manuscript accepted for publication at *JGR: Machine Learning and Computation* on May 6,2024

Cold Diffusion Model for Seismic Denoising

Daniele Trappolini[1,4], Laura Laurenti[1], Giulio Poggiali[2], Elisa Tinti[2,4], Fabio Galasso[5], Alberto Michelini[4], Chris Marone[2,3]

[1]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy
[2]Department of Earth Science, Sapienza University of Rome, Rome, Italy
[3]Department of Geosciences, Pennsylvania State University, University Park, Pennsylvania, USA
[4]Istituto Nazionale di Geofisica e Vulcanologia, Roma, Italy
[5]Department of Computer Science, Sapienza University of Rome, Rome, Italy

Key Points:

- We introduce a model for removing noise from seismograms using a Cold Diffusion (CDiffSD) model.
- Our technique is promising at making sense of earthquake data, even when the magnitude of background noise is almost as large as that of the earthquake signals.
- The CDiffSD model surpasses benchmark results, such as those achieved by DeepDenoiser.

Corresponding author: Daniele Trappolini, dtrappolini@diag.uniroma1.it

Abstract

Seismic waves contain information about the earthquake source, the geologic structure they traverse, and many forms of noise. Separating the noise from the earthquake is a difficult task because optimal parameters for filtering noise typically vary with time and, if chosen inappropriately, may strongly alter the original seismic waveform. Diffusion models based on Deep Learning (DL) have demonstrated remarkable capabilities in restoring images and audio signals. However, those models assume a Gaussian distribution of noise, which is not the case for typical seismic noise. Motivated by the effectiveness of "cold" diffusion models in speech enhancement, medical anomaly detection, and image restoration, we present a cold variant for seismic data restoration. We describe the first Cold Diffusion Model for Seismic Denoising (CDiffSD), including key design aspects, model architecture, and noise handling. Using metrics to quantify the performance of CDiffSD models compared to previous works, we demonstrate that it provides a new standard in performance. CDiffSD significantly improved the Signal to Noise Ratio (SNR) by about 18% compared to previous models. It also enhanced Cross-correlation by 6%, showing a better match between denoised and original signals. Moreover, testing revealed a 50% increase in the recall of P-wave picks for seismic picking. Our work show that CDiffSD outperforms existing benchmarks, further underscoring its effectiveness in seismic data denoising and analysis. Additionally, the versatility of this model suggests its potential applicability across a range of tasks and domains, such as GNSS, Lab Acoustic Emission, and DAS data, offering promising avenues for further utilization.

Plain Language Summary

Seismic waves contain clues about earthquakes and what's beneath the Earth's surface, but any recording of these waves is often mixed with unwanted sounds or disturbances to varying degrees. It's important to filter out these disturbances from the earthquake recordings to improve their clarity and, as a result, make any further analysis more accurate. However, this can be tricky because the nature of these disturbances can change over time, including their amplitude, or by analogy to audio: how loud they are and their pitch of high and low notes. Our work removes noise and thus cleans up recordings to make them more understandable. Recently, advanced computer methods that are good for improving images and sounds have shown promising results. But, these methods usually look for disturbances that follow a certain pattern, which does not always work for more complex disturbances found in earthquake data. To address this, we introduce a strategy called the Cold Diffusion Model for Seismic Denoising (CDiffSD). This strategy is tailor-made to deal with the specific kinds of disturbances found in earthquake data, and it does a better job than previous methods at removing noise and making the earthquake recordings clear again, providing a new standard in this area of study.

1 Introduction

Seismograms contain signals generated by earthquakes and by other unidentified sources categorized in general as 'noise' (e.g., oceanic waves, wind, vehicular traffic, sonic booms, quarry activities, and instrument malfunctions.). It is standard practice in seismology to denoise waveforms to improve the performance of the subsequent analyses, such as P- and S-wave onset picking, earthquake source moment tensor inversion, and techniques of exploration seismology. Most commonly and in routine analysis, denoising is performed through bandpass filtering. However recent works have proposed several more sophisticated schemes to "clean" seismic traces. These include methods based on the independent component analysis (ICA) (Comon, 1994; Cabras et al., 2010; Moni et al., 2012), beamforming methods (Gibbons et al., 2008; Boué et al., 2013; Brooks et al., 2009), and MUltiple SIgnal Classification(MUSIC) (Schmidt, 1986; Bear et al., 1999).

All of these methods, however, can fall short when the noise shares frequencies with the earthquake generated signal.

Denoising models have evolved to incorporate time-frequency methods, with techniques like the Wavelet transform (Gaci, 2014; Siyuan & Xiangpeng, 2005; W. Liu et al., 2016; Zhang & Ulrych, 2003; S. Cao & Chen, 2005; Mousavi & Langston, 2017), the Short-Time Fourier Transform (STFT) (Mousavi & Langston, 2016), the S-transform (Tselentis et al., 2012), and other transformation-decomposition methods (Hennenfent & Herrmann, 2006; Bekara & der Baan, 2009; Neelamani et al., 2008; Han & van der Baan, 2015; Y. Liu et al., 2013; Chen & Ma, 2014; Shan et al., 2009; Tang & Ma, 2011). These techniques have proven useful but the emergence of deep learning (DL) has provided new strategies with improved performance. A notable development in this area is the Deep Denoiser (DD) model (Zhu et al., 2019). The DD approach is based on a UNet architecture, which generates dual masks for seismic and noise signals, enhancing waveform extraction. Another notable approach is that of van den Ende et al. (2021) who employed DL to denoise Fiber-optic Distributed Acoustic Sensing (DAS) data. They demonstrate the potency of DL to enhance the quality of DAS and seismic data. Similarly, the Novoselov et al. (2022) project, utilizing a Dual-Path Recurrent Neural Network(DPRNN), led to another substantial stride in the application of deep learning for seismic signal denoising. These studies not only validate the efficacy of deep learning methods in seismic noise reduction but also pave the way for further innovations in this field.

Here we built on this topic, drawing parallels with techniques used in speech enhancement, a field closely related to seismic denoising. Speech enhancement has recently seen the use of models such as GANs (Pascual et al., 2017; Donahue et al., 2018; R. Cao et al., 2022; Kim et al., 2021) and VAEs (Fang et al., 2021; Leglaive et al., 2020, 2018; Bie et al., 2022). However, the recent trend points to the growing success of Diffusion Models (Sohl-Dickstein et al., 2015; Ho et al., 2020), which are now outperforming their predecessors (GAN & VAE) see (Lu et al., 2022; Richter et al., 2023). Using techniques like cold diffusion or Gaussian diffusion for denoising presents several advantages over approaches that use binary masks, especially in terms of flexibility, reconstruction quality, and the ability to handle complex noise; while binary generally retain advantages in terms of simplicity, speed, interpretability, and computational efficiency. Here, we investigate the application of diffusion models for seismic denoising. These models typically transform the input into an isotropic Gaussian distribution through the consistent addition of Gaussian noise. In the reverse process, diffusion probabilistic models aim to remove the anticipated noise from the corrupted input, thus recovering the original signal. A pioneering approach to seismic denoising using diffusion models with Gaussian noise was introduced by (Durall et al., 2023), specifically applied on shot gathers used for seismic imaging and exploration.

This challenge led us to explore the emerging Cold Diffusion model (Bansal et al., 2022; Yen et al., 2023), which adapts the diffusion process by replacing Gaussian noise with other types of noise and signal degradation processes. The Cold Diffusion model demonstrates how diffusion models can effectively restore signals impaired by various types of degradation. Its inherent properties make it particularly suitable for tasks such as speech source separation in practical settings with non-Gaussian noise. Building on this, our research aims to adapt the cold diffusion paradigm for seismic trace denoising. This adaptation involves specific modifications, primarily in the sampling algorithm, to suit the unique challenges of seismic data. The result is a Cold Diffusion Model for Seismic Denoising (CDiffSD).

Here we list the key points and novel aspects of our model:

- First to Utilize Cold Diffusion with Seismic Noise: This research pioneers the application of the Cold Diffusion model, adapting it to handle noise directly recorded from seismic stations.

- Promising Technique to facilitate downstream tasks: CDiffSD shows promising results in impoving downstream tasks, such as phase picking, even in scenarios where background noise levels are nearly as high as the earthquake signals themselves
- Thorough validation on reference benchmarks: The CDiffSD model not only introduces a new methodology but also demonstrates its capability to surpass existing strong baselines such as DeepDenoiser, a commonly-used benchmark which is a reference for denoising.
- Adaptation to Non-Gaussian Noise: Recognizing the limitations of traditional diffusion models that assume a Gaussian distribution of noise—which is often not the case in seismic applications—the paper introduces a "cold" variant of diffusion models. This adaptation is specifically tailored to restore clean and noisy seismic traces more effectively.

2 Methods

The model we propose is based on a generalization of diffusion models, termed the Cold Diffusion model. In this section we introduce the model. Additional details are found in Appendix A (Diffusion Model) and Appendix B (Cold Diffusion Model).

2.1 Proposed Method: Cold Diffusion Seismic Denoising (CDiffSD)

Problem formulation: The core of our CDiffSD (Cold Diffusion Model for Seismic Denoising) involves degrading a one-dimensional earthquake, in the form of a seismic record, x_0 (the target), with recorded seismic noise x_n, to produce x_T (noisy signal):

$$x_T = x_0 + x_n * NRF \tag{1}$$

Here, x_0 represents an earthquake recorded by a seismometer, while x_0 serves as a 'clean' sample in our context, it's important to note that it inherently contains some level of noise, given its real (non-synthetic) origin. Since we deal with normalized earthqaukes and noises in the range of $[-1, 1]$ (for more details on the normalization process see 2.1.1); the Noise Reduce Factor (NRF) is a key element in our specific analysis. It's responsible for calibrating the amplitude of the noise signal (x_n) in relation to the earthquake signal's amplitude, often indicated by the amplitude of S-waves in the data. By choosing a NRF value within the range 0.4 to 0.65, we ensure that the noise does not dominate the trace compared to the earthquake. We work with data from different stations that independently record noise and earthquake signals. It's worth mentioning that we mix earthquake x_0 and noise x_n recorded from different seismic stations, to improve generalizability and robustness.

Training: Regarding the specific operation of cold diffusion models, our approach is delineated using the improved training algorithm proposed by (Yen et al., 2023):

Concerning the forward diffusion process degradation see Appendix.A Diffusion Model, we can rephrase the degradation at time t as follows:

$$x_t = D_{x_T}(x_0, t) = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} x_T \tag{2}$$

where x_0 is the recorded earthquake, $x_t = x_0 + x_n * NRF$ and $\alpha \in [0, 1]$ is the parameter interpolation weight. α can also be regarded as the amount of information retained in the diffusion process, and it can alternatively be defined as $1 - \beta$, where beta represents the amount of noise introduced in the degradation, such parameters are defined a priori by a scheduler: $\{\beta_t \in (0, 1)_{t=1}^{T}\}$.

