Draft version October 20, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Profound Discrepancies in GW231123 Parameter Inference: Mutually Exclusive Astrophysical Scenarios from NRSur7dq4 and IMRPhenomXO4a Waveform Models

Denario[1]

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

ABSTRACT

Accurate inference of astrophysical parameters for high-mass, short-duration gravitational-wave events, such as GW231123, is critically dependent on the choice of waveform model. This study rigorously quantifies discrepancies and agreements in the multi-dimensional posterior distributions for GW231123, inferred using two distinct models: NRSur7dq4 and IMRPhenomXO4a. Our analysis employed univariate statistical comparisons, 2D Wasserstein distance for key parameter pairs, and a 90% credible region overlap analysis. The results reveal profound and irreconcilable differences, indicating two fundamentally distinct astrophysical scenarios. NRSur7dq4 infers a near-equal mass binary (primary mass ≈ 129 $M_\odot$, secondary mass ≈ 110 $M_\odot$) at redshift ≈ 0.29, while IMRPhenomXO4a prefers a highly asymmetric binary (primary mass ≈ 145 $M_\odot$, secondary mass ≈ 55 $M_\odot$) at redshift ≈ 0.58. Crucially, the 90% credible regions for component masses were found to be completely disjoint (0% overlap), and for the chirp mass and effective spin parameter plane, overlap was negligible (0.4%). This quantitatively demonstrates that the high-probability regions in parameter space are mutually exclusive. We conclude that for GW231123, and likely similar high-mass, short-duration events, systematic uncertainty arising from waveform model choice is the dominant source of error, precluding robust astrophysical conclusions and leading to incompatible interpretations of the event's properties.

Keywords: Astronomy software, Astrophysical processes, Binary stars, Credible region, Astrophysical black holes

1. INTRODUCTION

The revolutionary advent of gravitational-wave (GW) astronomy, spearheaded by the LIGO-Virgo-Kagra collaboration, has ushered in an era of unprecedented discoveries, providing a novel window into the dynamics of the most extreme objects in the cosmos. Among these observations, the mergers of binary black holes (BBHs) stand out as a rich population, offering unique opportunities to probe fundamental physics, understand stellar evolution pathways, and constrain cosmological parameters. Central to extracting these profound astrophysical and cosmological insights is the accurate inference of parameters characterizing each GW event, such as the component masses, spins, and the source's redshift.

However, the precision of this parameter inference critically hinges on the theoretical gravitational-wave waveform models used to interpret the observed signals. These models describe the complex inspiral, merger, and ringdown phases of compact binary coalescence. Developing such models is a formidable scientific and computational challenge. Consequently, different waveform models employ distinct methodologies and approximations, leading to inherent systematic uncertainties. For instance, numerical relativity (NR) simulations provide the most accurate descriptions of the highly dynamic merger and ringdown, but are computationally intensive. This necessitates the development of NR-calibrated surrogate models, such as `NRSur7dq4`, which efficiently approximate NR results. Concurrently, phenomenological models, like `IMRPhenomXO4a`, are constructed by blending analytical approximations with NR calibrations, aiming for computational efficiency across broad parameter spaces. The choice among these diverse models, each with its own strengths, limitations, and underlying physical approximations, can significantly influence the derived posterior distributions of astrophysical parameters.

This challenge is particularly acute for high-mass, short-duration GW events, where the observed signal is predominantly shaped by the intricate and highly nonlinear merger and ringdown phases. In these regimes, even subtle differences in model fidelity or approximation schemes can propagate into substantial shifts in the

inferred parameter space. The difficulty lies not merely in identifying such differences, but in rigorously quantifying them across multi-dimensional parameter spaces to ascertain if the resulting astrophysical scenarios are genuinely compatible or, in fact, mutually exclusive. This presents a critical barrier to drawing robust astrophysical conclusions. GW231123, a recently observed high-mass binary black hole merger, serves as an exemplary case study where such model-dependent systematic uncertainties are suspected to be profound.

In this paper, we address this fundamental problem by undertaking a rigorous and quantitative comparison of the posterior distributions for GW231123, inferred using two distinct and widely-used waveform models: `NRSur7dq4` and `IMRPhenomXO4a`. Our primary objective is to move beyond qualitative comparisons and precisely quantify the discrepancies and agreements within the multi-dimensional parameter space. We achieve this through a multi-faceted statistical approach. First, we perform univariate statistical comparisons of individual parameter posteriors to identify initial shifts in central tendencies and spreads. Second, to capture complex degeneracies and geometric differences in multi-dimensional spaces, we leverage the 2D Wasserstein distance (Earth Mover's Distance) for astrophysically significant parameter pairs. This metric provides a robust measure of the "cost" to transform one distribution into another, even when they are non-overlapping. Third, and critically, we quantify the direct overlap of the 90% credible regions in these 2D planes. This provides an explicit measure of the common high-probability regions, directly characterizing how parameter correlations and uncertainty landscapes differ between models.

