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

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

Denario[1]

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

ABSTRACT

Gravitational wave parameter estimation is susceptible to systematic uncertainties arising from the choice of waveform model, a challenge particularly acute for complex events like GW231123. We present a comprehensive, data-driven framework to systematically quantify, attribute, and mitigate these model-dependent discrepancies, aiming for more robust astrophysical inferences. Using posterior distributions for GW231123 derived from five distinct waveform models, we quantified discrepancies at both parameter-specific (Jensen-Shannon divergence) and global (Sliced Wasserstein Distance, UMAP) scales. Our core innovation is a feature-based diagnostic that correlates observed discrepancies with intrinsic model characteristics such as domain, calibration method, and treatment of precession or higher-order modes. This analysis revealed significant discrepancies, primarily linked to frequency-domain, phenomenological models (IMRPhenomXPHM and IMRPhenomXO4a), which notably lacked comprehensive higher-order mode or precession physics and exhibited the largest deviations from the numerical relativity surrogate. To provide a robust characterization of the source, we employed Bayesian Model Averaging, weighting each model's contribution by its approximate evidence. This yielded a definitive meta-posterior for GW231123, establishing its primary black hole mass at $134.9^{+24.0}_{-14.6}\,M_\odot$ and confirming strong evidence for significant spin-induced precession ($\chi_p = 0.79^{+0.13}_{-0.19}$). The merger formed an intermediate-mass black hole of approximately $221\,M_\odot$. Our findings underscore the critical role of waveform model features in influencing parameter estimates and provide a robust, uncertainty-quantified characterization of GW231123 as a high-mass binary in the pair-instability supernova mass gap, likely formed through dynamical pathways.

Keywords: Dimensionality reduction, Bayesian information criterion, Hierarchical models, Redshifted, Bayesian statistics

1. INTRODUCTION

The nascent field of gravitational-wave (GW) astronomy has opened an unprecedented window into the most extreme phenomena in the Universe, primarily through the direct detection of binary black hole (BBH) mergers. Interpreting these faint signals and extracting their astrophysical properties, such as component masses, spins, and orbital dynamics, fundamentally relies on highly accurate theoretical waveform models. These models are sophisticated numerical or analytical approximations of Einstein's General Theory of Relativity, developed through various approaches including post-Newtonian (PN) expansions, effective-one-body (EOB) theory, direct numerical relativity (NR) simulations, and phenomenological fits.

However, the immense complexity of the two-body problem in strong-field gravity, coupled with the computational demands of solving Einstein's equations, necessitates the use of diverse modeling strategies. Each approach involves distinct approximations, calibration methods, and inherent assumptions regarding the physical phenomena captured, such as spin-induced precession or the excitation of higher-order modes in the GW signal. While this diversity fosters continuous improvement and exploration of different physical regimes, it introduces a critical challenge: systematic uncertainties arising from the choice of waveform model. For many GW events, simpler models may suffice, yielding consistent parameter estimates. Yet, for complex or extreme events—such as those involving high masses, significant spin-induced precession, or prominent higher-order modes—the underlying approximations in different waveform models can lead to discernible and potentially significant discrepancies in the inferred source properties. These systematic biases are problematic be-

cause they can obscure true astrophysical features, distort population inference studies, and lead to an underestimation of the total statistical uncertainty in our measurements. Addressing these model-dependent discrepancies is therefore paramount for robust astrophysical inference and for guiding the development of next-generation waveform models.

The gravitational-wave event GW231123, identified as a high-mass binary black hole merger, serves as an ideal testbed for investigating such model discrepancies. Its inferred parameters place it squarely within the astrophysically significant pair-instability supernova mass gap, and the signal suggests the presence of significant spin-induced precession. These characteristics make GW231123 a complex event where the approximations inherent in different waveform models are likely to lead to substantial variations in parameter estimates. Given the profound astrophysical implications of GW231123's characterization, a thorough understanding and mitigation of waveform model systematics are essential for drawing reliable conclusions.

This paper presents a novel, data-driven framework designed to systematically quantify, attribute, and ultimately mitigate the impact of waveform model discrepancies on astrophysical parameter estimation. Our approach extends beyond merely comparing posterior distributions; the core innovation lies in our ability to understand *why* different models yield different results by linking observed discrepancies to specific, intrinsic features of the waveform models themselves. We leverage posterior distributions for GW231123 obtained from five distinct waveform models, encompassing a wide range of theoretical approaches, calibration methods, and treatments of complex physics.

Our methodology unfolds in three primary stages to address the problem. First, we perform a multi-scale comparison of the posterior distributions. We quantify parameter-specific discrepancies using metrics like the Jensen-Shannon (JS) divergence, which provides a bounded and symmetric measure of difference between one-dimensional marginal posteriors. Simultaneously, we assess global differences across the high-dimensional parameter space using techniques such as Uniform Manifold Approximation and Projection (UMAP) for qualitative visualization and the Sliced Wasserstein Distance (SWD) for a quantitative measure of overall posterior separation. Second, and central to our novel contribution, we develop a feature-based diagnostic framework. This involves systematically characterizing each waveform model by its fundamental properties, such as its domain (time or frequency), family (e.g., phenomenological, EOB, NR-surrogate), calibration method, and the specific inclusion or treatment of complex physics like spin-induced precession and higher-order modes. We then employ statistical techniques, specifically correlation analysis, to identify which of these intrinsic model features are most strongly correlated with the observed posterior discrepancies. This attribution step provides actionable insights into which aspects of waveform modeling contribute most significantly to systematic uncertainties, thereby directly guiding future model development efforts. Finally, to provide the most robust and uncertainty-quantified astrophysical characterization of GW231123, we employ Bayesian Model Averaging (BMA). This principled ensemble method combines the individual posteriors, weighting each model's contribution by its approximate evidence, thereby explicitly accounting for inter-model variability and yielding a definitive "meta-posterior" that marginalizes over waveform model uncertainty.