Algorithm 1 Cold Diffusion Enhanced Training

for $n = 1, ..., N_{iter}$ **do**
  Sample clean data x_0
  Sample $t \sim$ Uniform($\{1, ..., T\}$)
  $x_t \leftarrow D(x_0, t), \hat{x}_0 \leftarrow R_\theta(x_t, t)$
  Sample $t^{'} \sim$ Uniform($\{1, ..., t\}$)
  $\hat{x}_{t'} \leftarrow D(\hat{x}_0, t^{'}), \hat{\hat{x}}_0 \leftarrow R_\theta(\hat{x}_{t'}, t^{'})$
  Take gradient descent step on $\nabla_\theta(||\hat{x}_0 - x_0||_1 + ||\hat{\hat{x}}_0 - x_0||_1)$
end for

t: Represents a chosen random timestep within the predefined range $[1, T]$

t': Signifies an earlier timestep than t, facilitating a recursive learning process where the model iterates through noise addition and removal at progressively earlier moments.

x_t: Result of the degradation applied to the signal x_0 following the forward process as a function of t

$D(x_0, t)$: Refers to the degradation applied to the signal x_0 following the forward process as a function of t

$R_\theta(x_t, t)$: Refers to the reconstruction applied to the signal x_t and it results in producing: $\hat{x}_0$

$\hat{x}_0$: the result of the Restoration: $R_\theta(x_t, t)$

$\hat{x}_{t'}$: Result of the degradation applied to the signal $\hat{x}_0$ following the forward process as a function of t'

$\hat{\hat{x}}_0$: the result of the Restoration: $R_\theta(\hat{x}_{t'}, t^{'})$

This method improves the model's ability to learn during training, especially when dealing with unusual, non-Gaussian noise. In the training stage, the model picks a random timestep t from the range [1,T]. At this time, noise is added to the signal, followed by a cleaning step. This is important because it teaches the model to remove noise, mimicking the denoising process. The model's learning is deepened by repeating these steps at an earlier moment t', where $t' < t$. Here, the model works not with the original earthquake data but with the signal that was cleaned in the previous step. This signal is made noisy again up to the new time t' and cleaned once more. By doing this over different times, the model learns more effectively, getting better at handling the complex types of noise found in real data. The training approach is designed to be forgiving of mistakes that can happen when choosing moments to sample from. As outlined in Algorithm 1, the training includes using $\hat{x}_{t'}$, the cleaned signal. This introduces a way to deal with potential alignment mistakes that might happen during sampling.

Sampling: Regarding the sampling, our model employ the same approach as used in "Cold diffusion for speech enhancement". Specifically, we take the sampling algorithm 4 in Appendix.B and substitute how x_{t-1} is calculated. In detail, our new x_{t-1} becomes:

$$x_{t-1} \leftarrow \sqrt{\alpha_{t-1}}\hat{x}_0 + \frac{\sqrt{1-\alpha_{t-1}}}{\sqrt{1-\alpha_t}}(x_t - \sqrt{\alpha_t}\hat{x}_0) \tag{3}$$

2.1.1 Input Assumptions

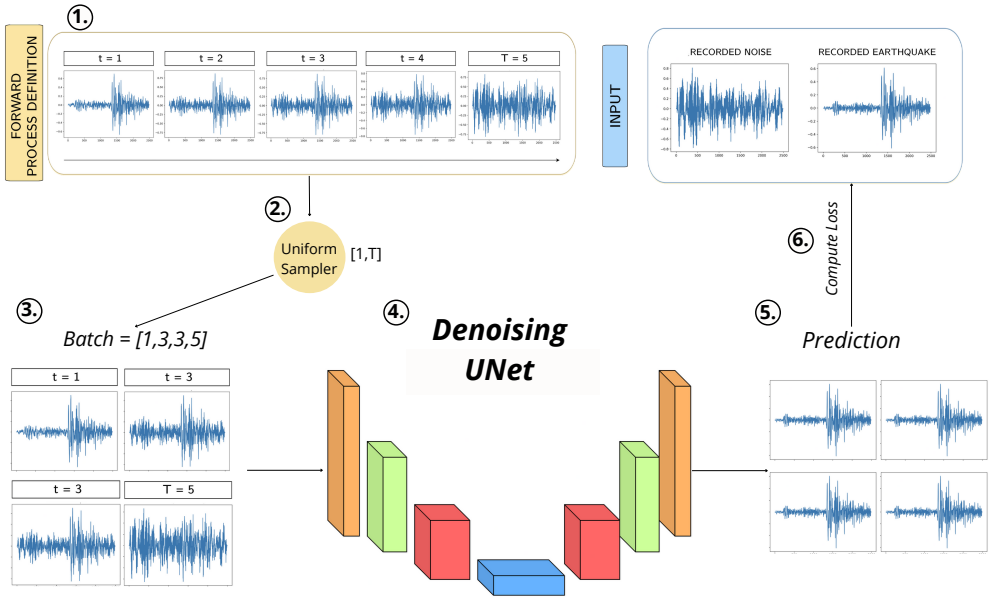

Figure 1. The figure represents the overall framework of our model. Specifically, it shows the handling of the input throughout the entire process: **1.Forward process definition** in the yellow box at the top left, we see an example of forward diffusion with T=5 (T = 5 for graphical reasons). **2.Uniform Sampler** involves uniform sampling [1,T]. **3.Batch** At the bottom left, we find an example of a drawn batch (batch size = 4). This batch provides us with different levels of noise at the extracted time T. The batch, noised according to the rules of the previously defined forward process, is passed to **4.Denoising UNet** model, which returns **5.Predicion**. This prediction is then **6.Compute Loss** compared with the original input, resulting in the calculation of the loss.

In our seismic denoising approach, we separately normalize the noise and earthquake data. We adopt a trace-specific method, normalizing each seismic trace (earthquake and noise) across its East-West (E), North-South (N), and Vertical (Z) channels. This normalization process aligns the maximum and minimum values within these channels, standardizing the data to a range of [-1,1]. Such an approach ensures that each component retains its relative amplitude, enabling precise and balanced analysis. This also enhances generalizability for each type of seismic trace that the end user wants to denoise.

In the training phase for each seismic trace, we begin by merging a normalized earthquake trace with a normalized noise trace. The noise component is scaled using the NRF, adjusting its intensity in the noisy signal before the forward process is applied.

The creation of the 'noisy signal' x_T, a combination of the earthquake and scaled noise signals, leads to the 'forward process' see Fig. 1. Here, a stochastic variable t, ranging from 0 to a predetermined maximum T, is chosen for further noise modulation. At $t = 0$, we have a recorded earthquake signal with no additional noise, whereas at $t = T$, the noise is at its full scale.

2.1.2 Model Configuration

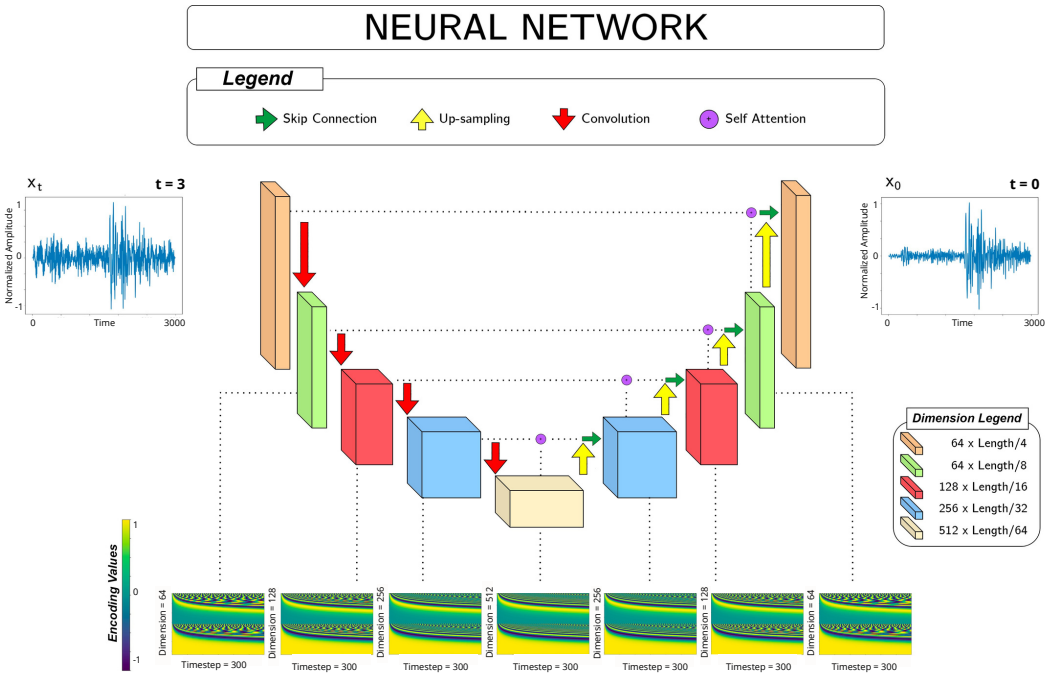

Figure 2. CDiffSD combines convolutional layers, ResNet blocks, attention mechanisms, and positional encoding (with positional encoding detailed in the bottom plots) for effective one-dimensional data processing.

In this subsection, we detail the model described in step 4, the Denoising UNet model, of Fig. 1. As a building block of the diffusion model, we adopt a neural network model inspired by the 1D U-Net (Ronneberger et al., 2015) design, for the processing of one-dimensional data streams such as time series or audio signals. The network begins with 1D convolutional layers, each equipped with 64 filters of a kernel dimension of 7, instrumental for the initial extraction of salient features. This is followed by the integration of temporal processing units that leverage sinusoidal positional encoding to effectively capture the temporal intricacies inherent within the data. These units then employ two linear layers dedicated to feature refinement and are paired with the Gaussian Error Linear Unit (GELU) (Hendrycks & Gimpel, 2016) activation function to instill the requisite non-linearity. As the architecture progresses, it introduces dimensionality manipulation layers consisting of ResNet modules (He et al., 2016) pivotal for feature conservation during downsampling and 1D convolutional layers for further data refinement. Post-downsampling, a series of upsampling layers are implemented, designed to elevate data dimensionality by merging ResNet blocks with dedicated upsampling operations. A noteworthy feature of our design is the mid-level blocks, each outfitted with dual residual units. They exploit attention mechanisms important for highlighting pertinent data characteristics. The network culminates with terminal residual blocks that are succeeded by 1D convolutional layers, making definitive outputs typically manifest as singular channels. The U-Net block is applied for each iteration of the diffusion model from each t_i to 0 and then again from t_{i-1} to 0 and so on until the end of the process.

We trained models with 3 configurations: $T = 20, T = 100, T = 300$. These diverse scheduler assumptions allowed us to evaluate how performance metrics vary with

increasing T, highlighting the trade-off between model performance and computation time, which is a crucial consideration in seismic monitoring room operations where balancing processing speed and precision is essential.

