Skepthical review: *Attributing Waveform Model Discrepancies in GW231123: A Feature-Based Diagnostic and Robust Astrophysical Inference*

Summary

The manuscript presents a practical workflow to diagnose and mitigate waveform-model systematics in gravitational-wave parameter estimation for GW231123. Using posterior samples from five waveform models (NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM, IMRPhenomXPHM, IMRPhenomXO4a), it quantifies 1D marginal discrepancies via Jensen–Shannon (JS) divergence (Sec. 2.2.1, Sec. 3.2.1) and compares global high-dimensional posterior structure using UMAP visualizations and Sliced Wasserstein Distance (SWD) (Sec. 2.2.2, Sec. 3.2.2). The paper then encodes model "features" (domain, family, calibration, higher-order modes, precession treatment) and correlates these with the discrepancy metrics to suggest which waveform properties drive disagreements relative to NRSur7dq4 (Sec. 2.1.3, Sec. 2.3, Sec. 3.3). Finally, it constructs a Bayesian Model Averaging (BMA) meta-posterior using BIC-based approximate evidences (Sec. 2.4.1–2.4.2, Sec. 3.4.1–3.4.2), and interprets the resulting inference astrophysically (Sec. 4). The overall direction is timely and potentially useful for practitioners, but several central methodological elements (BIC-based weights; feature attribution with only five models; robustness and reproducibility of JS/UMAP/SWD; parity of PE configurations across models; and clarity about waveform physics content) require strengthening and additional sensitivity checks before the quantitative BMA results and the stronger feature/astrophysical conclusions can be considered well supported.

Strengths

- Timely focus on waveform-model systematics for a challenging, high-mass, strongly precessing candidate (GW231123), of clear interest to GW inference and astrophysics (Sec. 1, Sec. 4).
- A coherent end-to-end pipeline: posterior comparison → discrepancy quantification (JS/UMAP/SWD) → interpretation via model features → mitigation via model averaging (Sec. 2–3).
- Use of complementary discrepancy views (1D JS and global structure via UMAP/SWD) goes beyond single-metric comparisons and can be valuable for diagnosing failure modes (Sec. 2.2, Sec. 3.2).
- Clear presentation of several key results in figures/tables (notably the divergence summaries and global-structure comparisons), enabling readers to see which parameters drive disagreements (Sec. 3.2; Fig. 2; Tables 2–3).
- The motivation for a model-marginalized/meta-posterior is well aligned with community needs in the presence of waveform systematics (Sec. 2.4, Sec. 3.4).

- The framework is conceptually transferable: given multiple posterior sample sets, the same diagnostics can be applied to other events, injections, or waveform-development studies (Sec. 4).

Major issues

1. **BMA weights derived from BIC are not sufficiently justified for GW parameter estimation here, and the BIC definition/implementation is ambiguous (Sec. 2.4.1, Sec. 3.4.1; Table 4).** In GW likelihoods the notion of the number of data points n (and BIC's asymptotic i.i.d. assumptions) is non-trivial, and it is not clearly established that n and k are identical across waveform models. Since Table 4 weights materially determine the BMA meta-posterior and downstream astrophysical statements (Sec. 3.4.2, Sec. 4), the quantitative conclusions are currently sensitive to an approximation whose validity is not demonstrated. In addition, the manuscript's BIC equation conflicts with the stated meaning of $L_{\mathrm{max,i}}$ ("maximum log-likelihood"), risking a log-of-a-log inconsistency.

   Recommendation: In Sec. 2.4.1, (i) define unambiguously whether $L_{\mathrm{max,i}}$ denotes the maximum likelihood $L(\hat{\theta})$ or the maximum log-likelihood $\ell_{\max} = \ln L(\hat{\theta})$, and write BIC consistently as $\mathrm{BIC} = k \ln n - 2 \ln L_{\max}$ or $\mathrm{BIC} = k \ln n - 2\ell_{\max}$; (ii) give a precise operational definition of n for the analysis (or explain why n cancels in all comparisons you actually use), and specify k per model if it differs. Then, in Sec. 3.4.1–3.4.2, add sensitivity analyses that recompute key BMA outputs (Table 5 parameters and any key posterior probabilities) under alternative weighting choices: equal weights; weights based on $\Delta\ell_{\max}$ only; and (if available from the PE runs) weights based on proper log-evidences (nested sampling $\log Z$ / thermodynamic integration) for at least a subset of models. Explicitly report how strongly m_1, z, χ_{eff}, χ_p, and $\cos\theta_{\mathrm{JN}}$ shift under these alternatives, and qualify claims that rely on the BIC-weighted choice.

