PAPER • OPEN ACCESS

GWAK: gravitational-wave anomalous knowledge with recurrent autoencoders

To cite this article: Ryan Raikman *et al* 2024 *Mach. Learn.: Sci. Technol.* **5** 025020

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 025020 https://doi.org/10.1088/2632-2153/ad3a31

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
20 September 2023

REVISED
11 January 2024

ACCEPTED FOR PUBLICATION
3 April 2024

PUBLISHED
25 April 2024

PAPER

GWAK: gravitational-wave anomalous knowledge with recurrent autoencoders

Ryan Raikman[1,5,*], **Eric A Moreno**[2,6,7], **Ekaterina Govorkova**[2,6,7], **Ethan J Marx**[1,2,7], **Alec Gunny**[1,2,7], **William Benoit**[3,7], **Deep Chatterjee**[1,2,7], **Rafia Omer**[3,7], **Muhammed Saleem**[3,7], **Dylan S Rankin**[4,7], **Michael W Coughlin**[3,7], **Philip C Harris**[2,6,7] and **Erik Katsavounidis**[1,2,7]

[1] MIT LIGO Laboratory, Cambridge, MA, United States of America
[2] Massachusetts Institute of Technology, Cambridge, MA, United States of America
[3] University of Minnesota, Minneapolis, MN, United States of America
[4] University of Pennsylvania, Philadelphia, PA, United States of America
[5] Carnegie Mellon University, Pittsburgh, PA, United States of America
[6] The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Cambridge, MA, United States of America
[7] The NSF Institute for Accelerated AI Algorithms for Data-Driven Discovery, Cambridge, MA, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: rraikman@mit.edu

Keywords: machine learning, semi-supervised learning, anomaly detection, gravitational-wave physics, autoencoders

Abstract

Matched-filtering detection techniques for gravitational-wave (GW) signals in ground-based interferometers rely on having well-modeled templates of the GW emission. Such techniques have been traditionally used in searches for compact binary coalescences (CBCs), and have been employed in all known GW detections so far. However, interesting science cases aside from compact mergers do not yet have accurate enough modeling to make matched filtering possible, including core-collapse supernovae and sources where stochasticity may be involved. Therefore the development of techniques to identify sources of these types is of significant interest. In this paper, we present a method of anomaly detection based on deep recurrent autoencoders to enhance the search region to unmodeled transients. We use a semi-supervised strategy that we name '*Gravitational Wave Anomalous Knowledge*' (GWAK). While the semi-supervised approach to this problem entails a potential reduction in accuracy compared to fully supervised methods, it offers a generalizability advantage by enhancing the reach of experimental sensitivity beyond the constraints of pre-defined signal templates. We construct a low-dimensional embedded space using the GWAK method, capturing the physical signatures of distinct signals on each axis of the space. By introducing signal priors that capture some of the salient features of GW signals, we allow for the recovery of sensitivity even when an unmodeled anomaly is encountered. We show that regions of the GWAK space can identify CBCs, detector glitches and also a variety of unmodeled astrophysical sources.

1. Introduction

Since the original observation of gravitational waves (GW) [1] by Advanced LIGO [2] and Advanced VIRGO [3], and with the recent introduction of KAGRA [4], more than 90 GW events [5–7] have been catalogued to date, fundamentally transforming our way of observing the Universe. While all of the detected signals thus far correspond to the coalescence of binary black hole (BBH), binary neutron star (BNS), or black hole—neutron star (BHNS) mergers [5, 8–11], there is growing interest in the detection of unmodeled transient signals, which are not described by any known theoretical waveforms, computationally prohibitive to simulate, or have a stochastic nature. Transient GWs are astrophysical phenomena characterized by short-duration signals, typically lasting from a few milliseconds to several seconds. These signals may originate from sources such as core-collapse supernovae (CCSN) [12], as well as exotic sources such as cosmic strings [13, 14], axion stars [15], neutron star glitches, or primordial black holes [16, 17]. Detection

of such sources can provide unique insights into the extreme and often violent processes in the Universe. For example, GWs from CCSN can offer valuable information about the dynamics of supernova explosions and the properties of dense nuclear matter [18].

Typical methods for detecting GWs rely on matched filtering techniques [19], which compare the observed data to a known signal template. These methods require precise knowledge of the signal prior, such as the waveform and the parameters of the source, to detect the signal. While matched filtering is a well-established and powerful method for detecting gravitational waves, it has certain limitations. For example, matched filtering is sensitive only to signals that match in a discrete grid of templates, and may miss signals with different waveforms or parameters. This makes matched filtering unsuitable for unmodeled transient signal searches, primarily due to their unpredictable nature and the diversity in their waveforms.

Unlike binary mergers, where the waveforms can be accurately predicted by general relativity, transient GWs such as those from CCSN are highly complex and not fully understood [20]. Moreover, many potential transient GW sources, like cosmic strings or primordial black holes remain challenging to define due to the intricate and speculative nature of these phenomena. Recent studies have delved into the dynamics and potential gravitational wave emissions from these sources, but the precise characterization of their signals is still an area of active research and exploration [21, 22]. Hence, there is a growing need to develop innovative detection strategies capable of identifying these elusive and potentially groundbreaking signals.

The GW community has developed several unmodeled approaches that do not rely on a specific waveform model. Some of these currently used by the international GW network (IGWN) include the coherent WaveBurst (cWB) [23, 24], which searches for and reconstructs GW transient signals without relying on a specific waveform model. cWB works by searching for coherent excess power in the data that is consistent with the expected waveform of a gravitational wave burst. The algorithm is sensitive to both unmodeled and modeled burst signals, making it versatile in detecting a range of potential sources. Another framework, oLIB [25] uses the Q transform to decompose data from GW observatories into several time-frequency planes of constant quality factors. The pipeline flags data segments containing excess power and searches for clusters of these segments to identify possible GW candidate events. MLy (read as 'Emily') [26] is a machine-learning-based search for generic sub-second-duration transient GW signals in the 20–500 Hz frequency band; MLy works by utilizing convolutional neural networks (CNNs) trained to recognize signals that are simultaneous and coherent between detectors.

In this paper, we explore the use of a method introduced in [27] deployed within the high energy physics community for the development of an anomaly detection pipeline for data collected by GW observatories. We introduce *'Gravitational Wave Anomalous Knowledge'* (GWAK), a strategy for anomaly search that combines deep learning (DL) techniques with prior information on potential signals to improve the sensitivity of detection. The GWAK algorithm is based on the intuition that unknown transient sources should loosely resemble known signals, as well as be coherent between present GW detectors. We apply the GWAK method to GW datasets, by introducing signal priors that capture some of the salient features of GW signatures, allowing for the recovery of sensitivity even when the observed signal mismatches the known priors.

Because GWAK does not rely on precise prior knowledge of the signal and can detect signals with unknown waveforms or parameters by matching incoming data streams with salient features (cross-correlation, oscillations, etc) that are generic to broad types of GWs, GWAK is more robust and powerful for the detection of unknown signals. As such, it can be used as a complementary approach to methods such as matched filtering for detecting transient GWs, and it has the potential to improve the sensitivity of GW detection systems to sources of this type.

This paper is organized as follows: in section 2, we provide a brief review of DL in GW detection and previous works in machine learning anomaly detection. In section 3.2, we describe the data used for this study. In section 3, we present the GWAK method for constructing embedded spaces for anomalous searches and the autoencoder architectures used to build these embedded spaces. In section 4, we discuss the performance of the GWAK method on real GW data. Conclusions and next steps are provided in section 5.

The code used to analyze data and generate results and plots can be found at[8].

2. Related work

DL approaches for GW detection are well explored [28–38]. However, these methods typically rely on supervised learning techniques, which provide efficiency competitive to templated searches by exploiting neural network nonlinearity and the information provided by ground-truth labels. By construction, these

[8] https://github.com/ML4GW/gwak.

methods rely on a realistic simulation of the signal generated by a specific kind of source, which is assumed upfront. With supervised DL approaches, there is no guarantee of generalizability to an out-of-training event, which we refer to as an 'anomaly'.

For reference, significant GW alerts by IGWN are sent out with a maximum expected false alarm rate (FAR) of 3.9×10^{-7} Hz (minimum one per month) for compact binary coalescence (CBC) sources, and 3.2×10^{-8} Hz (minimum one per year) for unmodeled burst signals that do not have a CBC signature[9].

Autoencoders are commonly implemented for a variety of GW applications. An autoencoder is a type of neural network that compresses from the input space into a significantly smaller latent space and projects back up to the input space. Both of these mappings are optimized to minimize the difference between the input and output, also referred to as reconstruction. This technique serves to efficiently encode salient features of the signal and provide a mechanism for recreating the signal based on those features. They can be used for non-linear subtraction of noise from incoming time-series of GW strain [39–41] in real-time [42]. Additionally, autoencoders are used in generative models, speeding up the computation of GW waveforms relative to very computationally expensive numerical relativity simulations [43]. Finally, due to the transient nature of single-detector artifacts, 'glitches', autoencoders can be used for glitch classification in an unsupervised manner [44].

There have also been explorations into unsupervised GW detection to enhance detection capabilities beyond signal templates and simulation. Initial studies with a one-dimensional CNN based autoencoder [45] and long-short term memory (LSTM) based autoencoder [46] show that unsupervised detection is possible. Both studies rely on learning the typical features of the background. Once a model is trained, it is then used to evaluate the similarity between background priors and new unseen data. To do this, the algorithms use reconstruction loss, computed by comparing the original signal and the signal outputted by the model trained on the background prior, as a detection statistic. By comparing the reconstruction error of new data with a threshold corresponding to an allowed FAR, data points that deviate enough from the normal pattern are identified as anomalous. This relies on the autoencoder's inability to reconstruct any potential signal that deviates from the background, or generally the prior on which the model was trained, triggering a high reconstruction loss. This approach has been shown to effectively detect out-of-training anomalies in GW datasets and has the potential to improve the performance of anomaly detection systems. However, for a specific signal, an algorithm trained with an unsupervised procedure on unlabeled data is typically less accurate than a supervised classifier trained on labeled data.

We build upon this approach by training several autoencoders; each with a different signal at training time. The signals that were chosen to be used in this paper are described in section 3.2. Unlike an earlier study [46] which used a simulated Gaussian background as a proof of concept, our method is trained with real background data. This is significantly more challenging than just simulated Gaussian noise, and therefore no direct comparison of the results presented here to the previous ones can be made.

Although the autoencoder architecture can implicitly learn the importance of the correlation between two detectors for signal identification, a related method explicitly leverages the correlation across detectors [47]. The method learns detector correlation from astrophysical signals from a white-noise burst (WNB) [48, 49] signal prior, which should be harder to detect than any other type of signal given its lack of a distinctive morphology. Similarly to [47], we choose not to rely on our autoencoders to learn the correlation between the two detector sites. Instead, we directly compute and include the correlation in our final metric that is used to find signal events as another axis in the GWAK embedded space. As such, the present version of our method applies to the case of aligned GW detectors. However, more off-plane GW detectors are being proposed, and a future area of work is generalizing the method to an arbitrary detector network.

3. GWAK algorithm

GWAK (reads: *guac*) builds on the concept of semi-supervision, pulling on concepts from both supervised and unsupervised learning. The semi-supervised method is manifested by using simulated signals as approximations for anomalous-unmodeled signals in the GWAK embedded space. A graph representation of this embedded space is shown in figure 1 for the case of two (left) and three (right) different dataset classes. We use five classes of datasets that can help us build an informative space to search for these unmodeled signals. A separate unsupervised autoencoder network is then trained on each class of samples separately, resulting in a low-dimensional 'GWAK' space consisting of the coherence metrics between the autoencoder inputs and outputs for each autoencoder, which is then used to search for anomalous signals. This approach is particularly useful for detecting new physics phenomena, where the signal prior is unknown but the

[9] https://emfollow.docs.ligo.org/userguide/analysis/index.html##alert-threshold.

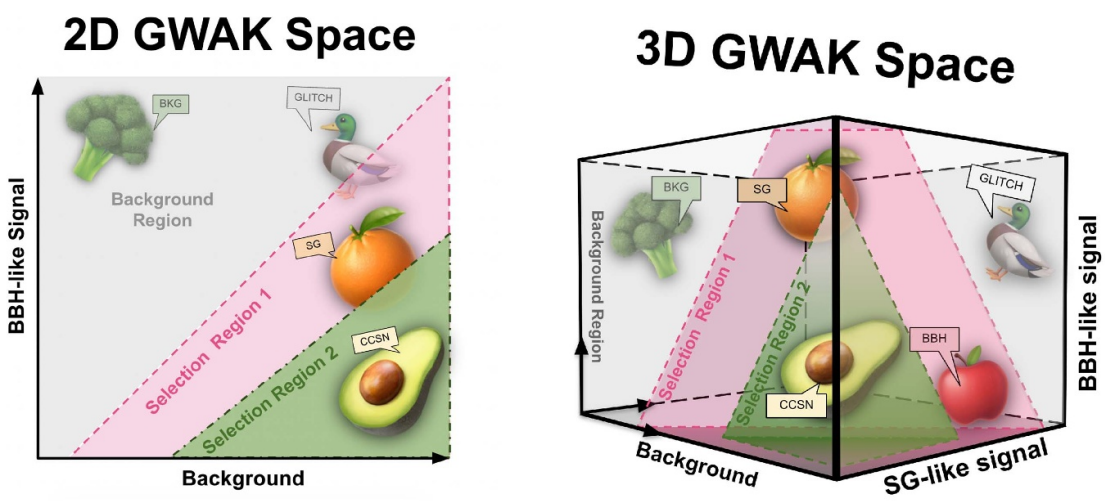

Figure 1. The 2D and 3D schematic depicting the main idea behind the GWAK algorithm. For plotting purposes we only show one background and two signal axes, while in this paper we are using 5 (2 for background and 3 for signals), and the main idea remains the same. All the fruits (orange, avocado and apple) are 'signals' that we would like to select, broccoli represents background that we would like to suppress and the duck represents true anomalies, such as detector malfunctioning, and glitches. Shaded pink and green region depicts selection regions that would correspond to different FARs. For higher FAR (pink shading) more signal-like anomalies would be picked up (orange, avocado and apple), while for a selection that corresponds to a lower FAR (green shading), only 'avocado signal' would be picked up.

simulation of some potential signal pattern, like BBH and sine-Gaussians (SGs), is available. The GWAK method results in classes of anomalous signals inhabiting different regions of the continuous GWAK space, each being reconstructed differently by the five autoencoders. Searches can then be performed on the lower-dimensional embedded space in the regions that anomalies are expected to inhabit.

