

Sequence modeling of higher-order wave modes of quasi-circular, spinning, non-precessing binary black hole mergers

Victoria Tiki^{1,2,3}, **Kiet Pham**⁴, **E. A. Huerta**^{1,2,3,5,6}

¹ Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois 60439, USA

² Department of Physics, University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA

³ NCSA, University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA

⁴ School of Physics and Astronomy, University of Minnesota, 55455 Minnesota, USA

⁵ Department of Computer Science, The University of Chicago, Chicago, Illinois 60637, USA

⁶ Department of Astronomy, University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA

E-mail: vtiki2@illinois.edu

Abstract. Higher-order gravitational wave modes from quasi-circular, spinning, non-precessing binary black hole mergers encode key information about these systems' nonlinear dynamics. We model these waveforms using transformer architectures, targeting the evolution from late inspiral through ringdown. Our data is derived from the NRHybSur3dq8 surrogate model, which includes spherical harmonic modes up to $\ell \leq 4$ (excluding $(4, 0)$, $(4, \pm 1)$ and including $(5, 5)$ modes). These waveforms span mass ratios $q \leq 8$, spin components $s_{1,2}^z \in [-0.8, 0.8]$, and inclination angles $\theta \in [0, \pi]$. The model processes input data over the time interval $t \in [-5000M, -100M]$ and generates predictions for the plus and cross polarizations, (h_+, h_\times) , over the interval $t \in [-100M, 130M]$. Utilizing 16 NVIDIA A100 GPUs on the Delta supercomputer, we trained the transformer model in 15 hours on over 14 million samples. The model's performance was evaluated on a test dataset of 840,000 samples, achieving mean and median overlap scores of 0.996 and 0.997, respectively, relative to the surrogate-based ground truth signals. We further benchmark the model on numerical relativity waveforms from the SXS catalog, finding that it generalizes well to out-of-distribution systems, capable of reproducing the dynamics of systems with mass ratios up to $q = 15$ and spin magnitudes up to 0.998, with a median overlap of 0.969 across 521 NR waveforms and up to 0.998 in face-on/off configurations. These results demonstrate that transformer-based models can capture the nonlinear dynamics of binary black hole mergers with high accuracy, even outside the surrogate training domain, enabling fast sequence modeling of higher-order wave modes.

1. Introduction

Artificial intelligence (AI) and machine learning (ML) are playing an increasingly central role in gravitational-wave (GW) astrophysics, offering powerful tools for tasks ranging from signal detection to population inference. Comprehensive reviews of these developments can be found in [1,2]. These approaches have enabled the development of low-latency search pipelines [3], robust glitch classification [4, 5, 6], and accurate denoising methods [7]. Deep learning techniques have matched or exceeded traditional matched-filter performance in detecting binary black hole mergers [8, 9, 10, 11, 12, 13], and enabled efficient parameter estimation [14,15] and population-level inference [16] orders of magnitude faster than classical methods.

Sequence modeling of gravitational waveforms has emerged as a promising direction for accelerating waveform generation in gravitational wave astrophysics. Several studies have explored machine learning approaches that map physical parameters to waveforms, particularly focusing on the merger and ringdown regimes. For example, Lee et al. [17] trained a fully connected network on SEOBNRv4 waveforms [18] to model non-spinning BBH mergers, achieving overlap scores above 99.9% for the $\ell = m = 2$ mode. This approach accelerated waveform generation and generalized well across parameter space, but was limited to non-spinning systems. Similarly, Khan and Green [19] used neural networks to build surrogate models of gravitational waveforms from parameters, achieving mismatches as low as 2×10^{-5} in aligned-spin systems. This work confirmed that simple feed-forward networks could rival interpolation-based surrogates in accuracy with reduced computational cost. These works exemplify the success of ML-based surrogate modeling but do not directly address the forecasting problem.

In contrast to parameter-to-waveform models, recent work has explored directly forecasting future gravitational wave signals from earlier parts of the waveform. This task is inherently sequential and lends itself to architectures such as recurrent neural networks or transformer-based architectures. A notable example is [20], which introduced a transformer-based model capable of forecasting the late inspiral, merger, and ringdown of signals from spinning, quasi-circular, non-precessing binary black hole mergers. The model was trained on over 1.5 million waveforms generated using the NRHybSur3dq8 surrogate [21], and learned to predict late inspiral, merger, and ringdown of the waveform with generally high overlap scores, though with overlaps as low as 0.6 across the test set. The approach was limited to the dominant $\ell = |m| = 2$ mode. A complementary approach was introduced in [22], where a lightweight LSTM network predicted the merger and ringdown using a multi-step sequence strategy. Despite using a much smaller training set based on the IMRPhenomD model [23], the LSTM [24] achieved over 99.6% mean overlap demonstrating that compact sequence models can generalize well from limited data—though their scope remained constrained to aligned spins and the dominant harmonic mode. More recently, BHP2NRMLSur [25] introduced a continuous-time neural network that maps perturbation theory waveforms to full NR waveforms across multiple modes and spins. While it achieves high accuracy, it differs

from ours in that it transforms approximate templates rather than forecasting waveform evolution from earlier signal segments.

In this work, we extend waveform forecasting to a significantly more complex regime: predicting both polarizations (h_+ , h_\times) across higher order harmonic modes, using only early inspiral information from quasi-circular, spinning, non-precessing binary black hole systems. This regime lies deep in the nonlinear sector of general relativity, where spin effects, mode coupling, and gravitational recoil are prominent, and analytical approximations break down. Higher-order modes are particularly important in systems with unequal masses, high spins, or edge-on orientations, where they enhance detection and parameter estimation accuracy [26, 27], but also introduce substantial modeling challenges [28]. Whereas previous studies [17, 19, 20, 22] focused on the quadrupole mode, we model both GW polarizations (h_+ , h_\times) and include all spherical harmonic

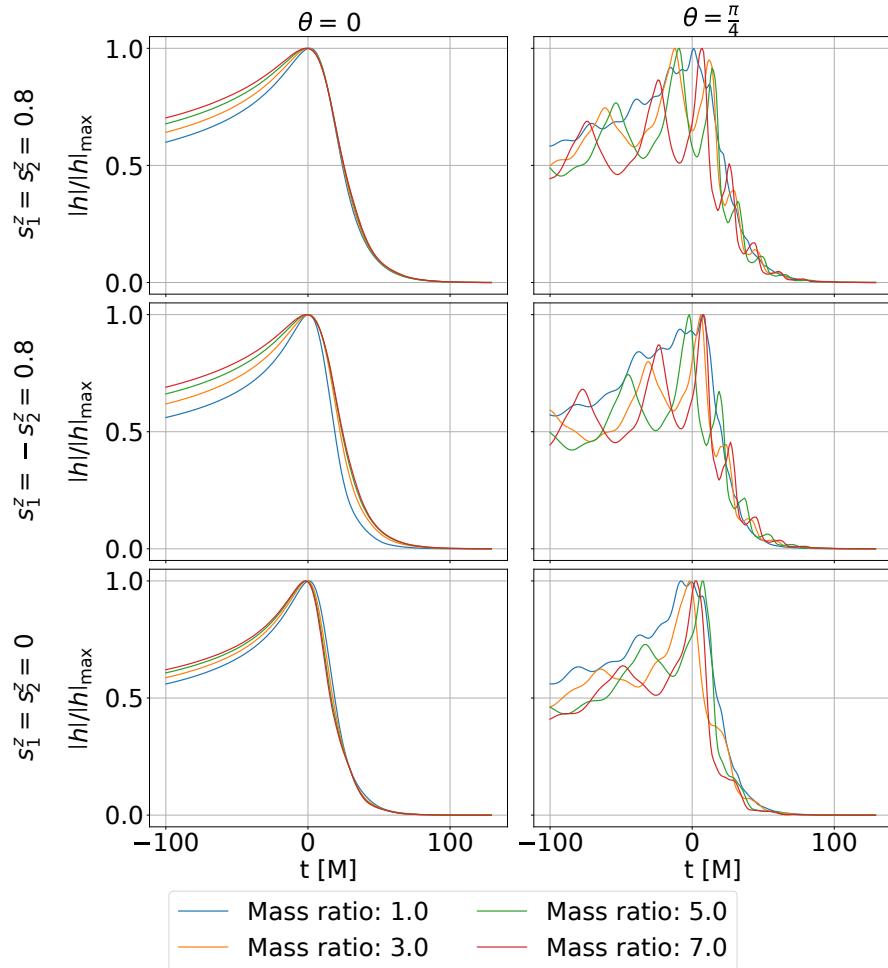


Figure 1: Normalized amplitude $|h| / |h|_{\max}$ of gravitational waves for different binary black hole configurations. Notice the impact of higher-order wave modes for inclination angles $\theta = 0$ (left column) and $\theta = \pi/4$ (right column) in terms of amplitude modulations, and the time at which the waveform amplitude peaks.

modes up to $\ell = 4$ (including $(\ell, m) = (5, 5)$ and excluding $(4, 0)$ and $(4, \pm 1)$, which are not reliably provided by the surrogate model used for training). Figure 1 illustrates the increased complexity introduced by these higher-order modes. The left panel shows representative waveforms from prior studies, where the amplitude evolves smoothly and is dominated by the quadrupole mode. In contrast, our target waveforms (right panel) show non-monotonic structure, strong modulation with inclination angle, and a shifted peak amplitude, demonstrating the richer dynamics that higher-order modes encode.

The key application of this transformer-based model is to serve as a fast surrogate for numerical relativity (NR) simulations, trained on physically realistic waveforms and designed to reflect key domain constraints such as polarization structure and sequence causality. Rather than functioning as a detection tool, it is designed for systematic modeling of higher-order modes, spin effects, and mode mixing across the BBH parameter space. This enables consistency tests of NR waveforms, waveform interpolation where NR is sparse or unavailable, and rapid hypothesis testing in gravitational theory. In doing so, it accelerates the development and validation of analytical waveform models and supports precision studies of strong-field gravity.

To implement this model, we introduce a set of architecture and training design choices that enable accurate and efficient sequence modeling of binary black hole waveforms:

1. Data generation and parameter space sampling We use the NRHybSur3dq8 surrogate model to generate gravitational waveforms with higher-order modes up to $\ell = 4$ (including $(5, 5)$ but excluding the $(4, 0)$, $(4, \pm 1)$ modes, which the surrogate does not reliably support). The parameter space covers mass ratios $q \in [1, 8]$, spins $s_{1,2}^z \in [-0.8, 0.8]$, and inclination angles $\theta \in [0, \pi]$. To match this domain, we developed HPC-enabled sampling tools to densely cover the space, yielding over 14 million waveforms for training.

2. Transformer architecture and distributed training. To handle the vanishing h_\times in edge-on mergers ($\theta = \pi/2$), we include conditionally activated processing layers that improve accuracy when $h_\times \approx 0$ due to the symmetry of the system. The model forecasts both GW polarizations, (h_+, h_\times) , from inspiral through ringdown as distinct channels, using a causal transformer that enforces smoothness and symmetry for physical consistency. Training using data parallelism across 16 NVIDIA A100 GPUs led to convergence within 15 hours on the Delta supercomputer.

3. Inference. Inference on 840,000 test waveforms was completed in under 5 hours using 1 NVIDIA V100 GPU on NCSA’s HAL system. The model achieved mean and median overlap scores of 0.996 and 0.997, respectively, when benchmarked against surrogate-model ground truth signals. When evaluated on a small set of numerical relativity waveforms beyond the training domain, the model achieves a mean overlap of 0.969 and a median of 0.902, which increased to a median of 0.998 and a minimum of 0.95 when NR waveforms were constrained to face-on binaries.

4. Exploratory interpretability studies. To better understand how the transformer achieves accurate predictions, we performed analyses examining the

influence of key binary parameters—mass ratio, spin, and inclination angle—on model behavior. These exploratory studies provide insight into which waveform features the model finds predictive and how physical symmetries may be reflected in its learned representations. All code and analysis scripts are available at GitHub [29] to ensure transparency and reproducibility.

This paper is structured as follows: Section 2 describes the data used to train, validate, and test the transformer model, and outlines the model design, training and inference methods. In Section 3, we present our findings. Future directions are discussed in Section 4. The numerical relativity waveforms employed in the NRHybSur3dq8 surrogate model use geometric units, where $G = c = 1$ and M represents the total mass of the binary black-hole merger system. Consequently, all waveform datasets follow this convention, with time measured in units of M , where M is the total mass and $1M_\odot = 4.93 \times 10^{-6}$ seconds.

2. Methods

Here we describe the datasets used for this work, the approaches used to design and train our transformer model, and methods to test its predictive capabilities for sequence modeling.