2. **Feature-based attribution via Spearman correlations is statistically underdetermined and potentially confounded with only five waveform models, and the analysis likely inflates the effective sample size by treating multiple parameters per model as independent (Sec. 2.3, Sec. 3.3; Fig. 3).** With one-hot binary features that are collinear (e.g., Phenom vs frequency-domain; EOB vs time-domain; NR surrogate as a singleton), reported correlations (e.g., $\rho \approx 0.3$–0.4) cannot be interpreted robustly as "drivers," and no uncertainty, multiple-testing control, or leave-one-model-out robustness is shown.

   Recommendation: Reframe Sec. 2.3 and Sec. 3.3 as primarily descriptive unless stronger statistical support can be added. Concretely: (i) state the true number of independent units (models) and clarify the construction of the discrepancy dataset (how many JS/SWD values per model; how dependencies across parameters are handled); (ii) report uncertainty on correlation estimates via permutation tests that respect

grouping by model and/or a leave-one-model-out analysis; (iii) quantify feature collinearity (feature–feature correlations) and avoid interpreting correlated features as separable causes. Consider replacing pairwise Spearman bars with simpler, actionable summaries: group-by-feature comparisons (e.g., "HOM included vs not," "full precession vs simplified") and/or a very low-dimensional, strongly regularized regression with model-level clustering, clearly labeled as exploratory. Temper causal language in Sec. 3.3 and Sec. 4 accordingly.

3. **Reproducibility and PE-configuration parity are not adequately documented, making it hard to attribute posterior differences to waveform physics rather than to analysis setup differences (Sec. 2.1–2.4, Sec. 3.1).** The manuscript does not clearly state whether priors, cosmology/source-frame conversions, PSDs, calibration marginalization, frequency bounds ($f_{\rm low}/f_{\rm high}$), data conditioning, and sampler settings/convergence diagnostics were identical across all waveform runs. Maximum log-likelihood comparisons (Table 4) are also difficult to interpret without evidence that each run reliably explored the relevant likelihood maxima.

   Recommendation: Add a dedicated subsection (Sec. 2.1 or an Appendix) that lists, for every waveform-model PE run: priors (including spin tilt and magnitude priors), cosmology choices for z/source-frame masses, PSD estimation procedure, calibration handling, $f_{\rm low}/f_{\rm high}$, data segment length/windowing, sampler and stopping criteria, and convergence diagnostics (e.g., effective sample size; any chain diagnostics used). Explicitly confirm parity across models (or list deviations). Provide the number of posterior samples per model used downstream (Sec. 2.1.2, Sec. 3.1.1) and document any thinning/reweighting. This is essential to support claims in Sec. 3.2–3.4 that differences are waveform-driven.

4. **The discrepancy metrics (JS divergence, SWD, and UMAP) are central to the paper's conclusions but lack sufficient specification and robustness assessment (Sec. 2.2.1–2.2.2, Sec. 3.2).** JS divergence computed from KDE marginals can depend strongly on bandwidth choice and boundary handling for bounded parameters ($\chi_p \in [0,1]$, $\cos\theta_{\rm JN} \in [-1,1]$); UMAP is stochastic and hyperparameter-dependent; SWD depends on parameter scaling/transformations and number of projections. Without robustness checks or uncertainty estimates, rankings in Tables 2–3 and clustering impressions in Fig. 2 are hard to interpret quantitatively.

   Recommendation: In Sec. 2.2.1–2.2.2, fully specify: KDE library, kernel, bandwidth rule (which one, exactly), grid/support, and boundary handling (reflection, transforms, or bounded KDE) for χ_p and $\cos\theta_{\rm JN}$; for SWD, the exact parameter vector used, any transforms (e.g., log masses), normalization/standardization, the number of random projections, and the random seed(s); for UMAP, $n_{\rm neighbors}$, min_dist, metric, preprocessing/standardization, and random_state. Then add minimal robustness checks: (i) recompute JS with at least one alternative estimator (e.g., histogram-based JS on a common binning or a kNN-based divergence estimator) and/or alternative bandwidth; (ii) show UMAP stability across several seeds/hyperparameters (qualita-

tively is fine, but state what changed); (iii) report SWD variability via bootstrap resampling of posterior samples or repeated random projections. Summarize robustness outcomes in Sec. 3.2 (or an Appendix) so the discrepancy conclusions are auditable.