By applying this comprehensive framework to GW231123, we verify that waveform model features significantly influence parameter estimates, with notable discrepancies primarily arising from models lacking comprehensive higher-order mode or precession physics. Our analysis provides a robust characterization of GW231123's properties, establishing its primary black hole mass at $134.9^{+24.0}_{-14.6}\,M_\odot$ and confirming strong evidence for significant spin-induced precession ($\chi_p = 0.79^{+0.13}_{-0.19}$). The merger formed an intermediate-mass black hole of approximately $221\,M_\odot$. These findings underscore the critical importance of systematically addressing model uncertainties and offer a blueprint for future robust astrophysical inferences in gravitational-wave astronomy, particularly for high-mass binaries in the pair-instability supernova mass gap, likely formed through dynamical pathways.

2. METHODS

This section details the comprehensive methodology employed to quantify, attribute, and mitigate waveform model discrepancies in the analysis of gravitational-wave event GW231123. Our framework is structured into four main stages: Data Ingestion and Exploratory Analysis, Multi-Scale Posterior Comparison, Model Discrepancy Attribution, and Robust Astrophysical Inference. This systematic approach aims to move beyond simple posterior comparisons by providing actionable insights into the sources of model-dependent uncertainties, ultimately yielding a more reliable characterization of the astrophysical source.

2.1. *Data Ingestion and Exploratory Analysis*

The initial phase of our analysis involved the acquisition, standardization, and preliminary examination of

the posterior samples for GW231123 derived from various waveform models. This step is crucial for establishing a consistent foundation for all subsequent comparisons.

2.1.1. *Data Loading and Unification*

Posterior samples for GW231123, generated using five distinct gravitational-wave waveform models, were obtained in CSV format. These models encompass a wide range of theoretical approaches and calibration methods, including Numerical Relativity (NR) surrogates, Effective-One-Body (EOB) models, and phenomenological models, as detailed in Table 2. Each CSV file, representing the posterior distribution from a specific model, was loaded into a unified data structure. During this process, a new categorical variable, `model_name`, was assigned to each sample, uniquely identifying its originating waveform model (e.g., 'NRSur7dq4', 'IMRPhenomXO4a', 'SEOBNRv5PHM', 'IMRPhenomXPHM', 'IMRPhenomTPHM'). This consolidation enabled a streamlined, comparative analysis across all models. Post-ingestion, the unified dataset underwent integrity checks to identify and address any missing values or anomalies, ensuring the reliability of the subsequent analyses.

2.1.2. *Exploratory Data Analysis and Summary Statistics*

To gain an initial understanding of the agreement and disagreement among the different waveform models, an exploratory data analysis (EDA) was performed. For each of the five models, key summary statistics were computed for the primary astrophysical parameters. Specifically, the median and the 90% credible interval (defined by the 5th and 95th percentiles) were calculated for the source-frame primary mass (m_1^{source}), the effective inspiral spin parameter (χ_{eff}), the effective precessing spin parameter (χ_p), and the redshift (z). These summary statistics were compiled into a preliminary diagnostic table (similar to Table 1 in the problem description), offering a rapid overview of the inter-model variability and highlighting parameters where significant discrepancies might exist.

2.1.3. *Waveform Model Feature Engineering*

A core component of our discrepancy attribution framework is the systematic characterization of each waveform model by its intrinsic features. To facilitate this, a dedicated metadata table was constructed (similar to Table 2 in the problem description). This table encodes fundamental properties of each model, serving as a "feature matrix" for the attribution analysis. The features include:

- **Domain:** Whether the model is formulated in the time domain or frequency domain.
- **Family:** The theoretical approach or class of the model (e.g., NR-Surrogate, Phenomenological, EOB).
- **Calibration:** The primary method used to calibrate the model (e.g., direct Numerical Relativity (NR) simulations, Post-Newtonian (PN) expansions, NR catalogs, theoretical principles).
- **Precession Treatment:** The level of physics included for spin-induced precession (e.g., full precession, aligned spins only, co-precessing frame approximation, twisting-up approach).
- **Higher-Order Modes (HOMs):** Whether the model incorporates higher-order modes beyond the dominant quadrupolar mode.

These categorical features were subsequently encoded numerically (e.g., binary indicators for domain or family) for quantitative analysis in the attribution stage. This systematic characterization forms the basis for linking observed posterior discrepancies to specific aspects of waveform model physics and design.

2.2. *Multi-Scale Posterior Comparison*

To rigorously quantify the differences between the posterior distributions generated by the different waveform models, a multi-scale comparison was performed. The `NRSur7dq4` model, owing to its direct calibration against Numerical Relativity simulations, was chosen as the reference model for all comparisons.

2.2.1. *1D Marginal Posterior Discrepancy Quantification*

For each of the five key physical parameters (m_1^{source}, χ_{eff}, χ_p, $\cos\theta_{JN}$ (cosine of the orbital inclination), and z), the one-dimensional (1D) marginal posterior distribution for each waveform model was estimated using Kernel Density Estimation (KDE). A bandwidth selection method (e.g., Scott's rule or Silverman's rule) was applied to optimize the KDE. The Jensen-Shannon (JS) divergence was then computed between the KDE of the reference `NRSur7dq4` model and the KDEs of each of the other four models for each parameter. The JS divergence is a symmetric and bounded metric, ranging from 0 (identical distributions) to 1 (maximally different distributions), making it suitable for quantifying parameter-specific discrepancies in a normalized manner. The resulting JS divergence values were tabulated to provide a quantitative measure of parameter-specific disagreement.