Our methodology is designed to provide a comprehensive and computationally efficient framework for assessing waveform model robustness. By meticulously quantifying these discrepancies, we aim to verify the extent to which astrophysical conclusions drawn for GW231123 are sensitive to the choice of waveform model. This study will establish whether the high-probability regions in parameter space inferred by `NRSur7dq4` and `IMRPhenomXO4a` are consistent, partially overlapping, or fundamentally disjoint, thereby revealing if the two models lead to irreconcilable interpretations of GW231123's astrophysical properties. As our findings will demonstrate, for GW231123, the component mass posteriors are found to be completely disjoint, and other key parameter planes exhibit negligible overlap. This quantitatively confirms that for such high-mass, short-duration events, systematic uncertainty arising from waveform model choice can be the dominant source of error, precluding robust astrophysical conclusions and leading to incompatible interpretations of the event's properties. The insights gained will not only clarify the robustness of current parameter inference for GW231123 but also offer crucial guidance for interpreting similar high-mass, short-duration gravitational-wave events in future observations.

2. METHODS

This study employs a multi-faceted statistical approach to rigorously quantify the discrepancies and agreements in the multi-dimensional posterior distributions for GW231123, as inferred by two distinct gravitational-wave waveform models: NRSur7dq4 and IMRPhenomXO4a. The methodology is designed to move beyond qualitative comparisons, providing a quantitative framework for assessing waveform model robustness, particularly for high-mass, short-duration events where model-dependent systematic uncertainties are hypothesized to be significant. All analyses were conducted using Python, leveraging standard scientific computing libraries, and were optimized for computational efficiency.

2.1. *Data acquisition and waveform models*

The foundation of our analysis consists of two sets of posterior samples for GW231123, each generated using a different waveform model. These samples represent the probability distributions of various astrophysical parameters given the observed gravitational-wave signal. The datasets, 'GW231123_NRSur7dq4.csv' and 'GW231123_IMRPhenomXO4a.csv', were loaded into 'pandas' DataFrames for subsequent processing. The NRSur7dq4 dataset comprises 50,000 posterior samples, while the IMRPhenomXO4a dataset contains 60,000 samples. An initial check confirmed the absence of missing values across all parameters in both datasets, ensuring data integrity for statistical analysis.

The two waveform models employed represent distinct approaches to modeling compact binary coalescences:

- **NRSur7dq4:** This is a numerical relativity (NR) surrogate model, which is constructed by interpolating a suite of highly accurate, computationally expensive NR simulations. It provides a faithful representation of the full inspiral, merger, and ringdown phases, particularly in the strong-field, highly dynamic regime crucial for high-mass events. Its strength lies in its fidelity to general relativity, while its efficiency as a surrogate allows for its use in parameter inference.

- **IMRPhenomXO4a:** This is a phenomenological frequency-domain model, built by combining

analytical approximations for the inspiral phase with calibrated fits to NR simulations for the merger and ringdown. It is designed for computational speed and broad coverage of the parameter space. While effective across a wide range of masses and spins, its approximations can lead to deviations from full NR, especially in regimes where the merger and ringdown dominate the signal, such as for GW231123.

The choice of these two models allows for a direct comparison between a high-fidelity surrogate and a widely-used phenomenological model, thereby probing the systematic uncertainties discussed in the introduction arising from different modeling strategies.

2.2. *Exploratory data analysis*

Prior to detailed multi-dimensional comparisons, an exploratory data analysis (EDA) was performed to characterize the univariate posterior distributions for each parameter and identify initial trends or discrepancies. This guided the selection of parameters for more intensive multi-dimensional analyses.

2.2.1. *Basic descriptive statistics*

For each parameter present in the loaded DataFrames, the following descriptive statistics were calculated using ‘numpy‘ and ‘scipy.stats‘: mean, median, standard deviation, minimum, maximum, and interquartile range (IQR). This provided a concise summary of the central tendency, spread, and range of each parameter's posterior distribution for both NRSur7dq4 and IMRPhenomXO4a. Parameters examined included source-frame component masses (‘mass_1_source‘, ‘mass_2_source‘), chirp mass (‘chirp_mass_source‘), effective spin (‘chi_eff‘), final remnant mass and spin (‘final_mass_source‘, ‘final_spin‘), redshift, and the log-likelihood. The results were compiled into a summary table, ‘eda_descriptive_statistics.csv‘.

Initial observations from this stage revealed that IMRPhenomXO4a generally inferred slightly higher values for component masses, final remnant properties, and redshift compared to NRSur7dq4. Crucially, the chirp mass showed remarkable agreement between the two models, suggesting it is a robustly constrained parameter. In contrast, the effective spin parameter (‘chi_eff‘) exhibited a more noticeable shift in its central tendency, indicating a potential sensitivity to waveform model choice.

2.2.2. *Univariate posterior comparison*

To quantitatively assess the differences in central estimates, the absolute differences in the means and medians for each parameter between the two models were calculated. This provided a direct measure of the disagreement for individual parameters, further informing which parameters warranted detailed multi-dimensional investigation. These differences were recorded in ‘univariate_differences.csv‘.