3.1. Network architectures

This analysis is a continuation of the study detailed in [46] where LSTM-based AE was used. We wanted to study the method of using signal AEs as well as the background ones, so we focused on this and applied minimal modifications to the architecture itself.

One of the main advantages of LSTM autoencoders is their ability to handle sequential data with temporal dependencies. This makes them suitable for anomaly detection in time-series data, such as GWs, speech signals, and sensor data. CNN autoencoders, on the other hand, only capture a smaller window at a time, therefore losing long-term memory. In testing, we saw that across a similar number of parameters, the LSTM models better learned the data as compared to CNN models. The LSTM autoencoder consists of an encoder and a decoder, where the encoder maps the input sequence to a fixed-length vector, and the decoder maps the vector back to the original sequence. We use a similar architecture to that optimized in [46].

For the signal classes, we used the LSTM autoencoder as described above, with the bottleneck size of 4, 8 and 8 for BBH, low and high frequency sine-gaussian (SG) signals, respectively, as shown in figure 2. For the background classes, which consist of data segments absent of signal or with a glitch in either detector, we used a fully connected dense model, as shown in figure 3. This was done as the signal classes have temporal behavior to exploit, whereas glitches have smaller-scale, localized features. In testing, glitch signals were fit better by dense networks than LSTM networks. The total number of trainable parameters for the BBH LSTM AE is 510324, for SG 64–512 Hz and SG 512–1024 Hz are 511672, for background AE 243352 and for glitches AE is 241302.

3.2. Data samples

The dataset used in this study was collected by the LIGO Hanford (H1) and LIGO Livingston (L1) [2] GW detectors during the first half of the third observing run (O3a), which took place between 1 April 2019 and 1 October 2019. We specifically used publicly available data between GPS times of 1238 166 018 and 1238 170 289, right at the beginning of the run. Next, the time-series data were downsampled from 16 384 Hz to 4096 Hz, and processed to remove and create a separate dataset of transient instrumental artifacts (glitches) using the excess power identification algorithm Omicron [50]. We used $Q_{\rm min} = 3.3166$, $Q_{\rm max} = 108$, and $f_{\rm min} = 32$ for the Omicron algorithm. We then took 4 s segments of the data without noise artifacts to serve as the baseline for the injection of signals. As such, we created five different classes of data as proxies for signals and background signatures:

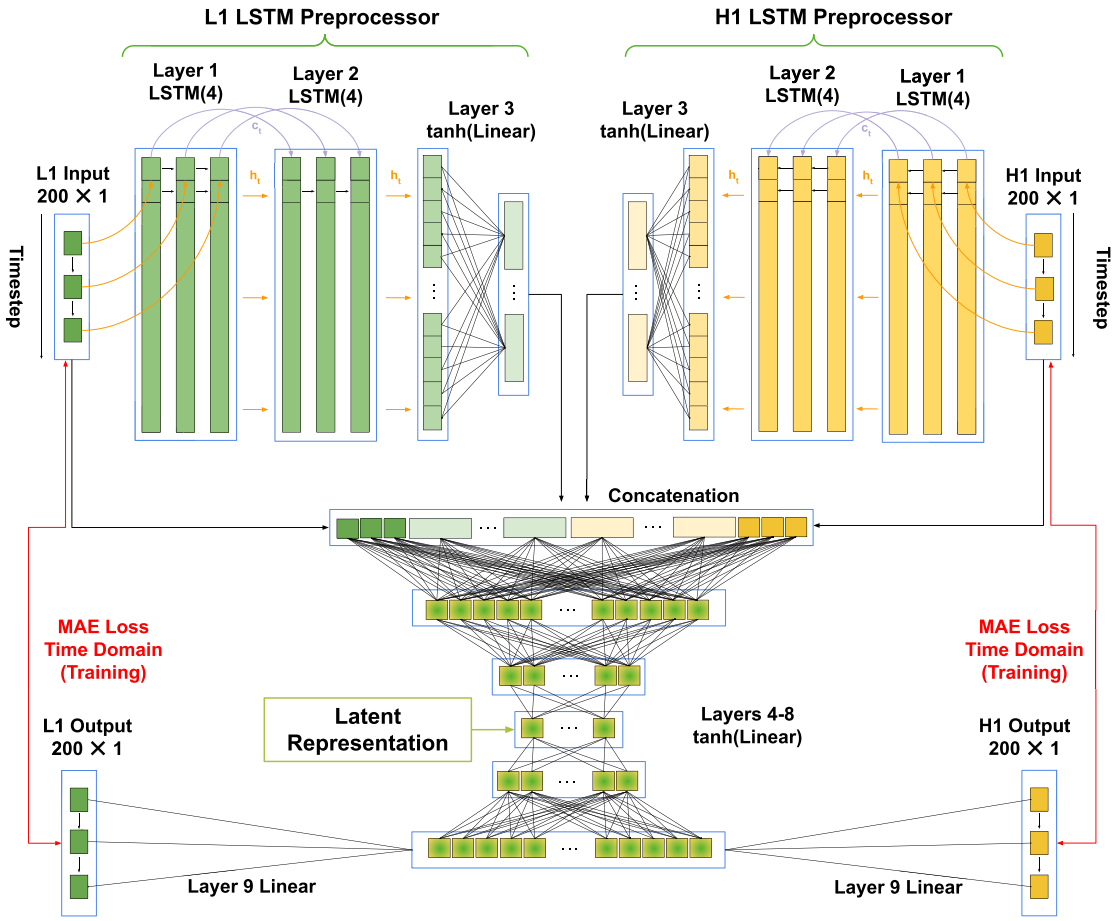

Figure 2. Graphical illustration of LSTM autoencoder, preserving the traditional encoder and decoder structure, which allows for a reconstruction loss detection statistic.

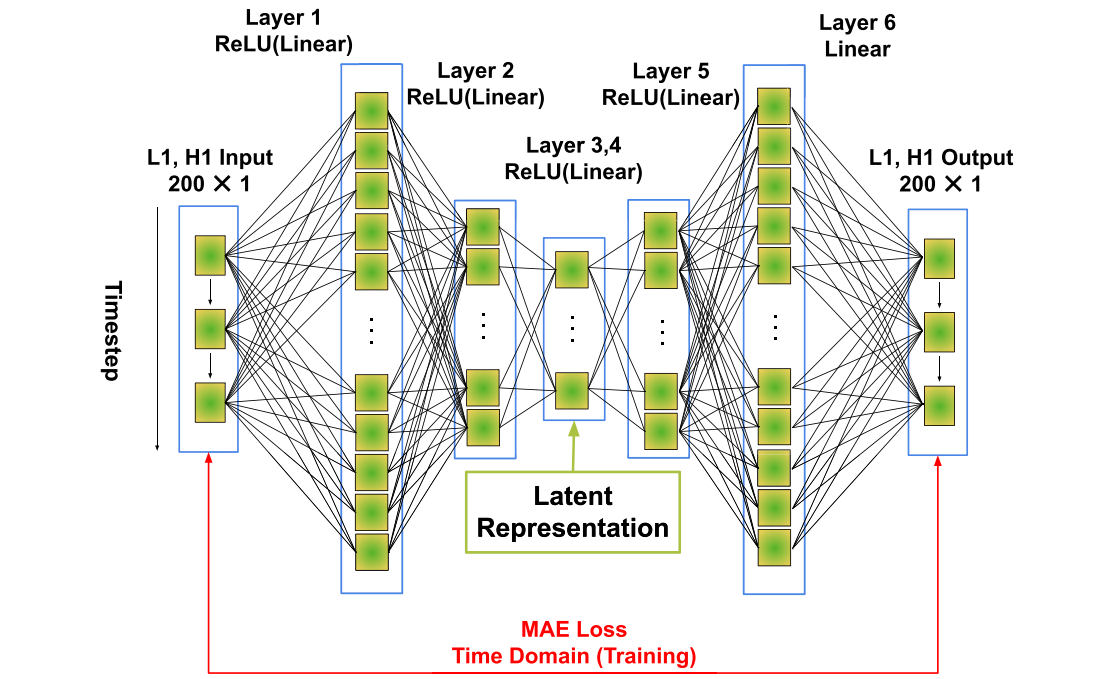

Figure 3. Graphical illustration of DNN autoencoder, preserving the traditional encoder and decoder structure, which allows for a reconstruction loss detection statistic.

Table 1. Sampling parameters and priors for BBH (top) and sine-Gaussian (bottom) injections.

	Parameter	Prior	Limits	Units
BBH	m_1	—	$(5, 100)$	$M_\odot$
	m_2	—	$(5, 100)$	$M_\odot$
	Mass ratio q	Uniform	$(0.125, 1)$	—
	Chirp mass M_c	Uniform	$(25, 100)$	$M_\odot$
	Tilts $\theta_{1,2}$	Sine	$(0, \pi)$	rad.
	Phase ϕ	Uniform	$(0, 2\pi)$	rad.
	Right Ascension	Uniform	$(0, 2\pi)$	rad.
	Declination δ	Cosine	$(-\pi/2, \pi/2)$	rad.
sine-Gaussian	Q	Uniform	$(25, 75)$	—
	Frequency	Uniform	$(64, 512)$ and $(512, 1024)$	Hz
	Phase ϕ	Uniform	$(0, 2\pi)$	rad.
	Right Ascension	Uniform	$(0, 2\pi)$	rad.
	Declination δ	Cosine	$(-\pi/2, \pi/2)$	rad.
	Eccentricity	Uniform	$(0, 0.01)$	—
	Ψ	Uniform	$(0, 2\pi)$	rad.

- **Binaries**—simulated BBH signals using IMRPhenomPv2 [51–53] injected into the real background noise, as shown in figure 5(top left). The simulation parameters are given in table 1. Other binary inspirals (BNS, neutron star black hole) were not used due to much longer characteristic signal time, of order seconds.
- **Background**—background from O3a with DQsecDB[10] state flag `DCS-ANALYSIS_READY_C01:1` applied and excess power glitches [50] and known GW-events removed, as shown in figure 4(bottom).
- **Sine-Gaussian 64 to −512 Hz**—generic low frequency signal model used to simulate generic GW sources, as shown in figure 5(middle left). These are denoted SGLF (sine-gaussian low frequency).
- **Sine-Gaussian 512 to −1024 Hz**—generic high frequency signal model used to simulate generic GW sources, as shown in figure 5(bottom left). These are denoted SGHF (sine-gaussian high frequency).
- **Glitches**—transient instrumental glitches (often of unknown origin) flagged by Omicron [50] as having excess power, as shown in figure 4(top).

To create samples of BBH and SG signals, we used numerical simulations to generate h_+ and $h_\times$ polarization modes. We then sampled sky localizations uniformly in the sky, projected the polarization modes onto the sky location, and injected the projected modes into the two LIGO detectors.

The BBH sample is generated with the parameters and priors [54] as shown in table 1, taken from `bilby` processing spins BBH prior, and the SG samples are generated with the parameters and priors as shown in table 1.

We employed a series of digital signal processing techniques to prepare data for training autoencoders in the context of GW detection. Specifically, we first applied a whitening filter to normalize the data to one hour of surrounding background data. This filter effectively suppressed frequency regions dominated by noise and reduced effects from spectral lines[11, 12]. Moreover, we implemented a band-pass filter within the frequency range of 30–1500 Hz to further attenuate noise outside of the most sensitive frequency range of GW instruments. After applying these filters, we removed 1 s intervals from each end of the data samples to eliminate any edge effects from preprocessing. The remaining 2 s samples, each containing either an injected signal, pure background, a low/high frequency SG, or a glitch artifact, were used to generate training data.

To obtain a set of windows suitable for training, we extracted 200 data points (total duration of 50 ms sampled at 4096 Hz) from each sample. Our experimentation revealed that training the autoencoders with input lengths greater than 200 data points would result in degraded performance, as the model would struggle to recreate these longer input signals. While reducing the number of data points below 200 could enhance computational efficiency, it reduces the autoencoder's ability to learn the evolution of a shorter-duration signal.

To optimize the data processing and facilitate learning by the network, the data are normalized to have a standard deviation of one on a sample-per-sample basis. This normalization was undertaken mainly for the reason that the neural networks struggled to learn with unnormalized samples. The strongest example of this is with the glitch dataset, where strain magnitude can reach amplitudes hundreds to thousands of times

[10] https://git.ligo.org/computing/dqsegdb/client.
[11] https://gwosc.org/s6speclines/.
[12] https://dcc.ligo.org/LIGO-T1500415/public.

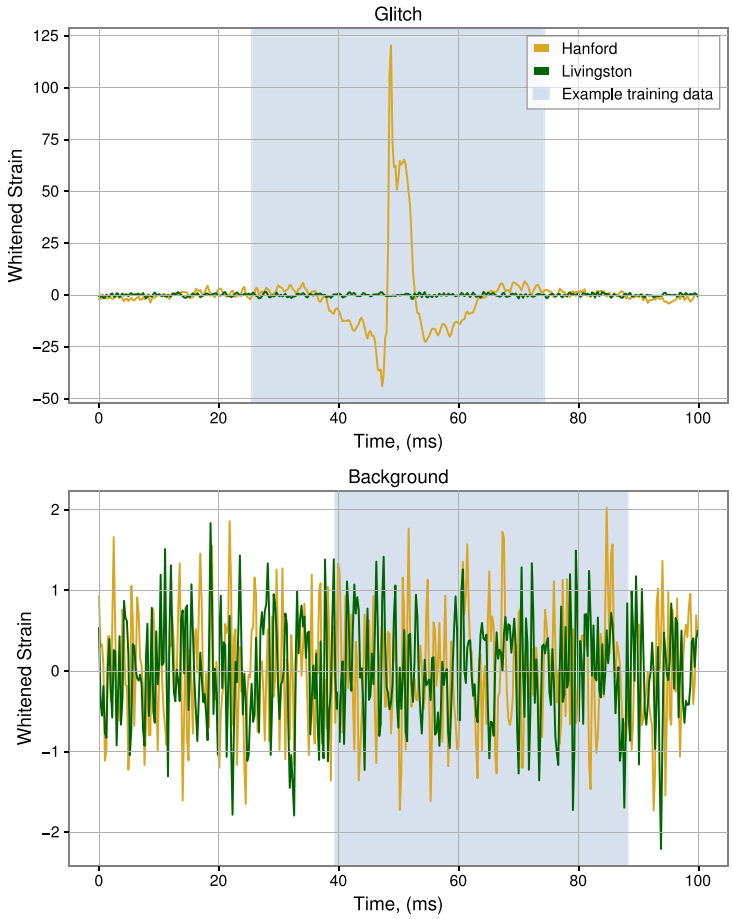

Figure 4. Example of GWAK classes: glitch (top) and background (bottom) strains. The light blue shading highlights an example region that is passed as input to the autoencoders for training.

above the background. The data was not normalized to have a mean of zero, as whitening should remove any constant component.

