Differentiating high frequency gravitational wave signals from thermal noise using Machine Learning

Christian Gottschlich

Bachelor's Thesis in Physics written at the

Physikalisches Institut

submitted to the der Faculty of Mathematics and Natural Sciences

at the

Rheinische Friedrich-Wilhelms-University Bonn

September 2025

Ich versichere, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die Zitate kenntlich gemacht habe.

Bonn, ...

Datum Unterschrift

1. Gutachter: Prof. Dr. Matthias Schott

Contents

Abstract. Traditionally, matched filtering in the frequency domain has been the standard approach for detecting gravitational wave signals in noisy data. In this thesis, an alternative solution is explored, applying Machine Learning to detect simulated high-frequency GW signals from mergers of two primordial black holes with masses between 10^{-13} and $10^{-6} M_{\odot}$, using time series data. The detection capabilities of such signals will be assessed for a variational auto-encoder using the mean-squared reconstruction error as the primary anomaly detection metric. It has been found that a range of Signal-to-Noise ratios from 1 up to 6.5 is required to detect signals with an efficiency of 95% and an upper bound of 3.26 $\times 10^5$ false positives per day. The detection performance is fully determined by how strongly injected signals alter the mean and variance of the noise, as explained by the bias–variance decomposition of the mean-squared error. The VAE effectively learns the point-wise mean of the noise distribution, which remains nearly constant.

1 Introduction

It should be noted that for writing this report, drafting support from ChatGPT as well as Gemini has been used.

On September 14 2015, the Laser interferometers of LIGO and Virgo [1] detected gravitational waves (GWs) that had been predicted by Albert Einstein in 1916 [2], for the first time. This marked the beginning of gravitational-wave astronomy and provided direct confirmation of a key prediction of General Relativity.

Traditionally, the state-of-the-art analysis method for detecting GW signals, employed by LIGO and Virgo [1], is matched filtering. Here, noisy detector data is correlated with a large bank of synthetic GW templates, which use waveform predictions based in General Relativity.

The goal of this thesis is to explore an alternative analysis method based on anomaly detection using Machine Learning (ML).

First, section 2 explains the origin of GWs and how high-frequency GWs (HFGWs) can be detected.

Afterwards, the Machine Learning analysis approach explored in this thesis is examined and explained in detail 3. In chapter 4 the approach used to simulate GW signals for testing purposes is stated. Section 5 explains the findings, and, lastly a summary 6, an outlook 7 and acknowledgments 8 conclude the thesis.

2 Gravitational Waves

2.1 Theoretical background

Note that Greek letters in the following will refer to the 4 space-time coordinates, with 0 indexing time and 1 to 3 space.

In the beginning of the 20th century, Albert Einstein found that the linearized field equations of General Relativity (GR) had wave solutions [2].

To describe gravitational waves theoretically, one typically resorts to a direct application of perturbation theory to Einstein's field equations. The central assumption is that the space-time metric $g_{\mu\nu}$ can, in first approximation, be expressed as the sum of the flat Minkowski metric $\eta_{\mu\nu}$ and a small perturbation $h_{\mu\nu}$,

$$g_{\mu\nu}(x) = \eta_{\mu\nu} + h_{\mu\nu}(x) \quad \text{where} \quad x = x^{\mu} \tag{1}$$

where $|h_{\mu\nu}|, |\partial_\rho h_{\mu\nu}| \ll 1$ [1]. A small perturbation here means that terms of order $\mathcal{O}(h^2)$ and higher can be neglected.

This leads to both the Christoffel symbols $\Gamma^{\kappa}_{\mu\nu}$ and the Ricci tensor $R_{\mu\nu}$ being composed of only linear terms in $h_{\mu\nu}$ and its derivatives. Furthermore, the Einstein tensor $G_{\mu\nu}$ becomes linear in $h_{\mu\nu}$ and its first and second derivatives: [2]

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}R = 8\pi G T_{\mu\nu} \tag{2}$$

with R the Ricci scalar, $T_{\mu\nu}$ the energy-momentum tensor, and G the gravitational constant.

To solve the linearized field equations one first notices that an infinitesimal coordinate transformation

$$x^{\mu} \longrightarrow x^{\mu} + \epsilon^{\mu}, \qquad |\epsilon^{\mu}| \ll 1, \tag{3}$$

leaves the field equations (2) invariant. This gauge freedom implies that the metric perturbation $h_{\mu\nu}$ contains redundant, unphysical degrees of freedom (dof), which makes the solutions not unique. Since $h_{\mu\nu}$ is symmetric, it initially has ten independent components. Choosing a gauge to work in, such as the Lorenz gauge

$$\partial^{\mu}\bar{h}_{\mu\nu} = 0, \tag{4}$$

where $\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h$ is the trace-reversed perturbation, allows us to reduce the dof of $h_{\mu\nu}$, associated with the four functions ϵ^{μ}, to six. Furthermore, the Lorenz gauge allows us to write the Einstein field equations as [3]

$$\Box\bar{h}_{\mu\nu} = -16\pi G T_{\mu\nu} \tag{5}$$

In addition, the residual freedom in $h_{\mu\nu}$ can be used to impose further convenient conditions, such as the transverse–traceless (TT) gauge [3], which leaves only two independent dynamical dof $h_{\mu\nu}$. These correspond to the two polarization states of GWs. Further details can be found in [3].

[1] ∂_ρ represents the partial derivative.

[2] Note that the speed of light c is set to one.

Physically, while the metric perturbations $h_{\mu\nu}$ themselves are gauge dependent, the physical content of the polarizations is observable through their gauge-invariant tidal effects. These effects can and have been directly observed in experiments (like the LIGO and Virgo interferometers [1]) as a time-dependent fractional length change $\delta L(t)$. The conventional quantity used to describe the length-changing effect of a GW is the dimensionless strain $h(t)$:

$$h(t) = \frac{\delta L(t)}{L} \tag{6}$$

which relates $\delta L(t)$ to the original distance L between two objects without any perturbation in the metric. It is important to note that $h(t)$ is, in general, a superposition of the two GW polarization states, dependent on the detector geometry. For further details, it is referred to [4].

To sum up, GWs are described by a perturbation away from flat Minkowski space-time. Physically, they are ripples in the fabric of space-time.

2.2 Sources of GWs and HFGWs

Generally, GWs are generated by accelerated masses with a time-varying quadrupole moment [3]. Low-frequency GWs (LFGWs) in the Hz to kHz range can originate from slowly merging or orbiting systems of several solar mass stellar objects. One category of objects causing LFGWs includes compact binaries which consist of: [5]

- Binary Black Holes (BBHs) - two black holes orbiting each other
- Binary Neutron Stars (BNSs) - two Neutron Stars orbiting each other
- Neutron Star-Black Hole Binary (NSBH) - a neutron star and a black hole orbiting each other

Thus far, the LIGO observatory has only detected signals, which fall into the category of *Compact Binary's*, though, in addition, there are two more categories outlined by LIGO in [5].

In contrast, high-frequency gravitational waves (HFGWs), in the MHz to GHz regime, cannot be produced by these astrophysical systems since their orbital and dynamical timescales are far too long [6]. Instead, HFGWs are expected to arise from much shorter length scales and earlier periods in the Universe. Candidate mechanisms include primordial phenomena such as preheating after inflation, phase transitions, and cosmic strings [6].

Of primary consideration, though, for this work are primordial black holes (PBHs) [7], hypothetical black holes formed from density fluctuations in the early Universe, possibly during the inflationary era. Unlike stellar black holes, PBHs could span a wide mass range, including sub-solar masses that would produce gravitational waves at much higher frequencies than those detectable by ground-based interferometers [6]. Their merger signatures and expected strain spectra have been further discussed in [7], making them a central target for HFGW searches.

2.3 Detection of HFGWs

Similar to axion haloscopes (see for example [8]), a possible measurement technique is the use of microwave cavities placed in a strong, external magnetic field, which exploits the coupling of GWs to the electromagnetic (EM) field, known as the Gertsenshtein effect [9].

As a GW passes through a cavity, it perturbs the geometry of the cavity and the external magnetic field lines. This causes the effective magnetic flux through the cavity's volume to oscillate, making the magnetic field spatially varying [3]. This magnetic flux modulation consequently creates a time-varying electric field that oscillates at the same frequency as the original GW. If the frequency of oscillation corresponds to a resonance mode of the cavity, the electric field can couple to a cavity mode, which results in the resonant enhancement of the oscillation amplitude. Subsequently, a current is induced in the cavity's antenna, which in turn can be measured. Further details can be found in [10] .

[3] Notice that it is non-trivial that the magnetic flux is altered through the cavity, since the GW alters both the magnets shape and the cavity's shape. For a rigorous derivation see Appendix A of [9]

The induced signal power P_{sig} thereby depends on several key cavity parameters and physical quantities [11]. The quality factor Q characterizes the sharpness of the cavity resonance, while ω_g denotes the angular frequency of the incoming gravitational wave. The effective cavity volume V_{cav} determines the mode overlap with the signal, and η_n represents the geometric form factor of the relevant mode. The gravitational wave strain amplitude is given by h_0, and B_0 denotes the external static magnetic field applied to the cavity. Altogether, these factors govern the expected signal power:

$$P_{\text{sig}} = \tfrac{1}{2} Q\, \omega_g^3 V_{\text{cav}}^{5/3} \left(\eta_n h_0 B_0\right)^2 . \tag{7}$$

For the case of HFGWs in the GHz range, direct digitization of the raw cavity signal is impractical. Instead, a local oscillator (LO) is used in a heterodyne mixing scheme to shift the signal to a lower intermediate frequency (typically in the MHz range). This preserves both amplitude and phase information while allowing efficient digitization. The resulting in-phase (I) and quadrature (Q) components can subsequently be stored as a two-channel time series (e.g. in a .tiq file).

This approach is employed by the SUPAX experiment [10], which was previously operated at the University of Mainz and provided the data analyzed in this report. **The sampling rate used is** 14 MHz.

A major challenge in detecting a HFGW signal with a cavity lies in identifying the signal in thermal noise.

Long-established, the state-of-the-art method for GW searches is matched filtering, where noisy detector data are correlated with a large bank of synthetic GW templates. However, this approach comes with several drawbacks in the present context:

- It *assumes Gaussian noise*, which is not valid for the amplitudes of the .tiq data (given by $\sqrt{I^2 + Q^2}$) used in this report. The noise amplitude distribution instead show Rayleigh-like heavy-tailed behavior shown in figure 2. Note that this is not true if working directly with the I and Q component, since both of those are Gaussian distributed (see figure 1).
- The weighting $\propto 1/\text{PSD}$ (PSD = Power Spectral Density) places more emphasis on low-frequency noise, making it ill-suited for the high-frequency signals considered here, and
- the exact signal waveform is generally unknown with a huge parameter space, making the creation of templates a major bottleneck.

For further reading, [12] is referenced.

The above mentioned limitations provide a strong motivation to explore Machine Learning based anomaly detection, where a trained model can perform amortized inference in a single forward pass and detect abnormalities in the noise.

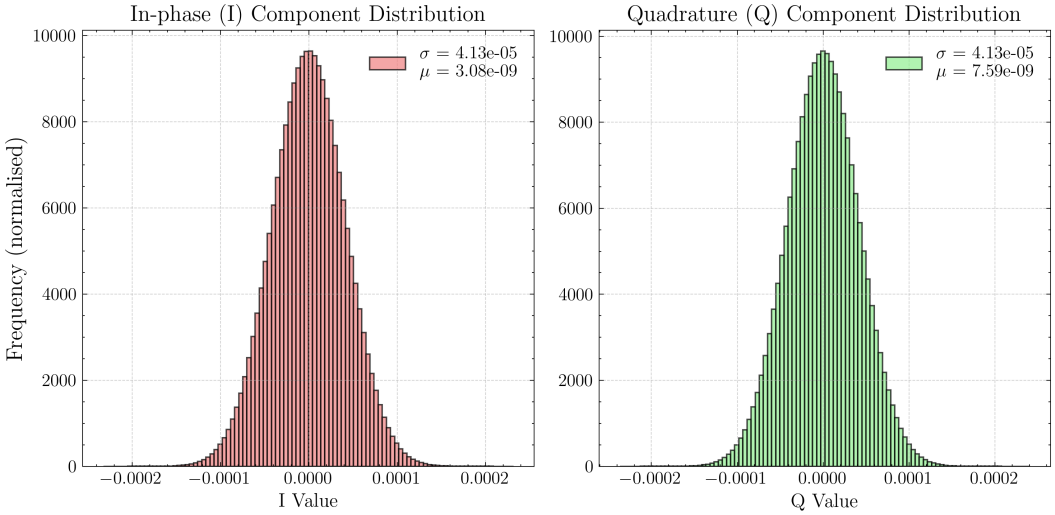

Figure 1: I and Q quadrature distributions used to compute amplitudes in figure 2. The legend show the distribution mean μ and standard deviation σ. Both I and Q components are gaussian distributed.