Particularly in the inference phase, understanding the impact of T on both model performance and computational efficiency is vital. For applications requiring rapid trace processing, like real-time seismic monitoring, a preference for speed may be necessary, though it could impact precision. Conversely, in tasks where accuracy is the priority, such as dataset cleaning, a greater emphasis on precision may be warranted, even at the expense of longer processing times.

We compared our approach using the same seismic dataset with DD, that we consider as benchmark. For this task, DD underwent comprehensive training for 400 epochs, while our model completed its training in just 150 epochs. This difference was due to our model's learning dynamics and efficiency. We initiated our model's training with a learning rate of 1e-3 and employed a scheduler to reduce this rate gradually, ensuring controlled and stable convergence.

2.1.3 Inference with Direct and Sampling Reconstruction

Cold diffusion models involve distinct methods to reconstruct the signal including the adoption of direct or sampling reconstruction. These methods represent approaches within the framework of diffusion models, each with unique operational mechanisms and implications for model performance. Understanding the nuances of these methods is vital for comprehending the overall efficacy and application potential of diffusion models.

For the range of configurations used in training our models ($T = [20, 100, 300]$), we applied these configurations to both direct and sampling reconstruction. In the context of diffusion models, the distinction between 'direct' and 'sampling' approaches is pronounced, marked by their differing operational mechanisms.

The **'direct'** method involves applying the reverse process using the U-Net architecture to transition from a specific timestep t_n directly to zero. Conversely, the **'sampling'** method incrementally applies this reverse transition from a specific timestep t_n to zero, but crucially, it traverses through all intermediate timesteps t_i, where $i \in [n-1, 0]$. This results in applying the U-Net architecture multiple times (n).

A key aspect of the cold diffusion paradigm is evaluating the effectiveness of the sampling procedure, which is hypothesized to outperform the direct approach. If the direct method, particularly using U-Net alone, yields comparable results, it would call into question the necessity of the complex training infrastructure typically associated with diffusion models. We provide a detailed comparison between the direct and sampling methods in section 4.

2.1.4 Metrics

For enhanced clarity, we define here the metrics used in our study now and then in Section 4 we provide a detailed commentary on the results.

1. **Signal to Noise Ratio (SNR)** is a measure used to compare the level of a signal (earthquake in this case) to the level of background noise. A higher SNR indicates that the seismic signal stands out clearly from the background noise, facilitating accurate analysis and interpretation. We defined SNR as in (Zhu et al., 2019):

$$10 \log_{10} \frac{\sigma_{signal}}{\sigma_{noise}}.$$

where σ_{noise} and σ_{signal} are the standard deviation of waveforms before and after the P arrival, respectively.

2. **Cross-correlation** is a widely used measure of similarity between two signals. We compute the zero-lag cross-correlation (CC) between the recorded earthquake signals (before noise is added) that represents our ground truth x_0 and the denoised ones to evaluate the performance of the different models in reconstructing the recorded waveform.
3. To evaluate the **picking** performances of the proposed method, we applied the deep learning phase picker PhaseNet (Zhu & Beroza, 2019) to the waveforms and compared the retrieved arrival times with the labeled picked phases of the catalog ($\sim 70\%$ of manually picked and $\sim 30\%$ of automatic picked). In this way we can assess the impact of the denoiser on P and S arrival determination, the accuracy of which enables the calculation of a well constrained location.
We evaluated picking performance by analyzing the distribution of time differences between picks identified by PhaseNet on the denoised traces and the labeled picks within the STEAD dataset. Additionally we employed a "recall" metric that is calculated as the number of picks falling within $\pm$ 50 samples of the nearest labeled pick, divided by the total number of labeled P or S arrivals.

3 Data Sources and Selection

In our study, we focus on a subset extracted from the STanford EArthquake Dataset (STEAD) (Mousavi et al., 2019). This section is dedicated to elucidating the composition of the subset, detailing the following components:

1. We selected specific seismic stations to gather earthquakes and others for noise, with some overlap, providing a clear trace of the data's origin for our analysis (Figure 4).
2. The distribution of seismic events across the globe (Figure 4) is mapped out, with these events sorted into training, validation, and test sets. This classification helps us to assess the model's effectiveness and its generalizability across different regions.
3. We applied constraints to the dataset, including the magnitude and proximity to the seismic stations.

STEAD features a significantly larger number of stations for earthquake data compared to those used for noise. Moreover, the majority of these stations are concentrated within the U.S. territory. In our study, we utilize a ratio of (1786/2613) stations for the extraction of earthquake data, representing a fraction of the total available. For seismic noise, we have selected a subset corresponding to 306 stations dedicated to noise recording.

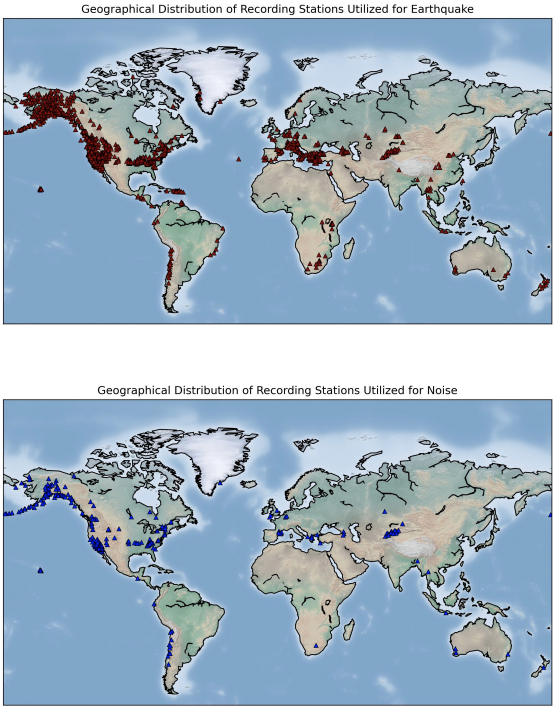

Figure 3. The maps show the subset of stations of the STEAD (STanford EArthquake Dataset) used for the recorded earthquake signal (upper) and the recorded noise (bottom).

Throughout our analysis, we consistently sample seismic traces of 30-second durations, based on the following criteria:magnitude > 2,earthquake-station distance < 100

km, and P-wave arrival after 7 seconds. Figure 5 shows the frequency-magnitude statistics for our data set.

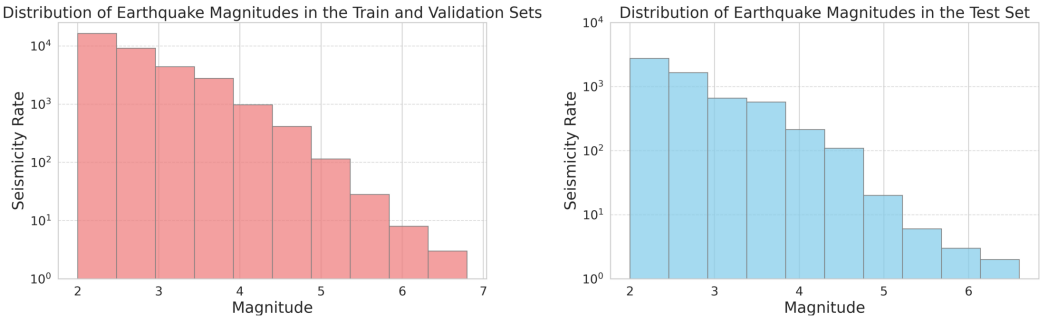

Figure 4. The histograms illustrate the frequency distribution of earthquake magnitudes within our dataset, with the left panel representing the training and validation sets and the right panel the test set.

We chose an inclusive approach for training, leveraging the full spectrum of available data, without any SNR selection criteria. While this might seem disadvantageous initially, a model that performs well under these conditions can be versatile across various scenarios. For researchers looking to retrain this model on their datasets, especially when specific datasets are limited, it may be advantageous not to put restrictive filters such as SNR.

Our dataset was divided into training (30491 traces), validation (3441 traces), and test (5994 traces) as illustrated in Figure 6. Such a division in machine learning ensures model reliability and generalizability. The training set aids the model's primary learning, the validation set is used for hyperparameter adjustments, and the test set objectively evaluates the model's performance on unseen data.

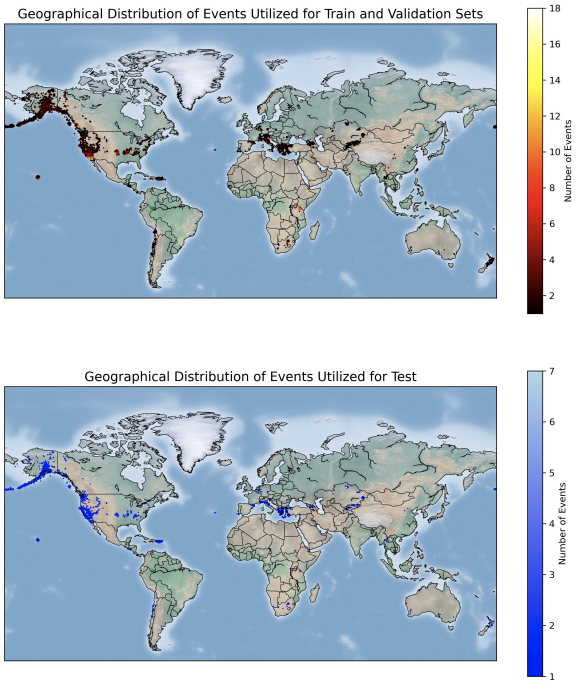

Figure 5. The image presents two maps of the geographical distribution of seismic events used in our study, with the upper map illustrating the events for the training and validation sets marked in red, and the bottom map showing the events for the test set in blue. The color intensity on each map corresponds to the number of events, with darker shades indicating a lower concentration of events in that location.

For more details on the specific train, validation, and test configurations, please refer to our GitHub repository (Trappolini, 2024a). Additionally, the dataset used in this study is available on Zenodo at (Trappolini, 2024b), which includes all necessary data and configurations.

4 Results

In the following we present our results and discuss the validity of our model by adopting quantitative and qualitative catagories. The metrics used for each are provided in Section 2.1.4.

4.1 Quantitative Results

4.1.1 Signal to Noise Ratio (SNR)

A comparison of the SNR metric for the denoised waveforms obtained with different models and configurations is shown in Fig. 6. Note that Figure 7 includes the same metric for the original earthquake signals (labeled "earthquake") and those with added noise (labeled "eqk + noise"). The latter are the inputs to the denoiser algorithm. The performances of the different models appear aligned, with DD differing by a slightly lower median but greater variability in output SNR. In Fig. 7 we classified the noisy observations as a function of the SNR before denoising to highlight the effectiveness of our models in cleaning the seismic traces. The performance of our CDiffSD are consistently superior with respect to DD in low SNR scenarios. This aspect is crucial, given that low SNR conditions correspond to more complex and heavily noisy seismic traces precisely where an effective denoising solution is most needed. The high-quality performance of our model in these low SNR environments is demonstrated in Fig. 7. We note in particular model reliability and efficacy in extracting correct signals from noisy data. This proficiency is important in real-world seismological applications, especially for discovering lower magnitude earthquakes often hidden in the noise.

