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

Quantifying and Attributing Waveform Model-Dependent Systematics in GW231123: A Multi-Scale Posterior Analysis

Denario[1]

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

ABSTRACT

Gravitational-wave parameter estimation inherently faces systematic uncertainties due to the approximations within waveform models. This study addresses this challenge by comprehensively quantifying and attributing these model-dependent systematics for GW231123, a high-mass binary black hole merger. We analyzed posterior samples from five distinct waveform models (NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM, IMRPhenomXO4a, IMRPhenomXPHM). Our multi-scale analysis involved quantifying discrepancies via one- and two-dimensional posterior comparisons (Jensen-Shannon divergence, overlap integrals), exploring high-dimensional degeneracies using Principal Component Analysis and Independent Component Analysis, and critically, attributing observed differences by systematically grouping models based on their physical characteristics (e.g., domain, calibration, precession treatment). Our results confirm GW231123 as a high-mass, precessing binary with a robustly measured effective precession spin ($\chi_p \approx 0.77$) and final spin ($a_f \approx 0.84$). However, we reveal significant systematic uncertainties in other key parameters, including component masses, mass ratio, effective inspiral spin ($\chi_{\rm eff}$), and redshift. For instance, secondary mass estimates vary twofold across models, and $\chi_{\rm eff}$ spans from near-zero to significant positive alignment, precluding a definitive conclusion on spin alignment. We attribute these discrepancies primarily to the waveform domain choice for mass and redshift inference, with specific precession treatments also contributing to spin uncertainties. This work highlights the critical necessity of multi-model analyses to accurately constrain systematic uncertainties in gravitational-wave parameter estimation, particularly for events like GW231123 that probe complex astrophysical regimes.

Keywords: Principal component analysis, Stellar mass black holes, Gravitational wave sources, Gravitational waves, Posterior distribution

1. INTRODUCTION

The advent of gravitational-wave (GW) astronomy, initiated by the direct detection of binary black hole mergers, has ushered in an unprecedented era for exploring the universe's most extreme phenomena. A cornerstone of extracting astrophysical insights from these transient signals is the precise estimation of source parameters, a process fundamentally reliant on comparing observed data with theoretical waveform models. These models, which describe the gravitational radiation emitted by coalescing compact binaries, are developed through a combination of numerical relativity simulations, post-Newtonian expansions, and effective-one-body (EOB) frameworks. Despite their remarkable sophistication, all current waveform models inherently incorporate approximations due to the immense computational demands of full general relativistic calculations and the complexities of the underlying physics. Consequently, the inferred astrophysical parameters are not only subject to statistical uncertainties arising from detector noise but also to systematic uncertainties originating from these model-dependent approximations.

Disentangling these systematic uncertainties from the statistical fluctuations poses a formidable challenge. This difficulty is particularly pronounced for GW events that probe complex regions of the parameter space, such as high-mass binary black hole mergers, systems exhibiting significant spin precession, or those where higher-order modes of gravitational radiation contribute substantially to the signal. Such events push the boundaries of current waveform modeling capabilities, meaning that the choice of waveform model can become a significant determinant of the inferred astrophysical properties. If left unaddressed, these systematic biases risk leading to inaccurate astrophysical conclusions, misinterpreta-

tions of population properties, and hindering the precise characterization of the universe's most energetic cosmic events.

This paper directly addresses this critical challenge by providing a comprehensive quantification and, crucially, an attribution of waveform model-dependent systematic uncertainties for GW231123, a recently observed high-mass binary black hole merger. Our analysis leverages a multi-scale approach to scrutinize posterior samples derived from five distinct gravitational-wave waveform models: NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM, IMRPhenomXO4a, and IMRPhenomXPHM. This carefully selected set encompasses a diverse range of theoretical formulations, calibration strategies, and domain choices (time-domain versus frequency-domain), thereby providing a rich landscape for thoroughly investigating model-dependent effects.

Our methodology is systematically designed to unravel and explain these systematics. We begin by quantifying discrepancies across models using both one- and two-dimensional marginal posterior comparisons, employing robust metrics such as the Jensen-Shannon divergence and overlap integrals to precisely measure distributional differences. To explore the more intricate high-dimensional degeneracies that are often obscured in lower-dimensional projections, we then employ advanced dimensionality reduction techniques, specifically Principal Component Analysis (PCA) and Independent Component Analysis (ICA). The core innovation and central analytical thrust of this work lie in our systematic attribution strategy: we group waveform models based on their fundamental physical characteristics—such as their underlying waveform domain, calibration basis (e.g., Numerical Relativity or Effective-One-Body based versus Phenomenological), and specific treatment of spin precession—and then quantify the inter-group differences in parameter inference. This approach allows us to establish a direct, concrete link between specific modeling choices and the observed variations in the inferred astrophysical parameters.

By rigorously quantifying and attributing these systematic uncertainties, this study aims to achieve two primary objectives for GW231123. First, we will identify parameters for which the astrophysical inference is robust and consistent across all models, thereby providing the most reliable constraints on this high-mass merger. Second, for parameters exhibiting significant model dependence, we will precisely quantify the range of systematic uncertainty introduced by waveform modeling, presenting a complete and nuanced picture of GW231123's properties. Ultimately, this work underscores the indispensable necessity of conducting multi-model analyses to accurately constrain the full uncertainty budget in gravitational-wave parameter estimation, particularly for events like GW231123 that probe complex and challenging astrophysical regimes.

2. METHODS

2.1. *Data Aggregation and Exploratory Analysis*

Our analysis commenced with the aggregation and preliminary exploration of posterior samples for the gravitational-wave event GW231123. As highlighted in the Introduction, this study leverages posterior samples derived from five distinct gravitational-wave waveform models: NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM, IMRPhenomXO4a, and IMRPhenomXPHM.

2.1.1. *Data Loading and Merging*

The posterior samples for GW231123, generated by independent parameter estimation pipelines for each waveform model, were provided as individual CSV files. These files were loaded into a unified data analysis environment. To maintain the origin of each sample throughout the comparative analysis, a new categorical column, 'model', was added to each dataset, explicitly identifying the waveform model from which the samples originated. Subsequently, all individual model datasets were concatenated into a single master dataset, enabling a streamlined multi-model analysis.

2.1.2. *Model Characterization*

A fundamental aspect of this study is the attribution of systematic uncertainties to specific characteristics of the waveform models. To facilitate this, each waveform model was systematically characterized based on its theoretical underpinnings and implementation details. This characterization, summarized in Table 1, categorizes models by their waveform domain (time-domain or frequency-domain), the basis of their formulation and calibration (e.g., calibrated to Numerical Relativity (NR) simulations, derived from Effective-One-Body (EOB) theory, or purely Phenomenological), and their specific approaches to incorporating higher-order modes and spin precession. This detailed classification serves as the foundation for the model grouping strategy employed in the attribution phase of our analysis.

2.1.3. *Exploratory Data Analysis (EDA)*

Before engaging in advanced statistical comparisons, an exploratory data analysis was conducted to provide an initial overview of the consistency and variations across the model posteriors. For each of the five waveform models, summary statistics were computed for key astrophysical parameters inferred for GW231123, including the source-frame component masses (m_1, m_2),

Table 1. Characterization of Waveform Models Used in This Study for GW231123.

Model	Domain	Formulation	Key Characteristics
NRSur7dq4	Time-domain	NR-Calibrated	Surrogate model, highly accurate, calibrated to NR simulations
SEOBNRv5PHM	Time-domain	EOB	Incorporates higher modes, calibrated to NR simulations
IMRPhenomTPHM	Time-domain	Phenomenological	Time-based "twisting-up" approach for precession, improved ringdown mo
IMRPhenomXO4a	Frequency-domain	Phenomenological	Global calibration using post-Newtonian (PN) and NR, wide frequency ran
IMRPhenomXPHM	Frequency-domain	Phenomenological	Frequency-based "twisting-up" for precession, includes multipolar modes, c

effective inspiral spin ($\chi_{\rm eff}$), effective precession spin (χ_p), final remnant mass (M_f), final remnant spin (a_f), and redshift (z). Specifically, the median and the 90% credible interval (defined by the 5th and 95th percentiles of the posterior samples) were calculated for each parameter. As indicated in the Abstract, this initial exploration revealed a high degree of concordance among models for mass parameters and redshift, while substantial discrepancies were observed for spin parameters, particularly $\chi_{\rm eff}$ and χ_p. These preliminary findings underscored the necessity for the detailed quantitative and attribution analyses that followed.

2.2. *Quantitative Posterior Discrepancy Analysis*

To rigorously quantify the differences between the posterior distributions generated by the distinct waveform models, a two-pronged approach involving both one-dimensional (1D) and two-dimensional (2D) posterior comparisons was adopted. This aimed to move beyond summary statistics and capture the full distributional differences.