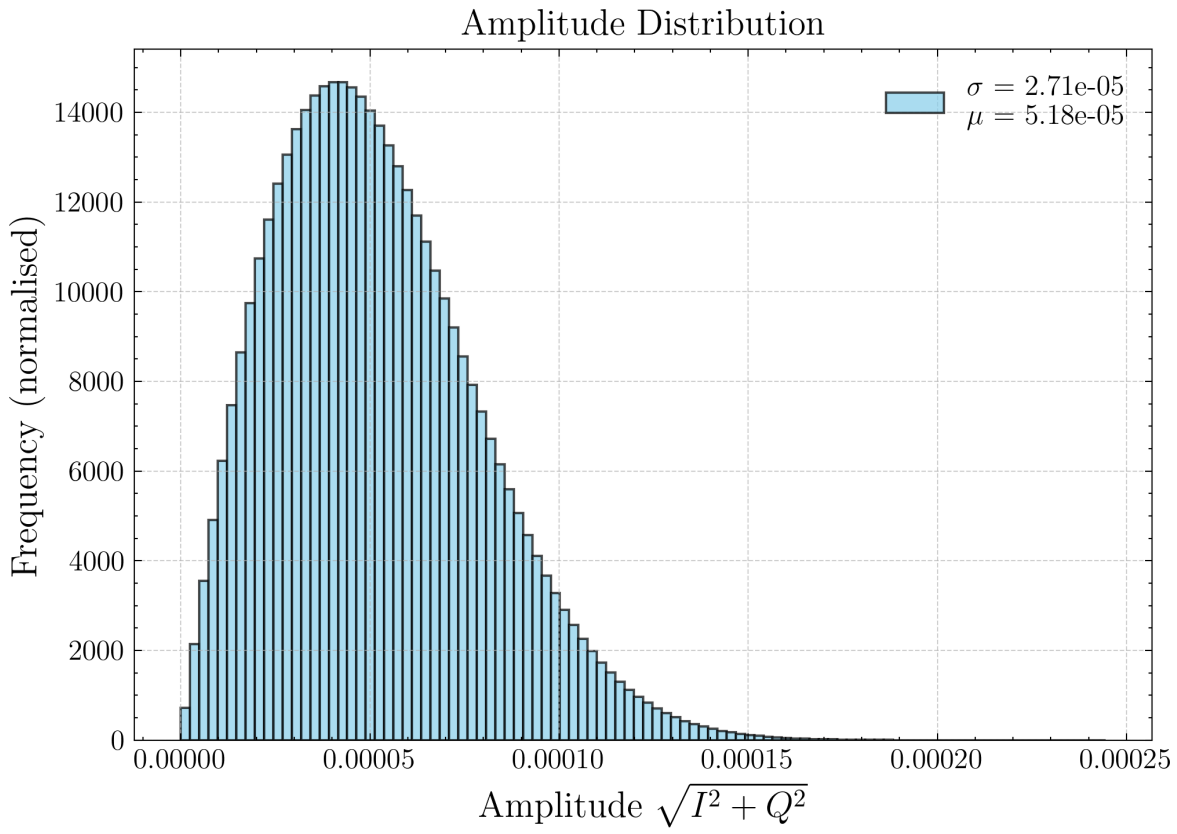

Figure 2: Example amplitude distribution of SUPAX noise data files.

3 Machine Learning

Machine Learning (ML) uses algorithms and statistical models to create a computer system which is able to learn patterns without explicit instructions to make inferences. Considering the task of signal detection inside of noise, a possible approach is the use of **anomaly detection**:

> Training a model such that deviations caused by signals alter a model's behavior and result in one or more anomalous metrics (suitable metrics must be chosen), indicating the presence of a signal.

In the following, three architectures are briefly introduced together with their advantages and drawbacks in the context of anomaly detection.

Recurrent Neural Networks (RNNs). RNNs are designed to capture *long-range dependencies*, which can in principle make them suitable for learning signal shapes inside of noise.

Advantages:

- Modeling of long-range dependencies

Drawbacks:

- They suffer from the vanishing gradient problem (for vanilla RNNs which were considered).
- For anomaly detection they are not well-suited, since noise ideally has no long-range dependencies, requiring training on noise + signal.

Transformers. Transformers use an attention mechanism that weights the importance of each input relative to all others, excelling at modeling long-range dependencies.

Advantages:

- Modeling of long-range dependencies and context

Drawbacks:

- High computational cost of $\mathcal{O}(N^2)$, with N the number of samples.
- Large amounts of data are required, which is challenging given limited hardware (e.g. MacBook Pro with M3 Pro).
- For anomaly detection they are not well-suited, since noise ideally has no long-range dependencies, requiring training on noise + signal.

Variational Autoencoders (VAEs) with Convolutional Neural Network (CNN) as encoder/decoder. A VAE encodes input windows into a latent space, creating an abstract representation of their characteristics, which is then used by the decoder to reconstruct the input. Through the choice of kernel size and window length, both local and global patterns can be captured.

Advantages:

- Both local and global features can be modeled by tuning the receptive field.
- Latent space provides a memory-like representation, enabling interpretability.
- Often cheaper to train and scalable, with considerable freedom in architecture choice.
- Naturally Bayesian: starts with a prior belief and continuously updates to approximate the posterior distribution during training.

Drawbacks:

- Vulnerable to adversarial events (e.g. high noise amplitudes can trigger large errors), though this can be mitigated by threshold calibration.

Comparing all three architectures mentioned above, a CNN-based VAE offers the greatest flexibility in terms of being able to model both local and global dependencies as well as having a latent space which offers additional degrees of freedom for choosing anomaly metrics (e.g. latent variables) and interpretability by linking these metrics to hidden generative factors. In comparison, RNNs and Transformers inherently do not provide such freedom.

Furthermore, VAEs are Bayesian networks, meaning they start with a prior belief about the distribution of latent variables, which represent hidden factors underlying the data, and update this belief during training to approximate the posterior distribution. The posterior then reflects the most likely latent causes of the observed data by assigning probabilities to different values of the latent variables. This probabilistic treatment of the latent space not only increases the flexibility of the model but also underpins its ability to extract hidden generative factors of an input.

In addition, both training on pure noise and noise + signal seems plausible. Training only on noise allows the VAE to learn the background distribution, such that injected signals manifest as deviations in either the reconstruction error or the latent representation. Training on noise + signal, in contrast, enables the model to build a richer latent space in which noise and signal occupy distinct regions, potentially improving interpretability and the range of anomaly metrics that can be defined. This dual possibility underlines the flexibility of VAEs in comparison to sequential models.

Contrasting this with RNNs and Transformers, both model global dependencies, which makes training on pure noise insensible, since noise ideally doesn't have any long-range dependencies [4]. Even if there are subtle dependencies, it is unclear if learning those will assist in anomaly detection.

3.1 VAE architecture

Note that in this section, notation is largely adopted from the VAE introduction paper [13] by Kingma and Welling. In addition, to build an understanding of the matter in [13] ChatGPT was used extensively.

A VAE is a neural network exploiting the variational Bayesian (VB) approach [13] . It models data x through hidden latent variables z, which function as the input to a decoder $p_\theta(x \mid z)$ [5] representing a probabilistic model of the data given z. In essence, z represents the generative factors that underlay the input data, such as amplitude, frequency, and phase in the case of a sine wave.

In addition, a prior distribution $p(z)$, typically a standard normal $\mathcal{N}(0, I)$ for simplicity, is chosen as a regularization assumption of how the z's are distributed a priori.
By Bayes' theorem (see Eq. (8)), inference over the latent variables requires the posterior $p(z \mid x)$, which acts as an updated belief of the prior $p(z)$ given data, since $p(z \mid x)$ tells us which z most likely generated the observed data. A more accurate knowledge of $p(z \mid x)$ results in a more efficient extraction of generative factors. This subsequently enlarges the likelihood $p_\theta(x \mid z)$ of accurate reproduction by the decoder:

$$p_\theta(z \mid x) = \frac{p_\theta(x \mid z)\, p(z)}{p_\theta(x)} \quad \text{with} \quad p_\theta(x) = \int dz\, p_\theta(x \mid z) p(z). \tag{8}$$

The challenge in finding a representation of $p(z \mid x)$ is that the denominator in Eq.(8) is generally intractable [13].

As a way to instead approximate the posterior, VAEs use variational inference, adding a family of distributions $q_\phi(z \mid x)$ over the latent variables and optimizing the set of variational parameters ϕ such that $q_\phi(z \mid x)$, the so-called *approximate posterior*, is close to the true posterior $p(z \mid x)$. Concretely, $q_\phi(z \mid x)$ parametrizes an encoder. A VAE architecture schematic is shown below in figure 3.

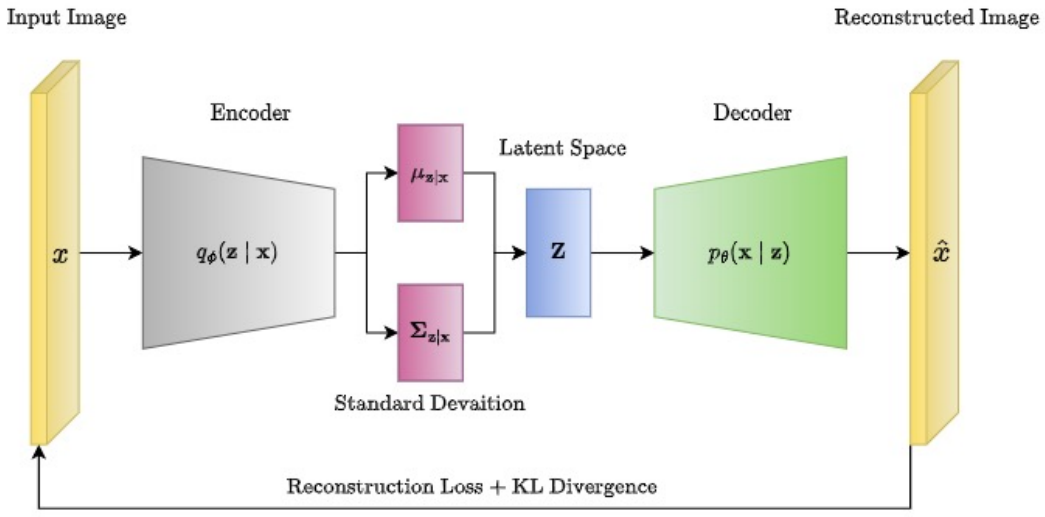

Figure 3: VAE architecture schematic using images as input examples [14] .

[4] Note that this is the ideal case and effects like electronic drift etc. can introduce subtle dependencies.
[5] θ represents the set of weights and biases of the decoder.

The addition of $q_\phi(z \mid x)$ lets us re-write the marginal likelihood $p_\theta(x)$ in Eq.(8) as an expectation value.

$$p_\theta(x) = \int dz\, p_\theta(x \mid z)p(z) = p_\theta(x) = \int dz\, q_\phi(z \mid x)\, \frac{p_\theta(x \mid z)p(z)}{q_\phi(z \mid x)} = \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x \mid z)p(z)}{q_\phi(z \mid x)}\right] \tag{9}$$

To make Eq. (9) more easily solvable, one can apply a log-function to $p_\theta(x)$ for

1. computational stability, turning products into sums and
2. to exploit Jensen's inequality, which states: $\log \mathbb{E}[X] \geq \mathbb{E}[\log X]$ where $\mathbb{E}$ represents the expectation value and X some function,

giving

$$\begin{aligned} \log p_\theta(x) = \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x \mid z)p(z)}{q_\phi(z \mid x)}\right] \geq & \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x \mid z)p(z)}{q_\phi(z \mid x)}\right] \\ & = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] + \mathbb{E}_{q_\phi(z|x)}\left[\log p(z) - \log q_\phi(z \mid x)\right] \end{aligned} \tag{10}$$

The so-found expression in Eq. (10) is known as the Evidence Lower Bound (ELBO).

To approximate the true posterior $p_\theta(z \mid x)$ with $q_\phi(z \mid x)$ and, at the same time, maximize the intractable marginal log-likelihood $\log p_\theta(x)$, one in theory maximizes the ELBO derived in Eq. 10. The exact reasoning why maximizing the ELBO pushes $q_\phi(z \mid x)$ close to $p_\theta(z \mid x)$ is outlined in Appendix A1.

In practice, however, it is more convenient to minimize the negative ELBO, because minimization of a quantity aligns with the fact that most ML libraries, e.g. TensorFlow, are designed to minimize losses using variants of gradient descent.

As a result of this, the Loss function $\mathcal{L}(\theta, \phi; x)$ of a VAE is chosen to be the negative ELBO,

$$\begin{aligned} \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}\Bigg[\underbrace{-\log p_\theta(x \mid z)}_{\text{NLL}}\Bigg] + \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log q_\phi(z \mid x) - \log p(z)\right]}_{\text{KL-divergence}} \\ = \mathbb{E}_{q_\phi(z|x)}\left[\text{NLL}\right] + D_{\text{KL}}(q_\phi(z \mid x)||p(z)) \end{aligned} \tag{11}$$

with the first term being the expectation value of the Negative Log-Likelihood (NLL), which can be interpreted as the reconstruction loss, and the second the Kullback-Leibler divergence (KL divergence). The KL-loss measures the distance between the prior and approximate posterior and acts as a regularizing term for the latent space. The minimization of the ELBO maximizes the reconstruction likelihood, while minimizing the difference between the approximate and true posterior.

Equation (11) reformulates the task of approximating the posterior and reconstructing the input into a single objective: minimizing the negative ELBO. However, in this current form the loss function $\mathcal{L}(\theta, \phi; x)$ is not directly differentiable w.r.t. ϕ. The reason is that the operation of sampling from $q_\phi(z \mid x)$ itself is stochastic and non-differentiable. This prevents gradients from being propagated through the sampling step. A detailed explanation of this issue is provided in Appendix A2.

To bypass this issue, the *reparametrization trick* is introduced. Here samples z from $q_\phi(z \mid x)$ are expressed as deterministic transformations g_ϕ that are differentiable in ϕ [6],

$$z = g_\phi(\varepsilon, x) \quad \text{where} \quad \varepsilon \sim p(\varepsilon) \tag{12}$$

where ε is a variable drawn from a parameter-free noise distribution $p(\varepsilon)$. This isolates the randomness in ε, thereby allowing for the flow of gradients through the sampling step.

