Skepthical review: *3. RESULTS*

Summary

The manuscript presents a systematic framework to diagnose waveform-model systematics for GW231123 by comparing parameter-estimation posteriors from five waveform models (NRSur7dq4, IMRPhenomXO4a, SEOBNRv5PHM, IMRPhenomXPHM, IMRPhenomTPHM). After mapping all results to a common 13-parameter set (Sec. 2.1, Sec. 2.3.1), the authors quantify univariate differences using Jensen–Shannon divergence (JSD) and 1-Wasserstein distance (Sec. 2.2), visualize high-dimensional structure with UMAP (Sec. 2.3), and introduce a "Physics-Informed Discrepancy Decomposition" that partitions parameters into physically motivated subspaces (mass+distance, effective spin, individual spins+orientation, and remnant properties) and computes multivariate JSDs per subspace (Sec. 2.4, results in Sec. 3.3). The results indicate substantial model dependence in key inferred parameters (Sec. 3.1), a core-vs-outlier clustering pattern in UMAP space (Sec. 3.2), and the largest multivariate disagreements in the individual spin & orientation subspace (Sec. 3.3). Using threshold-based robustness criteria (Sec. 2.5, Sec. 3.4; Table 3), the paper concludes that no key astrophysical parameter is "robust" for GW231123 under the adopted thresholds, implying waveform systematics dominate statistical uncertainties for this event. The problem is timely and the decomposition is intuitively useful, but several aspects require strengthening for reliability and reproducibility: (i) posterior provenance and prior/configuration consistency across models, (ii) the statistical stability of multivariate KDE+JSD in up to 6D (including uncertainty estimates), (iii) calibration/justification of robustness thresholds and the interpretation of "no robust parameters," (iv) UMAP stability and its proper evidentiary role, and (v) validation/tempering of causal attributions to specific waveform-physics ingredients.

Strengths

- Timely and high-impact target: GW231123 is an excellent case study where waveform systematics plausibly dominate for short, high-mass, potentially precessing BBH signals (Sec. 1, Sec. 4.2–4.3).
- The Physics-Informed Discrepancy Decomposition is a conceptually appealing way to localize model disagreements in physically interpretable parameter groupings (Sec. 2.4, Sec. 3.3).
- The workflow combines complementary tools—univariate divergences (JSD/Wasserstein), high-dimensional visualization (UMAP), and multivariate subspace divergences—yielding a coherent diagnostic suite rather than a single metric (Sec. 2.2–2.4, Sec. 3.1–3.3).
- The paper clearly demonstrates that different waveform models can yield materially different inferences for GW231123, which is an important message for downstream astrophysical interpretation (Sec. 3.1, Sec. 3.4; Table 3).

- Overall organization is logical (data → metrics → decomposition → results → implications), and several figures/tables effectively convey cross-model differences when readable (e.g., Fig. 3 conceptually; Table 3).

Major issues

1. **Reproducibility and provenance: the paper does not provide publication-grade details on how each posterior sample set was produced and whether all five analyses are comparable beyond the waveform approximant choice (Sec. 2.1–2.2).** The use of internal filesystem paths obscures provenance. Critically, cross-model posterior differences can be caused (or amplified) by differences in priors, calibration version, PSD estimation, data segments, detector network, reference frequency, parameterization choices, or sampler settings—not only waveform physics. The current description does not convincingly isolate "waveform-model systematics" from "analysis-configuration differences."

   Recommendation: Rewrite/expand Sec. 2.1 into a reproducible data+PE configuration specification. For each model, explicitly state: (i) source of posteriors (LVK public release DOI/URL vs your own PE), (ii) detectors used, GPS time window, PSD method, calibration version/marginalization, (iii) priors (including mass, distance/redshift, spin magnitude/tilt, orientation), (iv) sampler and settings, and (v) final number of (effective) posterior samples used after burn-in/thinning/reweighting. Add a compact table summarizing "what is held fixed vs what changes" across models. If priors differ in any way, reweight posteriors to a common prior before computing divergences (or explicitly quantify the effect by comparing divergences pre/post reweighting). Also explicitly confirm cosmology/source-frame conversions are identical across models when comparing source-frame masses and redshift.