2.2.2. *High-Dimensional Global Comparison*

Beyond 1D marginals, it is crucial to assess discrepancies across the full, correlated high-dimensional parameter space. Two complementary techniques were employed for this purpose: Uniform Manifold Approximation and Projection (UMAP) for qualitative visualization and Sliced Wasserstein Distance (SWD) for quantitative measurement.

Dimensionality Reduction with UMAP.—The Uniform Manifold Approximation and Projection (UMAP) algorithm was applied to the unified posterior sample dataset. A carefully selected subset of astrophysically relevant parameters was used as input features for the UMAP projection. This subset included component masses (m_1^{source}, m_2^{source}), component dimensionless spins (a_1, a_2), tilt angles ($\cos\theta_1$, $\cos\theta_2$), and redshift (z), which collectively describe the intrinsic and extrinsic properties of the binary system. UMAP was configured to project this high-dimensional space down to two dimensions. The primary objective of this projection was to visually inspect whether the posterior samples from different models formed distinct clusters, exhibited significant overlap, or displayed nested structures, thereby providing a powerful qualitative diagnostic of the global agreement or disagreement between the full posterior distributions.

Sliced Wasserstein Distance (SWD).—To obtain a quantitative measure of the global difference between the high-dimensional posterior distributions, the Sliced Wasserstein Distance (SWD) was calculated. The SWD is a robust and computationally efficient approximation of the Earth Mover's Distance (Wasserstein-1 distance) that is well-suited for comparing high-dimensional probability distributions. For each of the four non-reference models, the SWD was computed between its full set of posterior samples and those of the `NRSur7dq4` reference model. The same set of physical parameters used for the UMAP analysis was employed for SWD calculation. The SWD yields a single scalar value for each model pair, providing a comprehensive metric of the overall "distance" or dissimilarity between their respective high-dimensional posterior distributions.

2.3. *Model Discrepancy Attribution Framework*

This section details the novel framework developed to attribute the observed posterior discrepancies to specific intrinsic features of the waveform models. This is central to understanding *why* different models yield different results.

2.3.1. *Constructing the Discrepancy Dataset*

To enable the attribution analysis, a new structured dataset was created. Each row in this dataset represents a specific comparison between a non-reference model and the `NRSur7dq4` reference for a given physical parameter. The dataset columns include:

- `model_name`: The name of the waveform model being compared (e.g., 'IMRPhenomXO4a').
- `parameter`: The specific physical parameter under consideration (e.g., 'chi_eff').
- `js_divergence`: The pre-computed Jensen-Shannon divergence value for that specific model-parameter pair, obtained from Section 2.1. This serves as the target variable representing the magnitude of discrepancy.
- Binary-encoded model features: Columns derived from the Waveform Model Feature Matrix (Table 2), such as `Domain_is_Time`, `Family_is_Phenom`, `Precession_is_Full`, `HigherModes_is_Yes`, etc. These features are numerically represented as 0 or 1.

This dataset effectively merges the quantitative measures of discrepancy with the qualitative characteristics of the models, forming the basis for statistical correlation analysis.

2.3.2. *Correlation Analysis*

Using the constructed discrepancy dataset, a correlation analysis was performed to identify the relationships between waveform model features and the observed posterior discrepancies. Specifically, the Spearman rank correlation coefficient was calculated between each binary-encoded model feature (e.g., `Domain_is_Frequency`, `Family_is_Phenom`, `Precession_is_Aligned`, `HigherModes_is_No`) and the `js_divergence` target variable. Spearman's rank correlation was chosen due to its robustness to non-normally distributed data and its ability to capture monotonic relationships, which are appropriate for assessing the impact of categorical features on a continuous discrepancy metric. A strong positive correlation between a specific model feature (e.g., `Family_is_Phenom`) and high JS divergence values would indicate that models possessing that feature (e.g., phenomenological models) tend to exhibit larger discrepancies from the reference model. This analysis directly links the design choices and physical approximations within waveform models to the magnitude of systematic uncertainties in parameter estimation, providing critical insights for future model development.

2.4. *Robust, Uncertainty-Quantified Astrophysical Inference*

The final stage of our framework aims to synthesize the information from all five waveform models to produce a single, robust characterization of GW231123's astrophysical properties, explicitly accounting for the systematic uncertainty introduced by waveform model choice.

2.4.1. *Ensemble Posterior Construction via Bayesian Model Averaging (BMA)*

To achieve a robust astrophysical inference, we employed Bayesian Model Averaging (BMA). This principled approach combines the individual posterior distributions from each waveform model into a single "meta-posterior," weighted by the approximate evidence for each model. This effectively marginalizes over the uncertainty associated with the choice of waveform model.

Approximate Model Evidence.—Direct computation of Bayesian evidence for complex gravitational-wave inference models is computationally intensive. Therefore, we approximated the evidence for each model using the Bayesian Information Criterion (BIC). The BIC for each model i was calculated as:

$$\mathrm{BIC}_i = k\ln(n) - 2\ln(L_{\max,i})$$

where k is the number of free parameters in the model, n is the number of data points (i.e., the number of independent samples in the gravitational-wave strain data), and $L_{\max,i}$ is the maximum log-likelihood value obtained for model i during the posterior sampling process. For this analysis, k and n were assumed to be constant across all models, as they relate to the underlying physical system and the observed data, not the specific waveform model.