For each of the non-coherent classes (background and glitch), the dataset is split into three parts: 80 000 training samples (80%), 10 000 validation samples (10%), and 10 000 test samples (10%). For each signal class (BBH, low frequency SG and high frequency SG), we generate 5 sub-datasets, each with a specified signal-to-noise ratio (SNR) injection range, as shown in figure 6. Each of these sub-datasets has an identical splitting procedure: 80 000 training samples (80%), 10 000 validation samples (10%), and 10 000 test samples (10%). The training and validation datasets were used for the autoencoders to create loss values on which to update and score the networks respectively. The test samples were used for the recreation plots, as in figure 7, as well as to show the GWAK feature space, figure 10.

Signal events needed to build the GWAK space were created injecting simulated GWs on top of artifact-free detector noise. This provides an analogous situation to a real GW, in which the strain from the incoming wave is recorded in combination with the detector noise. We do not explore the case of coincident detector artifacts with transient astrophysical signals. The injection of signal also accounts for the difference in GW time-of-arrival at each detector owing to light travel time from the sky localization of the signal, which is significant at a sampling rate of 4096 Hz, corresponding to a maximum of 40 samples.

3.3. Autoencoder training

To train an autoencoder, the input sequence is passed through the encoder and the decoder, and the model is optimized to minimize the reconstruction error between the input and the output sequence. Once the model is trained, it can be used to identify data points that deviate from the normal pattern by comparing the reconstruction error of new data with a threshold.

To train our five autoencoders, each corresponding to one of the data classes, we used two different schemes. To train the glitch and background classes, we use the same dataset for all 200 epochs, the Adam [55] optimizer, and mean absolute error (MAE) loss, computed between the autoencoder input and autoencoder reconstruction. To train the BBH, SG low frequency and high frequency classes, we used a

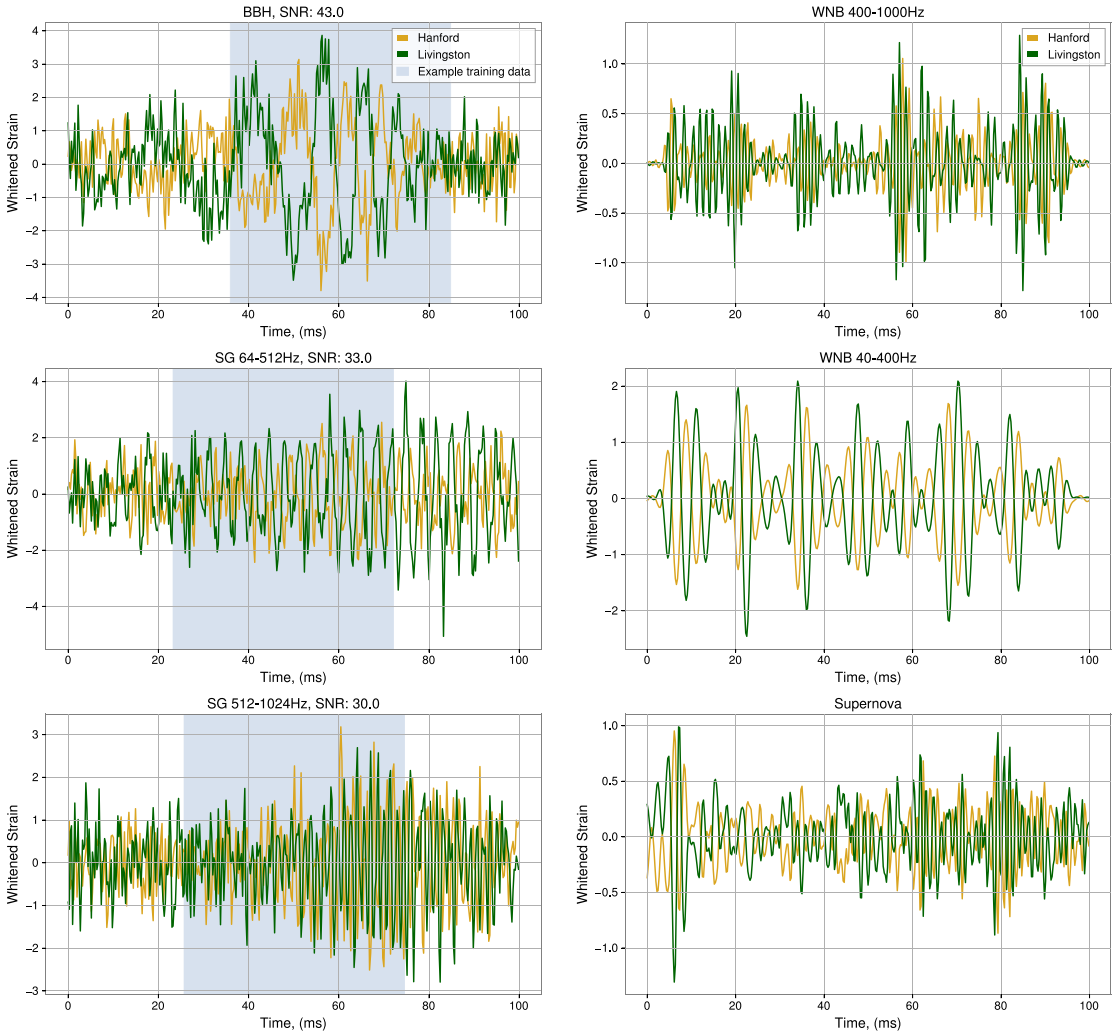

Figure 5. Example of signal-like classes: BBH (top left), WNB (top, middle right), sine-gaussian (middle, bottom left) and Supernova (bottom right) strains. The light blue shading highlights an example region that is passed as input to the autoencoders for training.

curriculum scheme. A curriculum scheme dictates a change in provided training data as a function of epoch. Generally, the model begins with learning an easier dataset, with louder signals, and progresses to quieter signals. When training directly on the lowest SNR dataset, convergence would take more epochs and be less stable. Over 200 epochs, we used 5 different datasets, each with identical injections but different uniform SNR priors: $U[192, 384]$ (epochs 1-40), $U[96, 192]$ (epochs 41-80), $U[8, 96]$ (epochs 81-120), $U[24, 48]$ (epochs 121-160), $U[12, 24]$ (epochs 161-200), as shown in figure 6. Here, we used the Adam optimizer, reset after each step of the curriculum, and computed error between the autoencoder output of a noisy input (injection into the real background with specified SNR) and the 'clean,' noise-less template. This was intended to train the autoencoders to learn to reconstruct the signal itself, without any noise. To compute the validation loss at each epoch, we used a subset of the $U[12, 24]$ SNR dataset for each curriculum. This was intended to provide a fair computation of the validation loss across each curriculum. The MAE loss $\mathbf{L}$ between original data $\mathbf{D}$ and reconstruction $\mathbf{R}$ is given by

$$\mathbf{L} = \frac{1}{N}\frac{1}{2}\frac{1}{200}\sum_{i=1}^{N}\sum_{j=1,2}\sum_{k=1}^{200}|\mathbf{D}_i\,[j,k] - \mathbf{R}_i\,[j,k]\,|.$$

$\mathbf{D}_i\,[j,k]$ represents the ith sample of the original data, taken from the jth detector at the kth timestep, and likewise $\mathbf{R}_i\,[j,k]$ represents the ith sample of the autoencoder output, taken from the jth detector at the kth timestep. The loss curves are shown in figure 6. The example of autoencoder reconstruction obtained with pre-trained autoencoders is shown in figure 7, and recreation samples of other training classes are shown at the end of the paper.

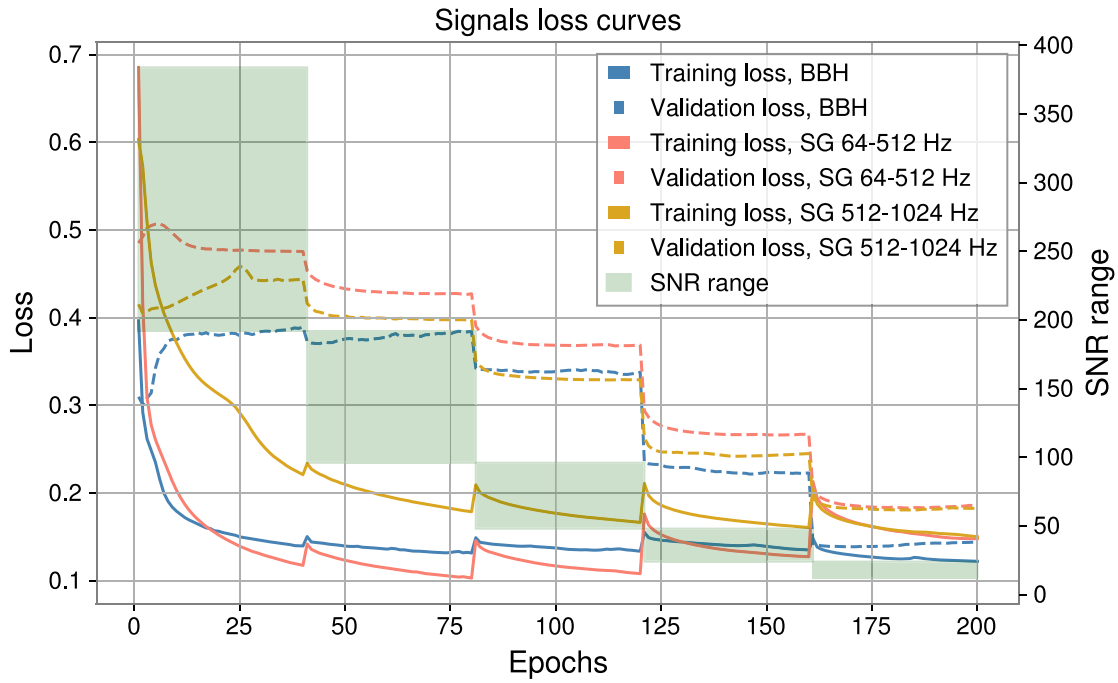

Figure 6. Autoencoder training and validation losses for signal classes, using curriculum learning to progressively reduce validation loss. The validation loss for each training SNR range is computed on the validation data from the last SNR step. In solid/dashed colors are the training/validation losses for BBH (blue), SG 64–512 Hz (salmon) and SG 512–1024 Hz (dark yellow). A light green shaded region depicts the SNR range for each step of training, spanning the range of injected SNR corresponding to each curriculum. For example, the first curriculum contains injections in the SNR range [192, 384].

In figure 6, we see that at the transition between curricula, there is a small spike in the training loss, especially for the SG 512–1024 Hz model. This is due to the transition to a 'more difficult' training dataset, as the lower SNR leads to a less distinguishable signal. With the transition between curricula, we see a sharp drop in validation loss throughout a few epochs. This is because the validation set is maintained for each autoencoder through training as belonging to the lowest—$U[12, 24]$ SNR range. With the transition to a new curriculum with a lower SNR range, the training data for that curriculum will more closely match the validation set, explaining the rapid drop. In testing, we confirmed that the model does not struggle on high SNR data after being trained on low SNR data.

To ensure that autoencoders were sized and trained properly, where they would struggle to recreate anomalies but be able to recreate their trained input, we looked for inflection points in training loss as compared to the overall model size. A model too small is unable to capture enough information about the trained input, whereas a model too large will simply approach an identity function and be able to recreate anything. The inflection point represented the point at which decreasing model size led to significantly poorer performance in the background classes. As such, model sizes were chosen right above that inflection point.

3.4. Feature extraction

While the MAE loss was used in training, at evaluation time we opted for a frequency-domain-based features, computed between the input and autoencoder output. The intuition for choosing to compute features in the frequency domain is based on the fact that signal autoencoders were trained with clean, noiseless targets. Since the models were trained to fit noise-less signals, the outputs solely represent the signal aspects of the input. Within a 50 ms window, signals occupy relatively localized regions of frequency space, i.e. a narrow bandwidth, while the whitened background noise stretches the entirety of the frequency band. The MAE between the original input and reconstructed output is not used as a feature directly because the presence of noise outside the frequency ranges of the signal will inflate the loss. This is not present at training time, the MAE loss is computed between the output and a noiseless signal. To bypass this, we chose to compute a dot product in frequency space. In the frequency regime where the true signal exists, both original and reconstructed signals will be similar, and as such the dot product will yield a high value. In the other frequency regimes where the signal is not present, the reconstructed signal is close to zero, and as such these noisy contributions get removed.

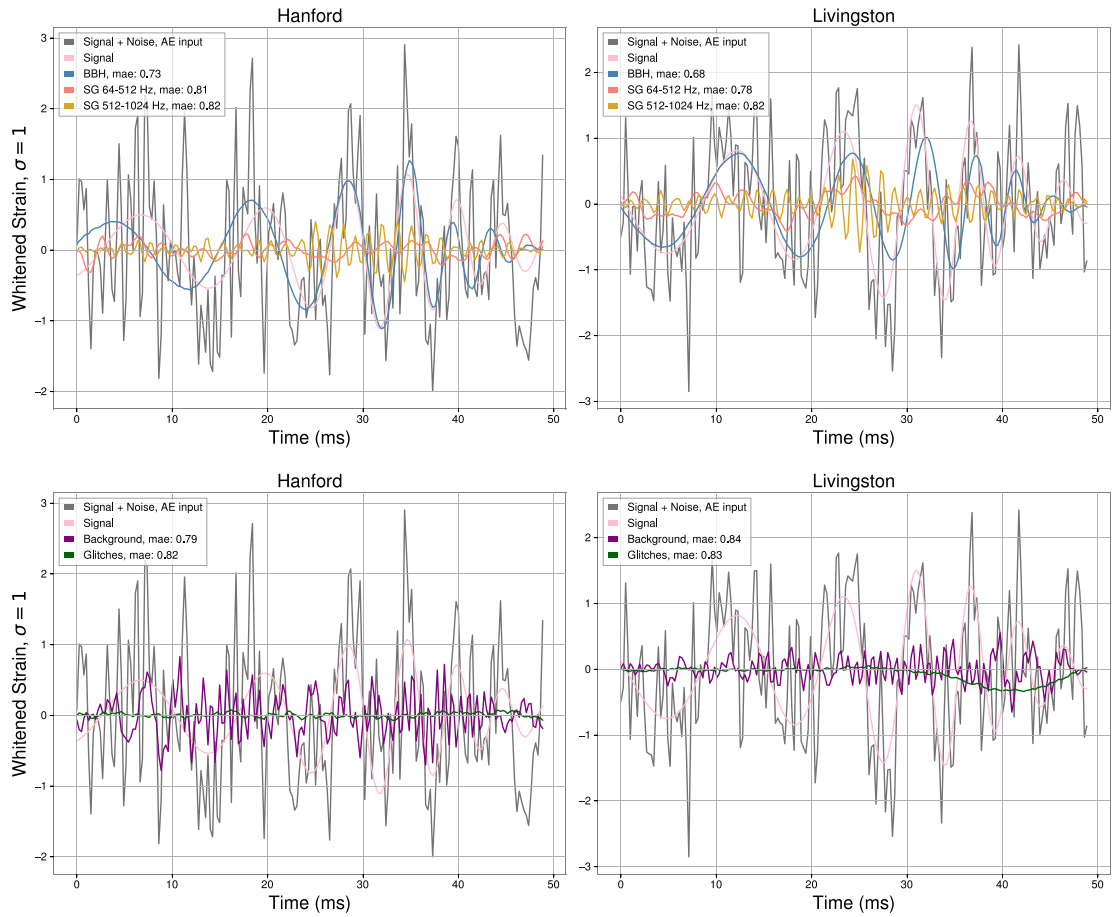

Figure 7. Example of recreation on injected BBH signal, with the noise-less template also shown. The recreation of the BBH autoencoder (in blue) follows closely the original signal injection (in pink) as expected, since it was trained to reconstruct the noiseless input. While background (in purple), glitches (in green), SG 64–512 Hz and SG 512–1024 Hz fail to reconstruct the injected BBH signal, as expected. In the Hanford recreation, a phase offset can be seen between the pure signal and reconstruction. This can be due to the relatively lower amplitude signal in the Hanford detector, leading to more difficulty in separating background from noise.

