

# Tension in GWTC-3 Dark-Siren Cosmology: A Calibrated Search for Modified GW Propagation\*

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

<sup>1</sup>Independent Researcher

We test modified gravitational-wave propagation with 36 GWTC-3 dark sirens using a host-incompleteness-marginalized galaxy-catalog likelihood. A fixed damping model,  $d_L^{\text{GW}} = R(z) d_L^{\text{EM}}$ , is scored against internal GR ( $R \equiv 1$ ) with posterior-predictive log scores, explicit selection normalization, and PE distance-prior removal. In an updated O3 rerun with an injection-trained logistic selection model, we find  $\Delta\text{LPD}_{\text{tot}} = +3.670$  ( $\Delta\text{LPD}_{\text{data}} = +2.670$ ,  $\Delta\text{LPD}_{\text{sel}} = +1.000$ ). Sky rotations give  $\langle \Delta\text{LPD}_{\text{rot}} \rangle = +3.017$  (sd 0.091) and  $P(\Delta\text{LPD}_{\text{rot}} \geq \Delta\text{LPD}_{\text{real}}) = 0.45$ , indicating the signal is not driven by unique host alignments. GR-truth injections (512 replicates) return mean  $-0.839$ , sd 0.240, and max  $+0.076$  (none  $\geq 3$ ). Tested GR-truth systematics remain far below the real-data score (max  $\leq +0.678$ ). The result disfavors a generic numerical artifact but does not yet uniquely identify modified gravity, because residual catalog/selection mismodeling can still mimic a spectral-channel preference.

**Keywords:** gravitational waves; cosmology: observations; methods: statistical; catalogs.

dominant distance-distribution (spectral) channel. The main result is a statistically interesting tension that can reflect either modified propagation or residual catalog/selection mismodeling.

## I. INTRODUCTION

In General Relativity (GR), gravitational waves (GWs) propagate such that the GW luminosity distance equals the electromagnetic (EM) luminosity distance for the same background expansion history. In many beyond-GR scenarios, however, the GW amplitude can experience an additional friction-like term during propagation, yielding a redshift-dependent ratio

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

A broad class of effective-field-theory constructions predicts  $R(z) \neq 1$  through an evolving effective Planck mass  $M_*(z)$  (equivalently, an evolving effective Newton coupling),

$$R(z) = \frac{M_*(0)}{M_*(z)}. \quad (2)$$

In the minimal running- $M_*$  embedding used here, the reconstructed horizon-entropy slope deformation  $\mu(A) \equiv G_{\text{eff}}(A)/G_N$  implies  $M_*^2(z) \propto 1/\mu(A(z))$  and therefore  $R(z) = \sqrt{\mu(A(z))/\mu(A(0))}$  (see, e.g., Belgacem et al. 2, Nishizawa 4).

Statistical dark sirens (no unique host identification) provide an out-of-sample probe of  $R(z)$  by comparing the GW distance posterior to a host-galaxy catalog and a selection-corrected population model. Here we report a posterior-predictive score comparison between a fixed propagation history inferred from an external reconstruction and an internal GR baseline, together with a GR-truth catalog-injection calibration that stress-tests the

## II. DATA AND METHODS

### A. Dark-siren sample and galaxy-catalog mixture likelihood

We analyze  $N_{\text{ev}} = 36$  GWTC-3 dark sirens (BBH-dominated), using public LVK parameter-estimation (PE) posterior samples [1]. For each event  $i$ , we evaluate a galaxy-catalog (GLADE+; Dálya et al. 3) mixture likelihood that marginalizes host-catalog incompleteness,

$$p(d_i | \theta, \mathcal{M}) = (1 - f_{\text{miss}}) p_{\text{cat}}(d_i | \theta, \mathcal{M}) + f_{\text{miss}} p_{\text{miss}}(d_i | \theta, \mathcal{M}), \quad (3)$$

where  $f_{\text{miss}}$  is the missing-host fraction marginalized on a fixed grid in the production configuration. The missing-host term adopts a comoving-uniform redshift prior  $p(z) \propto dV_c/dz$  on  $z \in [0, 0.3]$  (matching the production configuration), ensuring a conservative host-marginalized likelihood contribution when the catalog is incomplete.

### B. PE-prior-aware likelihood evaluation

Public PE samples satisfy  $p(\vartheta | d) \propto \mathcal{L}(d | \vartheta) \pi_{\text{PE}}(\vartheta)$  and therefore encode a PE distance prior. To avoid importing the PE prior into the propagation score, we reweight the released samples and divide by an analytic approximation to the PE distance prior (“PE-analytic” removal), yielding a Monte Carlo estimate of the likelihood ratio required by the mixture likelihood. This procedure is applied identically in the real-data analysis and in the GR-truth injection calibration.

