (Methods).

At a high level, the score rewards models that assign high probability to the observed set of GW events while accounting consistently for selection. This matters because a model can appear to fit better simply by shifting probability mass into regions where the detector is more sensitive, unless the selection normalisation is handled correctly. We therefore treat selection as part of the model: the likelihood for each event is evaluated together with a selection normalisation, and the same construction is used for real data and for simulated catalogues.

The observed catalogue yields $\Delta\mathrm{LPD}_{\mathrm{tot}} = +3.670$ in favour of the modified-propagation model. In this fixed scoring setup, $\exp(\Delta\mathrm{LPD})$ acts as a proxy for how much more predictive the modified model is than GR, corresponding here to a factor of about 39. Interpreting this number as a universal evidence ratio would be inappropriate: it is a score within a specific pipeline and event selection, designed to enable calibrated comparisons to matched null simulations rather than to arbitrary model classes. The strength of the result therefore comes from explicit calibration.

The score also decomposes naturally into a data term and a selection term. In the updated O3 rerun, the observed preference splits into $\Delta\mathrm{LPD}_{\mathrm{data}} = +2.670$ and $\Delta\mathrm{LPD}_{\mathrm{sel}} = +1.000$ (Methods). This decomposition is important for interpretation: it shows that the preference is not a pure selection artefact, but it also quantifies that selection contributes at an order-unity level. Any attempt to explain the anomaly within GR must therefore reproduce both the data-fit component and the selection component of the score.

We calibrate the GR false-alarm behaviour with a matched injection ensemble: 512 GR-consistent injection catalogues processed through the same scoring pipeline and event ensemble definition. Under this calibrated GR-truth generator, the score distribution is centred at -0.839 with width 0.240 and has maximum $+0.076$ (Methods). None of the 512 GR-consistent catalogues reaches the observed scale (0/512 with $\Delta\text{LPD} \geq 3$), which makes the observed value difficult to attribute to a routine statistical fluctuation within the tested null.

Figure 2 shows the calibration directly: the observed score lies far outside the simulated GR ensemble. This is the central result of the paper because it converts an otherwise ambiguous model preference into a calibrated tension. Figure 1 provides a complementary, more physical visualisation. We reconstruct effective distance-modulus residuals, $\Delta\mu \equiv \mu_{\text{GW,obs}} - \mu_{\text{EM}}$, for each event using the galaxy-weighted redshift distribution and the GW distance posterior, and compare them to the model-implied propagation residual band. These points are not independent direct measurements of $R(z)$; they are a compact diagnostic that makes it easier to see whether the preferred propagation history is qualitatively consistent with the data.

The residual representation also clarifies what the modified-propagation hypothesis is asserting. In friction-like propagation models, the GW amplitude decays more strongly than in GR, so events appear dimmer and therefore farther at fixed redshift. That corresponds to $\Delta\mu > 0$ at higher redshift. The reconstructed residual pattern in Fig. 1 is qualitatively consistent with that behaviour within the broad uncertainties of dark sirens, which helps to rule out a purely numerical explanation in which the score is driven by a pathological subset of the likelihood implementation.

Channel tests and stress tests

Dark sirens conflate several ingredients: sky localisation, a galaxy catalogue, distance posteriors, and a selection function. A raw preference for one model over another is therefore not informative unless it can be localised to a plausible physical channel and shown to survive stress testing. We perform two complementary mechanism controls.

These controls address a practical concern: because dark sirens use galaxy catalogues, a model preference could be driven by a small number of unusual lines of sight, by a particular overdensity in the catalogue, or by accidental alignment between broad GW sky maps and dense regions of the galaxy catalogue. In that case a model could appear preferred even if the underlying distance–redshift relation were unmodified. Mechanism controls therefore do not aim to explain the signal; they aim to rule out specific misleading mechanisms.