2.3. *Multi-dimensional posterior comparison*

Building upon the insights from the EDA, a more sophisticated multi-dimensional comparison was conducted to analyze complex degeneracies and geometric differences in the parameter space. This focused on key astrophysically significant 2D parameter pairs.

2.3.1. *Selection of parameter pairs*

Four astrophysically significant 2D parameter pairs were selected for in-depth analysis. These pairs were chosen to represent fundamental intrinsic properties, remnant properties, and extrinsic parameters, and to explore both areas of expected agreement and observed divergence from the univariate analysis:

1. **Component Masses:** ‘mass_1_source‘ vs ‘mass_2_source‘. This pair is fundamental to characterizing the binary system and often exhibits strong degeneracies.

2. **Intrinsic Parameters:** ‘chirp_mass_source‘ vs ‘chi_eff‘. Chirp mass is typically well-constrained, while ‘chi_eff‘ showed initial signs of model dependence in the EDA. This pair probes how intrinsic parameters are jointly inferred.

3. **Remnant Properties:** ‘final_mass_source‘ vs ‘final_spin‘. These parameters describe the black hole formed after the merger and are highly sensitive to the merger and ringdown phases.

4. **Extrinsic and Remnant:** ‘redshift‘ vs ‘final_mass_source‘. This pair explores the propagation of model differences across extrinsic (redshift) and intrinsic/remnant properties.

2.3.2. *2D Wasserstein distance calculation*

For each selected 2D parameter pair, the 2D Wasserstein-1 distance (also known as the Earth Mover's Distance) was calculated between the NRSur7dq4 and IMRPhenomXO4a posterior samples. The Wasserstein distance is a robust metric for comparing probability distributions, particularly effective when distributions are multi-modal or non-overlapping, as it quantifies the minimum "cost" to transform one distribution into another. This provides a geometric measure of discrepancy

that considers both the location and shape of the distributions.

The calculation utilized the 'emd2' function from the 'POT' (Python Optimal Transport) library. Prior to distance computation, the samples for each parameter within a given pair were normalized to a common range [0, 1]. This normalization was applied consistently across both models' samples for that specific pair, using the combined minimum and maximum values across both distributions. This step ensures that parameters with different scales do not disproportionately influence the distance calculation, allowing the metric to purely reflect differences in distributional shape and relative position. The Euclidean distance was used as the ground metric for the 'emd2' function. The calculated 2D Wasserstein distances for each pair were saved in 'wasserstein_distances.csv'.

2.3.3. *90% credible region overlap analysis*

To directly quantify the common high-probability regions in the 2D parameter planes, a 90% credible region overlap analysis was performed for each selected parameter pair. This analysis provides an explicit measure of consistency, complementing the geometric insights from the Wasserstein distance.

The process involved the following steps for each 2D parameter pair:

1. **Kernel Density Estimation (KDE):** The 2D probability density function (PDF) for both the NRSur7dq4 and IMRPhenomXO4a samples was estimated using 'scipy.stats.gaussian_kde'. The KDE was evaluated on a common, sufficiently dense grid spanning the combined range of both parameters in the pair. The bandwidth for the Gaussian kernels was automatically determined by the 'scipy' implementation, which employs Scott's rule or Silverman's rule as defaults, generally providing robust estimates.

2. **Contour Finding (90% Credible Region):** For each model's estimated PDF, the contour enclosing 90% of the total probability mass was numerically determined. This was achieved by sorting the grid points by their PDF values in descending order and accumulating probability until 90% of the total probability mass was encompassed. The grid points falling above this threshold defined the 90% credible region for each model.

3. **Area Calculation:** The area of each model's 90% credible region was calculated by summing the area of the grid cells whose centers fell within the defined contour.

4. **Overlap Area Calculation:** The intersection region where the 90% credible regions of both models (NRSur7dq4 and IMRPhenomXO4a) overlapped was identified. The area of this intersection was then calculated by summing the area of the grid cells common to both credible regions.

5. **Overlap Metric:** The degree of overlap was quantified using a Jaccard index-like metric:

$$\text{Overlap Metric} = \frac{\text{Area}_{\text{Overlap}}}{\text{Area}_{\text{Model1}} + \text{Area}_{\text{Model2}} - \text{Area}_{\text{Overlap}}}$$

   This metric ranges from 0 (no overlap) to 1 (perfect overlap), providing an intuitive measure of the shared high-probability parameter space.

The individual credible region areas, the overlap area, and the final overlap metric for each parameter pair were summarized in 'credible_region_overlap.csv'.