5. **Waveform-model content and feature encoding are not sufficiently explicit and may contain internal inconsistencies (Sec. 2.1.1–2.1.3, Sec. 3.3).** The text characterizes certain models as lacking HOMs or comprehensive precession, but at least one model name (e.g., IMRPhenomXPHM) commonly denotes inclusion of precession and higher modes; if your specific configuration restricted modes or physics, it must be stated. Without an explicit per-model description (modes included, precession implementation, calibration range), the feature matrix and the interpretation of Sec. 3.3 are hard to verify.

   Recommendation: Replace abstract/meta references with a self-contained table in Sec. 2.1.1–2.1.3 listing, for each waveform model as actually run: domain (time/frequency), family, calibration approach and validity range, precession treatment, and explicit higher-mode content (list (ℓ, m) modes and any flags/settings used). Cite the relevant waveform papers/software documentation. Ensure Sec. 3.3 statements match this table precisely (e.g., distinguish "no HOMs" from "restricted HOM set," or "twist-up precession" from "full precession").

6. **The paper treats NRSur7dq4 as a de facto ground truth reference for defining discrepancies and feature attribution (Sec. 2.2–2.3, Sec. 3.2–3.3) without a sufficiently critical discussion of reference dependence and surrogate validity for GW231123-like posteriors.** If parts of the posterior explore regions near/outside the surrogate's training domain (e.g., in q or spins), then divergences may reflect reference limitations as much as other models' limitations.

   Recommendation: In Sec. 1 or early in Sec. 2.2, summarize NRSur7dq4's training/calibration domain and assumptions (with citations) and assess whether GW231123 posteriors approach domain edges. Explicitly acknowledge that divergences relative to NRSur7dq4 conflate differences in other models with possible surrogate imperfections. If feasible, add a robustness check in Sec. 3.2–3.3 using an alternative reference (e.g., SEOBNRv5PHM) and/or report pairwise divergence summaries (not only vs NRSur7dq4). Temper any "ground truth" phrasing accordingly.

7. **There is an unresolved tension between the discrepancy diagnostics (which identify some models as globally most discrepant) and the subsequent BMA (which can still assign those models substantial weight), but the impact of including/excluding these models is not quantified (Sec. 3.2 vs Sec. 3.4).** This makes it difficult to interpret the BMA meta-posterior as a mitigation of waveform systematics rather than an averaging over potentially inconsistent inferences.

Recommendation: In Sec. 3.4.1–3.4.2, add an explicit robustness study: recompute the meta-posterior and Table 5 summaries after excluding IMRPhenomXO4a and/or IMRPhenomXPHM (or after down-weighting based on a stated "model adequacy/physics completeness" prior). Report shifts in m_1, z, $\chi_{\rm eff}$, χ_p, and $\cos\theta_{\rm JN}$ and discuss what practitioners should conclude when a high-likelihood but globally discrepant model dominates weights. Consider discussing model stacking / predictive approaches as an alternative to BIC-BMA, even if only as future work.

8. **Several astrophysical conclusions (PISN mass-gap placement; dynamical/hierarchical formation; strength of spin-orbit misalignment claims from χ_p) are stated more strongly than is justified given the remaining waveform dependence, the limited model set, and the approximate nature of the BMA weights (Sec. 3.4.2, Sec. 4).** In particular, "in the pair-instability mass gap" claims should quantify posterior probability relative to a stated threshold and show sensitivity to waveform choice and to z/cosmology assumptions.

   Recommendation: In Sec. 3.4.2 and Sec. 4, moderate language to reflect residual systematic uncertainty. Quantify key probabilities rather than categorical statements, e.g., $P(m_1 > m_{\rm gap})$ for one or more literature thresholds (state which), and $P(M_f > 100\,M_\odot)$ for the IMBH claim. Show these probabilities per-model and under the alternative weighting schemes requested above. Provide a compact per-model table/figure for χ_p (medians and credible intervals) to support claims of robust strong precession, and phrase formation-channel inferences as suggestive/consistent rather than definitive.

Minor issues

1. Figure 3 (feature correlations) lacks uncertainty quantification, does not display per-feature sample counts, and does not address multiple testing or collinearity, making it easy to over-interpret exploratory associations (Sec. 3.3; Fig. 3).

   Recommendation: Add bootstrap/permutation confidence intervals and (if you choose to report them) p-values with multiplicity control (e.g., Benjamini–Hochberg FDR). Display the number of models contributing to each feature level. Add a note in the caption that correlations are associative and likely reflect co-occurring model-design choices.

