MG-truth Causal Closure of the O3 Dark-Siren Score Anomaly

Aiden B Smith Independent Researcher **E-mail:** aidenblakesmithtravel@gmail.com

14 February 2026

Abstract

The O3 GWTC-3 dark-siren catalogue exhibits a calibrated preference for a fixed modified gravitational-wave (GW) propagation history, quantified by a posterior-predictive score shift $\Delta\mathrm{LPD}_{\mathrm{tot}} \simeq 3.67$ relative to an internal General Relativity (GR) propagation baseline in the production scoring framework. Here we perform a causal-closure test: assuming the same modified-propagation template is the truth, does a GR-baseline analysis naturally generate a score excursion on the observed scale when filtered through the identical catalogue and selection machinery? We generate synthetic dark-siren catalogues using the O3 event ensemble and the same injection-calibrated selection normalisation as the real-data analysis, draw redshifts from the same host-redshift support structures used by the pipeline, and generate synthetic distance likelihoods centred on either GR-truth or modified-propagation-truth distances. Across 256 MG-truth replicates, $\Delta\mathrm{LPD}_{\mathrm{tot}}$ has mean 3.50 and yields $P(\Delta\mathrm{LPD}_{\mathrm{tot}} \geq 3.6699) = 0.219$; under the matched GR-truth generator the mean is 3.31 with $P(\Delta\mathrm{LPD}_{\mathrm{tot}} \geq 3.6699) = 0.0195$. A dose–response dial in the forward generator, $R_\alpha(z) = 1 + \alpha[R(z) - 1]$, produces a monotone increase in the tail probability at the observed threshold as α is increased from 0 to 1. When the catalogue likelihood geometry is deliberately removed via structureless redshift weighting, the large-score scale collapses (mean $\Delta\mathrm{LPD}_{\mathrm{tot}} \simeq 0.64$). These results support the inference-bias mechanism as a coherent explanation of the observed score scale, while remaining conditional on the forward-generator assumptions and the tested nuisance family.

Keywords: gravitational waves, standard sirens, cosmology, inference bias, catalogues, selection effects

1 Introduction

Dark-siren cosmology combines gravitational-wave (GW) luminosity-distance posteriors with galaxy catalogues to infer statistical host-redshift information [?, ?]. In General Relativity (GR), the GW luminosity distance equals the electromagnetic luminosity distance for a given expansion history. In many modified-gravity constructions, however, the GW amplitude can experience additional effective "friction" during propagation, producing a redshift-dependent mapping $d_L^{\mathrm{GW}}(z) = R(z) d_L^{\mathrm{EM}}(z)$ with $R(z) \neq 1$ [?, ?].

The companion O3 dark-siren analysis in this codebase reports a posterior-predictive preference for a fixed modified-propagation history over an internal GR propagation baseline, quantified by $\Delta\mathrm{LPD}_{\mathrm{tot}} \simeq 3.67$ in the production scoring framework, with injection-based GR-truth calibrations that disfavour a generic numerical artifact [?, ?]. If such a propagation deviation is physical, then applying a GR ruler to a modified-propagation Universe introduces an inference bias: the analysis assumption (GR propagation) is not neutral, and can manufacture tension-scale parameter wedges when distances are inverted.

This paper isolates one narrow question that can be addressed pre-O4 using public data and the existing scoring pipeline: does the modified-propagation "truth" used in the Hubble-tension transfer analysis causally imply a score anomaly of the observed magnitude when analysed under a GR baseline? We answer this by constructing

MG-truth and GR-truth synthetic catalogues that preserve the key catalogue and selection structures, scoring them with the unmodified production code, and comparing the resulting ΔLPD distributions to the observed score.

2 Formalism and experiment design

2.1 Predictive score

The production analysis compares a fixed propagation hypothesis to an internal GR baseline using a joint posterior-predictive log score LPD evaluated across the event set with explicit selection normalisation. We report the score difference

$$\Delta\mathrm{LPD}_{\mathrm{tot}} \equiv \mathrm{LPD}(\mathrm{prop}) - \mathrm{LPD}(\mathrm{GR}), \tag{1}$$

where "prop" is the fixed modified-propagation template and "GR" sets $R \equiv 1$. Larger $\Delta\mathrm{LPD}_{\mathrm{tot}}$ indicates that the fixed template assigns higher predictive density to the catalogue under the scoring construction.