\* aidenblakesmithtravel@gmail.com

### C. Posterior-predictive scoring and selection normalization

We compare a fixed propagation model to an internal GR baseline using the joint posterior predictive density (PPD) over the full event set. Let  $\theta$  denote background/propagation parameters drawn from an external reconstruction posterior  $p(\theta | d_{\text{recon}})$ . For a model  $\mathcal{M}$ , define the joint score

$$\text{LPD}(\mathcal{M}) \equiv \log \left[ \frac{1}{N_s} \sum_{j=1}^{N_s} \exp \left( \sum_{i=1}^{N_{\text{ev}}} \log p(d_i | \theta_j, \mathcal{M}) - N_{\text{ev}} \log \alpha(\theta_j, \mathcal{M}) \right) \right], \quad (4)$$

where  $\{\theta_j\}_{j=1}^{N_s}$  are draws from  $p(\theta | d_{\text{recon}})$  and  $\alpha(\theta, \mathcal{M})$  is the standard selection normalization (detection efficiency) computed from an injection-calibrated selection model. We report Intuitively, LPD is a joint predictive-fit score across all events: larger values mean the model assigns higher probability density to the observed dataset.

$$\Delta \text{LPD}_{\text{tot}} \equiv \text{LPD}(\text{prop}) - \text{LPD}(\text{GR}). \quad (5)$$

A +1 shift in  $\Delta \text{LPD}$  corresponds to a multiplicative predictive-density ratio of  $\exp(1) \approx 2.7$ . For diagnostic bookkeeping we also use the decomposition

$$\Delta \text{LPD}_{\text{tot}} = \Delta \text{LPD}_{\text{data}} + \Delta \text{LPD}_{\text{sel}}, \quad (6)$$

where  $\Delta \text{LPD}_{\text{data}}$  is computed by omitting the  $\alpha$  term and  $\Delta \text{LPD}_{\text{sel}}$  isolates the contribution from the selection normalization.

## III. REAL-DATA TENSION AND MECHANISM CONTROLS

### A. Real-data score and sky-rotation null

On the  $N_{\text{ev}} = 36$  GWTC-3 sample, the updated injection-trained logistic-selection rerun yields

$$\Delta \text{LPD}_{\text{tot}} = +3.670, \quad \exp(\Delta \text{LPD}_{\text{tot}}) \approx 39, \quad (7)$$

which indicates a statistically interesting preference for the propagation phenomenology over the internal GR baseline under the PPD construction. Here  $\exp(\Delta \text{LPD})$  is used as a predictive-score Bayes proxy under this fixed scoring setup, not as a full marginal-likelihood evidence ratio over unrestricted model classes. A key diagnostic is a sky-rotation null: we randomly rotate each event's sky localization relative to the galaxy catalog while preserving its distance posterior and re-score the dataset. Under rotations we obtain a distribution of scores with  $\langle \Delta \text{LPD}_{\text{rot}} \rangle = +3.017$  (sd 0.091) and  $P(\Delta \text{LPD}_{\text{rot}} \geq \Delta \text{LPD}_{\text{real}}) = 0.45$ . Thus, the real-data preference is typical under rotations and is not driven by unique host-galaxy alignments.

As a direct robustness check on the selection term implementation, we reran the same O3 configuration with an injection-trained logistic detection model for  $\alpha(\theta, \mathcal{M})$  (“injection\_logit”), replacing the SNR-binned proxy. This rerun gives  $\Delta \text{LPD}_{\text{tot}} = +3.670$  (data +2.670, selection +1.000), showing the positive O3 anomaly persists under a more explicit injection-derived selection model. For continuity with earlier calibration suites, the legacy production configuration (SNR-binned selection) gave  $\Delta \text{LPD}_{\text{tot}} \simeq +3.03$ .

### B. Spectral-only vs. sky-only controls

To isolate the dominant channel behind the preference, we implement two controls. In a spectral-only control we retain the distance/posterior and selection machinery but remove sky information, whereas in a sky-only control we retain sky weighting but suppress distance/redshift leverage. We find that spectral-only retains most of the preference ( $\Delta \text{LPD}_{\text{spectral}} \simeq +2.995$ ), while sky-only is much smaller ( $\Delta \text{LPD}_{\text{sky}} \simeq +0.969$ ). This localizes the anomaly to population-level distance-redshift consistency coupled to selection/incompleteness modeling, rather than to sky-localized host associations.

### C. Hero-event concentration and selection sensitivity

Jackknife removal tests show that the total score is concentrated in a small subset of high-leverage events, led by GW200308\_173609 and then GW200220\_061928. We also find order-unity shifts in  $\Delta \text{LPD}_{\text{tot}}$  under plausible changes to the selection/population modeling (e.g., detection-model hyperparameters and population priors), motivating conservative interpretation and targeted stress-injection campaigns (Section V).