[6] To be more precise, g_ϕ is a differentiable function that maps the noise distribution $p(\varepsilon)$ to the target approximate posterior $q_\phi(z \mid x)$.

In practice, a common and computationally convenient choice is to assume both the prior $p(z)$ and the approximate posterior $q_\phi(z \mid x)$ to be Gaussian with diagonal covariance, e.g.

$$q_\phi(z \mid x) \sim \mathcal{N}(\mu_\phi(x), \sigma^2_\phi(x)) \quad , \quad p(z) \sim \mathcal{N}(0, I) \tag{13}$$

This assumption allows both the reparametrization trick as well as the KL-divergence to be written in a closed form [7]. Let $\mu_\phi(x)$ and $\sigma_\phi(x)$ be the parameters of $q_\phi(z \mid x)$, then the sample z can be expressed as

$$z(x) = \mu_\phi(x) + \varepsilon \odot \sigma_\phi(x) \quad \text{with} \quad \varepsilon \sim \mathcal{N}(0, I) \tag{14}$$

where $\odot$ indicates point-wise multiplication.

In this setting, the KL-divergence admits the following closed-form expression. Let J be the dimension of the latent space, μ_j the mean and σ_j the standard deviation of the j-th latent dimension, then for a given input x, the KL-divergence can be written as [13]

$$D_{\mathrm{KL}}(q_\phi(z \mid x)||p(z)) = -\frac{1}{2}\sum_{j=1}^{J}\left(1 + \log \sigma_j^2 - \mu_j^2 - \sigma_j^2\right) \tag{15}$$

Combining the reparametrization trick with the closed expression of the KL-divergence, the remaining expectation value in Eq. (11) can be estimated using Monte Carlo (MC) sampling:

$$\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] = \mathbb{E}_{p(\varepsilon)}\left[\log p_\theta(g_\phi(\varepsilon, x)\right] \simeq \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(g_\phi(\varepsilon^{(l)}, x) \tag{16}$$

Given a dataset $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$ consisting of N data points, and a minibatch $\mathcal{B} \subset \mathcal{D}$ of size $M < N$, $\mathcal{B} = \{x^{(i)}\}_{i=1}^{M}$, the loss can be approximated as follows. Let L denote the number of ε sampled from a standard normal and let $z^{(i,l)}$ be the l-th latent sample corresponding to the i-th data point $x^{(i)}$. Then the loss function for $x^{(i)}$ can be expressed as

$$\mathcal{L}(\theta, \phi; x^{(i)}) = \frac{1}{L}\sum_{l=1}^{L} - \log p_\theta(x^{(i)} \mid z^{(i,l)}) + D_{\mathrm{KL}}(q_\phi(z \mid x^{(i)})||p(z)) \tag{17}$$

With this, the total loss function $\mathcal{L}(\theta, \phi; \mathcal{D})$ of the whole dataset $\mathcal{D}$ can be approximated in terms of randomly drawn mini-batches $\mathcal{B}$ [8].

$$\mathcal{L}(\theta, \phi; \mathcal{D}) = \sum_{i=1}^{N} \mathcal{L}(\theta, \phi; x^{(i)}) \simeq \frac{N}{M}\sum_{i=1}^{M} \mathcal{L}(\theta, \phi; x^{(i)}) \tag{18}$$

Kingma and Welling [13] found that a value of $L = 1$ is sufficient in practice, provided the batch size is large enough. Here, the batch size refers to the number of samples processed simultaneously by a model before computing gradients and updating weights. This estimation technique used in Eq. (18) is referred to as Stochastic Gradient Variational Bayes (SGVB) estimation.

3.2 Training and inference pipeline

The data used to train and do inference with is pre-processed into fixed *window size* intervals which act as the input to the model. This means a single input data point $x^{(i)}$ for time series data is in fact a set of window size data points.

[7] See the original VAE paper [13] by Kingma and Welling for further details.

[8] Note that the factor of N/M ensures an unbiased estimate of the **total dataset loss**. In comparison, most ML libraries compute the **average loss per-data point** over the entire dataset for a more meaningful metric.

Furthermore, the pre-processing stage consists of using min-max normalization of all data, i.e. the normalization to the interval $[0, 1]$, using the values $\text{Min} = 8.0912 \times 10^{-9}$ and $\text{Max} = 2.5262 \times 10^{-4}$ of the training data sets respectively. This is done to avoid instabilities during training and inference and ensure all data is scaled to the same order of magnitude, such that no artifacts arise from order of magnitude differences.

In addition, all testing is done by normalizing new data by the same min and max mentioned above to ensure consistency in the data range a model learned from.

For the latent structure, both the prior $p(z)$ and the approximate posterior $q_\phi(z \mid x)$ were modeled as Gaussian distributions with diagonal covariance matrices, as introduced in Eq.(13). This choice is computationally convenient because the associated KL divergence admits a closed-form expression, shown in Eq.(15), which avoids costly numerical integration.

In terms of the reconstruction objective, the mean-squared error (MSE) was found to be the most suitable anomaly metric, which is discussed further in section 5. This choice corresponds to assuming that the likelihood $p_\theta(x \mid z)$ follows a Gaussian distribution with fixed variance $\sigma^2 = 1/2$, such that the decoder is trained to predict the mean μ of the normalized noise amplitude distribution.

With these distributional choices, the loss function across a single minibatch $\mathcal{B}$ takes the explicit form

$$\mathcal{L}(\theta, \phi; \mathcal{B}) = \frac{1}{M}\sum_{i=1}^{M}(x^{(i)} - \mu^{(i)})^2 + \beta \cdot \frac{1}{M}\sum_{i=1}^{M} D_{\text{KL}}(q_\phi(z \mid x^{(i)}), |, p(z)). \tag{19}$$

where β sets the impact of the KL-divergence loss.

The overall loss to be minimized is the average loss per data point across the full dataset $\mathcal{D}$ with $B = N/M$ minibatches, i.e.

$$\mathcal{L}(\theta, \phi; \mathcal{D}) = \frac{1}{B}\sum_{j=1}^{B}\mathcal{L}(\theta, \phi; \mathcal{B}j) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\theta, \phi; x^{(i)}). \tag{20}$$

Further technical details of the model architecture and chosen Hyperparameters are summarized in Appendix D.

4 Gravitational wave signal simulation

In order to assess the anomaly detection capabilities of the chosen VAE architecture, high-frequency GW signals must first be simulated. Since microwave cavities are used as detectors, both their resonance curve $B(f)$ and the chirp spectrum $\tilde{c}(f)$ have to be considered. The measured signal corresponds to the convolution of the chirp signal and the cavities' response in time domain. To reduce computational cost, the convolution can be performed in the frequency domain by multiplying the chirp spectrum with the cavity's resonance curve, after which the time-domain signal is obtained via an inverse Fast Fourier Transform (iFFT).

4.1 Breit-Wigner resonance curve

The cavity response is modeled by a normalized Breit-Wigner (BW) distribution, parametrized by a resonance frequency f_0 and the curve width Γ, in frequency domain.

For the specific simulation performed in this report, $f_0 = 5\,\text{GHz}$ and $\Gamma = 100\,\text{kHz}$ are used. Since $\Gamma \ll f_0$ the resonance curve is sharply peaked around f_0 and only sensitive to a small range of frequencies $f \approx f_0$. Concretely, a frequency range of $f_0 - 10\Gamma \leq f \leq f_0 + 10\Gamma$, consisting of $N = 32768$ bins, is chosen for the simulation. Under these two assumptions, $\Gamma \ll f_0$ and $f \approx f_0$ the BW-resonance curve can be approximated as

$$B(f) = \frac{\Gamma}{2\pi}\frac{1}{(f - f_0)^2 + \left(\frac{\Gamma}{2}\right)^2} \tag{21}$$

where $\frac{\Gamma}{2\pi}$ is the normalization constant and Γ now represents the Full-Width-at-Half-Max (FWHM) of the BW curve.

4.2 Chirp spectrum

For the chirp spectrum, five simplifying assumptions are made:

1. The GW signal is considered to be a chirp signal $c(t)$, i.e. in general a complex exponential with amplitude A and time varying phase $\varphi(t)$

$$c(t) = A\, e^{-i\varphi(t)} \tag{22}$$

2. The amplitude A is constant across all considered frequencies [9].

3. The spectrum of $c(t)$ can be approximated using the stationary phase approximation (SPA), which is motivated by the fact that the phase of the complex exponential in Eq. (23) is rapidly varying with time and therefore mainly the stationary points contribute to the integral. The accuracy of the SPA is shown in [15]. Let $\tilde{c}(f)$ denote the Fourier transform (FT) of $c(t)$, then according to the SPA one obtains

$$\begin{aligned} \tilde{c}(f) &= \int dt\, A\, e^{-i\varphi(t)}\, e^{-i\omega t} \quad , \quad \Phi(t) = \varphi(t) - \omega t \\ &= \int dt\, e^{-i\Phi(t)} \quad \approx \quad A\, e^{-i\Phi(t_s)} \sqrt{\tfrac{2\pi}{|\ddot{\Phi}(t_s)|}}\; e^{\,i\frac{\pi}{4}\,\mathrm{sgn}(\ddot{\Phi}(t_s))}. \end{aligned} \tag{23}$$

   The additional factor $e^{i\pi/4\,\mathrm{sgn}(\ddot{\Phi}(t_s))}$ corresponds to a constant, global phase shift that is independent of frequency. Since only the frequency evolution of the signal is of interest, this factor will be disregarded.

4. A linear chirp is considered, i.e. the frequency $f = \frac{\dot{\varphi}(t)}{2\pi}$ varies linearly with time t.

$$2\pi f(t) = \dot{\varphi}(t) = 2\pi f_{\mathrm{start}} + k\, t, \tag{24}$$

   where f_{start} is the starting frequency and k the chirp rate. Integrating yields the time-domain phase

$$\varphi(t) = \frac{1}{2} k t^2 + \omega_{\mathrm{start}} t + \varphi_0, \tag{25}$$

   with $\omega_{\mathrm{start}} = 2\pi f_{\mathrm{start}}$ and φ_0 an initial phase.

   The stationary time t_s follows from the SPA condition $\frac{d}{dt}\Phi(t) = \dot{\Phi}(t_s) = 0$, i.e.

$$\dot{\Phi}(t) = \dot{\varphi}(t) - 2\pi f = 0 \quad \Longrightarrow \quad t_s = \frac{f - f_{\mathrm{start}}}{k}. \tag{26}$$

   Evaluating the total phase $\Phi(t)$ at t_s gives the SPA frequency-domain phase

$$\Phi(t_s) = -\frac{\pi}{k}\left(f - f_{\mathrm{start}}\right)^2, \tag{27}$$

5. Furthermore, the chirp rate k is considered, for two PBH of equal mass m_{PBH} in the inspiral phase of a merger, to be [7]

$$k = 4.62 \times 10^{11}\,\mathrm{Hz}^2 \left(\frac{m_{\mathrm{PBH}}}{10^{-9} M_\odot}\right)^{5/3} \left(\frac{f_{\mathrm{start}}}{\mathrm{GHz}}\right)^{11/3} \quad \text{with} \quad f_{\mathrm{start}} = f_0 - 10\Gamma \tag{28}$$

   where $M_\odot$ is the solar mass.

[9] Note that this is a pure simplification and is planned to be changed in future works.

4.3 Simulated signal

Combining the Breit–Wigner response $B(f)$ from Eq. (21) with the chirp spectrum $\tilde{c}(f)$ obtained via the SPA in Eq. (23) gives the full frequency-domain signal. Applying an inverse Fourier transform (iFFT) yields the corresponding time-domain signal. In addition, since the signal duration can be orders of magnitude shorter than the full time window (depending on the number of frequency bins used), the time-domain waveform is cropped by retaining only those samples whose magnitude exceeds 1×10^{-3} of the maximum amplitude [10] . This procedure isolates the relevant waveform while discarding zero values. It should be noted that only the time-domain signal shape is of interest here; no further frequency-domain analysis is performed on the cropped data. Two exemplary signals obtained in this way are shown in Figs. 4 and 5.

.

An important concept for interpreting the simulated signal shapes is the response time of the cavity [11] .

Formally, the response time τ is defined as the characteristic timescale over which the amplitude of the cavity's electric field builds up (when driven on resonance) or decays (after the driving stops) to $1/e$ of its maximum value. Physically, this means that the cavity cannot instantaneously react to a driving signal but instead requires a finite time to establish or lose its stored energy.

The relevance of τ becomes apparent when compared with the time Δt a chirping GW signal is in resonance with the cavity. Since the frequency of the GW evolves linearly with time, the interval during which it stays within the resonance bandwidth of the cavity is

$$k = \frac{\Delta f}{\Delta t} \quad \Longleftrightarrow \quad \Delta t = \frac{\Delta f}{k}, \tag{29}$$

where Δf is the frequency interval considered. For the cavity, Δf can be set to the full width at half maximum (FWHM), $\Gamma = 100\,\mathrm{kHz}$.

The comparison of Δt and τ determines the observed signal shape:

- If $\Delta t > \tau$, the GW drives the cavity long enough for the field to build up and oscillate for several cycles before decaying, leading to signals with many visible oscillations (see figure 4, and figure 11 in Appendix B). It should be noted that a signal was also simulated for a mass of $m_{\mathrm{PBH}} = 1 \times 10^{-13} M_{\odot}$. While this signal likewise exhibits oscillatory behavior, its duration is too long for the oscillations to be resolved clearly in a figure and is therefore not shown explicitly. Its shape, however, closely resembles that of figure 4, with a greater number of oscillations and an extended overall duration.
- If $\Delta t \lesssim \tau$, the GW passes through resonance faster than the cavity can respond. In this case, the cavity only produces a single impulse-like response, resembling a short exponentially rising and decaying pulse (see figure 5). This behavior sets in for masses around $m_{\mathrm{PBH}} = 1 \times 10^{-9} M_{\odot}$. For this reason, although simulations were performed across the full mass range $1 \times 10^{-13} M_{\odot} \leq m_{\mathrm{PBH}} \leq 1 \times 10^{-6} M_{\odot}$, only one representative example is shown in figure 5 for the higher-mass cases.