2.2 Forward generators and causal closure

The causal-closure experiment generates synthetic catalogues that preserve the real pipeline's dominant ingredients: (i) the same O3 event ensemble definition; (ii) the same injection-calibrated selection normalisation; and (iii) the same host-redshift support structures used by the catalogue likelihood in the dominant spectral channel. For each replicate, we draw a redshift for each event from its cached host-redshift support and compute a corresponding electromagnetic luminosity distance $d_L^{\mathrm{EM}}(z)$ under the baseline background history used by the production run. We then generate a synthetic distance likelihood centred on either

$$d_L^{\mathrm{GW,\,true}}(z) = d_L^{\mathrm{EM}}(z) \qquad \text{(GR truth)}, \tag{2}$$

or

$$d_L^{\mathrm{GW,\,true}}(z) = R_\alpha(z)\, d_L^{\mathrm{EM}}(z), \qquad R_\alpha(z) = 1 + \alpha\,[R(z) - 1] \qquad \text{(MG truth)}, \tag{3}$$

with $\alpha \in [0, 1]$ used as a dose–response dial in the forward generator. The synthetic catalogues are then scored with the unmodified production scoring code to obtain $\Delta\mathrm{LPD}_{\mathrm{tot}}$ and its data/selection decomposition.

2.3 Controls

We include two controls designed to validate mechanism alignment rather than to claim a comprehensive systematics closure. First, we evaluate multiple GR-truth generator variants (e.g. altering the truth missing-host mixture fraction) to test whether the GR-tail probability at the observed threshold is stable under plausible GR-truth forward modelling choices. Second, we remove the catalogue likelihood geometry by replacing the galaxy-catalogue term with a deliberately structureless redshift weighting in both the truth generator and the analysis model; this tests whether the large-score scale depends on catalogue/selection geometry rather than being a free property of the score.

3 Results

3.1 Baseline MG-truth and GR-truth distributions

The observed real-catalogue score is $\Delta\mathrm{LPD}_{\mathrm{tot}} = 3.6699$. Figure **??** shows the $\Delta\mathrm{LPD}_{\mathrm{tot}}$ distributions from 256 MG-truth and 256 GR-truth synthetic catalogues.

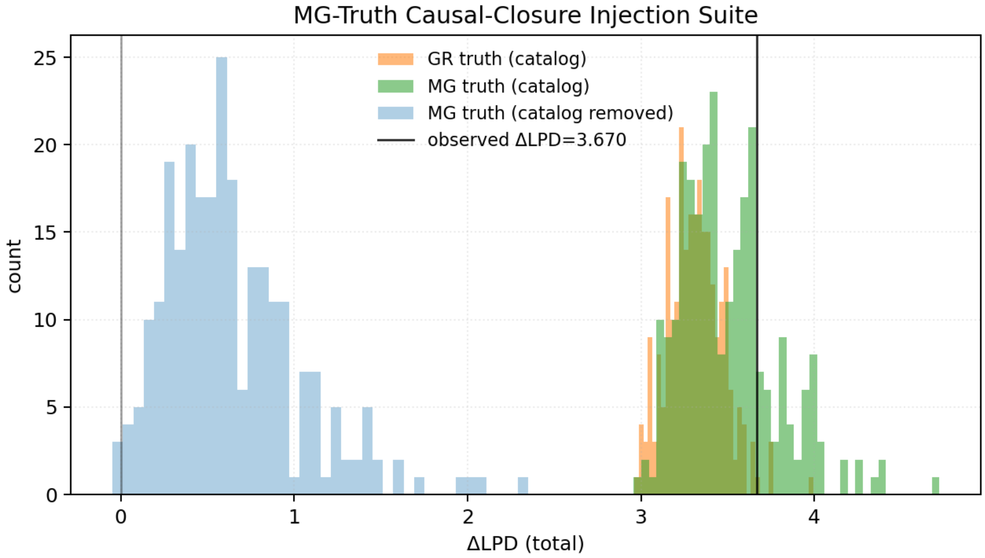

Figure 1. Causal-closure overlay. Histogram of $\Delta\mathrm{LPD}_{\mathrm{tot}}$ for MG-truth (blue) and GR-truth (orange) synthetic catalogues, together with the "catalog removed" control (grey). The vertical line marks the observed real-catalogue value.

