Skepthical review: *Dissecting Multi-Model Posterior Landscapes of GW231123: Unveiling Intrinsic Degeneracies via Mode-Finding and Shared Manifold Analysis*

Summary

The manuscript presents a multi-model posterior comparison for the gravitational-wave event GW231123 using five waveform models (NRSur7dq4, IMRPhenomXO4a, SEOBNRv5PHM, IMRPhenomXPHM, IMRPhenomTPHM). The analysis proceeds in a clear pipeline: (i) quantify inter-model differences via Jensen–Shannon (JS) divergence on 1D/2D marginalized posteriors (Sec. 2.1, Sec. 3.1); (ii) diagnose intra-model multimodality with HDBSCAN clustering in the full 14D feature space (Sec. 2.2, Sec. 3.2); and (iii) build a joint UMAP embedding ("degeneracy manifold") of all posterior samples and compare models via KDE densities and JS divergence in the embedded space (Sec. 2.3–2.4, Sec. 3.3–3.4). The headline qualitative outcome is that all models support a high-mass system with strong precession (high χ_p), while mass ratio, $\chi_{\rm eff}$, and viewing angle $\cos\theta_{jn}$ show pronounced model dependence, with the UMAP visualization suggesting three distinct solution "islands" broadly aligned with waveform-family groupings. The core idea—using a unified embedding to compare high-dimensional posterior structure across waveform approximants—is timely and potentially broadly useful. However, several key methodological details are currently missing (posterior provenance, priors/likelihood consistency, JS/KDE implementation, HDBSCAN/UMAP hyperparameters), and robustness checks are insufficient to support strong quantitative interpretations of JS values, clustering-derived "modes," and UMAP "islands." Strengthening reproducibility, sensitivity analyses, and the linkage between embedding structure and physically/likelihood-meaningful differences would substantially improve the paper's reliability and impact.

Strengths

- Addresses an important and timely problem: diagnosing waveform-systematics and high-dimensional degeneracy structure for a complex BBH event (Sec. 1).
- A sensible, layered comparison pipeline (marginal divergences → full-dimensional clustering → shared manifold visualization) that goes beyond standard corner-plot comparisons (Sec. 2, Sec. 3).
- Novel and intuitive use of a *joint* UMAP embedding to visualize and compare posterior topology across multiple waveform models (Sec. 2.3–2.4, Sec. 3.3).
- HDBSCAN-based exploration usefully highlights potential inclination-related multimodality in some waveform families and differences in how other models resolve degeneracies (Sec. 3.2).
- Clear narrative focus on the physically salient parameters (q, $\chi_{\rm eff}$, χ_p, $\cos\theta_{jn}$) and their relevance to formation-channel interpretation (Sec. 3.1–3.4, Sec. 4).

- Figures 2–4 (conceptually) provide a compact way to summarize complex inter-model differences in posterior structure, which could become a useful diagnostic template if accompanied by stronger robustness documentation.

Major issues

1. **Insufficient description of the *parameter-estimation provenance and inference context* makes it hard to attribute differences to waveform systematics rather than analysis configuration. The manuscript operates on posterior samples but does not clearly document (i) where they come from (public release vs in-house), (ii) whether *priors are identical across models* (mass, spin magnitude/tilt, distance/orientation), (iii) whether the same likelihood/PSD/calibration marginalization/data segment were used, and (iv) sampler settings and convergence diagnostics (Sec. 2.1).** If priors differ even slightly, JS divergences and "model-dependent" conclusions can be confounded.

   Recommendation: Add a dedicated subsection in Sec. 2.1 (e.g., "Parameter-estimation inputs") documenting for each waveform model: data segment, detector network context, PSD estimation, calibration uncertainty treatment, likelihood form, reference frequency and parameter conventions, sampler and settings, convergence metrics (ESS / stopping), and the *exact priors* used. State explicitly whether priors are identical across models; if not, either (i) reweight posteriors to a common prior where feasible, or (ii) quantify and discuss the expected impact on comparisons (especially for χ_p, $\chi_{\rm eff}$, and orientation). Provide citations/run IDs/links for the posterior files.

2. **The JS divergence calculations are central but not specified with enough rigor to audit, reproduce, or interpret quantitatively (Sec. 2.1.2, Sec. 2.3.2, Sec. 3.1, Sec. 3.3).** Missing items include: the exact JS definition implemented (continuous vs discretized; log base and range), KDE implementation details (bandwidth rule, covariance structure), evaluation grid/bounds/resolution, normalization after discretization, treatment of zero-probability bins (regularization ϵ), and boundary effects for bounded parameters (e.g., $\cos\theta_{jn} \in [-1, 1]$, spin magnitudes in $[0, 1]$). As written, strong statements based on values like JS $\sim$ **0.6–0.7** are difficult to calibrate or trust without estimator uncertainty and bias checks.