Additionally, the impulse shape of heavier mass signal can be explained by the number of oscillations induced in the cavity, which can be estimated as

$$N_{\mathrm{cycles}} \approx \Delta t \cdot f_0, \tag{30}$$

As Δt is longer than the oscillation time $1/f_0$ per cycle, both the real and imaginary parts of a signal will exhibit oscillatory behavior. Once $\Delta t < 1/f_0$, the time the GW signal is in resonance will be too short to create any oscillation.

[10] The value of 1×10^{-3} has been experimentally found to isolate the signal well and not overly cut away signal samples.
[11] Note that, ChatGPT was used for building an understanding of the concept of response time.

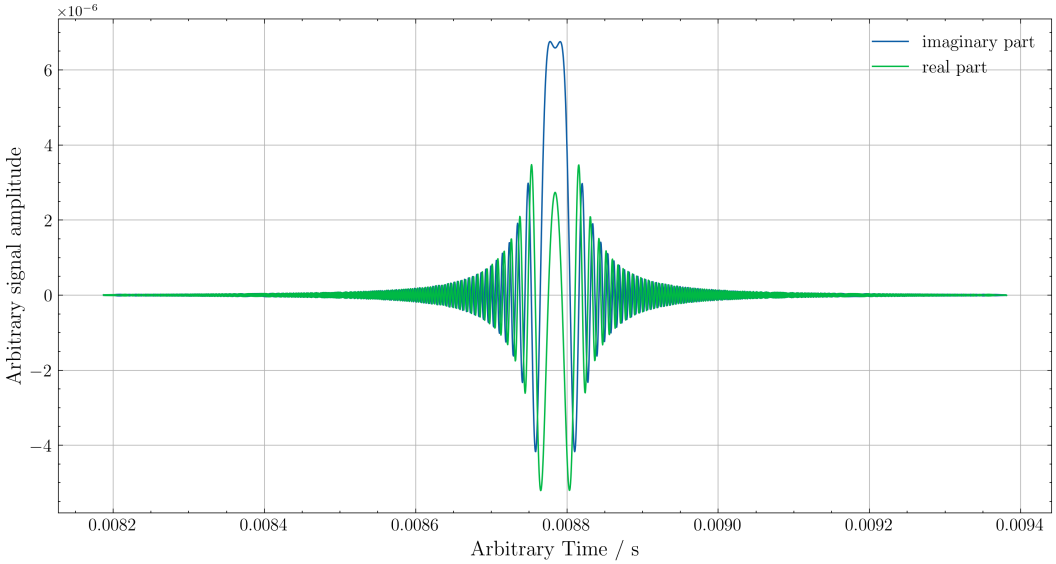

Figure 4: A simulated high-frequency GW signal of two merging PBHs of mass $m_{\mathbf{PBH}} = 1 \times 10^{-12} M_\odot$ each, measured with a microwave cavity. The number of frequency bins is $N = 32768$ and the chirp rate is $k \approx 1.69 \times 10^9 \, \frac{\mathrm{Hz}}{\mathrm{s}}$. The effective resonance time is $\Delta t \approx 5.92 \times 10^{-5}\,\mathrm{s}$, which is longer than both the response time $\tau = 1 \times 10^{-5}\,\mathrm{s}$ and the oscillation period $1/f_0 = 2 \times 10^{-10}\,\mathrm{s}$ set by the resonance frequency $f_0 = 5\,\mathrm{GHz}$. As a result, the electric field inside the cavity exhibits several oscillations in the real (green) and imaginary part (blue). The absolute time scale on the x-axis is not physically relevant; only the interval between successive zeros defines the effective signal duration.

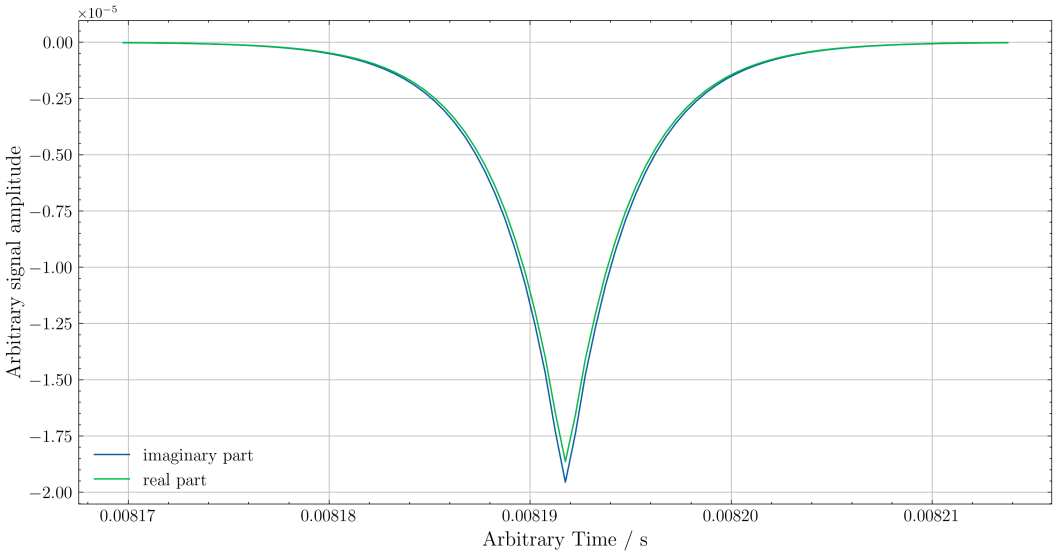

Figure 5: A simulated high-frequency GW signal of two merging PBHs of mass $m_{\mathbf{PBH}} = 1 \times 10^{-9} M_\odot$ each, measured with a microwave cavity. The number of frequency bins is $N = 32768$ and the chirp rate is $k \approx 1.69 \times 10^{14} \, \frac{\mathrm{Hz}}{\mathrm{s}}$. The effective resonance time is $\Delta t \approx 5.92 \times 10^{-10}\,\mathrm{s}$, which is shorter than the response time $\tau = 1 \times 10^{-5}\,\mathrm{s}$, but longer than the oscillation period $1/f_0 = 2 \times 10^{-10}\,\mathrm{s}$ set by the resonance frequency $f_0 = 5\,\mathrm{GHz}$. As a result, the electric field inside the cavity merely exhibits an exponential rise and exponential fall-off, with no oscillations. The absolute time scale on the x-axis is not physically relevant; only the interval between successive zeros defines the effective signal duration.

4.4 The impact of the chosen number of frequency bins

It is worth noting that the specific choice of the number of frequency bins N directly affects the shape of the simulated time-domain signal. This is because the distance Δf between two adjacent frequency bins is directly dependent on N and given by [12]

$$\Delta f = \frac{f_{\text{span}}}{N-1} = \frac{f_{\max} - f_{\min}}{N-1} \tag{31}$$

where $f_{\max}$ is the maximum and $f_{\min}$ the minimum grid frequency.

Altering N will directly result in a different Δf, thereby changing the iFFT basis function's phase, and with that the interference behavior. This is evident by ...

$$s(t_n) = \frac{1}{N}\sum_{k=0}^{N-1} Y(f_k)\, e^{2\pi i f_k\, t_n} = \frac{1}{N}\, e^{2\pi i f_{\min}\, t_n} \sum_{k=0}^{N-1} Y(f_{\min} + k\cdot\Delta f)\, e^{2\pi i\, k\cdot\Delta f\, t_n} \tag{32}$$

where f_k defines the frequency grid by

$$f_k = f_{\min} + k\cdot\Delta f \quad \text{with} \quad k = 0, 1, 2, \ldots, N-1 \tag{33}$$

Different interference behavior can directly lead to changes in the time-domain signal waveform such as polarity flips in the real or imaginary parts or both, and shifts from constructive to destructive interference or vice versa. For example, the impact of differently chosen N is shown in Appendix B.

With this in mind, the results presented in section 5, which are based on $N = 32768$, should be interpreted with the above considerations in mind.

5 Results

The results discussed below are based on the distribution configuration, examined in section 3.2 and the use of MSE as the anomaly score.

For testing purposes, signals were generated for a range of PBH masses of $1\times10^{-13} M_\odot \leq m_{\text{PBH}} \leq 1\times10^{-6} M_\odot$ using the procedure described in section 4. The injection into the pure noise time series is done probabilistically, meaning signals are injected in the full time series, choosing the start and end index of a signal randomly (but inside the time series and end index > start index). Afterwards, the windowing described in section 3.2 is applied. Further, the results in figure 6 were obtained by injecting 1000 simulated signals into 1.399 million windows of size 64 and 128, respectively. For each configuration, the detection efficiency was measured over a range of SNR values from 9.0 down to 1.0. The resulting efficiency curves were then fitted with a sigmoid function of the form

$$s(x) = \frac{L}{1 + e^{-k\cdot(x-x_0)}}, \tag{34}$$

as shown for exemplary cases in Appendix C.

Since the testing procedure is essentially a binomial process of success or failure, Wilson confidence intervals [16] were used. These provide a more accurate characterization of the uncertainty, especially in the regimes of low efficiencies near 0 and high efficiencies near 1 .

From the fitted sigmoid parameters, the SNR corresponding to a detection efficiency of $\epsilon = 95\%$ was determined, including its uncertainty derived from propagation of the fit parameter errors. It should be noted, however, that for a mass of $m_{\text{PBH}} = 1\times10^{-13}\, M_\odot$, shown in the lower panel of figure 6, the uncertainty in the extracted SNR is particularly large. This originates from the sharp transition in the efficiency curve, where the sigmoid fit becomes poorly constrained. The corresponding curve is shown in figure 13 in Appendix C.

[12] Note that this assumes both endpoints of the frequency range which spans the grid are included.

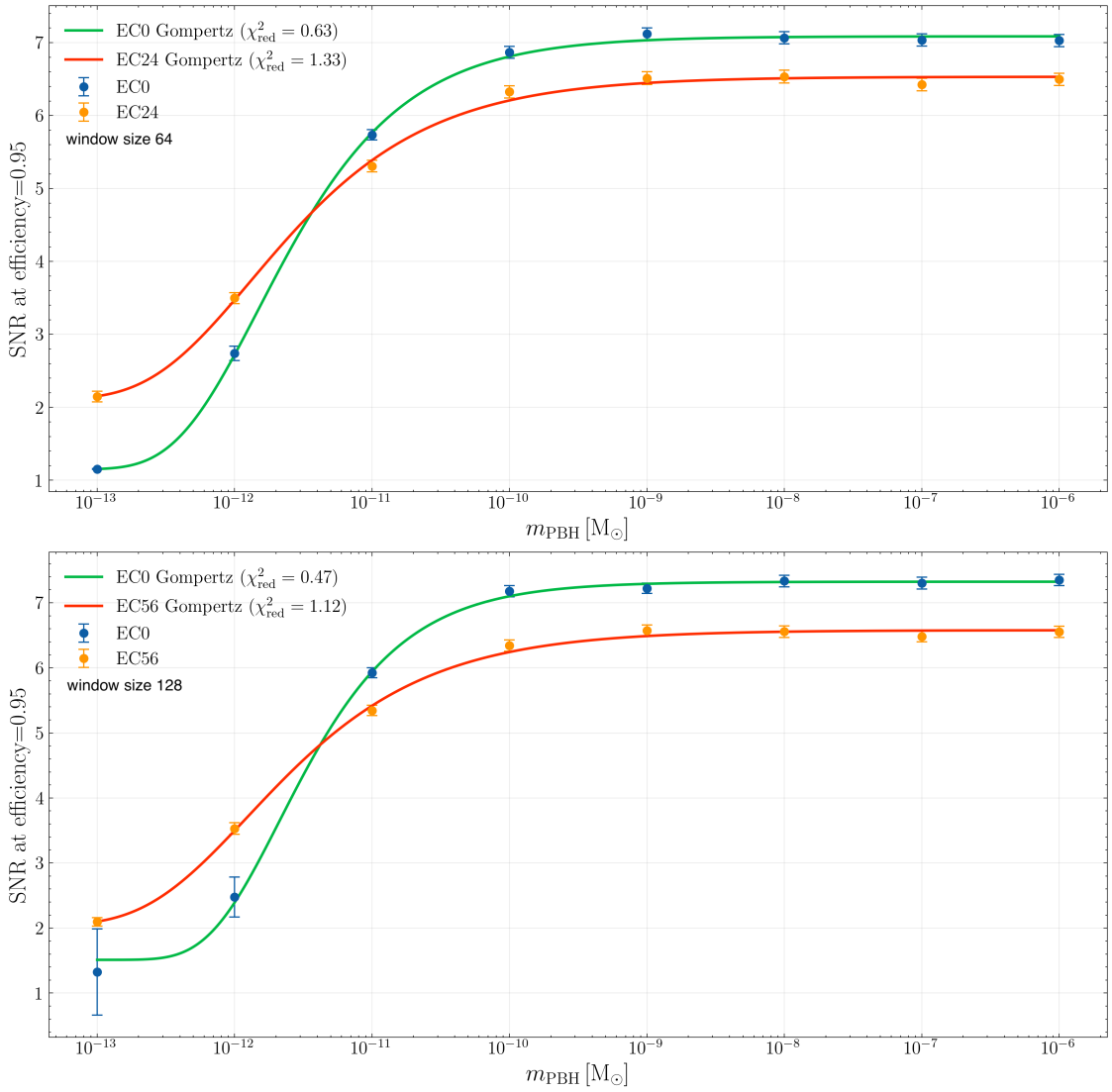

Figure 6: Required Signal-to-Noise Ratio (SNR) to achieve $\epsilon = 95\%$ detection efficiency for simulated gravitational-wave signals from merging PBHs. The top panel corresponds to a model trained with window size 64, the bottom to a model with window size 128. Detection efficiency is defined such that a signal is counted as detected if at least one of the windows it spans exceeds a prior set MSE threshold. EC stands for *Edge Cropped*, with the number next to EC indicating how many samples a window was cropped by on both the left and right side (here both 64 and 128 samples were cropped to the center 16 samples, which were then used for MSE calculation). The efficiencies were achieved with a false positive rate (FPR) of 0 FP for both window sizes on the test data set of 1.39 million window. The resulting upper bound for false positives (FP) per day is 3.26 $\times 10^5$ FP/day by the rule of three.

