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,

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

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.

IV. FALSIFICATION OF THE GR NULL HYPOTHESIS

We compute the GR false-alarm behavior directly with 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 1 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.

VI. INTERPRETATION

The analysis supports a clear statement: *within the tested null and nuisance families, the GWTC-3 dark-siren population shows a calibrated anomaly relative to GR.* This is stronger than a generic software artifact claim, because the anomaly survives explicit null simulations and stress tests.

At the same time, this is not yet a unique microphysical identification. The dominant channel is distance–redshift/selection, so unmodeled catalog or selection effects outside the tested family can still contribute. The correct interpretation is therefore a high-priority cosmological tension that motivates expanded systematics-truth campaigns and independent datasets.

VII. CONCLUSION

GWTC-3 dark sirens show a large predictive-score tension with the GR distance–redshift relation in this framework ($\Delta\mathrm{LPD}_{\mathrm{tot}} = +3.670$). A 512-run GR-null ensemble, matched to the same analysis pipeline, does not reproduce this amplitude. Mechanism tests localize the effect to the distance–redshift/selection sector, and tested GR-consistent systematics remain sub-critical. The result is a robust calibrated anomaly relative to the tested GR null, with physical interpretation limited primarily by remaining unmodeled selection and catalog systematics.

ACKNOWLEDGMENTS

This work used public GWTC-3 products and publicly available galaxy-catalog resources. Code and analysis artifacts are archived at Zenodo (DOI: 10.5281/zenodo.18604204).

[1] R. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration), *Phys. Rev. X* **13**, 041039 (2023), 10.1103/PhysRevX.13.041039.
[2] E. Belgacem, Y. Dirian, S. Foffa, and M. Maggiore, *Phys. Rev. D* **98**, 023510 (2018), 10.1103/PhysRevD.98.023510.
[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), 10.1051/0004-6361/201833886.

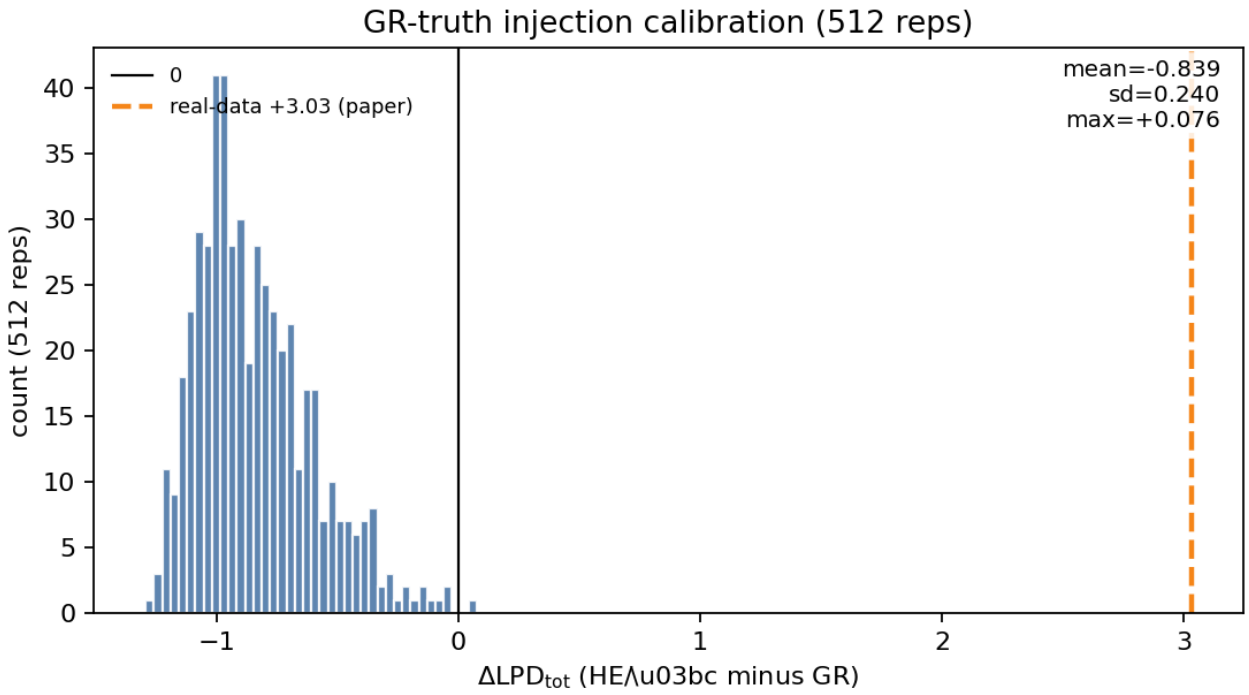

FIG. 1. Calibrated GR-null ensemble (512 injections). Blue histogram: expected score distribution under GR for this analysis pipeline. Dashed orange: observed GWTC-3 value. The observed point lies far outside the GR-null range found in these injections (none with $\Delta\text{LPD} \geq 3$).

