NRSurNN3dq4

A Deep Learning Powered Numerical Relativity Surrogate for Binary Black Hole Waveforms

Osvaldo Gramaxo Freitas[1,2], Anastasios Theodoropoulos[1], Nino Villanueva[1,3], José A. Font[1], Antonio Onofre[2], Alejandro Torres-Forné[1], José D. Martin-Guerrero[1,3]

[1] *Universitat de València, València, Spain*
[2] *Universidade do Minho, Braga, Portugal*
[3] *ValgrAI, València, Spain*

GRASS 2024, Trento, 01/10/2024

General Idea of NRSurNN3dq4

- NRSurNN3dq4 is a fast surrogate model that predicts binary black hole (BBH) merger waveforms using deep learning.
- Makes use of a pretraining step on approximating data before fine-tuning on NR data
- Predicts low-dimensional representations of waveform data, allowing fast inference.
- Leverages parallelization ability to generate large numbers of accurate waveforms very quickly.

Outline

Outline

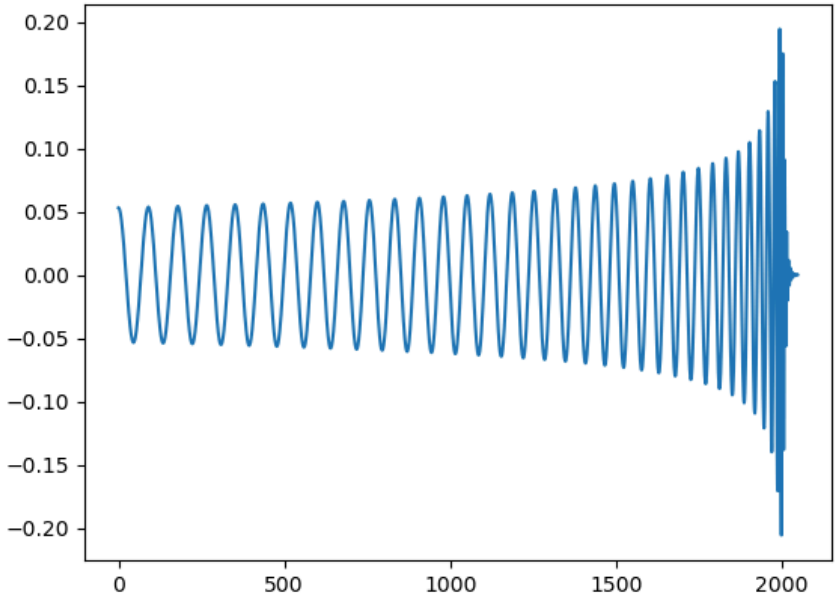
Surrogate models: Why/What

- Gravitational-wave (GW) astronomy involves many computationally intensive waveform generation tasks.
- Numerical relativity simulations are time-consuming but accurate.
- Approximants provide faster alternatives for waveform generation with slight accuracy trade-offs.
- Surrogate models aim to reduce this gap using interpolation on NR data.

Outline

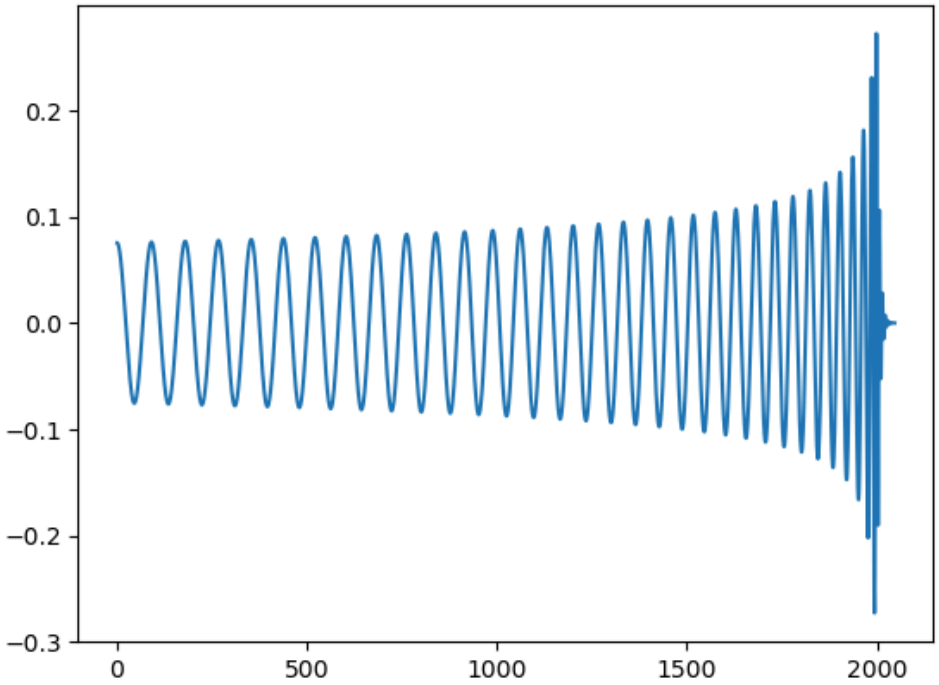
Approximant Data

- Generated using NRSur7dq8 surrogate model.
- 1,024,000 waveforms created with uniform mass ratios $q \in [1, 6]$ and aligned spins $\chi_1, \chi_2 \in [-0.99, 0.99]$.
- Amplitude and time in natural units.
- Duration: $4096 M_\odot$ with sampling rate of $2 M_\odot$



Numerical Relativity Data

- Sourced from SXS collaboration's simulation catalogue.
- Non-precessing waveforms filtered for consistency with approximant data, resulting in 381 simulations.