The descriptive fit curves in figure 6 are obtained by using a Gompertz function, which is defined as

$$g(x) = A + (L - A) \cdot e^{-e^{-k \cdot (x - x_0)}} \tag{35}$$

Comparing the cropped window results with the un-cropped windows, it is apparent that the cropped curves yield lower SNR values required to reach $\epsilon = 95\%$ detection efficiency for heavier masses in the range $1 \times 10^{-11} \leq m_{\text{PBH}} \leq 1 \times 10^{-6}$. On the other hand, though, the un-cropped curve undercuts the cropped one for masses $m_{\text{PBH}} = 1 \times 10^{-12}$ and $1 \times 10^{-13}\, M_\odot$.

Physically, this behavior can be understood by noting that smaller PBH masses generate longer GW signals on the order of $10^3 - 10^5$ samples. Such long signals alter the statistical properties of the noise less strongly, i.e. they induce only small changes in the variance of the noise-dominated data (**the role of statistical changes will further be discussed below**). In contrast, heavier masses produce shorter and sharper signals, to be more precise, 90 sample long signals, that stand out more clearly against the noise background. As a result, higher SNRs are required to reliably distinguish signal from noise [13].

Furthermore, both orange fit lines, obtained by just the center 16 samples, look equivalent.

This equivalence is quantitatively supported by the fact that all Gompertz fit parameters for EC24 and EC56 agree within their 1-σ uncertainty intervals (see table 1), whereas the EC0 parameters deviate by several σ, indicating a systematic difference.

Table 1: Fitted Gompertz parameters for different cropping configurations and window sizes. The near-identical values of EC24 and EC56 support the equivalence of cropped window size 64 and 128 models, while EC0 shows clear differences.

Configuration	**Window size**	A	L	k	x_0
EC0	64	1.15(1)	7.09(4)	1.66(8)	−11.83(3)
EC0	128	1.21(2)	7.02(5)	1.61(9)	−11.84(3)
EC24	64	2.11(9)	6.53(5)	1.37(10)	−11.88(4)
EC56	128	2.16(8)	6.50(5)	1.35(9)	−11.87(4)

This near-perfect agreement between these sets of parameters strengthens the conclusion that cropping renders the two models equivalent in the sense of obtaining equivalent MSE distributions and, as a result, equivalent detection efficiencies. The main implication is

- It is possible to choose a bigger window size, crop the window to the center 16 samples, calculate the MSE, and still obtain equivalent efficiency results[14]. This strategy can drastically reduce the expected number of false positives (FPs). Assuming a constant per-window false alarm probability p_{FA}, the expected number of false positives scales linearly with the total number of windows n as

$$\mathbb{E}[\mathrm{FP}] = n \cdot p_{\mathrm{FA}} \tag{36}$$

  Here, given S total samples, n is determined by

$$n = \frac{S - \text{window size}}{\text{step size}} + 1 \quad \text{with} \quad \text{step size} = \frac{\text{window size}}{m} \quad , \quad m \in \mathbb{N} \setminus \{0\}, \tag{37}$$

  where m denotes the overlap factor. For fixed S and overlap ratio, increasing the window size reduces n approximately proportional to 1/window size. Cropping then restores the detection efficiency, making this approach advantageous both in terms of accuracy and statistical stability.

5.1 Mechanism of Detection via MSE

Based on the use of MSE for anomaly detection and the fact that the decoder learns to output the point-wise mean $\mu^{(i)}$, the MSE is a direct measure of the average (squared) distance between the mean and all other data points inside an input window. A deviation in the MSE caused by a present signal inside an input window can thus be the result of the following

1. The amplitude content inside a window is altered sufficiently by an injected signal, while the mean reconstruction doesn't differ between pure noise and noise + signal windows
2. The point-wise mean reconstruction drifts in signal windows compared to pure noise windows, while the amplitude content of a window remains roughly the same

[13] Note that other cropping values have been tried, but cropping to the center 16 samples was found to have the strongest impact on lowering the needed SNR for detection.

[14] It is important to note though, that the anomaly threshold on MSE values must be calibrated explicitly for each model.

3. Both the amplitude content and the point-wise mean differ in signal + noise windows compared to pure noise windows

This can quantitatively be expressed via the bias–variance decomposition of the MSE loss for input $x^{(i)}$, which admits to [17]

$$\mathrm{MSE}_i = \underbrace{\mathrm{Var}[x^{(i)}]}_{\text{spread of amplitudes}} + \underbrace{\mathrm{Var}[\hat{\mu}^{(i)}]}_{\text{spread of decoder reconstruction}} + \underbrace{\left(\mu^{(i)} - \hat{\mu}^{(i)}\right)^2}_{\text{squared bias of the learned mean}} \tag{38}$$

where $\mu^{(i)}$ is the true window mean.

Hence, when the decoder accurately tracks the mean ($\hat{\mu}^{(i)} \simeq \mu^{(i)}$) and the reproduction variance is sufficiently low compared to the variance of the amplitude data, the MSE is essentially the *amplitude variance*. Any variance of $\hat{\mu}^{(i)}$ or $\mu^{(i)}$ or departure of both will add a positive bias term and raise the MSE.

Variance in the decoder reconstruction $\hat{\mu}^{(i)}$ Figure 14 in Appendix C plots the distribution of the decoder's *per-window mean* on pure unseen test noise. The learned mean by the decoder clusters tightly around $\hat{\mu}^{(i)} \approx 2.0458 \times 10^{-1}$ with fluctuations only at the 4th–5th decimal place, indicating small relative variability in the decoder reconstruction. For reference, Figure 15 in Appendix C shows the global amplitude distribution of the normalized test set with mean $\mu = 2.049 \times 10^{-1}$ and standard deviation $\sigma = 1.071 \times 10^{-1}$; i.e., the intrinsic spread of samples is orders of magnitude larger than the drift observed in $\hat{\mu}^{(i)}$.

Furthermore, performing inferences with the model several times shows only output distribution drifts on the order of 10^{-5}.

As a consequence, the variance term in Eq. (38) can be neglected due to its small size.

Squared bias of the learned mean $\left(\mu^{(i)} - \hat{\mu}^{(i)}\right)^2$ Figure 7 shows the distribution of window means before being passed into the VAE for both pure noise and signal-containing windows.

The difference illustrates that the mere injection of signals already alters the mean distribution inside an input window. Since the decoder output $\hat{\mu}^{(i)}$ is essentially constant, the squared bias term is only changed by a different per-window mean $\mu^{(i)}$, mostly giving a non-zero contribution to the MSE.

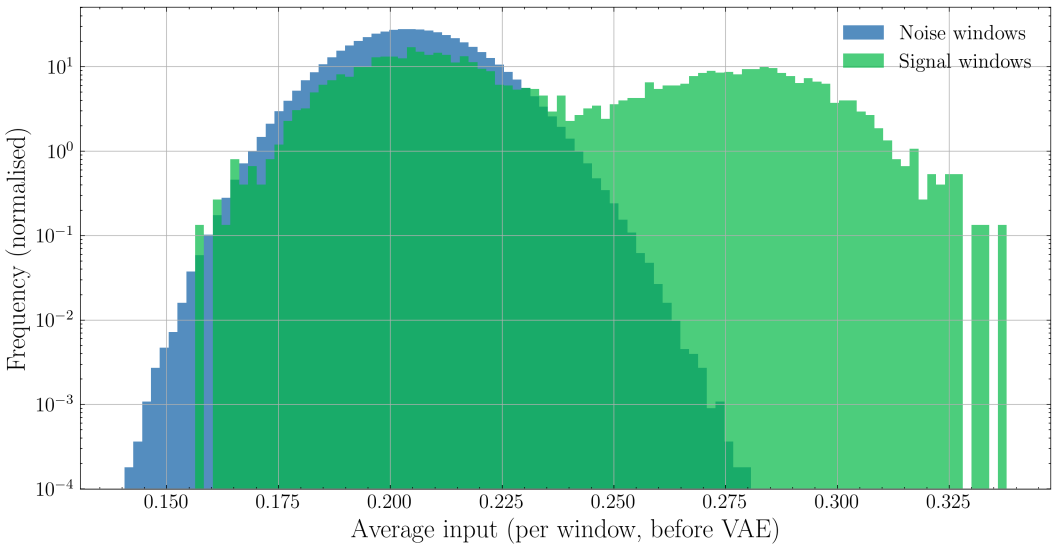

Figure 7: Distribution of the mean per window (before VAE) for pure noise and noise + signal windows. The window size 64 model was used for EC0 cropping and 200 signals injected into 2.79 million windows. A step size of 8 samples was used.

Variance contribution $\mathrm{Var}[x^{(i)}]$ Complementarily, figure 8 displays the per-window variance of each input window before the VAE. It is apparent that injecting a signal alters the per-window variance to a significant extent. Comparing this figure 7, the effect of an injected signal on the per-window variance is more pronounced than the effect on the per-window mean. Consequently, the dominant distribution to the MSE boils down to the change in per-window variance.

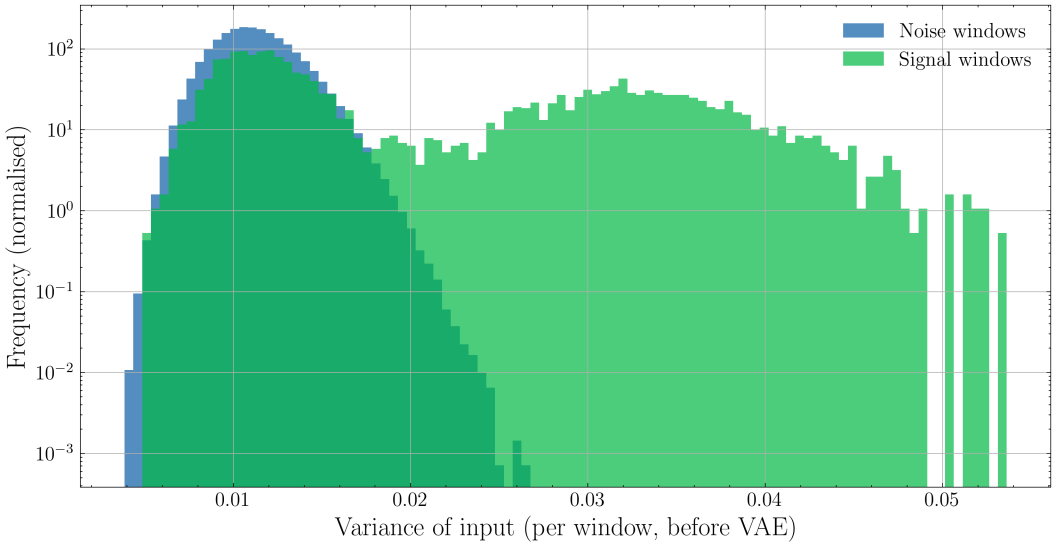

Figure 8: Distribution of the variance per window (before VAE) for pure noise and noise + signal windows. The window size 64 model was used for EC0 cropping and 200 signals for $m_{\mathrm{PBH}} = 10^{-6} M_{\odot}$ with an SNR = 7 were injected into 2.79 million windows. A step size of 8 samples was used.

MSE Based on the observations in figure 7 and 8 as well as Eq. (38), the total maximal MSE can be estimated by summing up the maximal possible value of the per-window variance $\sigma^2_{(i)} \approx 0.052$ and the squared-bias term $\left(\mu^{(i)} - \hat{\mu}^{(i)}\right)^2$, giving

$$\mathrm{MSE}_i = \sigma^2_{(i)} + \left(\mu^{(i)} - \hat{\mu}^{(i)}\right)^2 \approx 0.067 \quad \text{with} \quad \hat{\mu}^{(i)} \approx 0.205 \quad , \quad \mu^{(i)} \approx 0.330 \tag{39}$$

Comparing the above result with the actual MSE distribution, after the input was passed through the VAE, shown in figure 9, the maximal value of 0.062 is roughly in accordance with the estimation of Eq. (39). Note that one, it is a very rough estimate to assume that the highest $\sigma^2_{(i)}$ always corresponds to the highest $\hat{\mu}^{(i)}$, and two, the values used for estimation are estimates based on the histogram, introducing another error source. Thus, the discrepancy between estimated max MSE = 0.067 and real max MSE = 0.062.

The main implications are

- The interplay of the squared bias and amplitude variance, governed by the bias-variance decomposition in Eq. (38), determines the detectability of varying-length signals, when using the MSE as the anomaly metric.
- The contribution of the VAE is consistently accurately predicting the mean of pure noise windows, with low variance, indicating great ability in this.