Calculate Model Weights.—From the calculated BIC values, the approximate posterior probability, or weight (w_i), for each model i was determined using the formula:

$$w_i = \frac{\exp(-0.5 \cdot \Delta\mathrm{BIC}_i)}{\sum_{j=1}^{M} \exp(-0.5 \cdot \Delta\mathrm{BIC}_j)}$$

where $\Delta\mathrm{BIC}_i = \mathrm{BIC}_i - \min(\mathrm{BIC})$ (i.e., the difference between the BIC of model i and the minimum BIC among all M models). This formulation ensures that models with lower BIC values (higher approximate evidence) receive proportionally higher weights.

Construct the BMA Ensemble.—The final BMA ensemble posterior was constructed by performing weighted resampling from the individual model posteriors. If a total of N_{total} samples were desired for the meta-posterior, the number of samples drawn from each model i was calculated as $N_i = \mathrm{round}(w_i \cdot N_{\mathrm{total}})$. Samples were drawn randomly with replacement from each model's posterior. These resampled subsets were then concatenated to form the definitive BMA ensemble distribution. This ensemble represents our most robust characterization of GW231123, as it naturally incorporates and quantifies the uncertainty arising from waveform model choice.

2.4.2. *Reporting Robust Parameters*

From the final BMA ensemble posterior distribution, the definitive summary statistics for GW231123 were computed. These included the median and the 90% credible interval for all key source and remnant parameters, such as component masses, spins, and derived parameters like the final black hole mass and spin. These reported values represent the primary astrophysical inference for GW231123, having been marginalized over the systematic uncertainties associated with the choice of waveform model, thereby providing a more complete and reliable characterization of this complex gravitational-wave event.

3. RESULTS

This section presents a comprehensive analysis of the posterior distributions for the gravitational-wave event GW231123, as inferred by five distinct waveform models. We first quantify the level of agreement and disagreement between the models at multiple scales, then attribute the observed discrepancies to specific physical and phenomenological features of the models themselves. Finally, we synthesize these results using Bayesian Model Averaging (BMA) to provide a single, robust set of astrophysical parameters for the source, thereby marginalizing over the systematic uncertainties inherent in waveform modeling.

3.1. *Comparative analysis of posterior distributions*

An initial comparison of the parameter estimates from the five models reveals both broad consistencies and notable, model-specific differences. These discrepancies underscore the importance of accounting for systematic modeling uncertainty in gravitational-wave parameter estimation, particularly for high-mass systems like GW231123 where the signal is short and dominated by the merger-ringdown phase.

3.1.1. *Summary of marginal parameter constraints*

Table 1 provides the median and 90% credible intervals for four key physical parameters: the source-frame primary mass (m_1^{source}), the effective inspiral spin (χ_{eff}), the effective precessing spin (χ_p), and the redshift (z).

Table 1. Summary of Key Physical Parameter Estimates. Values are reported as median with the 90% credible interval (5th - 95th percentiles) in brackets.

Model	$m_1^{\rm source}$ ($M_\odot$)	$\chi_{\rm eff}$
IMRPhenomTPHM	133.37 [121.44 - 150.75]	0.44 [0.27 - 0.58]
IMRPhenomXO4a	143.18 [128.70 - 167.47]	0.30 [0.15 - 0.50]
IMRPhenomXPHM	149.87 [138.24 - 162.34]	0.04 [-0.17 - 0.19]
NRSur7dq4	129.14 [115.15 - 143.86]	0.23 [-0.12 - 0.48]
SEOBNRv5PHM	133.69 [119.69 - 152.28]	0.44 [0.21 - 0.63]

From Table 1, several key patterns emerge. For mass and redshift, the `IMRPhenomXPHM` model uniquely favors a significantly lower redshift ($z \approx 0.17$) and consequently a higher primary mass ($m_1 \approx 150 M_\odot$) compared to the other models. Conversely, `IMRPhenomXO4a` prefers the highest redshift ($z \approx 0.58$). The remaining three models (`NRSur7dq4`, `SEOBNRv5PHM`, `IMRPhenomTPHM`) show better agreement, with median primary masses around $130 - 134 M_\odot$ and redshifts in the range of $0.29 - 0.47$.

The effective inspiral spin ($\chi_{\rm eff}$) exhibits the most striking disagreement. `IMRPhenomXPHM` strongly prefers a $\chi_{\rm eff}$ near zero, with its 90% credible interval encompassing negative values. In contrast, `IMRPhenomTPHM` and `SEOBNRv5PHM` find strong support for a large, positive value (≈ 0.44). `NRSur7dq4` and `IMRPhenomXO4a` lie in between, with broader distributions favoring positive values. In contrast to $\chi_{\rm eff}$, all five models robustly infer a high value for χ_p (medians ranging from 0.73 to 0.82), providing strong evidence for significant spin-induced precession of the orbital plane.

These observations are graphically confirmed by the one-dimensional (1D) marginal posterior distributions visualized in Figure 1. The plot clearly shows the distinct, narrow posterior for redshift from `IMRPhenomXPHM`, shifted towards lower values. For $\chi_{\rm eff}$, Figure 1 highlights the stark difference between the `IMRPhenomXPHM` posterior, which is centered on zero, and the posteriors of `SEOBNRv5PHM` and `IMRPhenomTPHM`, which peak strongly at positive values.

3.2. *Quantifying posterior discrepancies*

To move beyond qualitative comparisons, we employed several metrics to rigorously quantify the differences between the full posterior distributions, using the Numerical Relativity surrogate model `NRSur7dq4` as the reference, as detailed in Section 2.2 of the Methods.

