Skepthical review: *Quantifying and Attributing Waveform Model-Dependent Systematics in GW231123: A Multi-Scale Posterior Analysis*

Summary

This manuscript presents a posterior-level multi-model comparison of waveform-model systematics in gravitational-wave parameter estimation for the high-mass, strongly precessing binary black hole event GW231123. Using posterior samples from five waveform models (NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM, IMRPhenomXO4a, IMRPhenomXPHM), the authors quantify inter-model differences with 1D/2D discrepancy measures (KDE-based Jensen–Shannon divergence and overlap integrals; Sec. 2.2) and explore high-dimensional structure via PCA/ICA (Sec. 2.3; Sec. 3.2). They then attempt to attribute discrepancies by grouping models according to waveform domain, calibration basis, and precession treatment (Sec. 2.4; Sec. 3.3) and propose "systematic-inclusive" intervals defined as the envelope (union) of per-model 90% credible intervals (Sec. 2.5.2; Sec. 3.4). The topic is timely and the diagnostic pipeline is conceptually appealing; however, the current manuscript has (i) critical reproducibility gaps about posterior provenance and run configuration (Sec. 2.1.1), (ii) corrupted/non-informative waveform-model characterization that undermines the grouping/attribution narrative (Table 1; Sec. 2.1.2, 2.4, 3.3), (iii) incomplete/truncated numerical summary tables (Table 2/5) that prevent verification of key claims, and (iv) insufficient specification and validation of the KDE-based JS/overlap computations and PCA/ICA pooling/weighting choices (Sec. 2.2–2.3). In addition, several "driver/attribution" conclusions read too causally given confounding between grouping labels (domain/family/calibration) and small group sizes (Sec. 3.3). Addressing these issues—especially sample provenance/priors, table integrity, metric implementation details with robustness checks, and a more careful attribution framing—would substantially strengthen the credibility and impact of the results.

Strengths

- Timely focus on waveform-model systematics for a challenging, high-mass and strongly precessing event (Sec. 1; Sec. 3).
- Use of multiple waveform families (NR surrogate, EOB, phenomenological) and multiple discrepancy diagnostics (JS divergence, 2D overlap, PCA/ICA) to probe both marginal and high-dimensional differences (Sec. 2.2–2.3; Sec. 3.1–3.3).
- Clear overall pipeline structure: quantify discrepancies → explore structure → attempt attribution via model grouping → summarize with systematic-inclusive intervals (Sec. 2; Sec. 3).
- The "systematic-inclusive" (envelope) interval idea is a practically useful reporting device for communicating model-to-model spread in difficult events, provided it is clearly labeled and interpreted (Sec. 2.5.2; Sec. 3.4).

- The manuscript highlights an important qualitative takeaway likely to be robust: GW231123 is inferred to be strongly precessing across models, while some intrinsic parameters (masses, $\chi_{\rm eff}$) and distance/redshift can be substantially model-dependent (Sec. 3.1; Sec. 3.4).
- Figures appear to be designed for direct visual comparison across models and groups (consistent overlays and grouped views), which is valuable once captions/method details are made fully transparent (Sec. 3; Figs. 1–21).

Major issues

1. **Posterior-sample provenance and parameter-estimation configuration are insufficiently documented, making it unclear whether differences are purely waveform-driven or partly due to non-waveform analysis differences (Sec. 2.1.1; also impacts Sec. 3.1–3.4).** The paper does not clearly state the PE framework(s) used (Bilby/LALInference/RIFT/…), whether priors and likelihood settings are identical across waveform models (mass/spin/distance priors; reference frequency; parameterization), the detector network and data segment, PSD estimation, calibration-uncertainty treatment, low-frequency cutoff, marginalizations, sampler settings/convergence diagnostics, and whether all runs used matched settings.

   Recommendation: Expand Sec. 2.1.1 with a reproducibility-grade description of the PE runs and posterior sources. Include a compact configuration table listing, for each waveform model: PE code and version, likelihood and priors (explicitly confirming they are identical across models or enumerating differences), data segment duration, detector network, $f_{\rm low}$, PSD method, calibration treatment, marginalizations, sampler type and key settings, and convergence/quality diagnostics (e.g., ESS; multiple chains where applicable). If posteriors come from public LVK/GWOSC releases, provide DOIs/URLs and document any post-processing. Briefly discuss (Sec. 1 or Sec. 4) how residual non-waveform differences could bias the interpreted "waveform systematics."