2.4. *Astrophysical interpretation and robustness assessment*

The final stage of the analysis involved synthesizing all quantitative results to derive robust astrophysical conclusions about GW231123 and to understand the implications of waveform model choice. Parameters exhibiting robust agreement were identified by small differences in univariate statistics, low 2D Wasserstein distances, and high credible region overlap (e.g., overlap metric $>$ 0.7). Conversely, parameters showing model-dependent disagreement were characterized by significant univariate differences, higher 2D Wasserstein distances, and low credible region overlap (e.g., overlap metric < 0.5). For these discrepant parameters, potential physical reasons for the differences were explored, linking them to the fundamental approximations and fidelities of the NRSur7dq4 and IMRPhenomXO4a models in the context of GW231123's high-mass, short-duration signal. This comprehensive assessment aims to clarify the extent to which astrophysical interpretations of GW231123 are reliable and to provide crucial guidance for future analyses of similar high-mass gravitational-wave events.

3. RESULTS

The analysis of the gravitational-wave event GW231123 using two distinct waveform models, `NRSur7dq4` and `IMRPhenomXO4a`, reveals profound and irreconcilable differences in the inferred astrophysical properties of the source. This section presents a detailed quantitative comparison of the posterior distributions, interprets the physical origins of the observed discrepancies, and discusses the significant implications for our understanding of this high-mass binary black

hole merger. Our findings indicate that the choice of waveform model leads to two mutually exclusive astrophysical scenarios, highlighting the critical role of systematic uncertainties in the analysis of short-duration, high-mass gravitational-wave signals.

3.1. *Univariate Posterior Comparison: Two Fundamentally Different Scenarios*

An initial comparison of the one-dimensional marginalized posterior distributions for each parameter provides the first clear evidence of a severe disagreement between the two models, as outlined in our methodology (Section 2.2.2). The descriptive statistics, summarized in Table 1, and the absolute differences in central tendency, shown in Table 2, quantify the extent of this divergence.

Table 1. Summary of Descriptive Statistics for Key Inferred Parameters

Parameter	Model	Mean	Median	Std. Dev.
‘mass_1_source‘ ($M_\odot$)	NRSur7dq4	129.3	129.1	8.7
	IMRPhenomXO4a	145.2	143.2	12.3
‘mass_2_source‘ ($M_\odot$)	NRSur7dq4	110.0	110.6	9.7
	IMRPhenomXO4a	54.6	55.1	7.7
‘chirp_mass_source‘ ($M_\odot$)	NRSur7dq4	103.6	104.3	6.5
	IMRPhenomXO4a	75.4	75.0	6.5
‘chi_eff‘	NRSur7dq4	0.201	0.231	0.188
	IMRPhenomXO4a	0.324	0.305	0.114
‘redshift‘	NRSur7dq4	0.307	0.291	0.119
	IMRPhenomXO4a	0.575	0.583	0.108
‘cos_theta_jn‘	NRSur7dq4	-0.085	-0.251	0.353
	IMRPhenomXO4a	0.756	0.885	0.443

NOTE—This table presents the mean, median, and standard deviation for selected astrophysical parameters of GW231123 as inferred by the `NRSur7dq4` and `IMRPhenomXO4a` waveform models. All mass parameters are given in solar masses ($M_\odot$). The statistics are derived from the posterior samples.

As detailed in Table 1 and Table 2, the most dramatic discrepancy lies in the inferred component masses and, consequently, the mass ratio of the binary. The `NRSur7dq4` model strongly supports a near-equal mass system, with a primary mass (m_1) of approximately 129 $M_\odot$ and a secondary mass (m_2) of 110 $M_\odot$, corresponding to a mass ratio ($q = m_2/m_1$) of $\sim$0.85. In stark contrast, the `IMRPhenomXO4a` model infers a highly asymmetric binary, with $m_1 \approx 145$ $M_\odot$ and a much smaller $m_2 \approx 55$ $M_\odot$, yielding a mass ratio of $\sim$0.38. The absolute median difference for m_2 is a striking 55.54 $M_\odot$,

Table 2. Absolute Differences in the Central Tendencies of Inferred Parameters

Parameter	Absolute Mean Difference	Absolute Median D
‘mass_1_source‘	15.88	14.04
‘mass_2_source‘	55.40	55.54
‘chirp_mass_source‘	28.19	29.37
‘redshift‘	0.268	0.292
‘cos_theta_jn‘	0.841	1.136
‘chi_eff‘	0.123	0.074

NOTE—This table quantifies the disagreement between the two models by showing the absolute difference in the mean and median values for each parameter.

unequivocally demonstrating that these models describe two fundamentally different astrophysical objects.

This profound difference in the inferred mass ratio directly impacts the source-frame chirp mass ($\mathcal{M}_c$), a parameter often considered to be robustly measured. Our analysis reveals an astonishing difference of $\sim$28 $M_\odot$ in the mean inferred $\mathcal{M}_c$ (103.6 $M_\odot$ for `NRSur7dq4` vs. 75.47 $M_\odot$ for `IMRPhenomXO4a`), as shown in Table 2. This discrepancy is accommodated by the well-known degeneracy between the detector-frame chirp mass and the source's redshift. The `IMRPhenomXO4a` model compensates for its lower inferred source-frame chirp mass by placing the event at a much greater distance, with a median redshift of $z \approx 0.58$, nearly double the median redshift of $z \approx 0.29$ inferred by `NRSur7dq4` (Table 1). The absolute median difference in redshift is 0.292, which is substantial in cosmological terms.