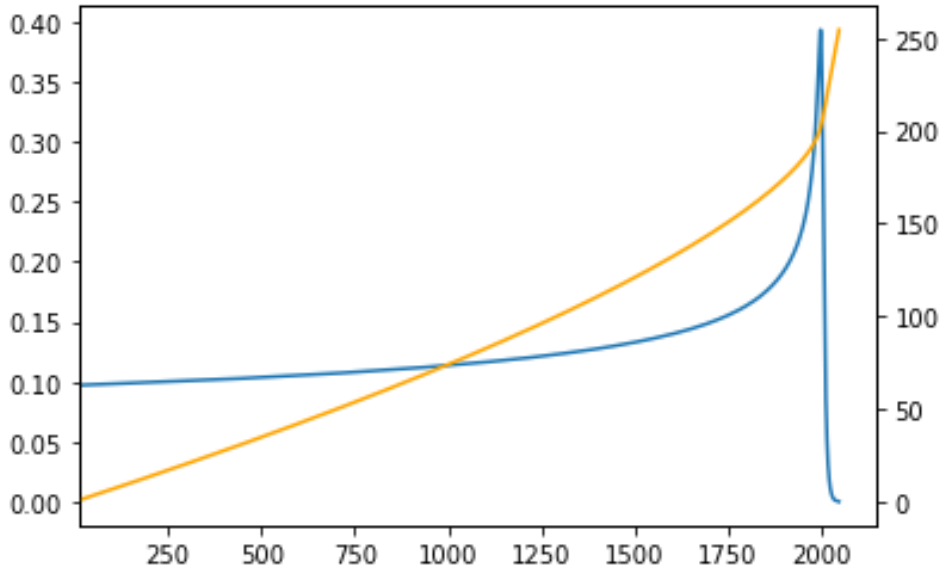
Outline

A primer on dimensionality

- When working with discretized data, waveforms of length N are effectively vectors in $\mathbb{R}^N$ whose components have certain restrictions
- The "curse of dimensionality" is the observation that an increase of the dimensionality of data leads to dramatic growth in the volume of the data space, which in turn leads to sparsity of data points (average distance scales with $\sqrt{N}$)

Dimensionality Reduction

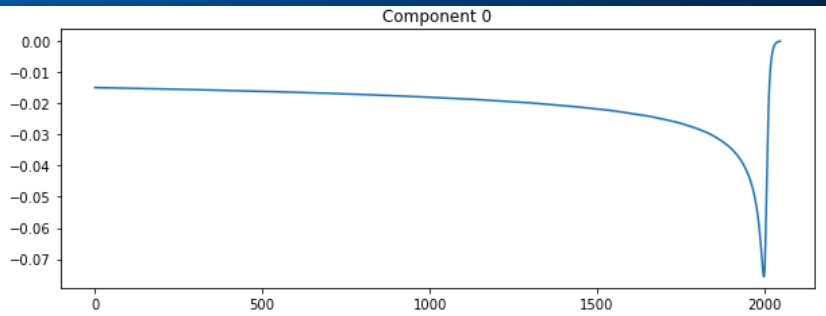
- High-dimensional waveforms can be simplified for efficient processing.
- There are simple ways of manipulating GW data in ways to tackle complexity:
  - Make sure the merger happens at the same point always
  - Set the initial phase to zero across the entire dataset
  - Represent the complex waveform as $h(t) = A(t)^{\phi(t)}$

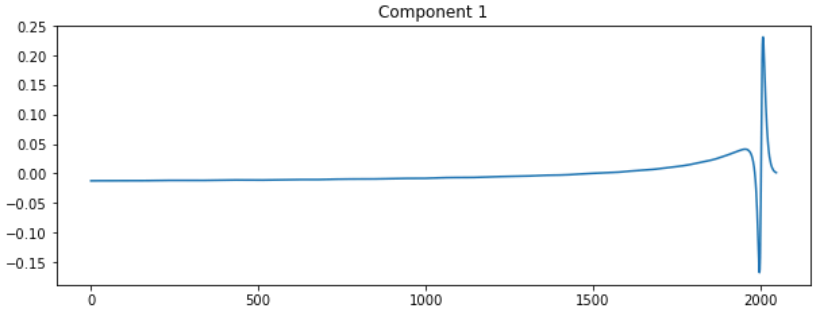


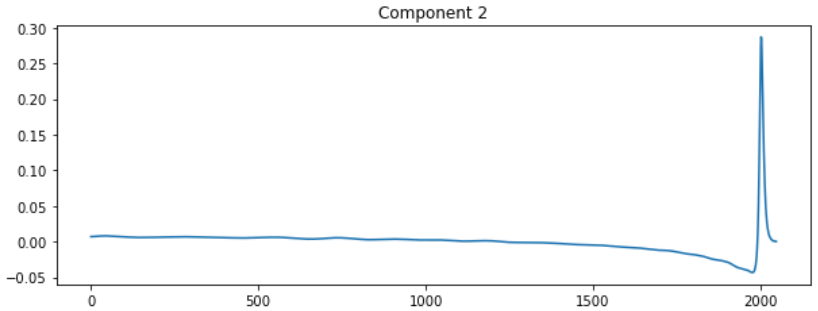
Dimensionality Reduction

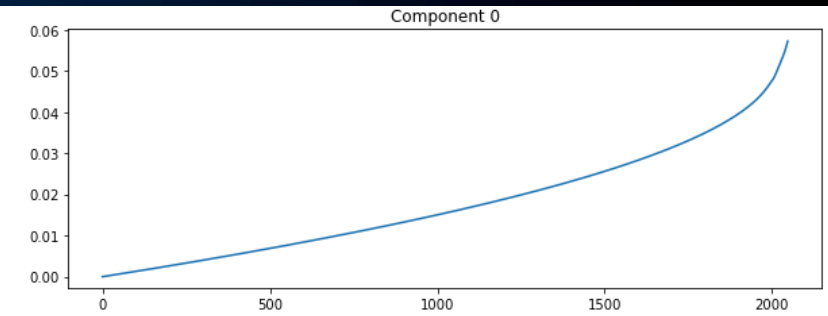
- We can further simplify data by using principal component analysis (PCA) to identify the most significant underlying patterns (principal components) within the data.
- PCA works by performing eigendecomposition on the covariance matrix of a subset of the approximant dataset
  - Eigenvectors will compose an orthonormal basis
  - Eigenvalues express the amount of variance explained
- We can then select a subset of the most relevant component (30 for amplitude, 70 for phase) and project the original dataset to this basis