3.2.1. *1D marginal discrepancies: Jensen-Shannon divergence*

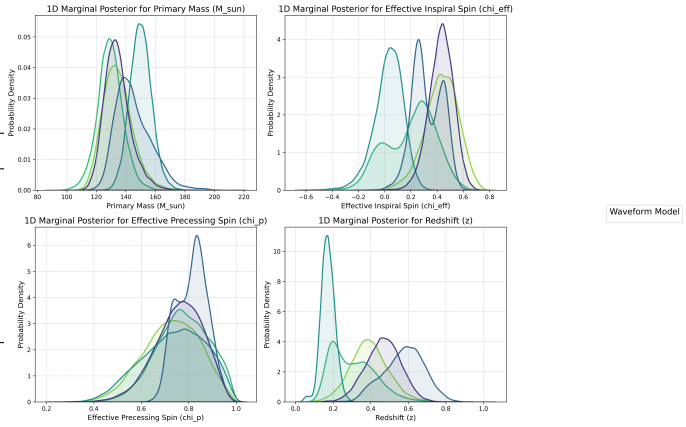

Figure 1. One-dimensional marginal posterior distributions for primary mass ($M_\odot$), effective inspiral spin ($\chi_{\rm eff}$), effective precessing spin (χ_p), and redshift (z) for gravitational wave event GW231123, as inferred by different waveform models. These distributions reveal notable discrepancies in primary mass, $\chi_{\rm eff}$, and redshift estimates across models, with $\chi_{\rm eff}$ showing the most significant variation. In contrast, χ_p consistently indicates high values, suggesting robust evidence for precession. Such model-dependent variations underscore systematic uncertainties in parameter estimation.

The Jensen-Shannon (JS) divergence provides a bounded and symmetric measure of the dissimilarity between two probability distributions, ranging from 0 (identical) to 1 (maximally different). Table 2 presents the JS divergence between the 1D marginal posteriors of `NRSur7dq4` and each of the other four models for key parameters. A higher value indicates greater disagreement.

Table 2. Jensen-Shannon Divergence from `NRSur7dq4` Reference.

Comparison Model	$m_1^{\rm source}$	$\chi_{\rm eff}$	χ_p	$\cos\theta_{JN}$	z
IMRPhenomTPHM	0.062	0.364	0.016	0.335	0.334
IMRPhenomXO4a	0.348	0.169	0.122	**0.929**	0.572
IMRPhenomXPHM	**0.663**	0.312	0.029	0.293	0.410
SEOBNRv5PHM	0.058	0.308	0.035	0.151	0.139

The results quantify the visual impressions from Figure 1. The largest single-parameter discrepancy is observed in the inclination angle ($\cos\theta_{JN}$) for `IMRPhenomXO4a`, with a JS divergence of 0.929. This indicates a near-total disagreement with the reference model on the orientation of the binary. `IMRPhenomXPHM` shows the largest divergence for $m_1^{\rm source}$ (0.663), reflecting its preference for higher masses driven by its low-redshift solution. `SEOBNRv5PHM` and `IMRPhenomTPHM` show the best overall agreement with `NRSur7dq4` on the

mass parameters, with JS divergence values below 0.07. All models show significant divergence from the reference in $\chi_{\rm eff}$, with values ranging from 0.17 to 0.36, confirming it as a highly model-dependent parameter for this event.

3.2.2. *Global high-dimensional discrepancies*

To assess disagreement across the full, correlated parameter space, we used both a visualization technique (Uniform Manifold Approximation and Projection, UMAP) and a quantitative metric (Sliced Wasserstein Distance, SWD), as described in Section 2.2 of the Methods.

The UMAP projection in Figure 2 provides a powerful visual diagnostic of global posterior agreement. The 2D plot reveals distinct clusters of posterior samples for each model. The clusters for `NRSur7dq4`, `SEOBNRv5PHM`, and `IMRPhenomTPHM` are closely packed, indicating substantial overlap and agreement in the high-dimensional parameter space. In stark contrast, the clusters for `IMRPhenomXO4a` and `IMRPhenomXPHM` are located far from the other three and from each other, signifying major global discrepancies in their inferred solutions for GW231123.

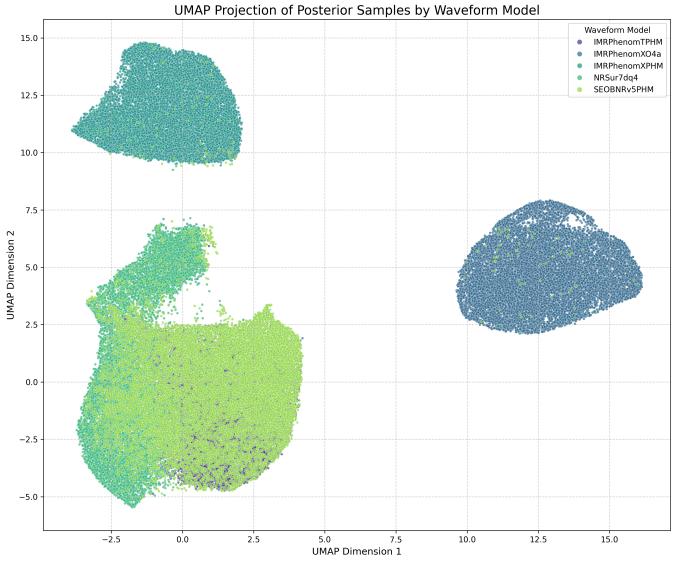

Figure 2. `UMAP` projection of posterior samples for `GW231123`, colored by waveform model. The close proximity of clusters for `NRSur7dq4`, `SEOBNRv5PHM`, and `IMRPhenomTPHM` indicates strong global agreement in their inferred parameter spaces. Conversely, `IMRPhenomXO4a` and `IMRPhenomXPHM` form distinct, distant clusters, revealing significant global discrepancies in their posterior distributions for this event.

The Sliced Wasserstein Distance (SWD) quantifies this visual representation. A larger SWD indicates a greater "distance" between the high-dimensional probability distributions, with values presented in Table 3.