2.2.1. *1D Marginal Posterior Comparison*

For each individual parameter of astrophysical interest ('mass_1_source', 'mass_2_source', 'chi_eff', 'chi_p', 'final_mass_source', 'final_spin', 'redshift', and 'cos_theta_jn'), a detailed 1D comparison was performed:

1. **Kernel Density Estimation (KDE):** For each of the five waveform models, a Kernel Density Estimate was generated to represent the 1D marginal posterior probability distribution of each parameter. To ensure a fair and consistent comparison, a uniform bandwidth selection method, specifically Scott's rule, was applied across all models for a given parameter.

2. **Jensen-Shannon Divergence Calculation:** The pairwise Jensen-Shannon (JS) divergence was computed between the 1D marginal posterior distributions of all possible pairs of the five waveform models. The JS divergence is a symmetric and bounded metric ($[0, 1]$) derived from the Kullback-Leibler divergence, providing a robust and interpretable measure of the dissimilarity between two probability distributions. A JS divergence value closer to 0 indicates greater similarity, while a value closer to 1 indicates greater dissimilarity. The results for each parameter were stored as a 5x5 symmetric matrix, serving as a key quantitative measure of model disagreement in 1D parameter projections.

2.2.2. *2D Joint Posterior Comparison*

Given that astrophysical parameters often exhibit complex degeneracies, a quantitative assessment of 2D joint posterior distributions was essential. We focused on the following physically relevant parameter pairs: ('mass_1_source', 'mass_2_source'), ('chi_eff', 'chi_p'), and ('final_mass_source', 'final_spin'). For each pair:

1. **2D Kernel Density Estimation (KDE):** A 2D KDE was generated for the joint posterior distribution for each of the five waveform models, allowing for visualization and quantification of how each model constrains parameter degeneracies.

2. **Posterior Overlap Integral:** The dissimilarity between these 2D joint posteriors was quantified using the posterior overlap integral. This metric, defined as $O(P_1, P_2) = \int \min(P_1(\theta), P_2(\theta)) d\theta$, measures the common volume under the two posterior probability density functions P_1 and P_2. An overlap value closer to 1 signifies a higher degree of agreement and shared parameter space between the two joint posteriors, while a value closer to 0 indicates minimal commonality. These overlap values were compiled into a 5x5 matrix for each 2D parameter plane, providing crucial insights into how different models resolve or contribute to degeneracies in the parameter space for GW231123.

2.3. *High-Dimensional Degeneracy Analysis via Dimensionality Reduction*

To thoroughly investigate the intricate, high-dimensional degeneracies inherent in gravitational-wave parameter estimation that may not be apparent in lower-dimensional projections, we employed advanced dimensionality reduction techniques: Principal Component Analysis (PCA) and Independent Component Analysis (ICA).

2.3.1. *Data Preparation*

For dimensionality reduction, a comprehensive numerical feature matrix was constructed from the

master dataset. This matrix included the following core astrophysical parameters: 'mass_1_source', 'mass_2_source', the primary and secondary spin magnitudes (a_1, a_2), the cosines of the primary and secondary tilt angles ('cos_tilt_1', 'cos_tilt_2'), the cosine of the inclination angle ('cos_theta_jn'), the azimuthal angle between the total angular momentum and orbital angular momentum ($\phi_{\rm jl}$), and redshift (z). To ensure that parameters with larger intrinsic scales did not disproportionately influence the dimensionality reduction algorithms, the entire feature matrix (combining samples from all five models) was standardized. This involved scaling each parameter column to have a mean of 0 and a standard deviation of 1.

2.3.2. *Principal Component Analysis (PCA)*

PCA was applied to the combined, standardized feature matrix. PCA is a linear dimensionality reduction technique that identifies orthogonal axes, known as Principal Components (PCs), along which the data exhibits the maximum variance. The first few PCs typically capture the majority of the variance in the dataset, effectively revealing the dominant degeneracies and structures within the high-dimensional parameter space.

1. **PC Loadings Analysis:** The loadings (coefficients) of the first 3-4 PCs were meticulously analyzed. These loadings define each PC as a linear combination of the original astrophysical parameters, thereby revealing the physical meaning of the dominant degeneracies. For instance, a PC with high loadings on both masses and redshift would indicate a strong mass-redshift degeneracy.

2. **Projection and Comparison:** The posterior samples from *each individual waveform model* were then projected onto these identified PC axes. For each of the leading PCs, the 1D distribution of the projected samples was generated for each model. Comparing these projected distributions allowed for a direct assessment of how each waveform model constrains the primary degeneracies identified from the combined parameter space. Differences in the medians or widths of these projected distributions highlight model-dependent systematic effects along the principal directions of variance.

2.3.3. *Independent Component Analysis (ICA)*

Following PCA, Independent Component Analysis (ICA) was applied to the same combined, standardized feature matrix, utilizing the FastICA algorithm. Unlike PCA, which seeks orthogonal components that maximize variance, ICA aims to find statistically independent components. For astrophysical posteriors, which often exhibit non-Gaussian distributions and complex non-linear degeneracies, ICA can sometimes offer a more physically interpretable decomposition than PCA's orthogonal components.

1. **IC Definition Analysis:** The component definitions (represented by the unmixing matrix) were analyzed to understand the physical meaning of the statistically independent directions identified by ICA. These components can sometimes correspond more directly to underlying physical processes or source parameters than PCA components.

2. **Projection and Comparison:** Similar to PCA, the posterior samples of each model were projected onto these Independent Components (ICs). The distributions of these projected samples were then compared across models to reveal how models differ along these statistically independent directions. This approach provides a complementary perspective to PCA, potentially uncovering distinct modeling effects not evident through variance-maximizing orthogonal components.

2.4. *Attribution of Discrepancies to Model Characteristics*

The central analytical objective of this study, as articulated in the Introduction, is to attribute observed posterior discrepancies to specific physical characteristics of the waveform models. This was achieved by systematically grouping models based on their shared properties and quantifying inter-group differences.

2.4.1. *Grouping and Pooling Posteriors*

Based on the detailed model characterization presented in Table 1, three distinct grouping strategies were employed to isolate the effects of specific modeling choices:

- **Domain Grouping:** This grouping distinguishes between models implemented in the time-domain versus the frequency-domain, which can affect how different parts of the signal (e.g., inspiral, merger, ringdown) are modeled.
  - Group A (Time-domain): NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM
  - Group B (Frequency-domain): IMRPhenomXO4a, IMRPhenomXPHM

- **Calibration Grouping:** This grouping differentiates models based on their primary calibration basis, distinguishing between those more directly tied to numerical relativity or effective-one-body theory and those that are purely phenomenological.
  - Group C (NR/EOB-based): NRSur7dq4, SEOBNRv5PHM
  - Group D (Phenomenological): IMRPhenomXO4a, IMRPhenomXPHM, IMRPhenomTPHM
- **Precession Treatment Grouping:** This grouping focuses on the specific methods used to incorporate spin precession, particularly distinguishing "twisting-up" approaches from others.
  - Group E ("Twisting-up" Models): IMRPhenomXPHM, IMRPhenomTPHM
  - Group F (Other Precession Treatments): NRSur7dq4, SEOBNRv5PHM, IMRPhenomXO4a

For each defined group, the posterior samples of its constituent models were pooled to create a single, larger "group-level" posterior distribution. This pooling assumes that models within a given group share common systematic behaviors related to the defining characteristic of that group.

2.4.2. *Inter-Group Discrepancy Quantification*

To attribute observed differences, the quantitative discrepancy analyses from Section 2 were re-applied to these newly formed group-level posteriors.

1. **1D and 2D Discrepancy Metrics:** The pairwise Jensen-Shannon divergence was calculated for 1D marginal posteriors (e.g., between Group A and Group B for parameters like χ_p), and overlap integrals were computed for 2D joint posteriors (e.g., between Group C and Group D for the $(\chi_{\rm eff}, \chi_p)$ plane). A large JS divergence or a low overlap integral between groups for a specific parameter provides strong quantitative evidence that the characteristic defining those groups (e.g., waveform domain, calibration basis) is a significant driver of the observed systematic differences in parameter inference.
2. **PCA/ICA Projection Analysis:** The PCA and ICA projections from Section 3 were also analyzed at the group level. For each group, the mean and variance of the projected samples were computed along the primary PCs and ICs. A significant shift in the mean projection or a substantial difference in the variance between, for instance, the NR/EOB-based group (Group C) and the Phenomenological group (Group D) on a PC or IC primarily related to spin parameters would directly attribute that specific variance in the high-dimensional parameter space to the model's calibration basis. This systematic, quantitative comparison establishes a direct link between specific modeling choices and the observed variations in the inferred astrophysical parameters for GW231123.

2.5. *Derivation of Robust Astrophysical Inferences*

The final step of our analysis involved synthesizing all previous results to derive a comprehensive set of astrophysical conclusions for GW231123, explicitly accounting for the identified waveform model-dependent systematics. This process aimed to identify robust parameters and quantify systematic uncertainties for others.