2.1. Data

Three independent datasets of modeled waveforms were produced for training, validation, and testing of the transformer using the numerical relativity surrogate model NRHybSur3dq8 [21]. These waveforms describe the gravitational emission from quasi-circular, spinning, non-precessing binary black hole mergers, encompassing the inspiral, merger, and ringdown phases. The datasets are constrained within the valid parameter space for this NR surrogate model, i.e., mass ratios $q \leq 8$ and individual spins $|s_{\{1,2\}}^z| \leq 0.8$. The gravitational wave strain time-series $h(t, \theta, \phi)$ can be expressed as a sum of spin-weighted spherical harmonic modes, h , on the 2-sphere [30]:

$$h(t, \theta, \phi) = \sum_{l=2}^{\infty} \sum_{m=-l}^{l} h_{lm}(t)^{-2} Y_{lm}(\theta, \phi), \quad (1)$$

where ${}^{-2} Y_{lm}$ are the spin-weight-2 spherical harmonics, θ is the inclination angle between the orbital angular momentum of the binary and the line of sight to the detector, and ϕ is the initial binary phase, which we set to zero. We include the higher-order wave modes $\ell \leq 4$ and $(\ell, m) = (5, 5)$, except for $(4, 0)$, $(4, -1)$ and $(4, 1)$, since NRHybSur3dq8 only provides reliable extrapolation for these specific higher-order modes. The waveforms cover the time span $t = [-5000 M, 130 M]$.

Training dataset The training dataset consists of 14,440,761 waveforms, generated by sampling the mass ratio $q \in [1, 8]$ in steps of $\delta q = 0.1$; individual spins $s_i^z \in [-0.8, 0.8]$

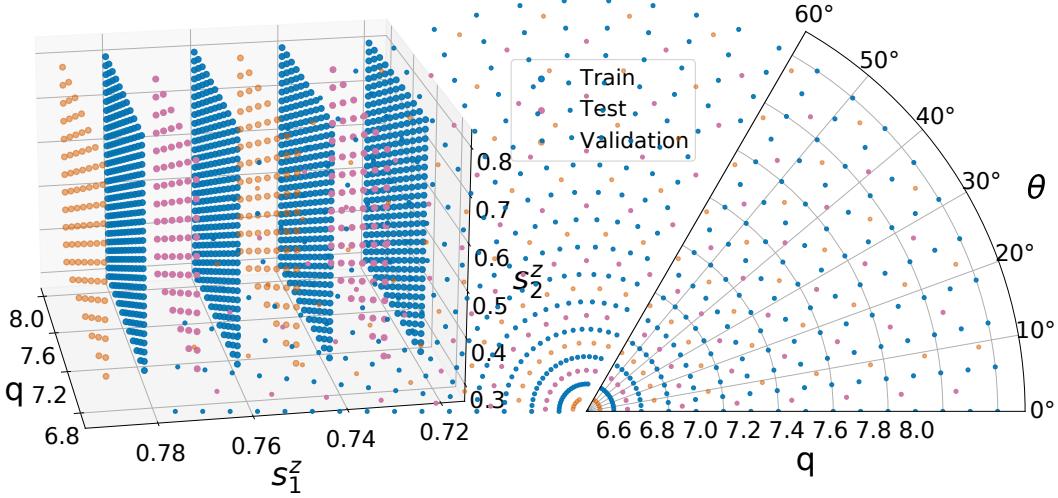


Figure 2: Sampling of the signal manifold defined by the mass ratio $q \in [1, 8]$, aligned spin components $\xi_{1,2}^z \in [-0.8, 0.8]$, and inclination angle $\theta \in [0, \pi]$ to generate three independent datasets for training, validation, and testing of our transformer models. The training set uses a uniform grid with step sizes $\delta q = 0.1$, $\delta \xi = 0.02$, and $\delta \theta = \pi/29$. The validation and test sets are constructed by interleaving values between training grid points to avoid overlap and evaluate model generalization. For visual clarity, the figure shows a reduced interval of the full signal manifold.

in steps of $\delta \xi^z = 0.02$; and inclination angle $\theta \in [0, \pi]$ in steps of $\delta \theta = \pi/29$.

Test and validation datasets Each of these datasets consists of 840,000 waveforms generated by sampling values spaced evenly between the training set values. Figure 2 illustrates the sampling strategy for generating the training, validation, and test sets, demonstrating that these datasets are independent and do not overlap.

Sampling and encoder-decoder split We reduce the temporal resolution of the data by subsampling the original input sequences, selecting every other timestep. This approach reduces computational complexity. The modeled waveform is then split such that the transformer’s encoder module receives the early inspiral sequence, consisting of time steps $t = [-5000 \text{ M}, -100 \text{ M}]$. The decoder module receives the inspiral, merger, and ringdown segments, made up of time steps $t = [-101 \text{ M}, 129 \text{ M}]$. The target waveform consists of the gravitational wave segment $t = [-100 \text{ M}, 130 \text{ M}]$, i.e., the decoder target is one timestep ahead of the decoder input, as shown in Figure 3.

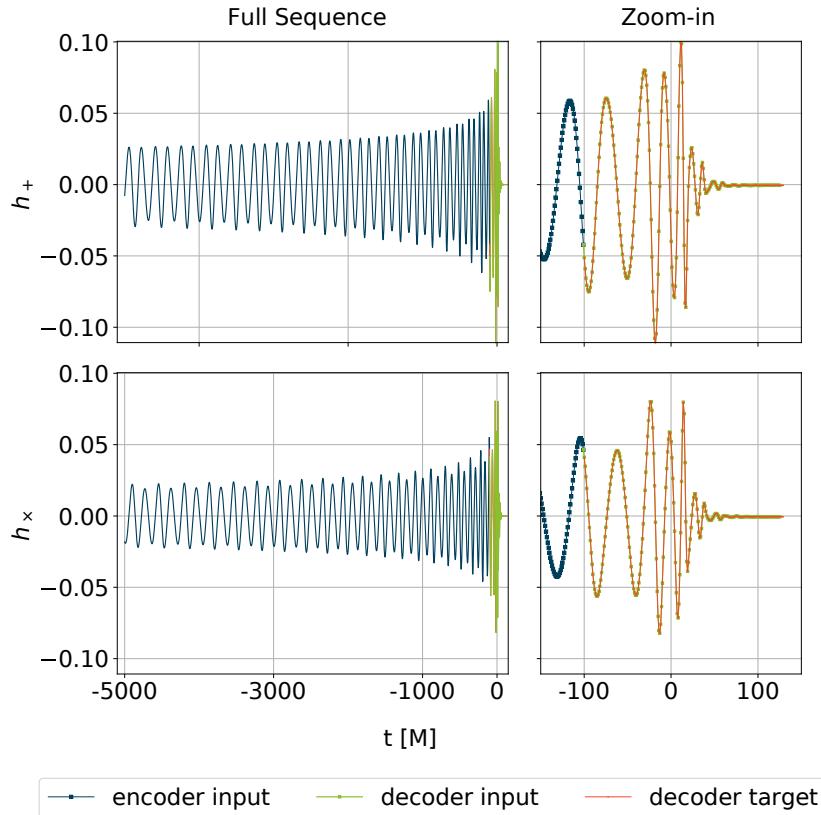


Figure 3: Gravitational wave sequence and separation into encoder input, decoder input, and decoder target for training. Note that the decoder target complements the encoder input exactly, reflecting the transformer’s intended function of completing the signal. The decoder input sequence is shifted to the left by one timestep $\Delta t = 1$ M with respect to the decoder target sequence and thus overlaps the encoder input by one timestep.

2.2. Architecture

We extend the original architecture of the transformer model introduced in Ref. [31], which leverages attention mechanisms to weigh different segments of the input sequence and improve performance in sequence prediction tasks. Transformer-based architectures have recently shown strong performance in physical sciences, including stellar light curve analysis [32], gamma-ray burst afterglow modeling [33], and turbulent flow prediction [34]. Here, we have modified the model to incorporate properties of gravitational waves. These modifications include an embedding scheme that encodes both polarizations, (h_+, h_x) , as well as an h mask that automatically activates additional layers for signals corresponding to edge-on mergers, which are difficult to resolve due to vanishing h .

Figure 4 schematically illustrates these modifications. Below, we describe the functionality of each component of the transformer as it processes the gravitational wave sequences. We represent the amplitudes of the input sequences for the encoder and decoder modules by p_x where p indicates the position within the sequence of length n . Note that the sequence length n varies for encoder and decoder inputs denoted as

n_{enc} and n_{dec} , respectively.

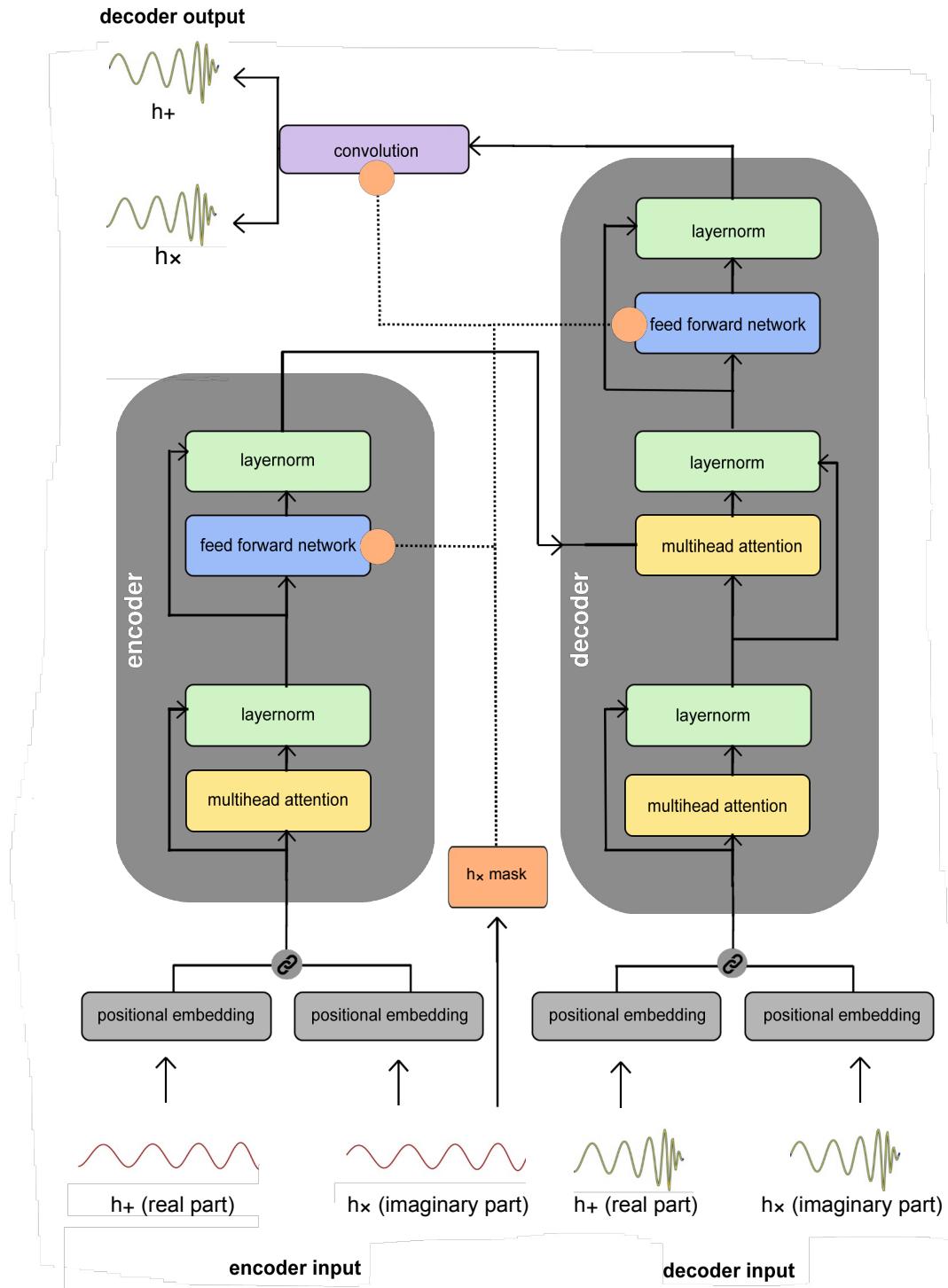


Figure 4: Schematic representation of transformer modules. The input time-series data for (h_+, h_x) is processed concurrently to generate corresponding predictions (h, h_x) . This approach ensures the model maintains physical consistency in forecasting both time-series data.

2.2.1. Embedding The transformer architecture is not intrinsically sensitive to the order of tokens within the input sequence. However, for sequence prediction, information about token order is critical. This issue is addressed by introducing non-trainable positional encoding, which embeds the input sequence into a higher-dimensional space. Through positional encoding, the input sequence x_p is transformed into a sequence of d_{embed} -dimensional vectors, each taking the form:

$$X_p = (x_p, P E(p, 1), P E(p, 2), \dots, P E(p, d_{\text{embed}})),$$

where $P E(p, i)$ denotes the positional encoding at position p and dimension i . This transformation carries information about the original token's position, and we represent this transformed set of vectors as a matrix $X \in \mathbb{R}^{n \times d_{\text{embed}}}$. The embedding function $P E$, where p is the position of any token within the original sequence and i is the dimension within the positional encoding vector, takes the form:

$$P E(p, 2i) = \sin \frac{p}{10000^{i/d_{\text{embed}}}}, \quad (2)$$

$$P E(p, 2i + 1) = \cos \frac{p}{10000^{i/d_{\text{embed}}}}. \quad (3)$$

As demonstrated in Ref. [31], this type of sinusoidal positional encoding enables the model to interpolate and extrapolate positional encodings for unseen positions, providing better generalization than other embedding schemes. Building upon previous transformer models [20] for gravitational wave forecasting, this work facilitates the processing of complex-valued waveforms: Each input sequence is split into real and imaginary components, corresponding to plus, h_+ , and cross, h_\times , polarizations, and independently embedded using the procedure outlined above. These embeddings are then concatenated along the embedding dimension, effectively doubling the input's dimensionality for subsequent processing by the encoder and decoder modules.