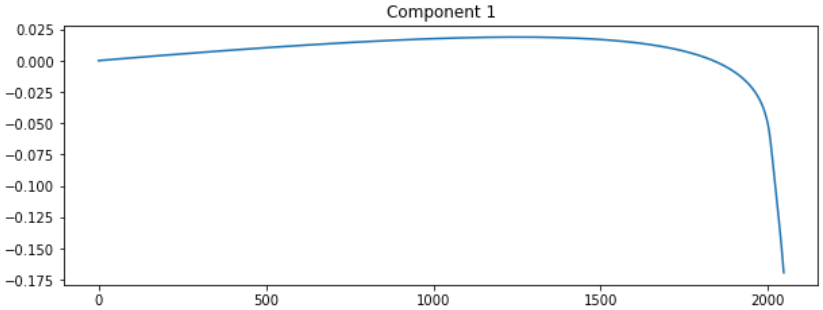
Dimensionality Reduction

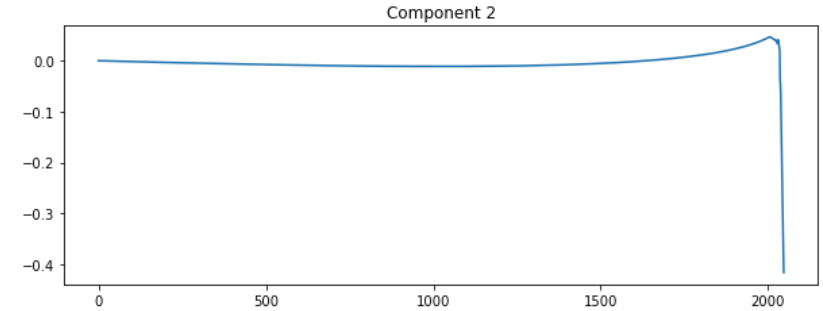








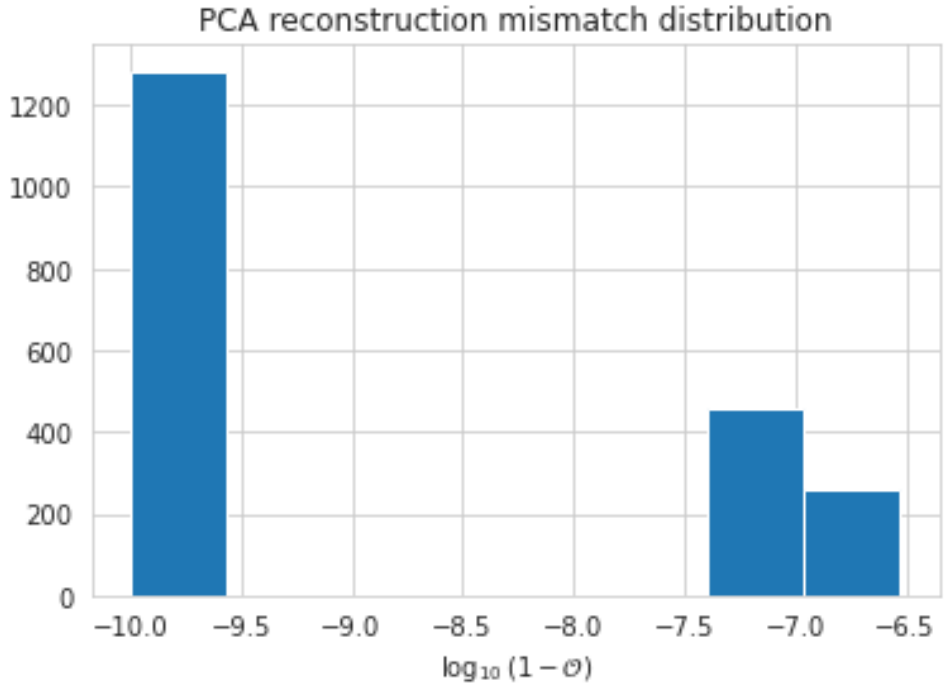




Reconstruction quality

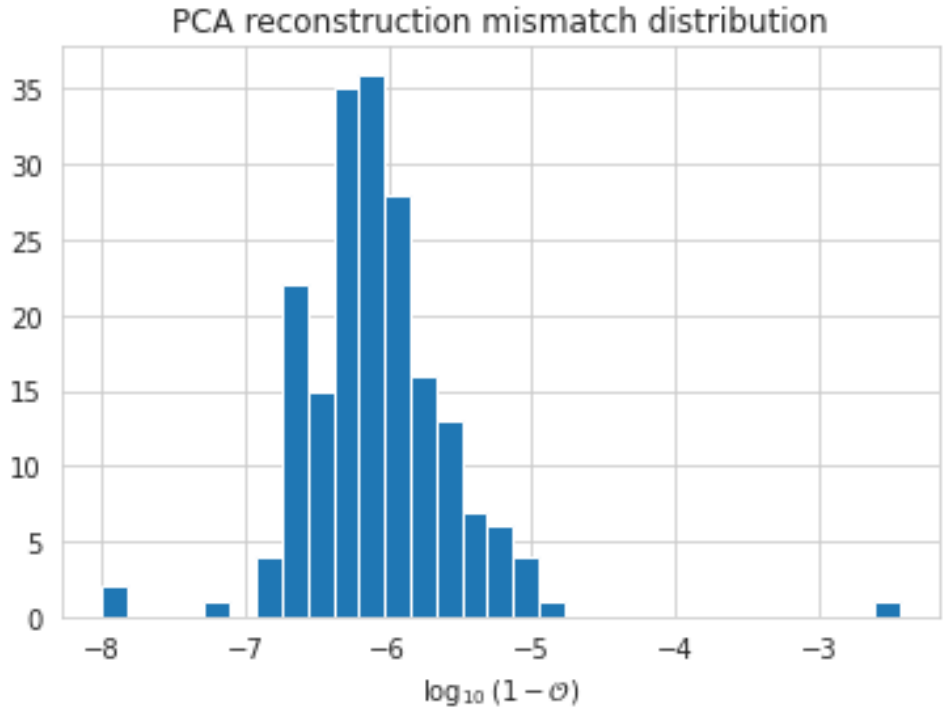
- To evaluate the quality of a reconstruction, we use the mismatch:

$$1 - \mathcal{O} = 1 - \frac{(h_1|h_2)}{\sqrt{(h_1|h_1)(h_2|h_2)}}$$



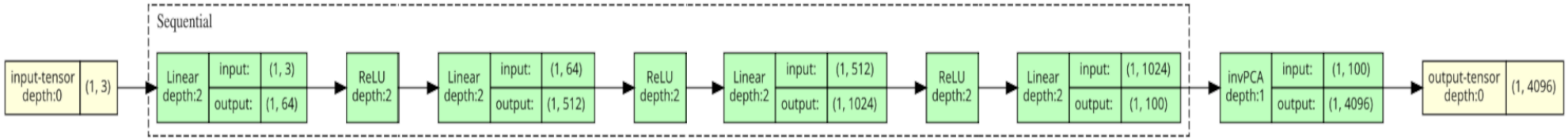
Reconstruction quality

We verify that the basis created on the approximant can largely reconstruct the NR data.



Network Architecture

- A multi-layer perceptron (MLP) with 3 hidden layers (64, 512, 1024 neurons) implemented in torch.
- ReLU activations.
- PCA basis as internal (frozen) model parameters.
- Integrated PCA projection transform and inverse.



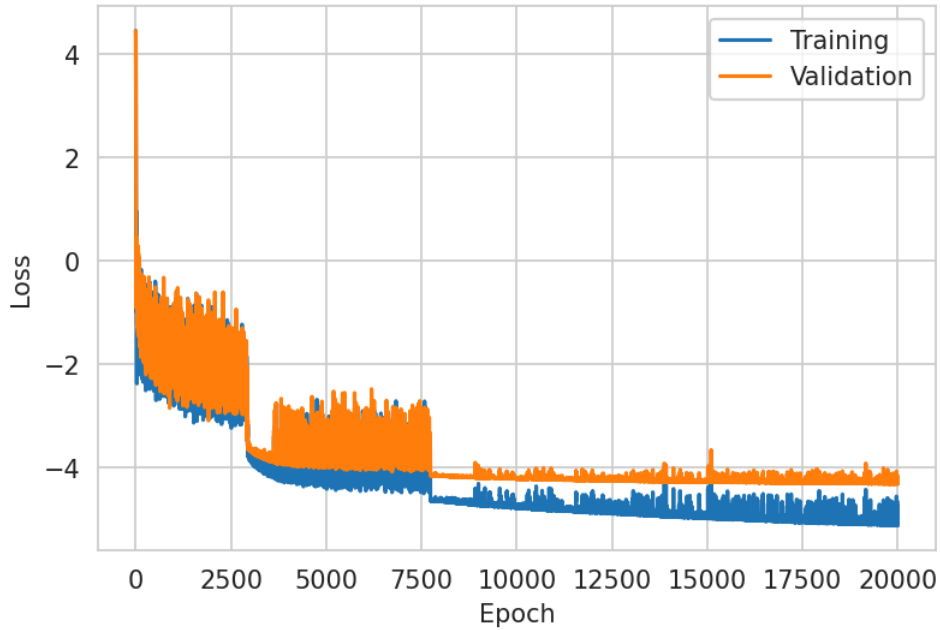
Loss Function:

- A combination of **PCA component loss** (mean absolute error) and **waveform overlap loss** (mismatch) is used to ensure both accurate PCA reconstruction and waveform fidelity.
- $L = L_1 + \log(L_2)$, where L_1 is the MAE on PCA components and L_2 is the mismatch between generated and actual waveforms.
- Key point: having the PCA transforms as part of the torch model allows us to use features from both the latent space and the waveform space together.

Outline

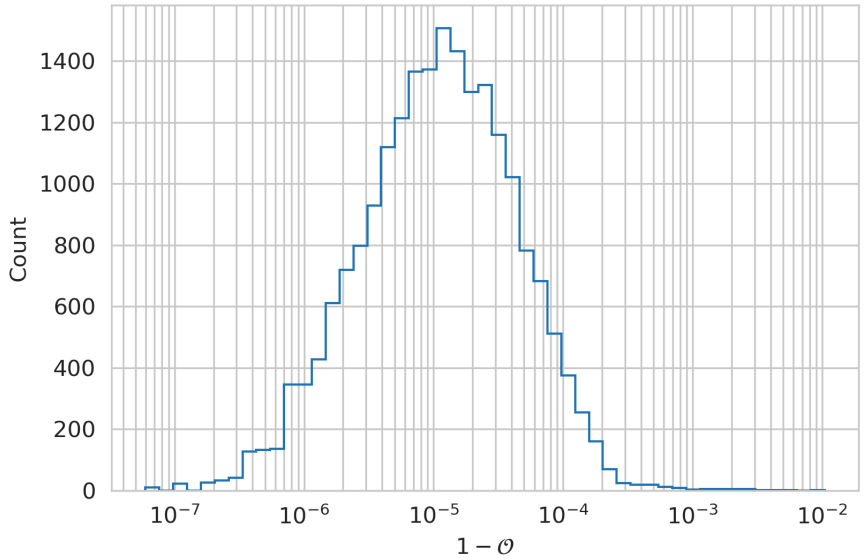
Pretraining Process

- **Pre-training on Approximant Data**:
- The neural network is pre-trained using 1 million waveforms generated by the NRSur7dq8 approximant model, with a 80/20 train-validation split
- **Sophia optimizer** with starting learning rate of 3×10^{-3}
- Learning rate scheduler reducing the learning rate after 1000 epochs without improved performance.
- Training lasts for 20,000 epochs, with learning rate adjustments based on validation performance.