2. **Waveform-model characterization is corrupted/non-informative (Table 1; Sec. 2.1.2, 2.4, 3.3), which undermines the central grouping/attribution analysis.** The current "Key Characteristics" entries contain filler text and omit essential details (domain, higher modes, precession implementation, calibration/training range, validity limits). As written, readers cannot assess whether the groupings are correct or whether GW231123 lies near/over model validity boundaries.

   Recommendation: Replace Table 1 with accurate, citation-backed model descriptions. For each of NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM, IMRPhenomXO4a, IMRPhenomXPHM, explicitly list: (i) time vs frequency domain; (ii) precession treatment (e.g., twisting-up vs fully precessing surrogate) and frame conventions; (iii) higher-mode content; (iv) calibration/training data and parameter-space coverage (q, spin

magnitudes, etc.); (v) known limitations relevant to GW231123 (extrapolation risk). Then ensure Sec. 2.4 (group definitions) and Sec. 3.3 (attribution statements) directly reference these documented properties.

3. **Key numerical summary tables are truncated/incomplete or inconsistent with the narrative, preventing verification of the main quantitative claims (Table 2, Table 5; Sec. 3.1, Sec. 3.4).** Examples include the truncated SEOBNRv5PHM column label/entries ("SE…") in Table 2 and truncated headers in Table 5; this makes it impossible to confirm reported spreads (e.g., in m_2, $\chi_{\rm eff}$, z) and the construction of systematic-inclusive intervals.

   Recommendation: Audit and correct all result tables (at minimum Table 2 and Table 5) directly from the underlying posterior samples. Ensure: complete medians and 90% credible intervals for all five models, untruncated headers, consistent parameter naming/units, and consistency between tables and text. Recompute Table 5 from the stated rule (Sec. 2.5.2: envelope/union of per-model 90% CIs) and explicitly cross-check that all quoted numerical ranges in Sec. 3.1 and Sec. 3.4 match the final tables.

4. **Discrepancy metrics (JS divergence; 1D/2D overlaps) are central evidence but the implementation is under-specified and lacks robustness/sensitivity validation (Sec. 2.2; Table 4; figures summarizing JS/overlap).** KDE-based JS can be sensitive to kernel choice, bandwidth, bounded-parameter leakage ($\chi_p \in [0,1]$, $\cos$ tilts $\in [-1,1]$), grid choice, and numerical integration. The JS bound claim ($[0,1]$) depends on log base and normalization, which is not fully specified.

   Recommendation: In Sec. 2.2.1–2.2.2, provide full computational details: KDE kernel, bandwidth method (Scott's rule is not sufficient alone—state kernel and any multivariate bandwidth strategy), boundary handling for bounded variables, grid/domain definitions (global vs per-model support), discretization and numerical integration scheme, and the JS definition including log base. Add sanity checks: overlap(model, model)≈ 1; stability under grid refinement; and at least a small sensitivity study for key parameters (e.g., m_2, $\chi_{\rm eff}$, z, χ_p) varying bandwidth and/or using an alternative divergence estimator (e.g., histogram with principled binning or kNN-based divergence). Report uncertainties (e.g., bootstrap) for representative JS/overlap values in Table 4 and/or an appendix.

5. **The grouping-based "attribution" claims are currently stronger than justified by the analysis design (Sec. 2.4; Sec. 3.3).** Pooling posteriors within groups (domain/calibration/precession) does not isolate a single modeling attribute because group labels are confounded with waveform family and implementation details; group sizes are small; memberships overlap; and intra-group variability is not systematically compared to inter-group separation. As a result, statements such as "domain is the primary driver" read as causal rather than associative.

Recommendation: Revise Sec. 2.4 and Sec. 3.3 to (i) explicitly state assumptions behind pooling and acknowledge confounding/overlap, (ii) soften causal language to "correlated with" unless controlled contrasts are demonstrated, and (iii) add quantitative tests comparing intra-group vs inter-group differences (e.g., variance-ratio/ANOVA-style summaries on key parameters or on PC/IC projections). Where possible, add within-family comparisons that reduce confounding (e.g., IMRPhenomXPHM vs IMRPhenomXO4a; IMRPhenomTPHM vs IMRPhenomX*; or EOB vs surrogate within similar modeling choices). Include bootstrap intervals for group-level metrics and perform sensitivity checks excluding potential outliers to see if trends persist.

6. **PCA/ICA analysis mixes samples across models in a way that can conflate inter-model shifts with physical degeneracy structure, and weighting/balancing across models is not clearly defined (Sec. 2.3; Sec. 3.2).** If one model contributes more samples or broader posteriors, it can dominate the standardization and component directions. ICA results are not fully reproducible because the scaling/whitening conventions and loading matrices are not provided.