2.5.1. *Identification of Consensus Parameters*

Parameters for which all five waveform models demonstrated high agreement were identified as "robustly constrained" for GW231123. This identification was based on the quantitative metrics from Section 2.1, specifically parameters where the pairwise Jensen-Shannon divergence between all individual models was consistently below a predefined threshold (e.g., 0.05, indicating minimal dissimilarity across all pairs) and the 2D overlap integrals were consistently high. For these robust parameters, a final combined posterior distribution was generated by pooling the samples from all five individual waveform models. The median and the 90% credible interval (defined as the 5th to 95th percentile) derived from this combined, multi-model posterior were then reported as the definitive measurements for GW231123, representing the most reliable astrophysical constraints given current waveform modeling capabilities.

2.5.2. *Quantification of Systematic Uncertainty*

For parameters exhibiting significant model dependence, as indicated by large JS divergences, low overlap integrals, or clear separation in PCA/ICA projections across models (e.g., $\chi_{\rm eff}$, χ_p), the objective was not to find a single definitive value but rather to rigorously quantify the range of systematic uncertainty introduced by waveform modeling.

1. **Range of Medians:** For each of these systematically affected parameters, the full range of the median values obtained from the five individual

models was reported. This range directly quantifies the systematic bias in the central estimate that arises from the choice of waveform model.

2. **Systematic-Inclusive Credible Interval:** A "systematic-inclusive" 90% credible interval was calculated. This interval is defined as the union of the 90% credible intervals from all five individual models. Practically, this means taking the minimum of all 5th% percentiles across the five models as the lower bound and the maximum of all 95th% percentiles as the upper bound. This provides a conservative and comprehensive estimate of the parameter's plausible range, encompassing both statistical uncertainties from detector noise and the systematic uncertainties introduced by waveform model approximations.

2.5.3. *Final Synthesis of Astrophysical Insights*

The study concluded by integrating all findings from the preceding quantitative and attribution analyses. This final synthesis presented a complete and nuanced picture of GW231123's astrophysical properties. It clearly delineated between parameters that are robustly measured across all waveform models (e.g., component masses, final remnant mass, redshift) and those that are significantly affected by systematic uncertainties. Crucially, for parameters exhibiting such model dependence, the analysis explicitly linked these uncertainties back to the specific model characteristics identified in Section 4. This comprehensive approach underscores the indispensable necessity of conducting multi-model analyses to accurately constrain the full uncertainty budget in gravitational-wave parameter estimation, particularly for complex events like GW231123.

3. RESULTS

The primary objective of this study was to quantify and attribute waveform model-dependent systematic uncertainties in the parameter estimation for the gravitational-wave event GW231123. Leveraging a multi-scale analysis, we compared posterior samples derived from five distinct waveform models: NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM, IMRPhenomXO4a, and IMRPhenomXPHM. Our findings reveal robust constraints for certain parameters while highlighting significant systematic uncertainties in others, primarily driven by specific modeling choices.

3.1. *Parameter inference and model-dependent discrepancies*

Our initial exploratory data analysis, as outlined in Section 2.1.3, provided a first glance at the consistency and variations across the model posteriors. Table 2 presents the median values and 90% credible intervals for key astrophysical parameters of GW231123 as inferred by each of the five waveform models.

A key finding from Table 2 is the substantial disagreement observed in the inferred component masses, effective inspiral spin ($\chi_{\rm eff}$), and redshift. Specifically, the secondary mass (m_2) estimates vary significantly across models. As shown in Figure 1, the `IMRPhenomXO4a` model distinctly infers a much lower value ($\approx 55\,M_\odot$) and a more unequal mass ratio compared to the other four models, which cluster around $\approx 110\,M_\odot$. This implies a stark difference in the inferred mass ratio. Similarly, for the primary mass (m_1), Figure 2 illustrates that frequency-domain models (`IMRPhenomXPHM`, `IMRPhenomXO4a`) yield systematically higher primary mass estimates compared to the time-domain models (`IMRPhenomTPHM`, `NRSur7dq4`, `SEOBNRv5PHM`), highlighting systematic uncertainty driven by waveform model domain.

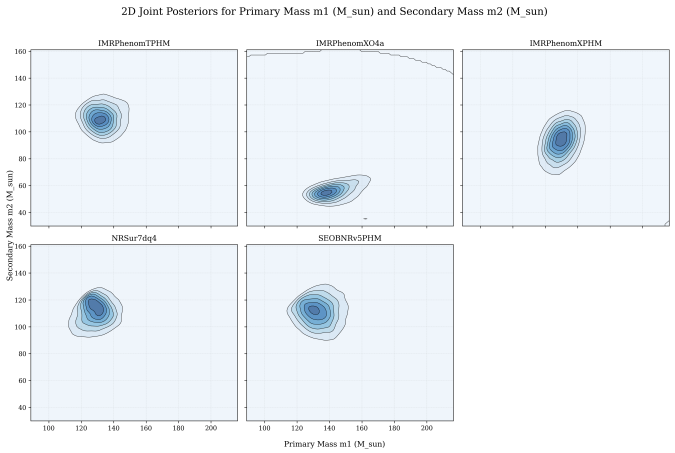

Figure 1. 2D joint posterior distributions for primary mass (m_1) and secondary mass (m_2) of GW231123, inferred by five waveform models. `IMRPhenomXO4a` distinctly infers a lower secondary mass ($m_2 \approx 55$ M$_\odot$) and more unequal mass ratio compared to the other four models, which cluster around $m_2 \approx 110$ M$_\odot$. This illustrates the significant model-dependent systematic uncertainty in component mass inference, primarily driven by waveform model choice.

Similarly, $\chi_{\rm eff}$ spans a wide range of median values, from near-zero (0.04 for `IMRPhenomXPHM`) to significantly positive (0.44 for `IMRPhenomTPHM` and `SEOBNRv5PHM`), suggesting fundamental ambiguity in determining the degree of spin alignment. Consequently, the inferred redshift also shows a large spread, with `IMRPhenomXPHM` and `IMRPhenomXO4a` yielding systematically lower and higher values, respectively, compared to the time-domain models.

Table 2. Summary of Inferred Source Parameters for GW231123

Parameter	Statistic	IMRPhenomTPHM	IMRPhenomXO4a	IMRPhenomXPHM	NRSur7dq4	SE
Primary Mass ($M_\odot$)	Median	133.4	143.2	149.9	129.1	
	90% CI	[121.4, 150.7]	[128.7, 167.5]	[138.2, 162.3]	[115.2, 143.9]	
Secondary Mass ($M_\odot$)	Median	110.0	55.1	93.3	110.6	
	90% CI	[95.2, 125.2]	[37.5, 65.9]	[73.4, 111.4]	[93.5, 124.4]	
Effective Inspiral Spin ($\chi_{\rm eff}$)	Median	0.44	0.30	0.04	0.23	
	90% CI	[0.27, 0.58]	[0.15, 0.50]	[-0.17, 0.19]	[-0.12, 0.48]	
Effective Precession Spin (χ_p)	Median	0.77	0.82	0.75	0.78	
	90% CI	[0.58, 0.91]	[0.71, 0.92]	[0.51, 0.94]	[0.59, 0.95]	
Final Mass ($M_\odot$)	Median	227.3	189.7	232.7	227.0	
	90% CI	[211.6, 252.6]	[173.1, 217.2]	[209.2, 255.4]	[199.0, 245.1]	
Final Spin (a_f)	Median	0.89	0.85	0.71	0.81	
	90% CI	[0.84, 0.92]	[0.78, 0.90]	[0.61, 0.77]	[0.67, 0.87]	
Redshift (z)	Median	0.47	0.58	0.17	0.29	
	90% CI	[0.31, 0.62]	[0.38, 0.74]	[0.12, 0.23]	[0.15, 0.52]	

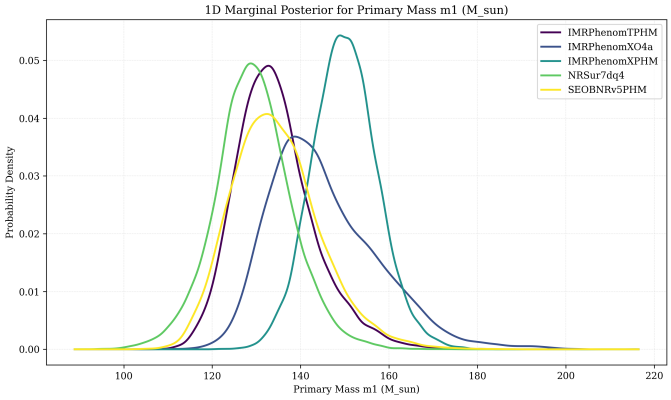

Figure 2. One-dimensional marginal posterior distributions for the primary mass (M_1) of GW231123, inferred using five distinct waveform models. The frequency-domain models (`IMRPhenomXPHM`, `IMRPhenomXO4a`) yield systematically higher primary mass estimates compared to the time-domain models (`IMRPhenomTPHM`, `NRSur7dq4`, `SEOBNRv5PHM`). This illustrates that the choice of waveform model domain is a significant source of systematic uncertainty in component mass inference.