2.2.2. Encoder The multi-head self-attention mechanism is designed to allow each position in the input sequence to attend to all positions within the same sequence. For each of the h ($= 10$) attention heads, the computation reads

$$\text{Attention}(Q^j, K^j, V^j) = \text{softmax} \frac{Q^j (K^j)^T}{\sqrt{d_k}} V^j. \quad (4)$$

Here, $Q^j \in \mathbb{R}^{n_{\text{enc}} \times d_q}$, $K^j \in \mathbb{R}^{n_{\text{enc}} \times d_k}$, and $V^j \in \mathbb{R}^{n_{\text{enc}} \times d_v}$ denote the query, key, and value matrices, respectively, defined as linear transformations of the input matrix X :

$$Q^j = XW_Q^j, \quad K^j = XW_K^j, \quad V^j = XW_V^j, \quad (5)$$

with weight matrices $W_Q^j \in \mathbb{R}^{d_{\text{embed}} \times d_q}$, $W_K^j \in \mathbb{R}^{d_{\text{embed}} \times d_k}$, and $W_V^j \in \mathbb{R}^{d_{\text{embed}} \times d_v}$. In this work, $d_k = d_q = d_v = d_{\text{embed}}/h$.

The matrix QK^T functions as a measure of similarity between keys and queries. The softmax activation operates on the rows of QK and returns weights corresponding to values V . The output of each attention head is concatenated along the last dimension and then linearly transformed:

$$\text{Attention}(Q, K, V) = \text{concat} \quad \text{Attention}(Q^1, K^1, V^1), \dots, \text{Attention}(Q^h, K^h, V^h) \quad W^0, \quad (6)$$

with $W^0 \in \mathbb{R}^{d_{\text{embed}} \times d_{\text{embed}}}$, enabling the model to integrate information across different representational subspaces. The resulting attention matrix has the same shape as the input X . After the attention mechanism, each position's output undergoes processing by a position-wise feed-forward network, which applies two linear transformations with ReLU activations. Each sublayer (self-attention and feed-forward network) is enclosed by a residual connection, followed by layer normalization. This configuration aids in stabilizing training.

2.2.3. Decoder The decoder receives both the encoder output and the decoder input, which is embedded using the same method as the encoder input. A masked multi-head self-attention mechanism is applied to the decoder input only. This mechanism prevents each position in the decoder input from attending to subsequent positions in the sequence, preserving causality and enabling autoregressive predictions during inference. The operation is defined as:

$$\text{MaskedAttention}(Q^j, K^j, V^j) = \text{softmax} \quad \frac{Q^j (K^j)^T + M}{\sqrt{d_k}} \quad V^j, \quad (7)$$

where M is a mask matrix that applies negative infinity to positions not to be attended to, ensuring that the softmax operation assigns them zero weight. As before, Q^j , K^j , and V^j are the query, key, and value matrices derived from linear transformations on the decoder's input. The dimensionalities of these matrices are the same as those of the encoder self-attention matrices, with the exception of the sequence length n , which varies for encoder inputs and decoder inputs. Outputs from each attention head are again concatenated and linearly transformed. These transformations leave the shape of the decoder input unchanged.

Following the self-attention layer, the cross-attention layer allows the decoder to focus on relevant positions in the encoder output sequence. The queries Q stem from the previous decoder layer (denoted as $Y \in \mathbb{R}^{n_{\text{dec}} \times d_{\text{embed}}}$), while the keys K and values V are derived from the encoder output X :

$$\text{CrossAttention}(Q^j, K^j, V^j) = \text{softmax} \quad \frac{Q^j (K^j)^T}{\sqrt{d_k}} \quad V^j, \quad (8)$$

with

$$Q^j = YW_Q^j \quad K^j = XW_K^j \quad V^j = XW_V^j. \quad (9)$$

This operation preserves the shape of the queries in this case the decoder input, ensuring the sequence length of the decoder output matches that of the decoder input. QK^T serves as a measure of similarity between keys and queries. The array returned by the softmax activation function is of shape $R^{n_{\text{dec}} \times n_{\text{enc}}}$ and serves as an array of weights for the value array V . Figure 12 presents an instructive visualization of these weights for a choice of input waveform and attention head.

The subsequent processing after the application of the attention mechanisms mirrors that of the encoder module: each position's output is fed to a feed-forward network, residual connections, and layer norms are applied after each attention layer and the feed-forward network to stabilize training. Finally, a 1D convolution is applied to the decoder output to produce predictions for the two time-series, (h^h_x) polarizations, respectively.

2.2.4. h_x Mask Empirically, we find that separating waveforms with null cross-polarization leads to slightly higher reconstruction accuracy. To facilitate this, we apply a mask to identify and filter waveforms with inclination angle $\theta = \pi/2$, as these exhibit null cross-polarization. This mask allows us to modify the final convolutional layer and feed-forward layers, which are implemented separately for these waveforms to better capture their distinct waveform structure. The remaining transformations in the model are consistent across all waveforms, ensuring uniform processing while accommodating the distinct features of the null cross-polarization case. The quantitative impact of this mask is outlined in section 3.5.

2.3. Training and Inference

For our transformer model, we used an embedding dimension of 160 (this includes both the contributions from the real and the imaginary part), a feed-forward dimension of 80, and 10 attention heads. The model is trained on a dataset of 14,440,761 waveforms, each consisting of 5,130 total timesteps, with only 362,722 trainable parameters. This results in a high data-to-parameter ratio, which may improve generalization capabilities and help prevent overfitting, as the model is exposed to orders of magnitude more training samples relative to its size.

This model was trained on the NCSA Delta system, using 4 quad NVIDIA A100 GPU nodes. We employed a batch size of 16 and trained the model using the Adam optimizer [35] with a mean squared error (MSE) loss metric. The initial learning rate was set to 0.001, with decay applied during training. Additional dropout or weight decay was not required. Training was distributed across all GPUs using PyTorch's DistributedDataParallel (DDP) framework [36]. The model was trained for 30 epochs and reached convergence in around 15 hours.

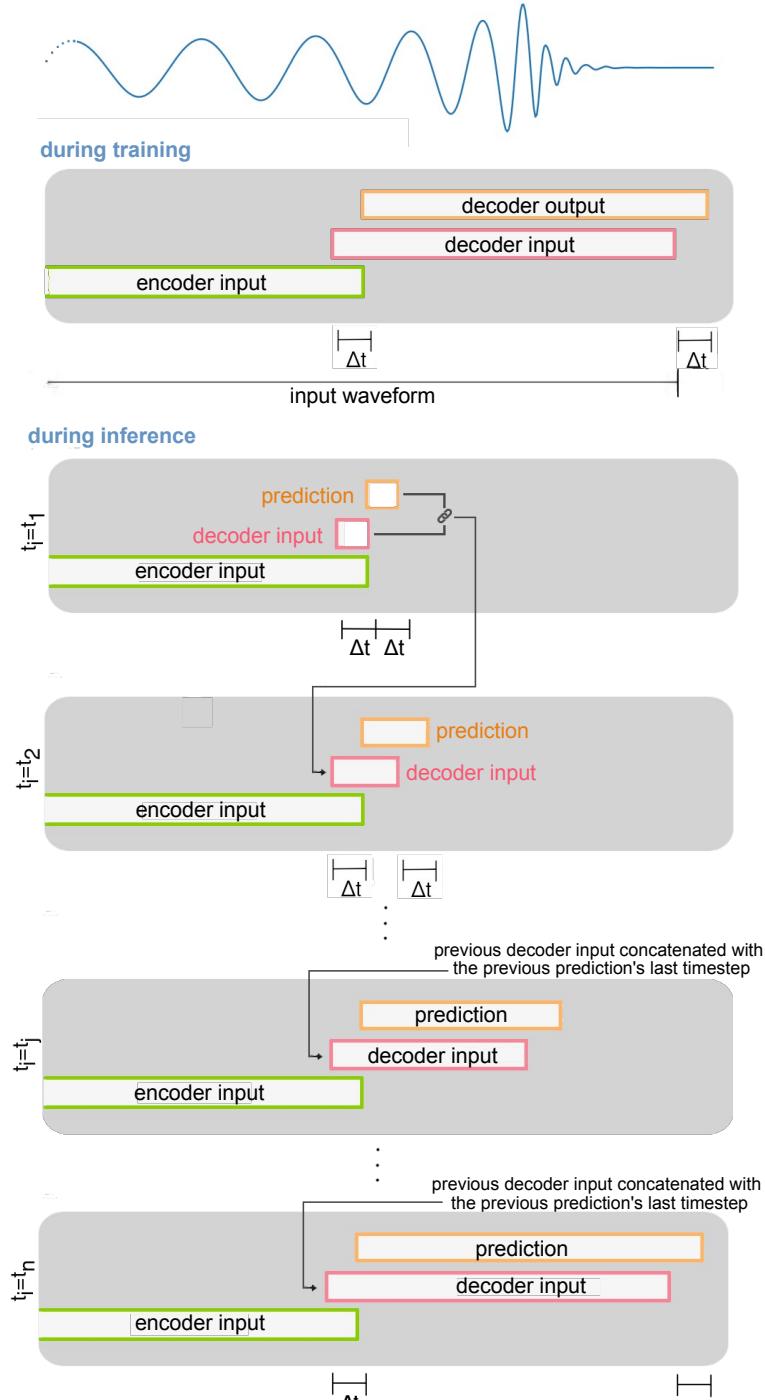


Figure 5: During training, a single decoder output sequence is generated. The attention mechanism uses an attention mask to ensure that each predicted timestep within the decoder output is based on the entire encoder input sequence and the preceding elements in the decoder input. During inference, autoregressive predictions are generated sequentially: at each step, the model produces an output sequence, from which only the last timestep is taken. This output is concatenated with the existing decoder input to form a new input for subsequent predictions. This iterative process continues until the desired sequence is fully generated.

Inference on 840,000 test waveforms was completed in under 5 hours using 1 NVIDIA V100 GPU on NCSA’s HAL system. During training, the model processes entire sequences at once¹, by contrast, proceeds via sequential autoregressive generation, with each output timestep used to predict the next timestep, which is in turn fed back into the transformer decoder module. The functional differences between training and inference are illustrated in Figure 4. During inference, the encoder outputs remain constant for each input signal and are therefore computed only once at the beginning of the autoregressive prediction. This approach substantially reduces processing time.

2.4. Forecasting Efficiency

To evaluate the inference performance of our transformer-based sequence model for gravitational wave signals, we benchmarked it against the surrogate waveform model NRHybSur3dq8, focusing on waveform generation latency. We selected 50 representative test points within the binary black hole parameter space using a Halton sequence, which ensures uniform coverage and reduces clustering compared to purely random sampling[37]. At each point, we generated waveforms 10 times with both models and recorded the per-point average. The surrogate model was run on a single AMD Milan CPU core on the NCSA Delta cluster¹, while the transformer model was benchmarked on a single NVIDIA A100 GPU on the same cluster. On average, the surrogate generated one waveform in 151 ms (6.6 waveforms/s), and the transformer completed inference for one waveform in 18 ms (55.6 waveforms/s).

The surrogate model reconstructs waveforms through interpolation of precomputed numerical relativity data, making it computationally lightweight and efficient within its training domain. In contrast, the transformer performs autoregressive sequence forecasting based on learned dynamics of strong-field spacetime evolution, enabling it to generalize beyond the training distribution. While this introduces greater computational complexity per forecast, it also allows the transformer to adapt to previously unseen input conditions. The transformer additionally benefits from GPU-optimized tensor operations and efficient batch processing, whereas the surrogate is more memory-bound and exhibits diminishing returns when parallelized beyond core saturation.

3. Results

We evaluated the predictive accuracy of our transformer-based sequence model on a test set comprising 840,000 waveforms spanning the full extent of the targeted signal

¹For the surrogate, performance remained consistent for waveform start times as early as -12000M, consistent with prior benchmarks [38].

