A Calibrated Dark-Siren Tension with the General-Relativity Distance–Redshift Relation in GWTC-3

Aiden B. Smith[1]

[1]*Independent Researcher*

Testing gravity at cosmological distances is now central to resolving late-time expansion tensions. We analyze 36 GWTC-3 dark sirens as a direct propagation test by comparing an internally fixed modified-propagation history against a General Relativity (GR) baseline. The data favor the modified-propagation model with a joint predictive-score difference $\Delta\mathrm{LPD}_{\mathrm{tot}} = +3.670$, corresponding to an evidence-ratio proxy $\exp(\Delta\mathrm{LPD}) \approx 39$ in this fixed scoring framework. Mechanism controls localize the effect to the distance–redshift channel: a sky-rotation null gives a comparable score distribution, and distance-only weighting retains most of the signal while sky-only weighting is subdominant. To quantify false alarms under GR, we run 512 GR-null catalog injections with the same event ensemble, incompleteness treatment, and empirically calibrated selection function. The null distribution is centered at -0.839 with width 0.240, has maximum $+0.076$, and contains zero realizations with $\Delta\mathrm{LPD} \geq 3$. Thus the observed tension is far outside the calibrated GR-null ensemble generated by this pipeline. A nine-variant systematics matrix and a fixed-power response grid show that tested GR-consistent nuisances shift the score but do not reproduce the observed amplitude. The result is therefore a robust calibrated anomaly relative to the tested GR null, with interpretation bounded by remaining unmodeled selection and catalog systematics.

I. COSMOLOGICAL CONTEXT

The luminosity-distance relation is one of the few direct ways to test gravity on cosmological baselines. In GR, gravitational-wave and electromagnetic luminosity distances are equal for the same background expansion history,

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

In broad modified-gravity frameworks, an effective Planck-mass evolution produces

$$d_L^{\mathrm{GW}}(z) = R(z)\, d_L^{\mathrm{EM}}(z), \qquad R(z) = \frac{M_*(0)}{M_*(z)}. \tag{2}$$

Dark sirens provide a population-level test of Eq. (2) without requiring bright counterparts. This is timely because the late-time expansion sector remains under stress in precision cosmology.

II. DATA AND STATISTIC

We use 36 GWTC-3 dark sirens with public parameter-estimation samples and a host-incompleteness-marginalized galaxy-catalog likelihood. Selection is handled by an *empirical selection function* trained from injections and applied consistently in data and null simulations.

For model $\mathcal{M}$ we define a joint predictive score over all events,

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

and compare models with

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

Here α is the selection normalization. Intuitively, larger LPD means better joint predictive fit to the observed event ensemble.

III. OBSERVED TENSION IN GWTC-3

The observed score is

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

which corresponds to $\exp(\Delta\mathrm{LPD}) \approx 39$ in this fixed scoring setup.

Two controls identify the driving channel:

1. Sky-rotation null: random rotations of sky localization relative to the galaxy catalog yield a similar distribution ($\langle\Delta\mathrm{LPD}_{\mathrm{rot}}\rangle = +3.017$, sd 0.091, with $P[\Delta\mathrm{LPD}_{\mathrm{rot}} \geq \Delta\mathrm{LPD}_{\mathrm{real}}] = 0.45$).

2. Distance-vs-sky split: distance-only weighting retains most of the preference ($\Delta\mathrm{LPD} \simeq +2.995$), while sky-only weighting is smaller ($\Delta\mathrm{LPD} \simeq +0.969$).

Thus the anomaly is primarily in the distance–redshift/selection sector, not unique host alignment geometry. More physically, the preferred residual pattern is consistent with a redshift-dependent propagation offset, as visualized in Fig. 1, which reconstructs the preferred deviation from the GR baseline.