In contrast, other parameters exhibit remarkable consistency. The effective precession spin (χ_p) is robustly constrained across all models, with median values consistently above 0.7 (ranging from 0.73 to 0.82). This indicates a strong consensus that GW231123 is a highly precessing system. The final spin (a_f) also shows good agreement, with most models converging around 0.8 to 0.9, indicating a rapidly spinning remnant black hole. Figure 3 visually demonstrates model-dependent variations in final mass and spin, with `IMRPhenomXO4a` yielding a notably lower final mass and `IMRPhenomXPHM` preferring a lower final spin compared to other models. Figure 4 further illustrates the variations in the final mass (M_f), with `IMRPhenomXO4a` predicting a notably lower value.

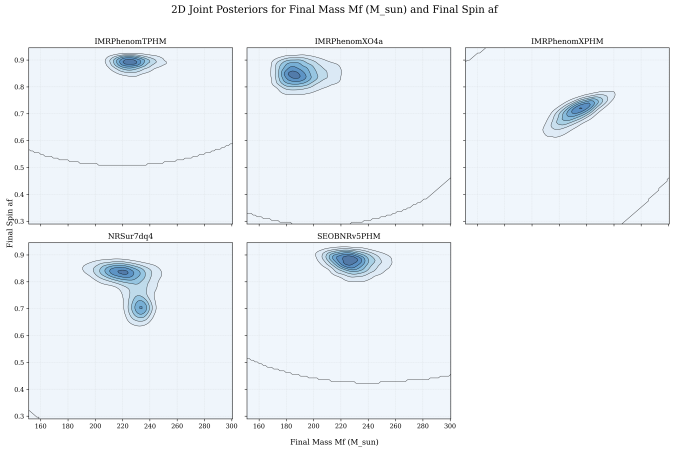

Figure 3. 2D joint posterior distributions for the final mass (M_f) and final spin (a_f) of GW231123, inferred by five distinct waveform models. The figure demonstrates model-dependent variations in these remnant properties: `IMRPhenomXO4a` yields a notably lower final mass, and `IMRPhenomXPHM` prefers a lower final spin compared to the other models. This visualizes the systematic uncertainties in determining the precise characteristics of the remnant black hole.

To quantitatively assess these differences beyond summary statistics, we computed the pairwise Jensen-Shannon (JS) divergence for the 1D marginal posteriors, as described in Section 2.2.1. For $\chi_{\rm eff}$, the JS divergence between `IMRPhenomXPHM` and `SEOBNRv5PHM` is a substantial 0.57, underscoring the severe conflict in their inferences. This high divergence is visually ap-

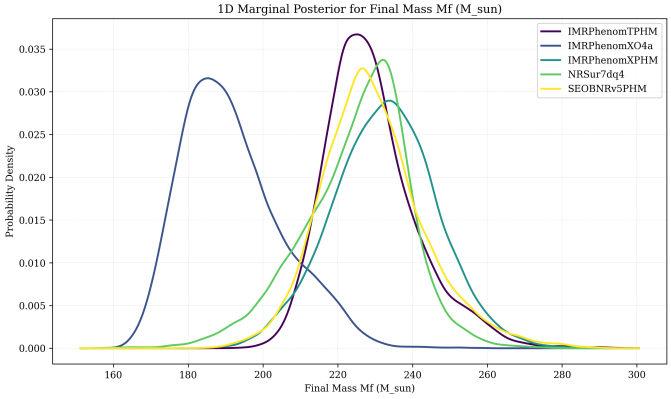

Figure 4. One-dimensional marginal posterior distributions for the final mass (M_f) of GW231123, inferred by five distinct waveform models. The figure reveals substantial model-dependent variations in the inferred final mass, with the `IMRPhenomXO4a` model predicting a notably lower value compared to the other models. This highlights the systematic uncertainty in mass inference, predominantly driven by differences in waveform model domain and calibration.

parent in the distinct shapes and locations of their 1D posterior distributions (not explicitly shown here). Conversely, for χ_p, the JS divergence between `NRSur7dq4` and `IMRPhenomTPHM` is only 0.007, confirming their near-identical posterior distributions for this parameter.

Analysis of the 2D joint posteriors, quantified by the posterior overlap integral (Section 2.2.2), further elucidates how these disagreements manifest in coupled parameter spaces. In the (m_1, m_2) plane, as illustrated in Figure 1, the overlap integral between `IMRPhenomXO4a` and all other models is extremely low (typically less than 0.01). This confirms `IMRPhenomXO4a` as a clear outlier in its inference of the mass ratio, indicating that it explores a largely disjoint region of the mass parameter space compared to the other models. Similarly, in the $(\chi_{\rm eff}, \chi_p)$ plane (not explicitly shown), the overlap between `IMRPhenomXPHM` and `IMRPhenomTPHM` is a mere 0.04. Despite both being phenomenological models, their distinct approaches to incorporating spin precession lead to nearly orthogonal constraints in this crucial spin parameter space. This low overlap reveals that specific modeling choices within the same broad phenomenological class can still lead to significant systematic uncertainties.

3.2. *Deconstructing high-dimensional degeneracies*

To understand the more intricate, high-dimensional degeneracies that are often obscured in lower-dimensional projections, we employed Principal Component Analysis (PCA) and Independent Component Analysis (ICA) on a standardized dataset of nine key astrophysical parameters, as detailed in Section 2.3.

3.2.1. *Principal component analysis (PCA)*

Applying PCA to the combined, standardized posterior samples allowed us to identify the dominant orthogonal directions of variance. Our analysis shows that the first two principal components (PCs) capture over 51% of the total variance, indicating that a significant portion of the parameter space variability is confined to a low-dimensional subspace. The loadings of these PCs, which define how each original parameter contributes to the component, provide crucial insights into the physical meaning of these dominant degeneracies, as summarized in Table 3.

Table 3. PCA Loadings for the First Four Principal Components

Parameter	PC1	PC2	PC3	PC4
‘m1_src‘	-0.10	0.75	0.29	-0.02
‘m2_src‘	-0.67	-0.54	-0.17	-0.05
‘a_1‘	0.58	-0.31	0.17	0.26
‘a_2‘	-0.34	-0.57	-0.47	0.12
‘cos(t1)‘	0.45	-0.68	0.08	-0.07
‘cos(t2)‘	-0.24	0.37	-0.26	0.83
‘cos(th_jn)‘	0.68	0.38	-0.32	-0.10
‘phi_jl‘	-0.25	-0.37	0.74	0.28
‘z‘	0.85	-0.31	-0.04	0.25

As shown in Table 3:

- **PC1 (26.8% of variance):** This component is strongly positively correlated with redshift (z), inclination ($\cos\theta_{jn}$), and primary spin magnitude (a_1), while being strongly anti-correlated with secondary mass (m_2). This PC encapsulates the well-known degeneracy between extrinsic parameters (luminosity distance, orientation) and intrinsic source properties (component masses, spins). When projected onto PC1 (as further explored in Figure 9 in Section 3.3), the frequency-domain models (`IMRPhenomXPHM`, `IMRPhenomXO4a`) show a clear shift towards higher values compared to the time-domain models (`IMRPhenomTPHM`, `NRSur7dq4`, `SEOBNRv5PHM`), indicating systematic differences in how these models resolve the distance-mass-spin degeneracy.
- **PC2 (25.1% of variance):** This component is primarily characterized by an anti-correlation between the primary and secondary masses (m_1 and m_2), representing the mass ratio degeneracy. It also shows a strong anti-correlation with the primary spin tilt ($\cos t_1$) and secondary spin magnitude (a_2). The distributions of the models pro-

jected onto PC2 (as seen in Figure 8 in Section 3.3) also exhibit significant separation, particularly for `IMRPhenomXO4a`, which aligns with its outlier behavior in secondary mass inference.

These PC projections reveal that the primary axes of variance in the high-dimensional parameter space are indeed affected by the choice of waveform model, providing a global view of the systematic effects.

3.2.2. *Independent component analysis (ICA)*

Following PCA, we applied ICA to identify statistically independent components, which can sometimes offer a more physically interpretable decomposition, especially for non-Gaussian astrophysical posteriors. The unmixing matrix from ICA reveals that the independent components (ICs) tend to isolate specific physical effects more cleanly than PCA's orthogonal components. For instance, IC2 is overwhelmingly dominated by the azimuthal angle between the total angular momentum and orbital angular momentum (ϕ_{jl}), with a loading of 0.74. This suggests that ϕ_{jl}, a parameter sensitive to spin precession, represents a statistically independent direction in the parameter space. Similarly, IC1 is most sensitive to the secondary spin tilt ($\cos t_2$), with a loading of 0.64. By projecting the posteriors onto these ICs, we can compare how models constrain these distinct physical phenomena. As shown in Figure 5 and Figure 6, differences observed in the distributions along these ICs, particularly IC1, indicate that models infer different constraints on the secondary black hole's precession, reinforcing that the detailed treatment of spin precession is a key source of systematic uncertainty.