   Recommendation: In Sec. 2.1.2 and Sec. 2.3.2, provide the explicit mathematical JS formula used (including log base) and the numerical procedure: KDE method, bandwidth selection, grid definition (shared support across model pairs), grid resolution, renormalization, and zero-handling. Add uncertainty/sensitivity estimates: bootstrap over posterior samples and/or vary KDE bandwidth (and, for bounded variables, demonstrate robustness via reflection/logit-transform/Beta-kernels or justify why boundary bias is negligible). If JS values are used as "high/low" indicators, define

thresholds relative to the JS maximum under your log base. Consider adding a non-KDE cross-check (e.g., histogram-based JS at multiple binning levels or a kNN-based divergence estimator) to demonstrate that qualitative conclusions are stable.

3. **UMAP is heavily relied upon to infer three "islands" and to compute model-to-model JS divergences in the embedded space, but embedding configuration, stability, and interpretability are not demonstrated (Sec. 2.3.1, Sec. 3.3–3.4).** UMAP can create apparent separations depending on hyperparameters ($n_{\text{neighbors}}$, min_dist, metric), random seed, and density differences across models; additionally, UMAP axes are not uniquely interpretable, and distances are not likelihood-preserving. The manuscript currently risks over-interpreting island structure and UMAP-space JS as "global posterior similarity."

   Recommendation: In Sec. 2.3.1, report *all* UMAP hyperparameters ($n_{\text{neighbors}}$, min_dist, metric, n_{epochs}, $random_state$/seed, any non-default settings) and justify choices. In Sec. 3.3 or an Appendix, add stability tests across multiple seeds and a reasonable hyperparameter grid; quantify stability (e.g., Procrustes-aligned embeddings; correlation of inter-point distance matrices; trustworthiness/continuity). Explicitly caution that UMAP axes are unitless/arbitrary, and quantify axis–parameter relationships via Spearman/Pearson correlations (or local regression) rather than deterministic claims. For UMAP-space JS (Sec. 3.3), either justify it with embedding-fidelity diagnostics or reframe it as heuristic, and add at least one complementary *high-dimensional* two-sample measure not relying on UMAP (e.g., classifier two-sample test AUC, MMD/energy distance) to corroborate which model pairs are genuinely distinguishable in 14D.

4. **HDBSCAN clustering underpins key claims about bimodality/unimodality and "statistically significant modes," but parameter choices, stability, and the interpretation of "noise" are not adequately documented (Sec. 2.2, Sec. 3.2).** The manuscript notes $min_cluster_size$ was "carefully tuned" but does not report the chosen values per model or explore sensitivity to $min_samples$/metric/standardization. Additionally, statements conflating "one cluster" with "all noise" make unimodality claims ambiguous; in HDBSCAN these are distinct outcomes, and "all noise" does not imply unimodality—only that density-based clusters were not identified under the chosen settings.

   Recommendation: In Sec. 2.2.2–2.2.3, report the full HDBSCAN configuration per model ($min_cluster_size$, $min_samples$, metric, $cluster_selection_method$, preprocessing). In Sec. 3.2 (or Appendix), provide a robustness study: vary $min_cluster_size$/$min_samples$ and (if applicable) distance metric; report number of clusters, noise fraction, cluster stability/persistence scores, and how key summaries (e.g., medians/90% CIs of $\cos\theta_{jn}$, χ_{eff}, χ_p, masses) change. Rewrite unimodality language to distinguish

"one robust cluster" from "no clusters found (all noise)." If inclination bimodality is a central physical claim, consider adding a direct bimodality diagnostic on $\cos\theta_{jn}$ (e.g., dip test/mixture fit) alongside HDBSCAN.

5. **The definition of the 14D feature space used for distances/clustering/UMAP is potentially inconsistent with the stated goal of comparing *physical* source-parameter structure (Sec. 2.1.1, Sec. 2.2, Sec. 2.3).** In particular: (i) including 'log_likelihood' as a coordinate can dominate geometry and create separations that reflect fit quality or sampler artifacts rather than physical degeneracy; (ii) periodic variables (e.g., phi_{jl}) appear to be z-score standardized linearly without circular handling, which can distort distances near wrap-around.