2. **Multivariate JSD via multivariate KDE is under-specified and potentially unstable in moderate dimension (up to 6D in the Individual Spin & Orientation subspace), yet it underpins key conclusions about which physics drives disagreements (Sec. 2.4.2, Sec. 3.3).** In 6D, KDE bandwidth choice and sample size can dominate density estimates and thus JSD; JSD values "near maximum" may reflect estimator artifacts rather than genuine separation. The manuscript also lacks uncertainty quantification for JSD estimates.

   Recommendation: In Sec. 2.4.2, provide the exact mathematical definition of the multivariate JSD used (including log base and normalization) and fully specify the estimator: KDE kernel, bandwidth selection rule (and whether shared across models), boundary handling (especially for bounded spins and angular variables), and numerical integration/Monte Carlo procedure used to compute JSD from KDEs. Report sample sizes and effective sample sizes per model per subspace. Add uncertainty and stability diagnostics: bootstrap over posterior samples to provide confidence intervals on each pairwise/subspace JSD, and a sensitivity study varying bandwidth (and/or

using subsampling) to demonstrate conclusions persist. Consider adding (even as a check in an Appendix) at least one alternative two-sample/divergence approach that is more robust in higher dimensions (e.g., kNN-based divergence estimators, MMD/energy distance, sliced Wasserstein, or classifier-based two-sample tests), and show that the ranking of "most discrepant subspace" is consistent.

3. **Robustness thresholds (max pairwise JSD and median-spread cutoffs) are central to the paper's headline conclusion ("no key parameter is robust"), but the thresholds are currently ad hoc/uncalibrated and inconsistently stated (0.01 vs 0.05) (Sec. 2.5.1, Sec. 3.4, Sec. 4.2–4.3; Table 3).** Without calibration, it is unclear whether the robustness classification is scientifically meaningful or overly stringent for this event class.

   Recommendation: In Sec. 2.5.1, state a single, unambiguous robustness decision rule and use it consistently everywhere (including Abstract, Sec. 3.4, Table 3, Sec. 4.2–4.3). Then calibrate/justify thresholds: (i) provide a sensitivity analysis over plausible cutoffs (e.g., JSD $\in \{0.02, 0.05, 0.1\}$ and median spread $\in \{5\%, 10\%, 20\%\}$) and report how many/which parameters change category; (ii) complement JSD with an effect-size metric tied to inference impact, e.g., median shifts in units of a pooled posterior standard deviation or overlap of 90% credible intervals. If possible, add an injection-based calibration (even one or two representative injections in the GW231123 regime) linking typical waveform-systematics-induced JSD values to known biases relative to truth. Temper the strength of the global conclusion if it hinges on one particular threshold choice.

4. **UMAP is used as a central diagnostic to claim clustering (core group vs isolated phenomenological models) and to motivate interpretation, but UMAP's stochasticity and hyperparameter dependence are not sufficiently documented or tested (Sec. 2.3, Sec. 3.2).** As presented, the embedding risks being interpreted as stronger evidence than warranted; distances and separations in 2 D may not correspond to meaningful separation in the original 13D space.

   Recommendation: In Sec. 2.3, fully document UMAP settings: parameter list, standardization (joint across all models vs per model), distance metric, handling of periodic/angular parameters (ideally via $\sin/\cos$ transforms before z-scoring), n_neighbors, min_dist, initialization, epochs, and random seed(s). In Sec. 3.2, add a stability analysis: repeat embeddings across multiple seeds and a small grid of (n_neighbors, min_dist), and show the qualitative grouping persists. Add at least one quantitative check: silhouette score by model label in the embedding and/or a two-sample distance/test computed directly in the original standardized space (to avoid over-reliance on 2D visuals). Rephrase interpretation to clearly distinguish "visual diagnostic" from "quantitative discrepancy metric," pointing readers to the divergence results (Sec. 3.1, Sec. 3.3) for the main quantitative claims.

5. **The manuscript frequently moves from "subspace posteriors differ" to causal attributions to specific waveform-physics ingredients (precession implementation, higher-mode content, merger–ringdown modeling), but the presented analyses primarily establish correlation/association, not causation (Sec. 2.4.3, Sec. 3.3.2–3.3.4, Sec. 4.2–4.3).** Without controlled comparisons, readers may overinterpret these as definitive mechanistic explanations.