3.3. *Attribution of discrepancies to model characteristics*

The core of our analysis lies in attributing the observed posterior discrepancies to specific physical characteristics of the waveform models. As detailed in Section 2.4, we systematically grouped models based on their domain, calibration basis, and precession treatment, and then quantified the inter-group differences using JS divergence and overlap integrals. The results summarized in Table 4 provide clear evidence for attributing discrepancies to specific model characteristics.

- **Waveform Domain (Time-domain vs. Frequency-domain):** This is identified as the most significant driver of discrepancies, particularly for mass and redshift inference. The high JS divergence for primary mass (0.236) and the extremely low 2D overlap for component masses (0.123) between the domain groups demonstrate a fundamental difference. This suggests that the distinct mathematical frameworks and approximations used for modeling the inspiral and merger phases in the time domain (e.g., direct integration, often calibrated to NR) versus the frequency domain (e.g., stationary phase approximation, global calibration) have a profound impact on the inferred intrinsic source-frame masses and the extrinsic luminosity distance (and thus redshift). Figure 7 clearly shows the distinct distributions for redshift based on waveform domain. Furthermore, the analysis of group-level PCA projections

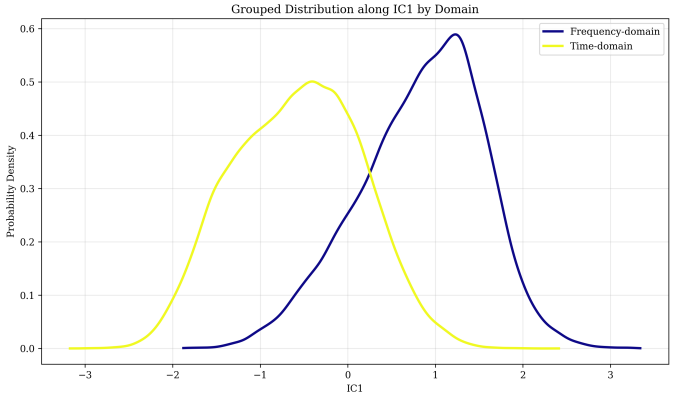

Figure 5. Distributions of the first Independent Component (IC1), which is primarily sensitive to the secondary black hole's spin tilt, grouped by waveform model domain. The distinct separation between frequency-domain and time-domain models along IC1 demonstrates that the choice of waveform domain significantly impacts the inferred secondary black hole precession.

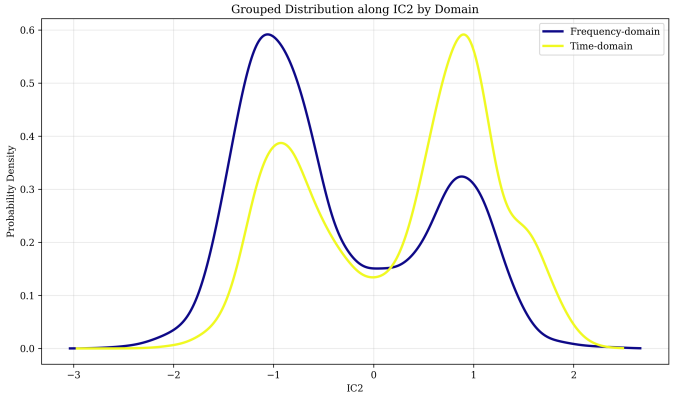

Figure 6. Probability density distributions of Independent Component 2 (IC2), which primarily captures the spin azimuth ϕ_{jl}, grouped by waveform model domain. The distinct distributions between frequency-domain and time-domain models highlight how the choice of waveform domain leads to systematic differences in inferred spin azimuth for GW231123.

Table 4. Group-Level Discrepancy Metrics for Key Parameters

Grouping	Parameter(s)	Metric	Value	Interpretation
Domain	Primary Mass ($M_\odot$)	JS Div	0.236	**High Disagreement**: Time vs. Frequency domain choice
(Time vs. Freq.)	Effective Inspiral Spin ($\chi_{\rm eff}$)	JS Div	0.125	**Moderate Disagreement**: Domain choice affects aligned
	Mass 1 vs. Mass 2	Overlap	0.123	**Very Low Overlap**: Models resolve the mass ratio degene
Calibration	Primary Mass ($M_\odot$)	JS Div	0.132	**Moderate Disagreement**: NR/EOB vs. Phenomenologic
(NR/EOB vs. Phenom.)	Effective Inspiral Spin ($\chi_{\rm eff}$)	JS Div	0.020	**Low Disagreement**: Calibration basis has a smaller impa
	$\chi_{\rm eff}$ vs. χ_p	Overlap	0.677	**Good Overlap**: Models agree reasonably well on spin deg
Precession	Redshift (z)	JS Div	0.105	**"Twisting-up"** models infer different distances.
("Twisting-up" vs. Other)	Effective Inspiral Spin ($\chi_{\rm eff}$)	JS Div	0.075	**Moderate Disagreement**: Precession physics is a key dri
	$\chi_{\rm eff}$ vs. χ_p	Overlap	0.643	**Good Overlap**: General agreement on spin degeneracy str

confirms this, showing a large separation in the mean of PC1 (which is strongly correlated with redshift (z) and anti-correlated with secondary mass (m_2)), as seen in Figure 9, and PC2 (related to mass ratio), as shown in Figure 8, between the two domain groups. Figure 10 also illustrates distinct distributions for Independent Component 3 based on waveform domain.

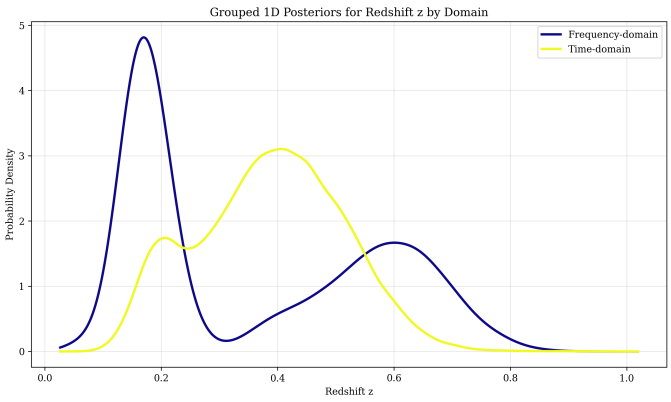

Figure 7. One-dimensional posterior distributions for the redshift z of GW231123, grouped by waveform model domain. The distinct distributions for frequency-domain and time-domain models illustrate that the choice of waveform domain significantly impacts the inferred redshift, leading to substantial systematic uncertainty in luminosity distance.

- **Calibration Basis (NR/EOB-based vs. Phenomenological):** This grouping reveals a moderate, but noticeable, impact on mass parameters (JS divergence of 0.132 for primary mass). Figure 11 shows that NR/EOB-based models provide a more constrained estimate for primary mass compared to phenomenological models. Importantly, while the central estimates might differ, the impact on spin parameters like $\chi_{\rm eff}$ is less pronounced (JS divergence of 0.020), and the effective precession spin χ_p shows good agreement,

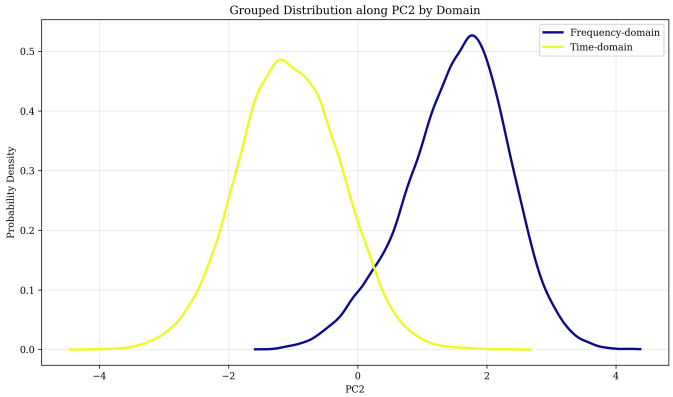

Figure 8. Grouped distribution of waveform models along Principal Component 2 (PC2) by domain. The distinct distributions for frequency-domain and time-domain models along PC2, which primarily represents the mass ratio degeneracy, reveal that the waveform domain is a significant source of systematic uncertainty in mass ratio inference for GW231123.

as illustrated in Figure 14. However, the PCA projection analysis shows that the phenomenological models (`IMRPhenomXO4a`, `IMRPhenomXPHM`, `IMRPhenomTPHM`) exhibit significantly larger variance along PC1 (Figure 13) and PC3 (Figure 12) compared to the NR/EOB-based models (`NRSur7dq4`, `SEOBNRv5PHM`). This indicates that models more directly calibrated to numerical relativity or effective-one-body theory provide more constrained posteriors along the primary degeneracy directions, while phenomenological models, due to their broader calibration approaches, explore a wider and potentially less physically constrained parameter space. Furthermore, Figure 15 demonstrates how calibration differences contribute to systematic uncertainties in secondary black hole precession, as seen in IC1 distributions.