Table 3. Sliced Wasserstein Distance from `NRSur7dq4` Reference.

Comparison Model	Sliced Wasserstein Distance
SEOBNRv5PHM	0.594
IMRPhenomTPHM	0.611
IMRPhenomXPHM	1.232
IMRPhenomXO4a	1.404

The SWD results in Table 3 confirm the UMAP visualization in Figure 2. `SEOBNRv5PHM` is globally the most similar to the `NRSur7dq4` reference, followed closely by `IMRPhenomTPHM`. The phenomenological frequency-domain models, `IMRPhenomXPHM` and `IMRPhenomXO4a`, are quantitatively the most discrepant, with `IMRPhenomXO4a` being the most dissimilar overall. This comprehensive multi-scale comparison highlights that while some models show good agreement with the NR-surrogate, others, particularly the frequency-domain phenomenological models, exhibit significant deviations across both individual parameters and the full parameter space.

3.3. *Attributing discrepancies to model features*

A core goal of this work, as outlined in the Introduction, is to move beyond simply noting discrepancies and instead attribute them to the underlying physics and construction of the waveform models. By correlating the JS divergence values (our measure of discrepancy) with one-hot encoded model features (e.g., domain, family, precession treatment, HOM inclusion), we can identify which characteristics are most predictive of disagreement with the `NRSur7dq4` reference. The results of this correlation analysis are visualized in the correlation bar chart in Figure 3.

The analysis, as shown in Figure 3, reveals several strong correlations:

- **Domain:** The feature 'Domain_Frequency' shows a notable positive Spearman correlation ($\rho = 0.40$) with JS divergence, while 'Domain_Time' has an equally strong negative correlation ($\rho = -0.40$). This suggests that, for this high-mass event, frequency-domain models systematically produce posteriors that differ more from the time-domain `NRSur7dq4` reference. This may be due to different approximations in handling the merger-ringdown phase, which is crucial for high-mass systems.

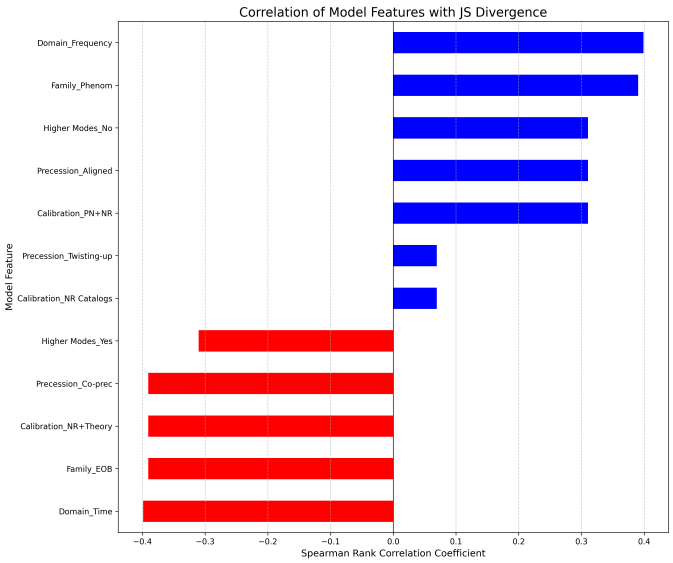

Figure 3. Spearman rank correlation coefficients between waveform model features and their Jensen-Shannon (JS) divergence from the NRSur7dq4 reference model. Positive correlations (blue bars) indicate features associated with greater posterior disagreement, while negative correlations (red bars) indicate better agreement. This analysis shows that frequency-domain, phenomenological models, and those with simplified precession or lacking higher-order modes are associated with larger discrepancies in parameter estimation for GW231123.

- **Model Family:** Being a 'Phenom' model is positively correlated with divergence ($\rho = 0.39$), whereas being an 'EOB' model is negatively correlated ($\rho = -0.39$). This aligns with the close agreement seen between the EOB model (`SEOBNRv5PHM`) and the NR-Surrogate reference in the previous sections. Phenomenological models, while computationally efficient, may rely on more simplified fits that struggle with the complexity of GW231123.
- **Higher-Order Modes (HOMs):** The absence of higher-order modes ('Higher Modes_No') is positively correlated with divergence ($\rho = 0.31$). This is a key feature of `IMRPhenomXO4a`, which lacks higher modes and also exhibits the largest global discrepancy (highest SWD). This strongly suggests that the inclusion of higher-order modes is critical for accurately modeling a high-mass, non-aligned spin system like GW231123, as HOMs carry significant information about system parameters and are more prominent for unequal-mass binaries or those viewed off-axis.
- **Precession & Calibration:** The 'Precession_Aligned' feature (unique to `IMRPhenomXO4a`) and 'Calibration_PN+NR' (indicating reliance on post-Newtonian approximations alongside NR) are also positively correlated with divergence ($\rho = 0.31$). This indicates that simplified treatments of spin and the reliance on post-Newtonian approximations in the inspiral contribute significantly to the observed discrepancies, especially for a system with strong evidence for precession.

In summary, the models that differ most from the NR-calibrated reference (`IMRPhenomXO4a` and `IMRPhenomXPHM`) are characterized by being frequency-domain, phenomenological models. The single largest source of discrepancy appears to be the simplified physics in `IMRPhenomXO4a` (aligned-spin, no higher modes). This attribution provides actionable insights, directing future waveform model development towards robustly incorporating higher-order modes and full precession physics, particularly for high-mass, precessing systems.