   Recommendation: In Sec. 2.4.3 and Sec. 3.3, explicitly label physical explanations as hypotheses ("consistent with") unless supported by additional tests. Add a concise table listing each model's salient physics ingredients (precession scheme, mode content, calibration region, ringdown treatment) and relate these to observed discrepancy patterns via structured comparisons (e.g., within-group vs between-group JSD among models sharing features). Strongly consider adding at least one validation exercise: a controlled injection study (or a small set) where the injected signal is known and where changing a single modeling ingredient is expected to predominantly affect a specific subspace; demonstrate the decomposition recovers this behavior. If injections are out of scope, narrow/temper causal language in Sec. 4.2–4.3.

6. **The "consensus posterior" construction via concatenating samples across models is not a generally valid inference operation unless explicitly framed as Bayesian model averaging with defined model weights; otherwise it can misrepresent uncertainty and can be misleading even if parameters appear 'robust' (Sec. 2.5.3).**

   Recommendation: Revise Sec. 2.5.3 to (i) explicitly define the mathematical object you intend (e.g., an equally weighted mixture over model-conditioned posteriors), (ii) state model weights and justify them (equal weights vs evidence-based weights vs prior model probabilities), and (iii) clarify limitations. If the intent is diagnostic, label it clearly as a visualization/summary tool, not a principled posterior. Optionally, add a short comparison to formal Bayesian model averaging: $p(\theta \mid d) = \sum_m p(\theta \mid d, m)\, p(m \mid d)$, and explain what is assumed/omitted (e.g., evidences not computed).

7. **Physical plausibility/consistency checks are not sufficiently developed given the extreme cross-model spreads reported for some derived quantities (e.g., large shifts in source-frame masses and redshift) (Sec. 3.1, Sec. 3.4).** These may be real for a short high-mass signal, but they also commonly indicate differences in priors, cosmology/source-frame conversion, or parameter-definition mismatches across posterior files.

   Recommendation: Add explicit consistency checks and documentation: confirm that (i) all posteriors use the same cosmology when reporting redshift/source-frame masses, (ii) detector-frame vs source-frame quantities are not mixed, and (iii) derived parameters (e.g., remnant properties) are computed using the same recipe across models. Provide a short "sanity check" subsection (end of Sec. 2.1 or start of Sec. 3.1) docu-

menting that prior supports match and that the observed posterior differences are not prior-driven (e.g., show prior overlays for a few key parameters, or report prior-to-posterior information gain per model).

Minor issues

1. Inconsistent definition/range of JSD across the manuscript (e.g., $[\mathbf{0}, \mathbf{1}]$ vs maximum $\approx \mathbf{0.693}$) can confuse interpretation and thresholding (Sec. 2.2.2, Sec. 2.4.2, Sec. 3.3.3).

   Recommendation: Define JSD once (equation + log base) and ensure all stated ranges and interpretations match (ln base-e: max $\ln \mathbf{2} \approx \mathbf{0.693}$; $\log_{\mathbf{2}}$: max $\mathbf{1}$). Update thresholds and narrative accordingly.

2. The exact list of the 13 parameters used for UMAP and comparisons is not explicitly enumerated, and treatment of angular variables is unclear (Sec. 2.3.1). This affects both interpretability and embedding validity.

   Recommendation: In Sec. 2.3.1, list the **13** parameters explicitly and justify inclusion/exclusion (e.g., omission of phase/time). For angles, document transformation (e.g., $\mathbf{\sin} / \mathbf{\cos}$) prior to z-scoring to avoid wrap-around artifacts.

3. Parameter subspaces in the discrepancy decomposition are physically motivated but not orthogonal; correlations across subspaces can cause 'leakage' of discrepancies (Sec. 2.4.1), which matters for interpretation ("which physics drives disagreement").

   Recommendation: Add a short discussion in Sec. 2.4.1 acknowledging non-orthogonality and interpret subspace attributions accordingly. Optionally test a slightly refined partition (e.g., separating spin magnitudes from angles, or separating inclination-related extrinsics) to check qualitative robustness.

4. The role of $\mathbf{1}$-Wasserstein distance is introduced (Sec. 2.2) but is not consistently reported/used in the Results; readers may be left unsure whether it corroborates JSD-based conclusions.

   Recommendation: Either (i) present the Wasserstein results alongside JSD for a few key parameters (and comment on agreement/disagreement), or (ii) remove/streamline Wasserstein to keep the narrative focused.