We choose two features per autoencoder to be the following: let H_O, L_O, H_R, L_R correspond to the original Hanford and Livingston signals and reconstructed Hanford and Livingston signals respectively, from a single autoencoder. Each are 200-datapoint segments, sampled at 4096 Hz. We then take the Fourier transform of each, yielding $\widetilde{H_O}, \widetilde{L_O}, \widetilde{H_R}, \widetilde{L_R}$. The two features, per autoencoder, are $|\widetilde{H_O} \cdot \widetilde{H_R}|$ and $|\widetilde{L_O} \cdot \widetilde{L_R}|$. We also used a general 'frequency space correlation' feature to compliment the Pearson correlation 3.5, defined by $|\widetilde{H_O} \cdot \widetilde{L_O}|$. $|\widetilde{A} \cdot \widetilde{B}|$ represents the magnitude of the dot product of two complex vectors, namely the Fourier transforms of A and B. Training an autoencoder to optimize directly on the $|\widetilde{H_O} \cdot \widetilde{H_R}|$, $|\widetilde{L_O} \cdot \widetilde{L_R}|$ features proved to be too simple of a task. Each autoencoder would consistently learn the largest feature in Fourier space, leading autoencoders to generalize too well, i.e. not being specific enough to their respective training class. By scoring the network's performance in the time domain, i.e. with MAE between the input and reconstructed output, the network must accurately recreate more details, such as temporal offsets, signal evolutions, as well as the corresponding frequency components, forcing the specificity to the class. Finally, by training with MAE, it provides us with a nice visual picture of the autoencoder output which we can easily compare against the input, but this would not necessarily be the case of using the frequency-domain features directly for loss.

3.5. Pearson cross-correlation

To derive a comprehensive metric for inference, we incorporated information regarding the cross-correlation between the two detector sites, in conjunction with the GWAK information. Given that any astrophysical signal will invariably manifest in both detector sites, the correlation of the measured strains is of critical significance for the signal search procedure. Although we utilized information from both detectors during the GWAK space training phase, we opted to directly incorporate cross-correlation information in our final metric.

To accomplish this, we employed the Pearson correlation coefficient [56]. Specifically, we computed the Pearson correlation coefficient between the Hanford and inverted Livingston sites by selecting the maximum correlation coefficient from all possible time shifts for a 200 datapoint window. Since the physical separation between the detectors is about 3000 km, corresponding to a time of flight of 10 ms, we must iterate over all possible temporal shifts within 10 ms to contain the correct time delay. At a sampling rate of 4096 Hz, this corresponds to a shift of 40 data points in either direction. This iteration over all possible time shifts is an advantage of the Pearson correlation coefficient over the frequency-domain correlation coefficient, and as such we chose to include both in our GWAK space. The source of this time delay is due to the sky location of the gravitational wave source. The Livingston detector is inverted, i.e. the output values are reflected over the x-axis, to account for the detectors' relative orientations [1]. The Pearson correlation coefficient is a widely used statistical measure that provides a measure of the strength of a linear relationship between two variables, in this case, the strain measurements from the two sites. The coefficient ranges from -1 to 1, with values close to 1 indicating a strong positive correlation, values close to -1 indicating a strong negative correlation, and values close to 0 indicating a lack of correlation between the two variables.

Given two detector streams, the Pearson cross-correlation at time t was computed via

$$\mathbf{P} = \max_{\Delta} \frac{\sum_{k=t-100}^{t+100} (H_k - < H >) \cdot (L_{k-\Delta} - < L >) \cdot (-1)}{\sqrt{\left(\sum_{i=t-100}^{t+100} (H_i - < H >)^2\right) \left(\sum_{j=t-100-\Delta}^{t+100-\Delta} \left(L_j - < L >\right)^2\right)}},$$

$$< H >= \frac{1}{200} \sum_{i=t-100}^{t+100} H_i, \; < L >= \frac{1}{200} \sum_{j=t-100-\Delta}^{t+100-\Delta} L_j$$

The presence of a multiplicative factor of -1 serves as the inversion of the Livingston detector due to the relative orientation. The range of Δ is within the maximum time of flight in units of data points, so $\Delta \in [-40, 40]$. In addition to the autoencoder frequency domain features, this yields 5 (autoencoders) $\times 2$ (features per autoencoder) $+1$ (pearson) $+1$ (frequency-domain correlation) $= 12$ overall features.

While the timeslides Pearson correlation metric was used in addition to the frequency-domain correlation feature at testing time, experimentation showed that the two values were often very similar, down to a constant scaling factor. The learned coefficients seemed to also prefer the frequency correlation (figure 8) to the pearson correlation, which is slightly surprising as the pearson feature had the advantage of being timeslides outside the 200 datapoint window. While we chose to maintain the Pearson feature in our results, it could be potentially removed to reduce runtime.

3.6. Artificial background with timeslides

To create background data we used a technique called timeslides. This involved temporally shifting the data from one detector relative to another by at least 10 ms, the light travel time between detectors, guaranteeing that there is no astrophysical correlation present. As a background dataset for 3.7, we computed 8 h worth of timeslides. To evaluate our entire algorithm and report false alarm rates for detections, we computed 1 year worth of timeslides.

3.7. Linear combination optimization

To construct a final metric that combines the information from the above 12 features, we opted for a linear combination of those values to produce one final metric value. Given that we are trying to generalize to unknown anomalous signals, opting for a simple linear model aims to reduce any bias towards known signal regions. To find optimal values for the parameters of the linear classifier, we used a simple linear support vector machine (SVM), which aims to optimize the binary classification of background and signal classes. The classification by a Linear SVM is simply given by $\vec{W}^T\vec{X} + b$, where $\vec{W}$ represents the learned weight vector, of the same dimensionality as the GWAK vector $\vec{X}$, which contains the 12 GWAK features. b represents a bias term. For the background dataset, we used 8 h worth of timeslides as described in 3.6. Also using this stretch of timeslides, we computed a set of normalization coefficients. For each of the 12 features, these were the mean value and standard deviation of that feature across 8 h of analysis. These were used to rescale each feature to mean zero and standard deviation one independently. This helped with training the linear classifier, as while the Pearson correlation coefficient is O(1), the frequency domain coefficients can be O(1000). For the signal dataset, we generated a new dataset, comprising of 6 classes—BBH, SG (64–512 Hz), SG (512–1024 Hz), low-frequency white noise burst (40–400 Hz), high frequency white noise burst (400–1000 Hz), and supernova [57] (85 $M_\odot$ progenitor mass, SFHo equation of state), each with an SNR prior of $U(10, 100)$. During training, signals were labeled as 0, and backgrounds were labeled as 1. As this model lacked any kind of normalizing layer, simply providing the distance from a learned hyperplane, final

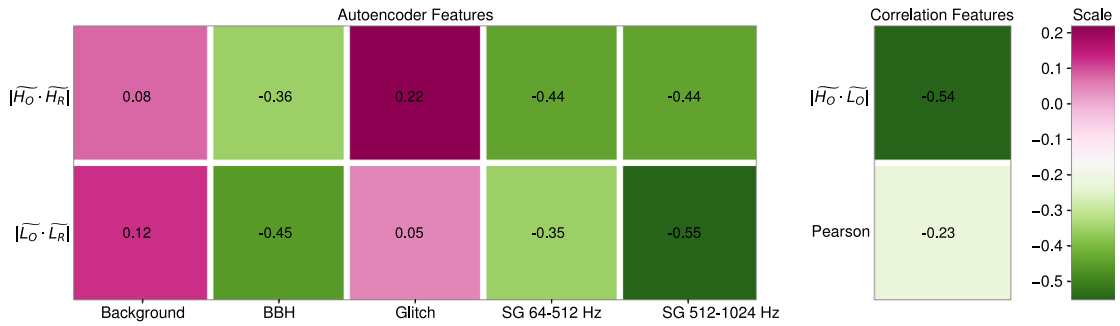

Figure 8. Learned coefficients of a linear combination of the 12 GWAK features. As expected, the features corresponding to the signal autoencoders have more negative values, as well as the correlation features, as a higher correlation score is more representative of an astrophysical event. While background and glitch features have more positive values since a stronger relationship to those classes will be less indicative of the signal.

output values were not limited to $[0, 1]$. This is useful, as it assigns a relative score, easily capable of classifying the anomalous 'strength'.

We trained this linear fit for 5000 epochs, with a learning rate of 0.01 and using the Adam optimizer. Figure 8 shows the learned coefficients for each of the 12 GWAK features. This serves as a sanity check that the impact of each feature on the final metric score is as we expect. Signals are classified via a more negative value, so the features corresponding to the signal autoencoders should have more negative values, as well as the correlation features, as a higher correlation score is more representative of an astrophysical event. On the contrary, autoencoders corresponding to non-astrophysical data (background or glitch) should have more positive values, since a stronger relationship to those classes will be less indicative of the signal. As a sanity check, these intuitions are all demonstrated via figure 8. The model also included a free parameter for the bias, which serves to enable optimal learning. Intuitively, we expect this value to be greater than zero, as almost all features (except for the pearson correlation) are positive. In testing, we found a learned bias value of $\sim$4, varying between training runs.

3.8. Smoothing for the final metric

As the GWAK algorithm assigns a single metric value to each 50 ms window, it does not naively carry sensitivity to signals longer than the 50 ms window. A longer signal will have multiple evaluation points, but to compute FAR only the lowest metric value is taken. To increase sensitivity by specifying the GWAK algorithm to signals greater than one window in length, we convolved the timeseries of final metric evaluations with uniform kernels of varying size, with the idea being that the sensitivity to a given anomalous signal is maximized when the kernel length is of order the signal length. We present the detection efficiency using this method in figure 15.

4. Results

This section describes how our proposed methodology can be used to discover anomalous events. Additionally, we evaluate the effectiveness of our semi-supervised approach in detecting several potential sources of GWs, without using information about these signals during the training phase.

4.1. Background and glitch mitigation

As the goal of this method is to identify anomalous signals in the background, the mitigation of non-astrophysical data being identified as signal-like improves sensitivity to real signals by reducing the corresponding false alarm rate. In particular, detector artifacts or 'glitches' can often pose a problem as they can have signal-like morphology [58, 59] as well as possess high SNR in a single detector. This serves as the motivation for training a glitch autoencoder, as it should be able to recognize single detector artifacts reliably. Upon 'recognition' of a glitch, one of the glitch autoencoder frequency domain features will yield a large positive value, therefore a positive contribution to the final metric score. Since signals are characterized by negative final metric scores, this serves to down-weight the significance of the glitch. In addition, as glitches are uncorrelated between detectors, occurring locally, the use of both correlation features also helps to mitigate false alarms caused by glitches. As there is no correlation between observations in one detector stream and another within the maximum light travel distance during a glitch, those features will correspondingly have small values, and similarly down-weight the glitch to have a less signal-like score.

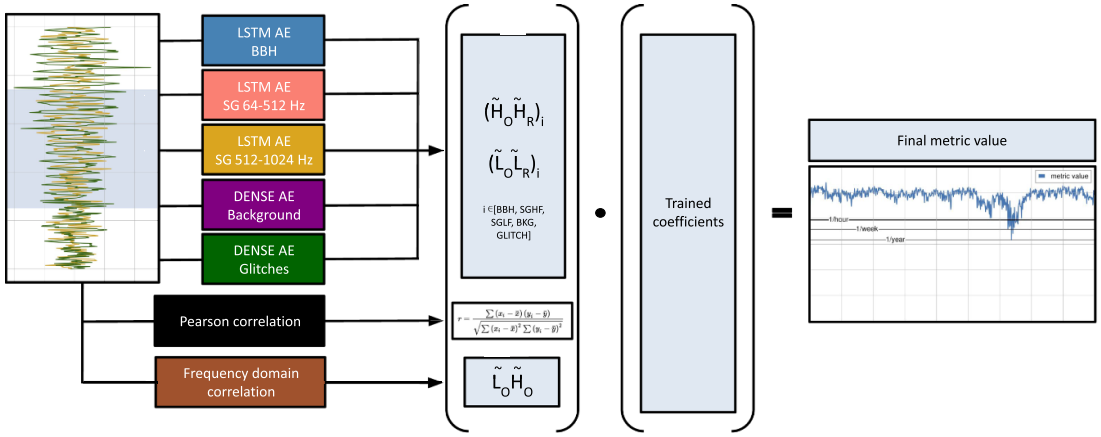

Figure 9. Example of metric calculation on a hypothetical event. The event is reconstructed with each of the five pre-trained autoencoders. Pearson and frequency domain correlation are computed on the given input and then each of the values is multiplied with a corresponding coefficient which arises from the pre-trained linear metric. The sum of all the features multiplied by their coefficients is referred to as the final metric and is used to decide on whether the event is signal-like.

4.2. Anomaly metric

We employ the linear combination method outlined in section 3.7, which includes the two frequency-domain features per autoencoder, the frequency-domain correlation, and the Pearson cross-correlation presented in section 3.5. The example of a full GWAK pipeline is shown in figure 9.