²Assuming ideal parallel scaling on one Delta node consisting of 64 AMD EPYC 7763 “Milan” processors with 64 cores and 4 NVIDIA A100 GPUs, these single-instance results suggest a maximum throughput of approximately 422 waveforms per second for the surrogate using all 64 CPU cores, and 222 waveforms per second for the transformer using all 4 NVIDIA A100 GPUs.

manifold. The accuracy of generated waveforms is quantified using the *overlap* metric, O , which captures the normalized inner product between predicted and true waveforms, maximized over a time shift:

$$O(h_{\text{true}}, h_{\text{pred}}) = \max_{t_c} \rho \frac{\langle h_{\text{true}}, h_{\text{pred}}[t_c] \rangle}{\sqrt{\langle h_{\text{true}}, h_{\text{true}} \rangle \langle h_{\text{pred}}[t_c], h_{\text{pred}}[t_c] \rangle}} , \quad (10)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product over the complex-valued waveform time series, and $h_{\text{pred}}[t_c]$ represents a version of the predicted waveform shifted by t_c . The overlap $O \in [-1, 1]$ serves as a surrogate for waveform similarity, with $O = 1$ indicating perfect agreement.

3.1. Visual Inspection of Predictions

Figure 6 presents a small sample of model outputs across a range of parameter regimes, compared to the surrogate-based ground truth. The figure includes: (i) two randomly selected waveforms, (ii) one with relatively low overlap, and (iii) one at $\theta = \pi/2$, where $h_x = 0$. These examples provide a visual reference for the range of prediction quality observed across the dataset. For an interactive gallery of more predicted and true waveforms, visit [29].

3.2. Overall Performance Across the Signal Manifold

Figure 7 (top panel) summarizes the distribution of overlap values across the surrogate-based test set. The model demonstrates high predictive fidelity: no overlap value falls below 0.85 across the full test set, fewer than 0.003% of waveforms yield overlaps below 0.90, and only 7.15% fall below 0.99, indicating that most predictions align closely with the ground truth. The mean and median overlap values are 0.996 and 0.997, respectively, reflecting high predictive accuracy across the test set.

3.3. Performance on Numerical Relativity Waveforms

We conducted an additional benchmark using numerical relativity waveforms from the SXS catalog[39, 40, 41]. Specifically, we selected all available quasi-circular, non-precessing black hole merger simulations, and reconstructed the full strain signal (excluding memory modes) using all available spherical harmonic modes. While our model was trained only on a subset of modes (see section 2.1) and within a limited parameter space ($q \leq 8$ and $|s_{\{1,2\}}^z| \leq 0.8$), the NR dataset spans a broader range (including systems with $q \leq 15$ and $|s_{\{1,2\}}^z| \leq 1.0$). We project each of the simulations using a random inclination angle $\theta \in [0, \pi]$. We retained only those waveforms that extend to at least $-5000M$ before merger to match the encoder input requirements, leaving 521 waveforms.

Despite the mismatch in training domain and harmonic content, the model generalizes reasonably well we find a median overlap of $O = 0.969$ and a mean overlap

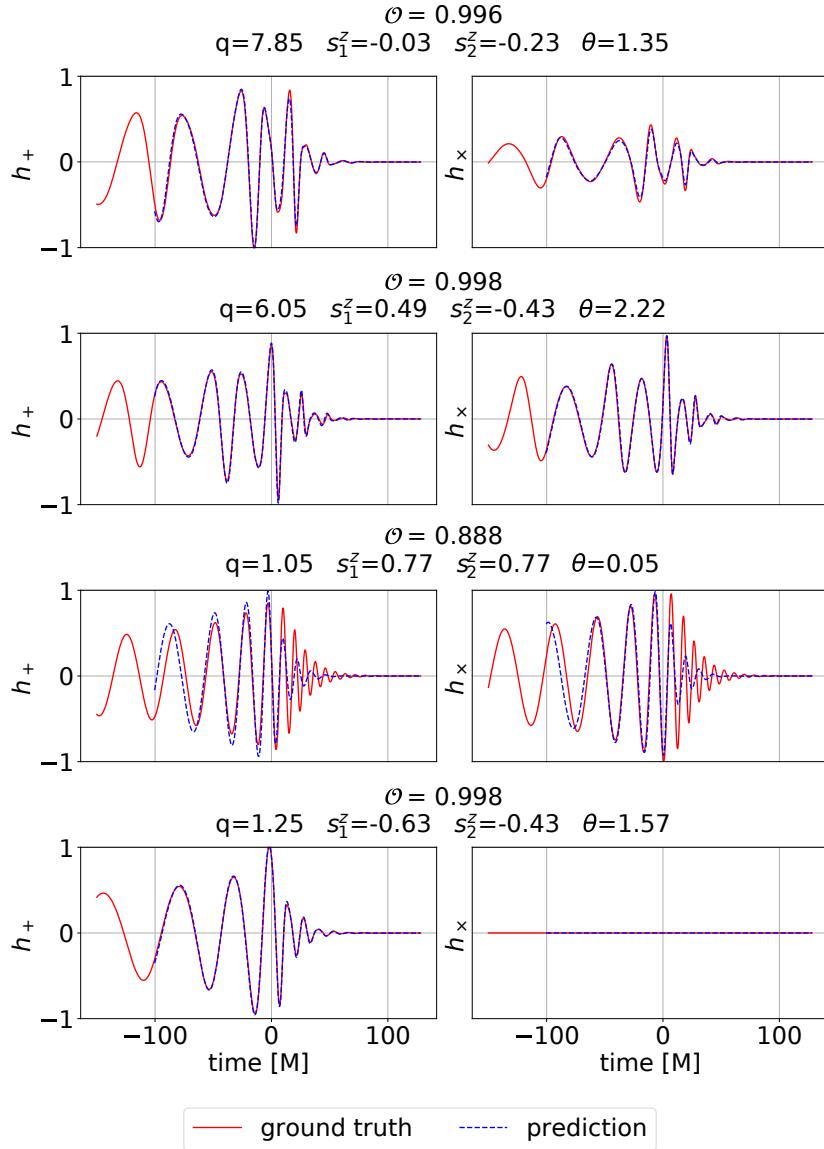


Figure 6: True and predicted waveforms for selected $(q, s_1^z, s_2^z, \theta)$ values. The first two rows represent random samples, the third a predicted waveform with one of the lowest overlap values in the test set, and the last a random sample with $\theta = \pi/2$.

of $O = 0.902$ between the transformer predictions and the full NR waveforms. As expected, these scores are lower than the surrogate-based test set, but remain high given the broader parameter coverage.

We find that low-overlap outliers are broadly distributed across inclination angles, but are absent near face-on or face-off configurations ($\theta \approx 0$ or π). Restricting to systems within $\pi/4$ of those orientations raises the median overlap to 0.990; further narrowing to $\pi/8$ yields a median of 0.998 and a minimum above 0.95—demonstrating exceptional fidelity in detector-favorable regimes. Notably, at any inclination angle, lower overlaps are not concentrated in out-of-distribution regions.

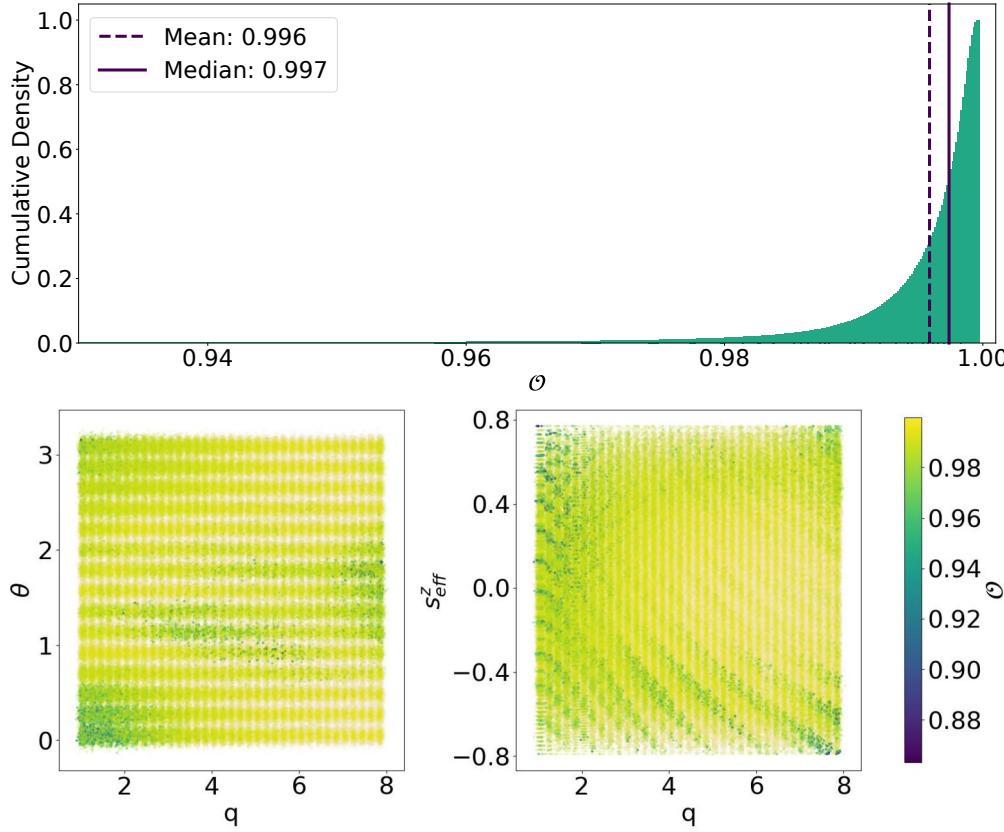


Figure 7: Top panel: Histogram of overlaps. Bottom panels: Overlap \mathcal{O} as a function of merger parameters. Point opacity is inversely proportional to overlap value, with overlaps close to 1 rendered fully transparent to highlight poor overlaps that might otherwise be obscured. A slight random jitter has been added for the same reason.

For a baseline comparison to our transformer-based forecasting model, we consider another machine learning model that reconstructs gravitational waveforms directly from source parameters. While most machine learning models for gravitational waveforms report performance on surrogate data or synthetic signals, few are evaluated directly on full numerical relativity waveforms. One exception is NRSurNN3dq4 [42], a neural network surrogate trained on the $(2, 2)$ mode of aligned-spin, non-precessing BBH waveforms with mass ratios up to $q = 6$ and dimensionless spin components $\in [-0.99, 0.99]$, using the NRSur7dq4 surrogate [43]. was then fine-tuned and evaluated on a filtered set of 381 SXS waveforms within the same parameter range, achieving a mean overlap of 0.995 and a minimum overlap of 0.979 on a held-out NR test set. Their model differs from ours in mode content, parameter space, and modeling objective (parameter-to-waveform mapping versus time-series forecasting).

Our results suggest that the model retains considerable generalization capacity even when applied to physical systems and mode structures beyond those it was explicitly trained on.

3.4. Dependence on Physical Parameters

To examine how predictive accuracy on the surrogate-based test set varies across the parameter space, we first visualize overlap scores as a function of key physical parameters. The bottom panels of Figure 7 display two scatter plots: one mapping overlap against mass ratio and inclination angle (q, θ), and the other against mass ratio and effective spin (q, s_{eff}^z). In both plots, point opacity is inversely scaled with overlap value, making regions of lower reconstruction accuracy visually prominent. These visualizations reveal several trends:

- **Mass ratio (q):** Degradations are observed at the edges of the parameter space for $q \sim \{1, 8\}$.
- **Inclination angle (θ):** Reduced performance is localized near $\theta \sim \{0, \pi/2\}$, corresponding to face-on and edge-on configurations, respectively.
- **Effective Spin (s_{eff}^z):** Defined as

$$s_{\text{eff}}^z = \frac{qs_1^z + s_2^z}{1 + q}, \quad (11)$$

a standard parameter used in gravitational-wave population modeling [44]. Highly anti-aligned spins (i.e., lower values of the spin component of the primary black hole) are associated with a reduction in reconstruction quality, with these regions visible as distinctive green bands in the bottom right panel of Figure 7.

We provide additional results in [Appendix A](#), where we study the behavior of the transformer by partitioning the signal manifold into fixed spin slices across the (q, θ) plane.

3.5. Importance of the $h \times$ Masking Mechanism

An important component of this model is the application of a conditional masking mechanism that affects edge-on orientations where the cross polarization $h \times \approx 0$. Ablation tests excluding this mechanism resulted in a decrease in average overlap to 0.994, and a drop in minimum overlap to 0.585 on the surrogate-based test set. These results underscore the importance of the $h \times$ mask for maintaining robustness across observational scenarios. Excluding this conditional masking mechanism yields a $\sim 15\%$ inference speedup, offering a trade-off that can be tuned based on application requirements.

3.6. Exploratory Interpretability Studies

These analyses aim to identify which waveform features the model relies on most during prediction. While interpretability results may reflect aspects of the input encoding or architecture—such as symmetry, vanishing components, or decoder proximity—many of the observed patterns align with known physical principles in gravitational wave

modeling. The following studies should thus be viewed as indicative of learned model behavior, with findings potentially offering insight into physically meaningful signal properties.

3.6.1. Obfuscation Experiments To understand which regions of the waveform most contribute to predictive performance, we conduct a systematic input obfuscation study. In each test, we zero out a fixed segment of the encoder input—effectively masking that region—and measure the resulting degradation in predictive performance. This allows us to estimate the relative importance of each input interval. For a gallery of predicted waveforms resulting from these obfuscation experiments, see [29].

Figure 8 illustrates the masking procedure: a segment of the encoder input (blue) from $t \in [-2000M, -1000M]$ is replaced with zeros (red), while the decoder target (orange) remains unaltered. This setup ensures that any change in output quality can be directly attributed to the missing input region.