5. Figures used to support key claims are difficult to read and/or under-documented (Fig. 1 and Fig. 3; Sec. 3.1–3.2). Overplotting and small fonts hinder verification; UMAP figure captions do not fully specify settings/panels.

   Recommendation: Improve Fig. 1 readability (larger panels or small multiples; show medians and $\mathbf{90\%}$ CIs; annotate with per-parameter JSD/W1). For Fig. 3, ensure caption matches all panels, report UMAP settings (including random seed), and reduce occlusion (smaller points, transparency, density contours, centroid markers).

6. Literature positioning is currently too brief given extensive prior work on waveform systematics, model-marginalized inference, and robustness diagnostics (Sec. 1, Sec. 4.2–4.3).

   Recommendation: Strengthen Sec. 1 and Sec. 4.2–4.3 with a structured comparison to prior approaches (LVK systematic studies, EOB vs Phenom comparisons, model-marginalized posteriors), clearly stating what is new here (subspace-based multivariate discrepancy attribution + visualization) and what is complementary rather than replacing existing methods.

7. Scope/generalization: at points the discussion reads as if conclusions extend broadly to 'high-mass precessing systems,' but the quantitative study is a single-event analysis (Sec. 1, Sec. 4.3).

   Recommendation: Add an explicit limitations paragraph (Sec. 4.3) clarifying which conclusions are event-specific vs plausibly general, and outline next steps (apply to other events/simulations; injection campaigns) required for general claims.

8. A cross-reference appears incorrect: Sec. 2.5.2 refers to the discrepancy decomposition as being in Sec. 4.3, although it is defined in Sec. 2.4 and applied in Sec. 3.3.

   Recommendation: Fix the cross-reference and quickly scan for other misnumbered internal references.

Very minor issues

1. Typos/formatting: split words (e.g., "substan tial"), missing spaces after periods, inconsistent math formatting for parameters, and inconsistent waveform-model naming across text/captions (Secs. 1–4; figure captions).

   Recommendation: Proofread and standardize: consistent LaTeX math notation for parameters, consistent model names (e.g., IMRPhenomXO4a), and fix spacing/line-break artifacts.

2. Bibliography formatting appears inconsistent (bullets/brackets, HTML entities, DOI/URL style). Some "physics-informed" citations may confuse readers if unrelated to the method.

   Recommendation: Normalize references to journal style, remove formatting artifacts, and ensure citations support the actual technical content (avoid implying a connection to PINNs unless relevant).

Key statements and references

- ✔ **The GW231123 event analyzed in this work is the same binary black hole merger previously identified and characterized by the LIGO–Virgo–KAGRA Collaboration as having a total mass in the range** 190–265 $M_\odot$**,**

and interpreted as a BBH merger in their dedicated discovery paper on GW231123.

- *Reference(s):* Collaboration, T. L. S., the Virgo Collaboration, & the KAGRA Collaboration. 2025
- *Justification:* Directly supported. The dedicated LIGO–Virgo–KAGRA discovery paper (Collaboration, T. L. S., the Virgo Collaboration, & the KAGRA Collaboration. 2025) explicitly analyzes GW231123, interprets it as a binary-black-hole merger, and infers a total mass between $\mathbf{190}\ M_\odot$ and $\mathbf{265}\ M_\odot$ (see title and Abstract; also Section 1: "We interpret GW231123 as a binary-black-hole merger and infer a total mass between $\mathbf{190}\ M_\odot$ and $\mathbf{265}\ M_\odot$").

- ⚠ **The posterior samples for GW231123 used in this study were originally generated in prior LVK and follow-up analyses employing five specific waveform models (NRSur7dq4, IMRPhenomXO4a, SEOBNRv5PHM, IMRPhenomXPHM, IMRPhenomTPHM), and the event itself has been proposed to originate from successive mergers of ~ 10 stellar-mass black holes in dense environments.**

- *Reference(s):* Li, Y.-J., Tang, S.-P., Xue, L.-Q., & Fan, Y.-Z. 2025
- *Justification:* Li, Y.-J., Tang, S.-P., Xue, L.-Q., & Fan, Y.-Z. 2025 state they reweight LVK posterior samples for GW231123 obtained from a Zenodo dataset and refer to a "Mixed sample," but they do not list or confirm the five specific waveform models (NRSur7dq4, IMRPhenomXO4a, SEOBNRv5PHM, IMRPhenomXPHM, IMRPhenomTPHM). They do directly argue that GW231123 likely arose from hierarchical mergers, estimating the primary and secondary components were assembled from $\sim \mathbf{6}$ and ~ 4 first-generation black holes ($\sim \mathbf{10}$ total) and discussing dense environments such as AGN disks or nuclear star clusters as plausible sites. Hence, only the second part of the statement is supported.