First, we randomise sky coordinates (a sky-rotation null) to destroy any special alignment between individual GW localisations and true large-scale structure, while preserving the distance information and the overall selection treatment. The resulting score distribution remains strongly positive, with mean $\langle \Delta\text{LPD}_{\text{rot}} \rangle \simeq +3.017$ (s.d. 0.091), indicating that the preference is not driven purely by a small number of privileged lines of sight. However, the fully updated observed score remains higher than this rotated ensemble, so the data are not simply indistinguishable from a sky-scrambled catalogue.

The key inference from the rotation test is qualitative: randomising sky position does not collapse the score to zero or to a GR-like value. That is hard to reconcile with an explanation based on a few particularly informative host lines of sight. Instead it points to a more global mismatch: even when the sky information is broken, the catalogue still prefers the same broad propagation shift, suggesting that the isotropic distance distribution and the selection treatment are doing most of the work. The fact that the real catalogue remains more extreme than the rotated ensemble leaves open a secondary contribution from genuine angular information or from an anisotropic systematic.

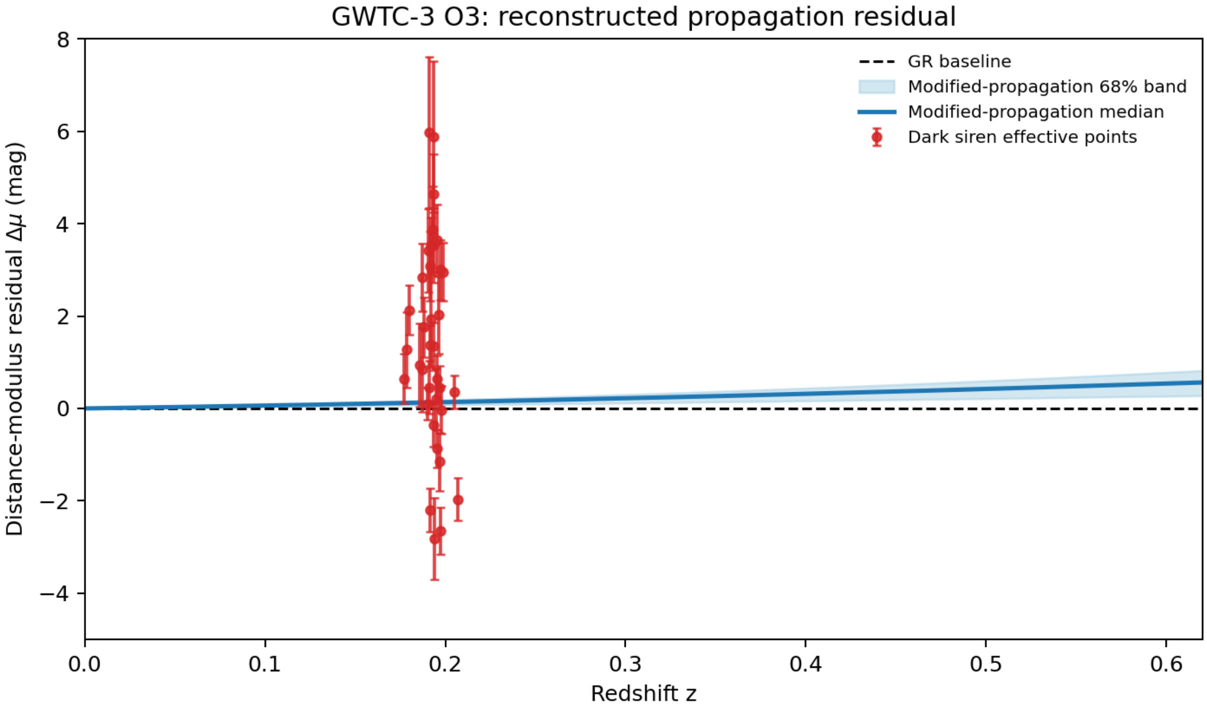

Figure 1: **Reconstructed propagation residuals.** Red points show event-level effective distance-modulus residuals, $\Delta\mu \equiv \mu_{\mathrm{GW,obs}} - \mu_{\mathrm{EM}}$, with propagated distance uncertainties. The blue band shows the 68% posterior range of the fixed modified-propagation history; the dashed line is the GR propagation baseline ($\Delta\mu = 0$). Positive $\Delta\mu$ corresponds to $d_L^{\mathrm{GW}} > d_L^{\mathrm{EM}}$ (sources appear dimmer or farther than GR propagation predicts).