## IV. GR-TRUTH CATALOG-INJECTION CALIBRATION (512 REPLICATES)

### A. Motivation and what is (not) tested

Because the real-data preference is largely sky-independent (Sections 3.1–3.2), the most important immediate question is whether the full analysis pipeline can accidentally generate a large positive  $\Delta \text{LPD}_{\text{tot}}$  under a calibrated GR null due to numerical, bookkeeping, or PE-prior-removal artifacts. We therefore construct a GR-truth catalog-injection suite designed to stress-test the dominant spectral/selection channel.

This calibration does not validate sky-host association physics: the injection generator uses a synthetic, sky-independent PE-like distance likelihood and the scoring is performed in the spectral-only channel. This design matches the empirically dominant mechanism, and the

TABLE I. Posterior-predictive score summary. Real-data and control scores compare the fixed propagation model to the internal GR baseline using the joint posterior-predictive definition in Eq. (4).

Configuration	$\Delta\text{LPD}$ summary
Real data (O3 re-run; injection\_logit selection model)	$\Delta\text{LPD}_{\text{tot}} = +3.670$ ( $\Delta\text{LPD}_{\text{data}} = +2.670$ , $\Delta\text{LPD}_{\text{sel}} = +1.000$ )
Sky-rotation null (distribution)	$\langle \Delta\text{LPD}_{\text{rot}} \rangle = +3.017$ (sd 0.091); $P(\text{rot} \geq \text{real}) = 0.45$
Spectral-only control	$\Delta\text{LPD}_{\text{spectral}} \simeq +2.995$
Sky-only control	$\Delta\text{LPD}_{\text{sky}} \simeq +0.969$
GR-truth catalog injection (512 reps; max 0.076 spectral/selection channel)	$\langle \Delta\text{LPD}_{\text{tot}} \rangle = -0.839$ (sd 0.240);
Fixed-power injection mean grid (5 scales, 256 (scale 0) to +0.562 (scale 2) reps/scale)	rises from $-0.495$ to $+0.562$ (scale 2)
GR-systematics matrix (9 variants, 128 near the legacy +3.03, and far below 3.670)	all variant maxima $\leq +0.678$ (none 185 reps/variant)
Hierarchical check-point (3 variants, 12 aligned reps)	$\langle \Delta\text{LPD}_{\text{tot}} \rangle = -0.548$ (sd 0.252); $\Delta\text{LPD}_{\text{tot}} = +3.027$ ; calibrated tail 0/12

sky-rotation null indicates that sky association is not the primary driver of the real-data score, but the calibration should not be over-interpreted as a full end-to-end validation of sky-localized host inference.

We perform a parametric-bootstrap-style calibration under GR truth ( $R_{\text{true}}(z) \equiv 1$ ), using the same event ensemble and the same posterior draws used in the production analysis. Per replicate: (i) we draw a “truth” background history from  $p(\theta | d_{\text{recon}})$ ; (ii) for each of the 36 template events we sample a true redshift from a cached event-specific redshift support histogram; (iii) we compute  $d_L^{\text{EM}}(z_{\text{true}})$  and set  $d_L^{\text{GW}} = d_L^{\text{EM}}$ ; (iv) we generate a synthetic PE-like distance likelihood with event-dependent width; and (v) we score the synthetic dataset under the propagation model and the GR baseline using the same incompleteness mixture and selection normalization as in Eq. (4).

### C. Calibration results

Figure 1 shows the GR-truth distribution of  $\Delta\text{LPD}_{\text{tot}}$  for 512 replicates, with the real-data value marked. Under GR truth we find mean  $-0.839$ , sd 0.240, and maximum  $+0.076$  in 512 replicates; none reach  $\Delta\text{LPD}_{\text{tot}} \geq 3$ . Figure 2 shows the decomposition into data and selection components: on average  $\langle \Delta\text{LPD}_{\text{data}} \rangle = -1.374$  and

$\langle \Delta\text{LPD}_{\text{sel}} \rangle = +0.534$ , so the selection term partially offsets the data term but does not reverse the net preference under GR truth.

## V. DISCUSSION

The GR-truth calibration materially reduces the likelihood that the real-data preference is a generic numerical artifact that would also appear under GR truth (e.g., PE-prior-removal bug, weight underflow/overflow, or selection-bookkeeping error). However, the calibration is a model-consistency test: it inherits the injection generator’s assumptions. If real data violate those assumptions—for example through catalog completeness mismodeling, selection-function mismatch to the true detector network, residual PE systematics, or redshift-support errors—a positive real-data score can still arise without new GW propagation physics.

The mechanism controls provide guidance for targeted next steps: (i) because the score is dominated by spectral/selection information, stress injections that perturb incompleteness and selection priors are likely to be the most discriminating systematics tests; (ii) hero-event concentration motivates per-event audits (including PE-prior sensitivity and selection-weight diagnostics) focused on the handful of events that dominate the joint score; and (iii) complementary non-GR truth injections can quantify statistical power and expected score distributions when  $R(z) \neq 1$ .

