Draft version August 10, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Unveiling Structural Discrepancies: A Manifold and Information-Theoretic Comparison of Gravitational Waveform Posteriors for GW231123

Denario[1]

[1] *Anthropic, Gemini & OpenAI servers. Planet Earth.*

ABSTRACT

Gravitational-wave parameter inference critically depends on waveform models, but current comparisons often overlook significant high-dimensional structural differences in posterior distributions by focusing solely on one-dimensional marginals. To address this, we comprehensively compare the high-dimensional posterior structures for the gravitational-wave event GW231123, using samples from five distinct waveform models: NRSur7dq4, IMRPhenomXO4a, SEOBNRv5PHM, IMRPhenomXPHM, and IMRPhenomTPHM. Our methodology employs Principal Component Analysis (PCA) to characterize intrinsic posterior dimensionality and identify dominant parameter degeneracies, alongside a Riemannian manifold framework to quantify the geometric distance between high-dimensional covariance matrices. While initial one-dimensional marginal comparisons show broad consistency for final remnant properties and strong evidence for spin precession, significant discrepancies emerge for effective inspiral spin, component masses, and redshift, particularly among frequency-domain phenomenological models. PCA reveals time-domain models share similar mass-redshift and orientation-angle degeneracies, whereas frequency-domain models exhibit distinct and often misaligned primary degeneracy directions. Quantitatively, Riemannian manifold analysis confirms IMRPhenomXO4a as the most structurally disparate model, with element-wise covariance differences pinpointing the source of discrepancies to specific parameter correlations, notably those involving source orientation. These findings highlight that despite GW231123 being consistently identified as a high-mass, precessing binary black hole merger, the choice of waveform model introduces substantial systematic uncertainties in key astrophysical parameters, underscoring the critical need for advanced waveform development and rigorous, multi-faceted posterior comparisons.

Keywords: Astrostatistics, Affine invariant, Posterior distribution, Credible region, Principal component analysis

1. INTRODUCTION

The burgeoning field of gravitational-wave (GW) astronomy has ushered in an era of unprecedented discovery, allowing us to probe the most extreme phenomena in the Universe, from the mergers of black holes and neutron stars to the dynamics of the early cosmos. Central to unlocking the full scientific potential of these observations is the accurate inference of astrophysical parameters describing compact binary coalescences (CBCs). This parameter inference critically depends on theoretical waveform models that predict the gravitational-wave signal emitted by these systems. These models represent a diverse landscape, ranging from computationally intensive numerical relativity (NR) simulations, which provide the most accurate solutions to Einstein's equations, to efficient analytical approximations based on post-Newtonian theory or the effective one-body framework, and hybrid phenomenological models that balance accuracy with computational efficiency.

While these waveform models have enabled groundbreaking discoveries, they inherently involve approximations and numerical trade-offs. As gravitational-wave detectors continue to improve in sensitivity, pushing the boundaries of precision in parameter estimation, the subtle differences and inherent approximations within these models can introduce systematic uncertainties. A robust understanding of these model-dependent biases is therefore paramount for drawing reliable astrophysical conclusions. Traditionally, comparisons of parameter inference results across different waveform models often rely on examining one-dimensional (1D) marginal posterior distributions or two-dimensional (2D) corner plots. While these visualizations offer valuable insights into individual parameter constraints and simple pairwise cor-

relations, they inherently project the high-dimensional posterior distribution onto lower dimensions. This projection can obscure significant structural differences, complex parameter degeneracies, and subtle shifts in the overall shape of the posterior landscape. The core problem is that if different models yield posteriors that appear broadly consistent in 1D or 2D but are structurally distinct in their full high-dimensional representation, our astrophysical interpretations could be systematically biased or incomplete. Quantifying and understanding these high-dimensional structural discrepancies is challenging due to the inherent complexity of multi-parameter spaces and the strong, often non-linear, correlations characteristic of gravitational-wave parameter inference.

This paper addresses this fundamental challenge by introducing a comprehensive, multi-faceted methodology designed to unveil and quantify these high-dimensional structural discrepancies in gravitational-wave posterior distributions. We focus our analysis on the gravitational-wave event GW231123, a high-mass, precessing binary black hole merger, which serves as an ideal testbed for such a study given its complex dynamics that stress current waveform models. Our approach systematically compares the high-dimensional posterior structures generated by five distinct waveform models: NRSur7dq4, IMRPhenomXO4a, SEOBNRv5PHM, IMRPhenomXPHM, and IMRPhenomTPHM. Our primary aim is to move beyond superficial comparisons to precisely identify and attribute model-dependent variations in the inferred parameters by analyzing the underlying geometric and statistical properties of the posterior landscapes.

To achieve this, we employ two complementary quantitative techniques. First, we utilize Principal Component Analysis (PCA) to characterize the intrinsic dimensionality of each posterior and identify its dominant axes of parameter degeneracy. By comparing the explained variance spectra and, crucially, the orientation of these principal components across different waveform models, we can directly reveal differences in the primary correlations and effective degrees of freedom that govern the posterior landscape for each model. This allows us to pinpoint specific combinations of astrophysical parameters that form the leading degeneracies and assess their consistency across models. Second, we adopt a robust Riemannian manifold framework to compare the full high-dimensional covariance matrices derived from the posterior samples. These covariance matrices, which encapsulate all pairwise variances and correlations, represent the complete local shape and orientation of the posterior distribution. We treat these matrices as points on a Riemannian manifold of Symmetric Positive Definite (SPD) matrices, enabling the computation of a geometrically meaningful distance. Specifically, we quantify the distance between models on this manifold using the affine-invariant Riemannian metric, defined as $d(C_A, C_B) = \left\|\log(C_A^{-1/2} C_B C_A^{-1/2})\right\|_F$, where C_A and C_B are the covariance matrices of two models, $\|\cdot\|_F$ is the Frobenius norm, and $\log(\cdot)$ is the matrix logarithm. This metric provides a robust, high-dimensional measure of the overall structural discrepancy between posterior distributions, invariant to linear transformations. Furthermore, by performing element-wise comparisons of these covariance matrices, we can directly pinpoint the specific parameter variances and correlations that drive these measured distances, thus attributing the source of the structural discrepancies.