   Recommendation: Decide and document a principled feature set for geometric comparisons. Prefer excluding 'log_likelihood' from the clustering/UMAP feature space (or analyze it separately as an outcome variable); if it is retained, explicitly motivate it and quantify its influence (e.g., repeat key results with/without it and show islands/clusters persist). For angular parameters (phi_{jl} and any others), use a circular embedding (sin/cos) or otherwise justify that wrap-around effects do not matter for the sampled support. Report the final feature list used for each method (JS/KDE, HDBSCAN, UMAP) in Sec. 2.

6. **Claims that χ_p is "robust" while $q/\chi_{\rm eff}/\cos\theta_{jn}$ are "model-dependent" are plausible but not supported by a consistent quantitative framework, nor clearly separated into data-driven findings vs hypothesized waveform explanations (Sec. 3.4, Sec. 4.1–4.3).** Additionally, the physical interpretation would be strengthened by showing whether different islands/modes have comparable fit quality (since 'log_likelihood' is available) and by demonstrating that high χ_p is not primarily prior-driven for this short/high-mass signal.

   Recommendation: In Sec. 3.4 and Sec. 4.1–4.2, add concise quantitative summaries across models for key parameters: range of posterior medians, 90% CI overlaps, and 1D JS distributions (e.g., median and interquartile range of pairwise JS). Define a criterion for "robust" vs "model-dependent" and apply it consistently. Add prior-vs-posterior comparisons (or KL to prior) for χ_p, $\chi_{\rm eff}$, and q for each model to substantiate robustness. Use 'log_likelihood' (or evidence if available) to check whether different UMAP islands correspond to similarly good fits; if one island systematically fits worse, state this explicitly and adjust interpretation. Expand Sec. 4.2–4.3 to connect observed discrepancies to specific known waveform-model differences (calibration ranges, precession/higher-mode treatments, time vs frequency domain approximations), with citations, clearly labeling hypotheses vs measured effects.

7. **The manuscript sometimes reads as if broad conclusions about waveform systematics and manifold-based comparison methods generalize beyond this case, but the study is a single-event analysis (Sec. 1, Sec. 4.3).** Given

GW231123's likely high-mass/short-duration nature (merger–ringdown dominated), it may be an especially challenging or atypical case; the generality and practical "workflow" implications for catalogs/populations are not yet established.

Recommendation: Reframe Sec. 1 and Sec. 4.3 more explicitly as a *case study* of GW231123. Add a short limitations-and-generalization paragraph (Sec. 4) describing what properties of GW231123 make it a strong stress-test (e.g., short signal, strong degeneracies) and what is needed to extend the approach: computational scaling, event selection criteria, and minimal robustness diagnostics to run routinely. Where you draw broader lessons, make clear which are methodological proposals vs empirically demonstrated general patterns.

Minor issues

1. Figures 2–4 are central to the argument but currently lack enough methodological annotation to be fully interpretable and comparable across models (Fig. 2: separate PCA per model and/or inconsistent limits; Fig. 3–4: missing UMAP/KDE settings; risk of over-reading axis meaning; unclear sample weighting across models).

   Recommendation: Improve figure actionability: (i) for Fig. 2, use a shared embedding (PCA trained on combined standardized data) or explicitly align axes/limits; report explained variance on axes; annotate cluster fractions and key parameter medians per cluster; (ii) for Fig. 3–4, state UMAP settings and whether samples are balanced/weighted per model; add per-model contours/legends on the manifold; standardize colormaps and colorbar ranges across parameter-color panels; and add a caption caveat that UMAP axes are arbitrary.

2. Key numerical results (JS matrices, UMAP-space JS matrix, clustering outcomes) are scattered in text rather than summarized in compact tables, reducing readability and reuse (Sec. 3.1–3.4).

   Recommendation: Add tables in Sec. 3 or an Appendix: (i) pairwise 1D JS matrices for χ_p, $\chi_{\rm eff}$, q (or component masses), $\cos\theta_{jn}$; (ii) the UMAP-space 2D JS matrix; (iii) per-model HDBSCAN summary (number of clusters, noise fraction, cluster fractions, cluster medians/CIs for a short list of parameters).

3. Terminology for waveform families is sometimes inconsistent (time-domain vs NR surrogate vs phenomenological frequency-domain), which can confuse readers (Sec. 2.1.1, Sec. 3.2–3.4).

   Recommendation: Add a short classification at the start of Sec. 2.1.1 listing each model and its family/characteristics (domain, precession and higher-mode treatment), and use that classification consistently when interpreting islands/clusters.

4. The connection to prior literature on cross-waveform comparisons, divergence measures in GW inference, and high-dimensional posterior visualization is relatively brief (Sec. 1).

   Recommendation: Add a short "Related work" paragraph/subsection in Sec. 1 (or early Sec. 2) situating JS/KL-style comparisons, waveform-systematics studies for high-mass/precessing BBHs, and prior uses of clustering/t-SNE/UMAP (or related) for degeneracy exploration; clearly state what is novel here (joint embedding across models + combined pipeline).

