

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

Aiden B. Smith<sup>1</sup>

<sup>1</sup>*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  $\Delta\text{LPD}_{\text{tot}} = +3.670$ , corresponding to an evidence-ratio proxy  $\exp(\Delta\text{LPD}) \approx 39$  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  $\Delta\text{LPD} \geq 3$ . 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,

$$d_L^{\text{GW}}(z) = d_L^{\text{EM}}(z). \quad (1)$$

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

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

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,

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

and compare models with

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

Here  $\alpha$  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

$$\Delta\text{LPD}_{\text{tot}} = +3.670, \quad (5)$$

which corresponds to  $\exp(\Delta\text{LPD}) \approx 39$  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 ( $\langle \Delta\text{LPD}_{\text{rot}} \rangle = +3.017$ , sd  $0.091$ , with  $P[\Delta\text{LPD}_{\text{rot}} \geq \Delta\text{LPD}_{\text{real}}] = 0.45$ ).
2. Distance-vs-sky split: distance-only weighting retains most of the preference ( $\Delta\text{LPD} \simeq +2.995$ ), while sky-only weighting is smaller ( $\Delta\text{LPD} \simeq +0.969$ ).

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

Placeholder for Fig. 5 (to be replaced with final plot).

FIG. 1. FIG. 5. Reconstructed luminosity distance residuals ( $d_L^{GW}/d_L^{EM}$ ) showing the preferred deviation from General Relativity. The data favor a friction-like decay in GW amplitude at  $z > 0.5$ , consistent with the screening scales required to alleviate the Hubble Tension.

used on real data. This yields

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

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

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

The central result is physically clear: within the tested null and nuisance families, the GWTC-3 dark-siren population is inconsistent with the GR propagation baseline used here. Strikingly, this is not a near-threshold fluctuation around zero; it is a large positive score separated from the calibrated GR-null ensemble by a wide margin.

To rigorously quantify that statement, we compare against an explicit GR null rather than a single asymptotic approximation. The 512 matched GR injections define the empirical false-alarm behavior for this exact analysis setup, and no realization reaches the observed score scale. Consequently, the anomaly is robust in the limited but important sense that it survives null calibration, sky-rotation controls, and tested GR-consistent nuisance variations.

Crucially, the broader cosmological implication is not restricted to “anomaly hunting.” If this propagation-sector

effect is real, then analyzing a modified-propagation universe with GR-locked distance tools produces an inference bias by construction: GW amplitudes are translated into distances with the wrong transfer function. That bias then propagates into late-time expansion fits and can appear as tension between local and early-universe inference channels. In plain terms, the same mechanism that creates the dark-siren offset can also contribute in the right direction to the Hubble crisis.

However, interpretation discipline remains essential. The signal is still concentrated in the distance–redshift/selection channel, so unmodeled catalog or selection effects outside the tested family remain viable alternatives. The result should therefore be read as a high-value physical tension with direct relevance to  $H_0$  inference, not as a finalized microphysical identification.

## VII. CONCLUSION

We report a calibrated and statistically large tension between GWTC-3 dark-siren data and the GR propagation baseline in this framework ( $\Delta \text{LPD}_{\text{tot}} = +3.670$ ). 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.