Approximant Accuracy

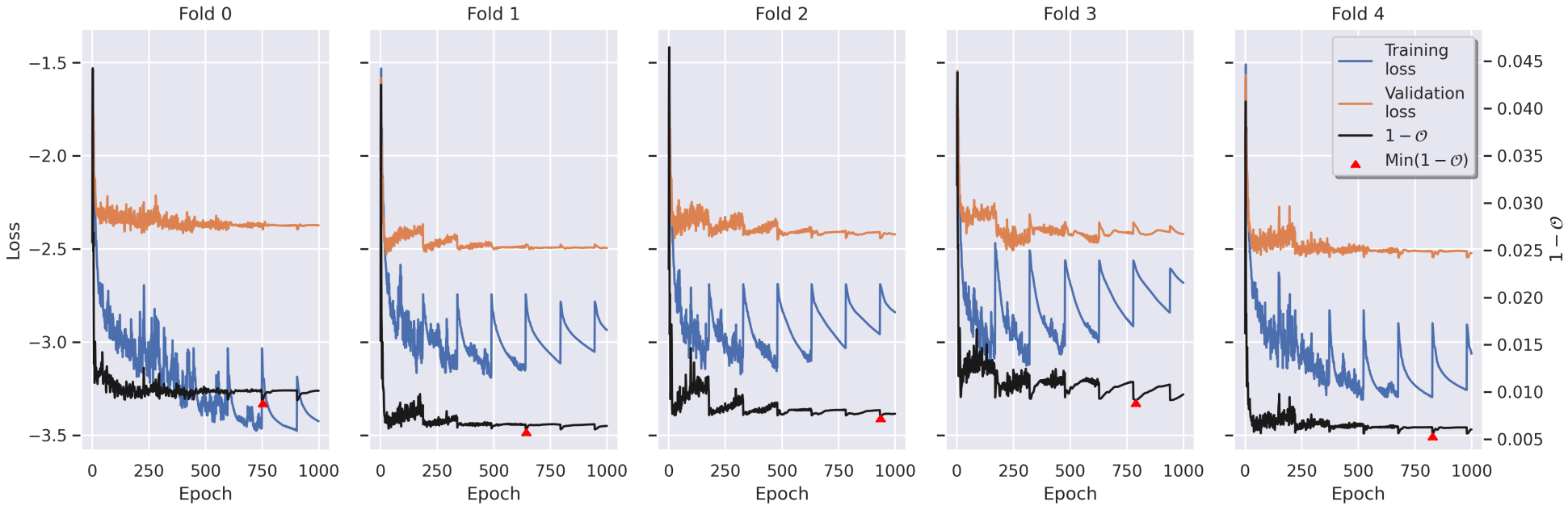
- The pretrained model achieves an average mismatch of about 10^{-5} on approximant data.
- The worst mismatch is of the order of 10^{-2}



Fine-tuning on NR Data

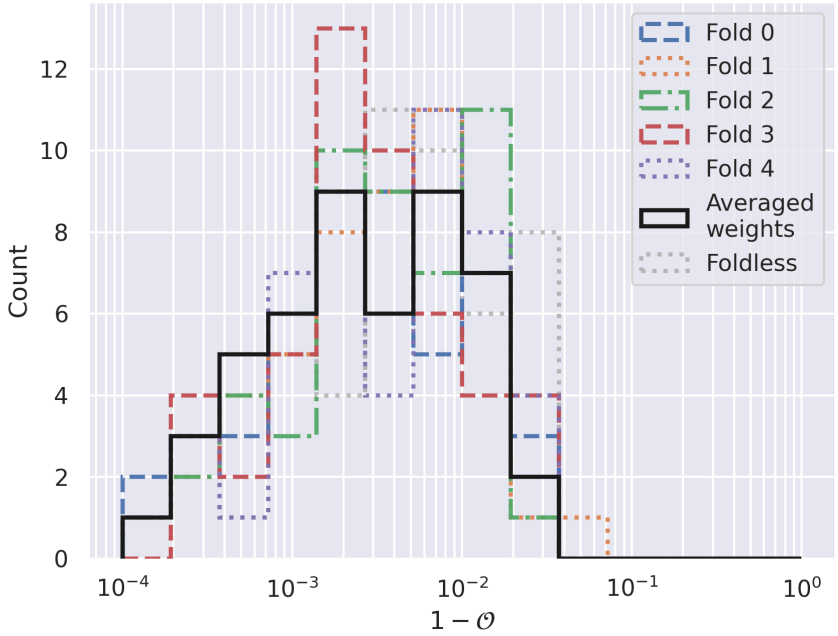
- The model is fine-tuned using 381 waveforms from the SXS collaboration.
- Starting learning rate of 1×10^{-4} keeps model in the weight-space region arrived at in pretraining
- 12.5% of the waveforms are set aside for testing purpose
- A **k-fold weight-averaging** strategy is employed to avoid overfitting and improve generalization, wich each of the 5 used folds having an 80/20 validation split.