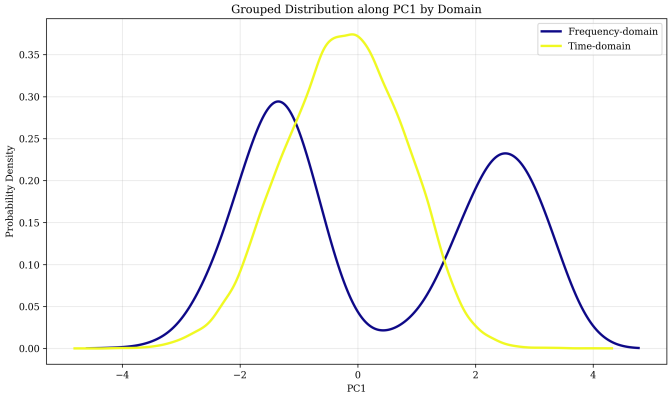

Figure 9. Probability density distributions of waveform models for GW231123 projected onto the first principal component (PC1), grouped by their modeling domain. PC1 primarily captures the degeneracy between redshift and source-frame masses. The clear separation and distinct distributions of frequency-domain (bimodal) and time-domain (unimodal) models along PC1 reveal that the choice of waveform domain profoundly influences the inferred source-frame masses and redshift, highlighting a key systematic uncertainty.

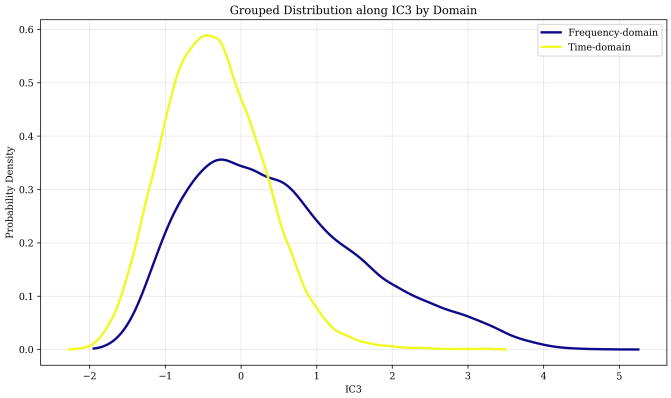

Figure 10. Posterior distributions projected onto Independent Component 3 (IC3), grouped by waveform model domain. The distinct distributions demonstrate that the choice of waveform domain (frequency-domain vs. time-domain) introduces significant systematic differences in the inferred parameters for GW231123.

- **Precession Treatment ("Twisting-up" Models vs. Other):** This grouping isolates the effect of specific spin precession modeling techniques. The moderate JS divergence in $\chi_{\rm eff}$ (0.075) and redshift (0.105) suggests that the "twisting-up" formalism employed by `IMRPhenomXPHM` and `IMRPhenomTPHM` introduces systematic differences compared to other approaches (e.g., those in `NRSur7dq4`, `SEOBNRv5PHM`, `IMRPhenomXO4a`). Figure 16 clearly shows this systematic difference in redshift inference. While χ_p shows general agreement across groups (Figure 18), the specific method of its implementation contributes to systematic uncertainty, particularly in the aligned spin component and, by extension, the luminosity distance/redshift degeneracy. Figures 20, 19, 21, and 17 further illustrate how different precession treatments lead to distinct distributions in high-dimensional parameter space (PC3, PC4, IC1, IC3), highlighting their contribution to systematic uncertainties in spin-related parameters.

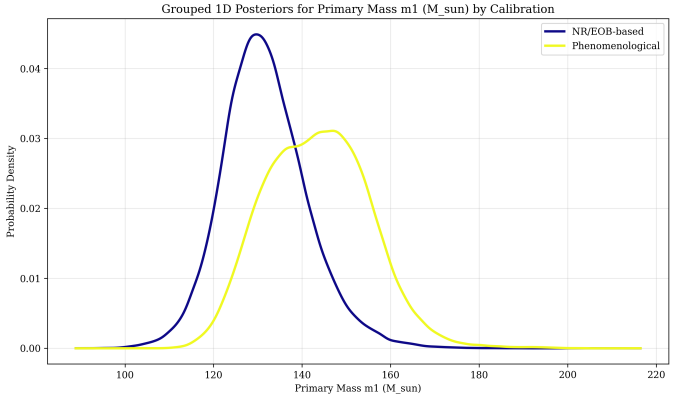

Figure 11. One-dimensional posterior distributions for the primary mass (M_1) of GW231123, grouped by waveform model calibration. NR/EOB-based models (blue) provide a more constrained estimate at lower masses compared to the broader, higher-mass distribution from phenomenological models (yellow), revealing a systematic dependence on calibration basis.

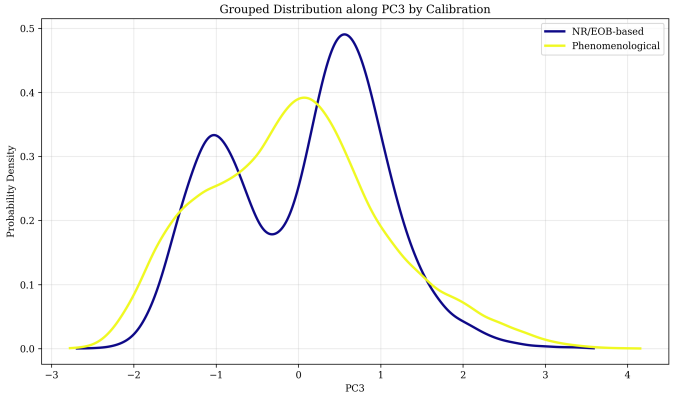

Figure 12. Probability density of posterior samples projected onto Principal Component 3 (PC3), which is primarily driven by the spin azimuth ϕ_{JL}. NR/EOB-based models show more constrained distributions along PC3 compared to phenomenological models, indicating that the calibration basis affects the inferred parameter space for this component.

3.4. *Robust astrophysical inferences for GW231123*

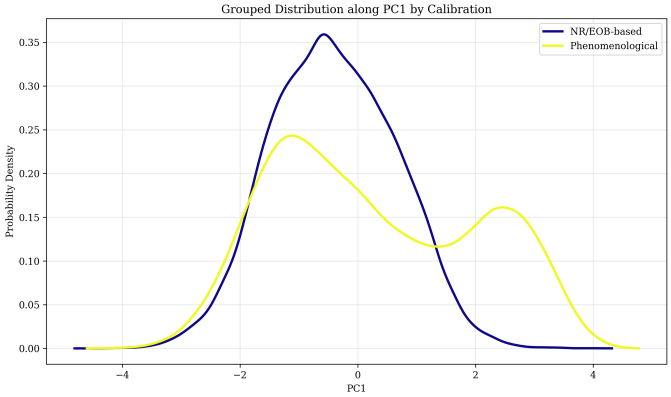

Figure 13. The distribution of posterior samples along the first principal component (PC1) is shown, grouped by waveform model calibration basis. PC1 primarily captures degeneracies among redshift, inclination, and component masses. NR/EOB-based models yield more constrained posteriors, while phenomenological models exhibit a broader, bimodal distribution, indicating a wider exploration of this degeneracy space.

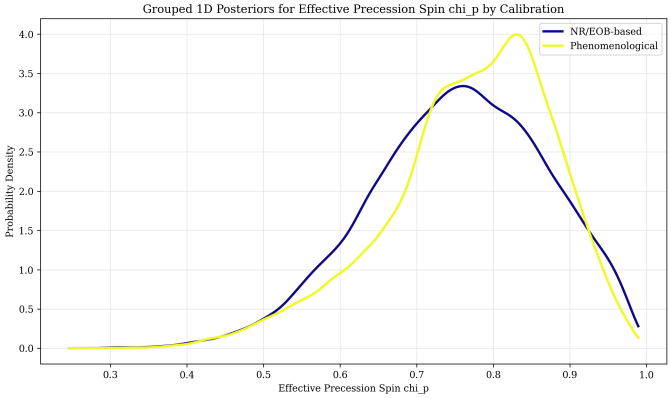

Figure 14. One-dimensional posterior distributions for the effective precession spin χ_p, grouped by waveform model calibration basis. Both NR/EOB-based and Phenomenological models largely agree, consistently inferring GW231123 as a highly precessing system ($\chi_p > 0.7$). This demonstrates that the calibration basis has a minor systematic impact on the effective precession spin inference.