5. Internal cross-references and placeholders appear inconsistent (e.g., references in Sec. 3.4 pointing to the wrong sections; mentions like "Table 1 (from the provided short results)") (Sec. 3–4).

   Recommendation: Perform a full cross-reference pass: correct all Sec./Figure/Table pointers; remove drafting placeholders; when citing a figure, specify panel(s) and briefly state what is plotted at first mention.

6. Reproducibility expectations would be better met with an explicit data/code availability statement; currently it is unclear whether posterior samples, derived JS matrices, and embeddings will be released (Sec. 4 / end matter).

   Recommendation: Add a Data/Code Availability section specifying: where posterior samples can be obtained; whether you will release the analysis scripts/notebooks; and whether derived artifacts (JS matrices, clustering labels, UMAP coordinates) will be archived with versioning. Include software package names/versions (hdbscan, umap-learn, KDE implementation) in Sec. 2 or end matter.

Very minor issues

1. Notation and naming inconsistencies (e.g., cos_theta_jn vs cos θ_{jn}; waveform model spelling/hyphenation; occasional line-break artifacts) reduce polish and can confuse readers (Sec. 1, Sec. 3–4; figure captions).

   Recommendation: Standardize parameter notation and waveform model names throughout text/figures; fix hyphenation/spelling; remove line-break/OCR artifacts.

2. Some prose is repetitive (pipeline re-described multiple times) and several sentences are long, especially in Sec. 1 and Sec. 4.

   Recommendation: Tighten Sec. 1 and Sec. 4 by removing repeated pipeline descriptions and splitting long sentences; keep one canonical pipeline description in Sec. 2 and refer back to it.

3. Figure captions sometimes mix extensive interpretation with description, and axis units/arbitrariness (especially UMAP) are not always stated (Fig. 3).

Recommendation: Edit captions to focus on what is shown; move interpretation to main text; label UMAP axes as arbitrary/unitless and ensure legends/encodings are redundant (not color-only).

4. Ambiguous quantifiers ("typically", "other models") and non-specific model references weaken checkability of some statements (Sec. 3.1, Sec. 3.4).

   Recommendation: Replace ambiguous phrasing with explicit counts/fractions (e.g., "8/10 pairs have JS < 0.1") and name the specific model(s) associated with quoted medians or JS values.

Key statements and references

- • **The time-domain waveform models NRSur7dq4, SEOBNRv5PHM, and IMRPhenomTPHM exhibit a bimodal posterior for the viewing angle ($\cos\theta_{jn}$), with HDBSCAN identifying two statistically significant modes corresponding to face-on and face-off solutions (e.g., for NRSur7dq4, Mode 0 with 25.7% of samples has median $\cos\theta_{in} = -0.37$ and Mode 1 with 11.7% of samples has median $\cos\theta_{in} = +0.40$), demonstrating that these models do not uniquely resolve the inclination-related degeneracy in the likelihood.**
- *Reference(s):* (none)
- • **In contrast to the time-domain models, the frequency-domain phenomenological models IMRPhenomXO4a and IMRPhenomXPHM yield unimodal posteriors in HDBSCAN clustering, with 100% of samples assigned to a single high-probability region and median $\cos\theta_{in} \approx 0.88$ (face-on) for IMRPhenomXO4a and near 0.0 (edge-on) for IMRPhenomXPHM, indicating that these models resolve the inclination degeneracy in mutually inconsistent ways.**
- *Reference(s):* (none)
- • **Baseline 1D marginalized posterior comparisons using Kernel Density Estimation and Jensen–Shannon divergence show that IMRPhenomXO4a infers a highly unequal mass ratio for GW231123, with a median secondary mass mass_2_source = 55.1\,M_\odot, whereas the other four models favor more comparable masses with mass_2_source medians between 93.3 $M_\odot$ and 111.1 $M_\odot$, and the JS divergence in mass_2_source between IMRPhenomXO4a and each other model exceeds 0.62, indicating minimal overlap of their posteriors.**
- *Reference(s):* (none)
- • **A unified 2D UMAP embedding of the combined 14-dimensional posterior samples from all five waveform models reveals three well-separated islands: a central island jointly populated by NRSur7dq4, SEOBNRv5PHM,**

and IMRPhenomTPHM, a right-hand island exclusively populated by IMRPhenomXPHM, and a top island exclusively populated by IMRPhenomXO4a; mapping physical parameters onto this manifold shows that UMAP Dimension 1 primarily correlates with effective aligned spin $\chi_{\rm eff}$ and Dimension 2 with viewing angle $\cos\theta_{jn}$, while pairwise JS divergences between models in different islands (e.g., $JS(\mathrm{IMRPhenomXPHM}, \mathrm{IMRPhenomTPHM}) = 0.69$) confirm that they correspond to fundamentally different solutions in the full parameter space.