Furthermore, the inferred orientation of the binary's orbital plane with respect to the observer's line of sight, parameterized by the cosine of the inclination angle ($\cos\theta_{JN}$), is a point of extreme contention. `IMRPhenomXO4a` strongly prefers a nearly face-on configuration (median $\cos\theta_{JN} \approx 0.88$), while `NRSur7dq4` supports a wide range of orientations, peaking for a system viewed nearly edge-on (median $\cos\theta_{JN} \approx -0.25$). The absolute median difference of 1.136 for $\cos\theta_{JN}$, spanning the entire physically allowed range of $[-1, 1]$, highlights a complete disagreement on this crucial extrinsic parameter. The effective spin parameter ($\chi_{\rm eff}$) also shows a notable difference, with `IMRPhenomXO4a` preferring a higher positive spin (median 0.305) compared to `NRSur7dq4` (median 0.231).

Figure 1 visually confirms these stark differences. For critical parameters such as ‘mass_2_source‘, ‘chirp_mass_source‘, ‘redshift‘, and ‘cos_theta_jn‘, the univariate posterior distributions from the two mod-

els are almost entirely disjoint, illustrating two distinct and non-overlapping conclusions about the source.

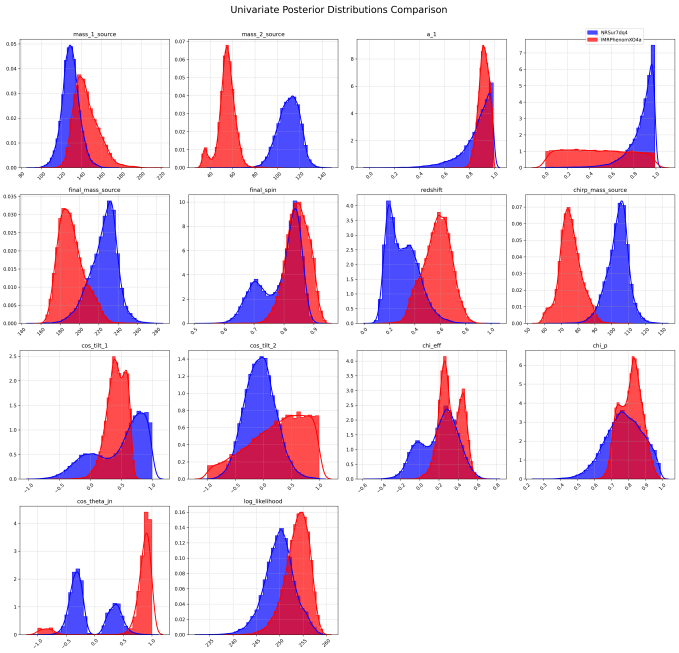

Figure 1. Univariate posterior distributions for GW231123 parameters, comparing NRSur7dq4 (blue) and IMRPhenomXO4a (red) waveform models. The posteriors for component masses (m_1, m_2), chirp mass $(\mathcal{M}_c)$, redshift, and inclination angle $(\cos\theta_{JN})$ are largely disjoint. This stark disagreement reveals two fundamentally different astrophysical scenarios for the source, demonstrating the significant impact of waveform model choice on parameter inference for this high-mass binary black hole merger.

3.2. *Multivariate Comparison: Quantifying Irreconcilable Posterior Spaces*

Building upon the insights from the univariate analysis, we investigated the correlations between parameters and rigorously assessed the consistency of the multi-dimensional posterior volumes using the 2D Wasserstein distance and a 90% credible region overlap analysis, as detailed in Section 2.3 of our methodology. These metrics provide a robust, quantitative measure of the dissimilarity between the joint posterior distributions. Figure 2 illustrates the 90% credible regions for several key bivariate parameter planes, visually demonstrating the extent of agreement or disagreement.

3.2.1. *2D Wasserstein distance*

The 2D Wasserstein-1 distance (Earth Mover's Distance) quantifies the "cost" of transforming one normalized probability distribution into another, offering