   Recommendation: Clarify and, if needed, revise the PCA/ICA procedure in Sec. 2.3: specify sample counts per model, whether you downsample/weight so each model contributes equally, the exact standardization convention (computed on pooled samples vs per-model then combined), and PCA→ICA whitening details, random seeds, convergence criteria, and software versions. Add: (i) explained-variance ratios for the first several PCs; (ii) the full loading matrix (or an appendix) for PCs used in figures; (iii) an ICA loading table (clearly defining whether values are mixing/unmixing loadings or correlations). Consider adding a stability check: repeat PCA/ICA with equal-per-model downsampling and report whether the qualitative conclusions in Sec. 3.2 and Sec. 3.3 are unchanged.

7. **Mass–redshift/distance degeneracy and cosmology assumptions are not discussed with enough care given that redshift is a headline systematic (Sec. 3.1; Sec. 3.4).** It is unclear what cosmology is used for converting luminosity distance to redshift, and whether the dominant disagreement is in distance/$(1+z)$ rather than detector-frame masses. Without showing detector-frame masses and luminosity distance, some "mass systematics" may be largely reparameterizations of distance/redshift differences.

   Recommendation: State explicitly the cosmology used for distance→redshift conversion and whether it is fixed across models. Add detector-frame mass posteriors (e.g., $M_{\mathrm{chirp,det}}$, total mass in detector frame) and luminosity distance posteriors alongside source-frame masses and z (either in Sec. 3.1 or an appendix). Discuss results in terms of the mass–distance–inclination/redshift degeneracy and clarify whether inter-model differences primarily appear in D_L/z or also in detector-frame intrinsic scales.

8. **One model (IMRPhenomXO4a) appears to behave as an outlier for key parameters (notably m_2 and remnant quantities; Sec. 3.1 and related figures), but the paper does not demonstrate that this is a well-sampled solution exploring the same likelihood mode rather than a sampling/pathology/prior-boundary issue.**

   Recommendation: Provide diagnostics showing the XO4a posterior is reliable and comparable: sampler convergence/ESS (or equivalent), checks for multimodality, and (if available) comparisons of maximum log-likelihood (or matched-filter/logL summaries) across waveform runs. Confirm identical priors/support (q, spins, distance) and identical waveform settings (modes included, $f_{\rm min}$, reference frequency). If XO4a is near a model-validity boundary or extrapolating, explicitly state this (Table 1 + Sec. 3.1/4) and temper interpretation accordingly.

Minor issues

1. Internal narrative inconsistency about which parameters are robust vs systematically uncertain (Abstract; Sec. 2.1.3 vs Sec. 3.1 and Sec. 3.4). Sec. 2.1.3 claims exploratory concordance for masses/redshift and discrepancies in spins, whereas later results emphasize substantial mass/redshift differences and more robust χ_p.

   Recommendation: Reconcile statements across Abstract/Sec. 1/Sec. 2.1.3/Sec. 3.1/Sec. 3.4/Sec. 4. Add a short "robust vs non-robust" parameter summary in Sec. 3.4 tied explicitly to the quantitative criteria in Sec. 2.5, and ensure all earlier/later text matches that summary.

2. Criteria for labeling parameters as "consensus/robust" are not presented in a single verifiable summary, and the choice of thresholds (e.g., JS$\approx$0.05) is not well justified (Sec. 2.5; Sec. 3.4).

   Recommendation: In Sec. 2.5, define the robustness classification rule precisely (JS threshold, overlap threshold if used, and how pairwise comparisons are aggregated). Add a compact table in Sec. 3.4 listing, for each key parameter, max pairwise JS and min pairwise overlap across models plus the resulting label. Briefly justify the threshold choice (e.g., via bootstrap sampling noise or a reference distribution).

3. Pooling/group posteriors is used in multiple places (Sec. 2.4; Sec. 2.5.1) but the mixture weighting is not clearly specified (equal weight per model vs per sample). This affects pooled medians/CIs and group-level JS/overlaps.

   Recommendation: Define pooled/group distributions explicitly as mixtures $P_{\rm group}(\theta) = \sum_k w_k P_k(\theta)$ and state w_k. If pooling samples, explain how unequal sample counts are handled (downsampling or weights) and ensure this is consistent across grouped comparisons and PCA/ICA.

4. Figure captions often lack the quantitative or methodological details needed for standalone interpretation (many figures in Sec. 3; Figs. 1–21): sample counts, KDE settings, normalization, and direct references to the relevant metric tables.