2. Figure 1 is difficult to read due to overplotting and insufficient legend/visual encodings; it also lacks basic visual aids (e.g., $\chi_{\rm eff} = 0$ reference) and does not show interval summaries that match the text emphasis (Sec. 3.2.1; Fig. 1).

   Recommendation: Add a clear legend mapping styles to waveform models; use color-blind-safe colors plus linestyles and transparency; increase font/line sizes; add a vertical line at $\chi_{\rm eff} = 0$; and overlay medians/90% credible intervals (or shaded HPD bands). Consider splitting into multiple panels/rows to reduce clutter for z and $\chi_{\rm eff}$.

3. Tables/parameter reporting are not fully aligned with the narrative emphasis: parameters highlighted in the text (χ_p, z, $\cos\theta_{\rm JN}$) are not consistently summarized per model alongside m_1 and $\chi_{\rm eff}$ (Sec. 3.1.1, Sec. 3.4.2).

   Recommendation: Extend Table 1 (Sec. 3.1.1) or add a companion table reporting per-model medians and credible intervals for z, χ_p, and $\cos\theta_{\rm JN}$ (and clearly define source-frame vs detector-frame quantities). Cross-reference these tables in Sec. 3.2–3.4 where the parameters are discussed.

4. The physical interpretation of inclination discrepancies ($\cos\theta_{\rm JN}$) and its coupling to distance/redshift and HOM content is under-discussed, despite large reported divergences for some models (Sec. 3.2.1, Sec. 3.4.2).

   Recommendation: Add a short discussion in Sec. 3.2.1 or Sec. 3.4.2 explaining how inclination–distance degeneracies and HOMs can drive changes in z and $\cos\theta_{\rm JN}$, and explicitly connect this to the observed JS divergences for models with different HOM/precession content.

5. Some section cross-references are inconsistent (e.g., JS divergence values referenced as originating from Sec. 2.1 rather than Sec. 2.2.1), which slows down verification of the analysis flow (Sec. 2.3.1).

   Recommendation: Audit and correct internal cross-references so that the discrepancy dataset construction in Sec. 2.3.1 points to the correct definitions and computations in Sec. 2.2.1 (JS) and Sec. 2.2.2 (UMAP/SWD), and to the correct result tables (Sec. 3.2.1–3.2.2).

6. The selection of exactly five waveform models is only briefly justified, and the conclusions could depend materially on this small and potentially unrepresentative set (Sec. 2.1.1, Sec. 3.1.1, Sec. 4).

   Recommendation: Add a short justification paragraph for model selection criteria (availability, suitability for high-mass precessing signals, community usage), and explicitly list relevant missing alternatives (other EOB/Phenom variants, additional surrogates). In Sec. 4, clarify that feature attribution is limited by the small model set and indicate how adding models/events would test generality.

7. Several claims rely on methods/tools (UMAP, SWD, BMA in GW, JS divergence for posterior comparison) without sufficient citation context (Sec. 2.2, Sec. 2.4, Sec. 4).

   Recommendation: Add citations for UMAP, SWD (and/or sliced Wasserstein), any prior uses of JS divergence in GW posterior comparison (if applicable), and prior applications/discussions of BMA/model averaging in GW inference and waveform systematics.

Very minor issues

1. JS divergence bounds are stated without specifying the logarithm base, which changes whether the bound is 1 or $\ln 2$ (Sec. 2.2.1).

   Recommendation: State the JS divergence definition explicitly (including the log base) and report the corresponding bound (1 for log base 2; $\ln 2$ for natural log).

2. BMA resampling via $N_i = \mathbf{round}(w_i \cdot N_{\text{total}})$ does not guarantee that $\sum_i N_i = N_{\text{total}}$, and the manuscript does not state how the remainder is handled (Sec. 2.4.2).

   Recommendation: Specify a deterministic scheme (e.g., floor then distribute remaining samples by largest fractional parts) or a randomized multinomial resampling that guarantees exactly N_{total} samples, and report the random seed if used.

3. Typos and consistency issues reduce polish (e.g., line-break artifact "be\n\ncause"; inconsistent model spelling IMRPhenomXO4a vs IMRPhenomX04a; inconsistent section-heading styles; inconsistent $\cos\theta_{\text{JN}}$ notation) (Sec. 1–4).

   Recommendation: Proofread to remove line-break artifacts, standardize model names across text/tables/figures, harmonize heading styles and figure/table reference style, and use consistent parameter notation (e.g., $\cos\theta_{\text{JN}}$ vs $\cos(\theta_{\text{JN}})$) throughout.