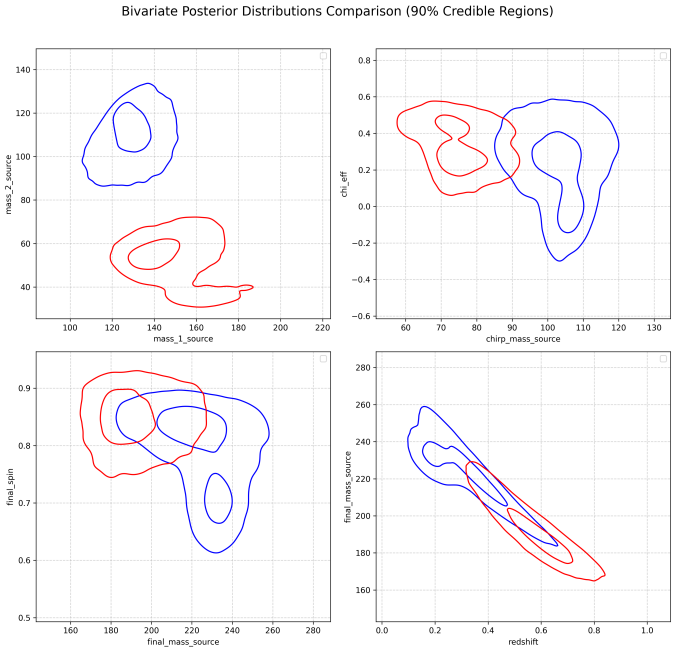

Figure 2. This figure shows the 90% credible regions for bivariate posterior distributions of GW231123, comparing NRSur7dq4 (blue) and IMRPhenomXO4a (red) waveform models. The posteriors for component masses (m_1, m_2) and chirp mass-effective spin $(\mathcal{M}_c, \chi_{\rm eff})$ are largely disjoint, revealing mutually exclusive astrophysical interpretations. For final mass and spin (M_f, χ_f) and redshift-final mass (z, M_f), partial overlap indicates some consistency. This highlights significant model-dependent systematic uncertainties in parameter inference for this event.

a geometric measure of discrepancy. The calculated distances for the selected astrophysically significant parameter pairs are presented in Table 3.

Table 3. 2D Wasserstein Distances for Selected Parameter Pairs

Parameter Pair	Wasserstein Distance
('mass_1_source', 'mass_2_source')	0.0807
('chirp_mass_source', 'chi_eff')	0.0396
('final_mass_source', 'final_spin')	0.0215
('redshift', 'final_mass_source')	0.0118

NOTE—The Wasserstein-1 distance quantifies the "cost" of transforming one normalized probability distribution into another. A larger value indicates greater dissimilarity. The results are derived from normalized posterior samples.

The Wasserstein distance is highest for the ('mass_1_source', 'mass_2_source') pair (0.0807), quantitatively confirming that the joint distribution of the component masses is the most dissimilar between

the two models. This substantial distance reinforces the severe disagreement in mass ratio observed in the univariate analysis and visually apparent in the top-left panel of Figure 2. The notable distance for the (‘chirp_mass_source‘, ‘chi_eff‘) pair (0.0396) further highlights the joint disagreement on these key intrinsic parameters. The distances for (‘final_mass_source‘, ‘final_spin‘) and (‘redshift‘, ‘final_mass_source‘) are lower, indicating somewhat greater agreement in these parameter planes compared to the component masses and intrinsic parameters, as also suggested by the partial overlap in the bottom panels of Figure 2.

3.2.2. *90% credible region overlap analysis*

The most definitive and compelling result of this study comes from the 90% credible region overlap analysis, which directly quantifies the common high-probability regions in the 2D parameter planes. The results, summarized in Table 4, provide an explicit measure of consistency, or lack thereof, and are visually supported by Figure 3.

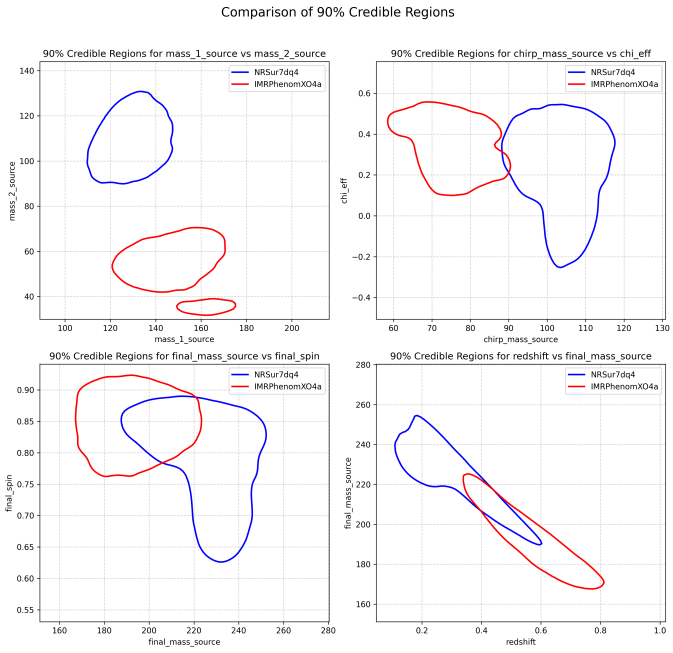

Figure 3. Comparison of 90% credible regions for GW231123 using NRSur7dq4 (blue) and IMRPhenomXO4a (red) waveform models. The component masses (m_1, m_2) and chirp mass $(\mathcal{M}_c)$ versus effective spin (χ_{eff}) show entirely disjoint or negligible overlap, demonstrating mutually exclusive astrophysical interpretations. In contrast, remnant properties (final mass and spin) and redshift versus final mass exhibit limited overlap, suggesting some consistency in the inferred final state despite differing initial conditions.