   Recommendation: Update captions to include: number of samples per model (or effective weights), KDE bandwidth rule and boundary handling, and (where appropriate) the relevant JS/overlap values (or a pointer to a table). Consider annotating medians/credible intervals directly on 1D plots for faster quantitative reading.

5. The "systematic-inclusive credible interval" is presented in a way that could be mistaken for a probability-calibrated 90% credible interval under a single model (Sec. 2.5.2; Sec. 3.4).

   Recommendation: Rename/label it consistently as an "envelope/range-over-models" and explicitly state it is not a 90% posterior region under any single generative model unless a discrete model index with weights is introduced. Optionally add (even in supplement) an equal-weight model-mixture posterior summary as a complementary report.

6. Event context is light for a case-study paper (Sec. 1): readers would benefit from basic properties (detector network, SNR, why the event is challenging for waveforms).

   Recommendation: Add a brief paragraph in Sec. 1 summarizing GW231123 context: observing run, detectors used, approximate network SNR (if available), and why high mass + strong precession (and/or higher modes) makes waveform systematics particularly relevant here.

Very minor issues

1. Typographical/formatting artifacts and truncated headers reduce readability and create ambiguity (Tables 1–5; various sections): broken words/line breaks, truncated column names ("SE", "Systematic-Incl"), inconsistent model spellings and parameter symbols.

   Recommendation: Proofread the manuscript source (not OCR) and fix all truncations and line-break artifacts. Standardize model names (NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM, IMRPhenomXO4a, IMRPhenomXPHM) and parameter notation (e.g., $\cos\theta_{\rm JN}$, $\phi_{\rm JL}$, $\chi_{\rm eff}$, χ_p, a_f, z) across text/tables/figures.

2. Notation/column-name inconsistency (e.g., mass_1_source vs $m_1^{\rm src}$; cos_theta_jn vs $\cos\theta_{\rm JN}$) can make the pipeline harder to follow (Sec. 2–3; Table 3).

   Recommendation: Add a short notation/variable mapping table (main text or appendix) mapping CSV column names to manuscript symbols and units, and use one canonical notation consistently.

3. Cross-references to sections/figures appear occasionally imprecise (Sec. 2.3; Sec. 2.5; Sec. 3.2–3.3).

   Recommendation: Systematically check and correct section/figure references so that definitions (Sec. 2.2–2.3) and uses (Sec. 3) are linked unambiguously.

4. Accessibility/presentation: some figures may be difficult in grayscale or for color-vision deficiency; axis labels/units and legend clarity are inconsistent across the many panels (Figs. 1–21).

   Recommendation: Adopt a colorblind-safe palette with redundant encodings (line styles/markers), ensure all axes include units/frames, increase font sizes where needed, and standardize legend ordering and model/group labeling across figures.

Key statements and references

- • **The first two principal components obtained from a PCA of the combined, standardized posterior samples for nine astrophysical parameters of GW231123 capture over 51% of the total variance, with PC1 accounting for 26.8% and being strongly positively correlated with redshift z, inclination $\cos\theta_{\rm jn}$, and primary spin magnitude a_1 while being strongly anti-correlated with secondary mass m_2, and PC2 accounting for 25.1% and primarily representing an anti-correlation between primary and secondary masses (m_1 and m_2) together with strong anti-correlation with primary spin tilt $\cos t_1$ and secondary spin magnitude a_2, thereby encapsulating the dominant mass–redshift–spin and mass-ratio degeneracies in the parameter space.**
- *Reference(s):* (none)
- • **Independent Component Analysis (ICA) applied to the same standardized feature matrix as the PCA yields an IC2 that is overwhelmingly dominated by the spin azimuth $\phi_{\rm jl}$ with a loading of 0.74 and an IC1 that is most sensitive to the secondary spin tilt $\cos t_2$ with a loading of 0.64, indicating that $\phi_{\rm jl}$ and the secondary spin tilt define statistically independent directions in the posterior and that different waveform models infer systematically different constraints along these ICs, particularly IC1, for GW231123.**
- *Reference(s):* (none)
- • **The pairwise Jensen–Shannon divergence between the 1D marginal posteriors for the effective inspiral spin $\chi_{\rm eff}$ inferred with IMRPhenomXPHM and SEOBNRv5PHM for GW231123 is 0.57, indicating severe disagreement between these two waveform models, whereas the JS divergence for**

the effective precession spin χ_p between NRSur7dq4 and IMRPhenomTPHM is only 0.007, demonstrating near-identical χ_p posteriors across these models.