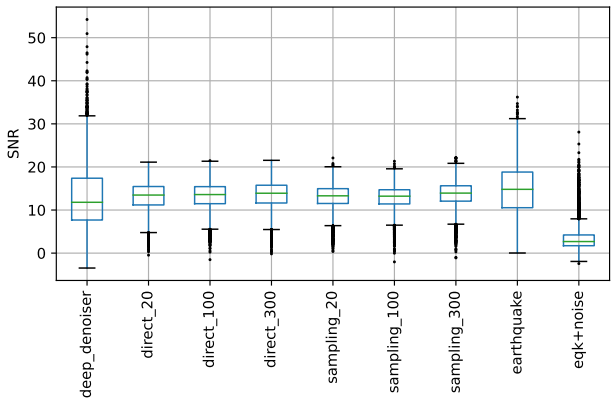

Figure 6. SNR comparisons using box-plots for various model configurations applied to the test set. The original signals (earthquake) and the ones with added noise (eqk+noise) have respectively the higher and lower SNR, as expected. The different denoising models appear overall aligned, with direct and sampling showing slightly higher median values and tighter distributions with respect to DD.

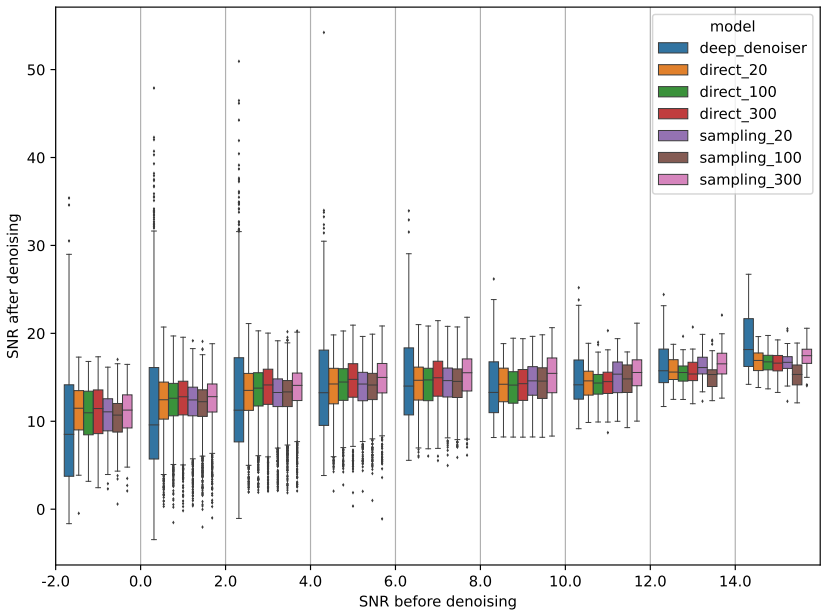

Figure 7. Distributions of SNR values of denoised waveforms for different ranges of input SNR. The SNR statistics after denoising are computed on 2dB wide ranges of input SNR. CDiffSD models show higher performances in low SNR scenarios, while DD is superior for the higher SNR signals. We study the range: $SNR < -2.0$ and $SNR > 16.0$, which covers 99% of data. Solid bars within each model (color) show the median value.

While the cold diffusion approach excels in low SNR scenarios, the binary mask-based method DD exhibits greater variability and tends to perform better in higher SNR conditions, benefiting from its ability to provide a clear-cut signal delineation (Fig. 6 and Fig. 7). In particular, DD shows improved performance when the input SNR is higher than $\sim$ 14 and is get worse at lower input SNR while our models remain consistently effective for a large range of input SNR. An example of high input SNR conditions can be found in the Supporting Information.

4.1.2 Cross Correlation

We evaluate the similarity between original signals and denoised signals, by showing the statistics of the computed CC values, in Fig. 8. A higher CC indicates a greater similarity between the denoised trace and the original signal. In this figure, we see that all CDiffSD models show similar performance and they are all consistently higher than DD. To better highlight the variability of CC values obtained from the different traces of the test set, in Fig. 9 we show the distribution of CC values between denoised and original traces as a function of CC of the noisy traces with original signals (x axis), that is, CC of traces before denoising is applied. The performances for both direct and sampling are higher than DD for every considered range of CC before denoising.

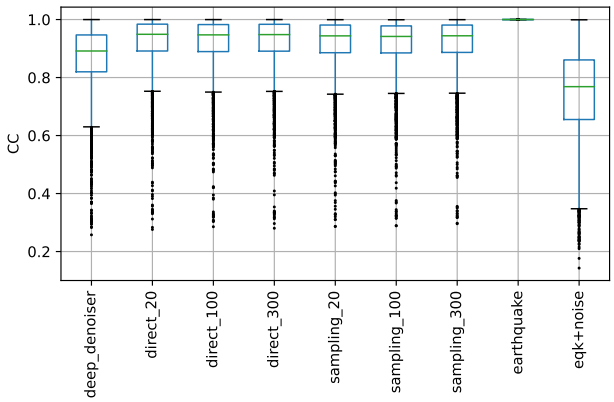

Figure 8. Cross-correlation (CC) comparisons for various model configurations applied to the test set. Higher CC values indicate greater similarity between the denoised trace and the original signal. All CDiffSD models show similar performance and that they are consistently higher than DD.

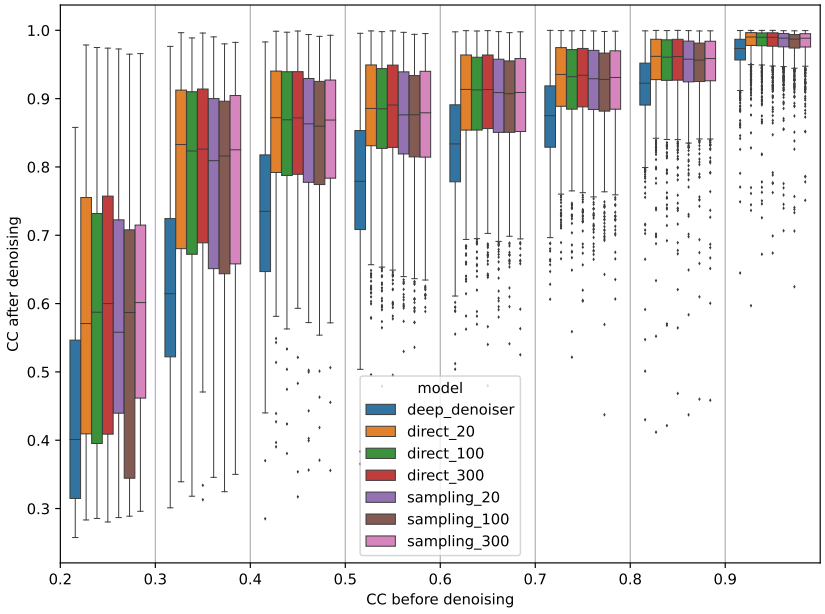

Figure 9. Distribution of CC values between original and denoised signals (y axis) as a function of CC before application of denoising (x axis). CDiffSD models outperform DD across all CC ranges. The difference is more noticeable especially at low pre-denoising CC values. Statistics are computed for ranges of 0.1. Note that the distribution of samples for 'CC before denoising' is identical to 'eqk+noise' in Fig. 8. Consequently, the 0.1-wide bins may encompass significantly varying numbers of samples.

For each model considered we see better performance, with higher values of CC after denoising (Fig. 9). Another noteworthy aspect is that at higher noise levels, thus lower

CC before denoising (values from 0.2 to 0.3), models with T = 300 outperform their counterparts. As expected, these performance disparities tend to converge with an increase in CC before denoising, corresponding to a relative reduction in noise compared to the signal.

4.1.3 Phase arrival picks

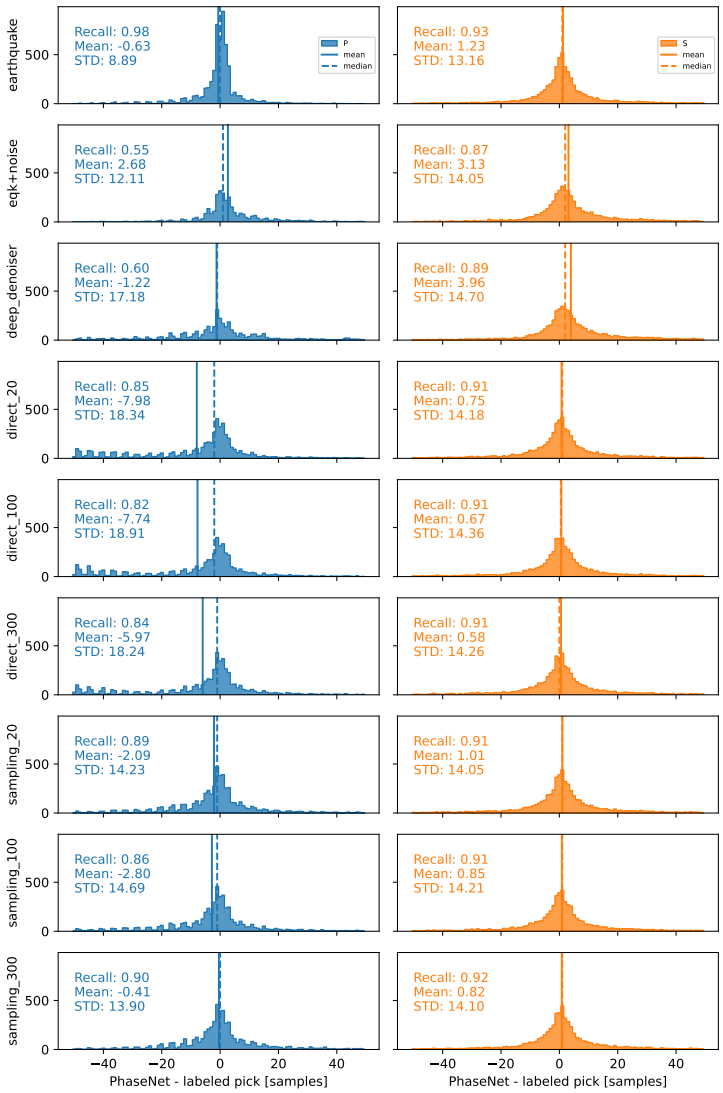

Figure 10. The histograms display the distributions of P-wave (blue) and S-wave (orange) arrival time differences between automated PhaseNet detections and label picks (in samples). The results obtained using the original seismograms, eqk+noise and DD are shown for comparison in the first, second and third row, respectively. The remaining rows show the results for different CDiffSD models applied to the same data subset, offering insights into the accuracy of wave arrival time detection by each model. Central tendency metrics, such as mean and median, are indicated in these histograms, highlighting any potential skewness in the distribution towards either early or late picks for both P and S waves.

The histograms in Fig. 10 provide a visual representation of the efficacy of different seismic signal denoising methods — "direct", "sampling", and DD — in retrieving a signal and preserve P- and S-wave onsets. The accuracy of automated P and S-wave arrival time picks by PhaseNet is compared to label picks. The histograms are organized by method and parameter variations, displaying the distribution of arrival time discrepancies measured in samples.