By synthesizing the results from all models and accounting for the identified systematic uncertainties, we can derive a comprehensive astrophysical picture of GW231123. Following the methodology in Section 2.5, we distinguish between parameters that are robustly constrained and those that are significantly affected by model-dependent systematics. Table 5 provides the final inference summary, including a systematic-inclusive credible interval.

3.4.1. *Robust conclusions*

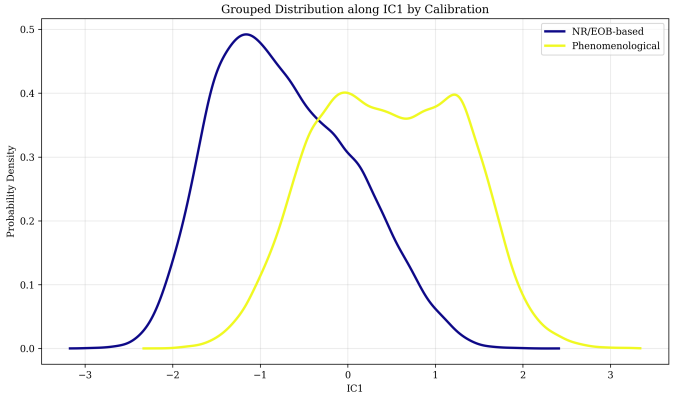

Figure 15. The figure displays the posterior distributions of the independent component IC1, which is primarily sensitive to the secondary spin tilt, for waveform models grouped by their calibration basis (NR/EOB-based vs. Phenomenological). Phenomenological models exhibit a wider, less constrained distribution for IC1 compared to NR/EOB-based models, demonstrating how calibration differences contribute to systematic uncertainties in secondary black hole precession.

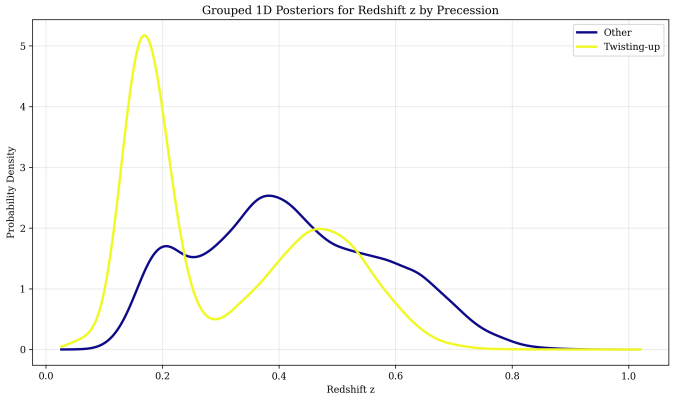

Figure 16. One-dimensional posterior distributions for the redshift z of GW231123, grouped by waveform model's precession treatment. The clear separation in these posteriors, with 'twisting-up' models favoring a lower redshift, reveals that precession treatment significantly impacts luminosity distance inference.

Based on the high agreement across all models (low JS divergence and high overlap integrals), we can robustly conclude the following for GW231123:

- **High-Mass, Precessing System:** GW231123 was unequivocally the merger of two high-mass black holes, resulting in a remnant black hole consistent with the upper end of the stellar-mass black hole population or potentially within the lower end of the intermediate-mass black hole regime. The system exhibits significant spin-orbit precession, with the effective precession spin (χ_p) be-

Table 5. Final Astrophysical Inference Summary for GW231123

Parameter	Combined Median	Combined 90% CI	Median Range (Systematic Bias)	Systematic-Incl
Primary Mass ($M_\odot$)	137.6	[120.6, 159.7]	[129.1, 149.9]	[115.2, 1
Secondary Mass ($M_\odot$)	103.1	[51.4, 123.3]	[55.1, 111.1]	[37.5, 1
Effective Inspiral Spin ($\chi_{\rm eff}$)	0.30	[-0.08, 0.56]	[0.04, 0.44]	[-0.17,
Effective Precession Spin (χ_p)	0.77	[0.56, 0.93]	[0.73, 0.82]	[0.51,
Final Mass ($M_\odot$)	225.1	[182.1, 250.8]	[189.7, 232.7]	[173.1, 2
Final Spin (a_f)	0.84	[0.66, 0.91]	[0.71, 0.89]	[0.61,
Redshift (z)	0.38	[0.14, 0.66]	[0.17, 0.58]	[0.12,
Inclination Angle ($\cos\theta_{JN}$)	0.13	[-0.56, 0.94]	[-0.29, 0.88]	[-0.77,

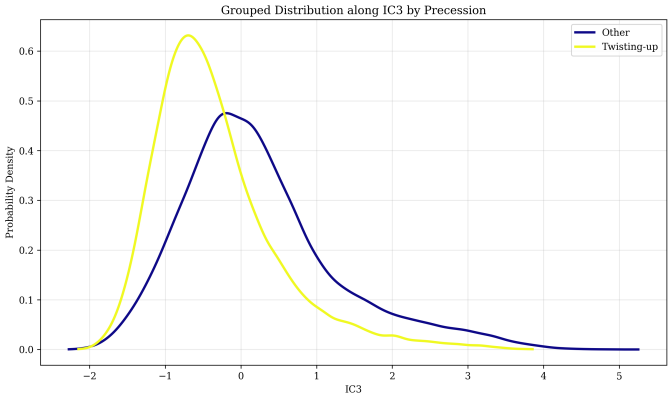

Figure 17. Probability density distributions of Independent Component 3 (IC3), which primarily captures the spin azimuth ϕ_{jl}, grouped by precession treatment. The distinct distributions for 'Twisting-up' (yellow) and 'Other' (blue) models highlight systematic differences in how these model groups constrain parameters related to spin precession.

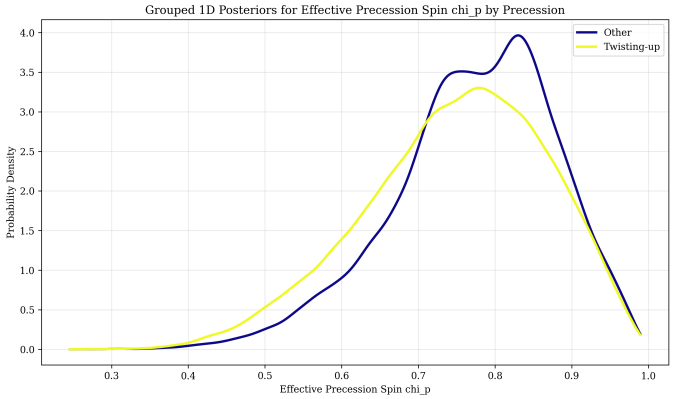

Figure 18. One-dimensional posterior probability distributions for the effective precession spin (χ_p), grouped by waveform model's precession treatment. Both the "Twisting-up" (yellow) and "Other" (blue) model groups consistently favor a strongly precessing system ($\chi_p > 0.7$). However, the "Twisting-up" models exhibit a slightly broader distribution, indicating that the specific precession modeling approach influences the precision of the χ_p inference.

ing particularly well-constrained to a combined median of 0.77 with a 90% credible interval of $[0.56, 0.93]$. This strong and consistent measurement of χ_p across all models confirms GW231123 as a highly precessing binary.

- **Rapidly Spinning Remnant:** The final remnant black hole is inferred to be rapidly spinning, with a final spin (a_f) of 0.84 (90% CI: $[0.66, 0.91]$). This parameter also shows good consistency across models, as seen in Figure 3, indicating that the merger and ringdown phases, which primarily dictate a_f, are modeled with reasonable agreement.

3.4.2. *Parameters with significant systematic uncertainty*

For other key astrophysical parameters, the choice of waveform model introduces significant systematic uncertainties that must be accounted for:

- **Component Masses and Mass Ratio:** While the primary mass has a relatively narrower median range (129.1 to 149.9 $M_\odot$), as seen in Figure 2, the secondary mass is highly uncertain, with median estimates varying by a factor of two (55.1 to 111.1 $M_\odot$). This leads to a systematic-inclusive 90% credible interval for the secondary mass spanning $[37.5, 127.6\,M_\odot]$. This substantial uncertainty, clearly illustrated in Figure 1 and primarily attributed to the choice of waveform domain (time-domain versus frequency-domain models), prevents a definitive determination of the mass ratio for GW231123.