- *Reference(s):* (none)
- **• In the 2D joint posterior for the component masses (m_1, m_2) of GW231123, the overlap integral between IMRPhenomX04a and each of the other waveform models is typically less than 0.01, showing that IMRPhenomX04a explores a largely disjoint region of mass parameter space and is a clear outlier in its inference of the mass ratio compared to the other models.**
- *Reference(s):* (none)
- **• Group-level discrepancy metrics between time-domain and frequency-domain waveform models for GW231123 show that the JS divergence for the primary mass is 0.236 and the 2D overlap for (mass_1_source, mass_2_source) is 0.123, while the JS divergence for $\chi_{\rm eff}$ is 0.125, demonstrating that the waveform domain choice (time vs. frequency) is the dominant driver of systematic differences in inferred component masses and contributes appreciably to aligned-spin uncertainties.**
- *Reference(s):* (none)
- **• When waveform models are grouped by calibration basis into NR/EOB-based (NRSur7dq4, SEOBNRv5PHM) and phenomenological (IMRPhenomXO4a, IMRPhenomXPHM, IMRPhenomTPHM) families, the group-level JS divergence for the primary mass is 0.132, the JS divergence for $\chi_{\rm eff}$ is 0.020, and the overlap in the $(\chi_{\rm eff}, \chi_p)$ plane is 0.677, indicating that calibration basis has a moderate impact on mass inference but only a minor effect on $\chi_{\rm eff}$ and the $\chi_{\rm eff}$–χ_p spin degeneracy for GW231123.**
- *Reference(s):* (none)

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper is primarily methodological/interpretive with a small number of explicit mathematical definitions (JS divergence description, overlap integral, interval construction) and linear-algebra outputs (PCA loadings). There are no multi-step derivations; the main audit points are definition precision, notation consistency, and whether stated properties (bounds, variance explained) align with the provided tables.

Checked items

1. ✔ **Posterior overlap integral definition** (Sec. 2.2.2, p.3)
    - **Claim:** Defines overlap between two 2D posteriors as $O(P_1, P_2) = \int \min(P_1(\theta), P_2(\theta)) d\theta$, where higher values indicate more agreement.
    - **Checks:** definition consistency, normalization/bounds, dimensional analysis
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** P_1 and P_2 are properly normalized probability density functions over the same parameter space, The integral is over the full support of θ (for 2D: the relevant 2D plane)
    - **Notes:** If P_1, P_2 are normalized densities, $\int \min(P_1, P_2)$ is dimensionless and lies in $[0, 1]$. The paper does not explicitly specify the 2D measure ($d\theta_1 d\theta_2$), but the definition itself is correct.
2. ⚠ **JS divergence boundedness statement** (Sec. 2.2.1, p.3)
    - **Claim:** States that Jensen–Shannon divergence is bounded in $[0, 1]$.
    - **Checks:** definition/bounds, missing specification check
    - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
    - **Assumptions/inputs:** The JS divergence is computed with a particular log base and a standard definition $JS(P, Q) = 0.5\, KL(P||M) + 0.5\, KL(Q||M)$, $M = 0.5(P + Q)$
    - **Notes:** The $[0, 1]$ bound is not guaranteed unless the definition specifies a normalization/log base (e.g., base-2 logs). The paper omits the explicit formula and log base, so the bound as stated is not mathematically pinned down.
3. ✔ **2D overlap interpreted as agreement metric** (Sec. 2.2.2, p.3; also used in Sec. 3.1, p.8)
    - **Claim:** Overlap closer to 1 implies agreement; closer to 0 implies minimal commonality.
    - **Checks:** sanity/limiting cases
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** P_1, P_2 are normalized and defined over the same space
    - **Notes:** Limiting cases are consistent: identical densities give overlap 1; disjoint supports give overlap 0.
4. ✔ **Systematic-inclusive credible interval construction** (Sec. 2.5.2, p.6)
    - **Claim:** Defines a 'systematic-inclusive' 90% credible interval as $[\min_k q_{0.05,k}, \max_k q_{0.95,k}]$ across models.
    - **Checks:** set/interval logic, definition consistency
    - **Verdict:** PASS; confidence: high; impact: moderate

- **Assumptions/inputs:** Each model posterior has defined 5^{th} and 95^{th} percentiles
- **Notes:** The interval is indeed the envelope (union in 1D) of the modelwise 90% intervals. It is conservative by construction. Minor terminology caution: it need not contain exactly 90% probability under any single combined distribution, but the stated construction itself is correct.