Under MG truth (with $\alpha = 1$), the distribution has mean 3.50 and yields $P(\Delta\mathrm{LPD}_{\mathrm{tot}} \geq 3.6699) = 0.219$. Under the matched GR-truth generator, the mean is 3.31 and $P(\Delta\mathrm{LPD}_{\mathrm{tot}} \geq 3.6699) = 0.0195$.

3.2 Dose–response dial

The forward-generator amplitude dial produces a monotone increase in both the mean score and the tail probability at the observed threshold. Figure **??** shows $P(\Delta\mathrm{LPD}_{\mathrm{tot}} \geq 3.6699)$ as a function of α for $\alpha = \{0, 0.25, 0.5, 0.75, 1\}$. In this run snapshot, the tail probability rises from 0.023 at $\alpha = 0$ to 0.219 at $\alpha = 1$.

3.3 Selection contribution and catalogue-removed control

Figure **??** visualises the decomposition of the score into data and selection components under MG truth. In this closure suite, the selection contribution remains of order unity and the dose response is primarily expressed in the data-fit term, consistent with the mechanism that the modified propagation accumulates in the distance–redshift channel.

When the catalogue likelihood geometry is removed by structureless redshift weighting, the large-score scale collapses: the "catalog removed" control has mean $\Delta\mathrm{LPD}_{\mathrm{tot}} \simeq 0.64$ and does not populate the observed tail (Figure **??**). This demonstrates that the score amplitude in the baseline closure runs is not arbitrary; it depends on the catalogue/selection compression of the distance posteriors.

4 Discussion

This causal-closure experiment is designed to answer a specific mechanistic question: whether the modified-propagation history used in the Hubble-tension transfer analysis can naturally generate a score excursion of the observed scale when analysed under a GR propagation baseline. Within the assumptions of the forward generator, the answer is affirmative: the observed $\Delta\mathrm{LPD}_{\mathrm{tot}}$ sits in a typical region of the MG-truth distribution and lies in the high tail of the matched GR-truth distribution, and the generator exhibits a clean monotone dose–response with α.