Second, we split the information content into a distance–redshift component and a sky-only component. Distance-only weighting retains most of the preference ($\Delta\mathrm{LPD} \simeq +2.995$), while a sky-only weighting gives a smaller contribution ($\Delta\mathrm{LPD} \simeq +0.969$). Taken together, these controls point to a dominant driver in the isotropic distance–redshift distribution and its selection calibration, rather than in rare angular host associations.

This split is also a useful sanity check. If a signal were driven primarily by sky localisation geometry, one would expect the sky-only channel to dominate. If it were driven primarily by a mismatch in how distance information is interpreted (for example through propagation, calibration, or distance-prior effects), one would expect the distance-only channel to dominate. The observed ordering therefore supports the interpretation that the anomaly is fundamentally a distance–redshift sector effect, with sky information contributing but not controlling the preference.

We then test whether standard GR-consistent nuisance choices can reproduce the observed amplitude. Two stress tests are particularly constraining. A fixed-power injection grid (five injected propagation strengths, 256 replicates per scale) shows a monotonic increase in mean score with injected propagation strength, validating that the statistic responds directionally to true propagation modifications rather than to numerical noise. A nine-variant GR-consistent systematics matrix (128 replicates per variant) shifts the score but does not approach the observed amplitude; the largest variant maximum is +0.678, far below +3.670.

These stress tests are not intended to be exhaustive. They are intended to answer a specific

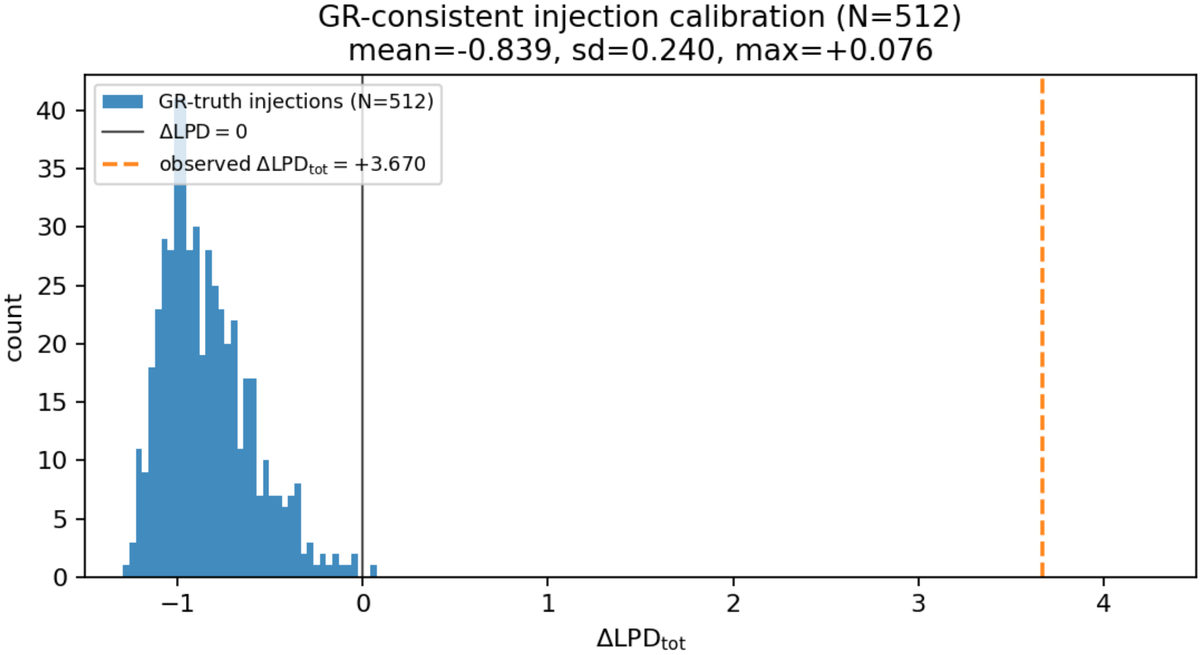

Figure 2: **Calibrated GR-null falsification.** Histogram of $\Delta\mathrm{LPD}_{\mathrm{tot}}$ for 512 GR-consistent injection catalogues processed through the same pipeline. The dashed line marks the observed value. None of the GR-consistent injections reaches the observed scale, rejecting the GR-null generator within the injection assumptions.