5. ⚠ **Pooling samples to form combined/group posteriors** (Secs. 2.4.1 and 2.5.1, pp.4–6)
    - **Claim:** Pooled samples from multiple models represent a group-level or combined posterior used for medians/CIs and for group comparisons.
    - **Checks:** implicit assumption identification, definition completeness
    - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
    - **Assumptions/inputs:** Pooling implies a mixture distribution over models, Weights across models are implicitly determined by pooling procedure
    - **Notes:** Mathematically, pooling corresponds to a mixture with some weights. The weights are not specified (equal per model vs proportional to sample count), so the resulting 'combined median/CI' and group-level metrics are not uniquely defined from the text.

6. ✔ **Standardization before PCA/ICA** (Sec. 2.3.1, p.4)
    - **Claim:** All features are standardized to mean 0 and standard deviation 1 to prevent scale dominance.
    - **Checks:** dimensional consistency, method appropriateness (analytic)
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Standardization is performed columnwise across the combined dataset
    - **Notes:** Columnwise z-scoring makes the PCA operate on a correlation-like scale; consistent with later variance/eigenvalue interpretation (total variance equals number of variables).

7. ✔ **PCA loadings vs explained variance (PC1)** (Sec. 3.2.1, Table 3 and PC1 description, p.8)
    - **Claim:** PC1 explains 26.8% of variance; Table 3 lists PC1 loadings across 9 standardized variables.
    - **Checks:** linear algebra consistency, variance/eigenvalue consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** The reported numbers are PCA 'loadings' (eigenvectors scaled by $\sqrt{\text{eigenvalue}}$) rather than unit-norm eigenvector coefficients
    - **Notes:** Sum of squared PC1 loadings ≈ 2.418, matching eigenvalue implied by $0.268 \times 9 \approx 2.412$. This supports that the table reports loadings in the common scaled convention and is internally consistent.

8. ✔ **PCA loadings vs explained variance (PC2)** (Sec. 3.2.1, Table 3 and PC2 description, p.8)
    - **Claim:** PC2 explains 25.1% of variance; Table 3 lists PC2 loadings.
    - **Checks:** linear algebra consistency, variance/eigenvalue consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Same loading convention as PC1
    - **Notes:** Sum of squared PC2 loadings $\approx \mathbf{2.252}$, matching eigenvalue implied by $\mathbf{0.251 \times 9 \approx 2.259}$ within rounding.
9. ✔ **Interpretation of PC1 as mass–redshift–inclination degeneracy direction** (Sec. 3.2.1, p.8)
    - **Claim:** PC1 is strongly correlated with z and $\cos(\theta_{\rm jn})$ and anti-correlated with m_2.
    - **Checks:** consistency with provided loadings
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Interpretation uses signs/magnitudes of loadings in Table 3
    - **Notes:** Table 3 shows $z = \mathbf{0.85}$ and $\cos(\theta_{\rm jn}) = \mathbf{0.68}$ with $m_2 = \mathbf{-0.67}$ on PC1, matching the narrative description.
10. ✔ **Interpretation of PC2 as mass-ratio direction** (Sec. 3.2.1, p.8–9)
    - **Claim:** PC2 represents an anti-correlation between m_1 and m_2 (mass ratio degeneracy).
    - **Checks:** consistency with provided loadings
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Interpretation uses loadings in Table 3
    - **Notes:** Table 3 has $m_{1,\rm src} = \mathbf{0.75}$ and $m_{2,\rm src} = \mathbf{-0.54}$ on PC2, consistent with the stated anti-correlation between masses.
11. ⚠ **ICA component dominance claims** (Sec. 3.2.2, p.9)
    - **Claim:** IC2 is dominated by $\phi_{\rm jl}$ with loading 0.74; IC1 is most sensitive to $\cos(t_2)$ with loading 0.64.
    - **Checks:** verifiability from provided content, definition completeness
    - **Verdict:** UNCERTAIN; confidence: high; impact: minor
    - **Assumptions/inputs:** A specific ICA scaling convention, Availability of unmixing/mixing matrices to verify dominance
    - **Notes:** No ICA loading/unmixing matrix is provided in the PDF excerpt, so the numerical dominance statements cannot be audited for internal mathematical correctness.

