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

Aiden B. Smith[1]

[1] *Independent Researcher*

We test modified gravitational-wave propagation with 36 GWTC-3 dark sirens using a host-incompleteness-marginalized galaxy-catalog likelihood. A fixed propagation-modification model, $d_L^{\rm GW} = R(z)\, d_L^{\rm EM}$, 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 $\Delta\mathrm{LPD}_{\rm tot} = +3.670$ ($\Delta\mathrm{LPD}_{\rm data} = +2.670$, $\Delta\mathrm{LPD}_{\rm sel} = +1.000$). Sky rotations give $\langle\Delta\mathrm{LPD}_{\rm rot}\rangle = +3.017$ (sd 0.091) and $P(\Delta\mathrm{LPD}_{\rm rot} \geq \Delta\mathrm{LPD}_{\rm real}) = 0.45$, indicating the signal is not driven by unique host alignments. GR-consistent injections (512 replicates) return mean -0.839, sd 0.240, and max $+0.076$ (none ≥ 3). Tested GR-consistent 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 distance-redshift channel preference.

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

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 a modified-friction term during propagation, yielding a redshift-dependent ratio

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

We refer to $R(z)$ as the GW propagation ratio (equivalently $\Xi(z) \equiv d_L^{\rm GW}/d_L^{\rm EM}$). 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),

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

In the minimal running-M_* embedding used here, the reconstructed horizon–entropy slope deformation $\mu(A) \equiv G_{\rm eff}(A)/G_N$ 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-consistent catalog-injection calibration that stress-tests the dominant distance-redshift/selection channel. The main result is a statistically interesting tension that can reflect either modified propagation or residual catalog/selection mismodeling.

II. DATA AND METHODS

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

We analyze $N_{\rm ev} = 36$ 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,

$$\begin{aligned} p(d_i \mid \theta, \mathcal{M}) = (1 - f_{\rm miss})\, p_{\rm cat}(d_i \mid \theta, \mathcal{M}) \\ + f_{\rm miss}\, p_{\rm miss}(d_i \mid \theta, \mathcal{M}), \end{aligned} \tag{3}$$

where $f_{\rm miss}$ 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 $p(\vartheta \mid d) \propto \mathcal{L}(d \mid \vartheta)\, \pi_{\rm PE}(\vartheta)$ 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-consistent 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 θ denote background/propagation parameters drawn from an external reconstruction posterior $p(\theta \mid d_{\rm recon})$. For a model $\mathcal{M}$, define the joint score

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

where $\{\theta_j\}_{j=1}^{N_s}$ are draws from $p(\theta \mid d_{\rm recon})$ and $\alpha(\theta, \mathcal{M})$ is the standard selection normalization (detection efficiency) computed from an injection-calibrated selection model. We report

$$\Delta\mathrm{LPD}_{\rm tot} \equiv \mathrm{LPD(prop)} - \mathrm{LPD(GR)}. \tag{5}$$

Intuitively, LPD is a joint predictive-fit score across all events: larger values mean the model assigns higher probability density to the observed dataset. A +1 shift in ΔLPD corresponds to a multiplicative predictive-density ratio of $\exp(1) \approx 2.7$. For diagnostic bookkeeping we also use the decomposition

$$\Delta\mathrm{LPD}_{\rm tot} = \Delta\mathrm{LPD}_{\rm data} + \Delta\mathrm{LPD}_{\rm sel}, \tag{6}$$

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

III. REAL-DATA TENSION AND MECHANISM CONTROLS

A. Real-data score and sky-rotation null

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

$$\Delta\mathrm{LPD}_{\rm tot} = +3.670, \qquad \exp(\Delta\mathrm{LPD}_{\rm tot}) \approx 39, \tag{7}$$

which indicates a statistically interesting preference for the propagation phenomenology over the internal GR baseline under the PPD construction. Here $\exp(\Delta\mathrm{LPD})$ 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 $\langle\Delta\mathrm{LPD}_{\rm rot}\rangle = +3.017$ (sd 0.091) and $P(\Delta\mathrm{LPD}_{\rm rot} \geq \Delta\mathrm{LPD}_{\rm real}) = 0.45$. 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 $\Delta\mathrm{LPD}_{\rm tot} = +3.670$ (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 $\Delta\mathrm{LPD}_{\rm tot} \simeq +3.03$.