- *Reference(s):* (none)

- • **Across all five waveform models, the precessing spin parameter χ_p is consistently inferred to be high ($\chi_p > 0.7$), with low JS divergences between models (e.g., 0.02 between NRSur7dq4 and IMRPhenomXPHM) and all models occupying regions of high χ_p in the shared UMAP manifold, establishing a robust, model-independent conclusion that GW231123 exhibits strong spin–orbit precession even though $\chi_{\rm eff}$, mass ratio, and viewing angle remain strongly model-dependent with JS divergences for $\chi_{\rm eff}$ often exceeding 0.6 (e.g., 0.57 between IMRPhenomXPHM and SEOBNRv5PHM).**

- *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 and descriptive; it uses standard statistical constructs (credible intervals, KDE, Jensen–Shannon divergence) and geometric ML procedures (standardization, HDBSCAN, UMAP) but provides essentially no explicit equations or derivations. The main auditability gap is that the paper does not specify the mathematical definition/approximation details for computing JS divergence from KDE-estimated continuous densities, especially in **2D** and on the UMAP manifold.

Checked items

1. ✔ **14D feature vector definition** (Sec. 2.1.1, p.2–3)
    - **Claim:** All subsequent analyses operate on 14-dimensional posterior samples with listed parameters.
    - **Checks:** definition consistency, symbol/notation consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** The parameter list is complete and intended to define the analysis feature space.

- **Notes:** The list contains **14** entries and later sections consistently refer to a **14D** space. However, one entry is 'log_likelihood', which is not a physical parameter; that concern is captured as a separate issue.

2. ✔ **Credible interval definition** (Sec. 2.1.2, p.3)
    - **Claim:** A **90%** credible interval is defined by the **5**th and **95**th percentiles.
    - **Checks:** definition consistency, basic probability sanity
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Posterior samples represent draws from the target posterior.
    - **Notes:** Definition matches a central **90%** quantile interval and is used consistently in the narrative.

3. ⚠ **JS divergence on 1D KDE PDFs is well-defined** (Sec. 2.1.2 (1D Marginal Posterior Analysis), p.3)
    - **Claim:** For each parameter and model pair, estimate PDFs via KDE and compute JS divergence between the two PDFs.
    - **Checks:** definition completeness, normalization/support consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: critical
    - **Assumptions/inputs:** KDE outputs are treated as continuous densities., JS divergence is computed in a way appropriate for continuous distributions.
    - **Notes:** No formula is given for JS divergence (continuous integral vs discrete approximation), no log base is specified, and no details are given on how the KDEs are evaluated on a common domain/grid and renormalized (if discretized). These details are necessary to verify the metric is computed consistently across model pairs.

4. ✔ **Symmetry and matrix structure of pairwise divergences** (Sec. 2.1.2, p.3; Sec. 2.3.2, p.4)
    - **Claim:** Divergences yield $\mathbf{5 \times 5}$ symmetric matrices (one per parameter for **1D**; one per selected pair for **2D**; one global matrix for UMAP-KDE).
    - **Checks:** algebra/structure sanity, consistency across sections
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** JS divergence is symmetric by definition/implementation.
    - **Notes:** Given JS is symmetric, reporting symmetric $\mathbf{5 \times 5}$ matrices is structurally consistent. Diagonal entries are not discussed but would typically be **0**.

5. ⚠ **JS divergence on 2D KDE density fields** (Sec. 2.1.2 (2D Marginal Posterior Analysis), p.3)

- **Claim:** Compute JS divergence between two **2D** KDE-estimated joint density fields.
- **Checks:** definition completeness, normalization/support consistency
- **Verdict:** UNCERTAIN; confidence: medium; impact: critical
- **Assumptions/inputs:** A valid **2D** JS divergence is computed between normalized densities on a common support.
- **Notes:** As with **1D**, details are missing: how the continuous densities are compared (integral vs grid sum), grid resolution, domain truncation, and renormalization. In **2D** these choices materially affect the computed divergence and must be specified for analytic auditability.

6. ✔ **Per-model standardization for clustering** (Sec. 2.2.1, p.3)

    - **Claim:** For each model, standardize each parameter by subtracting its mean and dividing by its standard deviation prior to clustering.
    - **Checks:** algebraic definition, well-posedness
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Each parameter has finite nonzero sample standard deviation.
    - **Notes:** The z-score transform is well-defined as stated, assuming no zero-variance columns. The paper does not discuss handling zero variance, but that is an edge case not indicated here.