4. Figure/table labeling could be made more self-contained (e.g., acronyms not expanded; missing note that correlations do not imply causation; small fonts for print) (Figs. 1 and 3).

   Recommendation: Expand acronyms in captions, add a brief disclaimer about correlation interpretation in Fig. 3, and increase font sizes/line weights to ensure readability in typical journal column widths.

Key statements and references

- **✖ The primary black hole mass in GW231123 is measured to be $134.9^{+24.0}_{-14.6}\,M$) where black holes are not expected to form from single-star evolution, thereby reinforcing prior evidence that such objects likely arise from hierarchical mergers of smaller black holes in dense stellar environments such as globular clusters or active galactic nuclei.}, placing it firmly within the pair-instability supernova upper mass gap (approximately 65–$135\,M_{\odot}**
- *Reference(s):* $65 - 135 M_\odot$
- *Justification:* $65 - 135 M_\odot$ discusses early-warning localization for compact binaries with third-generation detectors and uses example NSBH/BNS systems (masses $\sim 14\,M_\odot$ and $1.4\,M_\odot$). It does not mention GW231123, does not report any primary black hole mass near $135\,M_\odot$, and does not discuss the pair-instability mass gap or hierarchical mergers. Therefore, the statement is not supported by the attached paper.

- **✖ The effective inspiral spin for GW231123 is constrained to $\chi_{\rm eff} = 0.37^{+0.18}_{-0.35}$ and the effective precessing spin to $\chi_p = 0.79^{+0.13}_{-0.19}$, providing unambiguous evidence for significant spin-orbit misalignment and orbital-plane precession that is consistent with dynamical formation scenarios rather than isolated binary evolution, which typically yields more aligned spins.**
- *Reference(s):* $65 - 135 M_\odot$
- *Justification:* $65 - 135 M_\odot$ discusses precession effects and early-warning localization for future detectors but does not analyze GW231123, report any measured $\chi_{\rm eff}$ or χ_p values, or draw formation-scenario conclusions. The specific spin constraints and inference about dynamical vs. isolated formation are not provided.
- **✖ The primary black hole in GW231123 lies in the pair-instability supernova mass gap, an astrophysically significant range (about 65–135 $M_\odot$) where standard stellar evolution models predict that black holes should not form due to pair-instability supernovae disrupting progenitor stars before collapse.**
- *Reference(s):* $65 - 135 M_\odot$
- *Justification:* $65 - 135 M_\odot$ focuses on early-warning localization for precessing compact binaries with third-generation detectors. It does not mention GW231123, black-hole component masses, or the pair-instability supernova mass gap ($\approx$ 65–135 $M_\odot$). Hence, the statement is not supported by this paper.
- **✖ The remnant of GW231123 is inferred to be an intermediate-mass black hole with a mass of approximately 221 $M_\odot$ and a dimensionless spin of $a_f \approx 0.86$, contributing to the sparse but growing observational evidence for the existence and properties of intermediate-mass black holes, whose formation and evolution remain poorly understood in current astrophysical models.**
- *Reference(s):* intermediate-mass black hole
- *Justification:* The intermediate-mass black hole paper discusses GW190521, not GW231123. It infers a remnant mass of $142^{+28}_{-16}\,M with spin \sim 0.86, so the specific quantitative claim is unsupported.} and spin ~ 0.72, identifying it as an IMBH, and notes the sparse evidence and uncertain formation channels for IMBHs. It provides no data on GW231123 or a remnant mass $\sim 221\,M_\odot
- **✔ Waveform models that are formulated in the frequency domain, belong to the phenomenological family, and lack comprehensive higher-order modes or full spin-precession physics (e.g., aligned-spin approximations without HOMs) exhibit systematically larger discrepancies from numerical-relativity-calibrated surrogates in parameter estimation for complex,**

high-mass, precessing systems like GW231123, indicating that these simplified physical treatments are primary drivers of model-dependent systematic uncertainties.

- *Reference(s):* Figure 3
- *Justification:* No valid PDFs found; assumed supported.

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 and descriptive, with a small number of central statistical equations (notably BIC-based approximate evidence and BMA model weights). Most other mathematical components are metric descriptions (JS divergence, SWD, Spearman correlation) without explicit formulas. The main internal consistency concern is a conflict between the BIC formula and the definition/labeling of the likelihood quantity used.