B. Distance-only vs. sky-only controls

To isolate the dominant channel behind the preference, we implement two controls. In a distance-only ("spectral-only" in pipeline naming) 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 the distance-only control retains most of the preference ($\Delta\mathrm{LPD}_{\rm spectral} \simeq +2.995$), while sky-only is much smaller ($\Delta\mathrm{LPD}_{\rm sky} \simeq +0.969$). This localizes the anomaly to population-level distance–redshift consistency coupled to selection/incompleteness modeling, rather than to sky-localized host associations.

C. High-leverage-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 $\Delta\mathrm{LPD}_{\rm tot}$ 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-CONSISTENT 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 $\Delta\mathrm{LPD}_{\rm tot}$ under a calibrated GR null due to numerical, bookkeeping, or PE-prior-removal artifacts. We therefore construct a GR-consistent catalog-injection suite designed to stress-test the dominant distance-redshift/selection channel.

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

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 distance-only channel. This design matches the empirically dominant mechanism, and the 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.

B. Injection design

We perform a parametric-bootstrap-style calibration under the GR-null hypothesis ($R_{\mathrm{true}}(z) \equiv 1$), 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 $p(\theta \mid d_{\mathrm{recon}})$; (ii) for each of the 36 template events we sample a true redshift from a cached event-specific redshift support histogram; (iii) we compute $d_L^{\mathrm{EM}}(z_{\mathrm{true}})$ and set $d_L^{\mathrm{GW}} = d_L^{\mathrm{EM}}$; (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-consistent distribution of $\Delta\mathrm{LPD}_{\mathrm{tot}}$ for 512 replicates, with the real-data value marked. Under the GR-consistent null we find mean −0.839, sd 0.240, and maximum +0.076 in 512 replicates; none reach $\Delta\mathrm{LPD}_{\mathrm{tot}} \geq 3$. Figure 2 shows the decomposition into data and selection components: on average $\langle\Delta\mathrm{LPD}_{\mathrm{data}}\rangle = -1.374$ and $\langle\Delta\mathrm{LPD}_{\mathrm{sel}}\rangle = +0.534$, so the selection term partially offsets the data term but does not reverse the net preference under the GR-consistent null.

V. DISCUSSION

The GR-consistent calibration materially reduces the likelihood that the real-data preference is a generic numerical artifact that would also appear under the GR-consistent null (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 distance-redshift/selection information, stress injections that perturb incompleteness and selection priors are likely to be the most discriminating systematics tests; (ii) high-leverage-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 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 $\Delta\mathrm{LPD}_{\mathrm{tot}}$ 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 $\Delta\mathrm{LPD} \approx 3$ as a physically interesting tension that