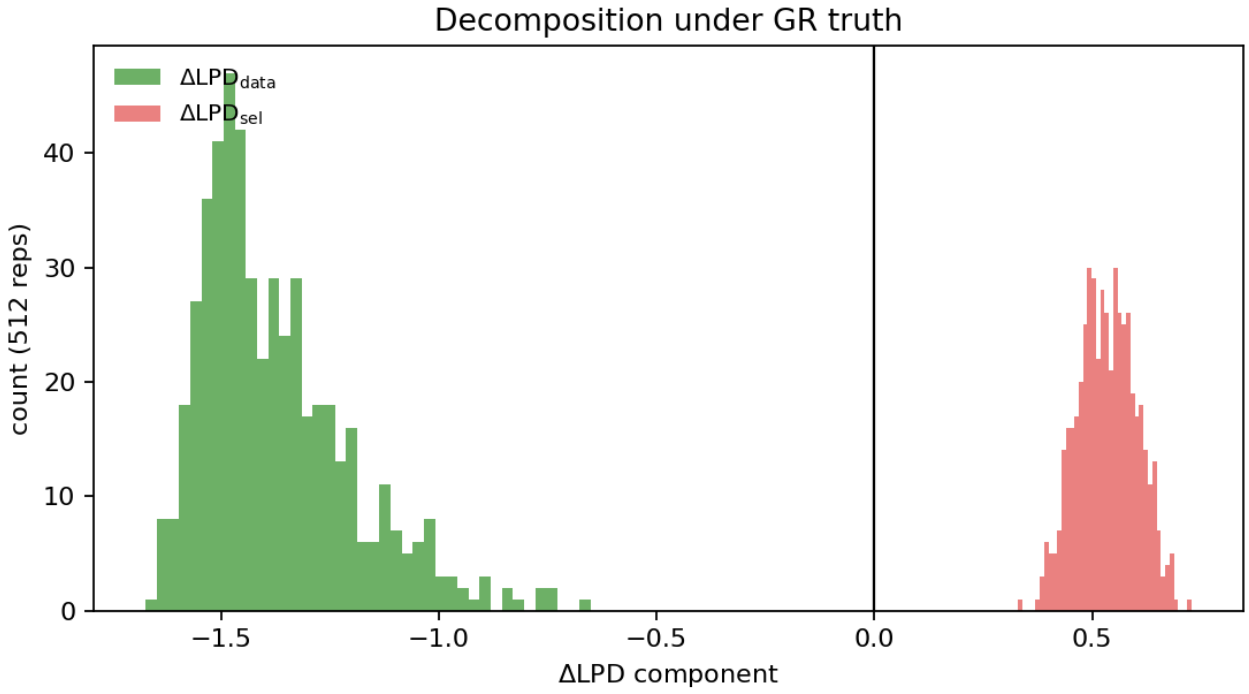

FIG. 2. Score decomposition in the GR-null ensemble: data term and selection term. The net GR-null score remains negative, while the observed real-data score is positive and large.