### A. How large a selection/systematics shift can move the score?

An auxiliary selection-normalization sensitivity sweep in the hierarchical PE channel (five EM seeds, cached likelihood stacks with varied selection-model assumptions) shows that mean  $\Delta\text{LPD}_{\text{tot}}$  can move from approximately  $-1.43$  to  $+2.14$  for moderate variant changes, and up to  $+6.92$  for intentionally aggressive weighting choices. In this auxiliary sweep, cached data-term likelihood stacks were held fixed while selection-model assumptions were varied. While this sweep is not a fully self-consistent replacement for the catalog-mixture production analysis, it demonstrates that order-unity to multi-unit score excursions are plausible under selection-model changes alone. This motivates interpreting  $\Delta\text{LPD} \approx 3$  as a physically interesting tension that requires dedicated systematics-truth injection tests before a modified-gravity claim.

### B. Completed fixed-power and systematics-truth suites

Fixed-power response under GR truth. Using a five-point injected log- $R$  grid (0, 0.5, 1.0, 1.5, 2.0; 256

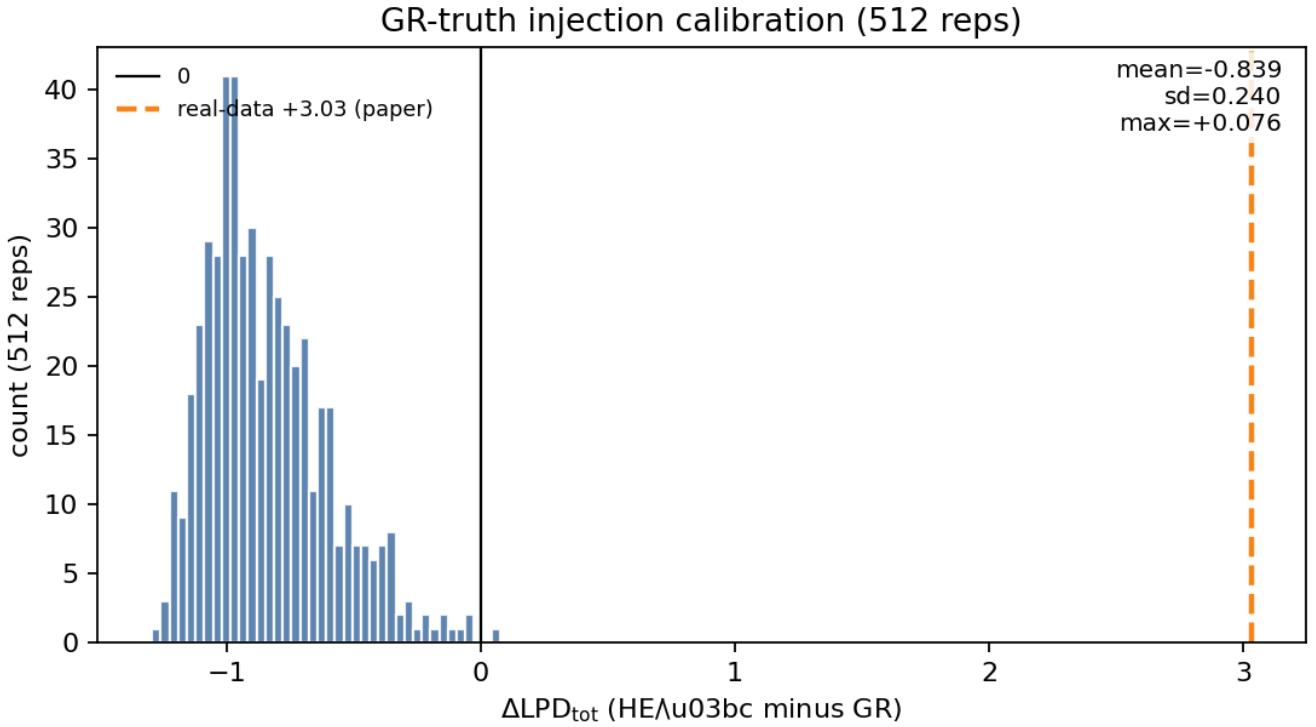


FIG. 1. GR-truth injection calibration (512 replicates): histogram of the posterior-predictive score difference  $\Delta\text{LPD}_{\text{tot}} = \text{LPD}(\text{prop}) - \text{LPD}(\text{GR})$  computed using the joint catalog-injection logmeanexp construction (Eq. 4). The vertical black line marks  $\Delta\text{LPD} = 0$ ; the dashed orange line marks the legacy real-data value  $\Delta\text{LPD}_{\text{tot}} \simeq +3.03$ . Under the calibrated GR-truth generator, the score distribution has mean  $-0.839$ , sd  $0.240$ , and maximum  $+0.076$  in 512 replicates.