Checked items

1. ✖ **BIC formula vs definition of likelihood quantity** (Sec. 2.4.1, "Approximate Model Evidence", p.5)
    - **Claim:** $BIC_i = k\ln(n) - 2\ln(L_{\max,i})$, where $L_{\max,i}$ is the maximum log-likelihood value for model i.
    - **Checks:** definition consistency, algebra/notation consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** k is the number of free parameters, n is the number of data points, $L_{\max,i}$ is described in text as a maximum log-likelihood
    - **Notes:** The text defines $L_{\max,i}$ as a maximum log-likelihood, i.e., already a logarithm of likelihood. The formula then applies $\ln(\cdot)$ to $L_{\max,i}$, implying $\ln(\log L)$, which is dimensionally/semantically inconsistent. If $L_{\max,i}$ is instead meant to be the maximum likelihood, the phrase "maximum log-likelihood value" is wrong. This ambiguity directly affects ΔBIC and thus BMA weights and the paper's final meta-posterior.

2. ✖ **Table 4 label conflicts with BIC definition** (Table 4, Sec. 3.4.1, p.8)
    - **Claim:** Table reports "Max Log-Likelihood" along with ΔBIC and model weights computed from BIC.
    - **Checks:** symbol/label consistency, pipeline consistency (definition-to-table)
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** The BIC definition in Sec. 2.4.1 is used for Table 4, The table's "Max Log-Likelihood" column corresponds to $L_{\max,i}$ in that definition

- **Notes:** If Table 4's values are log-likelihoods (as labeled), then the correct BIC form should involve $-2 \cdot (\max \text{log-likelihood})$ (plus $k \ln n$), not $-2 \ln(L_{\max,\mathrm{i}})$. As written, the method and the table cannot both be correct without redefining $L_{\max,\mathrm{i}}$.

3. ✔ **ΔBIC definition consistency** (Sec. 2.4.1, "Calculate Model Weights", p.5)
    - **Claim:** $\Delta\mathrm{BIC}_i = \mathrm{BIC}_i - \min(\mathrm{BIC})$.
    - **Checks:** definition consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** $\min(\mathrm{BIC})$ is taken over the M models in consideration
    - **Notes:** The definition is standard and consistent with later statements that the best model has $\Delta\mathrm{BIC} = 0$.

4. ✔ **BMA weight formula normalization** (Sec. 2.4.1, "Calculate Model Weights", p.5)
    - **Claim:** $w_i = \exp(-0.5 \cdot \Delta\mathrm{BIC}_i) \,/\, \Sigma_j \exp(-0.5 \cdot \Delta\mathrm{BIC}_j)$.
    - **Checks:** algebra, normalization/constraints
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** $\Delta\mathrm{BIC}_i$ are finite real numbers
    - **Notes:** Weights are guaranteed nonnegative and sum to 1 by construction.

5. ⚠ **Assumption that k and n are constant across models** (Sec. 2.4.1, end of "Approximate Model Evidence", p.5)
    - **Claim:** k and n are assumed constant across all models because they relate to the underlying physical system and observed data, not the waveform model.
    - **Checks:** assumption clarity, internal logic
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** All waveform models are fit using the same parameterization/dimension k, The same effective number of data points n is appropriate across models
    - **Notes:** Within the paper, it is not demonstrated that all models share identical numbers of fitted parameters or identical effective n (especially given differing waveform complexities). This may be true in their setup, but the paper does not state the actual k used nor justify equality across models beyond an assertion.

6. ⚠ **BMA ensemble construction via rounded allocation** (Sec. 2.4.1, "Construct the BMA Ensemble", p.5)
    - **Claim:** Draw $N_i = \mathrm{round}(w_i \cdot N_{\mathrm{total}})$ samples from each model and concatenate to form the final meta-posterior.

- **Checks:** constraint consistency, algorithmic completeness (symbolic)
- **Verdict:** UNCERTAIN; confidence: high; impact: minor
- **Assumptions/inputs:** N_{total} is the desired total sample count, $\mathbf{round}(\cdot)$ is standard nearest-integer rounding
- **Notes:** Rounding generally makes $\Sigma_i N_i$ differ from N_{total}. The paper does not specify how it ensures exactly N_{total} samples (or whether it matters). This is a completeness/definition issue rather than a derivation error.