Table 4. 90% Credible Region Overlap Analysis

Parameter Pair	Overlap Metric
(‘mass_1_source‘, ‘mass_2_source‘)	0.0000
(‘chirp_mass_source‘, ‘chi_eff‘)	0.0044
(‘final_mass_source‘, ‘final_spin‘)	0.1979
(‘redshift‘, ‘final_mass_source‘)	0.2057

NOTE—This table presents the areas of the 90% credible regions for each model, their intersection area, and a Jaccard-like overlap metric defined as ‘Overlap Area / Union Area‘. A value of 0 indicates no overlap, while 1 indicates perfect overlap.

The results are unequivocal and stark. For the joint distribution of component masses, (‘mass_1_source‘, ‘mass_2_source‘), the overlap metric is precisely zero (0.0000), as shown in Table 4. This indicates that the 90% credible regions inferred by the two models are completely disjoint, occupying entirely separate volumes in this parameter space. This is not merely a shift in central values, but a fundamental disagreement on the probable range of component masses, clearly depicted in the top-left panels of both Figure 2 and Figure 3. Similarly, for the (‘chirp_mass_source‘, ‘chi_eff‘) plane, the overlap is negligible at just 0.44%, meaning less than half a percent of the high-probability region is shared between the two models for these critical intrinsic parameters.

This is a profound statement on the model-dependency of the inference. It means that an astrophysicist using `NRSur7dq4` would conclude with 90% confidence that the component masses lie in a region that an astrophysicist using `IMRPhenomXO4a` would rule out with 90% confidence, and vice-versa. The models do not merely disagree on the central values; their high-probability regions are mutually exclusive.

Interestingly, the overlap is slightly better for the remnant and extrinsic properties. The (‘final_mass_source‘, ‘final_spin‘) and (‘redshift‘, ‘final_mass_source‘) pairs show approximately 20% overlap (Table 4). This suggests that despite the radically different inferred initial conditions (component masses, redshift), the models converge to a somewhat more consistent, albeit still highly discrepant, picture of the final state. This is likely because the merger-ringdown portion of the signal, which directly informs the final mass and spin, is the most prominent feature of the data for a high-mass system like GW231123. Both models incorporate numerical relativity information for this phase and are therefore more constrained here than in the inspiral phase. However, an overlap of ∼20% still indicates significant disagreement and precludes

robust conclusions without acknowledging the model dependence.

3.3. *Discussion: Physical Origins and Astrophysical Implications*

The existence of two distinct, high-likelihood solutions for GW231123 points to a fundamental challenge in analyzing short-duration signals from high-mass binaries. As highlighted in the Introduction, for such events, the observed signal is predominantly shaped by the intricate and highly non-linear merger and ringdown phases. This is precisely the regime where differences between waveform models, arising from their distinct methodologies and approximations, are most pronounced.

3.3.1. *Physical origin of the discrepancy*

The massive discrepancy in the inferred inclination angle ($\cos\theta_{JN}$), clearly visible in Figure 1 and quantified in Table 2, is likely the key to understanding this bimodal result. The observed gravitational waveform is a complex superposition of different spherical harmonic modes. For binaries that are not face-on or have unequal masses, higher-order modes (beyond the dominant $l = |m| = 2$ mode) become significant. The relative amplitude of these modes is strongly dependent on the inclination angle. The short duration of the GW231123 signal, characteristic of high-mass events, means there is insufficient information to robustly break the degeneracy between the mass ratio and the inclination angle.

The `IMRPhenomXO4a` solution, with its high mass ratio ($q \approx 0.38$) and near face-on orientation ($\cos\theta_{JN} \approx 0.88$), fits the data by invoking a specific combination of modes consistent with that geometry. This model, being a frequency-domain phenomenological model, blends analytical approximations for the inspiral phase with calibrated fits to numerical relativity (NR) simulations for the merger and ringdown. For a system with such high masses, the early inspiral portion of the signal, however short, might be less accurately represented by the analytical approximations, potentially guiding the parameter estimation towards a specific region of the likelihood space that favors a certain degeneracy resolution.

Conversely, the `NRSur7dq4` solution, with its near-equal masses ($q \approx 0.85$) and more inclined orientation ($\cos\theta_{JN} \approx -0.25$), finds an alternative combination of modes that also provides a good fit to the data. As an NR surrogate model, `NRSur7dq4` is constructed by interpolating a suite of highly accurate NR simulations across the full inspiral, merger, and ringdown. Its strength lies in its fidelity to general relativity, particularly in the strong-field, highly dynamic regime. It is plausible that `NRSur7dq4`'s more accurate representation of higher-order modes and merger dynamics for near-equal mass, inclined systems allows it to explore and favor this alternative parameter space more readily. The differing treatment of higher-order modes and their interplay with inclination and mass ratio, especially given the short signal duration, appears to be at the heart of the observed model divergence.