question: within the tested family of "reasonable" GR-consistent nuisances, can the observed score be reproduced without invoking a propagation deformation? The answer is no. This narrows the plausible explanations to two: either the modified-propagation history is picking up real propagation physics, or there exists an unmodelled catalogue or selection failure outside the tested family that can generate a shift of several log-units in predictive score. In a dark-siren context, such a failure could arise from subtle correlations between detector selection and galaxy-catalogue completeness, from redshift-dependent incompleteness that is not captured by the adopted marginalisation, or from a mismatch between the injection-trained selection model and the true detection process in the public catalogue.

Adversarial robustness. To narrow the remaining dismissal pathways, we extended the stress-testing programme with adversarial and split-coherence checks using cached-term re-scoring. A flexible nuisance deformation of the injection-calibrated selection function, implemented as bounded multiplicative splines in effective SNR and redshift with a mild mass tilt, does not erase the preference: across tested bounds up to ±60% in this nuisance family, the best-achieved score remains $\Delta\mathrm{LPD}_{\mathrm{tot}} \simeq +3.67$ and never falls below 1 or 0. In the dominant distance–redshift channel (spectral-only re-scoring, which retains most of the baseline preference), global photometric-redshift biases modelled as $\Delta z = b_0 + b_1 z$ can reduce the preference, but do not remove it on the tested grid: even an extreme choice ($b_0 = 0.03$, $b_1 = 0.2$) yields $\Delta\mathrm{LPD} \simeq +1.98$, while completeness-weight tilts at the ±20% level leave $\Delta\mathrm{LPD} \gtrsim +3.61$. Swapping public GWTC-3 parameter-estimation analysis groups for the two highest-leverage events (GW200308_173609 and GW200220_061928) leaves the global score stable at $\Delta\mathrm{LPD} \simeq +3.63$ across IMRPhenomXPHM, SEOBNRv4PHM and mixed