With the bias–variance decomposition in Eq.(38), the condition for detection using the MSE is that a signal must sufficiently modify the noise statistics. This behavior is illustrated by the efficiency curves in Fig.11 and Fig.13 in Appendix C, where the detection efficiency decreases as the signal induces progressively smaller changes in the noise mean and variance. This is a direct reflection of the fact that the signal window distribution in figure 9 becomes less separated from the noise window distribution as the amplitude drops, making detection via MSE anomalies not possible.

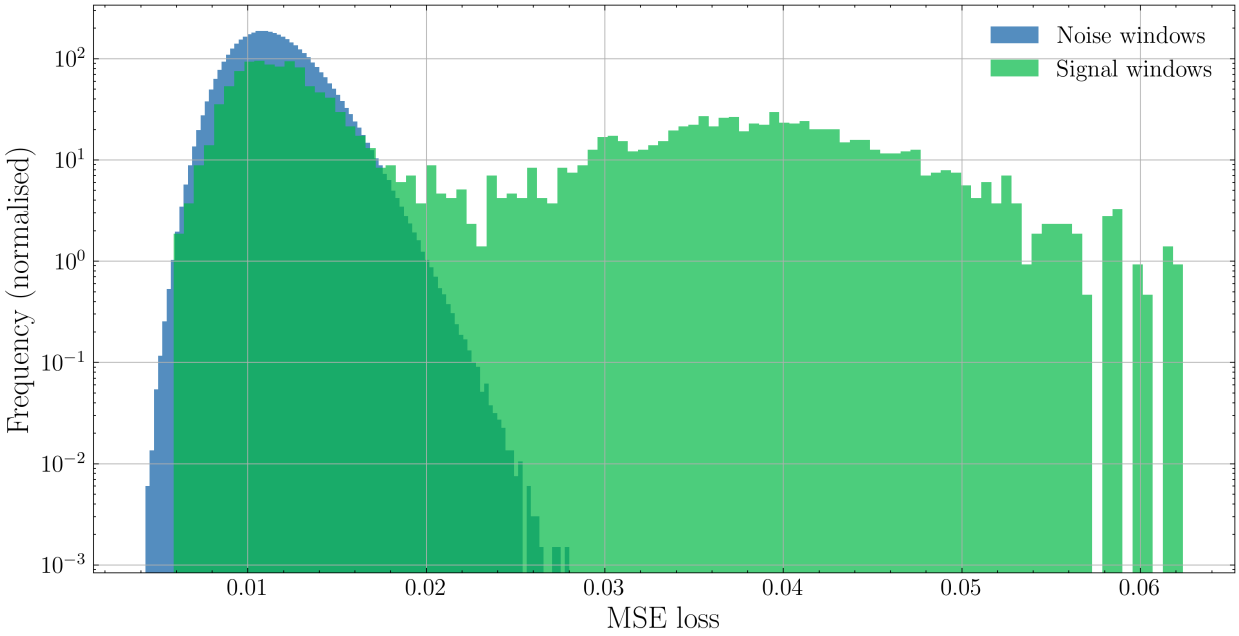

Figure 9: Distribution of the reconstruction error MSE after model inference. The window size 64 model was used for EC0 cropping and 200 signals for $m_{\mathrm{PBH}} = 10^{-6} M_{\odot}$ with an SNR = 7 were injected into 2.79 million windows. A step size of 8 samples was used.

It becomes apparent why detection performance differs between models trained with different window sizes. Despite the here-mentioned small window sizes of 64 and 128, larger ones, ranging from 1024 to 32768, have been tested as well.

For short signals, larger windows tended to perform worse. The reason is that a higher sample count per window leads to a lower variance in the mean of each prepared window. As a consequence, the per-window mean $\mu^{(i)}$ becomes mostly identical to the decoder reconstruction, such that the squared bias term $\left(\mu^{(i)} - \hat{\mu}^{(i)}\right)^2$ approaches zero.

At the same time, the per-window variance is only locally altered by the short signal, but since the noise contribution dominates the total variance in large windows, the effect of the signal is effectively washed out.

By contrast, longer signals can still be detected with large window sizes. This is because longer signals induce a global shift in both the mean and the variance of the data, leading to a measurable deviation in the MSE. In other words, the bias term no longer vanishes, just as it is the case for short signals and small window sizes.

This explains why, for example, signals from primordial black holes with mass $m_{\mathrm{PBH}} = 1 \times 10^{-13} M_{\odot}$, corresponding to signal lengths on the order of 3.2×10^5 samples, can be reliably detected with models trained on large windows. In parallel, shorter signals are best captured with smaller windows (e.g. 64 samples), since here the signal alters the local statistics sufficiently to yield an increased MSE, in agreement with the bias–variance decomposition of Eq. (38).

6 Summary

In summary, the detection capabilities can be attributed to the shift a GW signal induces in the mean and variance of the noise. The MSE anomaly metric is fully determined by these two quantities, described by the bias-variance decomposition,

$$\mathrm{MSE}_i \approx \underbrace{\mathrm{Var}[x^{(i)}]}_{\text{spread of noise amplitudes}} + \underbrace{\left(\mu^{(i)} - \hat{\mu}^{(i)}\right)^2}_{\text{squared bias of the learned mean}} \tag{40}$$

As demonstrated in the results, a VAE trained on pure noise effectively gives a near constant estimate $\hat{\mu}^{(i)}$ of the noise mean $\mu^{(i)}$. Consequently, because the MSE reflects only changes in mean and variance, anomaly detection using solely it imposes a fundamental limitation: GW signals can only be detected with high efficiency and simultaneously low false-positive rates when they cause sufficiently strong alterations of these noise statistics.

7 Outlook

Looking ahead, several directions appear promising for improving the presented detection method. First, finer threshold tuning on larger datasets to lower the current upper bound on the false positives per day, with a goal of less than 1 FP/year.

One possibility to do so, without needing a whole year of detector noise data being taken, is to plot the MSE threshold against the per-window FPR for a dataset and fit a function to the data. Subsequently, by means of extrapolation, the necessary threshold to achieve less than 1 FP/year can be found [15] .

Moreover, training the VAE not only on pure noise but also on noise + signal data, and incorporating the latent variables into anomaly metrics, may further enhance the sensitivity of the approach.

Furthermore, extending training a dedicated classifier trained on multiple features, e.g. latent variables, KL-loss, MSE, could provide an additional boost in detection efficiency.

Finally, considering an alternative approach of using RNNs or Transformers, trained on noise + signal to capture signal shapes, may give a more capable detection model beyond a VAE.

8 Acknowledgments

I want to thank Kristof Schmieden for the instructive conversations we had as well as the guidance he gave.

Additionally, I thank Matthias Schott and Maximilian Miles for their feedback and advice on the thesis formulation.

[15] Note that for the found threshold, the detection efficiencies probably change and therefore have to be assessed again

9 Appendix

A1 - Posterior approximation by maximizing the ELBO

The following elaborates on why maximizing the ELBO pushes the approximate posterior $q_\phi(z \mid x)$ towards the true posterior $p_\theta(z \mid x)$. A central identity for the derivation below is: [18]

$$p(x, z) = p(z \mid x)p(x) = p(x \mid z)p(z) \tag{41}$$

Consider the difference between the log-likelihood $\log p_\theta(x)$ and the ELBO, the result is the divergence between $q_\phi(z \mid x)$ and $p_\theta(z \mid x)$:

$$\begin{aligned}
\log p_\theta(x) - \text{ELBO} &= \log p_\theta(x) - \mathbb{E}_{q_\phi(z|x)} \log \left[\frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right] \\
&= \log p_\theta(x) - \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x, z) - \log q_\phi(z \mid x) \right] \\
&= \log p_\theta(x) - \underbrace{\mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x) \right]}_{\log p_\theta(x)} + \underbrace{\mathbb{E}_{q_\phi(z|x)} \left[\log q_\phi(z \mid x) - \log p_\theta(z \mid x) \right]}_{D_{\mathrm{KL}}(q_\phi(z|x) || p_\theta(z|x))} \\
&= D_{\mathrm{KL}}(q_\phi(z \mid x) \,||\, p_\theta(z \mid x)) \\
\iff \quad \log p_\theta(x) &= \text{ELBO} + D_{\mathrm{KL}}(q_\phi(z \mid x) \,||\, p_\theta(z \mid x))
\end{aligned} \tag{42}$$

Now, for a given fixed set of decoder parameters θ, which determine the reconstruction of an input, the log-likelihood is fixed. When now maximizing the ELBO, the only way the equality in Eq. (42) can hold is if the divergence between $q_\phi(z \mid x)$ and $p_\theta(z \mid x)$ shrinks, i.e. $q_\phi(z \mid x)$ approximates $p_\theta(z \mid x)$. This is due to the fact that the KL-divergence $D_{\mathrm{KL}}(q_\phi(z \mid x) \,||\, p_\theta(z \mid x))$ is greater than or equal to zero [13].

In practice, during model training both θ and ϕ are jointly updated. The optimizer used will find the joint optimum solution of generative + reconstruction ability by the decoder (θ set) and the approximation of the true posterior by the encoder (ϕ set). The relative dominance between the KL-loss used in training ($D_{\mathrm{KL}}(q_\phi(z \mid x) \,||\, p(z))$) and the reconstruction loss ($\mathbb{E}_{q_\phi(z|x)} \left[-\log p_\theta(x \mid z) \right]$) is therefore a crucial design choice, as it shapes the learned latent representation.

A2 - Sampling Bias in Gradient Estimation

When trying to estimate an expectation value, but the distribution you sample from is itself dependent on the parameter ϕ you are taking the derivative of, you won't be able to get an unbiased estimate of the real gradient. The reason is that the ϕ dependence adds a bias term consisting of the derivative of the distribution you sample from shown in Eq. (43).

Subsequently, drawing larger and larger amounts of random samples will never estimate the true gradient accurately. The extra term will always add a bias. This means, technically speaking, backpropagation through a random sampling step is possible, but there is no guarantee that sampling a large number of z will help with convergence towards the real gradient [19].

$$\begin{aligned}
\nabla_\phi \, \mathbb{E}_{q_\phi(z|x)}[f_\phi(z)] &= \nabla_\phi \left[\int_z dz \, q_\phi(z \mid x) f_\phi(z) \right] \\
&= \int_z dz \, \nabla_\phi \left[q_\phi(z \mid x) f_\phi(z) \right] \\
&= \int_z dz \, f_\phi(z) \, \nabla_\phi q_\phi(z \mid x) + \int_z dz \, q_\phi(z \mid x) \, \nabla_\phi f_\phi(z) \\
&= \underbrace{\int_z dz \, f_\phi(z) \, \nabla_\phi q_\phi(z \mid x)}_{\text{Bias}} + \mathbb{E}_{q_\phi(z|x)}[\nabla_\phi f_\phi(z)]
\end{aligned} \tag{43}$$

B - Simulated GW signal waveforms

Depicted below are several waveforms coming from the GW signal simulation discussed in section 4. In addition, the effect of the number frequency bins used, on the signal waveform, is shown.

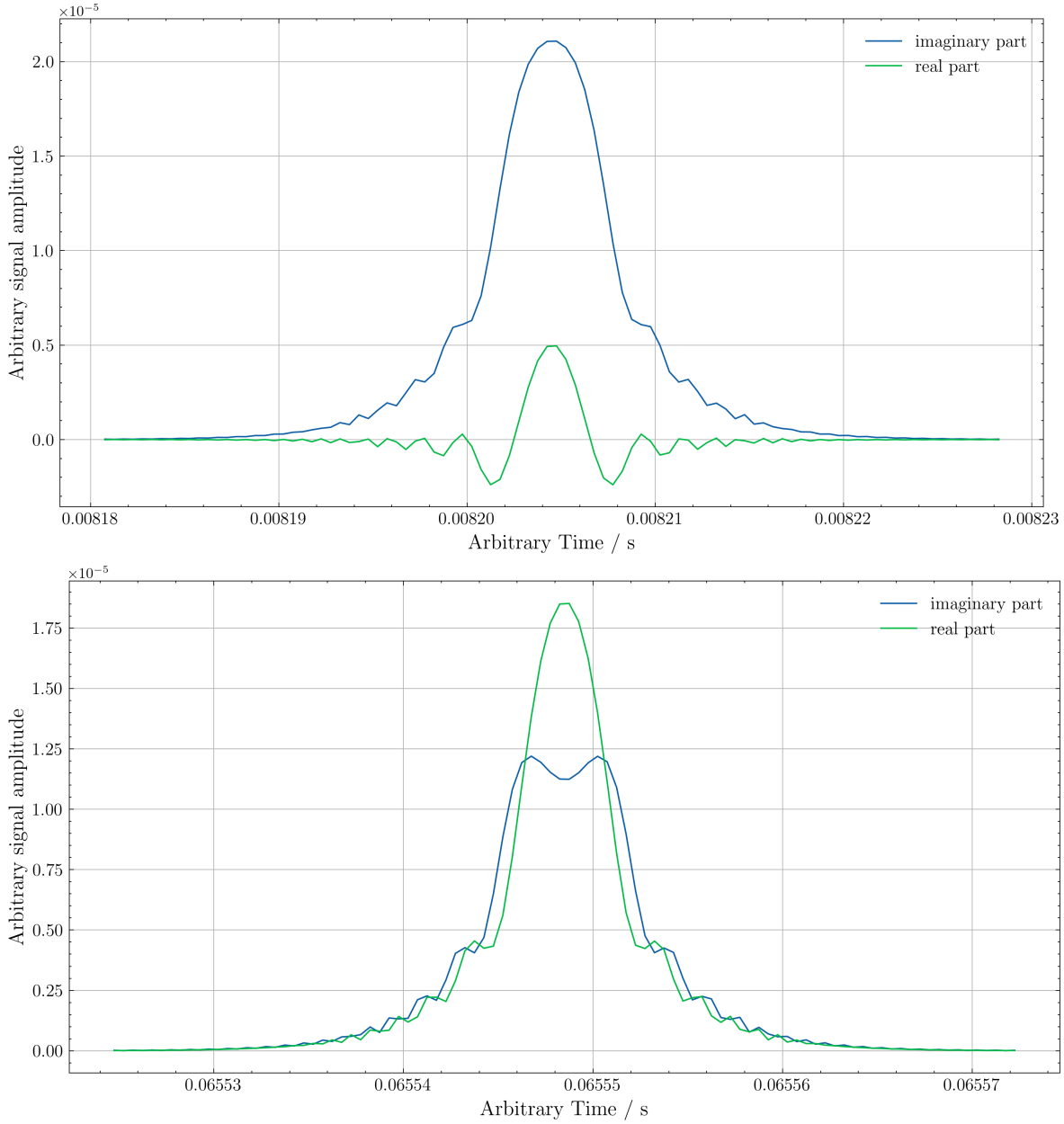

Figure 10: A simulated high-frequency GW signal of two merging PBHs of mass $m_{\mathbf{PBH}} = 1 \times 10^{-11} M_\odot$ each, measured with a microwave cavity. The top panel shows the case for $N = 32768$ frequency bins, while the bottom panel shows the same signal for $N = 262144$. The chirp rate is $k \approx 7.83 \times 10^{10}\ \frac{\text{Hz}}{\text{s}}$ and the effective resonance time $\Delta t \approx 1.27 \times 10^{-6}\,\text{s}$ is shorter than the response time $\tau = 1 \times 10^{-5}\,\text{s}$, but longer than the oscillation period $1/f_0 = 2 \times 10^{-10}\,\text{s}$ set by $f_0 = 5\,\text{GHz}$. As a result, the cavity field exhibits several oscillations in the real (green) and imaginary (blue) parts, though fewer than in figure 4. The absolute time scale on the x-axis is not physically relevant; only the interval between successive zeros defines the effective signal duration.