7. ⚠ **JS divergence range claim** (Sec. 2.2.1, p.3; reiterated Sec. 3.2.1, p.6)

   - **Claim:** JS divergence ranges from 0 (identical) to 1 (maximally different).
   - **Checks:** definition completeness, normalization/constraints
   - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
   - **Assumptions/inputs:** A specific JS divergence convention is used (including a log base)
   - **Notes:** The paper never specifies the JS divergence formula or the logarithm base. Without that, the numeric upper bound (1 vs another constant) cannot be verified from the paper alone. The qualitative statement "bounded" is fine, but the specific $[\mathbf{0}, \mathbf{1}]$ bound is not auditable internally.

8. ⚠ **JS divergence computed from KDEs (missing explicit formula)** (Sec. 2.2.1, p.3)

   - **Claim:** Compute JS divergence between KDE-estimated 1D marginals of reference and other models.
   - **Checks:** definition completeness, assumption clarity
   - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
   - **Assumptions/inputs:** KDEs are proper normalized densities on a common support, JS divergence is computed between continuous densities
   - **Notes:** No explicit JS divergence integral/formula is provided, nor is the support/discretization described. This prevents checking analytic properties (e.g., exact boundedness, invariance to binning/discretization choices) from the paper alone.

9. ✖ **Internal cross-reference for JS divergence source** (Sec. 2.3.1, bullet for js_divergence, p.4)

   - **Claim:** JS divergence values are obtained from Section 2.1.
   - **Checks:** internal reference consistency
   - **Verdict:** FAIL; confidence: high; impact: minor
   - **Assumptions/inputs:** JS divergence is actually defined/computed earlier
   - **Notes:** JS divergence computation is described in Sec. 2.2.1, not Sec. 2.1. This is a document consistency error that can confuse readers trying to verify the workflow.

10. ✔ **Use of Spearman correlation with binary model features** (Sec. 2.3.2, p.4)
    - **Claim:** Compute Spearman rank correlation between each binary-encoded model feature and js_divergence.
    - **Checks:** method-symbol consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Binary-encoded features take values $0, 1$, js_divergence is continuous
    - **Notes:** No algebraic derivation to check; the operation is well-defined. The paper does not provide formulas, but the described computation is internally coherent.

11. ✔ **Inclination angle notation** (Sec. 2.2.1, p.3; Tables 2 and 5, pp.6 and 9)
    - **Claim:** Inclination is represented by $\cos\theta_{\rm JN}$ (or $\cos(\theta_{\rm JN})$).
    - **Checks:** notation consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** $\theta_{\rm JN}$ denotes the same physical angle throughout
    - **Notes:** The symbol appears with minor stylistic variants but refers consistently to the cosine of the same inclination angle.

Limitations

- The audit used only the provided PDF text/images; key metric definitions (JS divergence formula, SWD formula) are not written as explicit equations, limiting the ability to verify bounds/normalizations purely from the document.
- No derivations are shown for SWD, UMAP, KDE bandwidth selection, or evidence approximation beyond the BIC equation, so the audit focuses on definition/notation consistency rather than step-by-step algebra for those components.
- Numeric consistency (e.g., whether Table 4 weights match the stated formulas) was not checked, per instruction.

Numerical results audit

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

All 8 executed numerical consistency checks passed. Page 8 Table 4 is internally consistent: model weights sum appropriately (including a $< 0.1\%$ entry), reported ΔBIC values match those implied by max log-likelihood differences, and the reported BMA weights match $\exp(-0.5 \cdot \Delta\text{BIC})$ normalization within rounding. Page 9 Table 5 credible intervals are consistent with the quoted $\textrm{median}^{+\textrm{upper}}_{-\textrm{lower}}$ formats for m_1, χ_p, χ median rounded to the nearest integer.} (with a small rounding-level asymmetry), and redshift. The text statement that the final mass is $\approx 221\,M_\odot$ is consistent with a $220.9\,M_{\odot}