Fine-tuning on NR Data

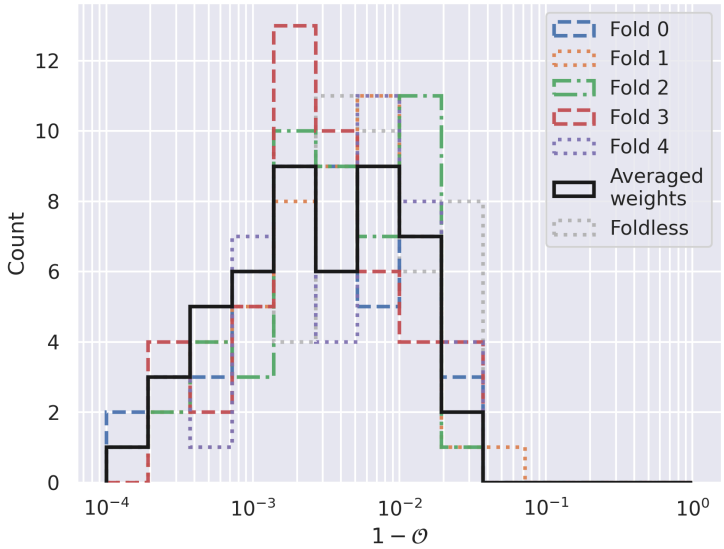


NR Accuracy

- The averaged weight distribution peaks around 5×10^{-3}
- The worst averaged weight mismatch is 2.1×10^{-2}.
- Weight averaging shows improvement over individual training attempts



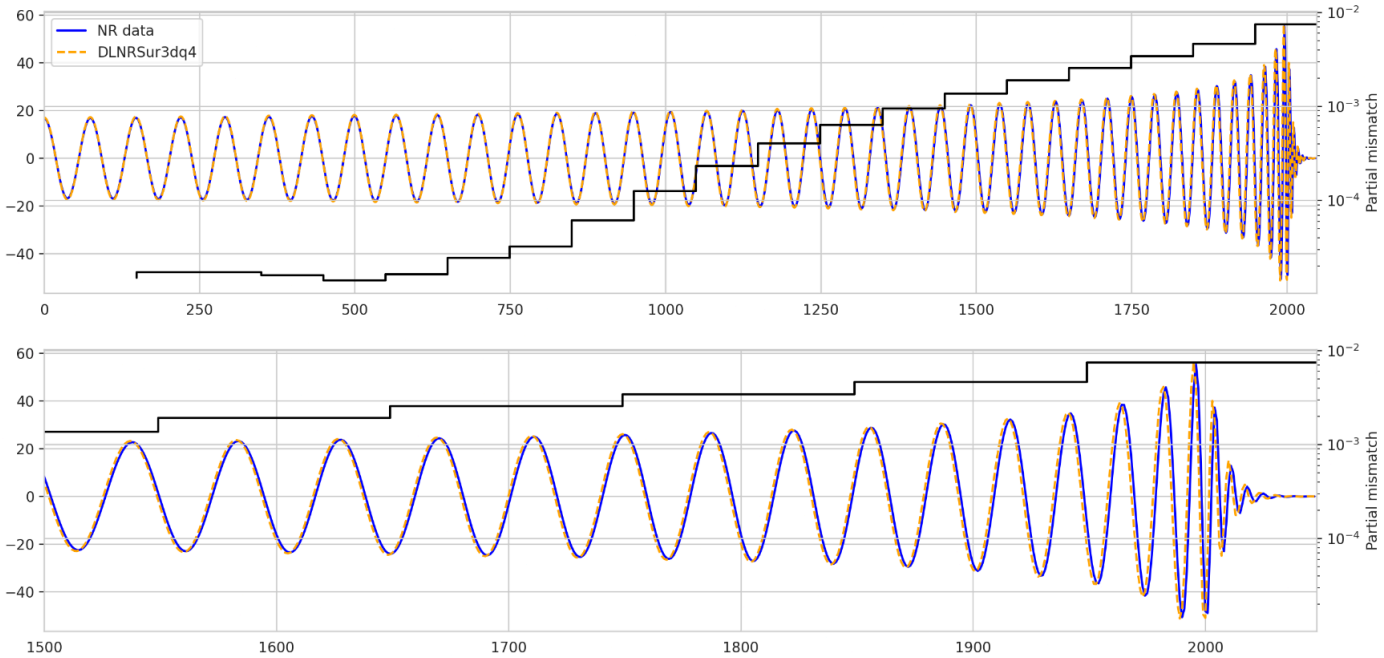
NR Accuracy



	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Avg. Weights	Foldless
Mean mismatch	5.57×10^{-3}	6.50×10^{-3}	5.72×10^{-3}	5.81×10^{-3}	7.15×10^{-3}	$\mathbf{5.43\times10^{-3}}$	8.30×10^{-3}
Max mismatch	2.57×10^{-2}	4.39×10^{-2}	2.16×10^{-2}	2.92×10^{-2}	3.16×10^{-2}	$\mathbf{2.10\times10^{-2}}$	3.31×10^{-2}