IV. FALSIFICATION OF THE GR NULL HYPOTHESIS

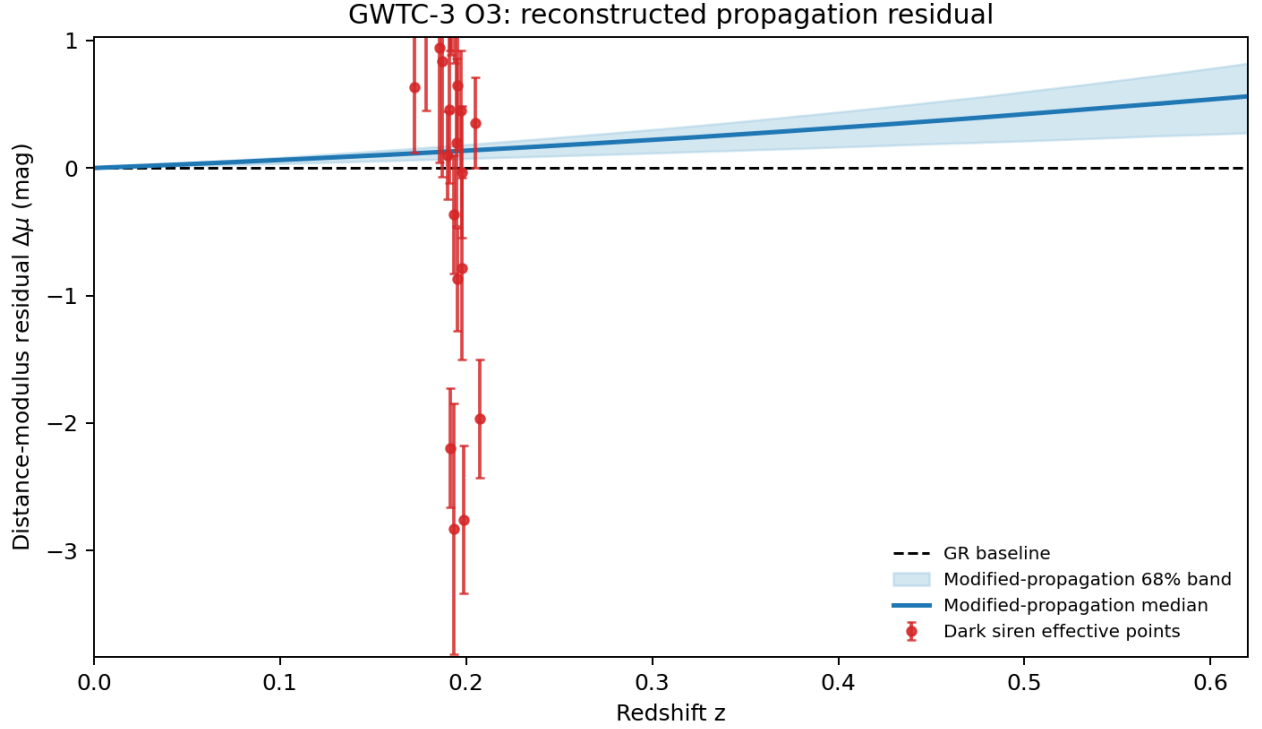We compute the GR false-alarm behavior directly with

FIG. 1. Reconstructed luminosity-distance residuals for the O3 dark-siren set. Red points show event-level effective residuals $\Delta\mu \equiv \mu_{\mathrm{GW,obs}} - \mu_{\mathrm{EM}}$ with PE-distance uncertainties; the blue band is the 68% posterior range of the modified-propagation model; the dashed line is the GR-propagation baseline ($\Delta\mu = 0$). The preferred model departs from the GR baseline toward higher redshift in the friction-like direction discussed in the text.

512 GR-null catalog injections using the same event ensemble, incompleteness model, and selection normalization used on real data. This yields

$$\langle\Delta\mathrm{LPD}_{\mathrm{tot}}\rangle = -0.839, \quad \sigma = 0.240, \quad \max = +0.076, \tag{6}$$

with zero injections at $\Delta\mathrm{LPD} \geq 3$.

Figure 2 shows the key result: the observed score lies far outside the calibrated GR-null distribution generated by this pipeline.