The resulting coefficients of the linear classifier, physically describing a hyperplane, are then used to project points from the 12-dimensional feature space down to a one-dimensional space via the dot product, or geometrically the perpendicular distance from that point to the classifying hyperplane. This one-dimensional real number is our final metric value. Since our classifier was trained to predict 0 for signals and 1 for backgrounds, a more negative final metric value corresponds to a more 'signal-like' or louder input, whereas a less negative/positive metric value indicates background or no signal of interest. Once we compute the final metric value for a hypothetical signal, we would like to know the corresponding false alarm rate. This corresponds to the frequency that a non-astrophysical input (glitch or otherwise) from the GW detectors would lead to a trigger of the same significance as the hypothetical signal. In figure 10, we show a lower-dimensional representation of the GWAK space. This is done by taking the two features corresponding to each trained class and adding them, yielding a total contribution from the given class. The correlation features are not shown. The relative values on the axis denote the contribution to the final metric value. As expected, signals occupy negative values, and backgrounds occupy positive values. We show each of the five data classes, taken from a validation dataset, in this space to highlight how strongly the signals are separated and grouped in this new, learned GWAK space.

4.3. Evaluation on core-collapse supernova

We use existing CCSN simulations to see how our approach extends to anomalous signals which GWAK has not seen during training. We use [57] (85 $M_{\odot}$ progenitor mass, SFHo equation of state). In the middle of figure 11, the evaluation of GWAK axes and Pearson correlation with time are shown. On the bottom, the total metric value and FAR are shown as an example of the algorithm's 'reaction' to unseen signals. Both BBH and SG losses drop at the time of the signal, which indicates that the strain at that moment is more signal-like. The Pearson correlation increases indicating a strong correlation between the two detector sides. The FAR at the moment of the event drops to a level consistent with one or fewer events per month, which means that even with strong trigger restrictions, detection of that type of event would be possible with our new proposed algorithm.

The bottom plot in figure 11(left) reflects the scan of minimum metric value for two different SNR injected core-collapse supernova models. As expected, with the higher SNR, the total metric is lower resulting in a lower FAR. If the trigger bandwidth would allow for up to 1 false event per hour being triggered, GWAK would be able to detect core-collapse supernova events with ∼22 SNR.

4.4. Evaluation on white noise bursts

Furthermore, we assessed the performance of our method on WNBs, which are signals characterized by the presence of h^{+} and $h^{\times}$ polarizations that are independent time series of Gaussian noise, which is whitened over a specific frequency range and multiplied by a sigmoid envelope. The bandwidth of each injected signal

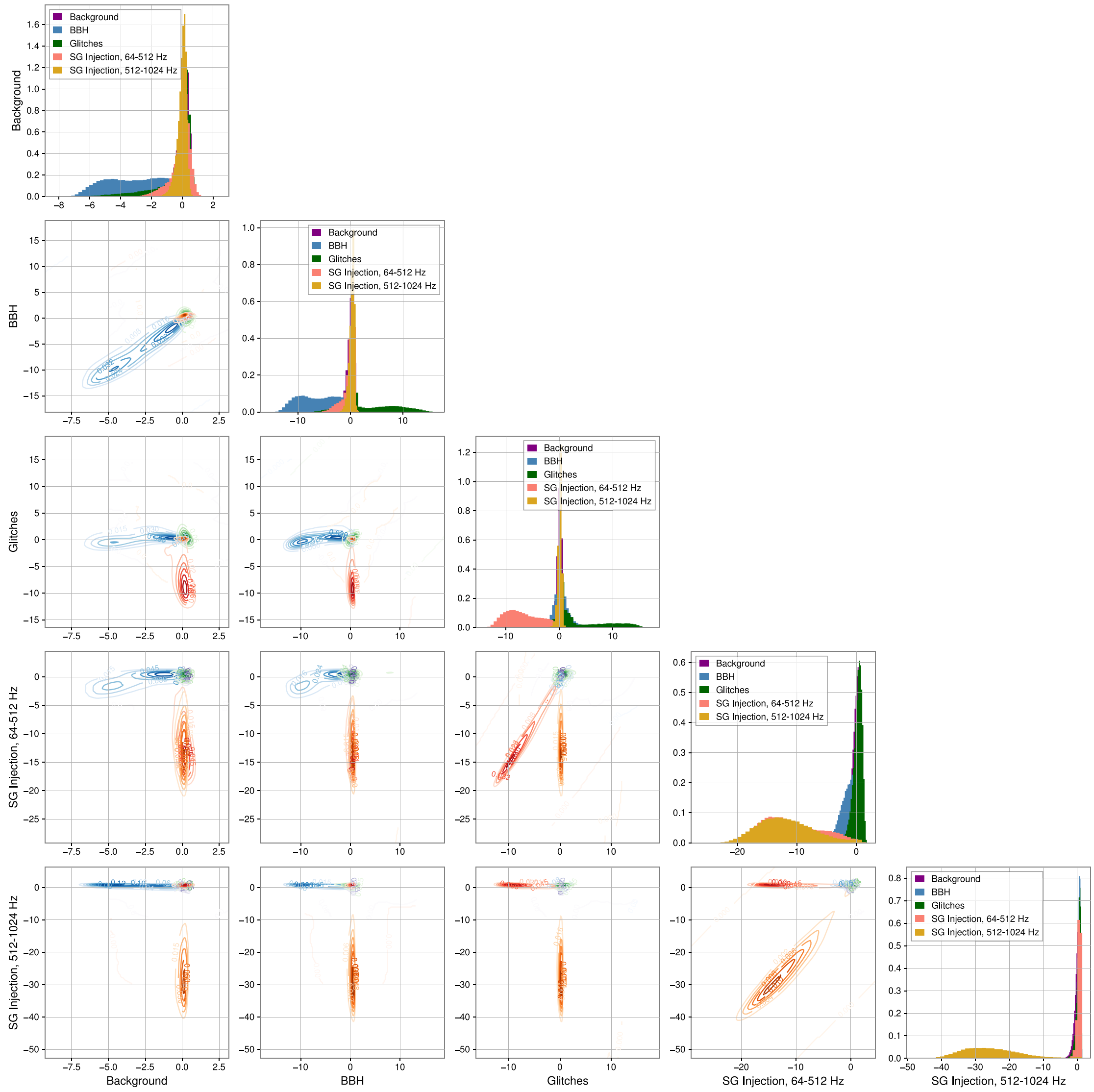

Figure 10. Trained GWAK space on five classes: BBH, background, glitch, and low and high frequency sine-gaussian; demonstrating that regions of the GWAK space correspond to different event classes.

is selected uniformly and randomly from a range spanning 40–400 Hz and 400–1000 Hz. These are denoted as WNBLF (white noise burst low-frequency) and WNBHF (white noise burst high-frequency) respectively. The duration of the signal is chosen to be 0.1 s. Theoretically, these would be the most difficult signals to detect with our algorithm, as their lack of distinctive morphology would render the signal autoencoder features useless. However, as shown in figure 11, the SG autoencoders were able to generalize to the WNBs.

To evaluate the performance of our algorithm, we generated these signals with SNRs uniformly distributed between 10 and 100. The average final metric value and corresponding standard deviation for various SNR ranges are shown in figure 12(right) and the lines corresponding to different false alarm rates.

The demonstration of the GWAK algorithm from strain to final metric is shown in figure 11. Starting with the whitened strain, we split up our data into 200 datapoint windows with a step of 5 datapoints between windows. For each window, the 10 autoencoder features are computed, along with the frequency domain correlation, and Pearson correlation, as described in section 3.5. Using the learned SVM weights as shown in figure 8, we reduce the 12-dimensional GWAK space to 7 dimensions and show those values at each timestep in the middle panel. For the correlation features, this is simply done by multiplying the value by the corresponding weight. For the autoencoder features, the same is done, but the $|\widetilde{H_O} \cdot \widetilde{H_R}|$ and $|\widetilde{L_O} \cdot \widetilde{L_R}|$ values are combined into a single value via addition. Finally, all of those weighted features are summed yielding the final metric value, which is shown in the bottom panel, along with various false alarm rate thresholds.

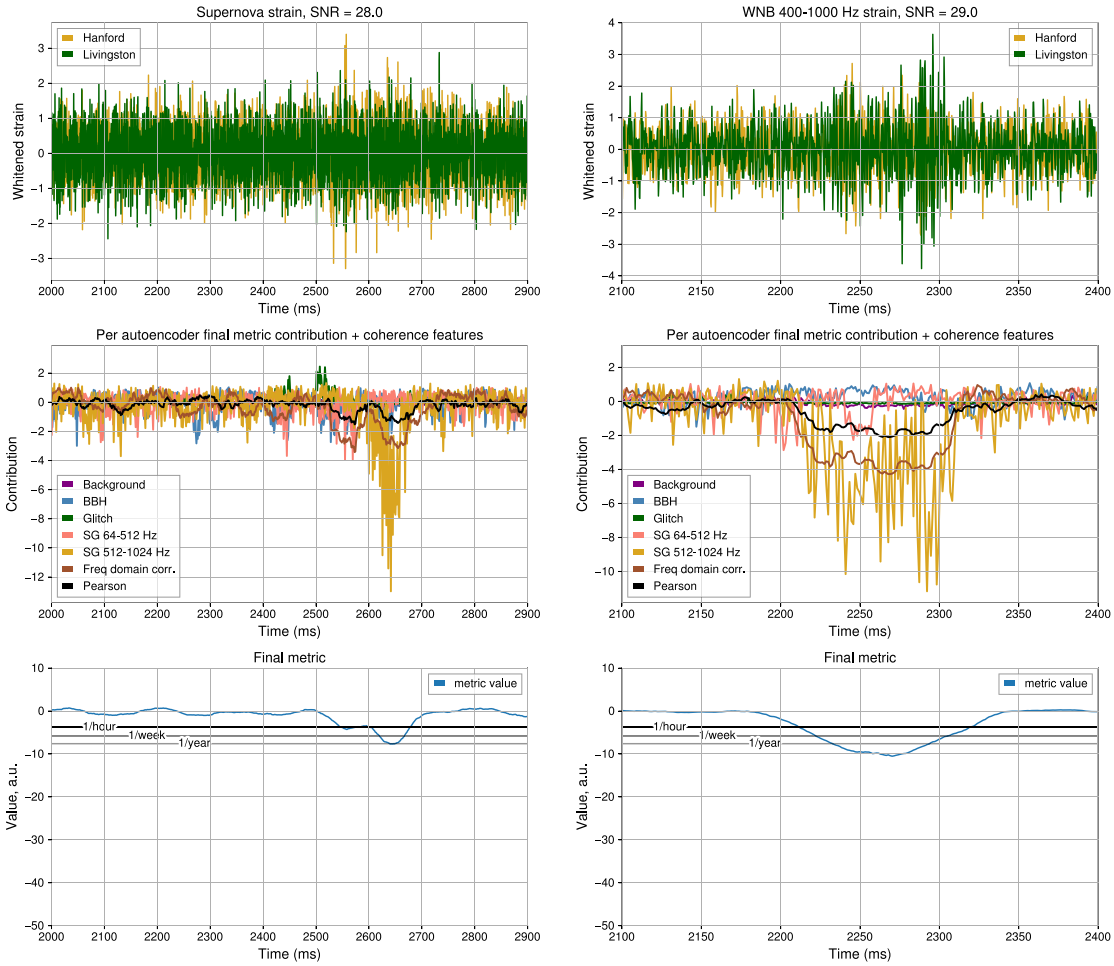

Figure 11. Strain (top), GWAK metric response (middle) and final metric response (bottom) for WNB (right) and Supernova (left). The evaluation of GWAK axes and Pearson correlation with time and on the top right total metric value and FAR are shown as an example of the algorithm's 'reaction' to unseen signals. Both BBH and sine-Gaussian losses drop at the time of the signal, which indicates that the strain at that moment is more signal-like. The Pearson correlation increases indicating a strong correlation between the two detector sides. At the moment of the event, the FAR drops to a level consistent with or below one event per year, which means that even with strong trigger restrictions, detection of that type of event would be possible with our new proposed algorithm.

The method performance on different signals at various SNR values is shown in figure 12. For each signal type, we generate 10 000 waveforms using the SNR prior $U[5, 50]$. We then run the full GWAK algorithm and obtain the minimum metric value achieved in each injection, as shown in figure 11. We then group the injections into SNR bins of width 5, and show the average metric value as well as the 1σ range for each bin, via the solid line and filled region respectively. The final metric values corresponding to certain false alarm thresholds are also shown. We first see that the three training classes—BBH, SG 64-512 Hz, SG 512–1024 Hz are most efficiently detected, which is expected as they have specific autoencoder models. Moving on to the anomalous signals—WNB 40–400 Hz, WNB 400–1000 Hz, and Supernova, we see that they are not detected as readily as the training signals, but are still able to achieve satisfactory false alarm rates around 1–2/month around 20–25 SNR.

We show the detection efficiency using a receiver operating characteristic (ROC) curve in figure 13 to compare it with other ML techniques. Here, we pick a fixed false alarm rate threshold of 1/year and compute what fraction of injections, for each signal, at each SNR, have detection statistics below the threshold. Similar to figure 12, we see that the signals corresponding to training classes can be more readily detected at lower SNRs, and the anomalous signals require slightly higher SNRs to be detected. This graph can be compared to the one presented by MLy [26], while for the BBH and SGs, we achieve similar performance, the efficiencies for WNBs and Supernovae are lower in our case. This is expected since the MLy algorithm was trained in a supervised manner, using WNBs during the training, while we only used those signals for finding the linear coefficients of the final metric but not as the GWAK axes.