7. ✔ **HDBSCAN applied in 14D standardized space** (Sec. 2.2.2, p.3)

    - **Claim:** HDBSCAN is applied to the standardized **14D** samples to identify modes.
    - **Checks:** method/space consistency
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Distance computations use the standardized coordinates as intended.
    - **Notes:** Consistent with Sec. 2.2.1. A separate issue remains about whether 'log_likelihood' and circular variables should be in that space, but the stated pipeline is internally consistent.

8. ✖ **Unimodal models: 'single cluster' vs 'all noise' ambiguity** (Sec. 3.2, point 2, p.6)

    - **Claim:** Frequency-domain models are unimodal: HDBSCAN classified **100%** of samples as belonging to a single region (or 'noise', meaning no separable clusters).
    - **Checks:** logical consistency of categorical outcomes
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** HDBSCAN output labels are interpreted correctly.

- **Notes:** The statement conflates two distinct HDBSCAN outcomes: (i) one cluster containing all points, versus (ii) zero clusters with all points labeled noise. The parenthetical does not resolve the ambiguity and is internally inconsistent with 'classified **100%** ... as belonging to a single ... region'.

9. ✔ **Global standardization before UMAP** (Sec. 2.3.1, p.4)
    - **Claim:** Concatenate all samples, compute global mean/std per parameter over the combined dataset, standardize, then run UMAP to **2D**.
    - **Checks:** definition consistency, well-posedness
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** All models share the same parameterization and units for each column.
    - **Notes:** This is internally consistent and well-defined given common columns. The paper does not discuss whether all models produce identically defined parameters, but within the document it is assumed.
10. ⚠ **KDEs in UMAP space defined on a common domain** (Sec. 2.3.2, p.4)
    - **Claim:** Perform **2D** KDE per model in UMAP space, producing density fields 'all defined on the identical set of UMAP coordinates', then compute pairwise JS divergence.
    - **Checks:** definition completeness, normalization/support consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: critical
    - **Assumptions/inputs:** A common evaluation grid/coordinate set exists for all model KDEs.
    - **Notes:** KDEs are continuous functions; to compare them pointwise you must specify the shared evaluation grid (or integration rule). The phrase 'identical set of UMAP coordinates' is ambiguous (data points vs grid), and no normalization-after-discretization procedure is described.
11. ✖ **Inclusion of log_likelihood as 'physical parameter'** (Sec. 2.1.1 (parameter list), p.2–3; referenced in Secs. 2.2–2.3, p.3–4)
    - **Claim:** The analysis is performed on **14** physical parameters including $log_likelihood$.
    - **Checks:** symbol/definition consistency, conceptual dimensional consistency
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** All **14** coordinates represent physical degrees of freedom of the source.
    - **Notes:** Within the document's own wording, $log_likelihood$ is not a physical source parameter but a sampling/fit-quality statistic; calling it physical is inconsistent. If included in clustering/UMAP, it changes the geometry of the space being analyzed.

12. ✔ **Parameter naming consistency for χ_{eff}, χ_p, $\cos\theta_{jn}$** (Throughout; e.g., Sec. 3.3–3.4, p.7–10 and figure captions)
    - **Claim:** Snake_case parameter names correspond to the typeset symbols used in interpretation.
    - **Checks:** notation consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** $chi_eff \equiv \chi_{\text{eff}}$, $chi_p \equiv \chi_p$, $cos_theta_jn \equiv \cos\theta_{jn}$.
    - **Notes:** The mapping is inferable from context and used consistently in narrative, though it is not explicitly defined in a notation table.

Limitations

- The provided PDF content contains essentially no explicit equations, equation numbers, or derivation steps; most mathematical operations are described in prose, limiting formal algebraic verification.
- Key definitions needed for an analytic audit of the central quantitative metric (JS divergence computed from KDEs) are omitted, making several checks necessarily UNCERTAIN.
- Figures are referenced for empirical relationships (e.g., UMAP axis correlations), but these are observational/visual claims that cannot be audited symbolically without explicit analytic definitions.

Numerical results audit

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

21 candidate numerical claims (Pages **5**–**9**, Secs. **3.1**–**3.4.2**) were set up for verification, covering threshold/range checks, recomputed medians, and **1D/2D** Jensen–Shannon (JS) divergences, plus HDBSCAN clustering fractions/medians and UMAP-space JS examples. All checks returned UNCERTAIN because the required numerical inputs (posterior samples and/or specific analysis settings/artifacts) were not available within the provided inputs, so no quantitative PASS/FAIL validation could be completed.