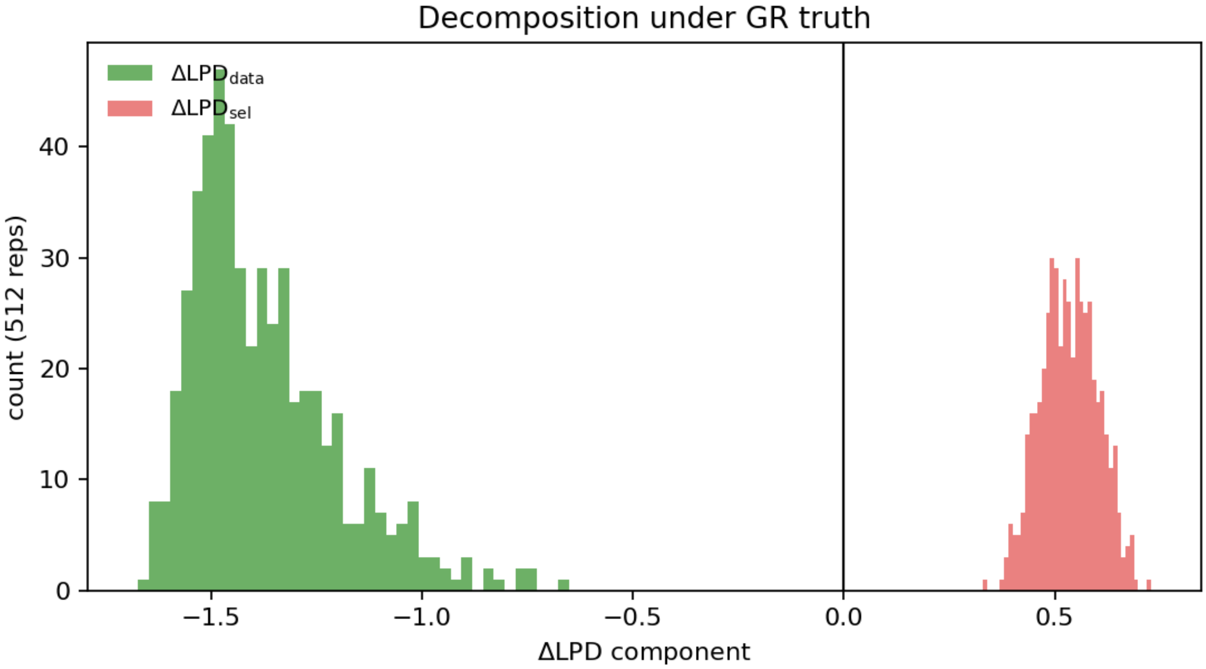

Figure 3: **Null decomposition.** Decomposition of the calibrated GR-null ensemble into data and selection components. The combined GR-null score remains negative even though the selection component contributes positively, illustrating that the null is not merely a selection artefact within the tested generator.

analyses. A binned-redshift scramble null in the same spectral-channel approximation weakens the preference but does not collapse it (null mean $\Delta\mathrm{LPD}_{\mathrm{tot}} \approx +3.38 \pm 0.15$ versus observed $+3.65$, empirical $p \approx 0.05$), indicating some dependence on coherent catalogue-redshift structure while leaving residual degeneracy with modelling assumptions. Finally, strictly non-circular splits show that the preference concentrates at larger inferred distances: in three equal-count bins of the GW luminosity-distance posterior median, $\Delta\mathrm{LPD} \approx -0.52$, $+0.57$ and $+3.55$ from low to high distance, consistent with a cumulative propagation effect.

Implications for late-time inference

Within the tested calibration suite, the simplest statement supported by the data is narrow: GWTC-3 dark sirens prefer the fixed modified-propagation history over the GR propagation baseline by a statistically large calibrated score. The immediate alternative explanation is also narrow: a catalogue or selection mismatch outside the tested nuisance family can still mimic a distance–redshift channel preference. Dark-siren cosmology is inherently exposed to these effects because it relies on a galaxy catalogue and a selection function, and because GW distance posteriors are broad.

For a general scientific audience, the important point is that propagation freedom is not an exotic embellishment. It is a parameterisation of whether the amplitude of a gravitational wave decays with distance exactly as GR predicts. If it does not, then a GW distance inferred under GR is systematically wrong, and that systematic error propagates into any cosmological inference that uses those distances. This is conceptually similar to using a miscalibrated standard candle: even if