In the case of earthquake (i.e., no noise added, top histogram), the P-wave pick difference distribution exhibits spreads that are narrower than those of the S-wave and this is in full agreement with the expected behavior.

When noise is introduced, the pick difference distributions for P-waves and S-waves tend to converge towards a more similar pattern. This convergence can be attributed to the primary impact of noise on P-waves, owing to their lower amplitude compared to S-waves. As a result, the performance with added noise on P waves detection is much more degraded than on S waves detection with the same level of noise because P-waves have also smaller amplitudes. This observation is further supported by the recall values for S waves, which remain greater than 0.85 not only for all the denoising methods, but also for the noisy traces (earthquake + noise). In contrast, the recall rate for P-waves is consistently lowered by the presence of noise (Fig. 11). For these reasons we focus our analysis on P-wave picks.

As seen in Fig. 10 the distribution of the "direct" methods show pronounced negative skews, with mean values far from 0. This indicates a tendency of PhaseNet to pick P-waves slightly before the labeled picks for the waveforms denoised with "direct" methods. The reason of this behavior is most likely to be attributed to noise remaining in the denoised traces processed with the "direct method". This in turn can mislead PhaseNet to an early detection (see the "direct" example in Fig. 12). This tendency, however, is mitigated completely by the CDiffSD "sampling" method, as shown in Fig. 12. In particular, we see that the "sampling" methods display recall rates that are consistently high for both P and S, especially the 300 configuration, indicating a good denoising performance and the ability to recover the labeled phases.

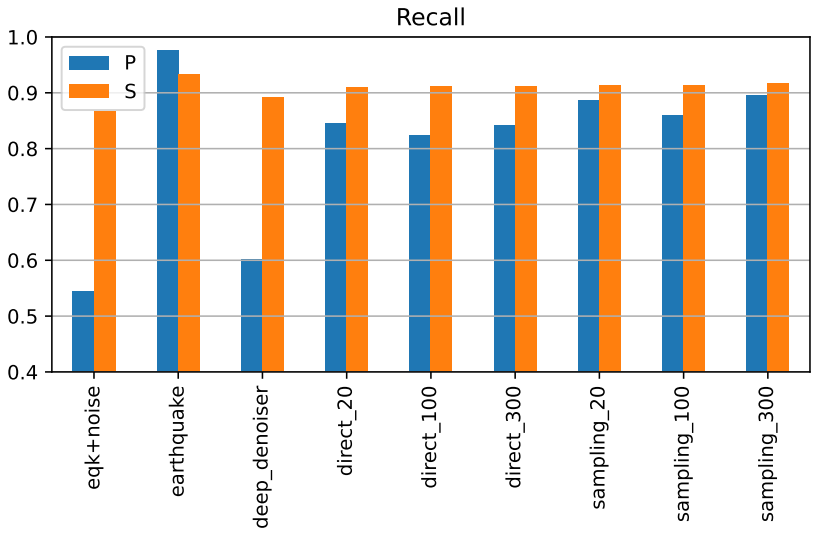

Figure 11. Comparison of recall rates for P and S waves between the different methods within a fixed window of 50 samples. S-waves recall rates are aligned almost for all models, indicating that the noise level is not enough to affect the S-waves because of the greater amplitude. P-waves recall rates instead show significant disparities between the DD approach and other methods, suggesting a lower performance of DD in preserving the P onset in these cases. The 'sampling-300' method is confirmed as the one with better performances.

From the comparison of the results obtained with the "direct", "sampling", and DD methods, it is evident that each method influences the automated pick accuracy differently. The "sampling" method, particularly at higher parameter settings, demonstrates a notable alignment with label picks, suggesting its superiority in mitigating noise and enhancing the precision of automated picking systems. It is also noteworthy that the recall values for P-waves shown in Fig. 11 are higher than DD for both "sampling" and "direct" methods, which suggests that in these cases DD does not preserve accurate P-wave onsets.

A comprehensive evaluation that considers all the proposed metrics in conjunction is essential for gaining a clear understanding of the various methods' performances. While examining each metric individually offers valuable insights into specific aspects of performance, a truly clear picture only emerges when we analyze these metrics together. For instance, SNR alone offers no insight into denoising quality. This metric turns out to be the least informative in our analysis, as evidenced by the lack of clear separation between methods. Conversely, CC assesses the similarity between original and denoised signals, providing a valuable but general measure of output quality. Here, CDiffSD's improvements are evident. Finally, the picking analysis tackles the crucial aspect of seismic wave onsets, focusing on the critical waveform portion where the noise-to-signal transition requires careful handling. Here the improvements of sampling versus direct and DD methods are clearly highlighted. A summary of the quantitative results discussed above is presented in Table 1.

Table 1. Summary of the metrics obtained with different denoising methods. Best score for each metric in bold.

Model	**SNR** (median)	**CC** (median)	**P picks diff** mean [samples]	**P picks diff** STD [samples]	**P picks recall**
DD	11.796	0.891	-1.22	17.18	0.60
CDiffSD direct20	13.476	0.949	-7.98	18.34	0.85
CDiffSD direct100	13.589	0.947	-7.74	18.91	0.82
CDiffSD direct300	13.900	**0.948**	-5.97	18.24	0.84
CDiffSD sampling20	13.298	0.943	-2.09	14.23	0.89
CDiffSD sampling100	13.216	0.941	-2.80	14.69	0.86
CDiffSD sampling300	**13.928**	0.943	**-0.41**	**13.90**	**0.90**

4.2 Qualitative Results

Qualitative factors are useful as side-by-side comparisons between the outcomes of different approaches. In the supplementary materials, we delve deeper into the analysis of seismic traces, examining the impact of denoising on picking. This includes a thorough examination of both the strengths and limitations of our model. We highlight instances where our model excels in denoising, as well as situations in which it does not perform optimally.

The examples below and those in the supplementary are organized with the same layout: in the top panel we compare the noisy signal (grey) with the denoised signal (black); in the middle panel we compare the original signal (green) with the denoised signal (black); the bottom panel is a zoom on the P-wave arrival.

4.2.1 Qualitative Picker Analysis: Direct Vs Sampling

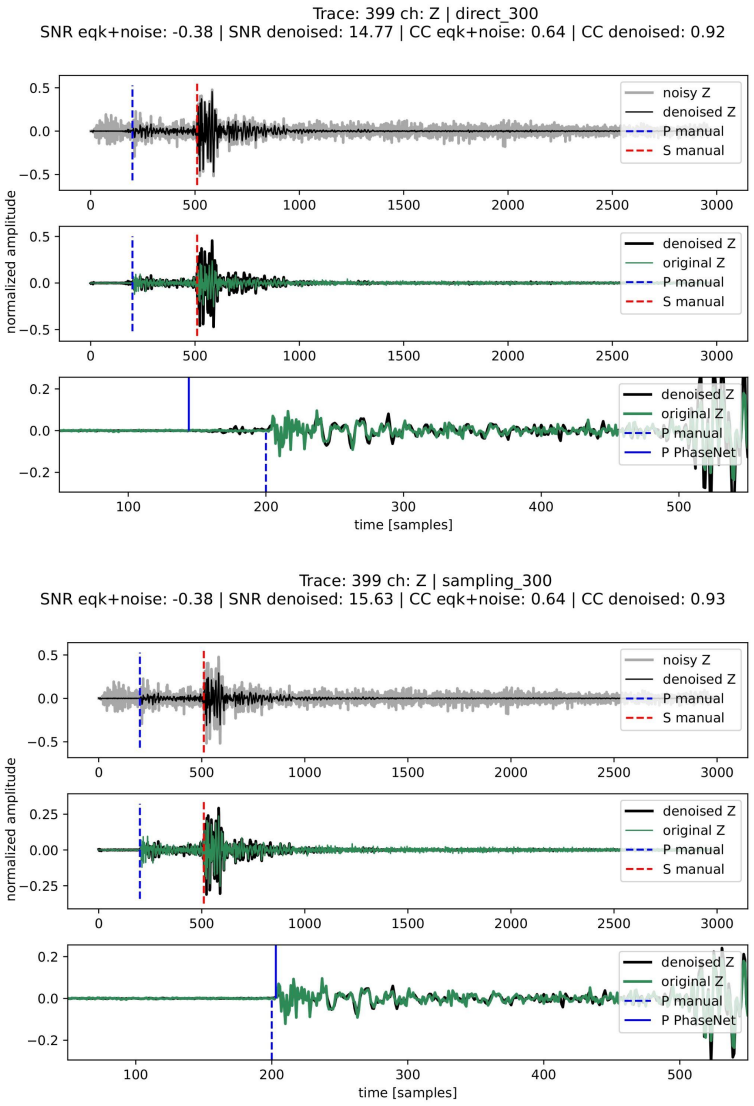

Figure 12. Comparison of a trace processed using 'direct_300' and 'sampling_300' methods. Notably, the 'direct_300' retains some of the noise preceding the P wave arrival, which is instead filtered out in the 'sampling_300' results. This noise before the P-wave retained with the the 'direct_300' method explains the tendency for this model to cause early picks (as seen in Fig. 10), as the residual noise can lead to earlier detections.

In the first example shown in Fig. 12 we compare the 'direct_300' and the 'sampling_300' methods. Here "sampling" method is found to be more effective than the "direct" method in denoising the seismic signal, and this is particularly evident from the middle and bottom panels, where the denoised signal in the "sampling" method match more closely the original signal. In contrast, the "direct" method shows more significant deviation from the original, especially before the P-wave arrival. This example is also

useful because it provides insight into the tendency of the "direct" methods to cause spurious early P-picks. The direct method, in fact, retains some pre-arrival noise, which can trigger an early pick in automatic approaches such as PhaseNet. This is less of an issue in the sampling method, as seen in the lower set of traces, where the denoised signal is cleaner, and the P-wave arrivals are closer to the labels. The implication for seismic processing is significant since the sampling method appears to produce cleaner signals and more accurate P-wave arrival times as a direct consequence. We note that this is crucial for various seismological applications such as earthquake location and tomographic imaging.