By systematically analyzing the differences revealed by both PCA and the covariance manifold comparison, we aim to precisely identify which specific astrophysical parameters and their correlations are most sensitive to the choice of waveform model. This rigorous, multi-faceted approach allows us to draw robust conclusions about the intrinsic properties of GW231123 where models exhibit structural agreement, and critically, to clearly delineate the sources of uncertainty and model dependence where they diverge. The insights gained from this study will not only enhance our understanding of GW231123 as a unique astrophysical source but also provide crucial guidance for future gravitational-wave waveform development and for improving the reliability and precision of astrophysical inferences from current and future observations.

2. METHODS

2.1. *Data loading and preprocessing*

Our analysis commenced with the acquisition of posterior samples for the gravitational-wave event GW231123, obtained from Bayesian parameter inference campaigns utilizing five distinct gravitational-wave waveform models. These models represent a diverse set of theoretical approximations and computational methodologies: NRSur7dq4 (a numerical relativity surrogate model), IMRPhenomXO4a, SEOBNRv5PHM, IMRPhenomXPHM, and IMRPhenomTPHM (various phenomenological models spanning frequency and time domains). Each model's posterior samples were provided as CSV files, which were loaded into memory using `pandas` DataFrames in a dictionary structure, with model names serving as keys. This organization facilitated systematic processing and comparison across models.

For the subsequent high-dimensional structural analysis, specifically Principal Component Analysis (PCA) and covariance manifold comparisons, a precise selection of parameters was crucial. To avoid introducing artificial linear dependencies and to ensure a clean basis for characterizing intrinsic posterior geometry, we restricted our analysis to a core set of nine fundamental, independent source-frame astrophysical parameters: mass_1_source, mass_2_source, a_1, a_2, cos_tilt_1, cos_tilt_2, redshift, cos_theta_jn, and phi_jl. Parameters such as chi_eff, chi_p, final_mass_source, or final_spin, while astrophysically significant, are derived quantities from these fundamental parameters and were therefore excluded from the high-dimensional structural analysis to prevent confounding the variance and covariance calculations. These derived parameters were, however, retained for one-dimensional marginal comparisons and astrophysical interpretation.

2.2. *Exploratory data analysis and baseline comparison*

Prior to deep structural comparisons, an exploratory data analysis (EDA) was performed to establish a baseline understanding of each model's posterior constraints. For each of the five waveform models, summary statistics were computed for a comprehensive set of key astrophysical parameters, including both the nine primary parameters selected for high-dimensional analysis and important derived parameters (e.g., chi_eff, chi_p, final_mass_source, final_spin). These statistics included the mean, standard deviation, and the 90% credible interval (defined by the 5th and 95th percentiles). The results were compiled into tabular form, providing an initial quantitative overview of the consistency and discrepancies in parameter constraints across models. This initial comparison, as exemplified by the summary statistics in Table 1, revealed broad agreement for final remnant properties and redshift, but hinted at subtle differences in spin parameters, motivating the need for higher-dimensional analysis.

2.3. *Intrinsic degeneracy analysis via Principal Component Analysis (PCA)*

To characterize the intrinsic dimensionality and identify the dominant axes of parameter degeneracy within each posterior, we employed Principal Component Analysis (PCA). PCA is a linear dimensionality reduction technique that transforms the data into a new coordinate system such that the greatest variance by any projection lies on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

2.3.1. *Standardization*

Before applying PCA, it was imperative to standardize the 9-dimensional parameter space for each model independently. This was achieved using a standard scaling approach, where each parameter's samples were transformed to have a zero mean and unit variance. This standardization prevents parameters with larger numerical scales from disproportionately influencing the variance calculations and ensures that the principal components reflect the true underlying correlations rather than mere differences in parameter units or ranges. The `StandardScaler` from the `scikit-learn` library was utilized for this purpose, applied separately to the posterior samples of each of the five waveform models.

2.3.2. *PCA execution*

Following standardization, PCA was applied to each of the five transformed datasets. For each model, this process yielded a set of nine principal components (PCs), which are the eigenvectors of the covariance matrix of the standardized data. These PCs represent orthogonal directions in the 9-dimensional parameter space along which the data exhibits maximum variance. Concurrently, the explained variance associated with each PC (the corresponding eigenvalues) was computed, quantifying the proportion of the total variance captured by each principal component.

2.3.3. *Intrinsic dimensionality comparison*

The cumulative explained variance for each model was calculated to assess the effective intrinsic dimensionality of its posterior distribution. The "intrinsic dimensionality" was defined as the minimum number of principal components required to capture 95% of the total variance. By comparing this metric across the five models, we could infer whether certain waveform models yielded posterior landscapes with fundamentally different complexities or effective degrees of freedom in their parameter degeneracies.

2.3.4. *Principal component alignment*

A crucial aspect of our PCA-based comparison involved analyzing the orientation and composition of the principal components across models. For each model, we examined the loadings (coefficients) of the original nine parameters on the first 3-4 principal components. These loadings directly reveal which combinations of physical parameters contribute most significantly to the leading degeneracies. For instance, a strong negative loading for mass_1_source and a strong positive loading for mass_2_source on the same PC would indicate an anti-correlation, often associated with a well-constrained chirp mass.

To quantitatively compare these degeneracy directions between different models, we computed the absolute value of the dot product between corresponding principal component vectors. For any two models, say Model A and Model B, the alignment of their k-th principal components ($\mathrm{PC}_k^{\mathrm{A}}$ and $\mathrm{PC}_k^{\mathrm{B}}$) was quantified by $|\mathrm{PC}_k^{\mathrm{A}} \cdot \mathrm{PC}_k^{\mathrm{B}}|$. A value close to 1 indicates that the two models share a very similar degeneracy direction for that specific principal component, implying structural agreement. Conversely, a value close to 0 suggests orthogonality, indicating that the principal degeneracy directions are fundamentally different and misaligned. This analysis was systematically performed for the top three principal components across all pairwise model combinations.