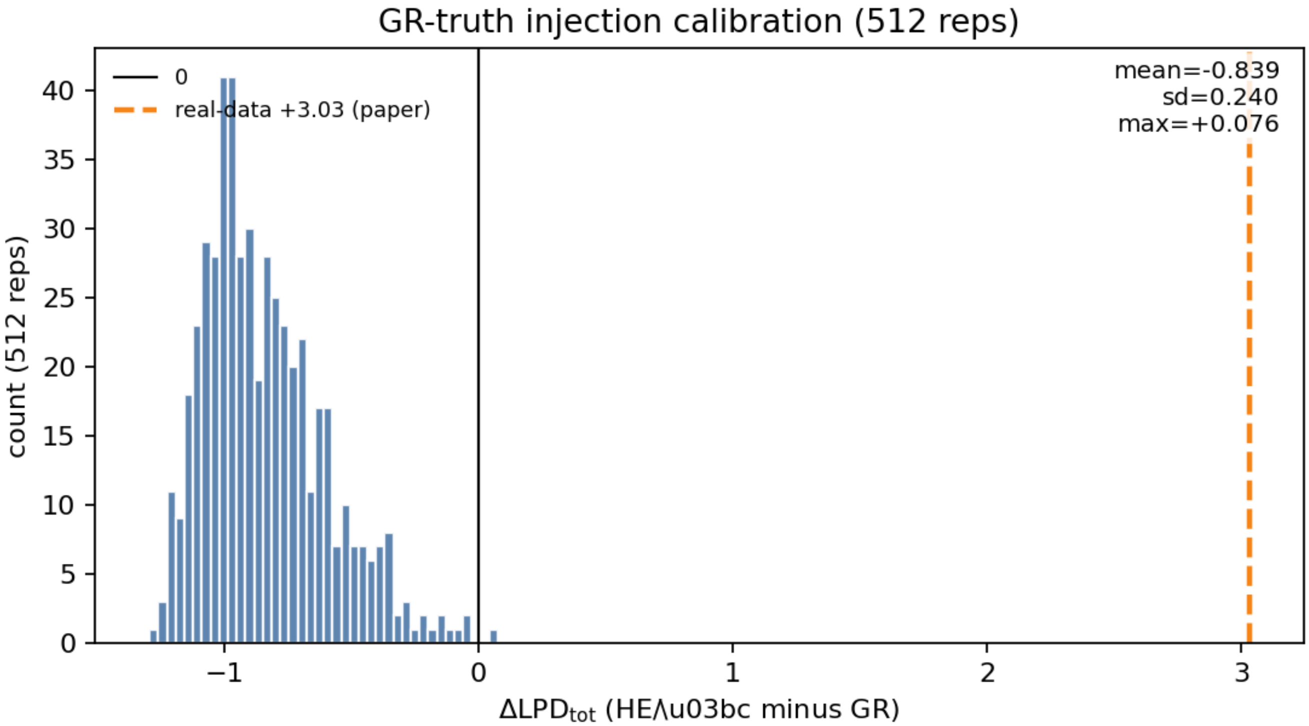

FIG. 1. GR-consistent injection calibration (512 replicates): histogram of the posterior-predictive score difference $\Delta\mathrm{LPD}_{\mathrm{tot}} = \mathrm{LPD(prop)} - \mathrm{LPD(GR)}$ computed using the joint catalog-injection logmeanexp construction (Eq. 4). The vertical black line marks $\Delta\mathrm{LPD}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

requires dedicated systematics-truth injection tests before a modified-gravity claim.

B. Completed fixed-power and systematics-truth suites

Fixed-power response under the GR-consistent null. Using a five-point injected log-R grid (0, 0.5, 1.0, 1.5, 2.0; 256 replicates each), the mean score increases monotonically: -0.495 ± 0.294, -0.271 ± 0.364, -0.019 ± 0.465, $+0.261 \pm 0.590$, and $+0.562 \pm 0.732$. This confirms that the implemented score has the expected directional sensitivity to progressively stronger injected propagation effects.

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 $\Delta\mathrm{LPD}_{\mathrm{tot}} \simeq +3.03$. Within this tested matrix, the real-data anomaly is therefore not reproduced by these perturbations of incompleteness and selection assumptions.

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_{miss} variants). The aligned GR-consistent replicate ensemble gives $\langle\Delta\mathrm{LPD}_{\mathrm{tot}}\rangle = -0.548$ with sd 0.252 ($n_{\mathrm{rep}} = 12$), while the fixed-weight real-data score is $\Delta\mathrm{LPD}_{\mathrm{tot}} = +3.027$ 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 9×128 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.