replicates each), the mean score increases monotonically:  $-0.495 \pm 0.294$ ,  $-0.271 \pm 0.364$ ,  $-0.019 \pm 0.465$ ,  
220  $+0.261 \pm 0.590$ , and  $+0.562 \pm 0.732$ . This confirms that the implemented score has the expected directional sensitivity to progressively stronger injected propagation effects.  
245

GR-systematics truth matrix. For the nine-variant systematics matrix (128 replicates per variant), all maxima stay below  $+1$  (largest observed maximum  $+0.678$ , in selection weight none). No tested GR/systematics variant approaches the real-data score  $\Delta\text{LPD}_{\text{tot}} \simeq +3.03$ . Within this tested matrix, the real-data anomaly is therefore not reproduced by these perturbations of incompleteness and selection assumptions.  
230

### C. Small-sample hierarchical checkpoint and reproducibility note

As an additional consistency check, we ran a three-variant hierarchical integration checkpoint in the same output tree (baseline, selection-threshold, and fixed-low- $f_{\text{miss}}$  variants). The aligned GR-truth replicate ensemble<sub>235</sub><sup>260</sup> gives  $\langle \Delta\text{LPD}_{\text{tot}} \rangle = -0.548$  with sd  $0.252$  ( $n_{\text{rep}} = 12$ ), while the fixed-weight real-data score is  $\Delta\text{LPD}_{\text{tot}} = +3.027$  with calibrated tail frequency  $0/12$ . This run

is directionally consistent with the larger suites but remains a small-sample confirmatory checkpoint, not a replacement for the 512-replicate and  $9 \times 128$  matrices.

During this checkpoint, we identified and fixed a resume-path aggregation bug in the hierarchical wrapper (`scripts/run_dark_siren_hier_selection_uncertainty.py`) that could omit completed variants when reconstructing the final combined summary after a restart. The fix does not change per-variant replicate files or real-data summaries; it restores correct final integration from already completed artifacts.

### D. Ancillary cross-probe checks (context, not primary evidence)

Two additional holdout probes were run as secondary context. First, a three-source void-prism run (BOSS DR12 voids with Planck lensing plus ACT DR6/SDSS kSZx  $\theta$  maps) gives very small same-sign shifts relative to its internal GR baseline:  $\Delta\text{LPD}_{\text{vs GR}} = [+0.0116, +0.0198, +0.0249, +0.0127, +0.0221]$  across five seeds (mean  $+0.0182$ ). The corresponding null batteries remain non-decisive in that setup, so this is at most a weak directional consistency hint. Second, a raw strong-lens re-inference with free post-Newtonian  $\gamma_{\text{PPN}}$