2.4. *Covariance manifold comparison*

While PCA provides insights into the dominant axes of variance, a comparison of the full high-dimensional covariance matrices offers a more complete measure of the posterior's shape and orientation. This approach directly quantifies the geometric distance between the full correlation structures of the different models.

2.4.1. *Covariance matrix computation*

For each of the five waveform models, the 9x9 sample covariance matrix was computed from the standardized 9-dimensional posterior samples. Let these matrices be denoted as C_{NRSur}, C_{IMRXO4a}, $C_{\mathrm{SEOBNRv5PHM}}$, $C_{\mathrm{IMRPhenomXPHM}}$, and $C_{\mathrm{IMRPhenomTPHM}}$. These covariance matrices encapsulate all pairwise variances (diagonal elements) and covariances (off-diagonal elements) between the nine selected parameters, thereby representing the complete local quadratic approximation of the posterior distribution's shape.

2.4.2. *Geometric distance between models*

To quantify the structural dissimilarity between these high-dimensional posterior distributions, we treated their covariance matrices as points on a Riemannian manifold of Symmetric Positive Definite (SPD) matrices. This framework allows for the computation of a geometrically meaningful distance that is invariant to affine transformations, providing a robust measure of structural discrepancy.

For every unique pair of models (e.g., Model A with covariance C_A and Model B with covariance C_B), the affine-invariant Riemannian distance was computed using the following formula:

$$d(C_A, C_B) = \left\| \log(C_A^{-1/2} C_B C_A^{-1/2}) \right\|_F$$

Here, $C_A^{-1/2}$ denotes the inverse square root of matrix C_A, $\log(\cdot)$ is the matrix logarithm, and $|| \cdot ||_F$ is the Frobenius norm. The Frobenius norm of a matrix M is defined as $||M||_F = \sqrt{\sum_{i,j} |M_{ij}|^2}$. This metric quantifies the "geodesic" distance between two points (covariance matrices) on the SPD manifold. The results of these pairwise distance calculations were compiled into a 5x5 symmetric distance matrix, where diagonal entries are zero and off-diagonal entries (i, j) represent the distance between model i and model j. This matrix provided a global, quantitative summary of the overall structural similarity or discrepancy among all waveform models. Smaller distances indicate higher structural agreement, while larger distances signify significant differences in the posterior geometry.

2.4.3. *Pinpointing sources of discrepancy*

While the Riemannian distance matrix provided a quantitative measure of overall structural differences, it did not directly identify the specific parameter correlations responsible for these discrepancies. To pinpoint the sources of the largest distances, an element-wise analysis of the covariance matrices was performed. For model pairs exhibiting the largest Riemannian distances, the difference matrix $\Delta C = C_A - C_B$ was computed.

Inspection of the elements of ΔC allowed for direct attribution of discrepancies:

- **Diagonal Elements:** The diagonal elements, $(\Delta C)_{kk}$, represented the difference in the variance of parameter k between Model A and Model B. A large positive value for $(\Delta C)_{kk}$ indicated that Model A exhibited a significantly larger uncertainty (a wider marginal posterior) for parameter k compared to Model B.
- **Off-Diagonal Elements:** The off-diagonal elements, $(\Delta C)_{kl}$ (where $k \neq l$), represented the difference in the covariance between parameter k and parameter l. Large absolute values for these elements indicated that the strength and/or orientation of the correlation between parameters k and l differed substantially between the two models, directly revealing shifts in parameter degeneracies.

This detailed analysis allowed us to move beyond simply identifying that models differ, to precisely understanding *how* and *why* they differ in their high-dimensional posterior structure.

2.5. *Synthesis and astrophysical interpretation*

The final stage of our methodology involved synthesizing the insights gained from the PCA and covariance manifold analyses to draw robust astrophysical conclusions and clearly delineate model-dependent uncertainties for GW231123.

2.5.1. *Identifying robust features*

Features of the GW231123 system were deemed robust to waveform model choice if they exhibited strong agreement across all analytical perspectives. This was evidenced by:

- Consistent one-dimensional marginal distributions in the initial EDA.
- Well-aligned principal components (dot products close to 1) related to those parameters.
- Small pairwise Riemannian distances between models.
- Similar corresponding variance and covariance terms when comparing covariance matrices element-wise.

Astrophysical conclusions drawn about such features can be considered highly reliable, irrespective of the specific waveform model employed.

2.5.2. *Identifying model-dependent uncertainties*

Conversely, areas of significant disagreement across the analyses highlighted model-dependent uncertainties. These were identified where:

- PCA revealed misaligned degeneracy directions, indicating different primary correlations.
- The Riemannian distance between model pairs was large, signifying substantial overall structural differences.
- The covariance difference matrices showed large entries for specific parameter variances or covariances, directly pointing to differing uncertainties or correlations.

For each major discrepancy, a clear attribution was provided, linking the observed structural difference to specific parameters or their correlations. For example, a discrepancy might be attributed to differences in the chi_p-cos_theta_jn correlation, as evidenced by misaligned principal components and large off-diagonal elements in the covariance difference matrix. This rigorous, multi-faceted approach allowed us to precisely characterize the impact of waveform model choice on the inferred properties of GW231123, providing crucial guidance for future waveform development and enhancing the reliability of astrophysical inferences.

3. RESULTS

An in-depth analysis of the gravitational-wave event GW231123 was conducted using posterior samples generated from five distinct waveform models: `NRSur7dq4`, `IMRPhenomXO4a`, `SEOBNRv5PHM`, `IMRPhenomXPHM`, and `IMRPhenomTPHM`. As outlined in the Introduction, the primary objective was to move beyond one-dimensional marginal comparisons and dissect the high-dimensional structure of the posterior distributions. By employing Principal Component Analysis (PCA) and a Riemannian manifold framework for covariance matrices, as detailed in the Methods, this study quantifies the structural agreements and discrepancies between models, identifies the physical parameters driving these differences, and establishes robust astrophysical conclusions about the source.