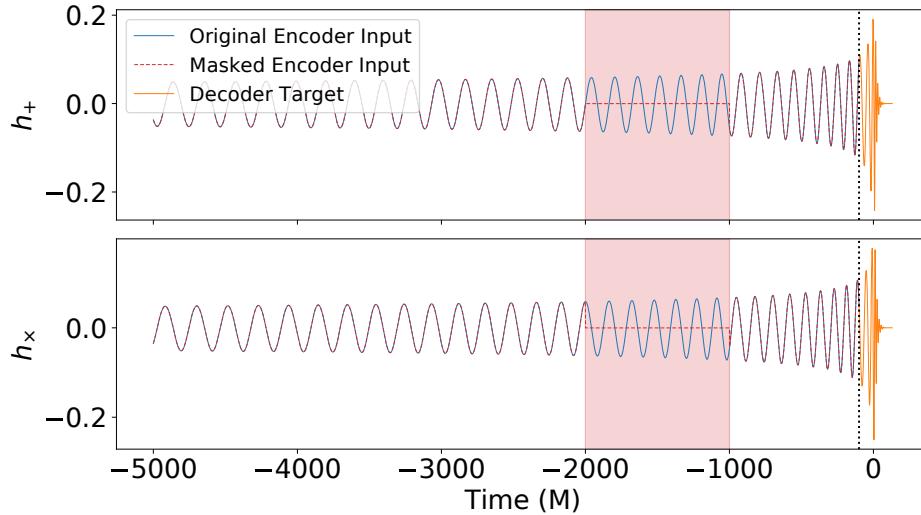


Figure 8: Example of obfuscation: a segment of the input waveform between $t \in [-2000M, -1000M]$ is zeroed out before inference.

We tested the following obfuscated intervals: $(-5000M, -4000M)$, $(-4000M, -3000M)$, $(-3000M, -2000M)$, $(-2000M, -1000M)$, $(-1000M, -200M)$, and $(-200M, -120M)$. All segments are approximately 1000M long to allow fair comparison except the final one, $(-200M, -120M)$, where we retain the last 20 timesteps to preserve continuity at the encoder-decoder boundary. Though shorter, this segment lies near the prediction start and, as later analysis shows, has an outsized effect on performance. This design helps ensure that differences in outcome reflect learned importance rather than artifacts of segment length.

To quantify performance degradation, we compute the overlap score O between predicted and ground truth waveforms. Figure 9 summarizes the distribution of O

values across albbfuscation intervals. Unsurprisingly, early inspiral segments have an overall limited impact on prediction quality, whereas regions approaching merger are disproportionately influential. Masking the final 200M results in a decline in overlap and an increase in variance, underscoring the role of late-time dynamics in informing accurate reconstructions.

To probe how this sensitivity varies with physical parameters, we calculate the Pearson correlation coefficient between the overlap drop (i.e., performance loss from masking) and each system's intrinsic properties mass ratio q , spin components (s_x , s_y , s_z), and inclination angle θ . Since inclination effects are approximately symmetric around $\theta = \pi/2$, we use the transformed variable $|\theta - \pi/2|$ to better reflect this structure. The resulting correlation matrix is shown in Figure 10. While correlation values provide a coarse summary of parameter importance, they can obscure nonlinear relationships. Appendix B shows binned overlap drops, revealing structure not captured by Pearson correlation alone.

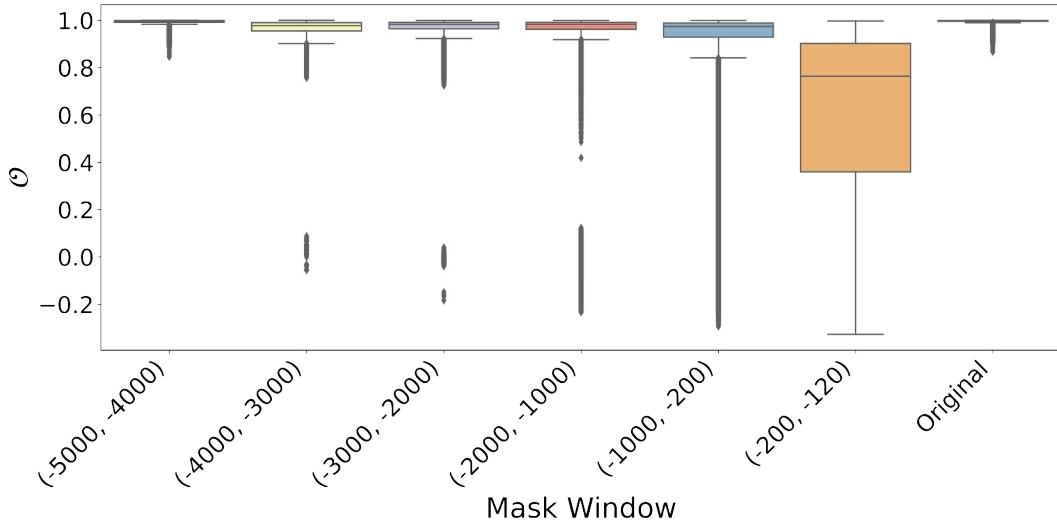


Figure 9: Distribution of overlap scores O for different obfuscated input intervals. Each box shows the interquartile range, with the median marked by a central line. Whiskers extend to $1.5 \times$ IQR, and outliers are shown as individual points.

The correlations reveal several noteworthy trends:

- **Mass Ratio (q):** A moderate negative correlation appears in the earliest segments, suggesting that equal-mass binaries carry more predictive utility in the early inspiral compared to asymmetric systems. While high mass-ratio systems exhibit richer harmonic content due to enhanced subdominant modes, equal-mass systems produce smoother, more symmetric signals dominated by the quadrupole mode [45]. These smoother signals may be more readily captured and generalized by the network in this regime. The features of asymmetric binaries become more prominent, and more predictive, closer to merger, as reflected in the steadily increasing match drop with q in later segments.

- **Spin of the More Massive Black Hole**: The correlation with overlap drop is strongest in the final input segment ($-200M$, $-120M$), suggesting that the model relies more heavily on late-time features which is aligned with the orbital angular momentum. This is in line with previous works which show that spin effects are most prominent near merger, where relativistic velocities enhance spin-orbit and spin-spin couplings. These contribute at higher post-Newtonian orders and thus have limited influence during early inspiral. Near merger, high spins affect phase evolution and waveform amplitude, as reflected in the amplitude's dependence on final spin and the emergence of spin-driven multipolar radiation [46, 47].
- **Spin of the Less Massive Black Hole**: Correlations with s_2^z are weak across all intervals, indicating a diffuse and relatively minor influence on waveform morphology. This is consistent with post-Newtonian predictions that attribute only a minor role in phase evolution to the spin of the smaller object [48].

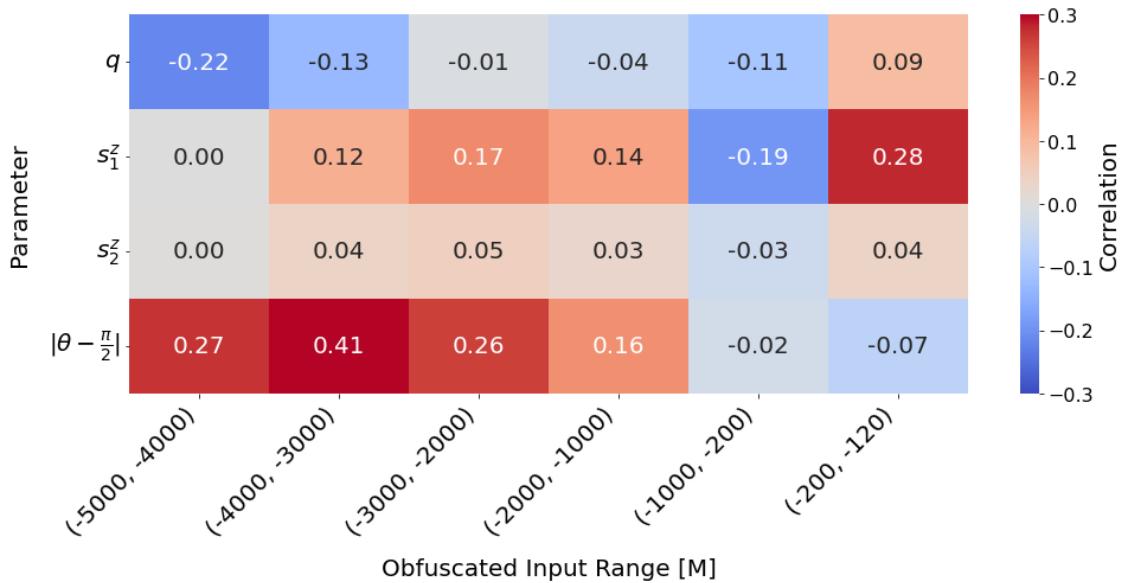


Figure 10: Correlation matrix showing how waveform parameters correlate with the drop in overlap when different parts of the input are obfuscated. Each row corresponds to a physical parameter: mass ratio q , spin components s_1^z and s_2^z , and the transformed inclination $|\theta - \frac{\pi}{2}|$, which captures the approximately symmetric structure observed around edge-on orientations. Each column corresponds to a time window in units of M .

- **Inclination Angle (θ)**: The inclination angle exhibits nontrivial correlation even in early inspiral segments, suggesting that the model leverages amplitude and polarization modulations induced by viewing geometry across long timescales. These modulations arise because inclination affects how spherical harmonic modes project onto the line of sight: face-on systems are dominated by the $(2, 2)$ mode, while edge-on systems excite a broader spectrum of harmonics. These geometric effects alter signal morphology across the entire waveform, as noted in [49]. Notably,

the match drop for edge-on binaries is near zero except in the final segments ($-1000M, -120M$) (see [Appendix B](#)), likely due to the suppressed cross-polarization component at $\theta \approx \pi/2$, which reduces the effect of masking rather than indicating model disinterest. As such, low match drops near edge-on inclination should be interpreted with caution, as they may reflect vanishing waveform content rather than reduced importance.

Overall, these results offer quantitative insight into the localized importance of waveform features across time. Across all binary configurations, aligned-spin systems (i.e., systems where the primary black hole's spin is aligned with the orbital angular momentum) and those with large $\theta \approx \pi$ are among the best resolved by the model and exhibit the highest information content in the late inspiral, as revealed by our obfuscation studies. In contrast, systems with anti-aligned primary spin or those with $\theta \approx 0$ carry less predictive information in the final input segments and correspond to regions of degraded performance. These patterns suggest a possible relationship between the informational richness of the late inspiral and the model's reconstruction accuracy: systems where specifically the late-time waveform evolution is more informative tend to be more successfully forecasted.

3.6.2. Mode Modified Datasets To evaluate the contribution of higher-order gravitational wave modes to model accuracy, we constructed two additional test sets with truncated mode content. The first includes only modes up to $\ell_{\max} = 3$, while the second is limited to $\ell_{\max} = 2$. These reduced-mode datasets are generated using the same procedure and cover the same temporal domain as the full dataset, $t \in [-5000M, 130M]$, but exclude subdominant modes. Unlike in the obfuscation experiments, the ground truth here is adjusted to match the ablated input. This ensures that performance differences reflect the absence of subdominant modes rather than inconsistencies between input and target.

For each dataset, we compute the overlap score O between model predictions and the corresponding reduced-mode ground truth. Summary statistics are presented in [Table 1](#). As anticipated, the full-mode dataset, $\ell_{\max} = 4$, yields the highest mean and median overlaps and the lowest standard deviation, whereas reducing mode content progressively decreases accuracy and increases variance.

Table 1: Overlap statistics for different mode cutoffs.

Dataset	Mean O	Median O	Std
$\ell_{\max} = 2$	0.9687	0.9890	0.0575
$\ell_{\max} = 3$	0.9864	0.9938	0.0226
$\ell_{\max} = 4$ (Full)	0.9959	0.9974	0.0052

To understand where these discrepancies originate, we analyze the performance drop relative to the full-mode dataset as a function of intrinsic binary parameters.

Figure 11 displays the loss in overlap for each reduced-mode variant across mass ratio q , the aligned spin of the heavier black hole \hat{s} , and inclination angle θ .