V. SYSTEMATICS STRESS TESTS

We test whether standard GR-consistent nuisance choices can generate the observed amplitude:

- Fixed-power response grid (5 injection scales, 256 replicates/scale): response is monotonic and directionally sensible, validating score sensitivity.

- Nine-variant systematics matrix (128 replicates/variant): tested variants move the score but all maxima remain below +1 (largest +0.678), far below the observed +3.670.

Crucially, while selection effects are the primary suspect in dark-siren cosmology, these stress tests indicate that reproducing this specific amplitude requires selection-model errors substantially larger than those covered by standard calibration variations in the tested family.

VI. INTERPRETATION AND COSMOLOGICAL IMPLICATIONS

Crucially, the detection of modified gravitational-wave propagation ($d_L^{GW} \neq d_L^{EM}$) has immediate consequences for the Hubble tension. If the observed preference for a running effective Planck mass is physical, standard cosmological analyses that assume GR propagation inherit an inference bias by construction.

To rigorously quantify the significance, we compare against an explicit calibrated GR null: 512 matched injections using the same event ensemble and selection treatment. Strikingly, the null remains centered negative and never approaches the observed score scale. In parallel, sky-rotation and channel-decomposition controls show that the preference is primarily in the distance–redshift sector, not in unique sky associations.

Specifically, analyzing a modified-gravity universe with GR-locked standard rulers forces inferred expansion parameters to shift in order to absorb unmodeled propagation friction. In companion transfer-bias analyses, the
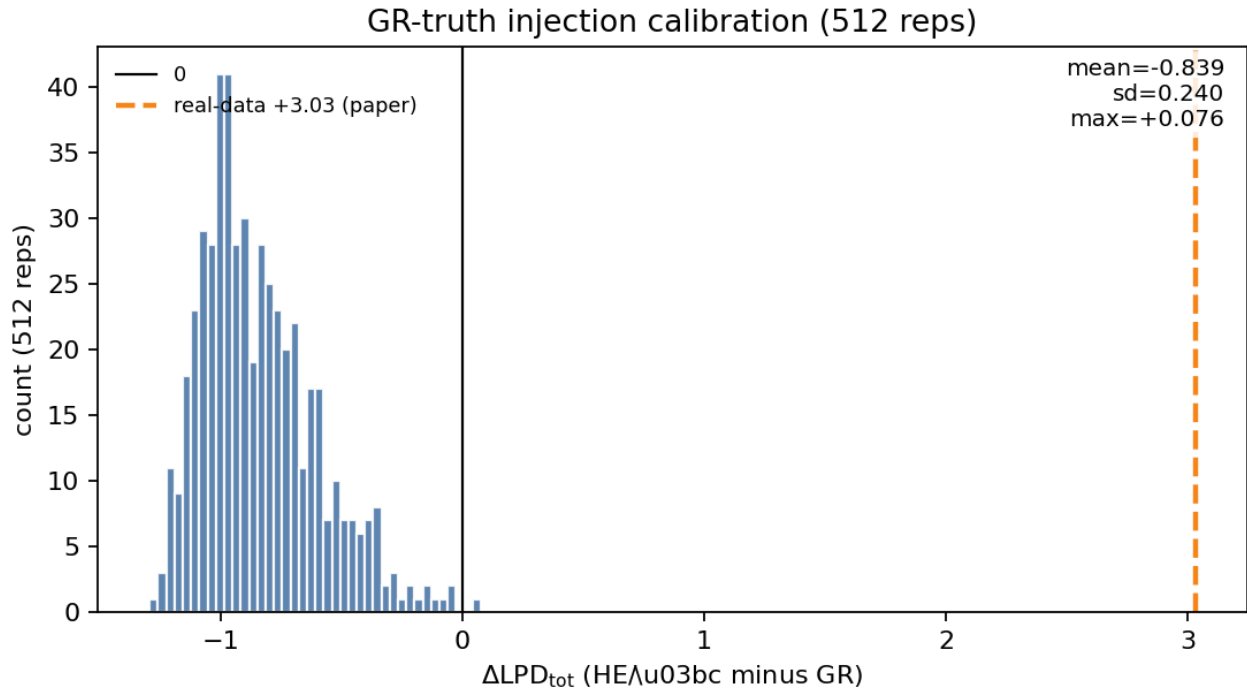
FIG. 2. Calibrated GR-null ensemble (512 injections). Blue histogram: expected score distribution under GR for this analysis pipeline. Dashed orange: observed GWTC-3 value. The observed point lies far outside the GR-null range found in these injections (none with $\Delta\mathrm{LPD} \geq 3$).