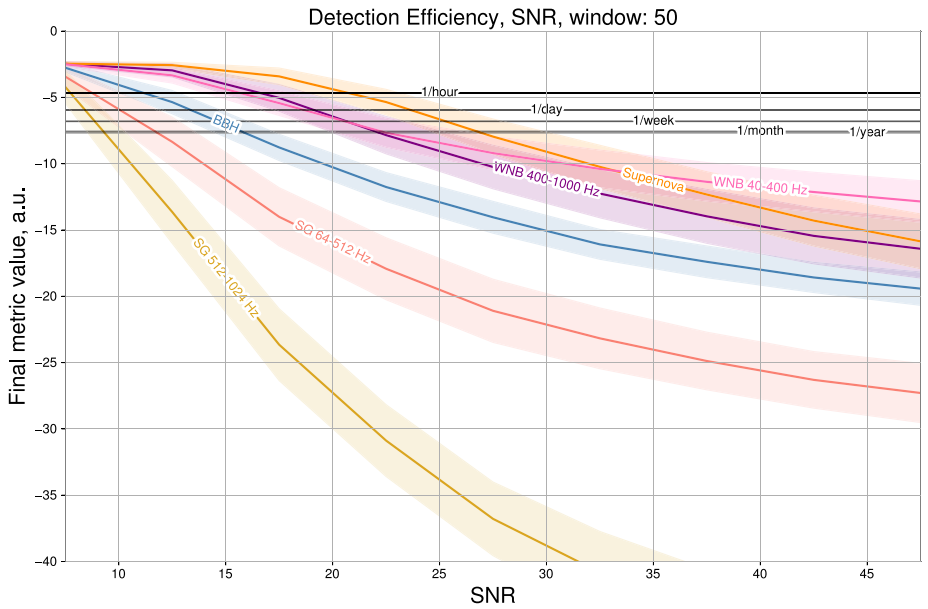

Figure 12. The final metric as a function of SNR for GWAK axes training signals, BBH (blue), SG 64–512 Hz (yellow), SG 512–1024 Hz (salmon) and for potential anomalies, WNB 40–400 Hz (pink), WNB 400–1000 Hz (purple), and Supernova (orange). The black lines of varied width correspond to different FARs, from the FAR of 1 per hour to 1 per year. For each of the lines, the events that are below that line would be detected. As expected, the signals used to train the GWAK axes are on average detected better, e.g. more events with a smaller SNR are detected for a given FAR threshold.

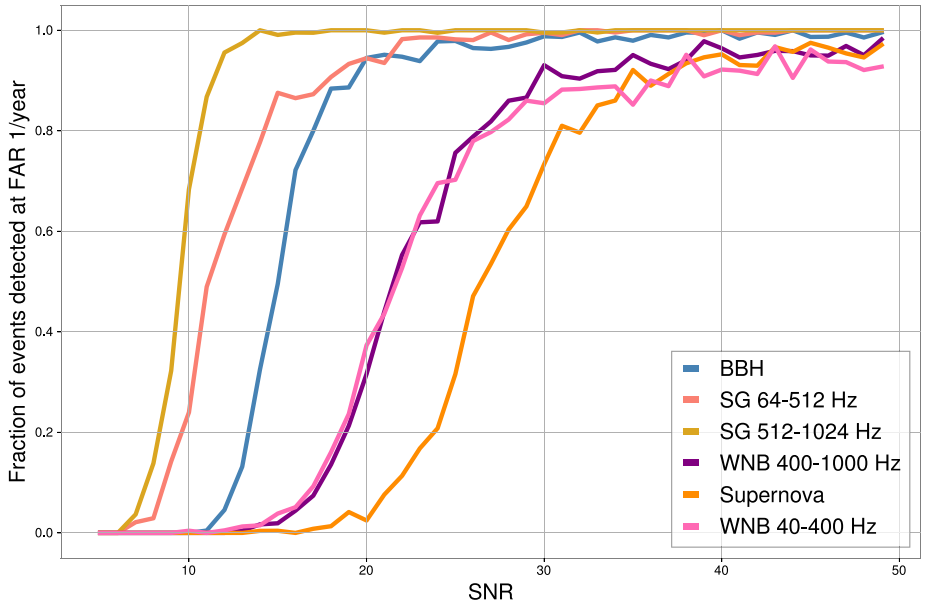

Figure 13. Detection efficiency as a function of SNR for GWAK axes training signals, and for potential anomalies. We see that the signals corresponding to training classes can be more readily detected at lower SNRs, and the anomalous signals require slightly higher SNRs to be detected.

4.5. Comparison to a supervised search

To quantify the loss of efficiency of the unsupervised GWAK method in comparison to a supervised search, we perform the following study. Firstly, we use pre-trained GWAK axes for the BBH search by using a small, dense network replacing the SVM for the final metric in the GWAK space. We train this network in a supervised manner, using BBH as the signal class and timeseries as the background class. In this way, we can quantify how much signal efficiency is lost when using a general linear combination instead of overfitting on a specific signal. The results are shown in figure 14. We observe that BBH detection efficiency surpasses that achieved with the linear final metric. However, as anticipated, the detection efficiency for all other signals significantly decreases. We omitted the use of a smoothing window, as it was determined to be the most

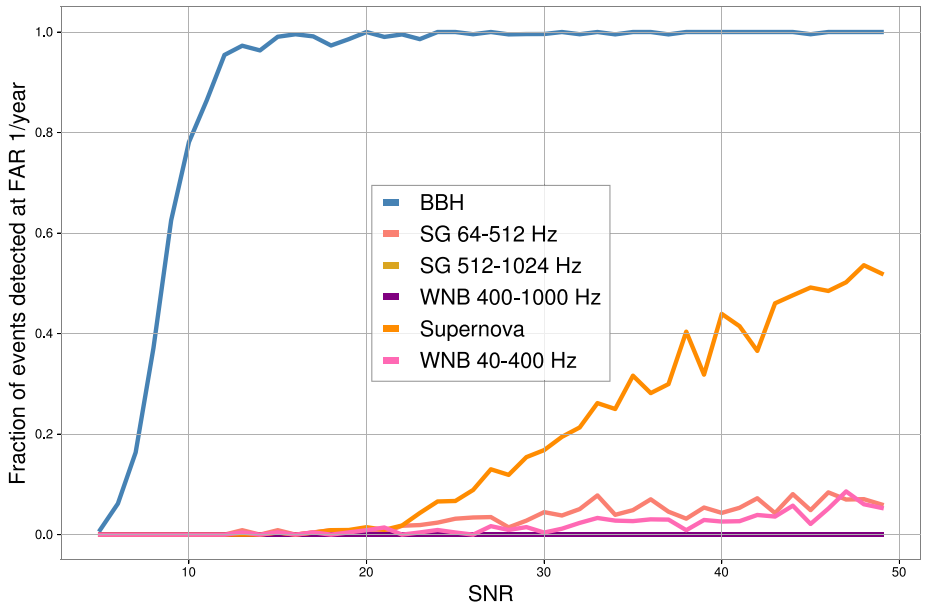

Figure 14. Detection efficiency for BBH and other signals and anomalies obtained using the final metric trained in a supervised manner on BBH signals pre-processed by the GWAK algorithm. This demonstrates that the GWAK method can approach a supervised search when given this specific task.

efficient for BBH-supervised search. Thus, we must compare it to the BBH ROC without the application of a smoothing window, as depicted in figure 15.

We demonstrate that, in general, enhancing the detection efficiency for a specific signal through supervised training is feasible. However, this improvement often incurs a noticeable decline in performance for other signals. Given our objective to develop an algorithm capable of detecting unknown signals, we lack the necessary information to train it in a supervised manner. Nevertheless, in follow-up works, we may consider exploring the adoption of a more sophisticated final metric function, though we must exercise caution to prevent overfitting the signals employed in optimizing this metric.

4.6. Comparison of different smoothing windows

In figure 15, we illustrate that the optimization of detection efficiency, both for known and anomalous signals, occurs when employing smoothing kernels with sizes approximately equal to the signal length. These findings affirm that utilizing smoothing within the final metric space is an effective strategy to adapt the GWAK algorithm to diverse signal lengths

5. Conclusions

In this study, we utilized the GWAK method to identify anomalies in datasets acquired by ground-based GW observatories. The GWAK method relies on the notion of introducing alternative signal priors that capture some of the salient features of new physics signatures, enabling the restoration of sensitivity even when the alternative signal is incorrect. We separately trained five unsupervised autoencoders on a dataset consisting of normal background noise, glitches, and a collection of simulated signals that incorporate the physical characteristics of a potential new physics signature. We then established a 12-dimensional GWAK space, composed of two reconstruction losses for each of the detector sites for each of the five autoencoders and two features representing the correlation between the detector sites. This GWAK space was used as a search region for anomalous signals. Finally, we combined all the 12 features into one final metric by multiplying with corresponding coefficients from figure 8 and summing the result.

Our findings indicate that the GWAK method efficiently detected anomalies in the GW datasets, particularly unmodeled sources like CCSN and white noise bursts. Additionally, the GWAK method could differentiate signal-like anomalies from anomalous events, such as those resulting from detector glitches.

Our proposed method demonstrates promising results, detecting these sources with high accuracy and without prior knowledge of their characteristics. These findings underscore the potential value of our approach in detecting new and unexpected sources of GW signals, while simultaneously reducing the dependence on labeled training data. In addition to serving as an unsupervised search, the machine-learning

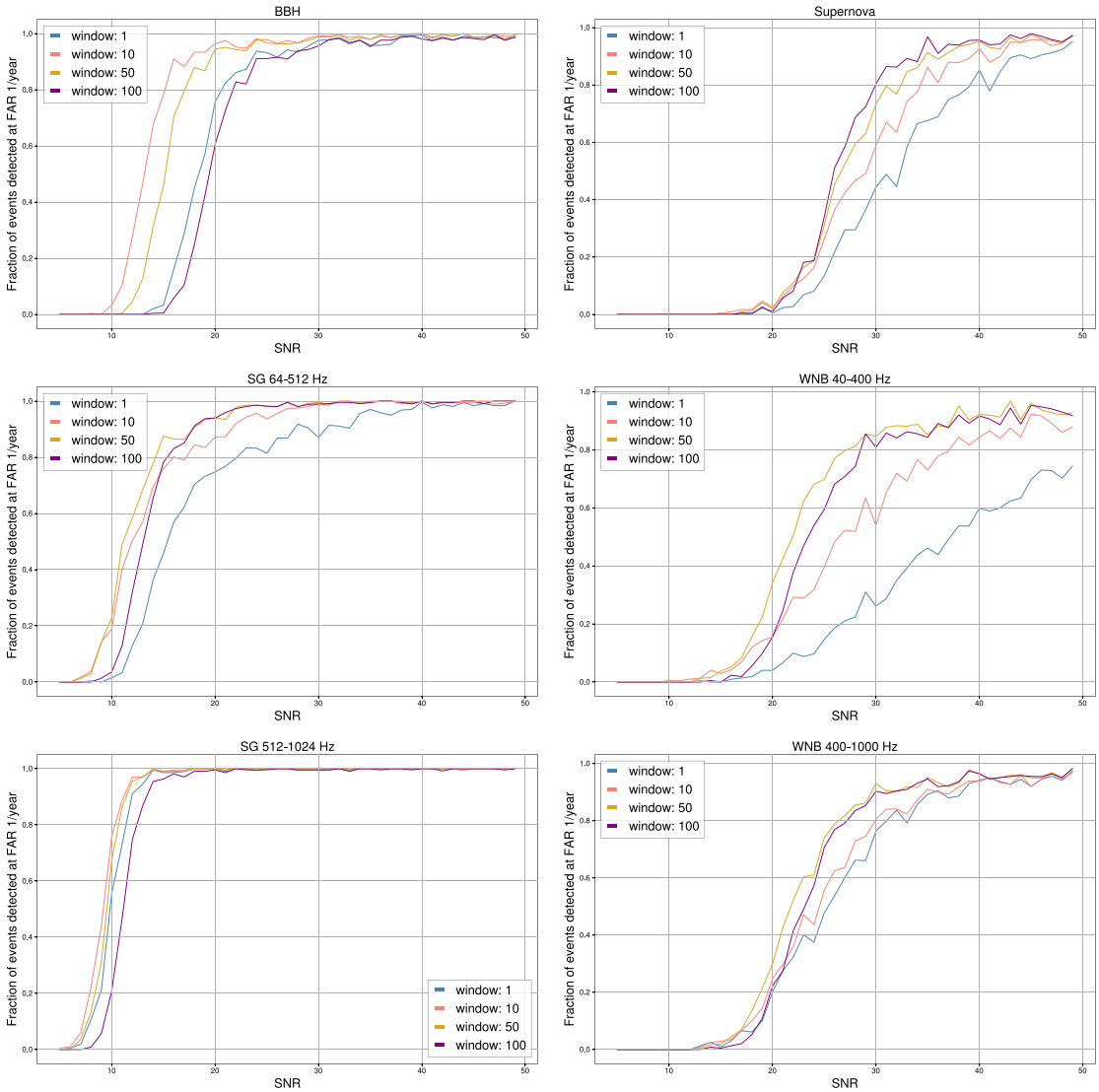

Figure 15. Different smoothing windows for training signals (left column) and anomalous signals (right column). The results of this plot demonstrate that the search for anomalies can be specified to signals of specific durations. For a shorter signal like a BBH, around 50 ms or 10 timesteps in the output space, we expect that smoothing the final metric values to this order gives the best sensitivity. This is because a 10 timestep window will, at some point, capture the entire signal, but no more. Indeed, the BBH detection sensitivity is optimal for a 10-datapoint window. With the supernovae, which are of order 500 ms, we see that the 100-datapoint window has the best detection statistic. Considering the relative durations of the other training and anomalous signals, we see that specifying the search to their duration via a smoothing window yields optimal detection capability.

based approach allows for the implementation of the GWAK algorithm as a low-latency search tool. By detecting anomalies rapidly, alerts can be sent to electromagnetic telescopes for follow-up.

In future work, there are many potential avenues to explore. One is using the recreation as a denoising tool instead of just a detection statistic, allowing for rapid parameter estimation. This is especially important with electromagnetic follow-up, as the telescopes need information on the source location for detailed observation. On the detection side of things, an idea is to use normalizing flows to learn the high dimensional manifolds on which signals lie and use the probability value as a score alternate to the less intuitive engineered autoencoder frequency domain features. There is also potential to improve the search by changing the way that the 12 GWAK features are reduced to a single value. While the linear fit served as a simple and explainable method and did not possess the potential to overfit the known signals and therefore decrease sensitivity to anomalies, it suffers the drawback that potentially useful patterns between features are not exploited. A simple example is with the features $|\widetilde{H_O} \cdot \widetilde{H_R}|$ and $|\widetilde{L_O} \cdot \widetilde{L_R}|$ for a single autoencoder. For a glitch event, you would expect only one of these values to increase, corresponding to the detector in which the glitch occurred, so you would expect an asymmetry between these features for a non-astrophysical event. The opposite is true for a BBH event, for example, as the fact that it is present in both detector channels means that there should be symmetry in the BBH autoencoder features. While this is just one intuitive

example, more complex relations could certainly exist. Yet another area to explore is modifying the network architecture to allow for a longer signal length to generalize the algorithm to various signal durations.