4.2.2 Qualitative Picker Analysis: DD Vs Sampling

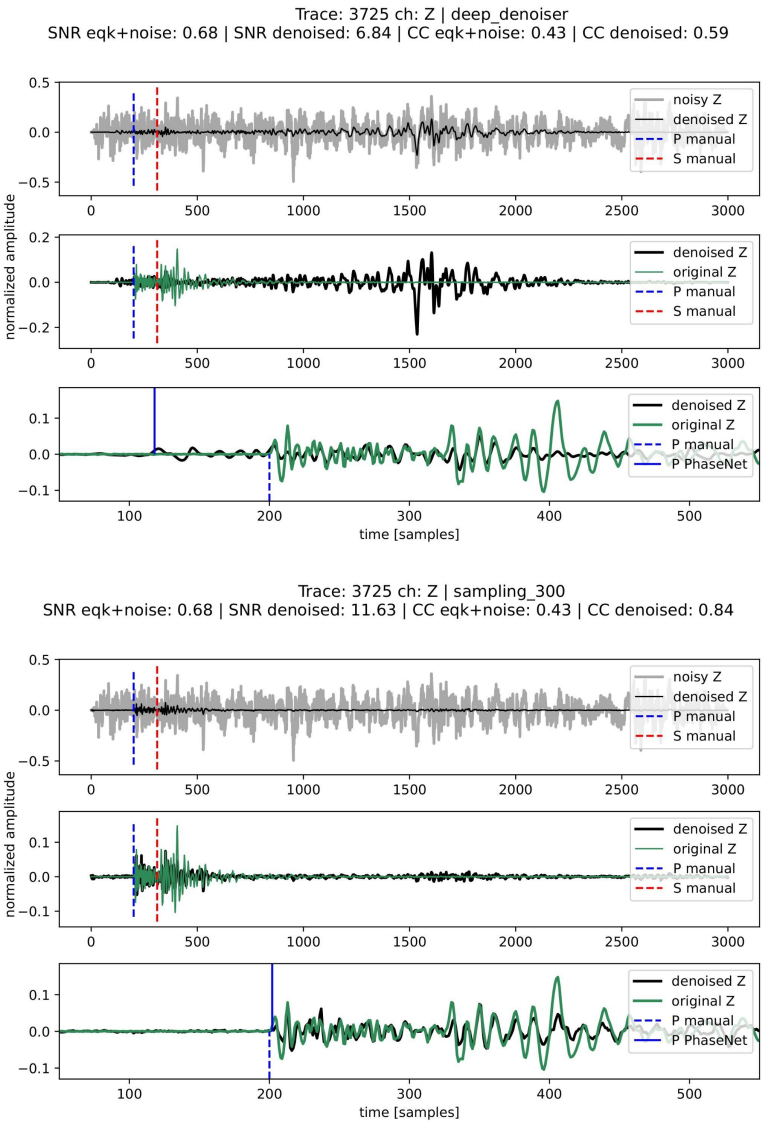

Figure 13. Comparison between a seismic trace processed with 'deep_denoiser' and 'sampling_300' methods. The 'sampling_300' method demonstrates a closer match to label phase picks and a more precise amplitude preservation, despite the substantial noise present in the original signal. DD also retains a high amplitude noise signal at around 1500 samples that 'sampling_300' manages to filter out almost completely.

Fig. 13 exemplifies the concepts previously discussed in Fig. 7, highlighting the performance of our model compared to that of the 'deep denoiser' in scenarios with very low Signal-to-Noise Ratio (SNR) before denoising. The figure demonstrates clearly how an extreme noise situation can lead to an error in phase picking for the 'deep denoiser', whereas the 'sampling' method is capable to reconstruct accurately the correct P wave arrival despite the presence of significant noise.

4.2.3 Qualitative Amplitude Analysis: Direct Vs Sampling

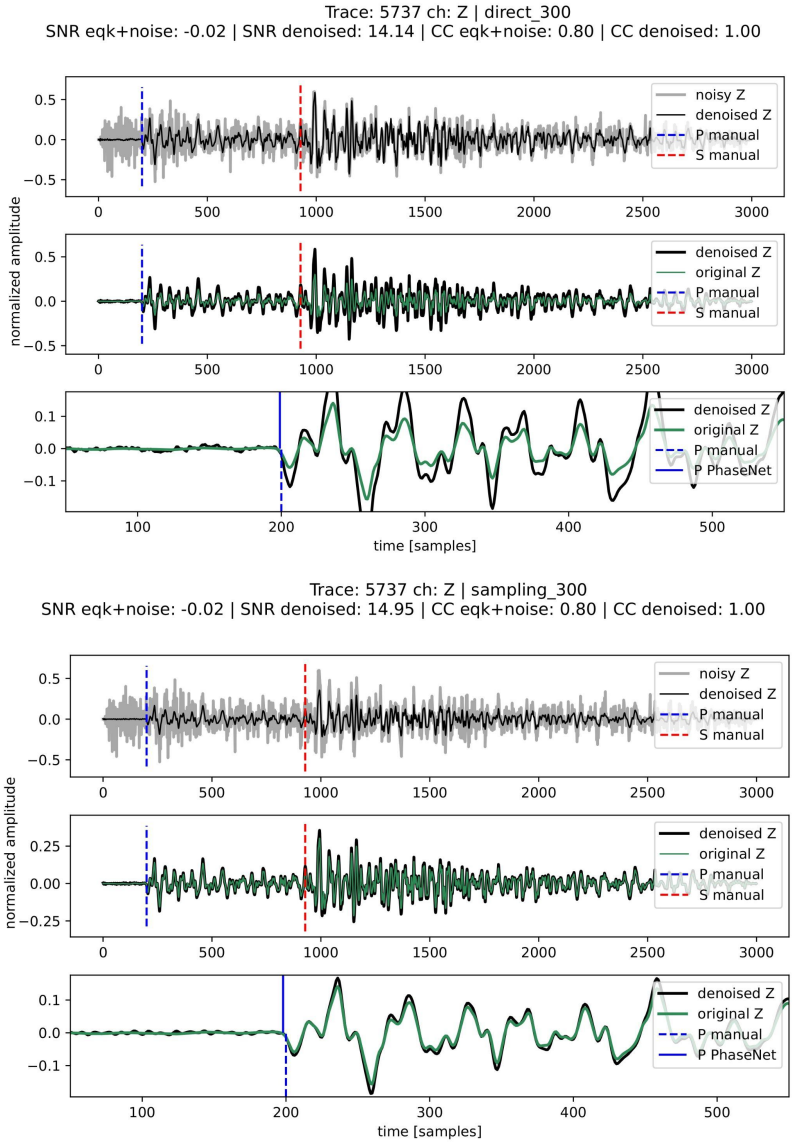

Figure 14. Comparison of a seismic trace processed with 'direct_300' and 'sampling_300' methods. It is particularly significant that the 'sampling_300' technique demonstrates an enhanced ability for amplitude reconstruction compared to the 'direct_300' method.

In Fig. 14 we show a comparative analysis of seismic signal denoising methods to investigate the importance of amplitude preservation. The Cold Diffusion Model employing a sampling strategy ('sampling_300') demonstrates a superior performance in maintaining the amplitudes of the seismic signal. In practice, the denoised signal aligns more accurately with the original waveform, preserving the integrity of the amplitude across the signal's duration. This is particularly evident in the detailed zoomed-in analysis, where the 'sampling_300' method displays remarkable congruence with the original signal, as evidenced by the minimal and consistent residuals. In contrast, the direct application

of a U-Net model ('direct_300') displays a slight but discernible attenuation in amplitude, most noticeable in segments with higher amplitude peaks. The increased residuals associated with the 'direct_300' method suggest a more significant alteration of the signal after the denoising process. Therefore, the Cold Diffusion Model with sampling stands out as the most effective method for seismic data denoising (amongst those tested here), especially where the preservation of amplitude is critically important.

5 Model assessment: Assessing the Impact of Exclusive Noise Input

In this section we aim to test the behaviour of the model in no-earthquake scenarios, i.e. with inputs containing only noise. This is done in order to verify whether the model doesn't generates any artifacts in the absence of signal generating false earthquakes.

Cold Diffusion is based on the model's ability to learn the broad data distribution during training, which generally includes a variety of seismic traces with different levels of noise. Therefore, the model should be able to generalize and identify traces that are entirely dominated by noise, even without direct exposure to specific types of earthquake samples where there is no earthquake signal. Based on these assumptions, we seek to verify if our results align with the theoretical expectations.

We have used the entire noise test set as input, without combining it with the earthquake data. Theoretically, with a perfect denoising, the expected output would be a trace composed exclusively of zeros, in the real context the trace should approach zero.

We applied the model without retraining, meaning the model's weights have never been exposed to the absence of earthquake traces as ground truth. To assess the correctness of the output we set an amplitude threshold between ± 0.02 to decide whether the output could resemble a trace of zeros. The direct and sampling methods have correctly reconstructed the expected signal in 60.3% and 88.6% of cases, respectively. This different performance highlights the sampling method's superior capability in recognizing the absence of earthquake signals and adapting to it.

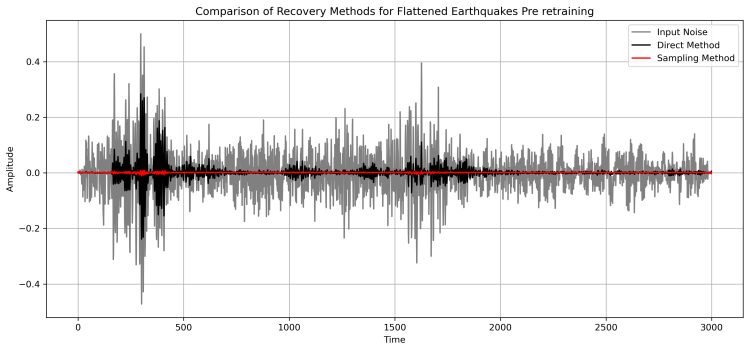

Figure 15. Example of the outputs of direct (in black) and sampling (in red) methods in case of a noise only input (in grey). No retraining is performed here, i.e. the models have never been exposed to zero-traces as ground truth for noise-only input. The direct method fails in recovering a zero-trace since it introduces artificial signals. In contrast, the sampling method reconstructs successfully an output that resembles a zero-trace.

Given the promising results just described, we further explored this scenario by retraining the model including no-signal traces as ground truth. We focused only on a sin-

gle channel for this test and incorporated 3% of the entire training set with zeroed traces to represent the absence of seismic events. The results align with our expectations, indicating an improvement in performance in the presence of noise alone. Specifically, the cases where zero traces are retrieved increases to 68.2% and 90.5% for direct and sampling methods, respectively. The direct method exhibits a more substantial improvement, starting from a lower baseline performance, whereas the sampling method shows a smaller increase, likely due to its performance already approaching saturation.

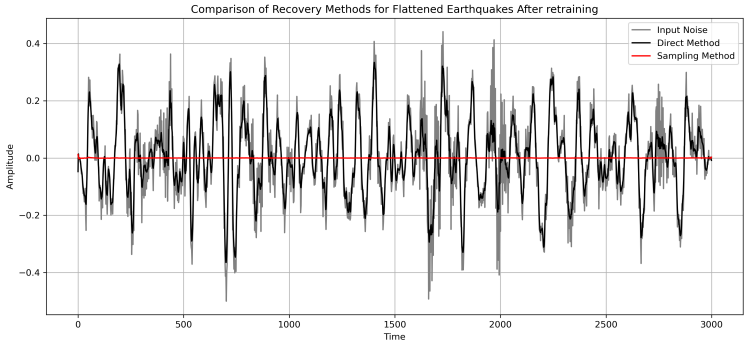

Figure 16. Example of the comparison between the sampling method (in red) and the direct method (in black), the input (in gray) for both methods is only noise. In this case the models have been retrained with zero-traces as ground truth for noise-only traces. The sampling method succeeds in reconstructing a zero-trace. On the other hand, the direct method outputs noise, indicating a less accurate reconstruction in this scenario.