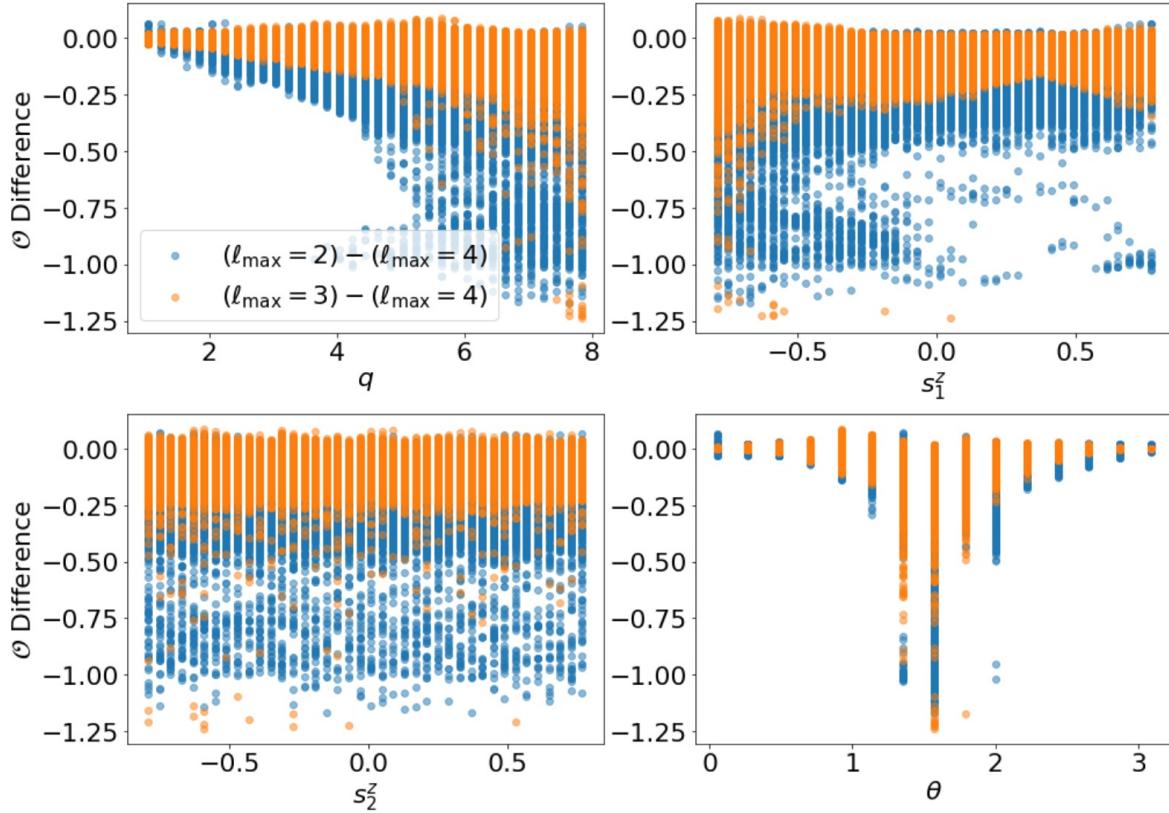


Figure 11: Each point compares either $O(\ell_{\max} = 2) - O(\ell_{\max} = 4)$ or $O(\ell_{\max} = 3) - O(\ell_{\max} = 4)$ as a function of binary parameters. The inclusion of higher order modes improves forecasting accuracy for binaries with large mass ratios, low spin values of the heavier black hole, \hat{s} , and edge-on orientations.

Our analysis reveals that higher-order modes improve accuracy in three specific regimes:

- **High Mass Ratio ($q \gtrsim 4$):** The exclusion of higher modes leads to significant overlap loss, particularly when reducing to $\ell_{\max} = 2$. This aligns with the increased contribution of subdominant harmonics in asymmetric mergers. As mass ratio deviates from 1, symmetry suppression of odd- m and higher- ℓ modes weakens, and subdominant modes grow in significance. Mode truncation most affects high- q systems, where the $(2, 2)$ mode no longer dominates [45].
- **Anti-Aligned Primary Spin ($s_1^z < 0$):** Systems where the primary black hole's spin is strongly anti-aligned with the orbital angular momentum show somewhat increased sensitivity to mode reduction, particularly in a small subset of cases with $s_1^z \ll 0$ when excluding $\ell > 3$. This trend is modest, and further analysis indicates it does not strongly correlate with spin-spin alignment or with effective

spin. The increased mode sensitivity may instead reflect a compensatory role of subdominant harmonics in systems with retrograde alignment, which are known to exhibit distinct merger dynamics and reduced inspiral duration compared to aligned configurations [50, 51].

- **Edge-On Orientations ($\theta \approx \pi/2$):** Orientation effects are especially pronounced in the transition from $\ell_{\max} = 4$ to $\ell_{\max} = 3$, but not between $\ell_{\max} = 3$ and $\ell_{\max} = 2$. This suggests that modes with $3 < \ell \leq 4$ encode angular structure essential for modeling polarization and amplitude variations in edge-on configurations. For such orientations, the dominant $(2, 2)$ mode contributes less power along the line of sight, while subdominant modes project more strongly. As a result, neglecting higher-order modes leads to the largest mismatches in edge-on cases, where accurate waveform modeling requires their inclusion [49].

In contrast, the spin of the lighter black hole s_2^z continues to play a secondary role across all mode configurations. The consistently weak dependence on s_2^z suggests that the model captures its influence through more global, lower-order features.

Overall, these results confirm that including higher-order modes improves model fidelity, particularly in regimes where subdominant harmonics contribute significantly to the waveform structure. Systems where higher-order modes are essential—such as high mass-ratio, edge-on mergers—exhibit both high sensitivity to mode removal and reduced baseline performance, suggesting that dependence on subdominant mode content correlates with reconstruction difficulty, likely due to the increased structural complexity and reduced dominance of the quadrupole mode in these regimes. The model’s sensitivity to angular effects and binary asymmetries further highlights the value of full mode content for achieving robust and generalizable predictions.

3.7. Attention Plots

In Figure 12, we provide visualizations of the attention weights derived from the two decoder attention modules. While primarily qualitative, these visualizations help form an intuitive understanding of how the attention mechanism operates across different input waveforms and attention heads.

As detailed in the model section, the attention weights are computed using the dot product of queries and keys, QK^T , which measures the similarity between them. The softmax function then transforms these similarities into a probability distribution, resulting in a matrix of attention weights with shape $R^{n_{\text{dec}} \times n_{\text{enc}}}$ in the case of the cross-attention module, or with shape $R^{n_{\text{dec}} \times n_{\text{dec}}}$ in the case of the self-attention module. These weights are used to weigh the contribution of the values V for the next representation.

In both cross- and self-attention plots, we include the corresponding input sequences alongside the x- and y-axes to help visualize the model’s focus. For self-attention, the plotted decoder input sequence represents both queries (x-axis) and keys (y-axis). For cross-attention, the plotted decoder input sequence corresponds to the queries (x-axis), while the encoder input sequence represents the keys (y-axis). Although the attention

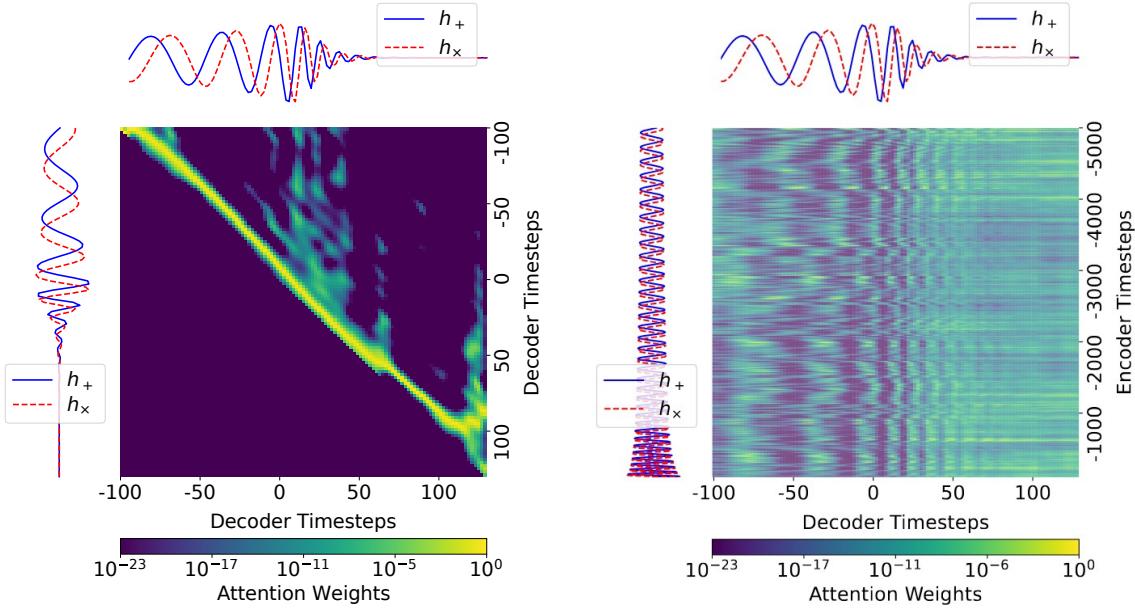


Figure 12: Visualization of self-attention (left) and cross-attention (right) weights for a random choice of attention head and input waveform. The raw attention weights have been transposed to obtain the visual representation achieved here to aid visualization, the corresponding input sequences have been plotted alongside the axes.

modules operate on embeddings or intermediate representations rather than the raw sequences. The plotted sequences provide a useful reference for interpreting the attention patterns. Note that each decoder timestep t_{dec} in the cross-attention plot (i.e., each column in the transposed weight matrix corresponding to some decoder timestep t_{dec}) contains contextual information about how each encoder timestep relates to the current decoder timestep, which is needed for the transformer to make the $t_{dec} + 1$ prediction in the decoder output. Due to the softmax operation, all values in each column sum up to 1.

The self-attention heatmap illustrates how each position in the decoder sequence attends to all other positions in the same sequence. For the self-attention mechanism in the decoder, each timestep can attend to all timesteps up to and including itself, but not to future timesteps, due to the causal attention mask. As a consequence of this property, there are no nonzero elements below the diagonal of the transposed weight matrix. As exemplified in Figure 12, the diagonal elements frequently tend to dominate in these self-attention weight plots, indicating that local context is especially important to the self-attention mechanism when making predictions. The cross-attention heatmap reveals the model’s focus on different parts of the encoder output sequence while decoding each position of the output sequence. The attention is more diffused than in the self-attention

case, as the model is integrating information from a broader range of encoder timesteps to inform each prediction in the decoder output.

Interactive visualizations of the attention patterns for selected waveforms are available at [29] to support further qualitative inspection. However, we caution against overinterpreting these attention plots as a reflection of the model’s reasoning. The patterns tend to correlate closely with the amplitude structure of the input waveforms, showing little abstraction or parameter-dependent structure across examples. These findings are consistent with prior work indicating that attention weights do not necessarily provide meaningful explanations for model behavior[52]. Accordingly, we present these visualizations as diagnostic and instructive tools rather than interpretability claims.

4. Discussion

This work demonstrates that transformer-based models can serve as fast, accurate surrogates for gravitational waveforms with higher-order modes, enabling waveform forecasting from quasi-circular, spinning, non-precessing binary black hole mergers across the late inspiral, merger, and ringdown phases. The model was trained on over 14 million waveforms using distributed training across 16 NVIDIA A100 GPUs on the Delta supercomputer, reaching convergence within 15 hours. Inference on 840,000 test waveforms was conducted using a single NVIDIA V100 GPU on the HAL system, with an average latency of 18 ms per waveform.

The model achieved high predictive accuracy across most of the signal manifold, with mean and median overlap scores of 0.996 and 0.997, respectively, and over 92% of samples out of the surrogate-based test set exceeding an overlap of 0.99. We also found that the model’s conditional masking mechanism for the \mathbf{p} polarization preserved accuracy in configurations where that signal component vanishes.

To assess generalization beyond the surrogate training domain, we benchmarked the model against numerical relativity waveforms from the SXS catalog. Despite including systems outside the training domain and additional harmonic content, the model achieved a median overlap of 0.969, with low-overlap outliers scattered only around $\theta = \pi/2$, indicating strong performance even on high-fidelity, out-of-distribution data. The high data-to-parameter ratio (40,000:1) and strong NR generalization suggest the model avoids overfitting.

Further analysis revealed that the model relies heavily on late-inspiral information in systems with high spin and mass asymmetry, while earlier inspiral features proved more important for near-equal-mass binaries. The model also appears to extract inclination angle-dependent structure well before mergeMode truncation experiments confirmed that higher-order modes significantly enhance accuracy, particularly in systems with large mass ratios and edge-on orientations.

While the model performs well across most of the parameter space, further improvement is needed for specific configurations, including low-mass-ratio, face-on

binaries and high-mass-ratio, edge-on mergers—cases in which waveform complexity and amplitude suppression may challenge autoregressive modeling. These edge cases represent a small fraction of the test set, with only 0.003% of waveforms yielding overlaps below 0.90.

This work provides a starting point for future research into higher-order wave modes, including those from spinning, precessing binary black hole mergers. This broader signal manifold offers an opportunity to systematically explore the capabilities of AI models and exascale computing in learning complex, nonlinear dynamics of compact binary systems. These directions should be pursued in future work to further deepen our understanding of gravitational wave signals.

Acknowledgements

This work was supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357 and by the Diaspora project of the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, under contract number DE-AC02-06CH11357. KP was in part supported by the NSF grant NRT-1922512. The work used resources of the Argonne Leadership Computing Facility, a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. EAH acknowledge support from National Science Foundation (NSF) award OAC-2209892. This research also used the Delta advanced computing and data resources, which is supported by the National Science Foundation (award OAC 2005572) and the State of Illinois. Delta is a joint effort of the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications. This research used the DeltaAI advanced computing and data resource, which is supported by the National Science Foundation (award OAC 2320345) and the State of Illinois. DeltaAI is a joint effort of the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications.