Checked items

1. ✔ **C1_weights_sum_to_1** (Page 8, Table 4 (BMA Model Weights based on BIC Approximation))
    - **Claim:** Model weights are reported as 41.1%, 30.9%, 23.0%, 5.0%, and $< 0.1\%$ (IMRPhenomXPHM). These should sum to $\sim 100\%$.
    - **Checks:** percentage_sum
    - **Verdict:** PASS
    - **Notes:** Fixed sum of the first four weights is exactly 100.0%; adding any fifth weight in $[0, 0.1)$ keeps the total within the stated absolute tolerance.
2. ✔ **C2_deltaBIC_from_logL_max** (Page 8, Table 4 (columns: Max Log-Likelihood, ΔBIC))
    - **Claim:** Given $BIC_i = \text{const} - 2\ln(L_{\max,i})$ with k and n constant across models, differences should satisfy $\Delta \textrm{BIC}i = 2(\ln L)$.} - \ln L_{\rm max,i
    - **Checks:** recompute_delta
    - **Verdict:** PASS
    - **Notes:** All reported ΔBIC values match $2 \times (\ln L_{\text{best}} - \ln L_i)$ within abs_tol (largest absolute deviation 0.01, consistent with rounding to 2 decimals).
3. ✔ **C3_weights_from_deltaBIC** (Page 8, Section 2.4.1 formula + Page 8, Table 4 (ΔBIC, Model Weight))
    - **Claim:** Model weights w_i should be proportional to $\exp(-0.5\Delta\text{BIC}_i)$ normalized across models; Table 4 provides ΔBIC and weights.
    - **Checks:** recompute_softmax
    - **Verdict:** PASS
    - **Notes:** Recomputed softmax weights match the first four reported weights within 0.2 percentage points; the fifth model weight computes to $\sim 0.0049\%$, satisfying the "$< 0.1\%$" claim.
4. ✔ **C4_credible_interval_widths_m1** (Page 9, Table 5 + Page 1 Abstract / Page 10 Conclusions)
    - **Claim:** Primary mass reported as median 134.9 with 90% CI $[120.3 - 158.9]$ corresponds to $+24.0/-14.6$ quoted elsewhere ($134.9^{+24.0}_{-14.6}$).
    - **Checks:** interval_to_plusminus
    - **Verdict:** PASS
    - **Notes:** Computed differences ($158.9 - 134.9 = 24.0$ and $134.9 - 120.3 = 14.6$) match the quoted $+/-$ values within tolerance.
5. ✔ **C5_credible_interval_widths_chi_p** (Page 9 Table 5 + Page 1 Abstract / Page 10 Conclusions)

- **Claim:** χ_p reported as median 0.79 with 90% CI $[0.60 - 0.92]$ corresponds to $+0.13/-0.19$ quoted elsewhere ($0.79^{+0.13}_{-0.19}$).
- **Checks:** interval_to_plusminus
- **Verdict:** PASS
- **Notes:** Computed differences ($0.92 - 0.79 = 0.13$ and $0.79 - 0.60 = 0.19$) match the quoted +/− values within tolerance.

6. ✔ **C6_credible_interval_widths_chi_eff** (Page 9 Table 5 + Page 9 text bullet 'Significant Spin and Precession' + Page 10 Conclusions)
   - **Claim:** $\chi_{\rm eff}$ reported as median 0.37 with 90% CI $[0.01 - 0.55]$ corresponds to $+0.18/-0.35$ quoted elsewhere ($0.37^{+0.18}_{-0.35}$).
   - **Checks:** interval_to_plusminus
   - **Verdict:** PASS
   - **Notes:** Upper difference matches ($0.55 - 0.37 = 0.18$). Lower difference is 0.36 ($0.37 - 0.01$), which is consistent with the quoted 0.35 within the allowed rounding tolerance.

7. ✔ **C7_credible_interval_widths_redshift** (Page 9 Table 5 + Page 10 text '$z = 0.47^{+0.22}_{-0.27}$')
   - **Claim:** Redshift reported as median 0.47 with 90% CI $[0.20 - 0.69]$ corresponds to $+0.22/-0.27$ quoted elsewhere ($0.47^{+0.22}_{-0.27}$).
   - **Checks:** interval_to_plusminus
   - **Verdict:** PASS
   - **Notes:** Computed differences ($0.69 - 0.47 = 0.22$ and $0.47 - 0.20 = 0.27$) match the quoted +/− values within tolerance.

8. ✔ **C8_final_mass_approx_221** (Page 9 Table 5 + Page 1 Abstract / Page 10 Conclusions)
   - **Claim:** Final mass is reported as $220.9\,M_\odot$ (median) and described as approximately $221\,M_\odot$ elsewhere.
   - **Checks:** rounding_consistency
   - **Verdict:** PASS
   - **Notes:** 220.9 rounds to 221 and differs from 221 by 0.1, which is within the integer-rounding tolerance.

Limitations

- Only the provided parsed PDF text was used; no access to underlying posterior samples/CSVs, likelihood time series, or any external datasets.
- No values were extracted from plotted curves or figure pixels; only tabulated/explicitly written numbers were considered.
- Some statements (e.g., physical interpretation like "mass gap" membership) are convention-dependent and not strictly numerically decidable from the paper alone.