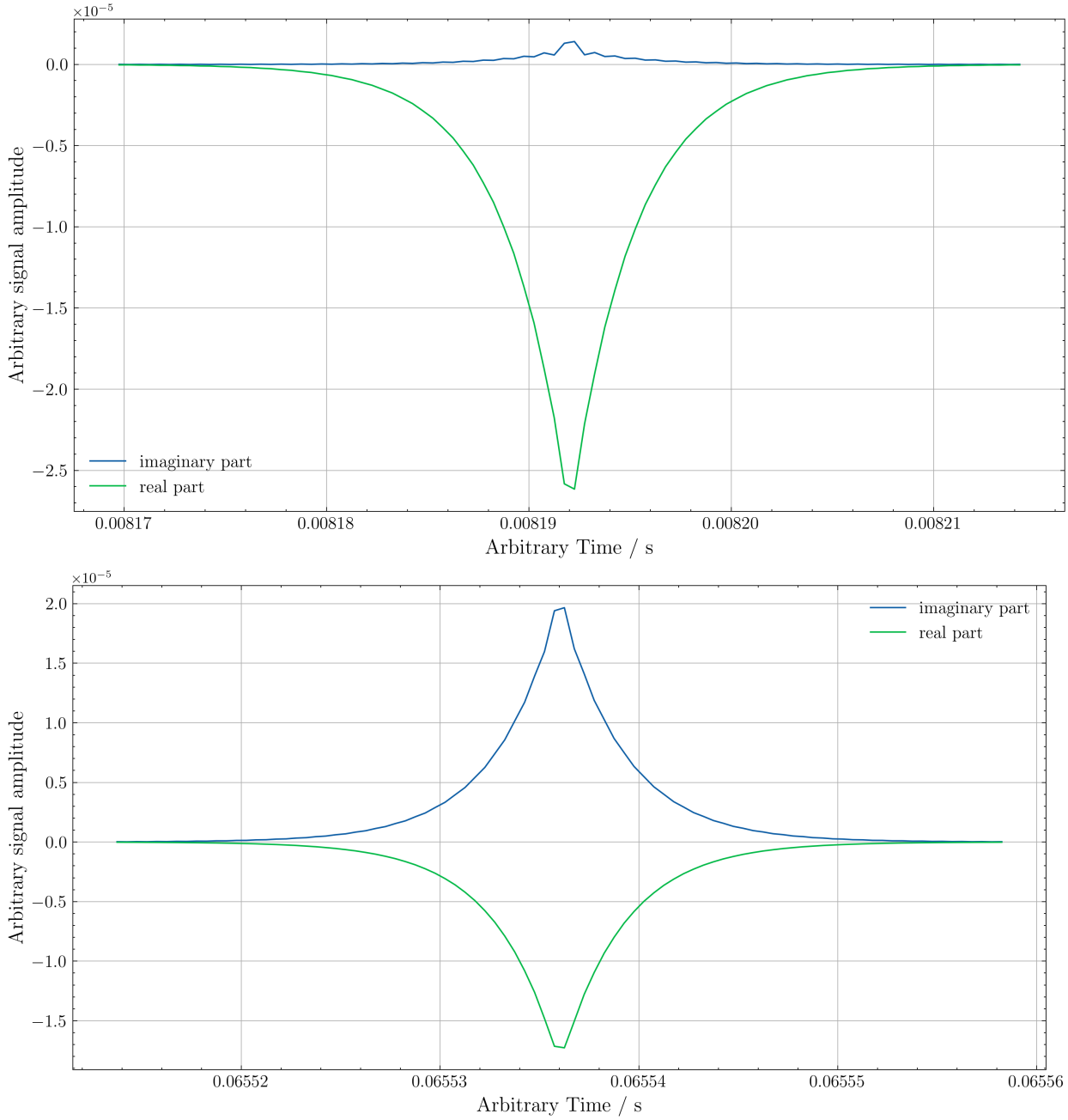

Figure 11: A simulated high-frequency GW signal of two merging PBHs of mass $m_{\mathbf{PBH}} = 1 \times 10^{-10} M_\odot$ each, measured with a microwave cavity. The top panel shows the case for $N = 32768$ frequency bins, while the bottom panel shows the same signal with $N = 262144$. The chirp rate is $k \approx 3.63 \times 10^{12}\,\frac{\mathrm{Hz}}{\mathrm{s}}$ and the effective resonance time is $\Delta t \approx 2.75 \times 10^{-8}\,\mathrm{s}$, shorter than the response time $\tau = 1 \times 10^{-5}\,\mathrm{s}$ but longer than the oscillation period $1/f_0 = 2 \times 10^{-10}\,\mathrm{s}$ set by $f_0 = 5\,\mathrm{GHz}$. As a result, the electric field inside the cavity exhibits only weak oscillatory behavior in the imaginary part (blue).

C - Detection Efficiency curves and Noise Characterization

Depicted are examples of efficiency curves used for the creation of figure 6. In addition, histograms of the decoder output of the window size 128 model 14 and the noise amplitude distribution 15 are shown.

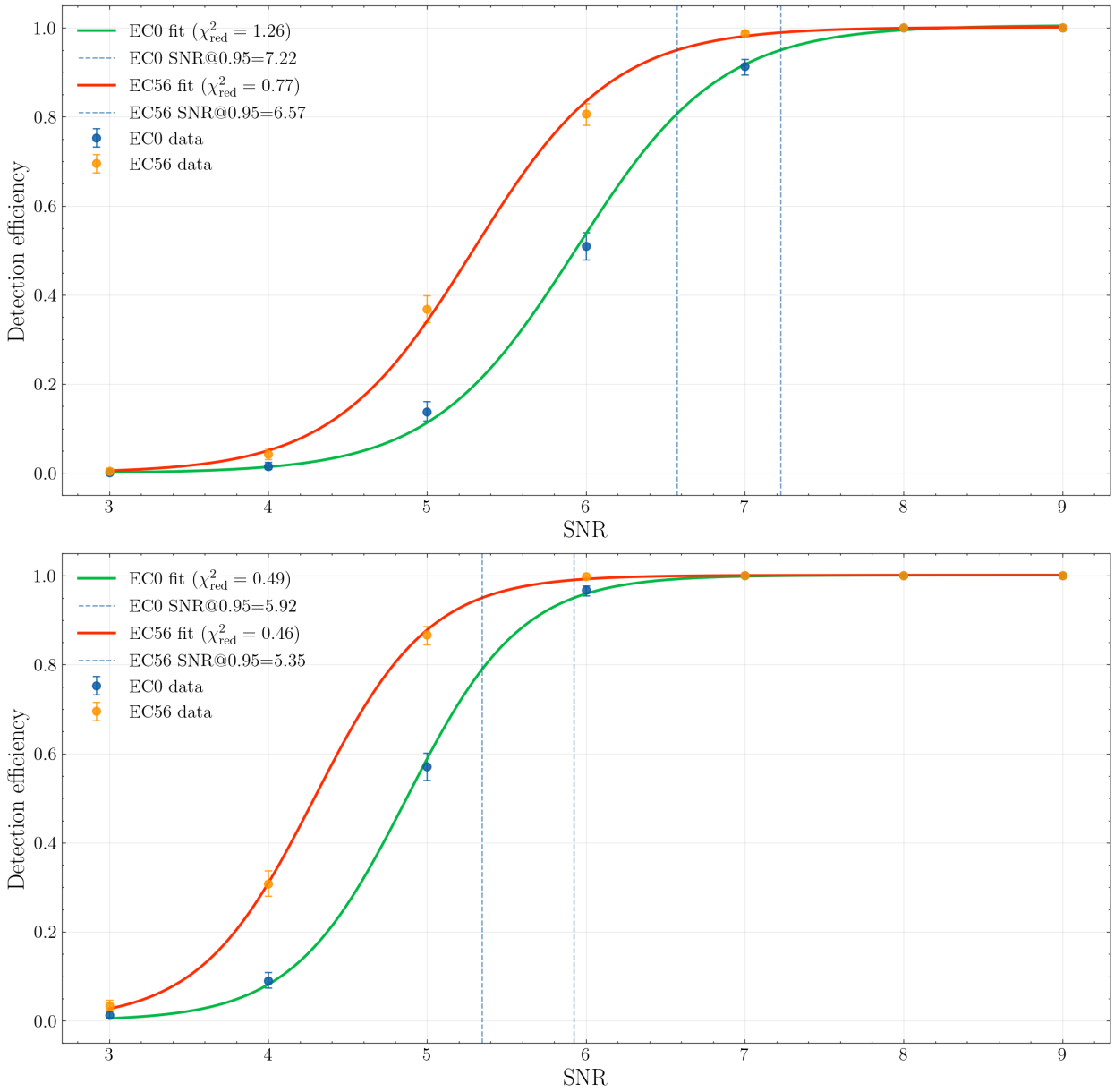

Figure 12: Detection efficiency curves for simulated GW signals. The top panel shows results for $m_{\text{PBH}} = 1 \times 10^{-9} M_\odot$, the bottom for $m_{\text{PBH}} = 1 \times 10^{-11} M_\odot$. Error bars are Wilson score intervals at 95% confidence. Each curve is based on 1.399 million windows, with no false positives observed in the test data set, using a window size 128 model.

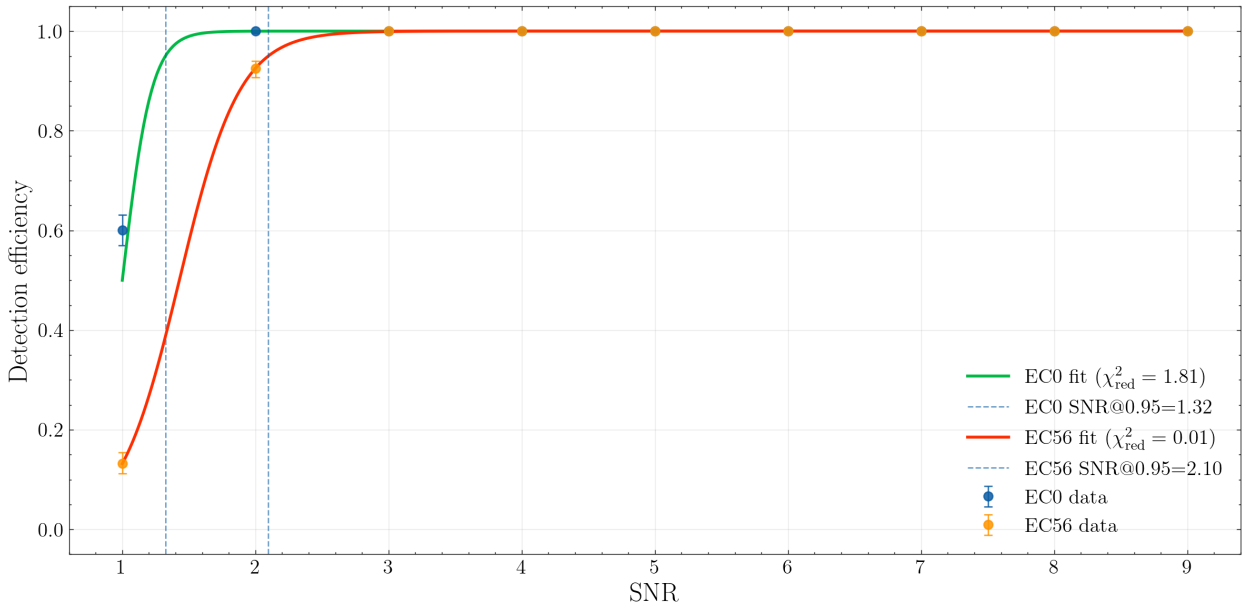

Figure 13: Detection efficiency curve for $m_{\mathbf{PBH}} = 1 \times 10^{-13} M_{\odot}$. Error bars represent Wilson score intervals at 95% confidence. The curve is based on 1.399 million test windows, with no false positives observed, using a window size 64 model.