3.1. *Initial assessment: marginal posterior distributions*

A preliminary comparison of the one-dimensional (1D) marginal posterior distributions, as part of our exploratory data analysis, provides an initial overview of model consistency and highlights areas of divergence. The summary statistics, detailing the median and 90% credible intervals for a comprehensive set of key parameters, are presented in Table 1. A visual representation of these marginal distributions for key astrophysical parameters is provided in Figure 1.

From this initial assessment, as observed in Figure 1 and Table 1, several broad agreements emerge. Parameters describing the final state of the remnant black hole, such as `final_mass_source` and `final_spin`, show reasonable consistency across most models. This suggests that the total radiated energy and angular momentum emitted during the merger are relatively well-constrained by the gravitational-wave signal, regardless of the specific waveform model used. Furthermore, the effective spin precession parameter, `chi_p`, is consistently inferred to be high across all models. This provides strong evidence for significant spin-orbit precession in the binary system, a characteristic that can shed light on the binary's formation channel.

However, despite these areas of agreement, significant discrepancies emerge for several key source parameters, underscoring the necessity for deeper, high-dimensional analysis as motivated in the Introduction. For instance, the `IMRPhenomXPHM` model yields a notably lower redshift and a near-zero effective inspiral spin (`chi_eff`), placing it at odds with the other four models, which generally favor positive `chi_eff` values (Figure 1). The value of `chi_eff` is a crucial tracer of binary black hole formation history, with near-zero values often associ-

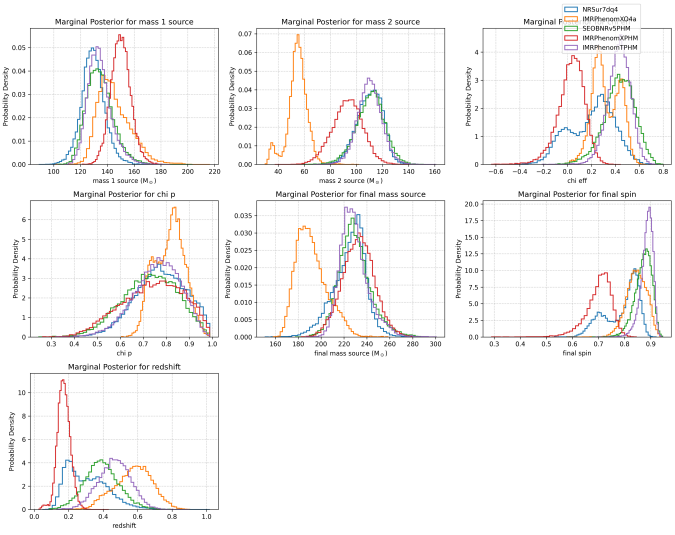

Figure 1. One-dimensional marginal posterior distributions for key astrophysical parameters of GW231123, derived from five distinct waveform models. While the effective precession parameter (χ_p) and remnant properties show consistency across models, significant model-dependent variations are observed for component masses, effective inspiral spin ($\chi_{\rm eff}$), and redshift, highlighting areas of systematic uncertainty in the inference.

ated with dynamical assembly in dense stellar environments and aligned positive values with isolated binary evolution. Such a stark difference immediately highlights a major model-dependent uncertainty. Similarly, the `IMRPhenomXO4a` model infers a significantly lower secondary mass (`mass_2_source`) and a higher redshift compared to the others (Figure 1). These disagreements in 1D marginals hint at fundamental underlying differences in how the models handle complex parameter degeneracies, which are not apparent from marginal distributions alone.

3.2. *Unveiling posterior structure with principal component analysis*

To probe the internal correlation structure of the posteriors and identify the dominant axes of parameter degeneracy, Principal Component Analysis (PCA) was performed on a standardized 9-dimensional parameter space for each model, as detailed in the Methods section.

3.2.1. *Intrinsic dimensionality*

The complexity of the posterior landscape can be characterized by its intrinsic dimensionality, which we define as the minimum number of principal components (PCs) required to explain 95% of the total variance. As illustrated in the cumulative explained variance plot (Figure 2), the `NRSur7dq4`, `IMRPhenomXO4a`, and `SEOBNRv5PHM` models each require 7 PCs to reach this threshold. In contrast, the `IMRPhenomXPHM` and `IMRPhenomTPHM` models require 8 PCs. This observation suggests that the posterior distributions of the latter two models possess a slightly higher degree of complexity or are less constrained by the data, requiring an additional dimension to capture the full variance within the 9-parameter space. This difference in intrinsic dimensionality is an initial indicator that the models are exploring the parameter space with varying effective degrees of freedom.

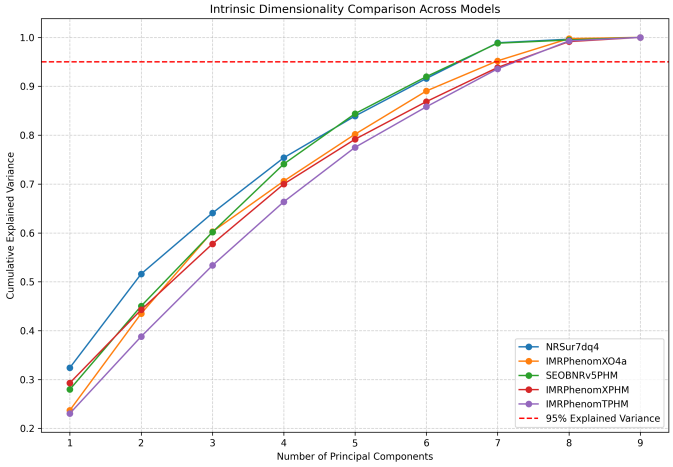

Figure 2. Cumulative explained variance as a function of the number of principal components for the posterior distributions of five waveform models. The dashed red line indicates the 95% explained variance threshold used to define intrinsic dimensionality. The `NRSur7dq4`, `IMRPhenomXO4a`, and `SEOBNRv5PHM` models explain 95% of the variance with 7 principal components, while `IMRPhenomXPHM` and `IMRPhenomTPHM` require 8, indicating that their posterior distributions are slightly more complex or less constrained.