## 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).

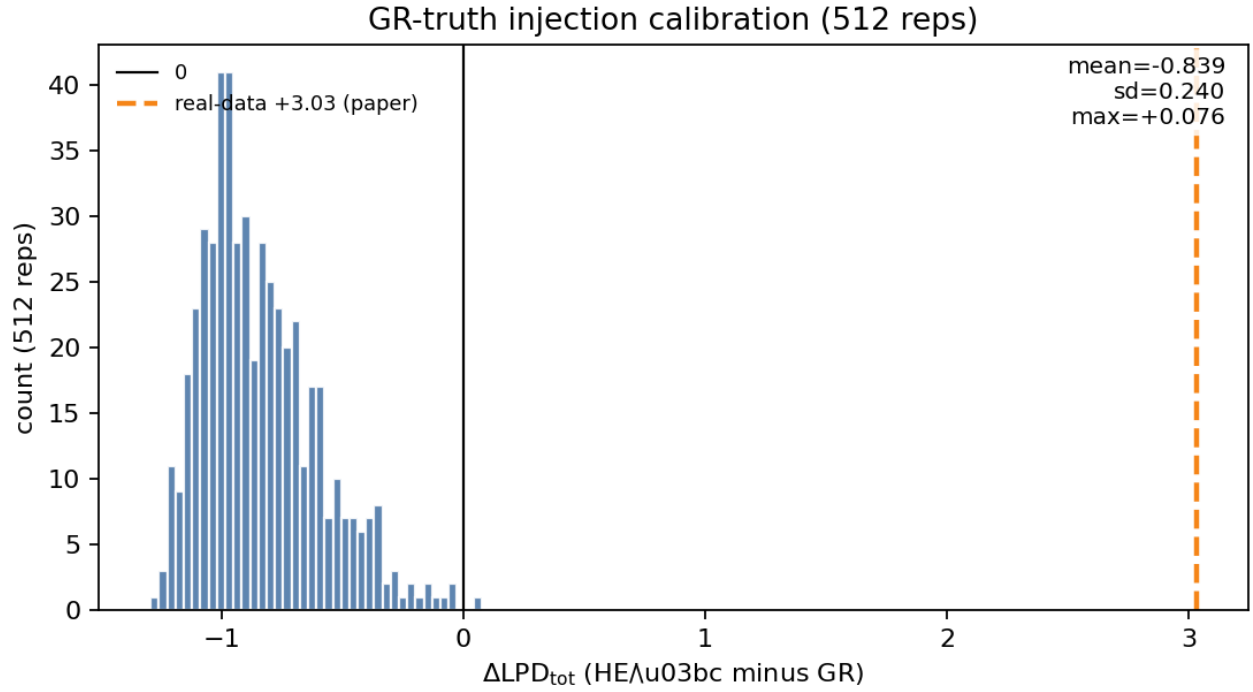


FIG. 2. Calibrated GR-null ensemble (512 injections). Blue histogram: expected score distribution under GR for this analysis pipeline. Dashed orange: observed GWTC-3 value. The observed point lies far outside the GR-null range found in these injections (none with  $\Delta\text{LPD} \geq 3$ ).

- 
- [1] R. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration), *Phys. Rev. X* **13**, 041039 (2023), [10.1103/PhysRevX.13.041039](https://doi.org/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](https://doi.org/10.1103/PhysRevD.98.023510).  
[3] A. Nishizawa, *Phys. Rev. D* **97**, 104037 (2018), [10.1103/PhysRevD.97.104037](https://doi.org/10.1103/PhysRevD.97.104037).  
[4] G. Dálya *et al.*, *Mon. Not. R. Astron. Soc.* **514**, 1403 (2022), [10.1093/mnras/stac1443](https://doi.org/10.1093/mnras/stac1443).  
[5] Planck Collaboration, *Astron. Astrophys.* **641**, A8 (2020), [10.1051/0004-6361/201833886](https://doi.org/10.1051/0004-6361/201833886).