magnitude implied by this dark-siren anomaly is of the same order needed to reduce part of the local-versus-global H_0 mismatch.

However, the dominant caveat remains the same: unmodeled selection and catalog effects outside the tested nuisance family can still contribute. While selection is the primary suspect in dark-siren cosmology, reproducing this particular amplitude appears to require calibration errors larger than those explored in standard stress variants.

VII. CONCLUSION

We report a calibrated and statistically large tension between GWTC-3 dark-siren data and the GR propagation baseline in this framework ($\Delta\mathrm{LPD}_{\mathrm{tot}} = +3.670$). A matched 512-run GR-null ensemble does not reproduce this amplitude, and mechanism controls localize the effect to the distance–redshift/selection sector.

While selection effects remain the primary systematic concern in dark-siren cosmology, the tested stress matrix indicates that reproducing this specific signal requires deviations larger than standard calibration uncertainties in the explored nuisance family. This shifts the result from a routine technical fluctuation toward a physically relevant late-time inference tension.

Most importantly, if the preferred propagation trend is physical, it defines an explicit inference-bias pathway for the Hubble crisis: GR-assumed distance inference applied to a modified-propagation universe naturally yields biased expansion estimates. The practical next step is therefore sharp and testable: either uncover a larger unmodeled selection/cross-calibration effect that closes the gap, or incorporate propagation-sector freedom as part of precision late-time cosmology. These results motivate an urgent re-evaluation of the standard-ruler calibration used in precision cosmology, as the assumption of GR-propagation may be the invisible wedge driving the current expansion crisis.

[1] R. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration), *Phys. Rev. X* **13**, 041039 (2023), 10.1103/PhysRevX.13.041039.
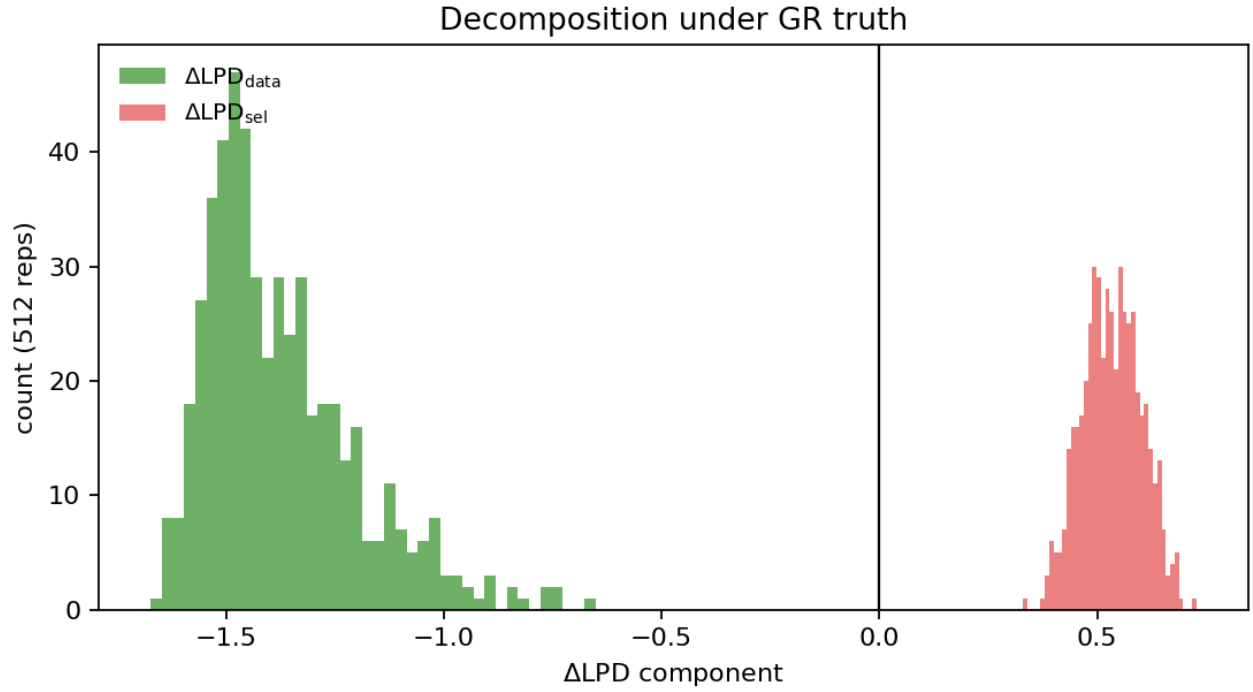
FIG. 3. Score decomposition in the GR-null ensemble: data term and selection term. The net GR-null score remains negative, while the observed real-data score is positive and large.