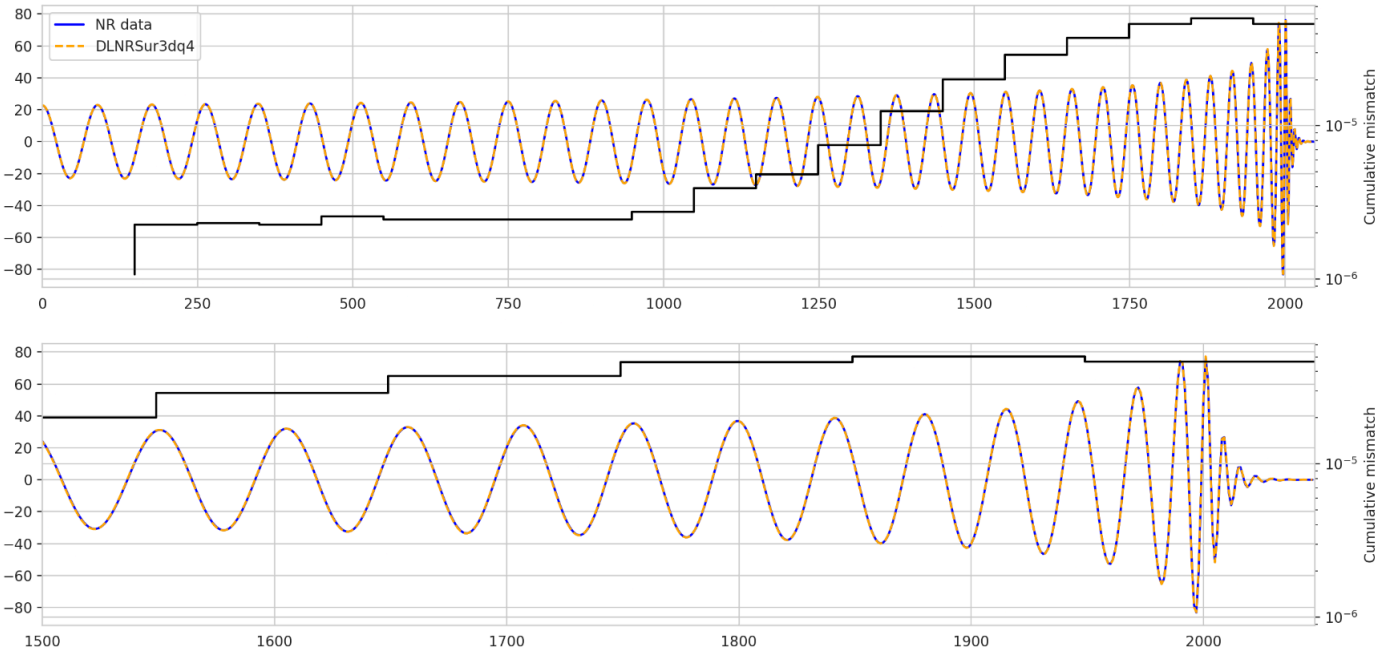
NR Accuracy

$$q = 3, \chi_1 = 0.0, \chi_2 = 0.3$$



NR Accuracy

$$q = 3, \chi_1 = 0.0, \chi_2 = 0.6$$

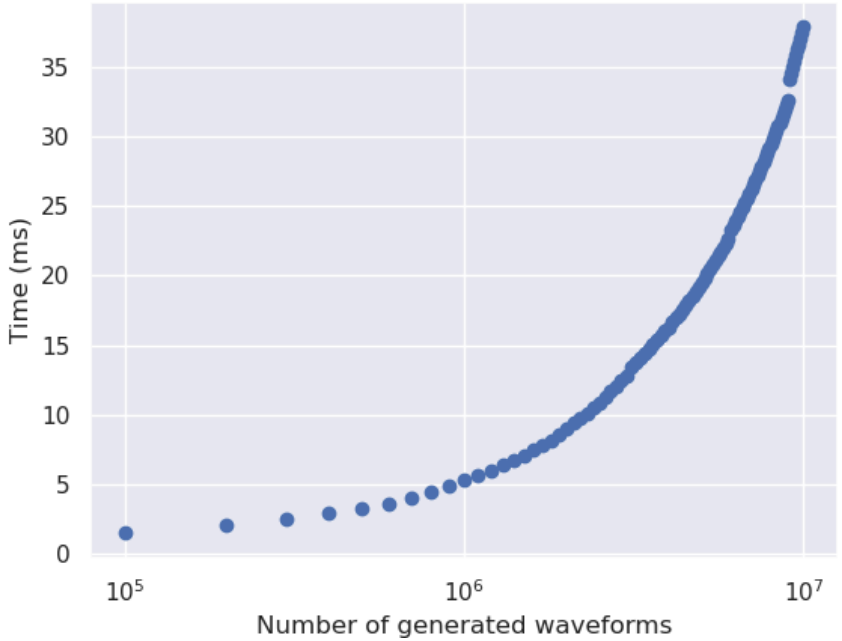


Speed

- DL approximant is highly parallelizable on the GPU.
- CUDA initialization processes take around 500 ms, after which waveform generation is extremely fast.
-