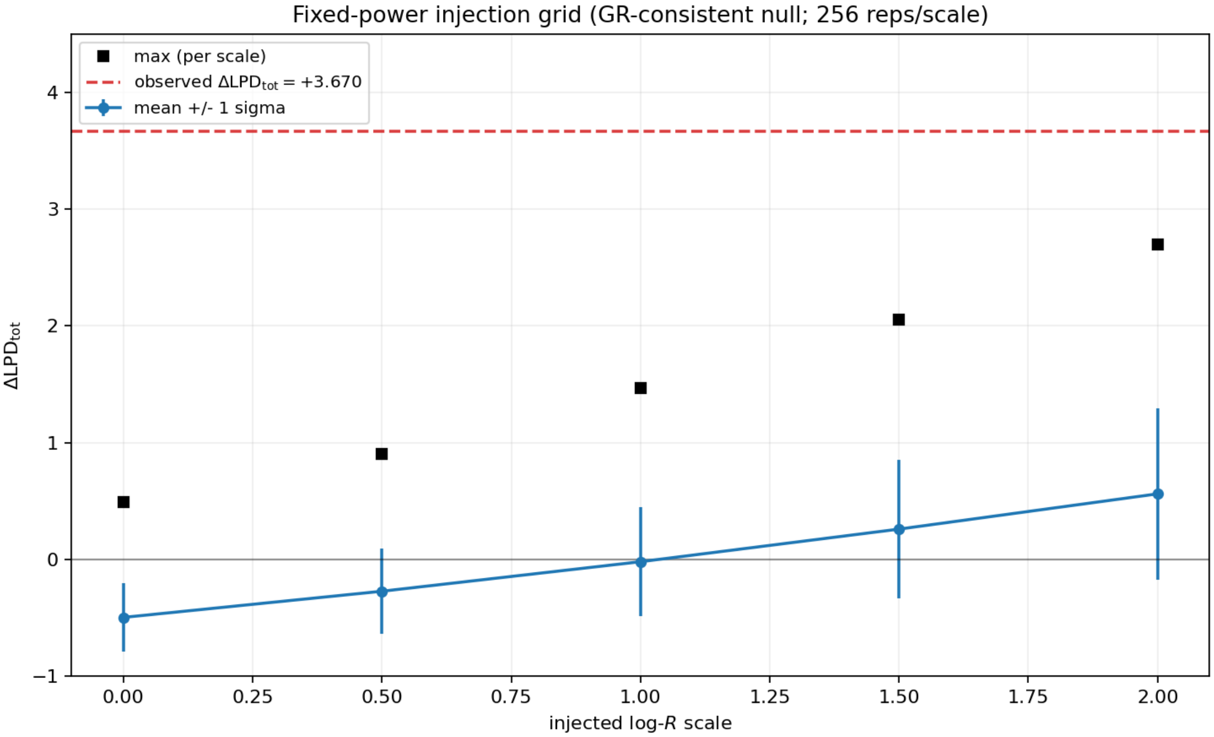

Figure 4: **Injected-power response.** Mean and scatter of $\Delta\mathrm{LPD}_{\mathrm{tot}}$ across a fixed-power injection grid under the GR-consistent generator. The monotonic trend validates score sensitivity to injected propagation strength.

the redshifts are perfect, the inferred expansion history will be biased because the distance scale is biased.

If the preference is physical, it has a direct and testable consequence for the Hubble tension. Many late-time inferences use GW distances either explicitly (standard sirens) or implicitly (cross-calibrations and consistency tests). If $d_L^{\mathrm{GW}}(z)$ is systematically offset from $d_L^{\mathrm{EM}}(z)$, then any analysis that assumes GR propagation will absorb that offset into inferred cosmological parameters. In that sense, GR-locked distance inference can behave like an "invisible wedge": it does not appear as an explicit nuisance parameter, but it distorts the mapping from observed amplitudes to expansion history.

This paper does not claim that GWTC-3 dark sirens resolve the Hubble tension on their own. The result is best viewed as a calibrated inconsistency that identifies a specific pathway by which a distance-inference assumption can bias late-time cosmology. Whether that pathway is realised in nature depends on future cross-checks. One class of cross-checks is internal: re-train the selection function on alternative injection sets, vary the galaxy catalogue and incompleteness modelling more aggressively, and confirm that the calibrated GR-null remains far below the observed score. Another class is external: apply the same pipeline, without retuning, to independent observing runs and independent catalogues. If the preference persists with different instruments, different selection conditions, and different galaxy data products, the interpretation shifts toward propagation physics.

There is also a positive scientific opportunity even if the signal ultimately proves to be systematic. A stress-tested calibrated anomaly is still a valuable audit result: it identifies where the current state of dark-siren cosmology is brittle, and therefore what must be improved to make dark sirens