Checked items

1. ⚠ **C1** (Page **5**, Sec. **3.1** (Baseline comparison))
    - **Claim:** “median values for ‘chi_p‘ consistently above **0.7** across all models”
    - **Checks:** threshold_check_against_reported_rule
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to compute per-model medians of chi_p.
2. ⚠ **C2** (Page **5**, Sec. **3.1**)

- **Claim:** “chi_p ... showed consistently low JS divergence values (typically $<$ 0.1) across most model pairs”
- **Checks:** matrix_fraction_below_threshold
- **Verdict:** UNCERTAIN
- **Notes:** Insufficient inputs to recompute pairwise JS divergences; also relies on an interpretation of “most”.

3. ⚠ **C3** (Page 5, Sec. 3.1)
   - **Claim:** “IMRPhenomXO4a ... median secondary source mass (‘mass_2_source‘) of 55.1 $M_\odot$”
   - **Checks:** recompute_median_from_samples
   - **Verdict:** UNCERTAIN
   - **Notes:** Insufficient inputs to compute median(mass_2_source) for IMRPhenomXO4a.

4. ⚠ **C4** (Page 5, Sec. 3.1)
   - **Claim:** “other four models ... ‘mass_2_source‘ medians ranging from 93.3$M_\odot$ to 111.1$M_\odot$”
   - **Checks:** range_check_across_models
   - **Verdict:** UNCERTAIN
   - **Notes:** Insufficient inputs to compute medians across the four specified models and compare min/max to the stated endpoints.

5. ⚠ **C5** (Page 5, Sec. 3.1)
   - **Claim:** “JS divergence for ‘mass_2_source‘ between ‘IMRPhenomXO4a‘ and all other models exceeded 0.62”
   - **Checks:** pairwise_threshold_all_pairs
   - **Verdict:** UNCERTAIN
   - **Notes:** Insufficient inputs to recompute JS_{1D}(mass_2_source) for XO4a vs each other model.

6. ⚠ **C6** (Page 5, Sec. 3.1)
   - **Claim:** “IMRPhenomXPHM ... median [chi_eff] of 0.04”
   - **Checks:** recompute_median_from_samples
   - **Verdict:** UNCERTAIN
   - **Notes:** Insufficient inputs to compute $median(chi_eff)$ for IMRPhenomXPHM.

7. ⚠ **C7** (Page 5, Sec. 3.1)
   - **Claim:** “other models ... favored significantly positive alignment (median ‘chi_eff‘ of 0.44)”
   - **Checks:** spot_check_specific_model_value

- **Verdict:** UNCERTAIN
- **Notes:** Insufficient inputs to compute non-IMRPhenomXPHM model medians; statement is also ambiguous about which model the **0.44** refers to.

8. ⚠ **C8** (Page **5**, Sec. **3.1**)
   - **Claim:** "JS divergence for 'chi_eff' ... (e.g., **0.63** against 'IMRPhenomTPHM')"
   - **Checks:** recompute_pairwise_js_value
   - **Verdict:** UNCERTAIN
   - **Notes:** Insufficient inputs to recompute $JS_{1D}(chi_eff)$ between IMRPhenomXPHM and IMRPhenomTPHM.

9. ⚠ **C9** (Page **5**, Sec. **3.1**)
   - **Claim:** "2D JS divergence of **0.69** between 'IMRPhenomXO4a' and 'NRSur7dq4' for ('mass_1_source', 'mass_2_source')"
   - **Checks:** recompute_2d_js_value
   - **Verdict:** UNCERTAIN
   - **Notes:** Insufficient inputs to recompute $2D$ KDE-based JS divergence on (mass_1_source, mass_2_source).

10. ⚠ **C10** (Page **6**, Sec. **3.2** (HDBSCAN results, NRSur7dq4 example))
    - **Claim:** "NRSur7dq4, Mode **0** (**25.7%** of the samples) ... Mode **1** (**11.7%** of samples)"
    - **Checks:** cluster_fraction_sum_and_labeling
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to reproduce HDBSCAN clustering and label fractions.

11. ⚠ **C11** (Page **6**, Sec. **3.2** (NRSur7dq4 mode medians))
    - **Claim:** "NRSur7dq4 ... Mode **0** ... median 'cos_theta_jn' of -0.37 ... Mode **1** ... median 'cos_theta_jn' of $+0.40$"
    - **Checks:** clusterwise_median_check
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to reproduce clustering labels needed to compute clusterwise medians.