- ✔ **Previous independent analyses of GW231123 have argued that the event may either be a binary black hole merger or a cosmic string signal, and have also explored interpretations in terms of primordial black holes and superradiance constraints, indicating that the astrophysical nature of GW231123 is currently debated in the literature.**

- *Reference(s):* Cuceu et al., 2025, Caputo et al., 2025, Yuan et al., 2025
- *Justification:* Cuceu et al., 2025 explicitly perform a Bayesian model comparison for GW231123 between a binary black hole merger and cosmic-string burst models (cusps/kinks), finding BBH favored but treating cosmic strings as a viable alternative under test. Yuan et al., 2025 investigates a primordial black hole origin for GW231123, modeling mass function, spins, rates, and constraints. Caputo et al., 2025 uses GW231123's high spins to place superradiance (axion) constraints based on the

event. Together these independent analyses consider cosmic strings, primordial black holes, and superradiance-related physics around the event, indicating active discussion about its underlying nature.

- ⚠ **Systematic biases from waveform modeling for binary black hole populations, especially in next-generation detectors, have been shown in earlier work to be significant, motivating the physics-informed discrepancy decomposition approach adopted here to attribute model disagreements in GW231123 to specific physical approximations.**

- *Reference(s):* Kapil et al., 2024
- *Justification:* Kapil et al., 2024 summarizes earlier studies showing significant waveform-modeling systematics for BBHs in next-generation detectors and presents its own population-level bias estimates. However, it does not analyze GW231123 nor adopt a 'physics-informed discrepancy decomposition' to attribute model disagreements to specific physical approximations. Thus only the first part of the statement is supported.

- ⚠ **The multi-dimensional Jensen–Shannon divergence metric employed in this paper to quantify discrepancies between posterior distributions is based on a generalized formulation of JSD and its JS-symmetrization of distances relying on abstract means, as developed in prior theoretical work on information-theoretic divergences.**

- *Reference(s):* Nielsen, 2022
- *Justification:* Nielsen, 2022 develops generalized Jensen–Shannon divergences and JS-symmetrizations based on abstract means (Definitions 3–4, Sec. 3) and gives multi-dimensional examples such as multivariate Gaussians (Sec. 4.1). However, it does not state use for posterior distributions nor explicitly describe a ‘metric’ employed in another paper. Thus only the basis of a generalized JSD via abstract means is directly supported; the application context is not.

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 statistical distance/divergence concepts (JSD, 1-Wasserstein), KDE-based density estimation, z-score standardization, and heuristic robustness thresholds, but provides few explicit equations/derivations. The main internal mathematical consistency issue is contradictory statements about the JSD range/maximum, which depends on the logarithm base.

Checked items

1. ✖ **JSD range claim (0 to 1)** (Sec. 2.2.2, step 3, p.3)
    - **Claim:** JSD is symmetric and finite, ranging from **0** (identical) to **1** (maximally divergent).
    - **Checks:** definition consistency, boundedness claim consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** JSD is computed between two PDFs estimated by KDE, No explicit formula or log base is stated
    - **Notes:** Later sections state the theoretical maximum is $\sim$ **0.693**, contradicting the stated **0–1** range. The range depends on the log base; the paper must fix one definition and make all range/maximum statements consistent.
2. ✖ **JSD theoretical maximum claim ($\sim$ 0.693)** (Sec. 3.3.3, p.9 (Individual spin & orientation subspace discussion))
    - **Claim:** JSD values approach the theoretical maximum of approximately **0.693**.
    - **Checks:** definition consistency, boundedness claim consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** Same JSD definition as earlier sections
    - **Notes:** This maximum conflicts with the earlier statement that JSD ranges up to **1**. One of these statements is wrong unless the log base is clarified and used consistently.
3. ⚠ **Multi-dimensional JSD via multi-D KDE (well-definedness)** (Sec. 2.4.2, pp.4–5)
    - **Claim:** Compute multi-dimensional JSD between joint PDFs estimated with multi-dimensional KDE to get one scalar divergence per subspace per model pair.
    - **Checks:** missing definition check, notation/assumption completeness
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** A multi-dimensional KDE yields normalized joint PDFs, JSD is computed from these PDFs
    - **Notes:** No explicit formula for the continuous/multi-D JSD is given (integral form, mixture distribution, log base). Without that, the stated bounds (e.g., max $\sim$ **0.693**) and even the exact quantity being computed cannot be analytically verified from the paper alone.
4. ✔ **1D JSD computed from KDE-estimated PDFs** (Sec. 2.2.2, steps 2–3, p.3)
    - **Claim:** Estimate PDFs via KDE with common bandwidth per parameter, then compute JSD between PDFs.