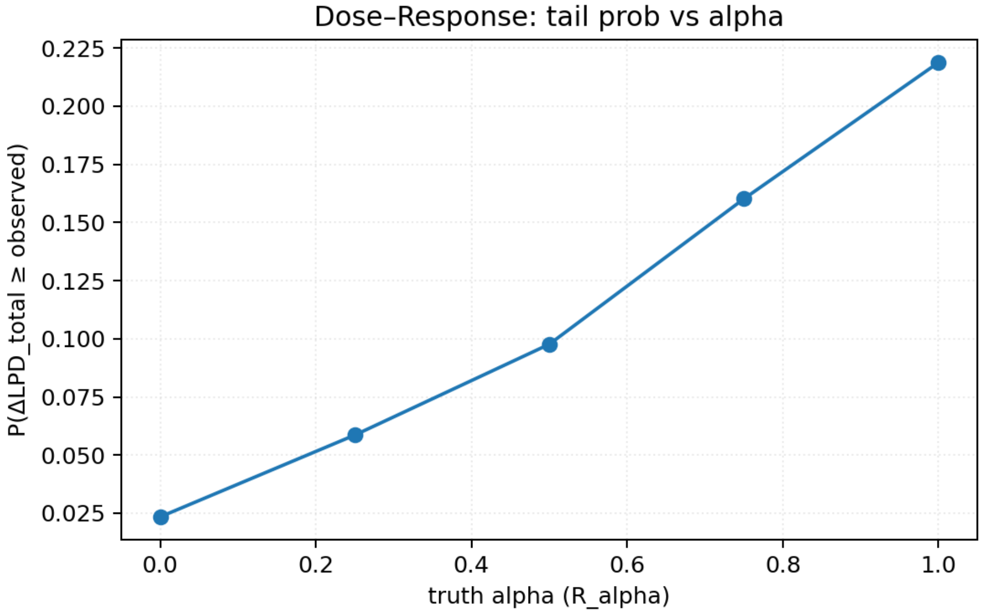

Figure 2. Dose–response in the forward generator. Tail probability $P(\Delta\mathrm{LPD}_{\mathrm{tot}} \geq 3.6699)$ as a function of the truth amplitude dial α in $R_\alpha(z)$.

The experiment remains conditional in two ways. First, the forward generator uses synthetic distance likelihoods tuned to the O3 event ensemble; it is not a full strain-level reanalysis, and it primarily targets the dominant spectral channel used by the production scoring. Second, the GR-tail probability is not immutable: GR-truth generator variants that substantially alter the truth redshift-drawing mixture can increase the probability of exceeding the observed threshold. For this reason, we interpret the closure result as a mechanism alignment test rather than as a standalone significance statement.

The broader implication is methodological: whenever a pipeline compresses catalogue structure and selection functions through a fixed modelling assumption, “anomalies” of tension scale can be generated by ruler mismatch (truth $\neq$ analysis assumption). This does not in itself diagnose the origin of any specific anisotropy claim, but it motivates forward-model closure tests and structure-destruction nulls as standard robustness requirements.

5 Conclusion

We have implemented a MG-truth causal-closure experiment within the O3 dark-siren scoring pipeline. Under the baseline forward generator, MG truth yields $P(\Delta\mathrm{LPD}_{\mathrm{tot}} \geq 3.6699) = 0.219$ whereas GR truth yields $P(\Delta\mathrm{LPD}_{\mathrm{tot}} \geq 3.6699) = 0.0195$ for the observed threshold, and a forward-generator amplitude dial produces a monotone dose–response. When catalogue geometry is removed, the large-score scale collapses. These results support the inference-bias mechanism as a coherent explanation of the observed score magnitude, while remaining explicit about forward-generator dependence and the residual space of catalogue/selection systematics.

Acknowledgements

The author used generative AI tools (large language models) for editorial assistance and code refactoring. All scientific claims, calculations and results were verified by the

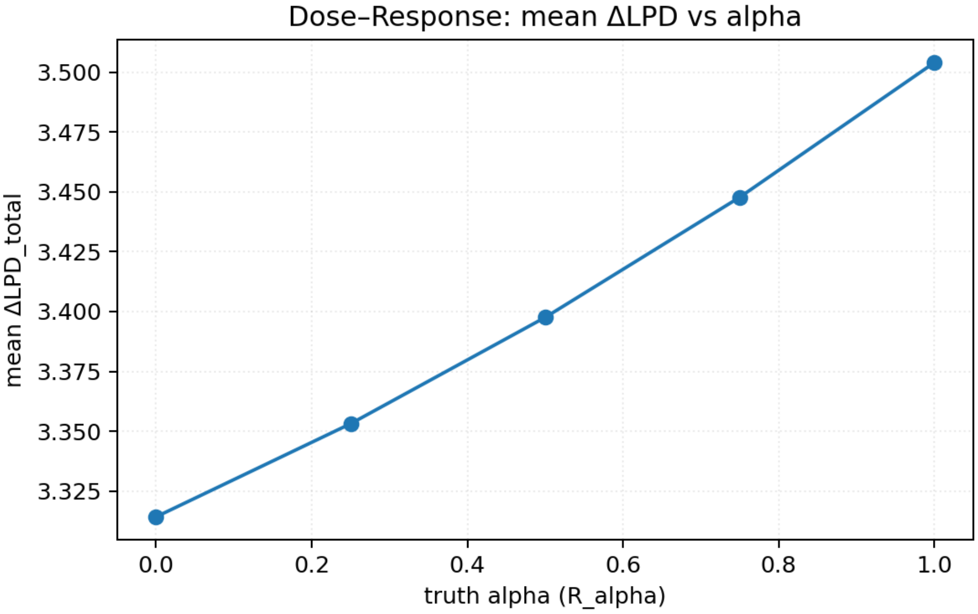

Figure 3. Dose–response in the forward generator. Mean $\Delta\mathrm{LPD}_{\mathrm{tot}}$ as a function of α.

author.