12. ⚠ **C12** (Page **6**, Sec. **3.2** (SEOBNRv5PHM mode medians))
    - **Claim:** "SEOBNRv5PHM (Mode **0**: median 'cos_theta_jn' of $+0.47$, Mode **1**: median of -0.44)"
    - **Checks:** clusterwise_median_check
    - **Verdict:** UNCERTAIN

- **Notes:** Insufficient inputs to reproduce clustering labels needed to compute clusterwise medians.

13. ⚠ **C13** (Page 6, Sec. 3.2 (frequency-domain models unimodal))
    - **Claim:** "IMRPhenomXO4a and IMRPhenomXPHM ... HDBSCAN classified 100% of their samples as belonging to a single ... region (or 'noise' ... meaning no separable clusters were found beyond the main body)"
    - **Checks:** cluster_count_and_fraction_check
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to reproduce HDBSCAN results (cluster counts and largest-assignment fractions).

14. ⚠ **C14** (Page 6, Sec. 3.2)
    - **Claim:** "IMRPhenomXO4a ... median 'cos_theta_jn' of 0.88"
    - **Checks:** recompute_median_from_samples
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to compute $median(cos_theta_jn)$ for IMRPhenomXO4a.

15. ⚠ **C15** (Page 6, Sec. 3.2)
    - **Claim:** "IMRPhenomXPHM ... median 'cos_theta_jn' near 0.0"
    - **Checks:** near_zero_check
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to compute $median(cos_theta_jn)$ for IMRPhenomXPHM.

16. ⚠ **C16** (Page 8, Sec. 3.3 (UMAP-manifold JS examples))
    - **Claim:** "JS('IMRPhenomXPHM', 'IMRPhenomTPHM') = 0.69" (JS on $2D$ KDEs in UMAP space)
    - **Checks:** recompute_js_on_umap_kdes
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to reproduce the UMAP+KDE pipeline and compute the UMAP-space JS divergence.

17. ⚠ **C17** (Page 8, Sec. 3.3 (UMAP-manifold JS examples))
    - **Claim:** "JS('SEOBNRv5PHM', 'IMRPhenomTPHM') = 0.42" (JS on $2D$ KDEs in UMAP space)
    - **Checks:** recompute_js_on_umap_kdes
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to reproduce the UMAP+KDE pipeline and compute the UMAP-space JS divergence.

18. ⚠ **C18** (Page 8, Sec. **3.4.1** (Robust conclusions))
    - **Claim:** "JS divergence for 'chi_p' was consistently low (e.g., **0.02** between 'NRSur7dq4' and 'IMRPhenomXPHM')"
    - **Checks:** recompute_pairwise_js_value
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to recompute $JS_{1D}(chi_p)$ between NRSur7dq4 and IMRPhenomXPHM.
19. ⚠ **C19** (Page 8, Sec. **3.4.2** item **1**)
    - **Claim:** "time-domain models ... medians around **0.23**–**0.44** [for chi_eff]"
    - **Checks:** range_check_across_models
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to compute chi_eff medians for the three time-domain models and compare to the stated range.
20. ⚠ **C20** (Page 8, Sec. **3.4.2** item **1**)
    - **Claim:** "JS ... (e.g., **0.57** between 'IMRPhenomXPHM' and 'SEOBNRv5PHM')"
    - **Checks:** recompute_pairwise_js_value
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to recompute $JS_{1D}(chi_eff)$ between IMRPhenomXPHM and SEOBNRv5PHM.
21. ⚠ **C21** (Page **9**, Sec. **3.4.2** item **3**)
    - **Claim:** "high JS divergence for 'cos_theta_jn' (e.g., **0.62** between 'IMRPhenomXO4a' and 'NRSur7dq4')"
    - **Checks:** recompute_pairwise_js_value
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to recompute $JS_{1D}(cos_theta_jn)$ between IMRPhenomXO4a and NRSur7dq4.

Limitations

- Only parsed text and page images provided; referenced tables (Table 1, Table 2) are not available in extractable numeric form here, limiting internal cross-checks against tabulated summaries.
- Several checkable quantities (JS divergences, HDBSCAN mode fractions/medians, UMAP-manifold JS) depend on unspecified implementation details (evaluation grid, KDE normalization, UMAP hyperparameters/seed, HDBSCAN $min_cluster_size$), so reproduction may be approximate unless the exact analysis code/settings are available within the PDF (not present in the provided excerpt).

- No checks were proposed that require reading numeric values from plots, per instruction; this excludes verifying many figure-based claims directly.
- All executed checks C1–C21 returned UNCERTAIN because the required numerical inputs (posterior samples and/or derived artifacts such as JS matrices and exact UMAP/HDBSCAN settings) were not available and file/network access was disallowed.