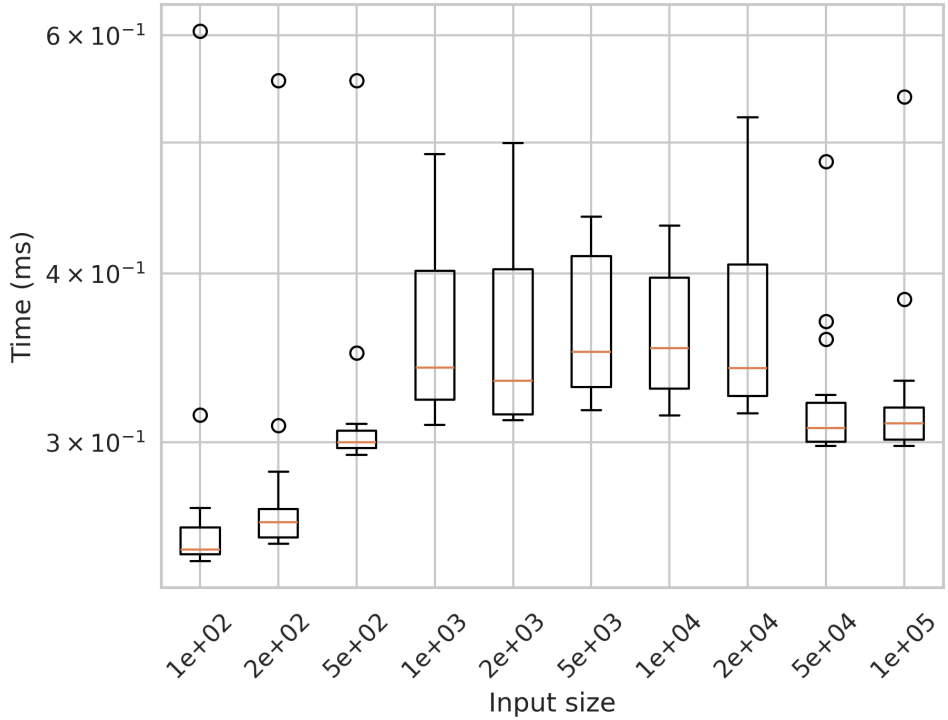
Speed

- DL approximant is highly parallelizable on the GPU.
- CUDA initialization processes take around 500 ms, after which waveform generation is extremely fast.
- Generating 10^7 waveforms (in batches of 10^5) takes under 40 ms.



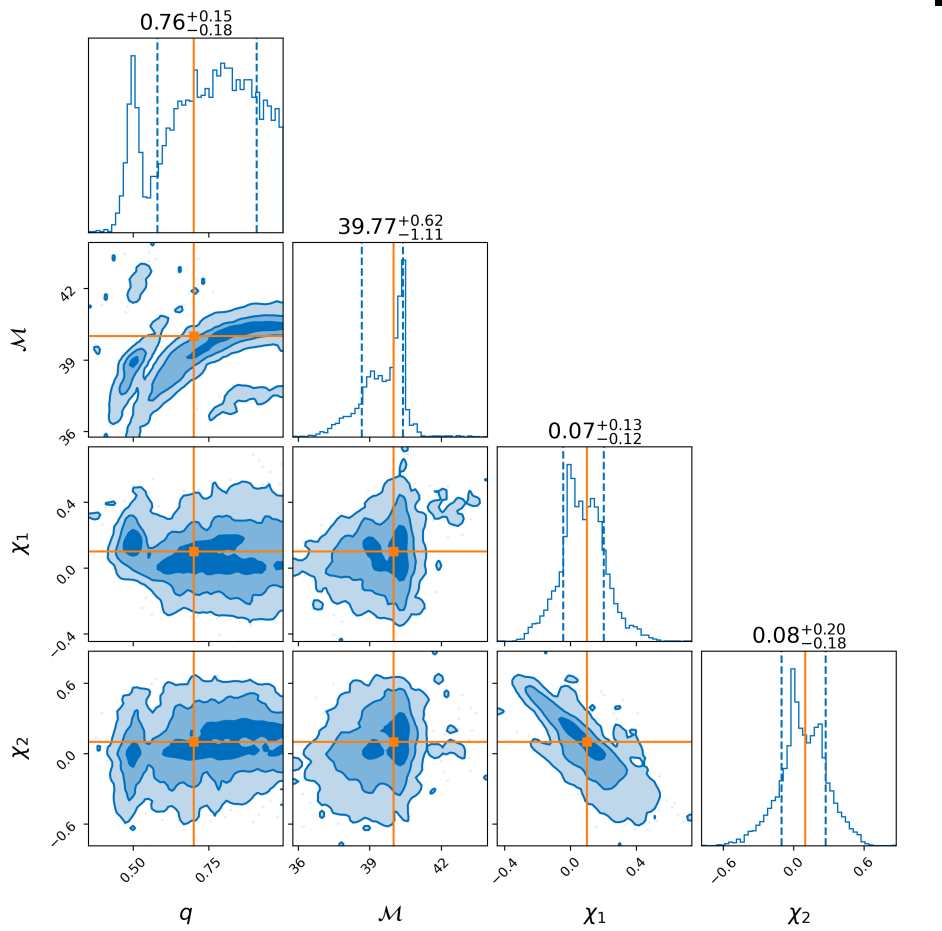
Speed

- DL approximant is highly parallelizable on the GPU.
- CUDA initialization processes take around 500 ms, after which waveform generation is extremely fast.
- Generating 10^7 waveforms (in batches of 10^5) takes under 40 ms.



Usage with bilby

- Usage of NRSurNN3dq4 with bilby is straight forward on the CPU (generation time ~5ms)
- Full GPU support will require deeper changes in how bilby samplers deal with parallelization
- NRSurNN3dq4 shows the ability to accurately recover parameters from an injected signal



Outline

Conclusions

- NRSurNN3dq4 is a deep learning-powered surrogate model that offers both high accuracy and fast generation speeds.
- Could be employed very effectively for the creation of large template banks.
- Through bilby, it is an effective tool for parameter estimation in gravitational-wave astronomy, though better GPU integration is possible.
- Future work: Extend to precessing systems and include higher-order modes for more complex waveforms.

Acknowledgements

OGF is supported by an FCT doctoral scholarship (reference UI/BD/154358/2022).

JAF and ATF are supported by the Spanish Agencia Estatal de Investigaci'on (grant PID2021-125485NB-C21) funded by MCIN/AEI/10.13039/501100011033 and ERDF A way of making Europe.

Further support is provided by the Generalitat Valenciana (grant CIPROM/2022/49), by the EU's Horizon 2020 research and innovation (RISE) programme H2020-MSCA-RISE-2017 (FunFiCO-777740), and by the European Horizon Europe staff exchange (SE) programme HORIZON-MSCA-2021-SE-01 (NewFunFiCO-101086251).

AO is partially supported by FCT, under the Contract CERN/FIS-PAR/0037/2021.

Computations have been performed at the Artemisa cluster (UV-CSIC) co-funded by the European Union through the 2014-2020 FEDER Operative Programme of Comunitat Valenciana, project IDIFEDER/2018/048.

Thank you