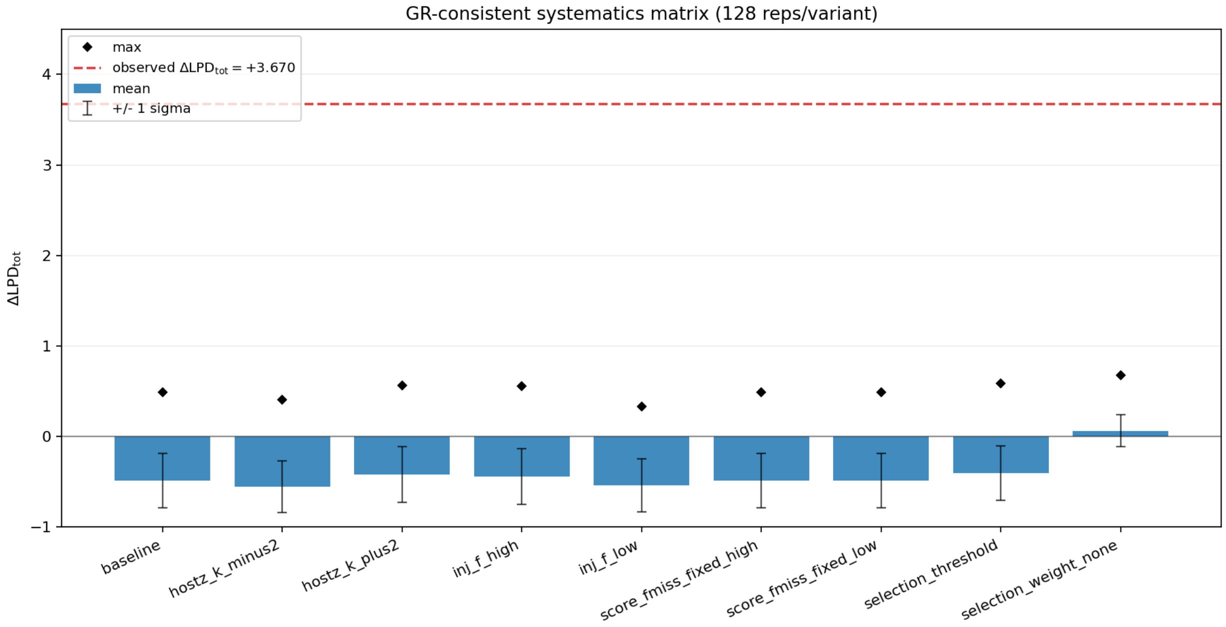

Figure 5: **Systematics matrix.** GR-consistent nuisance-variant matrix (nine variants). Tested nuisance choices shift $\Delta\mathrm{LPD}_{\mathrm{tot}}$ but do not reproduce the observed high-amplitude signal, motivating targeted expansion of the tested systematics family.

a decision-grade tool for late-time inference. The same machinery that detects a propagation-like preference can be turned into an adversarial diagnostic: what specific selection or catalogue perturbation is required to mimic the observed score under GR truth?

The path forward is therefore straightforward in principle. The same calibrated framework can be extended along two axes: (i) broaden the systematics stress family to include more aggressive and physically motivated selection and catalogue perturbations, and (ii) apply the test to independent event sets (for example O4 and beyond) and independent galaxy catalogues. If the signal persists under materially different selection training and catalogue construction, a propagation interpretation becomes harder to dismiss. If, instead, a plausible selection or catalogue perturbation can reproduce the observed amplitude under GR truth, the result becomes a concrete diagnostic for pipeline vulnerability rather than a claim of new gravitational physics.

In summary, GWTC-3 dark sirens already contain enough information to falsify a specific calibrated GR-null generator in this pipeline. The responsible next step is not rhetorical escalation, but targeted enlargement of the calibrated null family and replication on independent data. Either outcome is useful: it will either promote propagation-sector freedom from optional extension to required baseline in late-time inference, or it will deliver a concrete, quantified selection or catalogue failure mode that must be corrected before dark sirens can be used to arbitrate the Hubble tension.

Methods

Data selection

Event selection. This analysis is restricted to GWTC-3 (O1–O3) and explicitly excludes O4a data. O4a operated primarily as a two-detector network (LIGO Hanford and Livingston) because Virgo was absent and KAGRA had a limited duty cycle. In a two-detector configuration, sky localisation is typically degenerate, producing large annuli ("rings") rather than compact triangulated regions. This greatly inflates the galaxy-catalogue search volume and renders the dark-siren likelihood comparatively information-poor, relative to the three-detector triangulation available in GWTC-3.