Appendix A. Fine-Grained Analysis via Spin Slices

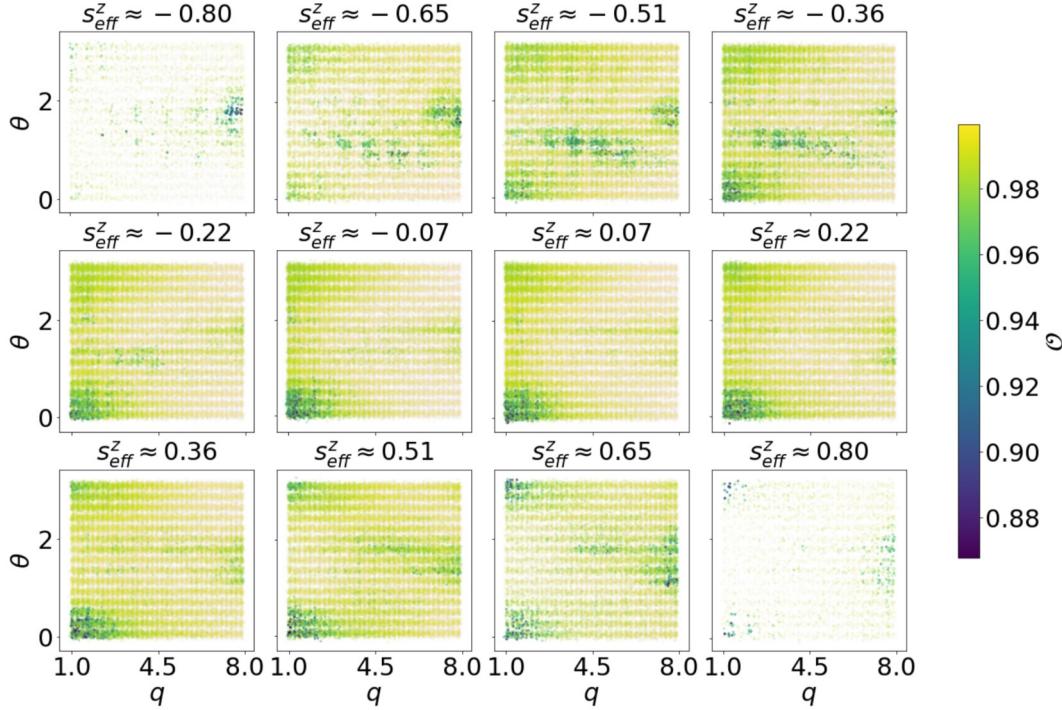


Figure A1: Reconstruction accuracy across mass ratio q , and inclination angle, θ , for several fixed values of the effective spin parameter s_{eff}^z .

Figure A1 shows the test set partitioned into fixed-spin slices across the (q, θ) plane for various values of effective spin, s_{eff}^z . This stratified analysis reveals that the transformer model consistently achieves high predictive accuracy for binaries with strongly aligned or anti-aligned spins, i.e., $s_{\text{eff}}^z \sim \pm 0.8$. Within these regimes, the model performs robustly across a wide range of mass ratios and inclination angles.

However, at intermediate spin values—particularly around $s_{\text{eff}}^z \sim 0$ —a localized degradation in performance is observed. Specifically, reduced overlap values are concentrated around systems with comparable mass ratios ($q \sim 1$) and face-on orientations ($\theta \sim 0$), suggesting a sensitivity of the model to configurations with low intrinsic spin alignment and minimal inclination-induced modulation.

As s_{eff}^z departs from zero in either direction, predictive accuracy systematically improves across the (q, θ) slices, highlighting the model’s effectiveness in capturing spin-modulated waveform features. Nonetheless, the model exhibits a consistent dip in accuracy for edge-on binaries ($\theta \sim \pi/2$), irrespective of spin, indicating a challenging regime associated with vanishing cross polarization and potential waveform degeneracies.

Finally, while the model attains near-optimal performance for $s_{\text{eff}}^z \sim \pm 0.8$, we identify a narrow region of reduced accuracy at $s_{\text{eff}}^z \sim -0.8$ for high-mass-ratio, edge-on systems ($q \sim 8$, $\theta \sim \pi/2$). This subtle degradation suggests a complex interplay between spin orientation, geometric projection effects, and mass asymmetry in these

extreme regions of the parameter space.

Appendix B. Binned Overlap Drop by Parameter for Obfuscation Study

To complement the correlation analysis in section 3.6.1, we provide binned heatmaps of the average overlap drop as a function of parameter value and obfuscated input range. These plots offer a finer-grained view of parameter influence on predictions, allowing us to assess possible nonlinear or symmetric effects that may be obscured by Pearson correlation.

Figure B1 shows match drop by mass ratio q . Most input segments exhibit a broadly monotonic increase in match drop with q , except in the $(-1000M, -200M)$ segment, where it decreases. Early segments show slight non-monotonicity, particularly near $q \approx 1$, explaining the low Pearson correlation in that region. Figure B2 displays match drop by primary spin s_1^z . The trend is mostly monotonic, with drop increasing with spin. A deviation appears for $\xi < -0.5$, where drop temporarily decreases. In the $(-1000M, -200M)$ segment, match drop instead decreases with spin.

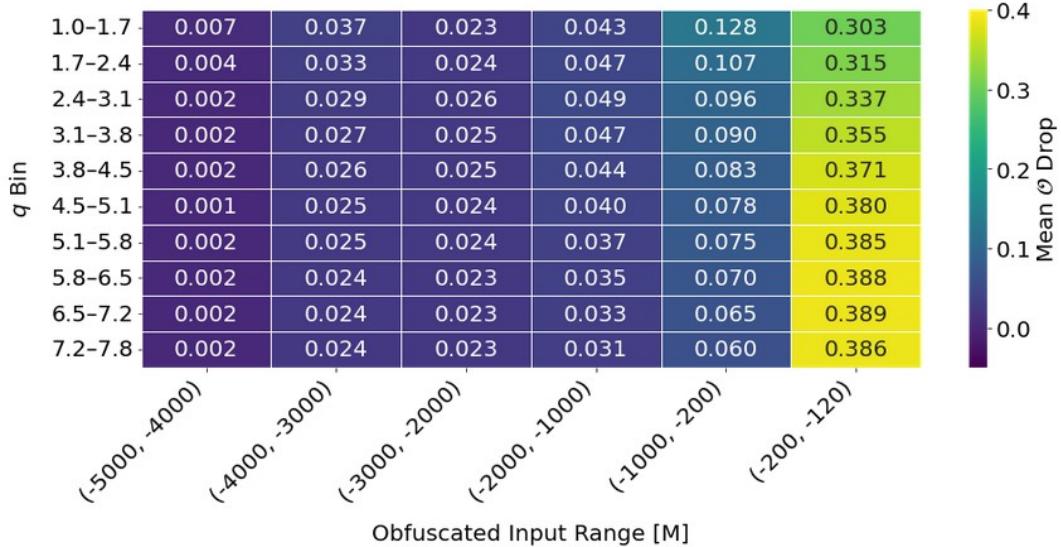


Figure B1: Binned overlap drop by mass ratio q . Most input segments show a broadly monotonic relationship, with match drop increasing for more asymmetric binaries (with the exception of the $(-1000M, -200M)$ range, where match drop decreases monotonically with q). Some early segments display small deviations from this monotonicity, particularly at low q . Notably, the low Pearson correlation observed in the earliest segment is largely driven by this localized effect around $q \approx 1$.

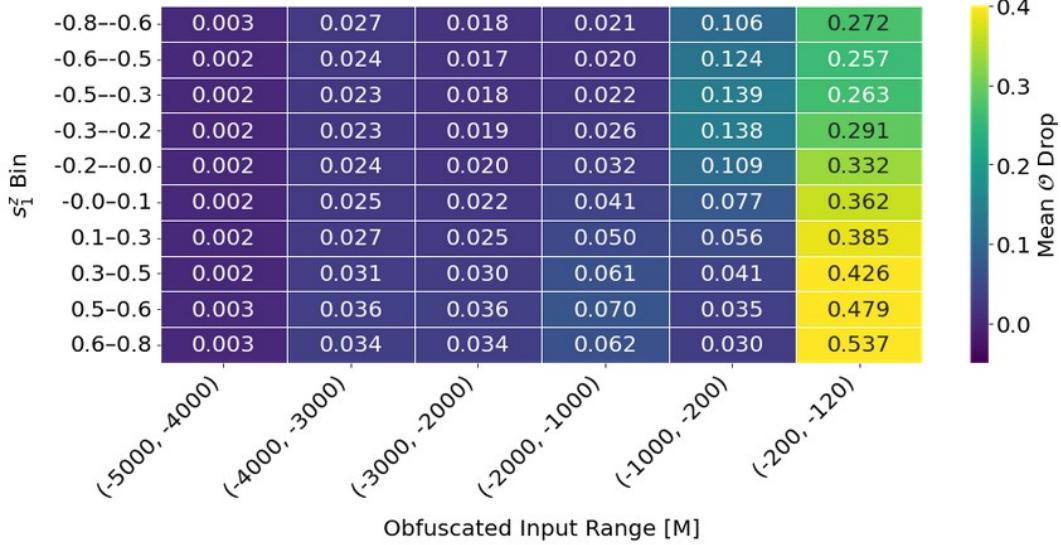


Figure B2: Binned overlap drop by primary spin ξ_1 . The overall trend is approximately monotonic across the spin range, with match drop generally increasing with spin (except in the $(-1000M, -200M)$ range, where match drop decreases with spin). A slight deviation from monotonicity is observed for very negative spin values ($s_1^z < -0.5$), where match drop temporarily decreases before continuing the broader trend.

Figure B3 shows match drop by secondary spin ξ_2 , which is relatively flat across all bins. Figure B4 plots match drop by inclination angle θ . Most segments show a symmetric dependence around $\theta \approx \pi/2$, justifying the use of $|\theta - \pi/2|$ in the main analysis. In the final segment, this symmetry breaks slightly, with drop increasing toward higher inclination.

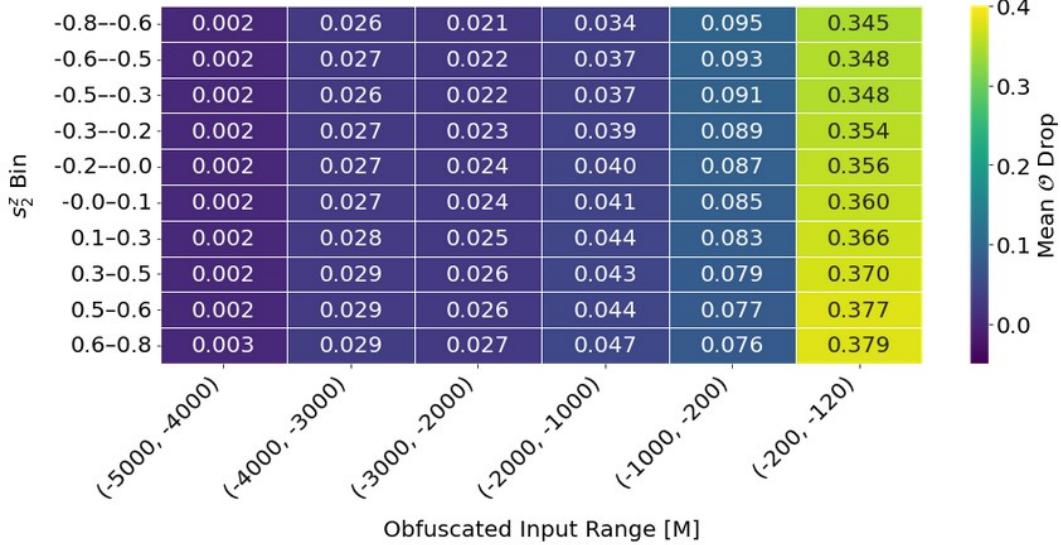


Figure B3: Binned overlap drop by secondary spin ξ_2 . The overlap drop is relatively flat across all bins (i.e., within each column).