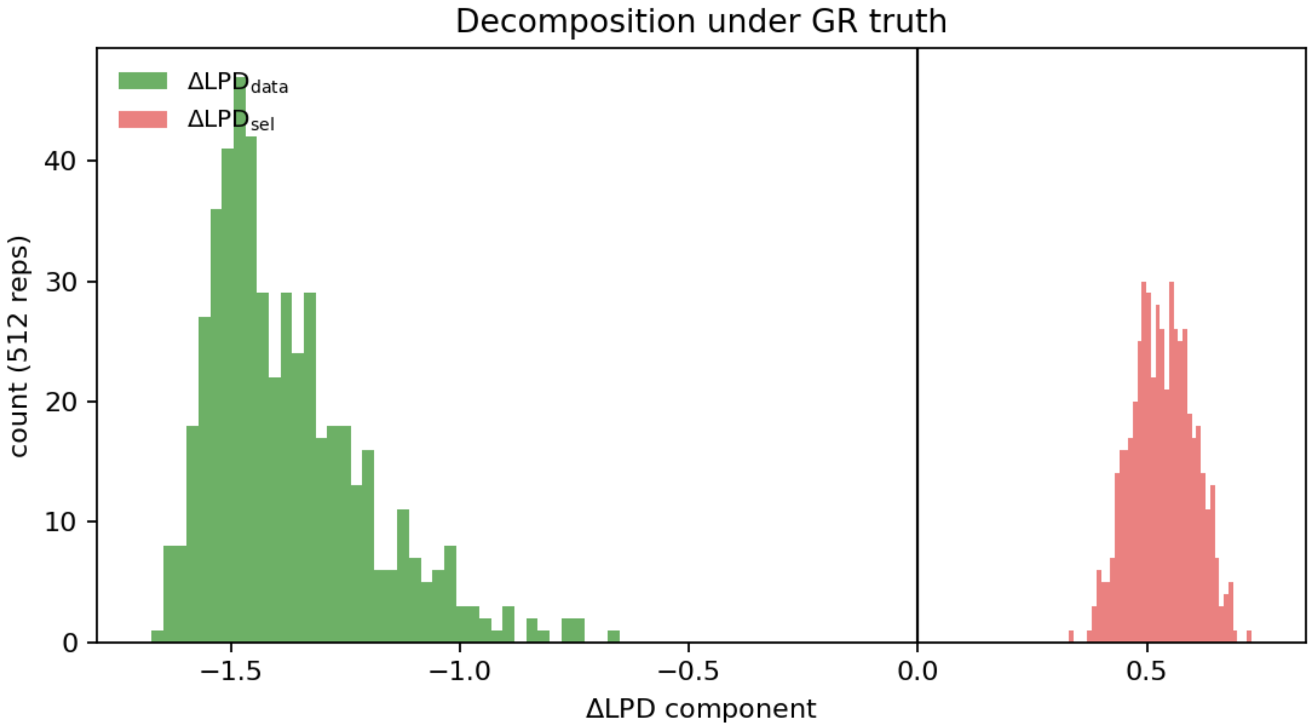

FIG. 2. Decomposition of the GR-consistent calibration scores into data and selection components, defined by toggling the selection normalization term in Eq. 4 and taking the difference (Eq. 6). The selection term partially offsets the data term on average ($\langle \Delta\mathrm{LPD}_{\rm data}\rangle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

D. Waveform-systematics checkpoint

As an additional robustness check focused on the two highest-leverage events, we evaluated waveform-family consistency across four random seeds per event. For paired comparisons between IMRPhenomXPHM and the corresponding XHM-labeled runs (executed with an IMRPhenomPv2 likelihood in this pipeline), the evidence shift is consistent with zero: $\langle \Delta \log Z_{\rm XPHM-Pv2}\rangle = -0.005$ across 8 pairs, with typical per-pair uncertainty $\sigma_{\Delta \log Z} \simeq 0.12$. Event-level means are -0.013 for GW200220_061928 and $+0.003$ for GW200308_173609. No coherent waveform-family preference is observed, indicating waveform-approximant choice is subdominant relative to the dark-siren anomaly scale.