Data. We use 36 GWTC-3 dark-siren events from the O3 catalogue and public parameter-estimation samples.[8] Host redshifts are treated statistically via a galaxy-catalogue likelihood based on GLADE+.[9]

Propagation hypotheses. In GR, $d_L^{\mathrm{GW}}(z) = d_L^{\mathrm{EM}}(z)$. In modified-propagation scenarios, the GW amplitude can be damped relative to GR, which can be parameterised by a ratio $R(z)$ such that $d_L^{\mathrm{GW}}(z) = R(z)\, d_L^{\mathrm{EM}}(z)$.[6,7] We compare an internal GR baseline to a fixed modified-propagation history (a specified $R(z)$ curve) derived upstream and held fixed during scoring.

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

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

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 $\Delta\mathrm{LPD}_{\mathrm{tot}} = \mathrm{LPD}(\mathrm{prop}) - \mathrm{LPD}(\mathrm{GR})$. This construction uses a log-mean-exp aggregation to avoid numerical underflow and to retain a consistent calibration interpretation.

Selection calibration. The selection normalisation α is calibrated empirically from injections using a logistic selection model trained on an injection set, and applied consistently in both real-data scoring and injection-based null tests.

GR-null injection calibration. We generate 512 GR-consistent injection catalogues matched to the real-data analysis settings and compute $\Delta\mathrm{LPD}_{\mathrm{tot}}$ for each replicate. Summary statistics quoted in the main text (mean, width and maximum) are computed directly from this ensemble.

Mechanism controls. The sky-rotation null applies random rotations to the sky coordinates of each event to destroy any true angular association with large-scale structure. The distance-only versus sky-only split isolates the relative contribution of the distance–redshift information versus angular information in the catalogue likelihood (implementation details follow the same scoring pipeline with alternative likelihood weightings).

Stress tests. The fixed-power grid injects a set of propagation strengths (five scales) and confirms that mean $\Delta\mathrm{LPD}_{\mathrm{tot}}$ increases monotonically with injected strength. The systematics matrix evaluates nine GR-consistent nuisance variants; for each variant we run 128 replicates and report means and maxima.

Code and data availability. Analysis artifacts and summary tables are provided with this repository and mirrored on the associated project archive (Zenodo DOI: 10.5281/zenodo.18635659).

References

References

[1] Verde, L., Treu, T. & Riess, A. G. Tensions between the early and late Universe. *Nat. Astron.* **3**, 891–895 (2019).

[2] Planck Collaboration. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.* **641**, A6 (2020).

[3] Riess, A. G. *et al.* A comprehensive measurement of the local value of the Hubble constant with 1 km s^{-1} Mpc^{-1} uncertainty. *Astrophys. J. Lett.* **934**, L7 (2022).

[4] Schutz, B. F. Determining the Hubble constant from gravitational wave observations. *Nature* **323**, 310–311 (1986).

[5] Abbott, B. P. *et al.* (LIGO Scientific Collaboration & Virgo Collaboration). A gravitational-wave standard siren measurement of the Hubble constant. *Nature* **551**, 85–88 (2017).

[6] Belgacem, E., Dirian, Y., Foffa, S. & Maggiore, M. Modified gravitational-wave propagation and standard sirens. *Phys. Rev. D* **98**, 023510 (2018).

[7] Nishizawa, A. Generalized framework for testing gravity with gravitational-wave propagation. *Phys. Rev. D* **97**, 104037 (2018).

[8] Abbott, R. *et al.* (LIGO Scientific Collaboration, Virgo Collaboration & KAGRA Collaboration). GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run. *Phys. Rev. X* **13**, 041039 (2023).

[9] Dálya, G. *et al.* GLADE+: An extended galaxy catalogue for multimessenger searches. *Mon. Not. R. Astron. Soc.* **514**, 1403–1411 (2022).