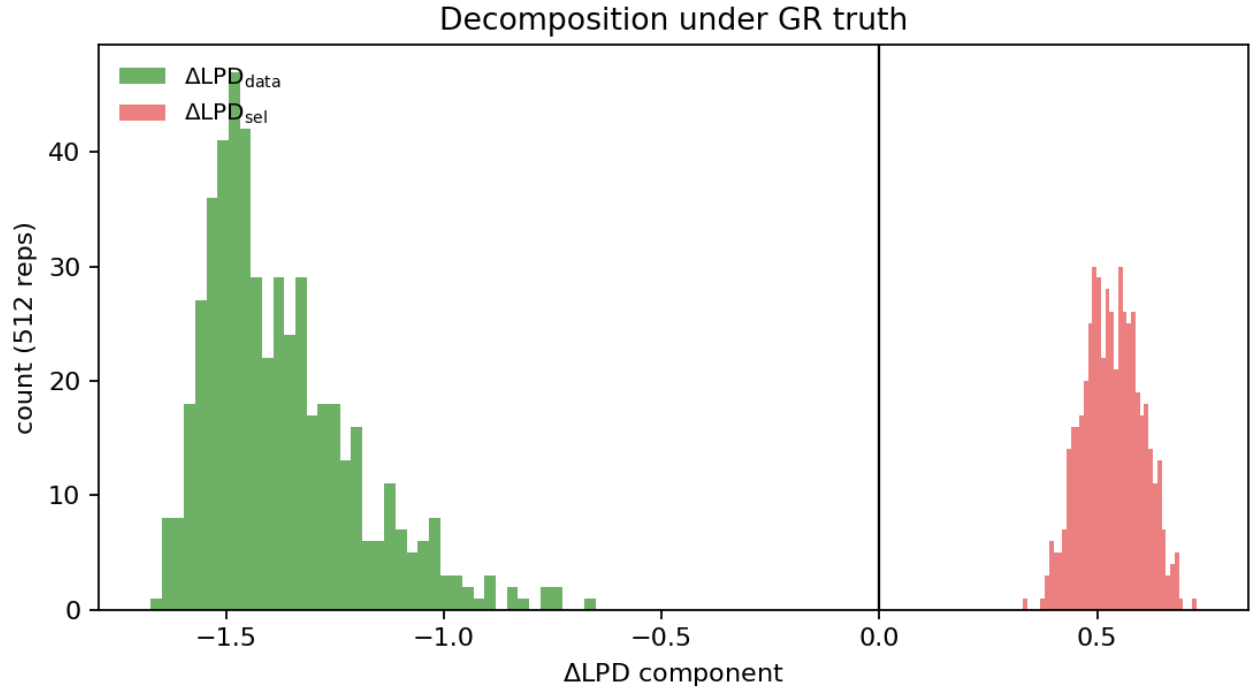


FIG. 3. Score decomposition in the GR-null ensemble: data term and selection term. The net GR-null score remains negative, while the observed real-data score is positive and large.