- **Checks:** logical consistency of procedure, normalization plausibility (symbolic)
- **Verdict:** PASS; confidence: medium; impact: minor
- **Assumptions/inputs:** KDE outputs integrate to 1 (proper PDFs), Common bandwidth is used across models for a given parameter
- **Notes:** The described procedure is internally coherent at the symbolic level (density estimates then divergence). The only definitional gap is the missing explicit JSD formula/log base (handled as a separate issue).

5. ✔ **1-Wasserstein distance description** (Sec. 2.2.2, step 4, p.3)
   - **Claim:** Compute the 1-Wasserstein (Earth Mover's) distance between empirical distributions as a distance between probability distributions.
   - **Checks:** conceptual definition sanity-check
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** Univariate case per parameter as described in the section
   - **Notes:** No explicit formula is given, but the qualitative description is internally consistent with using a distributional distance in 1D.

6. ✔ **Z-score standardization before UMAP** (Sec. 2.3.1, p.3)
   - **Claim:** Standardize each parameter by subtracting the mean and dividing by the standard deviation across the combined dataset.
   - **Checks:** dimensional/units sanity-check, symbolic transformation check
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Mean and standard deviation are computed per column over the pooled samples
   - **Notes:** Z-scoring yields dimensionless standardized variables and is described consistently.

7. ✔ **Subspace dimensionality consistency (spin & orientation)** (Sec. 2.4.1 item 3, p.4; Sec. 3.3.3, p.9)
   - **Claim:** The Individual Spin & Orientation subspace is 6-dimensional with variables (a_1, a_2, $\cos_{\text{tilt}_1}$, $\cos_{\text{tilt}_2}$, $\cos_{\theta_{\text{jn}}}$, ϕ_{jl}).
   - **Checks:** dimension counting, notation consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Each listed symbol denotes one coordinate/parameter
   - **Notes:** Six variables are listed and later referenced as 6D. Minor formatting differences ($a1$ vs a_1) appear but do not change the dimensionality.

8. ✖ **Robustness threshold inconsistency** (Sec. 2.5.1, p.5 vs Sec. 3.4, p.10)

- **Claim:** A parameter is robust if pairwise divergences are below a threshold; the paper states different thresholds/criteria in different places.
- **Checks:** definition consistency
- **Verdict:** FAIL; confidence: high; impact: moderate
- **Assumptions/inputs:** Sec. 2.5.1: example threshold 'e.g., JSD < 0.01' plus mention of Wasserstein, Sec. 3.4: uses max pairwise JSD < 0.05 and median-range $< 10\%$
- **Notes:** The robustness definition changes between methods and results, which makes the conclusion "no parameter is robust" dependent on an unclear criterion.

9. ⚠ **Consensus posterior construction by pooling samples** (Sec. 2.5.3, p.5)

    - **Claim:** Consensus constraints are obtained by combining posterior samples from all five models into one aggregated dataset, then computing median and credible interval.
    - **Checks:** definition completeness, implicit weighting check
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Pooling corresponds to forming some mixture distribution over models, No model weights are specified
    - **Notes:** Pooling implies an (often equally weighted) mixture over model-conditioned posteriors, but the weights are not defined. Without this, the mathematical meaning of the "consensus" distribution is underspecified.

Limitations

- The provided content contains no explicit equations for JSD, KL, or Wasserstein distances; the audit cannot verify derivations that are not shown.
- Some parameter names appear split/garbled in the parsed text (e.g., 'redshif t'), which may be an OCR/artifact; notation issues were assessed cautiously.
- Figures/heatmaps are referenced for specific JSD values, but the analytic audit does not (and cannot) validate the numeric magnitudes—only the definitional consistency around them.