## Decomposition under GR truth

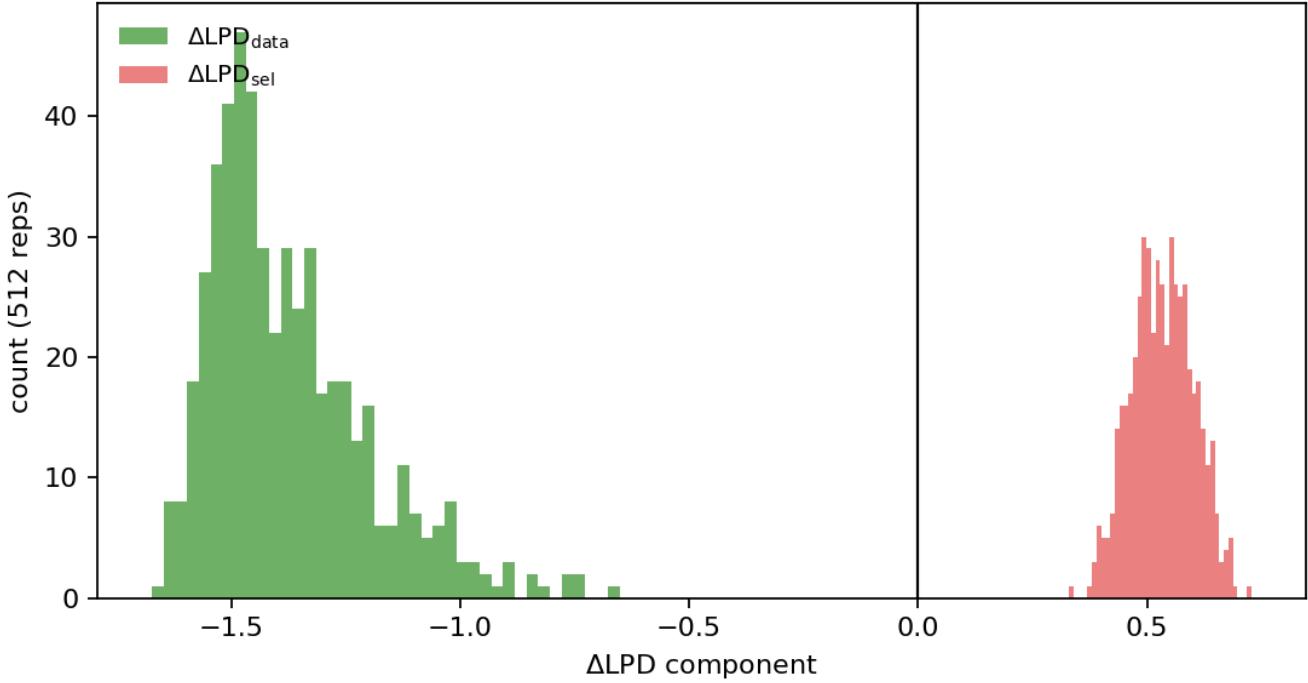


FIG. 2. Decomposition of the GR-truth calibration scores into data and selection components, defined by toggling the selection normalization term in Eq. 4 and taking the difference (Eq. 6). The selection term partially offsets the data term on average ( $\langle \Delta \text{LPD}_{\text{data}} \rangle = -1.374$ ,  $\langle \Delta \text{LPD}_{\text{sel}} \rangle = +0.534$ ), but the net GR-truth score remains negative.

over 8 public TDCOSMO/H0LiCOW lenses gives  $\gamma_{\text{PPN}}$  posterior quantiles ( $p16, p50, p84$ ) = (0.718, 0.968, 1.195),  
265 i.e., a mild sub-GR central value but with GR ( $\gamma_{\text{PPN}} = 1$ )  
290 still inside the credible interval. We therefore treat these ancillary probes as useful external stress checks, but  
295 not decisive model selectors at current data volume and calibration depth.

## VI. CONCLUSION

Using 36 GWTC-3 dark sirens, we find a posterior-predictive tension with the internal GR baseline, quantified by  $\Delta \text{LPD}_{\text{tot}} = +3.670$  in the updated O3 rerun  
275 with injection-trained logistic selection. Sky-rotation and spectral/sky mechanism controls localize the anomaly to the spectral/selection channel rather than unique host alignments. A GR-truth catalog-injection calibration (512 replicates) targeted at this dominant channel  
280 yields a centered-negative score distribution with maximum +0.076, placing both real-data scores (+3.03 legacy,  
305 +3.670 updated rerun) far outside the calibrated GR-truth ensemble under the injection-generator assumptions.  
285 The completed fixed-power grid further shows the expected monotonic score response to injected propagation strength, while the completed nine-variant GR-  
310 systematics matrix does not reproduce values close to

+3. These results substantially weaken the generic numerical-artifact explanation under tested assumptions, but they still do not uniquely identify modified gravity. The highest-priority next step remains expansion of the systematics-truth space (and independent catalogs/selection calibrations) to test whether unmodeled effects can bridge the remaining gap to  $\Delta \text{LPD} \sim 3$ . The new three-variant hierarchical checkpoint is consistent with this picture but is intentionally treated as a small-sample reinforcement only. An updated O3 rerun with an injection-trained logistic selection model gives  $\Delta \text{LPD}_{\text{tot}} = +3.670$ , confirming that the positive O3 signal survives this selection-model upgrade.

## DATA AND SOFTWARE AVAILABILITY

The source code and reproducibility materials for this analysis are archived on Zenodo at doi:10.5281/zenodo.18535331 (record title: “O3 Modified Gravity Tension Replication”). Core external sources used in this Letter include GWTC-3 PE products (doi:10.1103/PhysRevX.13.041039), GLADE+ (doi:10.1093/mnras/stac1443), and Planck 2018 lensing (doi:10.1051/0004-6361/201833886); additional ancillary-catalog source pointers are documented in the repository manifest. All figures in this Letter