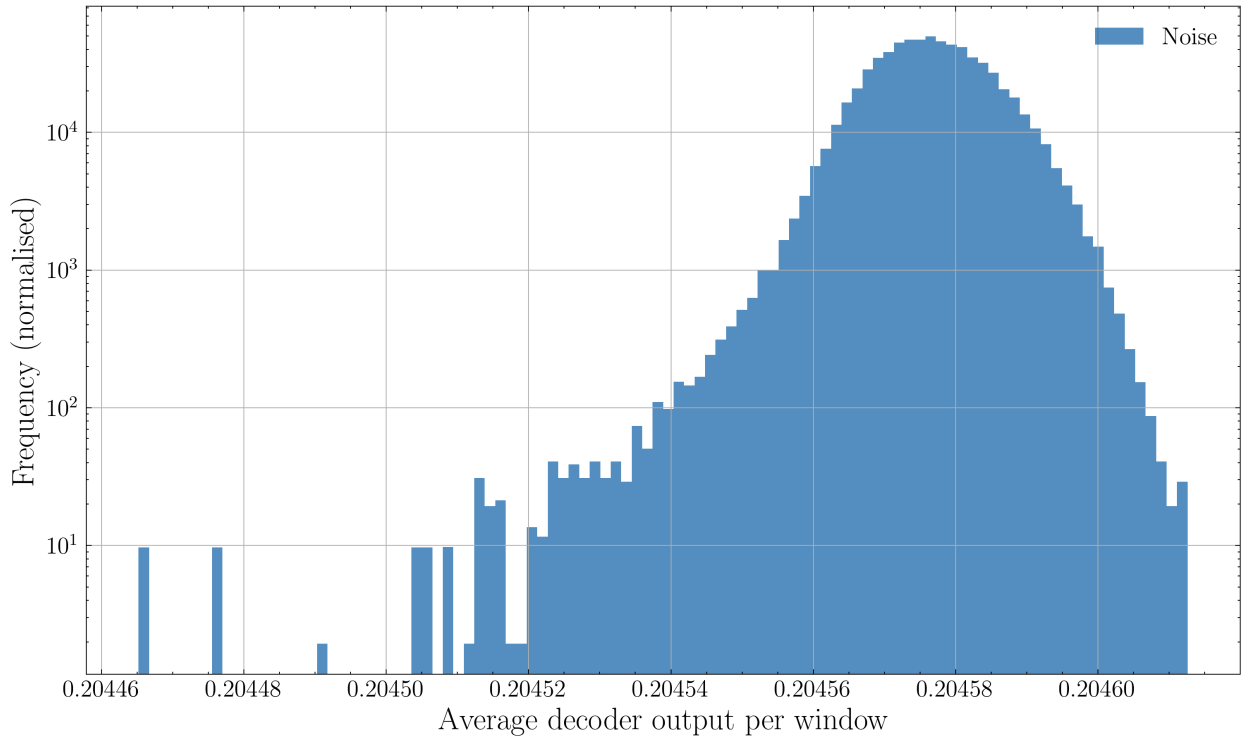

Figure 14: Distribution of the decoder's per-window mean on pure noise test data. Fluctuations occur only at the 10^{-4}–10^{-5} level. The distribution is based on 2.79 million windows with a step size of 8 samples, using a window size 64 model.

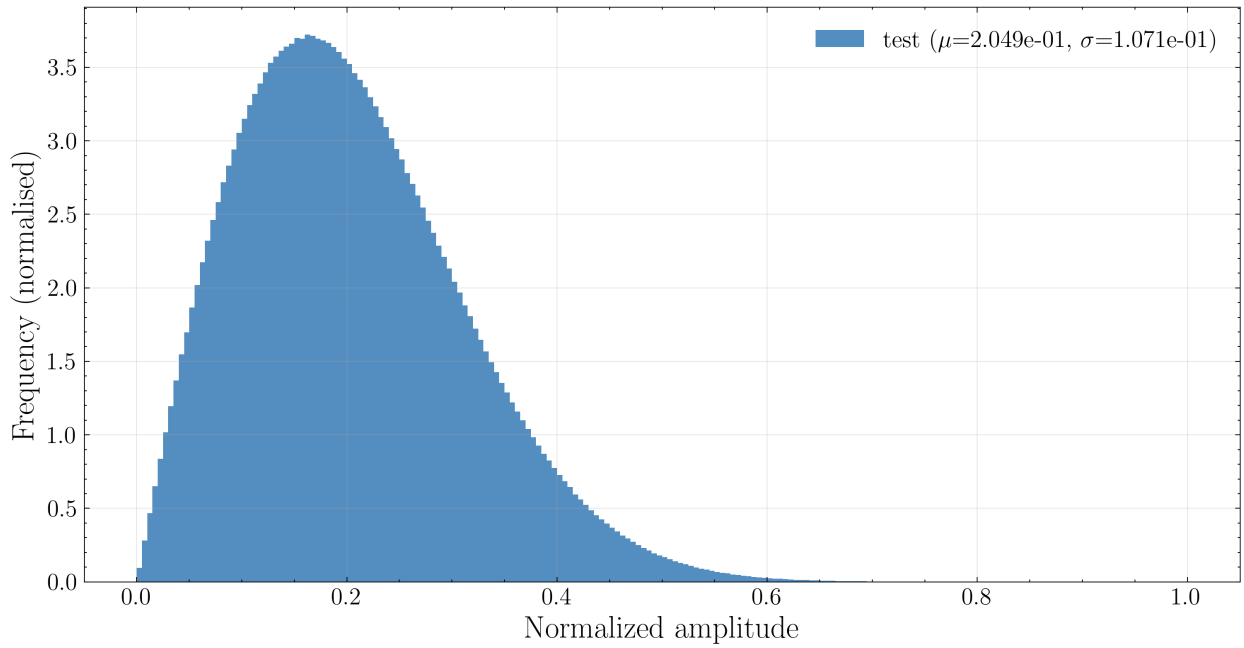

Figure 15: Normalized test-set amplitude distribution with mean $\mu = 2.049 \times 10^{-1}$ and standard deviation $\sigma = 1.071 \times 10^{-1}$). A total of 2.79 million windows with a step size of 8 samples was used.

D - Hyperparameters and specifications used for the CNN based VAE

Comparison of Model Configurations (Window Size = 64 vs. 128)

Common Parameters:

- Normalization of data: min-max
- Training files: `"IQDataFile-2024.04.18.19.22.48.163.tiq"`, `"IQDataFile-2024.04.18.19.22.56.276.tiq"`
- No signals injected during training
- Overlap: 93.33 %
- Step size: $\max(1, \text{round}(\text{window\_size} \cdot (1 - 0.9333)))$, where round means to round the result of $\text{window\_size} \cdot (1 - 0.9333)$
- Data precision: FP32
- Optimizer: Adam
- Dataset split: 0.80/0.10/0.10 (train/val/test)
- Batch size: 512
- Learning rate: 5×10^{-4}
- Latent dimension: 3
- Latent vector extraction: Apply a Conv1D layer with kernel size 1 to produce the latent channels, followed by Global Average Pooling (1D) to reduce each channel to a scalar.
- CNN layers in encoder and decoder: 3
- Number of filters in the first CNN layer: 16
- Filter increase factor after each layer: 2
- Padding mode: `valid`
- Reflect padding: enabled
- Decoder upsampling: Use Keras' `RepeatVector` to expand the latent vector to match the encoder's final tensor shape, then apply a Conv1D (kernel size 1) to increase feature count before mirroring the encoder's CNN layers in reverse.
- Decoder dilations list: [1, 1, 1] (no dilation)
- β (KL weight): 0.01
- Activation function (CNN): `GELU`

Model-Specific Parameters:

- Window size = 64
  - Epochs: 2
  - Kernel size: 9
  - Strides list: [2, 2, 1]
  - Dilations list: [1, 1, 1] (no dilation)
- Window size = 128
  - Epochs: 3
  - Kernel size: 15
  - Strides list: [1, 1, 1]
  - Dilations list: [1, 2, 4]

References

[1] B. P. Abbott et al. "Observation of Gravitational Waves from a Binary Black Hole Merger". In: *Physical Review Letters* 116.6 (2016), p. 061102. DOI: 10.1103/PhysRevLett.116.061102. arXiv: 1602.03837 [gr-qc].

[2] Albert Einstein. "Näherungsweise Integration der Feldgleichungen der Gravitation". In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)* (1916), pp. 688–696. URL: https://adsabs.harvard.edu/pdf/1916SPAW.....688E.

[3] Sean M. Carroll. *Lecture Notes on General Relativity*. Lecture notes covering manifolds, Riemannian geometry, Einstein's equations, gravitational radiation, black holes, and cosmology. 1997. arXiv: gr-qc/9712019 [gr-qc]. URL: https://doi.org/10.48550/arXiv.gr-qc/9712019.

[4] C. J. Moore, R. H. Cole, and C. P. L. Berry. "Gravitational-wave sensitivity curves". In: *Classical and Quantum Gravity* 32.1 (2015), p. 015014. DOI: 10.1088/0264-9381/32/1/015014. URL: http://dx.doi.org/10.1088/0264-9381/32/1/015014.

[5] LIGO Laboratory, a joint Caltech/MIT project. *Sources and Types of Gravitational Waves*. https://www.ligo.caltech.edu/page/gw-sources. Accessed: 2025-09-12.

[6] N. Aggarwal et al. "Challenges and Opportunities of Gravitational Wave Searches at MHz to GHz Frequencies". In: *Living Reviews in Relativity* 24.4 (2021). DOI: 10.1007/s41114-021-00032-5. arXiv: 2011.12414 [gr-qc]. URL: https://doi.org/10.48550/arXiv.2011.12414.

[7] Gabriele Franciolini, Anshuman Maharana, and Francesco Muia. "Hunt for Light Primordial Black Hole Dark Matter with Ultra–High–Frequency Gravitational Waves". In: *Physical Review D* 106.10 (2022), p. 103532. DOI: 10.1103/PhysRevD.106.103532. arXiv: 2205.02153 [astro-ph.CO]. URL: https://doi.org/10.48550/arXiv.2205.02153.

[8] Hojin Yoon et al. "Axion Haloscope Using an 18 T High Temperature Superconducting Magnet". In: *Physical Review D* 106.1 (2022), p. 012005. DOI: 10.1103/PhysRevD.106.092007. arXiv: 2206.12271 [hep-ex]. URL: https://arxiv.org/abs/2206.12271.

[9] Ashu Kushwaha, Sunil Malik, and S. Shankaranarayanan. "Gertsenshtein–Zel'dovich effect: A plausible explanation for fast radio bursts?" In: *arXiv preprint arXiv:2202.00032* (2022). Version accepted in MNRAS. DOI: 10.48550/arXiv.2202.00032. arXiv: 2202.00032 [astro-ph.HE]. URL: https://arxiv.org/abs/2202.00032.

[10] Kristof Schmieden and Matthias Schott. "The Global Network of Cavities to Search for Gravitational Waves (GravNet): A novel scheme to hunt gravitational waves signatures from the early universe". In: *arXiv preprint arXiv:2308.11497* (2023). arXiv: 2308.11497 [gr-qc]. URL: https://arxiv.org/abs/2308.11497.

[11] Asher Berlin et al. "Detecting High-Frequency Gravitational Waves with Microwave Cavities". In: *Phys. Rev. D* 105.11 (2022), p. 116011. DOI: 10.1103/PhysRevD.105.116011. arXiv: 2112.11465 [hep-ph]. URL: https://doi.org/10.48550/arXiv.2112.11465.

[12] Roberto Vio and Paola Andreani. "Everything you always wanted to know about matched filters (but were afraid to ask)". In: *arXiv preprint* (2021), pp. 1–23. arXiv: 2107.09378 [astro-ph.IM]. URL: https://doi.org/10.48550/arXiv.2107.09378.

[13] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *arXiv preprint arXiv:1312.6114* (2013). Version v11, last revised 10 Dec 2022. URL: https://arxiv.org/abs/1312.6114.

[14] Rushikesh Shende. *Autoencoders, Variational Autoencoders (VAE) and β-VAE*. Accessed: 2025-09-01. Apr. 2023. URL: https://medium.com/@rushikesh.shende/autoencoders-variational-autoencoders-vae-and-%CE%B2-vae-ceba9998773d.

[15] Serge Droz et al. "Gravitational waves from inspiraling compact binaries: Validity of the stationary-phase approximation to the Fourier transform". In: *Physical Review D* 59 (1999), p. 124016. DOI: 10.1103/PhysRevD.59.124016. arXiv: gr-qc/9901076 [gr-qc]. URL: https://doi.org/10.48550/arXiv.gr-qc/9901076.

[16] NIST/SEMATECH. *7.2.4.1. Confidence intervals*. Section discusses Wilson/Agresti–Coull intervals for proportions. NIST/SEMATECH e-Handbook of Statistical Methods. URL: https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm (visited on 09/10/2025).

[17] James Orlin. *Prediction: Bias-Variance Tradeoff*. MIT OpenCourseWare, 15.097 Prediction: Machine Learning and Statistics, Spring 2012. Accessed: 2025-09-10. 2012. URL: https://ocw.mit.edu/courses/15-097-prediction-machine-learning-and-statistics-spring-2012/dec694eb34799f6bea2e91b1c0MIT15_097S12_lec04.pdf.

[18] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN: 978-0387310732.

[19] Gregory Gundersen. *The Reparameterization Trick*. https://gregorygundersen.com/blog/2018/04/29/reparameterization/. Accessed: 2025-09-04. Apr. 2018.