12. ✖ **Notation/variable-name consistency across methods and tables** (Secs. 2.2–3.2; Tables 2–5, pp.3–13)
    - **Claim:** Uses consistent symbols/labels for parameters across KDE/JS/overlap/PCA/ICA and summary tables.
    - **Checks:** notation consistency
    - **Verdict:** FAIL; confidence: high; impact: minor
    - **Assumptions/inputs:** Same physical parameters are referenced by different labels
    - **Notes:** Multiple aliases are used without an explicit mapping (e.g., 'mass_1_source' vs $m_{1,\mathrm{src}}$; 'cos_tilt_1' vs $\cos(t_1)$; 'cos_theta_jn' vs $\cos\theta_{\mathrm{JN}}$). This is not an algebraic error but is an internal consistency/clarity defect that complicates auditing and replication.

Limitations

- The PDF contains few explicit equations/derivations; most mathematical procedures are described in prose, limiting the depth of step-by-step algebraic verification.
- The exact computational definitions for JS divergence (formula, log base) and for KDE-based approximations (grid/support/normalization) are not specified, preventing a fully determinate analytic audit of claimed bounds and metric properties as implemented.
- ICA numerical claims cannot be verified because the relevant matrices (unmixing/mixing) are not included.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Fifteen numeric checks were performed on medians, systematic-bias median ranges, systematic-inclusive 90% CI construction rules, and a PCA variance-sum statement. All checks passed within the stated tolerances, with no computed discrepancies against the reported reference endpoints/thresholds.

Checked items

1. ✔ **C1** (Page 7, Table 2 (Primary Mass medians))
    - **Claim:** Primary Mass ($M_\odot$) median values per model are: 133.4, 143.2, 149.9, 129.1, 133.7.
    - **Checks:** min_max_range_check
    - **Verdict:** PASS
    - **Notes:** Compared computed min/max of provided medians to reported endpoints.
2. ✔ **C2** (Page 7, Table 2 (Secondary Mass medians))

- **Claim:** Secondary Mass ($M_\odot$) median values per model are: 110.0, 55.1, 93.3, 110.6, 111.1.
- **Checks:** min_max_range_check
- **Verdict:** PASS
- **Notes:** Compared computed min/max of provided medians to reported endpoints.

3. ✔ **C3** (Page 7, Table 2 (χeff medians))
    - **Claim:** Effective Inspiral Spin (χ_{eff}) median values per model are: 0.44, 0.30, 0.04, 0.23, 0.44.
    - **Checks:** min_max_range_check
    - **Verdict:** PASS
    - **Notes:** Compared computed min/max of provided medians to reported endpoints.
4. ✔ **C4** (Page 7, Table 2 (χp medians))
    - **Claim:** Effective Precession Spin (χ_p) median values per model are: 0.77, 0.82, 0.75, 0.78, 0.73.
    - **Checks:** min_max_range_check
    - **Verdict:** PASS
    - **Notes:** Compared computed min/max of provided medians to reported endpoints.
5. ✔ **C5** (Page 7, Table 2 (Final Mass medians))
    - **Claim:** Final Mass ($M_\odot$) median values per model are: 227.3, 189.7, 232.7, 227.0, 228.2.
    - **Checks:** min_max_range_check
    - **Verdict:** PASS
    - **Notes:** Compared computed min/max of provided medians to reported endpoints.
6. ✔ **C6** (Page 7, Table 2 (Final Spin medians))
    - **Claim:** Final Spin (a_f) median values per model are: 0.89, 0.85, 0.71, 0.81, 0.87.
    - **Checks:** min_max_range_check
    - **Verdict:** PASS
    - **Notes:** Compared computed min/max of provided medians to reported endpoints.
7. ✔ **C7** (Page 7, Table 2 (Redshift medians))
    - **Claim:** Redshift (z) median values per model are: 0.47, 0.58, 0.17, 0.29, 0.39.

- **Checks:** min_max_range_check
- **Verdict:** PASS
- **Notes:** Compared computed min/max of provided medians to reported endpoints.