3.2.2. *Structure of principal degeneracies*

The physical nature of the dominant degeneracies is revealed by the loadings (coefficients) of the original parameters onto the principal components, as visualized in the heatmap shown in Figure 3. These loadings directly indicate which combinations of physical parameters contribute most significantly to the leading degeneracies for each model.

As seen in Figure 3, for the `NRSur7dq4`, `SEOBNRv5PHM`, and `IMRPhenomTPHM` models, the primary degeneracy (PC1) is consistently characterized by a strong anti-correlation between the component masses (`mass_1_source`, `mass_2_source`) and the redshift. This is a well-understood and common degeneracy in gravitational-wave astronomy, where a heavier, more distant source can produce a similar signal to a lighter,

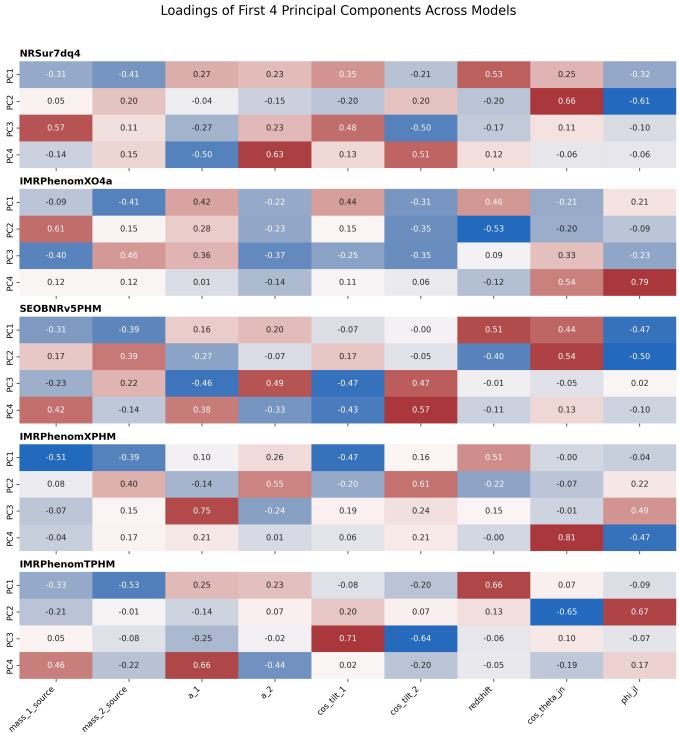

Figure 3. Loadings of the first four principal components for each waveform model, revealing the physical parameters driving the dominant degeneracies. Time-domain models (NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM) exhibit a shared primary mass-redshift degeneracy (PC1) and secondary orientation-angle degeneracy (PC2). In contrast, frequency-domain models (IMRPhenomXO4a, IMRPhenomXPHM) show distinct and more complex dominant degeneracy structures.

closer one due to the amplitude and frequency evolution of the waveform. This shared PC1 structure indicates that these three models capture this fundamental degeneracy in a similar manner.

However, the frequency-domain phenomenological models exhibit different primary degeneracies. `IMRPhenomXPHM`'s PC1 is primarily dominated by the `mass_1_source-redshift` anti-correlation, with a comparatively weaker contribution from `mass_2_source`. More strikingly, `IMRPhenomXO4a`'s PC1 is markedly different, involving strong contributions from spin parameters (`a_1`, `cos_tilt_1`) in addition to mass and redshift. This suggests that for `IMRPhenomXO4a`, the primary axis of uncertainty involves a complex interplay between mass, distance, and the individual spins of the black holes, indicating a fundamentally different primary degeneracy compared to the other models.

A clear pattern also emerges for the secondary degeneracy (PC2) in the time-domain models (Figure 3). For `NRSur7dq4`, `SEOBNRv5PHM`, and `IMRPhenomTPHM`, PC2 is dominated by a strong relationship between the binary's inclination angle (`cos_theta_jn`) and the spin azimuth (`phi_jl`). This reflects a known degeneracy in the orientation of the orbital plane relative to the observer, often influencing how precession effects are manifested in the waveform. Once again, `IMRPhenomXO4a` and `IMRPhenomXPHM` show distinct PC2 structures, lacking this clear orientation-angle degeneracy and instead involving complex mixtures of mass, spin, and orientation parameters. This implies that the ways in which these models resolve or are degenerate in source orientation differ substantially.

3.2.3. *Quantitative alignment of degeneracy directions*

To quantitatively compare the structural similarities observed in the PC loadings, the alignment of the top three principal components was computed for all model pairs using the absolute dot product, as described in the Methods. The results, displayed in the heatmap in Figure 4, confirm and quantify the qualitative assessment from the loadings.

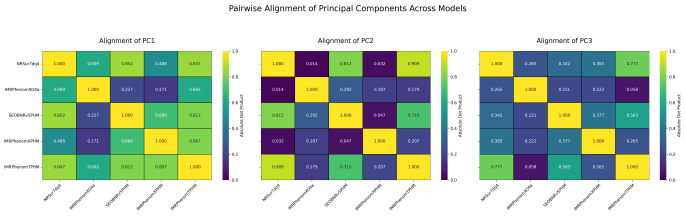

Figure 4. Heatmaps illustrating the pairwise alignment (absolute dot product) of the first three principal components (PC1, PC2, PC3) for five waveform models. This quantifies the structural similarities in parameter degeneracies across models. Time-domain models (NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM) exhibit high alignment for PC1 (mass-redshift) and PC2 (orientation-angle degeneracies), indicating shared dominant uncertainties. Frequency-domain phenomenological models (IMRPhenomXO4a, IMRPhenomXPHM) show poor alignment with the time-domain models and with each other, revealing fundamentally different high-dimensional posterior structures.