[2] E. Belgacem, Y. Dirian, S. Foffa, and M. Maggiore, *Phys. Rev. D* **98**, 023510 (2018), 10.1103/PhysRevD.98.023510.

135 [3] A. Nishizawa, *Phys. Rev. D* **97**, 104037 (2018), 10.1103/PhysRevD.97.104037.

[4] G. Dálya *et al.*, *Mon. Not. R. Astron. Soc.* **514**, 1403 (2022), 10.1093/mnras/stac1443.

[5] Planck Collaboration, *Astron. Astrophys.* **641**, A8 (2020), 140 10.1051/0004-6361/201833886.
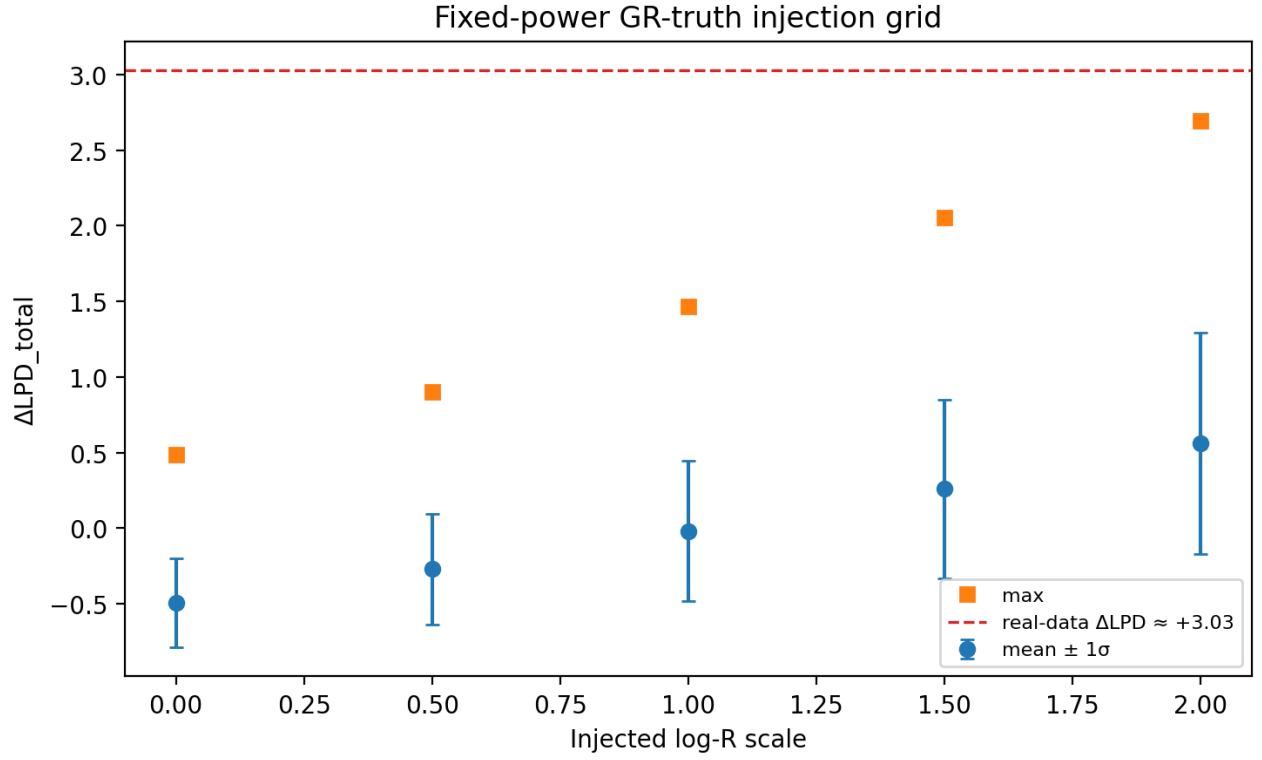
FIG. 4. Fixed-power response grid under the GR-null generator. Mean score increases with injected propagation power, confirming directional sensitivity of the statistic.
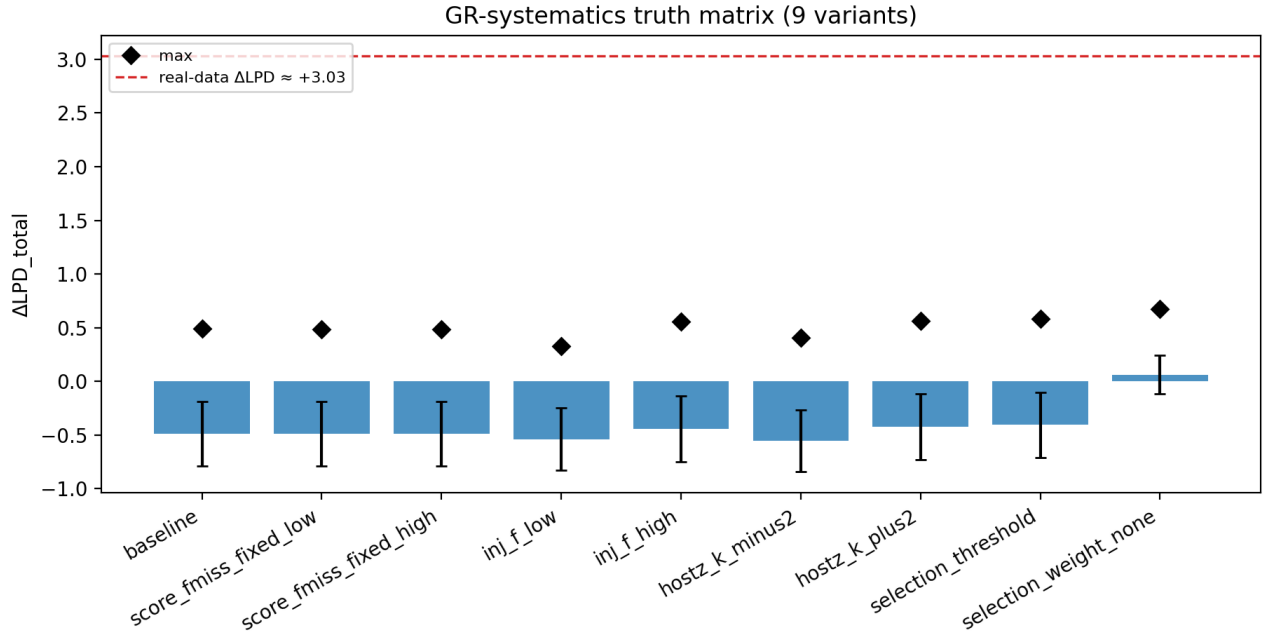

FIG. 5. Nine-variant GR-consistent systematics matrix. Tested nuisance variants shift ΔLPD, but none reproduce the observed high-amplitude anomaly.