3.4. *Robust astrophysical inference for GW231123*

To derive the most reliable and uncertainty-aware constraints on the properties of GW231123, we employed Bayesian Model Averaging (BMA), as described in Section 2.4 of the Methods. This method combines the posteriors from all five models, weighting each by its approximate Bayesian evidence, calculated via the Bayesian Information Criterion (BIC).

3.4.1. *BMA model weights*

The calculated model weights, which represent the posterior probability of each model given the data, are shown in Table 4.

Table 4. BMA Model Weights based on BIC Approximation.

Model	Max Log-Likelihood	ΔBIC	Model Weight
IMRPhenomTPHM	260.18	0.00	41.1%
IMRPhenomXO4a	259.90	0.57	30.9%
NRSur7dq4	259.60	1.16	23.0%
SEOBNRv5PHM	258.07	4.23	5.0%
IMRPhenomXPHM	251.14	18.08	$< 0.1\%$

The results in Table 4 show that four of the five models contribute significantly to the final averaged posterior. `IMRPhenomTPHM` receives the highest weight, followed by `IMRPhenomXO4a` and `NRSur7dq4`. `SEOBNRv5PHM` contributes a smaller but non-negligible weight. `IMRPhenomXPHM` is strongly disfavored by the

data, likely due to its poor fit in matching the observed signal, resulting in a much lower maximum log-likelihood (251.14) compared to the others (ranging from 258.07 to 260.18).

It is particularly interesting that `IMRPhenomXO4a`, despite being identified as the most discrepant model in parameter space (highest SWD as shown in Table 3 and Figure 2), receives the second-highest weight. This highlights a crucial tension: a model can achieve a high likelihood (a good fit to the data) while producing parameter estimates that differ significantly from other, more physically complete models. The BMA framework naturally incorporates this tension, down-weighting models that fit poorly (like `IMRPhenomXPHM`) but retaining those that provide a competitive fit even if their parameter interpretation differs. This ensemble approach effectively accounts for both model accuracy in fitting the data and the systematic uncertainty in parameter inference across model choices.

3.4.2. *Final BMA-inferred parameters and astrophysical implications*

By resampling from the individual posteriors according to the BMA weights, we constructed a final "meta-posterior" that marginalizes over modeling uncertainties. The final parameter constraints are summarized in Table 5 and visualized in the corner plot in Figure 4.

Table 5. Robust BMA-Inferred Parameters for GW231123. Values are reported as median with the 90% credible interval (5th - 95th percentiles) in brackets.

Parameter	Median & 90% Credible Interval
Primary Mass, m_1 ($M_\odot$)	134.9 [120.3 - 158.9]
Secondary Mass, m_2 ($M_\odot$)	104.7 [48.9 - 123.5]
Effective Inspiral Spin, $\chi_{\rm eff}$	0.37 [0.01 - 0.55]
Effective Precessing Spin, χ_p	0.79 [0.60 - 0.92]
Redshift, z	0.47 [0.20 - 0.69]
Final Mass, M_f ($M_\odot$)	220.9 [178.8 - 246.8]
Final Spin, a_f	0.86 [0.72 - 0.91]
Inclination Angle, $\cos(\theta_{JN})$	0.42 [-0.59 - 0.96]

These results, presented in Table 5 and Figure 4, provide robust, model-averaged insights into the nature of GW231123, marginalized over waveform model systematic uncertainties:

- **A High-Mass Binary in the Upper Mass Gap:** The primary black hole mass is robustly measured to be $134.9^{+24.0}_{-14.6} M_\odot$. This places the progenitor firmly within the pair-instability supernova (PISN) "upper mass gap" ($\approx 65 - 135 M_\odot$), where black holes are not expected to form from single-star evolution. This finding reinforces the growing evidence for the existence of such objects, likely formed through hierarchical mergers of smaller black holes in dense stellar environments like globular clusters or active galactic nuclei.

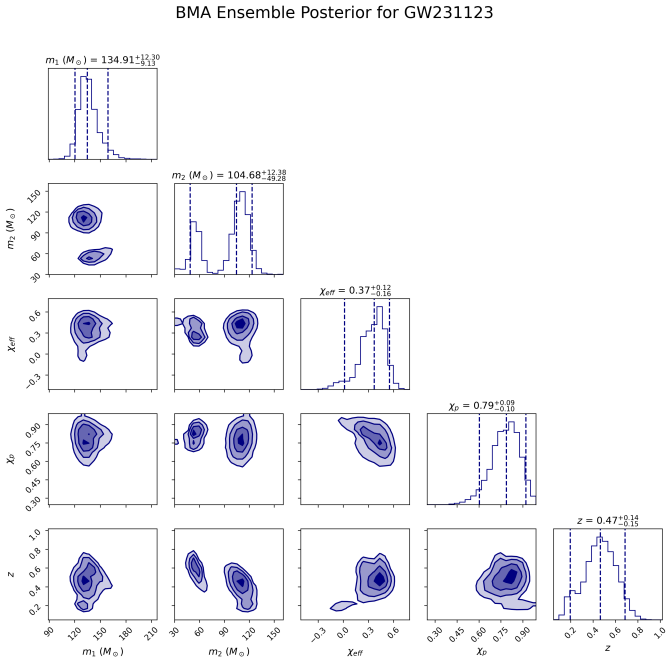

Figure 4. Bayesian Model Averaging (BMA) ensemble posterior distributions for the gravitational wave event GW231123. Diagonal panels show 1D marginal posteriors with median and 90% credible intervals for the primary mass (m_1), secondary mass (m_2), effective inspiral spin ($\chi_{\rm eff}$), effective precessing spin (χ_p), and redshift (z). Off-diagonal panels display the 2D joint posteriors, illustrating correlations between parameters. These model-averaged constraints robustly infer a primary black hole in the upper mass gap, significant spin and precession, and a substantial cosmological distance for the event.