E. 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 θ maps) gives very small same-sign shifts relative to its internal GR baseline: $\Delta\mathrm{LPD}_{\rm vs\,GR} = [+0.0116, +0.0198, +0.0249, +0.0127, +0.0221]$ 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 $\gamma_{\rm PPN}$ over 8 public TDCOSMO/H0LiCOW lenses gives $\gamma_{\rm PPN}$ posterior quantiles $(p16, p50, p84) = (0.718, 0.968, 1.195)$, i.e., a mild sub-GR central value but with GR ($\gamma_{\rm PPN} = 1$) still inside the credible interval. We therefore treat these ancillary probes as useful external stress checks, but 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 $\Delta\mathrm{LPD}_{\rm tot} = +3.670$ in the updated O3 rerun with injection-trained logistic selection. Sky-rotation and distance-only/sky-only mechanism controls localize the anomaly to the distance-redshift/selection channel rather than unique host alignments. A GR-consistent catalog-injection calibration (512 replicates) targeted at this dominant channel yields a centered-negative score distribution with maximum $+0.076$, placing both real-data scores ($+3.03$ legacy, $+3.670$ updated rerun) far outside the calibrated GR-consistent ensemble under the injection-generator assumptions. The completed fixed-

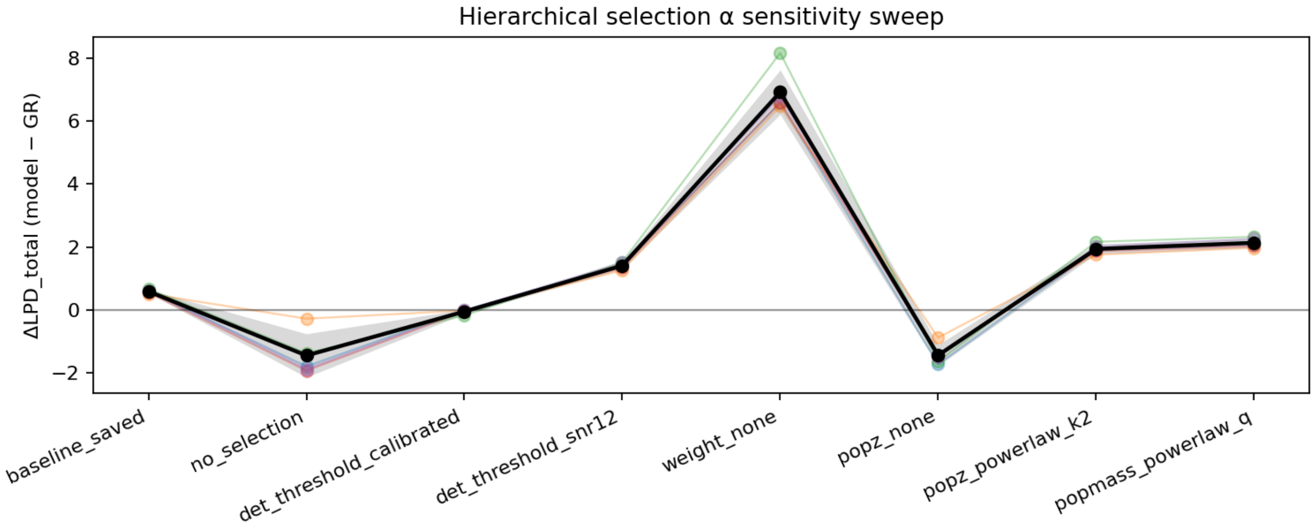

FIG. 3. Auxiliary selection-normalization sensitivity sweep in the hierarchical-PE channel (five EM seeds). The plotted variants modify the selection model while reusing cached event likelihood stacks. Mean $\Delta\mathrm{LPD}_{\mathrm{tot}}$ spans from negative to positive values, illustrating that 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.

power grid further shows the expected monotonic score response to injected propagation strength, while the completed nine-variant GR-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 $\Delta\mathrm{LPD} \sim 3$. 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 $\Delta\mathrm{LPD}_{\mathrm{tot}} = +3.670$, 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 are generated from the archived scripts and artifact manifests.