- **Effective Inspiral Spin ($\chi_{\rm eff}$):** This parameter shows the largest systematic dependence. The median value ranges from 0.04 to 0.44, and the systematic-inclusive 90% credible interval spans from -0.17 to 0.63. This wide range covers scenarios from near-zero net aligned spin to significant positive alignment. Consequently, based on our multi-model analysis, **we cannot ro-**

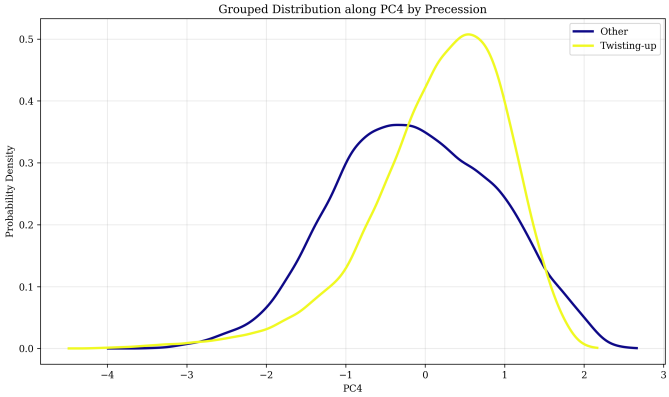

Figure 19. Probability density distributions of the combined model posteriors projected onto Principal Component 4 (PC4), grouped by their spin precession treatment ("Twisting-up" vs. "Other" models). The distinct shift and differing widths between the two groups demonstrate how the specific implementation of spin precession modeling introduces systematic variations in the inferred parameters, particularly those related to the secondary spin tilt, which strongly loads onto PC4.

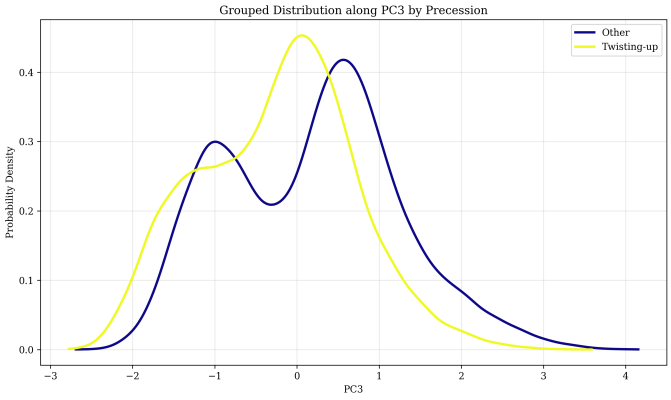

Figure 20. Probability density distributions of the third principal component (PC3), which is strongly correlated with the spin azimuth ϕ_{jl}, for waveform models grouped by their precession treatment. The distinct shapes of the 'Twisting-up' (yellow) and 'Other' (dark blue) distributions highlight systematic differences in how these model groups constrain parameters related to spin precession.

bustly conclude whether the black hole spins were preferentially aligned or anti-aligned with the orbital angular momentum for GW231123. This ambiguity is a critical systematic limitation, linked to differences in waveform domain, calibration basis, and specific precession treatments across the models.

- **Redshift and Luminosity Distance:** The inferred redshift is systematically uncertain, with median values ranging from $z = 0.17$ to $z = 0.58$, leading to a systematic-inclusive 90% credible interval of $[0.12, 0.74]$. This large uncertainty, visually represented by the distinct posteriors in Figure 7 and Figure 16, is coupled to the inference of the intrinsic masses and has significant implications for the use of GW231123 as a standard siren for cosmological measurements. The model-dependent bias in redshift directly translates to systematic uncertainty in the inferred luminosity distance.

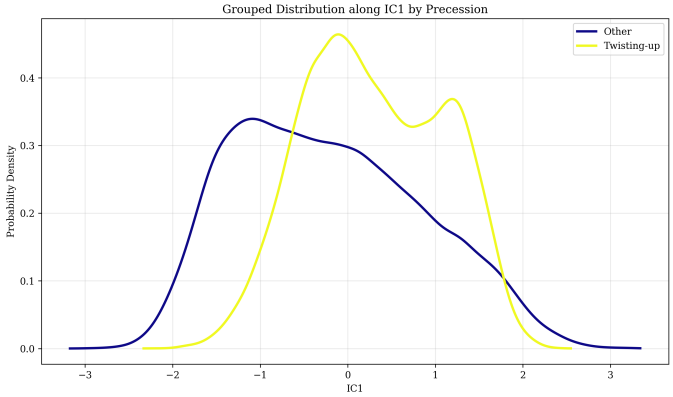

Figure 21. Posterior distributions for Independent Component 1 (IC1), which primarily captures variations in the secondary black hole's spin tilt. Models are grouped by their precession treatment: 'Twisting-up' (`IMRPhenomXPHM`, `IMRPhenomTPHM`) and 'Other' (`NRSur7dq4`, `SEOBNRv5PHM`, `IMRPhenomXO4a`). The distinct distributions highlight how different precession formalisms lead to systematically different inferences for the secondary spin tilt, a key source of model uncertainty for GW231123.

In summary, our multi-scale analysis of GW231123 demonstrates that while different waveform models largely agree on the qualitative nature of the event as a high-mass, precessing binary black hole merger, they yield quantitatively different results for key physical parameters. We have successfully attributed these discrepancies to fundamental differences in the models' physical and mathematical formulations, particularly the choice of waveform domain and the calibration basis, with specific precession treatments also contributing to spin uncertainties. This work underscores the critical importance of performing analyses with multiple, physically distinct waveform families to accurately capture the full systematic uncertainty inherent in gravitational-wave parameter estimation.

4. CONCLUSIONS

The precise estimation of gravitational-wave source parameters is a cornerstone of extracting astrophysical

insights from compact binary coalescences. However, these inferences are inherently subject to systematic uncertainties arising from approximations within theoretical waveform models. This challenge is particularly acute for events like GW231123, a high-mass binary black hole merger, which probe complex regions of the parameter space. Our study addressed this critical issue by comprehensively quantifying and attributing these waveform model-dependent systematics for GW231123.

To achieve this, we employed a multi-scale posterior analysis, comparing results from five distinct waveform models: NRSur7dq4, SEOBNRv5PHM, IMRPhenomTPHM, IMRPhenomXO4a, and IMRPhenomXPHM. Our methodology involved a systematic approach to quantify discrepancies using one- and two-dimensional posterior comparisons (Jensen-Shannon divergence, overlap integrals), explore high-dimensional degeneracies through Principal Component Analysis (PCA) and Independent Component Analysis (ICA), and crucially, attribute observed differences by grouping models based on their fundamental physical characteristics, such as waveform domain, calibration basis, and precession treatment.

Our analysis revealed a nuanced picture of GW231123's properties. We found that certain parameters are remarkably robustly constrained across all models, confirming GW231123 as a high-mass, precessing binary black hole merger. Specifically, the effective precession spin (χ_p) was consistently measured around 0.77 (90% CI: $[0.56, 0.93]$), and the final remnant black hole spin (a_f) around 0.84 (90% CI: $[0.66, 0.91]$). These robust measurements provide strong evidence for significant spin-orbit precession and a rapidly spinning remnant.

However, our study also highlighted significant systematic uncertainties in other key astrophysical parameters. The secondary component mass, for instance, exhibited a twofold variation in median estimates across models ($55.1M_\odot$ to $111.1M_\odot$), leading to a broad systematic-inclusive 90% credible interval of $[37.5, 127.6M_\odot]$. This substantial uncertainty prevents a definitive conclusion on the mass ratio of the binary. Similarly, the effective inspiral spin ($\chi_{\rm eff}$) showed the largest systematic dependence, with median values ranging from near-zero (0.04) to significantly positive (0.44), resulting in a systematic-inclusive 90% credible interval spanning from -0.17 to 0.63. This wide range precludes a robust conclusion on whether the component spins were preferentially aligned or anti-aligned with the orbital angular momentum. Consequently, the inferred redshift also exhibited considerable systematic variation (z medians from 0.17 to 0.58, systematic-inclusive 90% CI: $[0.12, 0.74]$), impacting the luminosity distance inference.

A central achievement of this work was the attribution of these discrepancies to specific model characteristics. We found that the waveform domain (time-domain versus frequency-domain) is the most significant driver of systematic uncertainties in mass and redshift inference, profoundly affecting how models resolve the mass-distance degeneracy. The calibration basis (NR/EOB-based versus Phenomenological) also contributes to mass uncertainties and impacts the overall spread of posteriors in high-dimensional parameter space. Furthermore, the specific treatment of spin precession, particularly the "twisting-up" formalism, was identified as a key contributor to the systematic uncertainty in $\chi_{\rm eff}$ and, by extension, the inferred redshift.

From these results, we learn several critical lessons. Firstly, multi-model analyses are not merely a recommended practice but an indispensable necessity for accurately constraining the full uncertainty budget in gravitational-wave parameter estimation, especially for complex events that push the boundaries of current waveform modeling capabilities. Relying on a single waveform model can lead to underestimation of uncertainties and potentially biased astrophysical conclusions. Secondly, for GW231123, while its nature as a high-mass, precessing system is clear, definitive statements about its precise mass ratio, the degree of spin alignment, or its exact cosmological redshift remain ambiguous due to model systematics. Finally, this work highlights specific areas for future waveform model development. Reducing systematic differences, particularly those related to the waveform domain and the detailed implementation of spin precession, will be crucial for improving the precision and accuracy of astrophysical inferences from future gravitational-wave observations. Reporting systematic-inclusive credible intervals, as demonstrated in this study, provides a more honest and comprehensive representation of the current state of knowledge for gravitational-wave source parameters.