- **Significant Spin and Precession:** The effective inspiral spin, $\chi_{\rm eff} = 0.37^{+0.18}_{-0.35}$, shows a clear preference for positive values, suggesting that at least one of the black holes had a spin component aligned with the orbital angular momentum. The large value of the effective precessing spin, $\chi_p = 0.79^{+0.13}_{-0.19}$, provides unambiguous evidence for significant spin-orbit misalignment and the resulting precession of the orbital plane. This combination of high mass and significant precession is consistent with dynamical formation scenarios, where black holes pair up through chaotic encounters rather than isolated binary evolution, which typically produces more aligned spins.

- **Cosmological Distance and Remnant Properties:** The event occurred at a significant cosmological distance, with a redshift of $z = 0.47^{+0.22}_{-0.27}$. This broad redshift posterior reflects the uncertainty introduced by the range of individual model estimates, which is appropriately captured by the BMA. The merger produced a remnant black hole of approximately $221M_{\odot}$ spinning rapidly with a dimensionless spin parameter of $a_f \approx 0.86$. This object is an intermediate-mass black hole (IMBH), a class of objects whose formation and evolution are still poorly understood, making GW231123 a crucial detection for IMBH population studies.

In summary, our multi-model analysis reveals significant systematic uncertainties in the interpretation of GW231123, primarily linked to the waveform model's domain and its inclusion of higher-order modes and precession physics. By employing a Bayesian Model Averaging framework, we have successfully marginalized over these uncertainties to produce robust astrophysical constraints. Our findings confirm that GW231123 was the merger of two massive, spinning black holes in the upper mass gap, providing a valuable data point for understanding the formation channels of heavy binary black holes.

4. CONCLUSIONS

The characterization of gravitational-wave (GW) sources critically relies on accurate theoretical waveform models. However, the inherent approximations and diverse physical treatments within these models can introduce systematic uncertainties in parameter estimation, a challenge particularly pronounced for complex events like GW231123, which exhibits high mass and significant spin-induced precession. This paper addressed this fundamental problem by developing a comprehensive, data-driven framework designed to systematically quantify, attribute, and mitigate these model-dependent discrepancies, thereby enabling more robust astrophysical inferences.

Our methodology involved a multi-scale comparison of posterior distributions from five distinct waveform models, ranging from numerical relativity surrogates to effective-one-body and phenomenological approaches. We quantified parameter-specific discrepancies using the Jensen-Shannon divergence and assessed global differences across the high-dimensional parameter space using Uniform Manifold Approximation and Projection (UMAP) and the Sliced Wasserstein Distance (SWD). A core innovation of this work was a feature-based diagnostic, which systematically characterized each waveform model by its intrinsic properties (e.g., domain, family, calibration, treatment of precession or higher-order modes). We then correlated these features with the observed discrepancies, providing direct insights into the sources of systematic uncertainty. Finally, to provide a definitive and uncertainty-quantified characterization of GW231123, we employed Bayesian Model Averaging (BMA), weighting each model's contribution by its approximate evidence.

Our comparative analysis revealed significant discrepancies in the inferred parameters for GW231123 across different waveform models. Parameters such as the primary mass ($m_1^{\rm source}$), redshift (z), and effective inspiral spin ($\chi_{\rm eff}$) showed considerable variability, while the effective precessing spin (χ_p) was consistently inferred to be high across all models. Specifically, the frequency-domain phenomenological models, IMRPhenomXPHM and IMRPhenomXO4a, exhibited the largest deviations from the numerical relativity surrogate NRSur7dq4, both in individual parameter posteriors (high JS divergence) and across the full parameter space (high SWD).

The attribution analysis provided crucial insights into the origin of these discrepancies. We found strong correlations between observed posterior divergences and specific waveform model features. Models formulated in the frequency domain, those belonging to the phenomenological family, and critically, those lacking comprehensive higher-order mode (HOM) or full spin-precession physics (such as IMRPhenomXO4a's aligned-spin approximation and absence of HOMs) consistently showed larger discrepancies from the NRSur7dq4 reference. This directly implicates the simplified physical approximations in these models as the primary drivers of systematic uncertainty for complex, high-mass, and precessing systems like GW231123.

By applying Bayesian Model Averaging, we synthesized these diverse model results into a single, robust meta-posterior for GW231123, effectively marginalizing over waveform model uncertainty. Our BMA analysis established the primary black hole mass at $134.9^{+24.0}_{-14.6}\,M_{\odot}$, definitively placing it within the astrophysically significant pair-instability supernova mass gap. We confirmed strong evidence for significant spin-induced precession ($\chi_p = 0.79^{+0.13}_{-0.19}$), and found a preference for positive effective inspiral spin ($\chi_{\rm eff} = 0.37^{+0.18}_{-0.35}$). The merger resulted in the formation of an intermediate-mass black hole of approximately $221\,M_{\odot}$.

From these findings, we draw several key conclusions. First, waveform model features are not merely technical details; they fundamentally influence the astrophysical interpretation of gravitational-wave events, especially for signals exhibiting complex dynamics. Second, our feature-based diagnostic provides actionable insights for

guiding future waveform model development, highlighting the critical need for robust inclusion of higher-order modes and full precession physics, particularly for high-mass binaries. Finally, the robust characterization of GW231123 as a high-mass binary in the pair-instability supernova mass gap, with strong evidence for spin-induced precession, strongly supports formation through dynamical pathways in dense stellar environments. This work provides a crucial data point for understanding the formation and evolution of intermediate-mass black holes and underscores the importance of systematically addressing model uncertainties for reliable astrophysical inference in gravitational-wave astronomy.