Numerical results audit

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

17 candidate numeric checks were executed: **15** PASS and **2** FAIL. Most narrative-to-table range/median consistency checks passed exactly or within the stated rounding tolerances. However, two failures (redshift medians match; UMAP_1 value match) show internal inconsistency in the reported discrepancy computations despite identical reported and computed values, so

those two items require re-verification. Several additional JSD-based claims could not be independently verified because the needed numeric matrices/values are not present as extractable text.

Checked items

1. ✔ **C1** (Page 6 (Sec. 3.1 text) vs Page 7 (Table 1, mass_2_source medians))
    - **Claim:** NRSur7dq4, SEOBNRv5PHM, and IMRPhenomTPHM infer mass_2_source medians ranging from **110.04** $M_\odot$ to **111.10** $M_\odot$.
    - **Checks:** range_consistency_from_table
    - **Verdict:** PASS
    - **Notes:** Computed min/max from the three Table 1 medians match the claimed bounds.
2. ✔ **C2** (Page 6 (Sec. 3.1 text) vs Page 7 (Table 1, mass_2_source median))
    - **Claim:** IMRPhenomXO4a predicts a median mass_2_source of only **55.08** $M_\odot$.
    - **Checks:** value_match_across_sections
    - **Verdict:** PASS
    - **Notes:** Narrative and Table 1 show the same numeric value.
3. ✔ **C3** (Page 6 (Sec. 3.1 text) vs Page 7 (Table 1, mass_2_source median))
    - **Claim:** IMRPhenomXPHM infers a lower secondary mass (**93.33** $M_\odot$).
    - **Checks:** value_match_across_sections
    - **Verdict:** PASS
    - **Notes:** Narrative and Table 1 show the same numeric value.
4. ✔ **C4** (Page 6 (Sec. 3.1 text) vs Page 7 (Table 1, chi_eff medians))
    - **Claim:** For χ_{eff}, medians span from **0.04** (IMRPhenomXPHM) to **0.44** (SEOBNRv5PHM and IMRPhenomTPHM).
    - **Checks:** min_max_and_membership_check
    - **Verdict:** PASS
    - **Notes:** Min=**0.04** and max=**0.44**; the max is attained by SEOBNRv5PHM and IMRPhenomTPHM as stated.
5. ✔ **C5** (Page 6 (Sec. 3.1 text) vs Page 7 (Table 1, chi_p medians))
    - **Claim:** χ_p shows a smaller spread in median values (from **0.73** to **0.82**).
    - **Checks:** min_max_check
    - **Verdict:** PASS
    - **Notes:** Table 1 chi_p medians have min **0.73** and max **0.82**.
6. ✖ **C6** (Page 6 (Sec. 3.1 text) vs Page 7 (Table 1, redshift medians))

- **Claim:** For redshift, IMRPhenomXPHM median is 0.17 and IMRPhenomXO4a median is 0.58.
- **Checks:** value_match_across_sections
- **Verdict:** FAIL
- **Notes:** Exec output shows reported and computed values both equal to ($0.17, 0.58$), yet the check reports a large discrepancy; this indicates an internal inconsistency in the check result that should be debugged/re-run.

7. ✔ **C7** (Page 10 (Sec. 3.4 definition) vs Page 5 (Sec. 2.5.1 example threshold))
   - **Claim:** Robustness criterion differs: Section 2.5.1 suggests JSD < 0.01 as an example, but Section 3.4 defines robust as max JSD < 0.05 plus median-range $< 10\%$.
   - **Checks:** threshold_consistency_check
   - **Verdict:** PASS
   - **Notes:** Distinct thresholds (0.01 vs 0.05) are present; this passes as a detection of inconsistency rather than an identity match.
8. ✔ **C8** (Page 10 (Sec. 3.4 text) vs Page 7 (Table 1, mass_2_source medians))
   - **Claim:** mass_2_source varies from $55.1\ M_\odot$ to $111.1\ M_\odot$.
   - **Checks:** rounded_range_from_table
   - **Verdict:** PASS
   - **Notes:** Table 1 min/max ($55.08, 111.10$) round to one decimal as ($55.1, 111.1$).
9. ✔ **C9** (Page 10 (Sec. 3.4 text) vs Page 7 (Table 1, chi_eff medians))
   - **Claim:** χ_{eff} ranges from 0.04 to 0.44.
   - **Checks:** min_max_check
   - **Verdict:** PASS
   - **Notes:** Table 1 chi_eff medians have min 0.04 and max 0.44.
10. ✔ **C10** (Page 11 (Table 3) vs Page 7 (Table 1, mass_1_source medians))
    - **Claim:** Table 3 reports mass_1_source range 129.1 - $149.9\ M_\odot$ based on model medians.
    - **Checks:** rounded_range_from_table
    - **Verdict:** PASS
    - **Notes:** Table 1 min/max ($129.14, 149.87$) round to one decimal as ($129.1, 149.9$).
11. ✔ **C11** (Page 11 (Table 3) vs Page 7 (Table 1, mass_2_source medians))
    - **Claim:** Table 3 reports mass_2_source range 55.1 - $111.1\ M_\odot$ based on model medians.
    - **Checks:** rounded_range_from_table