8. ✔ **C8** (Page 13, Table 5 vs Page 7, Table 2 (systematic-inclusive CI rule for Primary Mass))
    - **Claim:** Table 5 defines Systematic-Inclusive 90% CI as [min of all 5^{th} percentiles, max of all 95^{th} percentiles] across models. For Primary Mass, Table 5 gives [115.2, 167.5]. Table 2 lists 90% CIs: [121.4,150.7], [128.7,167.5], [138.2,162.3], [115.2,143.9], [119.7,152.3].
    - **Checks:** union_of_intervals_check
    - **Verdict:** PASS
    - **Notes:** Computed systematic-inclusive CI as min(lower bounds), max(upper bounds) and compared to reported Table 5 interval.
9. ✔ **C9** (Page 13, Table 5 vs Page 7, Table 2 (systematic-inclusive CI rule for Secondary Mass))
    - **Claim:** For Secondary Mass, Table 5 gives systematic-inclusive 90% CI [37.5, 127.6]. Table 2 90% CIs are: [95.2,125.2], [37.5,65.9], [73.4,111.4], [93.5,124.4], [91.6,127.6].
    - **Checks:** union_of_intervals_check
    - **Verdict:** PASS
    - **Notes:** Computed systematic-inclusive CI as min(lower bounds), max(upper bounds) and compared to reported Table 5 interval.
10. ✔ **C10** (Page 13, Table 5 vs Page 7, Table 2 (systematic-inclusive CI rule for χeff))
    - **Claim:** For χ_{eff}, Table 5 systematic-inclusive 90% CI is [-0.17, 0.63]. Table 2 90% CIs: [0.27,0.58], [0.15,0.50], [-0.17,0.19], [-0.12,0.48], [0.21,0.63].
    - **Checks:** union_of_intervals_check
    - **Verdict:** PASS
    - **Notes:** Computed systematic-inclusive CI as min(lower bounds), max(upper bounds) and compared to reported Table 5 interval.
11. ✔ **C11** (Page 13, Table 5 vs Page 7, Table 2 (systematic-inclusive CI rule for χp))
    - **Claim:** For χ_p, Table 5 systematic-inclusive 90% CI is [0.51, 0.95]. Table 2 90% CIs: [0.58,0.91], [0.71,0.92], [0.51,0.94], [0.59,0.95], [0.52,0.91].
    - **Checks:** union_of_intervals_check
    - **Verdict:** PASS
    - **Notes:** Computed systematic-inclusive CI as min(lower bounds), max(upper bounds) and compared to reported Table 5 interval.

12. ✔ **C12** (Page 13, Table 5 vs Page 7, Table 2 (systematic-inclusive CI rule for Final Mass))
    - **Claim:** For Final Mass, Table 5 systematic-inclusive 90% CI is [173.1, 255.4]. Table 2 90% CIs: [211.6,252.6], [173.1,217.2], [209.2,255.4], [199.0,245.1], [208.6,254.8].
    - **Checks:** union_of_intervals_check
    - **Verdict:** PASS
    - **Notes:** Computed systematic-inclusive CI as min(lower bounds), max(upper bounds) and compared to reported Table 5 interval.
13. ✔ **C13** (Page 13, Table 5 vs Page 7, Table 2 (systematic-inclusive CI rule for Final Spin))
    - **Claim:** For Final Spin a_f, Table 5 systematic-inclusive 90% CI is [0.61, 0.92]. Table 2 90% CIs: [0.84,0.92], [0.78,0.90], [0.61,0.77], [0.67,0.87], [0.81,0.92].
    - **Checks:** union_of_intervals_check
    - **Verdict:** PASS
    - **Notes:** Computed systematic-inclusive CI as min(lower bounds), max(upper bounds) and compared to reported Table 5 interval.
14. ✔ **C14** (Page 13, Table 5 vs Page 7, Table 2 (systematic-inclusive CI rule for Redshift))
    - **Claim:** For Redshift z, Table 5 systematic-inclusive 90% CI is [0.12, 0.74]. Table 2 90% CIs: [0.31,0.62], [0.38,0.74], [0.12,0.23], [0.15,0.52], [0.23,0.57].
    - **Checks:** union_of_intervals_check
    - **Verdict:** PASS
    - **Notes:** Computed systematic-inclusive CI as min(lower bounds), max(upper bounds) and compared to reported Table 5 interval.
15. ✔ **C15** (Page 8, Section 3.2.1 (PCA variance statement))
    - **Claim:** PC1 captures 26.8% of variance and PC2 captures 25.1% of variance; together they capture over 51% of total variance.
    - **Checks:** sum_of_percentages_check
    - **Verdict:** PASS
    - **Notes:** Checked that PC1+PC2 exceeds the claimed 'over' threshold.

Limitations

- Audit is limited to numeric values explicitly present in the provided PDF text/tables; posterior samples and underlying computations are not available.
- No verification that requires recomputing KDEs, Jensen–Shannon divergences, posterior overlaps, PCA/ICA fits, or reading curve values from plotted figures.

- Some properties (e.g., PCA orthonormality) are sensitive to rounding in reported tables, making strict numeric validation unreliable without unrounded values.
- Several numerical claims (JS divergence/overlap at pairwise or group level) cannot be verified without access to posterior samples or equivalent underlying data used to compute the metrics.
- Some claims depend on interpretive thresholds/wording (e.g., the meaning of 'twofold'), and are not strict cross-references to a separately reported numeric result.