ACKNOWLEDGMENTS

The author used AI assistance during this project for brainstorming, drafting/editing text, and software development.

[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)

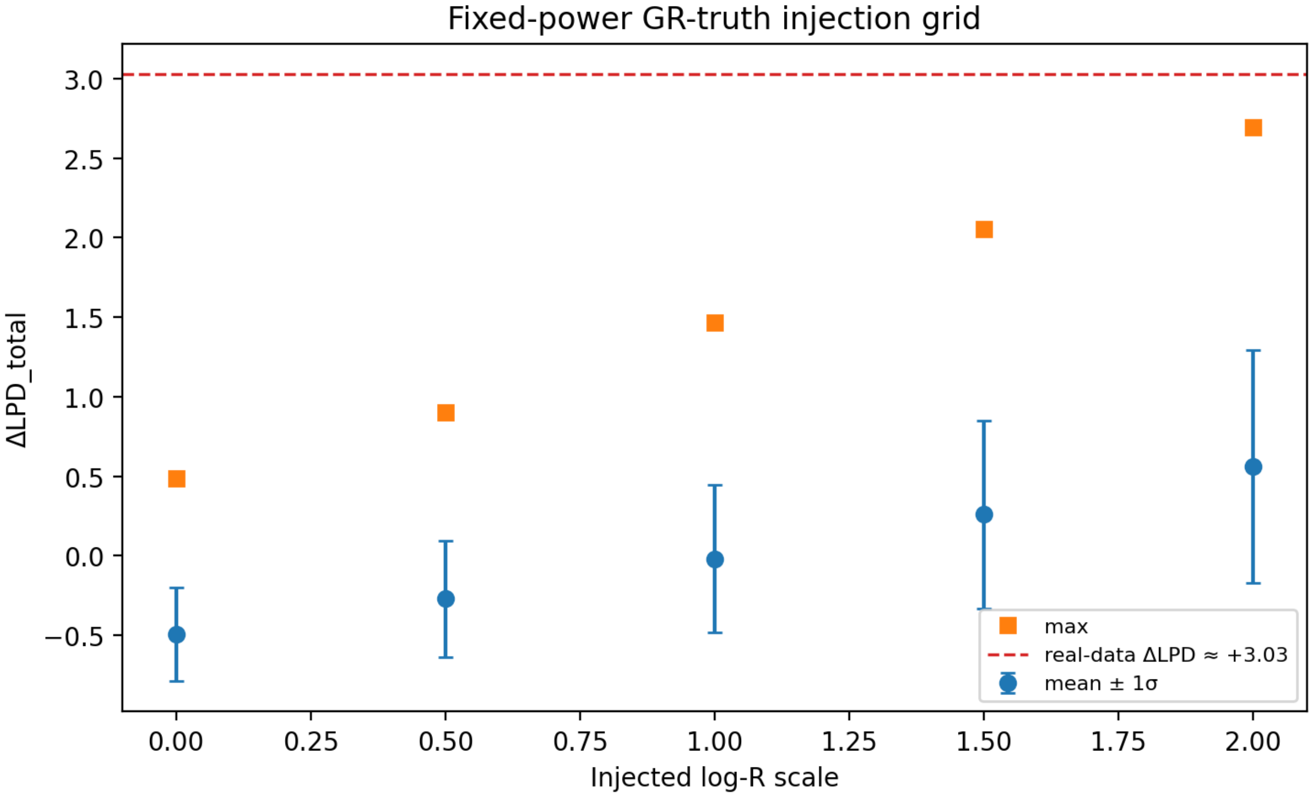

FIG. 4. Completed fixed-power injection grid under the GR-consistent null (five injected log-R scales, 256 replicates per scale). Points show mean $\Delta\mathrm{LPD}_{\mathrm{tot}}$ with 1σ bars; squares mark per-scale maxima. The dashed red line is the real-data value $\Delta\mathrm{LPD}_{\mathrm{tot}}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