- **Verdict:** PASS
- **Notes:** Table 1 min/max $(\mathbf{55.08}, \mathbf{111.10})$ round to one decimal as $(\mathbf{55.1}, \mathbf{111.1})$.

12. ✔ **C12** (Page 11 (Table 3) vs Page 7 (Table 1, chi_eff medians))
    - **Claim:** Table 3 reports χ_{eff} range $\mathbf{0.04 - 0.44}$ based on model medians.
    - **Checks:** min_max_check
    - **Verdict:** PASS
    - **Notes:** Table 3 range matches min/max from Table 1 chi_eff medians.
13. ✔ **C13** (Page 11 (Table 3) vs Page 7 (Table 1, chi_p medians))
    - **Claim:** Table 3 reports χ_p range $\mathbf{0.73 - 0.82}$ based on model medians.
    - **Checks:** min_max_check
    - **Verdict:** PASS
    - **Notes:** Table 3 range matches min/max from Table 1 chi_p medians.
14. ✔ **C14** (Page 11 (Table 3) vs Page 7 (Table 1, redshift medians))
    - **Claim:** Table 3 reports redshift range $\mathbf{0.17 - 0.58}$ based on model medians.
    - **Checks:** min_max_check
    - **Verdict:** PASS
    - **Notes:** Table 3 range matches min/max from Table 1 redshift medians.
15. ✔ **C15** (Page 9 (Sec. 3.3.3) internal numerical claim)
    - **Claim:** Theoretical maximum of the Jensen-Shannon Divergence is approximately $\mathbf{0.693}$.
    - **Checks:** constant_verification
    - **Verdict:** PASS
    - **Notes:** $\mathbf{0.693}$ is consistent with $\ln(2) = \mathbf{0.693147\ldots}$ within tolerance.
16. ✔ **C16** (Page 4 (Sec. 2.3.2) UMAP hyperparameters)
    - **Claim:** Initial values of n_neighbors = $\mathbf{50}$ and min_dist = $\mathbf{0.1}$ were used as a starting point.
    - **Checks:** simple_numeric_parse_check
    - **Verdict:** PASS
    - **Notes:** Parameter-domain sanity checks passed (positive integer; min_dist in $[0, 1]$).
17. ✖ **C17** (Page 8 (Sec. 3.2 text) vs Page 8 (Table 2))
    - **Claim:** IMRPhenomXPHM is positioned at UMAP_1 $\approx \mathbf{-3.86}$ and IMRPhenomXO4a at UMAP_1 $\approx \mathbf{11.42}$.
    - **Checks:** value_match_across_sections
    - **Verdict:** FAIL

- **Notes:** Exec output shows reported and computed values both equal to ($-3.86, 11.42$), yet the check reports a large discrepancy; this indicates an internal inconsistency in the check result that should be debugged/re-run.

Limitations

- Only the provided parsed-text content was used; numeric values embedded solely in figures/heatmaps were not extracted.
- Several claims depend on pairwise JSD/Wasserstein matrices computed from posterior samples; the underlying samples/CSVs and full matrices are not included in the PDF text, so those computations cannot be independently reproduced here.
- Table 1 in the provided parse lacks a final_mass_source section, preventing cross-checks for Table 3's final_mass_source range.
- Two executed checks (C6 and C17) returned FAIL with large discrepancies despite identical reported and computed values in the execution output, indicating an internal inconsistency in the automated checking results for those items.