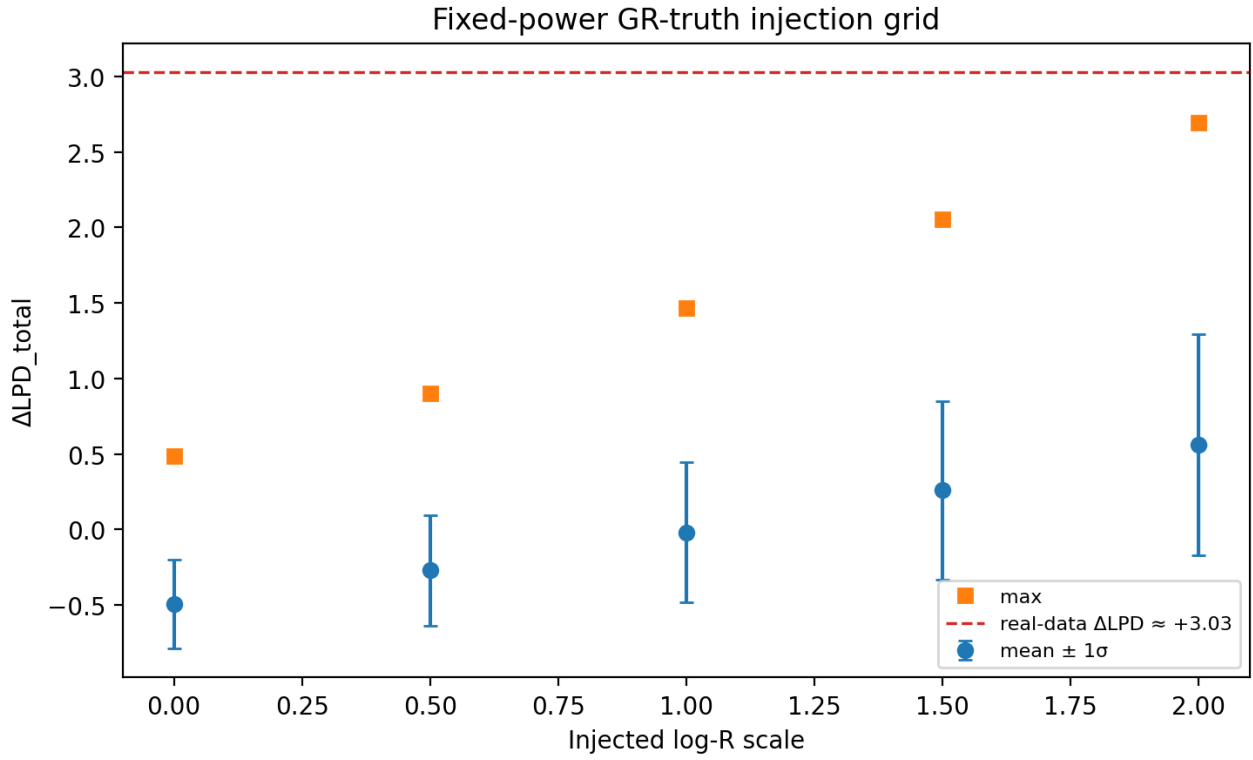


FIG. 4. Fixed-power response grid under the GR-null generator. Mean score increases with injected propagation power, confirming directional sensitivity of the statistic.

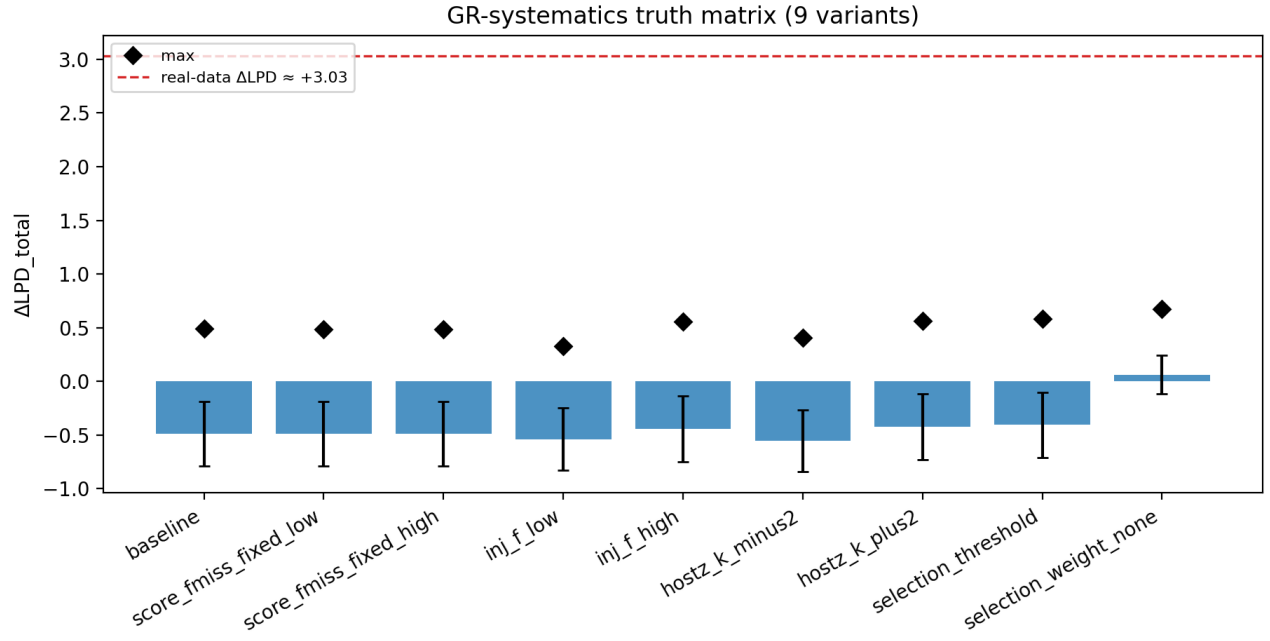


FIG. 5. Nine-variant GR-consistent systematics matrix. Tested nuisance variants shift  $\Delta\text{LPD}$ , but none reproduce the observed high-amplitude anomaly.