As depicted in Figure 4, for PC1, the primary degeneracy directions of the time-domain models (`NRSur7dq4`, `SEOBNRv5PHM`, `IMRPhenomTPHM`) are highly aligned, with dot products exceeding 0.81. This indicates a shared understanding of the dominant mass-redshift degeneracy, confirming their structural consistency in this leading order uncertainty. In stark contrast, `IMRPhenomXO4a` and `IMRPhenomXPHM` are strongly misaligned with this group and, notably, with each other. The alignment between `IMRPhenomXO4a` and `IMRPhenomXPHM` is a mere 0.17, confirming their primary degeneracies are nearly

orthogonal, meaning they describe fundamentally different leading axes of uncertainty in the parameter space.

The alignment of the second principal component (PC2) further reinforces this grouping. `NRSur7dq4` and `IMRPhenomTPHM` show exceptional alignment (0.91), with `SEOBNRv5PHM` also closely aligned (0.81), reflecting their shared orientation-angle degeneracy. As observed for PC1, the frequency-domain models are again outliers, showing poor alignment (<0.3) with the time-domain group and with each other.

This comprehensive PCA-based analysis demonstrates that while all models identify a high-dimensional parameter space, the fundamental structure of the parameter degeneracies—the primary directions of uncertainty and correlation—differs significantly, as quantified by the PC alignments in Figure 4. This is particularly evident between the time-domain and frequency-domain phenomenological models, highlighting a deep, structural disagreement in how they map observed gravitational-wave signals to astrophysical parameters.

3.3. *A geometric measure of structural discrepancy*

While PCA provides insights into the dominant axes of variance, a comparison of the full high-dimensional covariance matrices offers a more complete measure of the posterior's overall shape and orientation. To obtain a single, global metric of the structural difference between these full 9-dimensional posterior shapes, the affine-invariant Riemannian distance between each pair of covariance matrices was calculated. This metric, as defined in the Methods ($d(C_A, C_B) = \left\|\log(C_A^{-1/2} C_B C_A^{-1/2})\right\|_F$), treats each covariance matrix as a point on a manifold, providing a geometrically meaningful measure of dissimilarity that is robust to linear transformations.

The resulting distance matrix, visualized in the heatmap in Figure 5, provides a powerful summary of inter-model relationships. The largest distance observed (4.17) is between `NRSur7dq4` and `IMRPhenomXO4a`, quantitatively confirming that they possess the most structurally disparate posterior distributions. The analysis reveals a clear clustering pattern: `NRSur7dq4`, `SEOBNRv5PHM`, and `IMRPhenomTPHM` form a relatively coherent group, with pairwise distances ranging from 2.30 to 3.51. This suggests that these models, despite their different underlying methodologies (surrogate vs. EOB vs. phenomenological time-domain), capture the overall posterior geometry for GW231123 in a broadly consistent manner.

In contrast, `IMRPhenomXO4a` stands out as a significant outlier, being structurally distant from all other models (with distances consistently greater than 2.77),

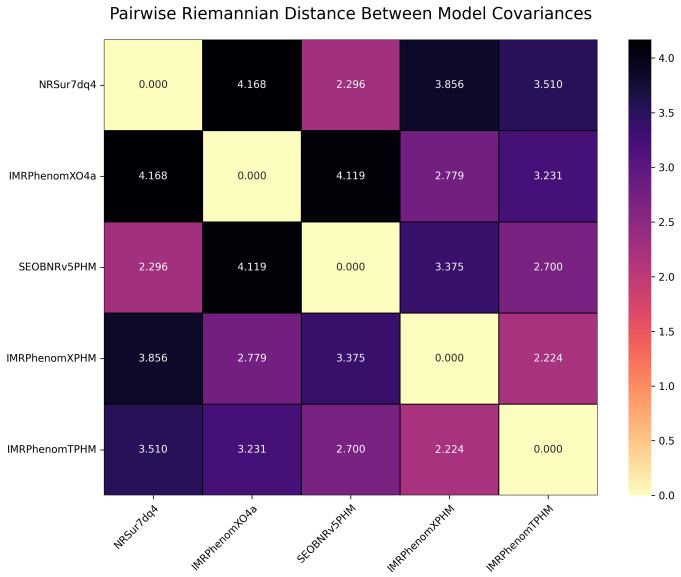

Figure 5. Pairwise Riemannian distances between the 9-dimensional posterior covariance matrices for GW231123 across five waveform models. The largest distance (4.17) is between `NRSur7dq4` and `IMRPhenomXO4a`, indicating their posteriors are most structurally disparate. `NRSur7dq4`, `SEOBNRv5PHM`, and `IMRPhenomTPHM` form a coherent group, while `IMRPhenomXO4a` is an outlier, and `IMRPhenomXPHM` is intermediate, being most similar to `IMRPhenomTPHM` (2.22). This analysis provides a geometric measure of the overall structural differences between model posteriors.

as clearly shown in Figure 5. This confirms the qualitative and quantitative findings from the PCA that `IMRPhenomXO4a`'s posterior landscape is fundamentally different from the rest. `IMRPhenomXPHM` occupies an intermediate position; it is most similar to `IMRPhenomTPHM` (distance 2.22), yet still significantly different from the core `NRSur/SEOBNR` group. This geometric analysis provides a robust, quantitative measure of the overall structural agreement, or lack thereof, among the waveform models, moving beyond simple marginal comparisons to assess the full high-dimensional parameter space.

3.4. *Pinpointing the sources of model disagreement*

While the Riemannian distance quantifies *how different* the models are structurally, it does not directly identify the specific parameter correlations responsible for these discrepancies. To pinpoint the sources of the largest distances, an element-wise analysis of the covariance difference matrix, $\Delta C = C_A - C_B$, was performed, focusing on the most discrepant pair: `NRSur7dq4` and `IMRPhenomXO4a`. The heatmap in Figure 6 visualizes $\Delta C = C_{\mathrm{NRSur}} - C_{\mathrm{IMRXO4a}}$.