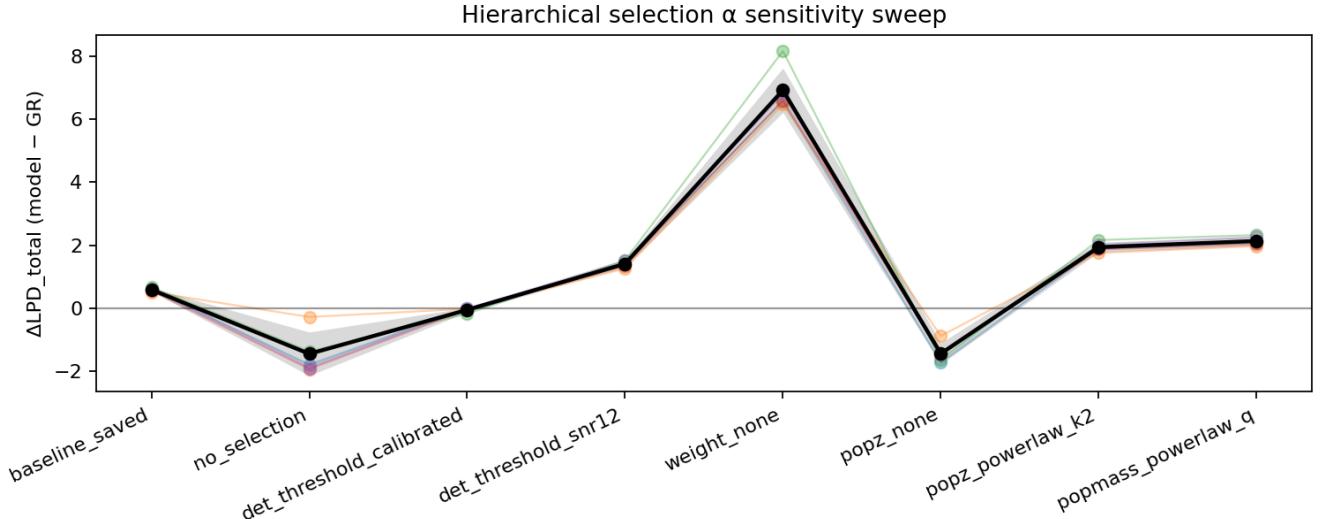


FIG. 3. Auxiliary selection-normalization sensitivity sweep in the hierarchical-PE channel (five EM seeds). The plotted variants modify the selection model while reusing cached event likelihood stacks. Mean  $\Delta \text{LPD}_{\text{tot}}$  spans from negative to positive values, illustrating that plausible selection assumptions can shift the score by order unity or larger. This is a scale-setting diagnostic for systematic sensitivity, not a substitute for full catalog-mixture stress injections.

are generated from the archived scripts and artifact manifests.

opment.

## ACKNOWLEDGMENTS

<sup>315</sup> The author used AI assistance during this project for brainstorming, drafting/editing text, and software devel-

- 
- [1] Abbott, R., et al. (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration) 2023, *Phys. Rev. X*, 13, 041039, doi:10.1103/PhysRevX.13.041039 (arXiv:2111.03606)
- [2] Belgacem, E., Dirian, Y., Foffa, S., & Maggiore, M. 2018, *Phys. Rev. D*, 98, 023510, doi:10.1103/PhysRevD.98.023510 (arXiv:1712.08108)
- [3] Dálya, G., et al. 2022, *Mon. Not. R. Astron. Soc.*, 514, 1403, doi:10.1093/mnras/stac1443
- [4] Nishizawa, A. 2017, *Phys. Rev. D*, 97, 104037, doi:10.1103/PhysRevD.97.104037 (arXiv:1710.04825)