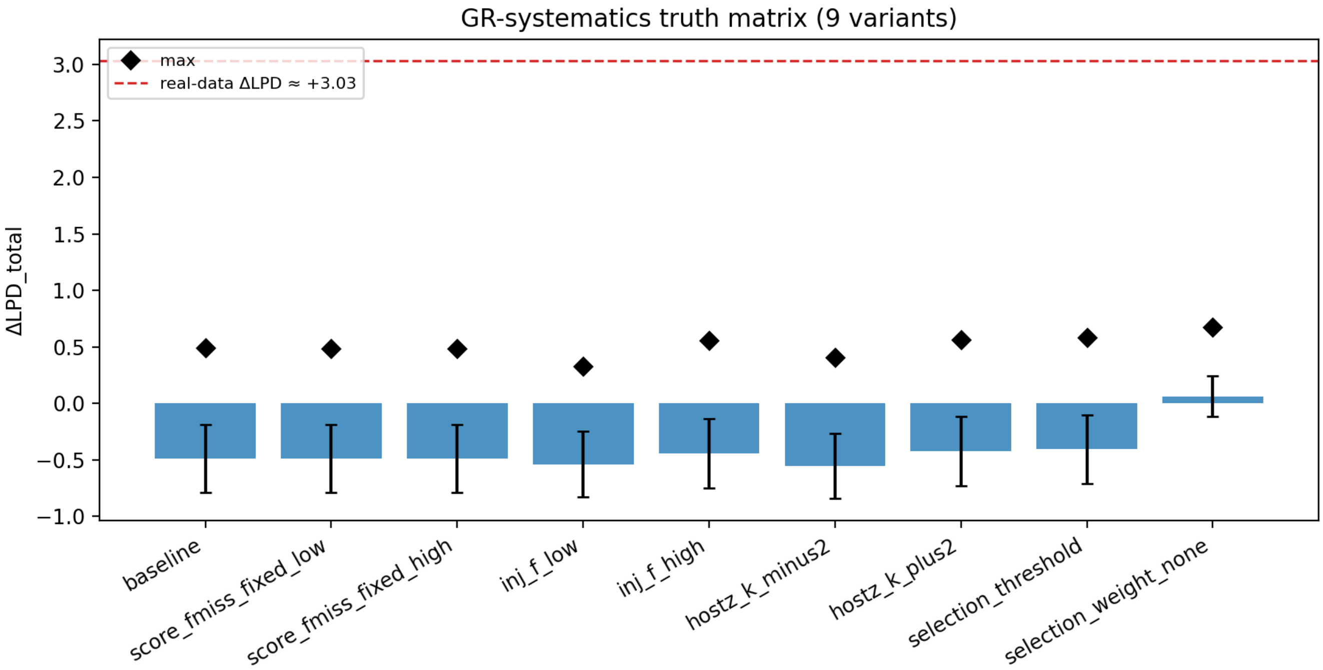

FIG. 5. Completed GR-systematics truth matrix (nine variants, 128 replicates each). Bars show mean $\Delta\mathrm{LPD}_{\mathrm{tot}}$ with 1σ bars; diamonds mark variant maxima. All tested variant maxima are $\leq +0.678$, well below the real-data $\Delta\mathrm{LPD}_{\mathrm{tot}} \simeq +3.03$ (dashed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