Regarding the results post-retraining, it should be noted that the output trace of the sampling method shown in Figure 16 is indeed close to the expected zero-trace. On the contrary, low amplitude noise was still present in the output of the non retrained-case shown in Figure 15. This highlights the importance of including flat traces during the training.
In this evaluation of the CDiffSD on these cases comprised solely of noise, we proved that it is not imperative to include such examples in training to accurately discern between noise and genuine seismic signals. However, including these kind of signals in training, improves the capability of effectively identifying traces that are comprised solely of noise.

6 Conclusion

Our study demonstrates promising results and affirms the validity of cold diffusion denoising for seismological applications. We employed three key metrics to quantify the enhancement brought by the CDiffSD model. Specifically, the CDiffSD model showcased a substantial improvement over DD in denoising seismic traces (see Table 1), enhancing the SNR by approximately 18%. Furthermore, we observed a 5.84% increase in Cross-Correlation, indicating a higher congruence between the denoised signals and the original ones. Finally, as a third metric, our approach significantly enhanced the accuracy of seismic event detection, achieving a 50% improvement in recall for P-wave picks, indicating a much better preservation of the P-onset even for noisy seismic data. Regarding the evaluation of different CDiffSD versions, it is important to highlight that, despite SNR and CC metrics aligning between the "direct" and "sampling" configurations, the "sampling" systematically demonstrates its superiority in applied contexts, such as P-

phase picking. Focusing on "direct" versus "sampling" we observe significant enhancements: e.g. comparing both configurations with $T = 300$, the average difference from labeled picks reduces significantly from -5.97 to -0.41 samples. Similarly, the standard deviation drops from 18.2 to 13.9. This trend of improved performance holds true across all levels of T. Moreover, "sampling" yields a notable 6.7% increase in P picks recall compared to the "direct" method. However, it is noteworthy that the computation time increases with an increase in T. A more detailed study is reported in the supplementary materials. Therefore, the size of T should be considered when using these models in real-time scenarios such as in a seismic monitoring room. We emphasize the importance of looking at the results as a whole. That is, while SNR and CC are important metrics for assessing, respectively, the raw denoising power and the quality of the reconstructed signal, in fact, the preservation of the integrity of the P- and S-wave arrivals is of critical importance for a reliable denoising technique. Among the models evaluated in Section 4, the one utilizing "sampling" with T=300 emerged as the most effective according to the three combined metrics. The model's fidelity in preserving seismic trace characteristics, especially at the signal-to-noise transition, highlights its practical advantages in real-world seismological applications.

We note however, that while our results provide an important advance, they should be regarded as a preliminary step towards addressing a broader spectrum of open questions and potential model enhancements. A significant direction for future advancement lies in applying these techniques to broader datasets. Our initial explorations aimed to establish the feasibility of these methods.

Moving forward we could potentially develop a more generalized model by retraining on large datasets such as INSTANCE (Michelini et al., 2021) and STEAD, which encompass several million traces compared to the $\sim$ 40k traces used in this study. The use of larger datasets would allow treatment of noise in a wide range of seismological contexts without the need for further retraining, thus significantly boosting model applicability and robustness across diverse seismic scenarios.

Our model exhibits significant potential for cleaning and enhancing seismic traces. Moreover, it holds promise for recovering earthquakes hidden by noise that may have eluded both human and automatic detection. Such capability could contribute to expanding seismic catalogs. While further refinements are conceivable, this method, which is borrowed from speech enhancement tasks, has proven its validity in the intricate domain of seismological analysis. This cross-disciplinary innovation underscores the model's versatility and suggests broader applicability in extracting and analyzing subtle seismic signals.

Acronyms

AttDD Attention Deep Denoiser
CC Cross Correlation
CDiffSD Cold Diffusion Model for seismic denoising
DAS Distributed Acoustic Sensing
DD Deep Denoiser
DL Deep Learning
DM Diffusion Model
DPRNN Dual-Path Recurrent Neural Network
E East-West
eqk Earthquake
ERC European Research Council
GAN Generative Adversarial Network
GELU Gaussian Error Linear Unit

ICA Independent Component Analysis
INGV Istituto Nazionale di Geofisica e Vulcanologia
INSTANCE Italian Seismic Dataset For Machine Learning
MUSIC MUltiple SIgnal Classification
N North-South
NRF Noise Reduce Factor
ResNet Residual Neural Network
SNR Signal to Noise Ratio
STEAD STanford EArthquake Dataset
STFT Short-Time Fourier Transform
VAE Variational Autoencoder
Z Vertical

Open Research Section

The STEAD dataset (Mousavi et al., 2019) (Seismological Tools for Earthquake Analysis and Detection) is openly accessible. For data manipulation, ObsPy, a Python library for processing seismological data, can also be used (for more information on ObsPy, see its documentation (Beyreuther et al., 2010)).

To replicate the data accurately, it is necessary to apply the filters described in Section 3 to chunk2 of the STEAD dataset. Furthermore, specific data related to this research are available in Zenodo with the identifier (Trappolini, 2024b). For additional details, see GitHub repository (Trappolini, 2024a).

Acknowledgments

We would like to thank the editor Yangkang Chen, the reviewer Martijn Van Den Ende and two anonymous reviewers for their helpful and detailed reviews that greatly enhanced the quality of our paper. This study was supported by MUR PNRR FAIR project (PE00000013), the INGV Pianeta Dinamico 2021 Tema 8 SOME project (CUP D53J1900017001) funded by the Italian Ministry of University and Research and by the European Research Council (ERC) under grant 835012 (TECTONIC).

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining.*

Bansal, A., Borgnia, E., Chu, H.-M., Li, J. S., Kazemi, H., Huang, F., ... Goldstein, T. (2022). Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*.

Bear, L. K., Pavlis, G. L., & Bokelmann, G. H. (1999). Multi-wavelet analysis of three-component seismic arrays: application to measure effective anisotropy at pinon flats, california. *Bulletin of the Seismological Society of America*, *89*(3), 693–705.

Bekara, M., & der Baan, M. V. (2009). Random and coherent noise attenuation by empirical mode decomposition. *Geophysics, vol. 74, no. 5, pp. V89–V98, 2009*.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., & Wassermann, J. (2010). Obspy: A python toolbox for seismology. *Seismological Research Letters*, *81*(3), 530–533.

Bie, X., Leglaive, S., Alameda-Pineda, X., & Girin, L. (2022). Unsupervised speech

enhancement using dynamical variational autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2993–3007.

Boué, P., Roux, P., Campillo, M., & de Cacqueray, B. (2013). Double beamforming processing in a seismic prospecting context. *Geophysics*, *78*(3), V101–V108.

Brooks, L. A., Townend, J., Gerstoft, P., Bannister, S., & Carter, L. (2009). Fundamental and higher-mode rayleigh wave characteristics of ambient seismic noise in new zealand. *Geophysical Research Letters*, *36*(23).

Cabras, G., Carniel, R., & Wasserman, J. (2010). Signal enhancement with generalized ica applied to mt. etna volcano, italy. *Bollettino di Geofisica Teorica ed Applicata*, *51*(1).

Cao, R., Abdulatif, S., & Yang, B. (2022). Cmgan: Conformer-based metric gan for speech enhancement. *arXiv preprint arXiv:2203.15149*.

Cao, S., & Chen, X. (2005). The second-generation wavelet transform and its application in denoising of seismic data. *Applied geophysics*, *2*, 70–74.

Chen, Y., & Ma, J. (2014). Random noise attenuation by f-x empiricalmode decomposition predictive filtering. *Geophysics, vol. 79, no. 3, pp. V81–V91, 2014*.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, *36*(3), 287–314.

Donahue, C., Li, B., & Prabhavalkar, R. (2018). Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5024–5028).

Durall, R., Ghanim, A., Fernandez, M. R., Ettrich, N., & Keuper, J. (2023). Deep diffusion models for seismic processing. *Computers & Geosciences*, *177*, 105377.

Fang, H., Carbajal, G., Wermter, S., & Gerkmann, T. (2021). Variational autoencoder for speech enhancement with a noise-aware encoder. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 676–680).

Gaci, S. (2014). The use of wavelet-based denoising techniques to enhance the first-arrival picking on seismic traces. *IEEE Trans. Geosci. Remote Sens., vol. 52, no. 8, pp. 4558–4563, Aug. 2014*.

Gibbons, S. J., Ringdal, F., & Kværna, T. (2008). Detection and characterization of seismic phases using continuous spectral estimation on incoherent and partially coherent arrays. *Geophysical Journal International*, *172*(1), 405–421.

Han, J., & van der Baan, M. (2015). Microseismic and seismic denoising via ensemble empirical mode decomposition and adaptive thresholding. *Geophysics, vol. 80, no. 6, pp. KS69–KS80, . doi: 10. 1190/geo2014-0423.1*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hennenfent, G., & Herrmann, F. J. (2006). Seismic denoising with nonuniformly sampled curvelets. *Comput. Sci. Eng., vol. 8, no. 3, p. 16, May 2006*.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Kim, H. Y., Yoon, J. W., Cheon, S. J., Kang, W. H., & Kim, N. S. (2021). A multi-resolution approach to gan-based speech enhancement. *Applied Sciences*, *11*(2), 721.

Leglaive, S., Alameda-Pineda, X., Girin, L., & Horaud, R. (2020). A recurrent variational autoencoder for speech enhancement. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 371–375).

Leglaive, S., Girin, L., & Horaud, R. (2018). A variance modeling framework based

on variational autoencoders for speech enhancement. In *2018 ieee 28th international workshop on machine learning for signal processing (mlsp)* (pp. 1–6).

Lim, J. S. (1990). *Two-dimensional signal and image processing*. Prentice-Hall, Inc.

Liu, W., Cao, S., & Chen, Y. (2016). Seismic time-frequency analysis via empirical wavelet transform. *IEEE Geosci. Remote Sens. Lett., vol. 13, no. 1, pp. 28–32, Jan. 2016*.

Liu, Y., Li, Y., Lin, H., & Ma, H. (2013). An amplitude-preserved time–frequency peak filtering based on empirical mode decomposition for seismic random noise reduction. *IEEE Geoscience and Remote Sensing Letters*, *11*(5), 896–900.

Lu, Y.-J., Wang, Z.-Q., Watanabe, S., Richard, A., Yu, C., & Tsao, Y. (2022). Conditional diffusion probabilistic model for speech enhancement. In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7402–7406).

Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., & Lauciani, V. (2021). Instance–the italian seismic dataset for machine learning. *Earth System Science Data*, *13*(12), 5509–5544.

Moni, A., Bean, C. J., Lokmer, I., & Rickard, S. (2012). Source separation on seismic data: Application in a geophysical setting. *IEEE Signal Processing Magazine*, *29*(3), 16–28.

Mousavi, S. M., & Langston, C. A. (2016). Adaptive noise estimation and suppression for improving microseismic event detection. *Appl. Geophys., vol. 132, pp. 116–124, Sep. 2016. doi: 10.1016/j.jappgeo.2016.06.008*.