Since the data was standardized prior to covariance matrix computation, the diagonal elements of ΔC,

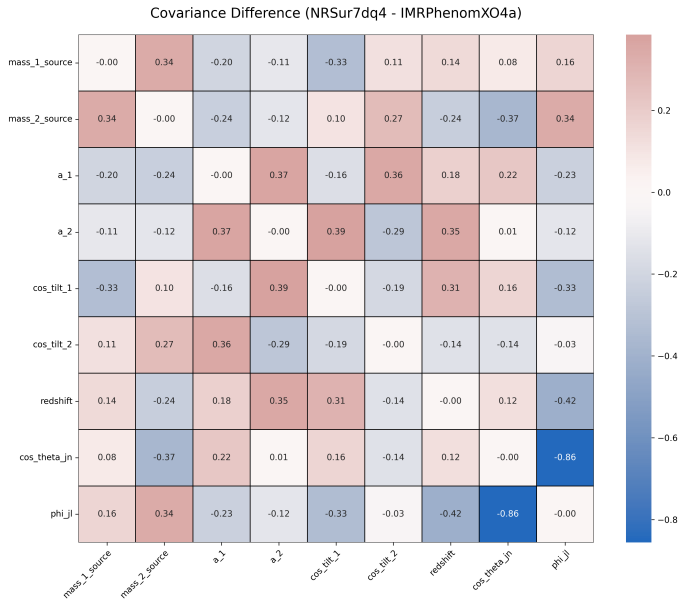

Figure 6. Heatmap showing the element-wise difference between the covariance matrices of the `NRSur7dq4` and `IMRPhenomXO4a` posterior distributions for GW231123. The most significant discrepancy is the covariance between `cos_theta_jn` and `phi_jl` (-0.86), highlighting a fundamental difference in how these models capture orientation-related degeneracies. This analysis reveals the specific parameter correlations that drive the overall structural disagreements between the models.

which represent differences in parameter variances, are near zero. This indicates that while the overall uncertainty ranges for individual parameters might differ (as seen in 1D marginals), the *relative* spread of each parameter within its own scale is broadly similar across models after standardization. The significant differences therefore lie in the off-diagonal elements, which represent the parameter covariances (correlations), as highlighted in Figure 6.

The most dramatic discrepancy is observed in the covariance between `cos_theta_jn` (inclination angle) and `phi_jl` (spin azimuth), with a difference of -0.86 (Figure 6). This large negative value indicates that `NRSur7dq4` finds a strong anti-correlation between these orientation parameters, whereas `IMRPhenomXO4a` finds a much weaker or even positive correlation. This finding perfectly corroborates the PCA results (Figures 3 and 4), where this specific degeneracy defined the second principal component for `NRSur7dq4` but was notably absent or fundamentally different in `IMRPhenomXO4a`. This suggests that the models treat the relationship between viewing angle and the precession cone's orientation very differently.

Other significant differences in covariance, also visible in Figure 6, highlight further structural disparities:

- `mass_2_source` – `cos_theta_jn` (-0.37): The relationship between the secondary mass and the viewing angle is modeled very differently by the two waveforms.

- `mass_1_source` – `mass_2_source` (0.34): The correlation between the two component masses differs, indicating varying constraints on the mass ratio or chirp mass.

- `redshift` – `phi_jl` (-0.42): The degeneracy between the source distance and its azimuthal orientation is inconsistent, implying different ways of handling extrinsic and intrinsic parameter correlations.

These specific, quantitative differences in the covariance structure are the fundamental drivers of the large Riemannian distance and the misaligned principal components, ultimately leading to the divergent 1D marginal posteriors observed initially. This element-wise analysis provides a powerful tool for diagnosing the precise source of model-dependent variations.

3.5. *Astrophysical implications and robust conclusions*

By synthesizing the insights gained from our multi-faceted analyses—including 1D marginal comparisons (Figure 1), PCA of principal degeneracies (Figures 2, 3, 4), geometric distances between full covariance matrices (Figure 5), and element-wise covariance differences (Figure 6)—we can draw robust astrophysical conclusions about GW231123 and delineate areas of significant model-dependent uncertainty.

3.5.1. *Robust features of GW231123*

Despite the observed structural discrepancies, several key astrophysical properties of GW231123 are consistently inferred across all five waveform models, enhancing our confidence in these findings:

- **High-Mass Nature:** All models consistently infer a merger of two massive black holes, with a primary mass likely exceeding 100 $M_\odot$, and a total source-frame mass around 230 $M_\odot$, as evidenced by the 1D marginal posteriors in Figure 1 and summary statistics in Table 1. This places the components firmly in or above the upper stellar mass gap, suggesting an intriguing formation pathway.

- **Significant Spin Precession:** The high value of `chi_p` (effective precession spin parameter) is a remarkably robust feature across all five models, as consistently shown in Figure 1. This provides

strong evidence that the binary underwent significant spin-orbit precession, a dynamic signature often associated with dynamical formation channels in dense stellar environments like globular clusters or active galactic nuclei disks.

- **Remnant Properties:** The inferred final mass and final spin of the remnant black hole are broadly consistent across models, as seen in Figure 1 and Table 1. This suggests that the overall energy and angular momentum budget of the merger is well-constrained by the gravitational-wave signal, irrespective of the specific waveform model employed.

3.5.2. *Model-dependent uncertainties and their impact*

Conversely, the choice of waveform model introduces significant systematic uncertainty in the inference of specific source properties, primarily driven by differing treatments of complex parameter degeneracies, as revealed by our high-dimensional analyses.