3.3.2. *Astrophysical implications of the bimodal inference*

The two solutions carry vastly different implications for stellar and binary evolution, making it impossible to draw a single, coherent astrophysical picture for GW231123:

- **The `NRSur7dq4` Scenario:** With inferred component masses of $\sim$129 $M_\odot$ and $\sim$110 $M_\odot$ (Table 1), both black holes would lie deep within the upper pair-instability mass gap (roughly 65–120 $M_\odot$). The formation of such high-mass objects is highly challenging for standard stellar evolution theory, which predicts that stars in the requisite mass range are completely disrupted by pair-instability supernovae, leaving no remnant. This scenario would strongly point towards a hierarchical merger origin, where these black holes are themselves the products of previous mergers in a dense stellar environment like a globular cluster or active galactic nucleus disk.

- **The `IMRPhenomXO4a` Scenario:** This scenario involves a primary black hole of $\sim$145 $M_\odot$ (firmly above the pair-instability gap) and a secondary of $\sim$55 $M_\odot$ (below the pair-instability gap), as shown in Table 1. While still requiring a formation mechanism for the massive primary, it represents a different type of merger with a very high mass ratio. Such systems are of great interest as they are efficient emitters of higher-order gravitational-wave modes, which can provide more detailed information about the binary's geometry. Furthermore, highly asymmetric mergers can produce significant recoil kicks on the final black hole, potentially ejecting it from its host galaxy or cluster.

These two interpretations represent fundamentally different pathways for black hole formation and evolution, highlighting the critical impact of waveform model choice on astrophysical conclusions.

The quantitative demonstration of completely disjoint 90% credible regions for component masses (Table 4, Figure 3), and negligible overlap for chirp mass and effective spin, provides compelling evidence that the systematic uncertainty arising from waveform model choice is the dominant source of error for GW231123. This

precludes robust astrophysical conclusions about its intrinsic properties and leads to incompatible interpretations of the event's nature. This finding aligns with the initial hypothesis outlined in our Introduction, emphasizing that for high-mass, short-duration events, even subtle differences in model fidelity can propagate into substantial shifts in the inferred parameter space.

4. CONCLUSIONS

The accurate inference of astrophysical parameters from gravitational-wave observations is fundamental to unlocking the universe's most extreme phenomena. This study rigorously investigated the sensitivity of parameter inference for the high-mass, short-duration binary black hole event GW231123 to the choice of waveform model. We employed a comprehensive statistical framework, comparing posterior distributions derived from two distinct models, NRSur7dq4 and IMRPhenomXO4a, through univariate analyses, 2D Wasserstein distances, and critically, 90% credible region overlap.

Our analysis revealed profound and irreconcilable discrepancies in the inferred astrophysical properties of GW231123. The two waveform models led to fundamentally distinct interpretations of the event:

- **NRSur7dq4** inferred a near-equal mass binary (primary mass ≈ 129 $M_\odot$, secondary mass ≈ 110 $M_\odot$) at a redshift of approximately 0.29, with a moderately inclined viewing angle.
- **IMRPhenomXO4a** preferred a highly asymmetric binary (primary mass ≈ 145 $M_\odot$, secondary mass ≈ 55 $M_\odot$) located at a significantly higher redshift of approximately 0.58, viewed nearly face-on.

These differences were not merely shifts in central estimates but represented entirely separate high-probability regions in the parameter space. Quantitatively, the 90% credible regions for the component masses were found to be completely disjoint (0% overlap), and for the chirp mass and effective spin parameter plane, overlap was negligible (0.4%). While some parameters related to the final remnant black hole and its relation to redshift showed slightly higher (though still low, $\sim$20%) overlap, the initial conditions of the binary were found to be mutually exclusive.

The primary physical origin of these discrepancies is likely rooted in the differing treatment of higher-order gravitational-wave modes and their degeneracy with the binary's mass ratio and inclination angle, particularly prominent in the short-duration, merger-dominated signal of GW231123. The distinct modeling approaches of NRSur7dq4 (NR surrogate with high fidelity across all phases) and IMRPhenomXO4a (phenomenological model blending analytical inspiral with NR-calibrated merger/ringdown) appear to resolve these degeneracies in fundamentally different ways.

From these results, we conclude that for GW231123, and by extension likely for similar high-mass, short-duration gravitational-wave events, the systematic uncertainty arising from the choice of waveform model is the dominant source of error. This systematic error transcends statistical uncertainties, leading to incompatible interpretations of the event's properties and precluding robust astrophysical conclusions. The two inferred scenarios for GW231123 carry vastly different implications for black hole formation pathways, with NRSur7dq4 favoring a hierarchical merger origin for objects within the upper pair-instability mass gap, and IMRPhenomXO4a suggesting a highly asymmetric merger with one component below the gap.

This paper highlights the critical need for continued development and rigorous validation of gravitational-wave waveform models, with a particular focus on accurately capturing higher-order modes and merger dynamics across the full parameter space, especially for challenging high-mass, short-duration signals. Our findings underscore the importance of explicitly quantifying and reporting systematic uncertainties due to model choice in all gravitational-wave parameter inference analyses to ensure reliable astrophysical interpretations. Without addressing these profound model dependencies, the ability to draw definitive conclusions about the nature and origin of such extreme astrophysical events will remain severely limited.