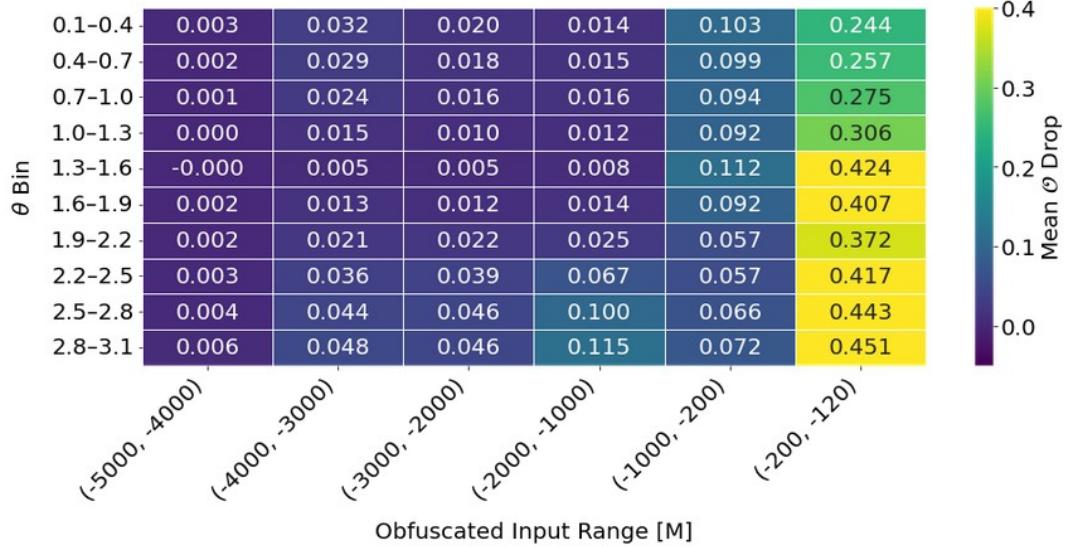


Figure B4: Binned overlap drop by inclination angle θ . Most time segments show a roughly symmetric dependence around $\theta \approx \pi/2$, motivating the use of the transformed variable $|\theta - \pi/2|$ in the main analysis. In the final segment ($-200M, -120M$), this symmetry breaks down slightly, with match drop increasing toward higher inclination angles.

References

- [1] Cuoco E, Cavagli`a M, Heng I S, Keitel D and Messenger C 2025 *Living Reviews in Relativity* **28** 2 (Preprint [2412.15046](https://arxiv.org/abs/2412.15046))
- [2] Cuoco E, Powell J, Cavagli`a M, Ackley K, Bejger M, Chatterjee C, Coughlin M, Coughlin S, Easter P, Essick R, Gabbard H, Gebhard T, Ghosh S, Haegel L, less A, Keitel D, Marka Z, Marka S, Morawski F, Nguyen T, Ormiston R, Puerrer M, Razzano M, Staats K, Vajente G and Williams D 2021 *Mach. Learn. Sci. Tech.* **2** 011002
- [3] Marx E, Benoit W, Gunny A, Omer R, Chatterjee D, Venterea R C, Wills L, Saleem M, Moreno E, Raikman R, Govorkova E, Desai M, Krupa J, Rankin D, Coughlin M W, Harris P and Katsavounidis E 2025 *Phys. Rev. D* **111**(4) 042010 URL <https://link.aps.org/doi/10.1103/PhysRevD.111.042010>
- [4] Bahaadini S, Noroozi V, Rohani N, Coughlin S, Zevin M, Smith J, Kalogera V and Katsaggelos A 2018 *Information Sciences* **444** 172–186 ISSN 0020-0255 URL <https://www.sciencedirect.com/science/article/pii/S0020025518301634>
- [5] Chaturvedi P, Khan A, Tian M, Huerta E A and Zheng H 2022 *Front. Artif. Intell.* **5** 828672 (Preprint [2201.11133](https://arxiv.org/abs/2201.11133))
- [6] George D, Shen H and Huerta E A 2018 *Physical Review D* **97** 101501 (Preprint [1706.07446](https://arxiv.org/abs/1706.07446))
- [7] Murali C and Lumley D 2023 *Phys. Rev. D* **108**(4) 043024 URL <https://link.aps.org/doi/10.1103/PhysRevD.108.043024>
- [8] Gabbard H, Williams M, Hayes F and Messenger C 2018 *Phys. Rev. Lett.* **120** 141103 (Preprint [1712.06041](https://arxiv.org/abs/1712.06041))
- [9] Koloniari A E, Koursoumpa E C, Nousi P, Lampropoulos P, Passalis N, Tefas A and Stergioulas N 2025 *Mach. Learn. Sci. Tech.* **6** 015054 (Preprint [2407.07820](https://arxiv.org/abs/2407.07820))
- [10] George D and Huerta E A 2018 *Physics Letters B* **778** 64–70 (Preprint [1711.03121](https://arxiv.org/abs/1711.03121))
- [11] Yan J, Avagyan M, Colgan R E, Veske D, Bartos I, Wright J, Márka Z and Márka S 2022 *Phys.*

Rev. D **105** 043006 (*Preprint* [2104.03961](https://arxiv.org/abs/2104.03961))

[12] Huerta E A, Khan A, Huang X, Tian M, Levental M, Chard R, Wei W, Heflin M, Katz D S, Kindratenko V, Mu D, Blaiszik B and Foster I 2021 *Nature Astronomy* **5** 1062–1068 (*Preprint* [2012.08545](https://arxiv.org/abs/2012.08545))

[13] Nousi P, Koliogianni A E, Passalis N, Iosif P, Stergioulas N and Tefas A 2023 *Phys. Rev. D* **108** 024022 (*Preprint* [2211.01520](https://arxiv.org/abs/2211.01520))

[14] Dax M, Green S R, Gair J, Macke J H, Buonanno A and Schölkopf B 2021 *Phys. Rev. Lett.* **127**(24) 241103 URL <https://link.aps.org/doi/10.1103/PhysRevLett.127.241103>

[15] Green S R, Simpson C and Gair J 2020 *Phys. Rev. D* **102**(10) 104057 URL <https://link.aps.org/doi/10.1103/PhysRevD.102.104057>

[16] Leyde K, Green S R, Toubiana A and Gair J 2024 *Phys. Rev. D* **109**(6) 064056 URL <https://link.aps.org/doi/10.1103/PhysRevD.109.064056>

[17] Lee J, Oh S H, Kim K, Cho G, Oh J J, Son E J and Lee H M 2021 *Phys. Rev. D* **103** 123023 (*Preprint* [2101.05685](https://arxiv.org/abs/2101.05685))

[18] Bohé A, Shao L, Taracchini A, Buonanno A, Babak S, Harry I W, Hinder I, Ossokine S, Pürrer M, Raymond V, Chu T, Fong H, Kumar P, Pfeiffer H P, Boyle M, Hemberger D A, Kidder L E, Lovelace G, Scheel M A and Szil’agyi B 2017 *Phys. Rev. D* **95**(4) 044028 URL <https://link.aps.org/doi/10.1103/PhysRevD.95.044028>

[19] Khan S and Green R 2021 *Phys. Rev. D* **103**(6) 064015 URL <https://link.aps.org/doi/10.1103/PhysRevD.103.064015>

[20] Khan A, Huerta E A and Zheng H 2022 *Phys. Rev. D* **105** 024024 (*Preprint* [2110.06968](https://arxiv.org/abs/2110.06968))

[21] Varma V, Field S E, Scheel M A, Blackman J, Kidder L E and Pfeiffer H P 2019 *Phys. Rev. D* **99** 064045 (*Preprint* [1812.07865](https://arxiv.org/abs/1812.07865))

[22] Luo C, Zhang Y and Zhu J 2024 A multi-step prediction algorithm for analysis of gravitational waves based on deep learning *ALLDATA 2024: The Tenth International Conference on Big Data, Small Data, Linked Data and Open Data* (Nice, France: International Academy, Research, and Industry Association (IARIA)) URL https://www.thinkmind.org/articles/alldata_2024_1_10_80002.pdf

[23] Pürrer M, Khan S, Ohme F, Birnholtz O and London L 2023 *IMRPhenomD: Phenomenological waveform model* *Astrophysics Source Code Library*, record ascl:2307.019

[24] Hochreiter S and Schmidhuber J 1997 *Neural Computation* **9** 1735–1780 ISSN 0899-7667 (*Preprint* <https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf>) URL <https://doi.org/10.1162/neco.1997.9.8.1735>

[25] Zhong X Y, Han W B and Sun L 2025 *Phys. Rev. D* **111** 064033 (*Preprint* [2503.04534](https://arxiv.org/abs/2503.04534))

[26] Abbott R *et al.* (LIGO Scientific, Virgo) 2020 *Phys. Rev. Lett.* **125** 101102 (*Preprint* [2009.01075](https://arxiv.org/abs/2009.01075))

[27] Ng K K Y *et al.* 2023 *Phys. Rev. D* **107** 024041

[28] Hu Y and Veitch J 2022 *Phys. Rev. D* **106** 044042

[29] Victoria Tiki 2025 Sequence modeling of higher-order wave modes of binary black hole mergers https://github.com/victoria-tiki/transformer_complex/blob/main/README.md

[30] Newman E T and Penrose R 1966 *Journal of Mathematical Physics* **7** 863–870 (*Preprint* <https://doi.org/10.1063/1.1931221>) URL <https://doi.org/10.1063/1.1931221>

[31] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 *Advances in neural information processing systems* **30**

[32] Pan J S, Ting Y S and Yu J 2024 *Monthly Notices of the Royal Astronomical Society* **528** 5890–5903

[33] Boersma O M, Ayache E H and van Leeuwen J 2024 *RAS Tech. Instrum.* **3** 472–483 (*Preprint* [2406.12515](https://arxiv.org/abs/2406.12515))

[34] Drikakis D, Kokkinakis I W, Fung D and Spottswood S M 2024 *Physics of Fluids* **36** 065113

[35] Kingma D P and Ba J 2014 *arXiv preprint arXiv:1412.6980*

[36] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J and Chintala S 2019 *arXiv e-prints* arXiv:1912.01703 (*Preprint* [1912.01703](https://arxiv.org/abs/1912.01703))

- [37] Halton J 1960 *Numerische Mathematik* **2** 84–90 URL <http://eudml.org/doc/131448>
- [38] Field S E, Galley C R, Hesthaven J S, Kaye J and Tiglio M 2014 *Phys. Rev. X* **4**(3) 031006 URL <https://link.aps.org/doi/10.1103/PhysRevX.4.031006>
- [39] Boyle M, Mitman K, Scheel M and Stein L 2025 The sxs package URL <https://zenodo.org/doi/10.5281/zenodo.15412737>
- [40] SXS Collaboration 2025 The third sxs collaboration catalog of binary black-hole simulations URL <https://arxiv.org/abs/2505.13378>
- [41] SXS Collaboration 2025 The sxs catalog of simulations v3.0.0 URL <https://zenodo.org/doi/10.5281/zenodo.15415231>
- [42] Gramaxo Freitas O, Theodoropoulos A, Villanueva N, Fernandes T, Nunes S, Font J A, Onofre A, Torres-Forné A and Martin-Guerrero J D 2024 (*Preprint* [2412.06946](https://arxiv.org/abs/2412.06946))
- [43] Varma V, Field S E, Scheel M A, Blackman J, Gerosa D, Stein L C, Kidder L E and Pfeiffer H P 2019 *Phys. Rev. Research.* **1** 033015 (*Preprint* [1905.09300](https://arxiv.org/abs/1905.09300))
- [44] Abbott B and et al (LIGO Scientific Collaboration and Virgo Collaboration) 2016 *Phys. Rev. Lett.* **116**(24) 241102 URL <https://link.aps.org/doi/10.1103/PhysRevLett.116.241102>
- [45] Berti E, Cardoso V, Gonzalez J A, Sperhake U, Hannam M, Husa S and Bruegmann B 2007 *Phys. Rev. D* **76** 064034 (*Preprint* [gr-qc/0703053](https://arxiv.org/abs/gr-qc/0703053))
- [46] Burke O, Gair J R and Simón J 2020 *Phys. Rev. D* **101** 064026 (*Preprint* [1909.12846](https://arxiv.org/abs/1909.12846))
- [47] Dong-Páez, Chi An, Volonteri, Marta, Beckmann, Ricarda S, Dubois, Yohan, Trebitsch, Maxime, Mangiagli, Alberto, Vergani, Susanna D and Webb, Natalie A 2023 *A&A* **673** A120 URL <https://doi.org/10.1051/0004-6361/202346295>
- [48] Blanchet L 2014 *Living Rev. Rel.* **17** 2 (*Preprint* [1310.1528](https://arxiv.org/abs/1310.1528))
- [49] Pekowsky L, Healy J, Shoemaker D and Laguna P 2012 *Physical Review D* **87**
- [50] Campanelli M, Lousto C O and Zlochower Y 2006 *Phys. Rev. D* **74** 041501 (*Preprint* [gr-qc/0604012](https://arxiv.org/abs/gr-qc/0604012))
- [51] Hannam M, Husa S, Ohme F, Müller D and Brügmann B 2010 *Phys. Rev. D* **82**(12) 124008 URL <https://link.aps.org/doi/10.1103/PhysRevD.82.124008>
- [52] Jain S and Wallace B C 2019 *arXiv e-prints* arXiv:1902.10186 (*Preprint* [1902.10186](https://arxiv.org/abs/1902.10186))

Data availability statement

We produced training, validation and test datasets using the numerical relativity surrogate model NRHybSur3dq8 [21] which is available at the following URL/DOI: <https://github.com/sxs-collaboration/gwsurrogate/tree/master>. Additionally, we benchmarked the model's predictions against waveforms from the SXS catalog of full numerical relativity simulations, available at [41].

Code availability

The scientific software to reproduce our results, including trained transformer models, and a tutorial for their use is available on GitHub [29].