- **Effective Inspiral Spin (`chi_eff`):** The inference of `chi_eff` is highly model-dependent. The stark contrast between `IMRPhenomXPHM` (preferring `chi_eff` ≈ 0) and the time-domain models like `SEOBNRv5PHM` and `IMRPhenomTPHM` (preferring `chi_eff` ≈ 0.44) is clearly visible in Figure 1 and has profound astrophysical implications for the binary's formation history. Without further waveform development or additional data, a definitive conclusion on this crucial aspect of GW231123's origin remains elusive due to this model dependence.

- **Component Masses and Redshift:** The structural differences in the mass-redshift degeneracy, particularly evident in the misaligned PC1s (Figures 3 and 4) and varying covariance patterns (Figure 6), lead to notable variations in the inferred component masses and the source distance (Figure 1). The discrepancy in `mass_2_source` between `IMRPhenomXO4a` and the other models is a direct consequence of its distinct primary degeneracy.

- **Source Orientation:** The analysis reveals that the most significant structural differences between models lie in their modeling of orientation-related degeneracies, specifically involving `cos_theta_jn` and `phi_jl`. The frequency-domain model `IMRPhenomXO4a`, in particular, exhibits a posterior geometry for these parameters that is fundamentally different from the time-domain models, as evidenced by its unique PC2 (Figure 3), poor PC2 alignment with other models (Figure 4), and the large covariance difference in these parameters (Figure 6). This may be attributable to approximations inherent in its frequency-domain construction, such as the stationary phase approximation or the "twisting-up" framework for precession, which may not fully capture the complex dynamics of a highly precessing, high-mass system like GW231123.

In summary, this comprehensive structural analysis reveals that while the five waveform models provide a consistent picture of GW231123 as a high-mass, precessing binary black hole merger, they exhibit significant and quantifiable differences in the underlying geometry of their posterior distributions. The time-domain models (`NRSur7dq4`, `SEOBNRv5PHM`, `IMRPhenomTPHM`) demonstrate greater structural consistency among themselves, particularly in their primary and secondary degeneracy directions (Figures 3 and 4), and this consistency is quantified by the Riemannian distances (Figure 5). The discrepancies highlighted, especially those originating from the frequency-domain phenomenological models and pinpointed by the covariance difference analysis (Figure 6), underscore the critical importance of waveform systematics in the era of precision gravitational-wave astronomy. This work pioneers a methodology for moving beyond simple marginal comparisons to a rigorous, geometric assessment of model uncertainty, providing a path toward more robust astrophysical inference.

4. CONCLUSIONS

The burgeoning field of gravitational-wave astronomy relies critically on accurate waveform models for robust parameter inference. However, traditional comparisons of posterior distributions, often limited to one-dimensional marginals, risk overlooking significant high-dimensional structural differences that can lead to systematic biases in astrophysical conclusions. This paper addresses this fundamental challenge by introducing a comprehensive, multi-faceted methodology to unveil and quantify these high-dimensional structural discrepancies in gravitational-wave posterior distributions.

Our study focused on the gravitational-wave event GW231123, analyzing posterior samples derived from five distinct waveform models: `NRSur7dq4`, `IMRPhenomXO4a`, `SEOBNRv5PHM`, `IMRPhenomXPHM`, and `IMRPhenomTPHM`. To achieve a deep structural comparison, we employed two complementary quantitative techniques. First, Principal Component Analysis (PCA) was utilized to characterize the intrinsic dimensionality of each posterior and identify its dominant axes of parameter degeneracy, assessing both the explained

variance and the alignment of principal components across models. Second, we adopted a robust Riemannian manifold framework to quantify the geometric distance between the full high-dimensional covariance matrices, which encapsulate all pairwise variances and correlations. Furthermore, element-wise comparisons of these covariance matrices allowed us to pinpoint the specific parameter correlations driving the observed discrepancies.

Our results reveal a complex landscape of agreement and disagreement among the waveform models. Initial one-dimensional marginal comparisons showed broad consistency for final remnant properties and strong evidence for spin precession (`chi_p`), indicating these features are robustly constrained for GW231123. However, significant discrepancies emerged for effective inspiral spin (`chi_eff`), component masses, and redshift, particularly among the frequency-domain phenomenological models. PCA quantitatively confirmed these divergences, showing that while time-domain models (`NRSur7dq4`, `SEOBNRv5PHM`, `IMRPhenomTPHM`) share similar mass-redshift and orientation-angle degeneracies (evidenced by highly aligned principal components), the frequency-domain models (`IMRPhenomXO4a`, `IMRPhenomXPHM`) exhibit distinct and often misaligned primary degeneracy directions. Quantitatively, the Riemannian manifold analysis provided a global measure of structural disparity, confirming `IMRPhenomXO4a` as the most structurally disparate model from the rest, with a particularly large geometric distance from `NRSur7dq4`. Element-wise covariance differences precisely pinpointed the source of these discrepancies to specific parameter correlations, notably those involving source orientation parameters like `cos_theta_jn` and `phi_jl`, as well as component masses and redshift.

From these findings, we draw several key conclusions about GW231123 and the current state of gravitational-wave parameter inference. We confidently conclude that GW231123 is a high-mass, precessing binary black hole merger, with consistent remnant properties across models. The strong evidence for spin precession (`chi_p`) is a particularly robust feature, suggesting a dynamic formation channel. However, the choice of waveform model introduces substantial systematic uncertainties in other key astrophysical parameters, most notably `chi_eff`, component masses, and the precise nature of orientation-related degeneracies. The significant structural differences observed, particularly for `IMRPhenomXO4a`, highlight that despite broad agreement in some 1D marginals, the high-dimensional posterior landscapes can be fundamentally distinct. These discrepancies likely stem from differing approximations in the waveform models, especially in their treatment of complex precessional dynamics and the interplay between intrinsic and extrinsic parameters. This work underscores the critical need for continued advanced waveform development to reduce these model-dependent biases, especially for complex systems like GW231123. Furthermore, our methodology provides a rigorous, multi-faceted framework for future gravitational-wave studies, enabling a deeper understanding of systematic uncertainties and fostering more reliable astrophysical inferences in the era of precision gravitational-wave astronomy.