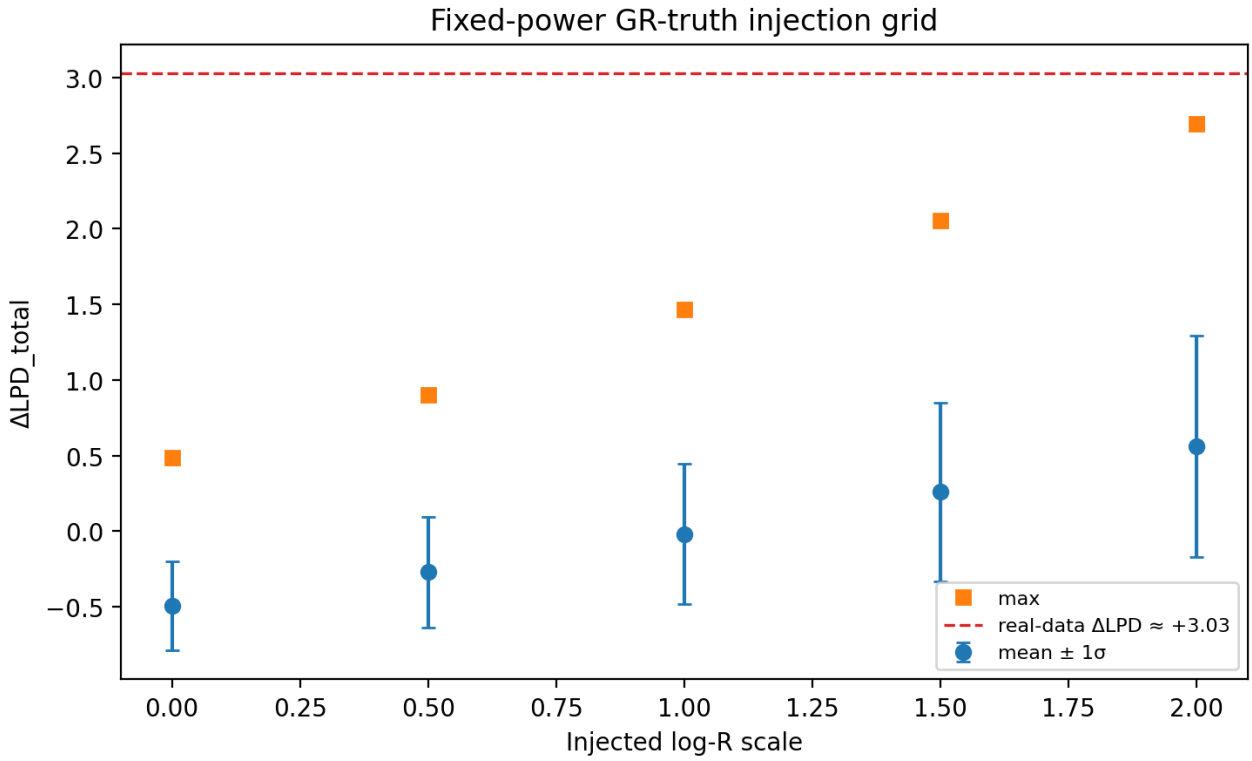

FIG. 3. Fixed-power response grid under the GR-null generator. Mean score increases with injected propagation power, confirming directional sensitivity of the statistic.

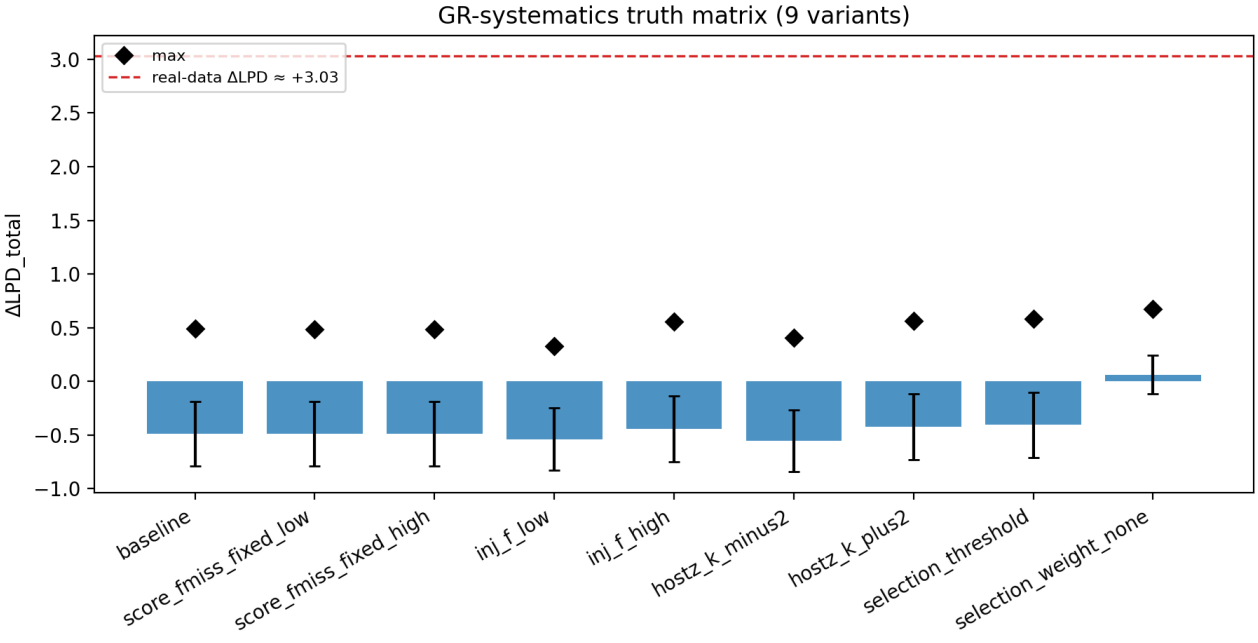

FIG. 4. Nine-variant GR-consistent systematics matrix. Tested nuisance variants shift ΔLPD, but none reproduce the observed high-amplitude anomaly.