Declarations

Conflict of interest

The author declares no competing interests.

Data availability

The code and archived artefacts required to reproduce this experiment are provided with the repository and mirrored on Zenodo (code DOI: 10.5281/zenodo.18640608). The companion O3 dark-siren scoring release is archived at Zenodo DOI: 10.5281/zenodo.18640507.

Funding

The author received no specific funding for this work.

References

References

[1] Schutz B F 1986 *Nature* **323** 310–1

[2] Abbott B P *et al* (LIGO Scientific Collaboration & Virgo Collaboration) 2017 *Nature* **551** 85–8

[3] Belgacem E, Dirian Y, Foffa S and Maggiore M 2018 *Phys. Rev. D* **98** 023510

[4] Nishizawa A 2018 *Phys. Rev. D* **97** 104037

[5] Abbott R *et al* (LIGO Scientific Collaboration, Virgo Collaboration & KAGRA Collaboration) 2023 *Phys. Rev. X* **13** 041039

[6] Dálya G *et al* 2022 *Mon. Not. R. Astron. Soc.* **514** 1403–11

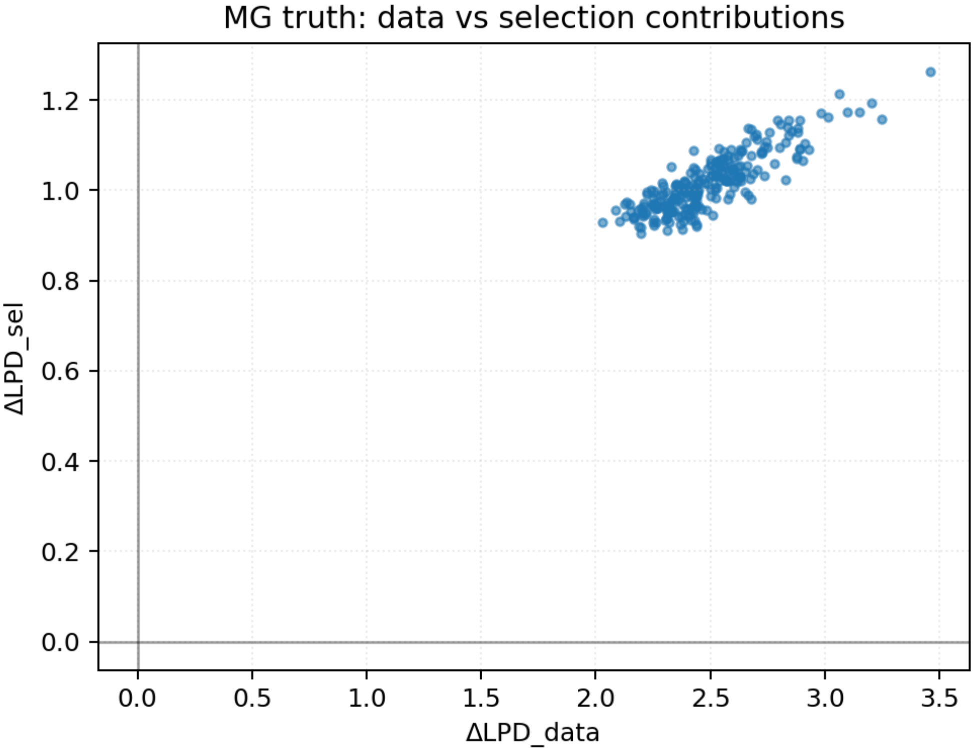

Figure 4. MG-truth decomposition. Scatter of $\Delta\mathrm{LPD}_{\mathrm{data}}$ versus $\Delta\mathrm{LPD}_{\mathrm{sel}}$ across MG-truth replicates (diagnostic for whether selection dominates the score).