To conclude, the GWAK method displays potential as a powerful tool for detecting anomalies in GW datasets and has the potential to enhance the performance of GW anomaly detection systems.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

The authors are immensely grateful to Tino Tibaldo for his exceptional creativity and expertise in developing the 3D figures and representations, which greatly enhanced the visual appeal and clarity of our work. The authors gratefully acknowledge William Patrick McCormack and Jeffery Krupa for their invaluable contributions during the QUAK discussions, which significantly enhanced the depth and quality of our research. The authors extend their sincere gratitude to Rachel Smith for her insightful and fruitful discussions. The authors acknowledge support from the National Science Foundation with Grant Numbers OAC-2117997 and CSSI-1931469. This research was undertaken with the support of the LIGO computational clusters. M W C and S M also acknowledge support from the National Science Foundation with Grant Number PHY-2010970. E M acknowledges support from the National Science Foundation with Grant Number GRFP-2141064. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation.

Appendix

A.1. Backgrounds training curve

For completeness, in figure 16 we show the training and validation losses for the background and glitch autoencoders.

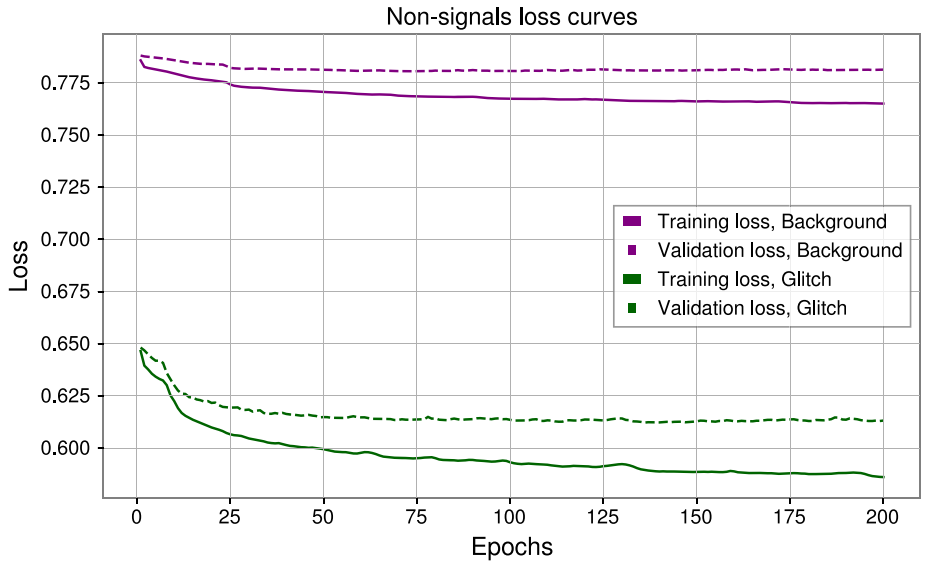

Figure 16. Autoencoder training and validation losses for background classes. The training/validation losses are in bold/dashed for background (in purple) and for glitches (in green).

A.2. Recreations

For completeness, in figures 17–20 we show recreation plots for the low frequency SG, high frequency SG, glitches and background correspondingly. For each of the dataset types, we can see that the corresponding autoencoder is the best in reconstructing the input, as expected.

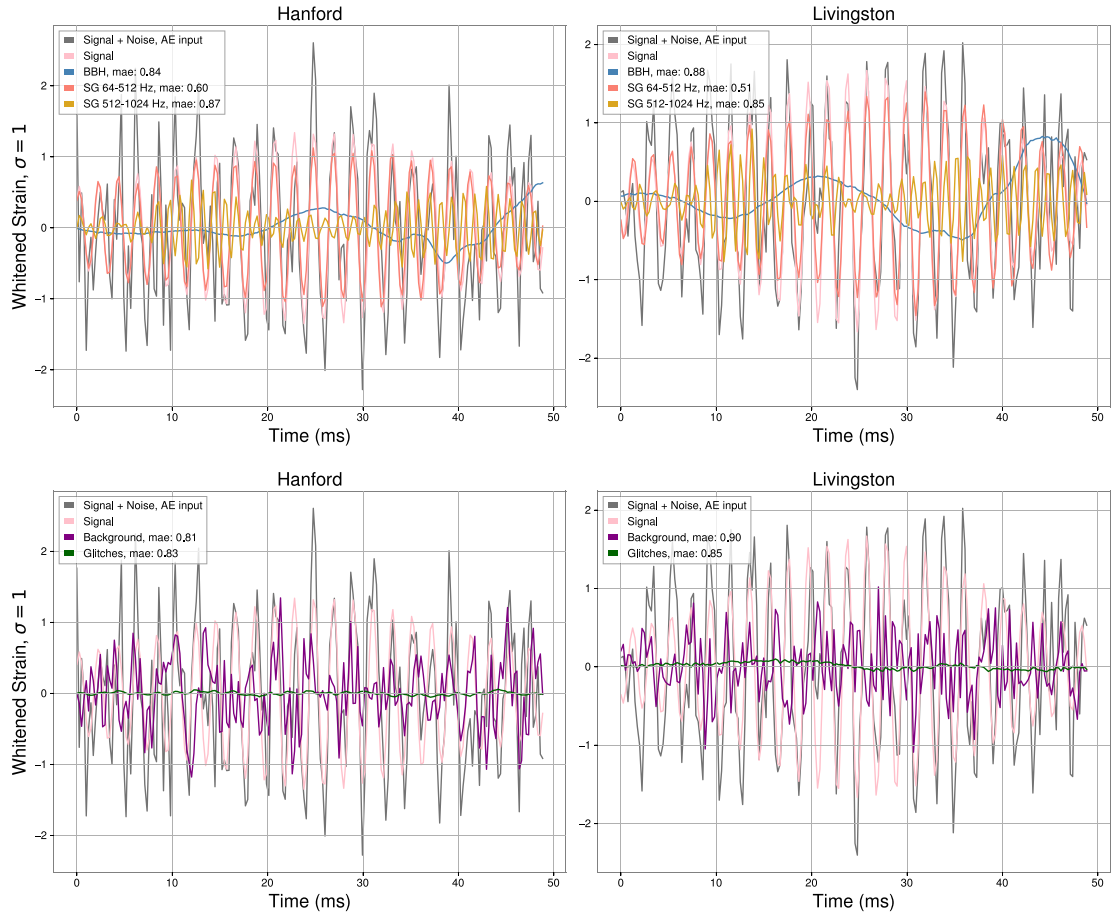

Figure 17. Example of recreation plots obtained with all five pre-trained autoencoders on low frequency sine-gaussian dataset.

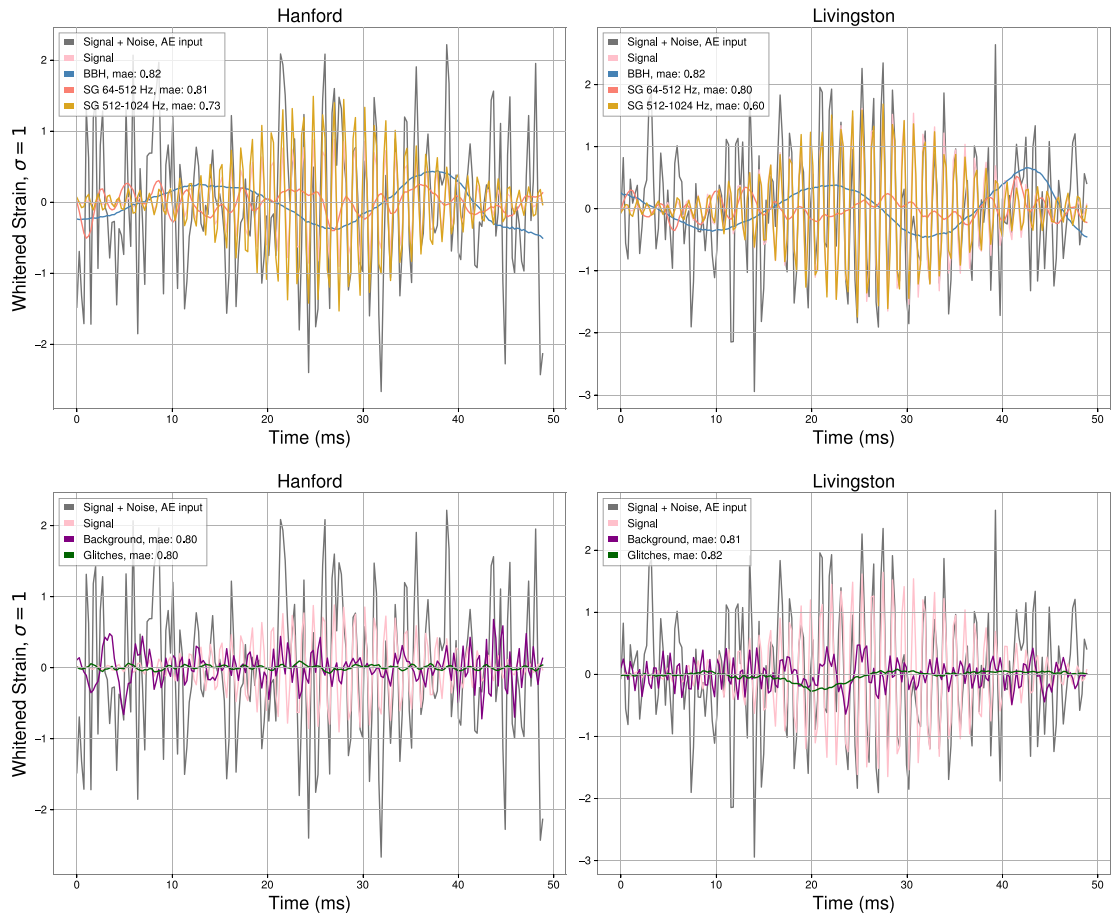

Figure 18. Example of recreation plots obtained with all five pre-trained autoencoders on high frequency sine-gaussian dataset.

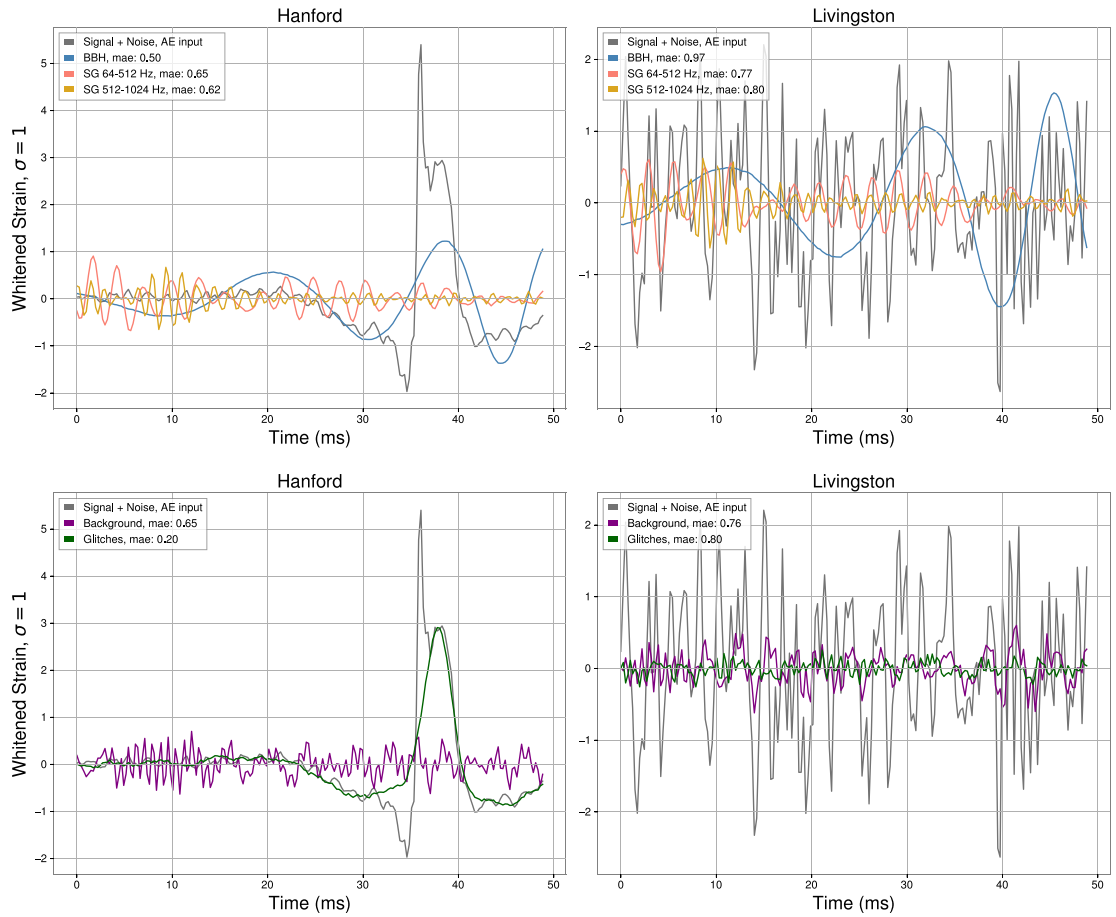

Figure 19. Example of recreation plots obtained with all five pre-trained autoencoders on glitch dataset.

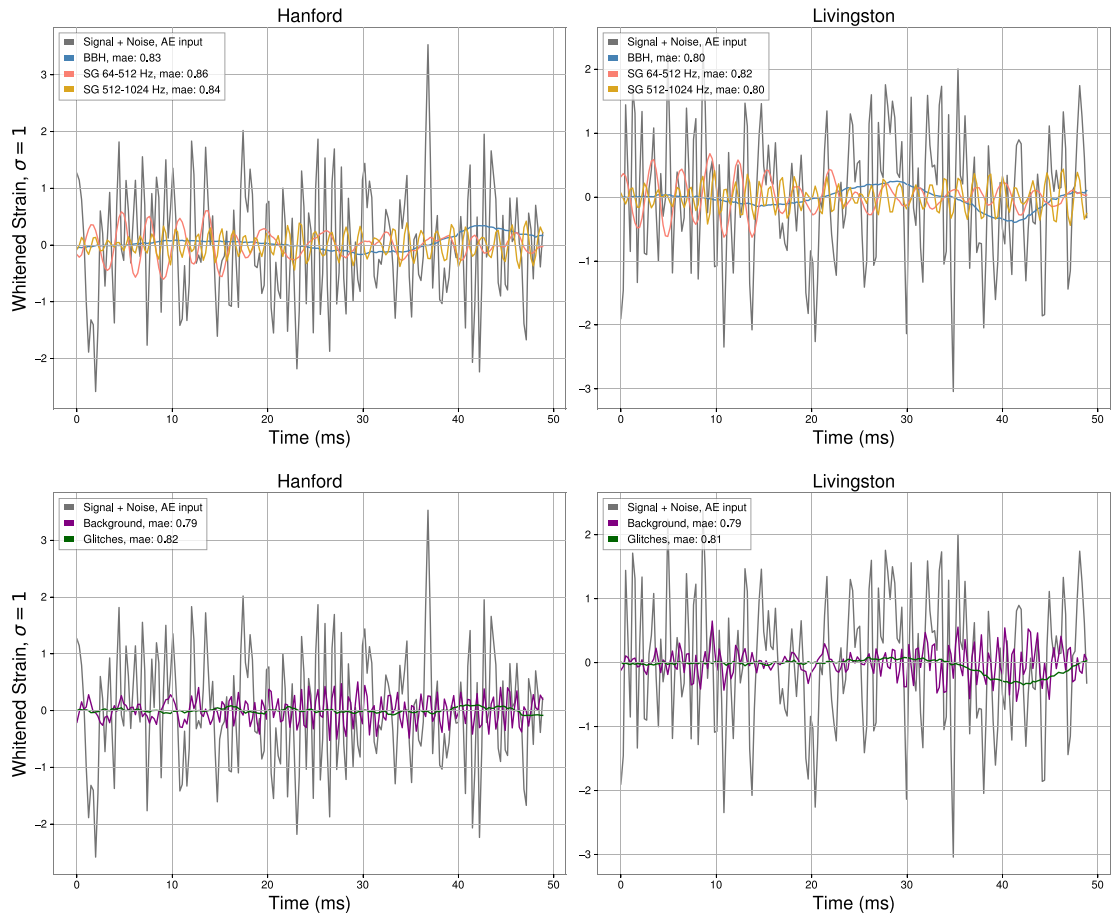

Figure 20. Example of recreation plots obtained with all five pre-trained autoencoders on background dataset.