Mousavi, S. M., & Langston, C. A. (2017). Automatic noise-removal/signalremoval based on general cross-validation thresholding in synchrosqueezed domain and its application on earthquake data. *Geophysics, vol. 82, no. 4, pp. V211–V227, 2017. doi: 10.1190/geo20160433.1*.

Mousavi, S. M., Sheng, Y., Zhu, W., & Beroza, G. C. (2019). Stanford earthquake dataset (stead): A global data set of seismic signals for ai [dataset]. *IEEE Access*, *7*, 179464–179476.

Neelamani, R., Baumstein, A. I., Gillard, D. G., Hadidi, M. T., & Soroka, W. L. (2008). Coherent and random noise attenuation using the curvelet transform. *The Leading Edge*, *27*(2), 240–248.

Novoselov, A., Balazs, P., & Bokelmann, G. (2022). Sedenoss: Separating and denoising seismic signals with dual-path recurrent neural network architecture. *Journal of Geophysical Research: Solid Earth*, *127*(3), e2021JB023183.

Pascual, S., Bonafonte, A., & Serr'a, J. (2017). Segan: Speech enhancement generative adversarial network. *in Proc. Interspeech, 2017*.

Richter, J., Welker, S., Lemercier, J.-M., Lay, B., & Gerkmann, T. (2023). Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–miccai 2015: 18th international conference, munich, germany, october 5-9, 2015, proceedings, part iii 18* (pp. 234–241).

Schmidt, R. (1986). Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, *34*(3), 276–280.

Shan, H., Ma, J., & Yang, H. (2009). Comparisons of wavelets, contourlets and curvelets in seismic denoising. *Appl. Geophys., vol. 69, no. 2, pp. 103–115, Oct. 2009. doi: 10.1016/j.jappgeo.2009.08.002*.

Siyuan, C., & Xiangpeng, C. (2005). The second-generation wavelet transform and its application in denoising of seismic data. *Appl. Geophys.,vol. 2, no. 2, pp. 70–74, Jun. 2005*.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (pp. 2256–2265).

Tang, G., & Ma, J. (2011). Application of total-variation-based curvelet shrinkage for three-dimensional seismic data denoising. *IEEE Geosci. Remote Sens. Lett., vol. 8, no. 1, pp. 103–107, Jan. 2011*.

Trappolini, D. (2024a). *Diffusion model for earthquake.* GitHub. Retrieved from `https://github.com/Daniele-Trappolini/Diffusion-Model-for-Earthquake`

Trappolini, D. (2024b). *Stead subsample 4 cdiffsd.* Zenodo. Retrieved from `https://zenodo.org/record/10972601` doi: 10.5281/zenodo.10972601

Tselentis, G.-A., Martakis, N., Paraskevopoulos, P., Lois, A., & Sokos, E. (2012). Strategy for automated analysis of passive microseismic data based on s-transform, otsu's thresholding, and higher order statistics. *Geophysics*, *77*(6), KS43–KS54.

van den Ende, M., Lior, I., Ampuero, J.-P., Sladen, A., Ferrari, A., & Richard, C. (2021). A self-supervised deep learning approach for blind denoising and waveform coherence enhancement in distributed acoustic sensing data. *IEEE Transactions on Neural Networks and Learning Systems*.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... others (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, *17*(3), 261–272.

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., ... Soto, H. (2022). Seisbench—a toolbox for machine learning in seismology. *Seismological Research Letters*, *93*(3), 1695–1709.

Yen, H., Germain, F. G., Wichern, G., & Le Roux, J. (2023). Cold diffusion for speech enhancement. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Zhang, R., & Ulrych, T. J. (2003). Physical wavelet frame denoising. *Geophysics*, *68*(1), 225–231.

Zhu, W., & Beroza, G. C. (2019). Phasenet: A deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, *216*(1), 261–273.

Zhu, W., Mousavi, S. M., & Beroza, G. C. (2019). Seismic signal denoising and decomposition using deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(11), 9476–9488.

Appendix

A.Diffusion Model

In this appendix, we will briefly explain how diffusion models work. Diffusion models are inspired by the physical process of diffusion, where particles spread out from areas of higher concentration to areas of lower concentration over time (Sohl-Dickstein et al., 2015). In the context of generative modeling, this process is simulated in a reverse manner. The model starts with a distribution of random noise and gradually refines this noise into a coherent sample from the target distribution over a series of steps. The theoretical foundation of diffusion models is rooted in stochastic differential equations (SDEs) and involves two key phases: the **forward diffusion** (or noise addition) process and the **reverse diffusion** (or denoising) process.

Forward diffusion process: In this phase, the model incrementally adds noise to data from the original distribution over a series of steps, transforming it into a distribution of pure noise. Mathematically, this can be represented as a Markov chain that gradually transitions the data distribution $p(x_0)$ to a noise distribution $p(x_T)$, where T is the total number of diffusion steps and x_0 to x_T represents the data at each step of the forward diffusion process. The step sizes are controlled by a variance schedule $\{\beta_t \in (0,1)_{t=1}^{T}\}$.

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{4}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{5}$$

A nice property of the process above is that we can reparametrizes x_t in terms of ϵ (i.e. the added gaussian noise), which is independent of the model parameters, allowing the gradient of the loss function to be backpropagated through the deterministic part of the model. We can perform this using the reparametrization trick. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha_t} = \prod_{i=1}^{t} \alpha_i$:

$$\begin{aligned} x_t &= \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} \\ &= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\epsilon}_{t-2} \\ &= ... \\ &= \sqrt{\bar{\alpha_t}} x_0 + \sqrt{1-\bar{\alpha_t}}\epsilon \\ q(x_t|x_0) &= N(x_t; \sqrt{\bar{\alpha_t}} x_0, (1-\bar{\alpha}_t)I) \end{aligned} \tag{6}$$

Reverse diffusion process: The reverse process involves learning to denoise the data, starting from the noise distribution and progressively reconstructing the data distribution through a series of learned denoising steps. The goal of the model during this phase is to learn the conditional distribution $p(x_{t-1}|x_t)$.

$$p_\theta(x_{0:T}) = p(x_T) \prod p_\theta(x_{t-1}|x_t) \tag{7}$$

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{8}$$

For a more detailed discussion, please refer to the paper: (Ho et al., 2020).

Training: Training diffusion models involves optimizing the parameters of the reverse diffusion process to minimize the difference between the original data distribution and the distribution of the generated samples. This is typically achieved through variational inference, where the model learns to predict the noise that was added at each step of the forward process, thereby allowing it to reverse the diffusion.

Algorithm 2 Diffusion Model Training

repeat
  $x_0 \sim q(x_0)$
  $t \sim$ Uniform$(\{1, ..., T\})$
  $\epsilon \sim N(0, I)$
  Take gradient descent step on $\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha_t}} x_0 + \sqrt{1-\bar{\alpha_t}}\epsilon, t)\|^2$
until converged

The training of diffusion models begins with the application of the forward diffusion process to the training data, which generates noisy versions of the data at various timesteps, sampled using the uniform distribution in 2. Following this initial step, the model begins the phase of noise prediction for each noisy sample. It attempts to accurately predict the specific noise that was added at each timestep. The accuracy of this prediction is measured against the actual noise used during the forward process, utilizing typically the mean squared error (MSE) as the loss function. Once the loss has been calculated, it is then backpropagated through the model to update its parameters. It is during this phase that the reparametrization trick plays an important role, as it allows for the gradients to flow through the stochastic sampling of noise, thus enabling the optimization process to proceed (letting the ϵ parameter to be a trainable parameter).

Sampling: After the model has been trained to reverse diffusion process. It can generate new samples starting from noise, or denoise new inputs. This process is the inverse of the forward diffusion process, where noise is gradually added to the data. Instead, starting with a purely noisy distribution, the model iteratively generates/denoise data that increasingly resembles the target distribution. The algorithm 3 give a closer look at how the sampling process unfolds. The sampling process begins with an initial

Algorithm 3 Diffusion Model Sampling

$x_T \sim N(0, I)$
for $t = T, ..., 1$ **do**
  $z \sim N(0, I)$ if $t > 1$, else $z = 0$
  $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)) + \sigma_t z$
end for
return x_0

noise vector sampled from a Gaussian distribution. This noise vector x_T represents the final state of the forward diffusion process and serves as the starting point for generation. From this initial state, the model iteratively applies the learned reverse diffusion steps to reduce the noise and move closer to the data distribution. At each step t, the model uses its parameters to estimate the cleaner version of the current state x_{t-1} from x_t. This is based on the conditional probability learned during training, which models how to reverse the noise addition for that particular step. After the final reverse diffusion step, the output is a sample that closely resembles/denoises the target data distribution. This sample is the model's "best guess" at a real data point, having transformed from pure noise (for generation) or a noised input (for denoising) to structured data through the reverse diffusion process. The sampling process in diffusion models exemplifies how structured data can emerge from randomness (for generation) or from noisesness (for denoising) through iterative refinement, helped by the complex statistical relationships learned during training.

B.Cold Diffusion Model

Cold diffusion models are very recent designs, and currently, there are limited works implementing such architecture. The original concept of diffusion models (Bansal et al., 2022) involves extending and generalizing degradation using non-Gaussian noise. This becomes achievable due to enhancements in the sampling algorithm. In particular, as stated in (Bansal et al., 2022), they start from a simple assumption: the original sampling: Algorithm 3 works well when the restoration operator is perfect. This means that:

$$R(D(x_0, t), t) = x_0 \text{ for all } t. \tag{9}$$

With restoration operator is p_θ in Eq. 8 that here is reffered as R. However, in the scenario where the restoration is imperfect, this causes the model to make errors, leading it to deviate from $D(x_0, s)$, D stands for Degradation operator hence: $q(x_{1:T}|x_0)$ in Eq. 4. The implemented sampler possesses excellent mathematical capabilities that are not detailed in this work (for further details, refer to 3.3 Properties of the Algorithm in (Bansal et al., 2022)), enabling the accurate reconstruction of the signal even in cases where the restoration operator R fails to completely invert D.
As a starting point to address our task, we have taken animorphosis as a reference (for further details, see Section 5.3 "Generation using other transformations" (Bansal et al., 2022)). In this context, a "clean" sample (an image of a person) is systematically subjected to a series of transformations resulting in an out-of-domain "degraded" sample (an image of an animal). However, it's important to highlight that our approach deviates from this process. Our degraded sample retains the underlying information of the

Algorithm 4 Improved Sampling for Cold Diffusion

Input: A degraded sample x_t
for $s = t, t-1, ...1$ **do**
  $\hat{x}_0 \leftarrow R(x_s, s)$
  $\hat{x}_{s-1} = x_s - D(\hat{x}_0, s) + D(\hat{x}_0, s-1)$
end for

clean sample, as our degradation process now introduces an out-of-domain sample in conjunction with the clean sample. Such a degradation does not correspond to any of those addressed in (Bansal et al., 2022).