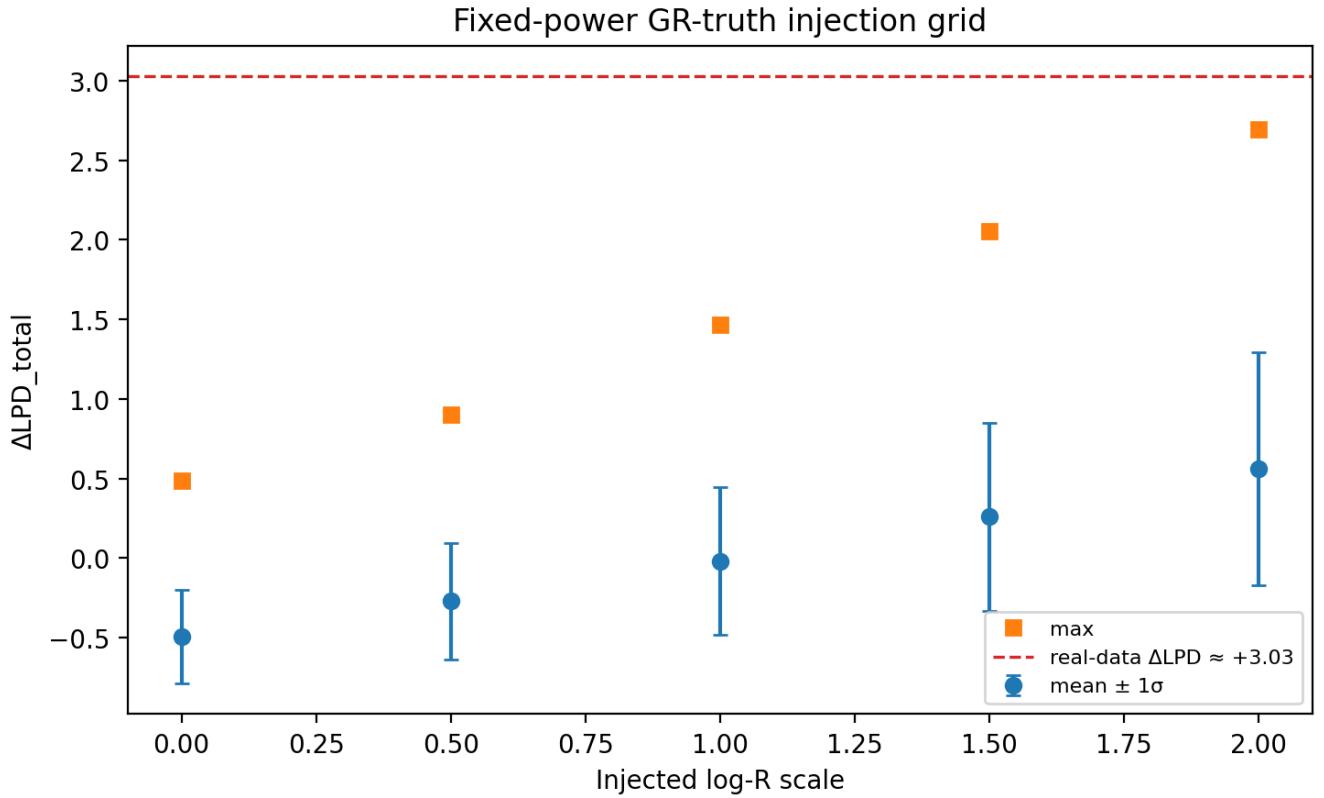


FIG. 4. Completed fixed-power injection grid under GR truth (five injected log- $R$  scales, 256 replicates per scale). Points show mean  $\Delta\text{LPD}_{\text{tot}}$  with  $1\sigma$  bars; squares mark per-scale maxima. The dashed red line is the real-data value  $\Delta\text{LPD}_{\text{tot}} \simeq +3.03$ . The monotonic upward trend validates directional score sensitivity to injected propagation strength.

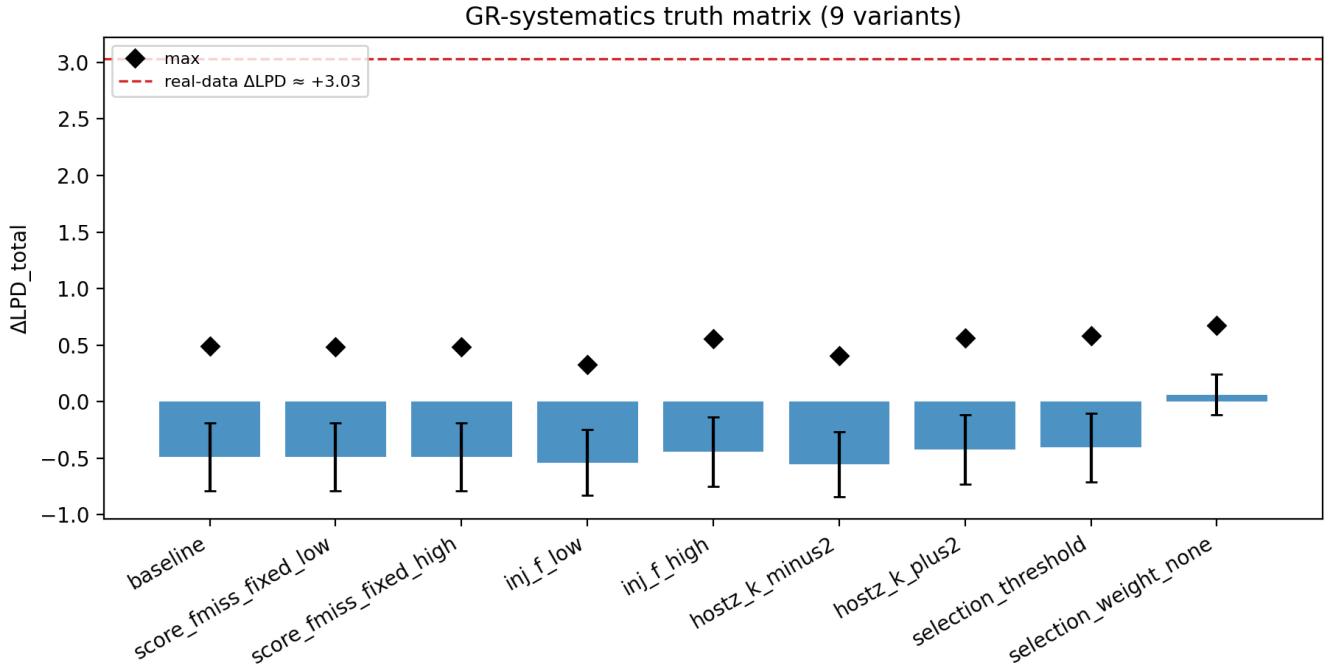


FIG. 5. Completed GR-systematics truth matrix (nine variants, 128 replicates each). Bars show mean  $\Delta\text{LPD}_{\text{tot}}$  with  $1\sigma$  bars; diamonds mark variant maxima. All tested variant maxima are  $\leq +0.678$ , well below the real-data  $\Delta\text{LPD}_{\text{tot}} \simeq +3.03$  (dashed red line).