A.3. Method validation

In this section, we perform a check on the significance of signal autoencoders for signal detection efficiency. To do this, we evaluate efficiencies using a final metric linear combination of two correlation features, the Pearson coefficient and frequency domain correlation. The results obtained with such metric are shown in figure 21, the signal efficiencies for all the potential anomalous signals are much worse than when using the signal autoencoders in the final metric, as shown in figure 12. Therefore, we show that it is beneficial to add several potential signals to the final metric.

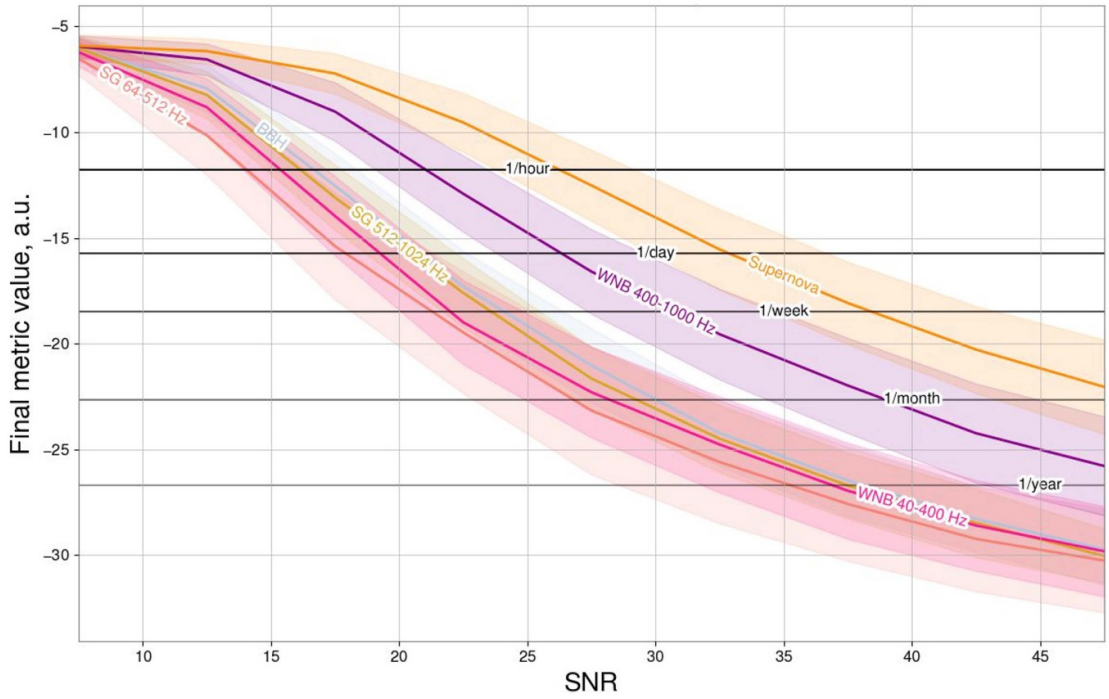

Figure 21. Detection efficiencies using only a linear combination of correlation features (Pearson and frequency domain correlation) as a final metric.

ORCID iDs

Ryan Raikman https://orcid.org/0000-0003-4083-6390
Eric A Moreno https://orcid.org/0000-0001-5666-3637
Ekaterina Govorkova https://orcid.org/0000-0003-1920-6618
Deep Chatterjee https://orcid.org/0000-0003-0038-5468
Philip C Harris https://orcid.org/0000-0001-8189-3741

References

[1] Abbott B P *et al* 2016 Observation of gravitational waves from a binary black hole merger *Phys. Rev. Lett.* **116** 061102
[2] Aasi J *et al* (The LIGO Scientific Collaboration) 2015 Advanced LIGO *Class. Quantum Grav.* **32** 074001
[3] Acernese F *et al* 2014 Advanced Virgo: a second-generation interferometric gravitational wave detector *Class. Quantum Grav.* **32** 024001
[4] Akutsu T *et al* 2021 Overview of KAGRA: detector design and construction history *Prog. Theor. Exp. Phys.* **2021** 05A101
[5] Abbott B P *et al* 2019 GWTC-1: a gravitational-wave transient catalog of compact binary mergers observed by LIGO and Virgo during the first and second observing runs *Phys. Rev.* X **9** 031040
[6] Abbott R *et al* 2021 GWTC-2: compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run *Phys. Rev.* X **11** 021053
[7] Abbott R *et al* 2021 GWTC-3: compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run (arXiv:2111.03606)
[8] Abbott B P *et al* 2016 GW151226: observation of gravitational waves from a 22-solar-mass binary black hole coalescence *Phys. Rev. Lett.* **116** 241103
[9] Abbott B P *et al* 2017 GW170104: observation of a 50-solar-mass binary black hole coalescence at redshift 0.2 *Phys. Rev. Lett.* **118** 221101
[10] Abbott B P *et al* 2017 GW170814: a three-detector observation of gravitational waves from a binary black hole coalescence *Phys. Rev. Lett.* **119** 141101

[11] Abbott B P et al 2017 GW170817: observation of gravitational waves from a binary neutron star inspiral Phys. Rev. Lett. 119 161101
[12] Abdikamalov E, Pagliaroli G and Radice D 2021 Gravitational waves from core-collapse supernovae Handbook of Gravitational Wave Astronomy (Springer) pp 1–37
[13] Hindmarsh M B and Kibble T W B 1995 Cosmic strings Rep. Prog. Phys. 58 477–562
[14] Damour T and Vilenkin A 2000 Gravitational wave bursts from cosmic strings Phys. Rev. Lett. 85 3761–4
[15] Braaten E and Zhang H 2018 Axion stars (arXiv:1810.11473)
[16] Franciolini G 2021 Primordial black holes: from theory to gravitational wave observations (arXiv:2110.06815)
[17] Eroshenko Y N 2018 Gravitational waves from primordial black holes collisions in binary systems J. Phys.: Conf. Ser. 1051 012010
[18] Ott C D 2009 The gravitational wave signature of core-collapse supernovae Class. Quantum Grav. 26 063001
[19] Allen B, Anderson W G, Brady P R, Brown D A and Creighton J D E 2012 FINDCHIRP: an algorithm for detection of gravitational waves from inspiraling compact binaries Phys. Rev. D 85 122006
[20] Gossan S E, Sutton P, Stuver A, Zanolin M, Gill K and Ott C D 2016 Observing gravitational waves from core-collapse supernovae in the advanced detector era Phys. Rev. D 93 042002
[21] Vilenkin A, Levin Y and Gruzinov A 2018 Cosmic strings and primordial black holes J. Cosmol. Astropart. Phys. JCAP11(2018)008
[22] Ghoshal A, Gouttenoire Y, Heurtier L and Simakachorn P 2023 Primordial black hole archaeology with gravitational waves from cosmic strings J. High Energy Phys. JHEP08(2023)196
[23] Klimenko S et al 2016 Method for detection and reconstruction of gravitational wave transients with networks of advanced detectors Phys. Rev. D 93 042004
[24] Klimenko S, Mohanty S, Rakhmanov M and Mitselmakher G 2005 Constraint likelihood analysis for a network of gravitational wave detectors Phys. Rev. D 72 122002
[25] Lynch R, Vitale S, Essick R, Katsavounidis E and Robinet F 2017 Information-theoretic approach to the gravitational-wave burst detection problem Phys. Rev. D 95 104046
[26] Skliris V, Norman M R K and Sutton P J 2022 Real-time detection of unmodelled gravitational-wave transients using convolutional neural networks (arXiv:2009.14611)
[27] Park S E, Rankin D, Udrescu S-M, Yunus M and Harris P 2021 Quasi anomalous knowledge: searching for new physics with embedded knowledge J. High Energy Phys. JHE06(2021)030
[28] Baker P T, Caudill S, Hodge K A, Talukder D, Capano C and Cornish N J 2015 Multivariate classification with random forests for gravitational wave searches of black hole binary coalescence Phys. Rev. D 91 062004
[29] George D and Huerta E A 2018 Deep neural networks to enable real-time multimessenger astrophysics Phys. Rev. D 97 044039
[30] Kapadia S J, Dent T and Dal Canton T 2017 Classifier for gravitational-wave inspiral signals in nonideal single-detector data Phys. Rev. D 96 104015
[31] George D and Huerta E A 2018 Deep learning for real-time gravitational wave detection and parameter estimation: results with advanced LIGO data Phys. Lett. B 778 64–70
[32] Gabbard H, Williams M, Hayes F and Messenger C 2018 Matching matched filtering with deep networks for gravitational-wave astronomy Phys. Rev. Lett. 120 141103
[33] Miller A L et al 2019 How effective is machine learning to detect long transient gravitational waves from neutron stars in a real search? Phys. Rev. D 100 062005
[34] Jadhav S, Mukund N, Gadre B, Mitra S and Abraham S 2021 Improving significance of binary black hole mergers in advanced LIGO data using deep learning: confirmation of GW151216 Phys. Rev. D 104 064051
[35] Huerta E A et al 2021 Accelerated, scalable and reproducible AI-driven gravitational wave detection Nat. Astron. 5 1062–8
[36] Jiang L and Luo Y 2022 Convolutional transformer for fast and accurate gravitational wave detection 2022 26th Int. Conf. on Pattern Recognition (ICPR) pp 46–53
[37] Chatterjee C, Wen L, Diakogiannis F and Vinsen K 2021 Extraction of binary black hole gravitational wave signals from detector data using deep learning Phys. Rev. D 104 064046
[38] Beveridge D, Wen L and Wicenec A 2023 Detection of binary black hole mergers from the signal-to-noise ratio time series using deep learning (arXiv:2308.08429)
[39] Ormiston R, Nguyen T, Coughlin M, Adhikari R X and Katsavounidis E 2020 Noise reduction in gravitational-wave data via deep learning Phys. Rev. Res. 2 033066
[40] Bacon P, Trovato A and Bejger M 2022 Denoising gravitational-wave signals from binary black holes with dilated convolutional autoencoder (arXiv:2205.13513)
[41] Saleem M et al 2023 Demonstration of Machine Learning-assisted real-time noise regression in gravitational wave detectors (arXiv:2306.11366)
[42] Gunny A, Rankin D, Krupa J, Saleem M, Nguyen T, Coughlin M, Harris P, Katsavounidis E, Timm S and Holzman B 2022 Hardware-accelerated inference for real-time gravitational-wave astronomy Nat. Astron. 6 529–36
[43] Liao C-H and Lin F-L 2021 Deep generative models of gravitational waveforms via conditional autoencoder Phys. Rev. D 103 124051
[44] Sankarapandian S and Kulis B 2021 β-annealed variational autoencoder for glitches (arXiv:2107.10667)
[45] Morawski F, Bejger M, Cuoco E and Petre L 2021 Anomaly detection in gravitational waves data using convolutional autoencoders Mach. Learn.: Sci. Technol. 2 045014
[46] Moreno E 2021 Source-agnostic gravitational-wave detection with recurrent autoencoders: BBH dataset Zenodo (https://doi.org/10.5281/zenodo.5772814)
[47] Skliris V, Norman M R K and Sutton P J 2020 Real-time detection of unmodelled gravitational-wave transients using convolutional neural networks (arXiv:2009.14611)
[48] Sutton P J et al 2010 X-pipeline: an analysis package for autonomous gravitational-wave burst searches New J. Phys. 12 053034
[49] Abbott B P et al 2017 All-sky search for short gravitational-wave bursts in the first Advanced LIGO run Phys. Rev. D 95 042003
[50] Robinet F, Arnaud N, Leroy N, Lundgren A, Macleod D and McIver J 2020 Omicron: a tool to characterize transient noise in gravitational-wave detectors SoftwareX 12 100620
[51] Santamaría L et al 2010 Matching post-Newtonian and numerical relativity waveforms: systematic errors and a new phenomenological model for nonprecessing black hole binaries Phys. Rev. D 82 064016
[52] Husa S, Khan S, Hannam M, Pürrer M, Ohme F, Forteza X J and Bohé A 2016 Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal Phys. Rev. D 93 044006
[53] Khan S, Husa S, Hannam M, Ohme F, Pürrer M, Forteza X J and Bohé A 2016 Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era Phys. Rev. D 93 044007

[54] Nitz A et al 2020 gwastro/pycbc: Pycbc release v1.16.9 Zenodo (https://doi.org/10.5281/zenodo.3993665)
[55] Kingma D P and Ba J 2017 Adam: a method for stochastic optimization (arXiv:1412.6980)
[56] Karl P 1895 VII. Note on regression and inheritance in the case of two parents Proc. R. Soc. 58 240–2
[57] Powell J, Müller B and Heger A 2021 The final core collapse of pulsational pair instability supernovae Mon. Not. R. Astron. Soc. 503 2108–22
[58] Abbott B P et al 2016 Characterization of transient noise in Advanced LIGO relevant to gravitational wave signal GW150914 Class. Quantum Grav. 33 134001
[59] Cabero M et al 2019 Blip glitches in Advanced LIGO data Class. Quantum Grav. 36 15